# Development of shRNA screens to identify effectors of three complex traits: neighbour suppression of tumour growth and proliferation and protection from lipotoxicity in β-cells.

Submitted by Lorena Boquete Vilariño to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies in August 2016.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

***Abstract***

RNA interference (RNAi) is a natural mechanism of cellular defence against exogenous double stranded RNA (dsRNA). The discovery of small dsRNA molecules which can be processed by the RNAi pathway in mammalian cells was one of the key advances in the study of functional genomics. These molecules can be designed to downregulate the expression of specific genes. Collections or libraries of dsRNA molecules targeting an extensive number of genes are now available. Using these libraries, numerous studies have implemented high-throughput screens for the study of molecular effectors of numerous phenotypes.

The process of designing an RNAi screen requires the consideration of several critical factors during both the experimental and analysis phases. The experimental screen should aim to reproduce the biological phenomenon studied as closely as possible by choosing an adequate model and screening conditions. Phenotype evaluation and assessment of knockdown effects need careful consideration. The results obtained from large-scale RNAi screens are often complex. An analysis pipeline should be implemented which integrates the biological basis of the phenomenon and facilitates the interpretation of the data.

This project designed and implemented an unbiased shRNA screen in two *in vitro* models relevant to carcinogenesis and diabetes. The first screen implemented used a model of neighbour suppression to study the molecular effectors of the response in tumorigenic cells to growth suppression cues from the surrounding tissue, a cellular interaction relevant in early tumorigenesis. The second screen studied two phenotypes relevant to diabetes: proliferation and resistance to lipotoxicity of β-cells in a reversibly immortalised β cell line. An integrative analysis pipeline was also developed to apply network biology and functional enrichment analysis methods for the interpretation of the data obtained from both screens.

*List of Contents*

**ABSTRACT** ..... **3**

**LIST OF CONTENTS** ..... **5**

**ACKNOWLEDGMENTS** ..... **13**

**ABBREVIATIONS** ..... **14**

**LIST OF FIGURES** ..... **16**

**LIST OF TABLES** ..... **23**

**CHAPTER 1. INTRODUCTION** ..... **31**

1.1. FUNCTIONAL GENOMICS ..... 31

1.2. FORWARD FUNCTIONAL GENOMICS SCREENS ..... 32

1.3. REVERSE FUNCTIONAL GENOMICS SCREENS ..... 33

1.4. GAIN-OF-FUNCTION STUDIES ..... 34

1.5. LOSS-OF-FUNCTION STUDIES ..... 34

*1.5.1. Homologous recombination* ..... *34*

*1.5.2. Genome editing* ..... *35*

1.5.2.1. Zinc finger nucleases ..... 35

1.5.2.2. Meganucleases ..... 36

1.5.2.3. TALENS ..... 36

1.5.2.4. CRISPR/Cas9 ..... 36

*1.5.3. RNA interference* ..... *37*

1.5.3.1. siRNA ..... 38

1.5.3.2. shRNA ..... 38

1.5.3.3. RNAi high throughput screens (HTS) ..... 39

1.5.3.4. Arrayed screens ..... 39

1.5.3.5. Pooled screens ..... 40

1.6. SUMMARY ..... 42

**CHAPTER 2. MATERIAL AND METHODS** ..... **45**

2.1 CELL CULTURE ..... 45

*2.1.1 List of materials* ..... *45*

*2.1.2 Cell lines* ..... *47*

*2.1.3 General tissue culture procedures* ..... *47*

*2.1.4 Sub-culturing cells* ..... *48*

*2.1.5 Freezing cells ..... 48*
*2.1.6 Thawing cells ..... 49*
*2.1.7 Counting cells ..... 49*
*2.1.8 Puromycin kill curve ..... 51*
*2.1.9 Puromycin selection of transduced cells. ..... 51*
*2.1.10 Harvesting cells for DNA extraction ..... 51*
2.2 MOLECULAR BIOLOGY ..... 52
*2.2.1. List of materials ..... 52*
*2.2.2. Phenol:chloroform extraction of DNA from cell lines ..... 53*
*2.2.3. Quantification of nucleic acids ..... 53*
*2.2.4. PCR amplification of DNA ..... 53*
*2.2.5. Preparation of agarose gels ..... 55*
*2.2.6. Agarose gel electrophoresis of DNA ..... 55*
*2.2.7. DNA extraction from agarose gel ..... 55*
*2.2.8. Sequencing ..... 56*
2.3 CLONING ..... 57
*2.3.1 List of materials ..... 57*
*2.3.2 Bacterial growing medium ..... 57*
*2.3.3 LB agar plates ..... 57*
*2.3.4 Preparation of competent bacteria ..... 57*
*2.3.5 Heat-shock transformations ..... 58*
*2.3.6 Plasmid isolations ..... 58*
2.4 NKI MOUSE SHRNA LIBRARY ..... 59
2.5 RE-ANNOTATION OF THE NKI SHRNA LIBRARY ..... 61
2.6 PREPARATION OF PLASMIDS WITH CUSTOM SHRNA CONSTRUCTS ..... 61
2.7 ANALYSIS ..... 65
*2.7.1 Building gene networks ..... 65*
*2.7.2 Analysis of gene interactions ..... 65*
*2.7.3 Enrichment in gene interactors ..... 65*

**CHAPTER 3. A POOLED GENOME-WIDE SHRNA SCREEN OF NEIGHBOUR SUPPRESSION EFFECTORS IN TUMORIGENIC CELLS ..... 67**

3.1. INTRODUCTION ..... 67
*3.1.1. Cancer ..... 67*
*3.1.2. Mechanisms of resistance to carcinogenesis ..... 67*

*3.1.3. Neighbour suppression* ... *68*
3.2. MATERIAL AND METHODS ... 71
*3.2.1. Cell lines* ... *71*
*3.2.2. Production of retroviral particles using Phoenix cells* ... *71*
*3.2.3. Transduction of SVE-6 cells* ... *72*
*3.2.4. Seeding monolayers* ... *72*
*3.2.5. Plating SVE-6 cells on monolayers and plastic* ... *72*
*3.2.6. Restriction enzyme digestion of genomic DNA* ... *72*
*3.2.7. PCR amplification and purification of shRNA constructs* ... *75*
*3.2.8. Analysis of sequencing data* ... *78*
*3.2.9. Recursive partitioning analysis of tumour expression data* ... *80*
*3.2.10. Gene set enrichment analysis* ... *82*
*3.2.11. Design of the screening process* ... *82*
3.3. RESULTS ... 84
*3.3.1. Description of the screening process* ... *84*
*3.3.2. Overview of the NKI shRNA library used* ... *84*
*3.3.3. Optimization of multiplicity of infection (MOI)* ... *85*
*3.3.4. Optimization of co-cultures* ... *88*
*3.3.5. Optimization of shRNA construct recovery from SVE-6 cells* ... *92*
*3.3.6. Reproducibility of the screening process* ... *103*
*3.3.7. Running the genome-wide shRNA screening process* ... *121*
*3.3.8. Sequencing output of the screening process* ... *121*
*3.3.9. Selection of potential effectors of neighbour suppression in the sequencing data* ... *127*
*3.3.10. Network analysis of the sequencing output* ... *130*
3.3.10.1. Network analysis using STRING ... 130
3.3.10.2. Network analysis using FunCoup ... 135
3.3.10.3. Prioritization of potential effectors of neighbour suppression in the network analysis data ... 143
*3.3.11. Functional enrichment analysis of GOI* ... *152*
3.3.11.1. Functional enrichment analysis using STRING ... 152
3.3.11.2. Functional enrichment analysis using FunCoup ... 162
3.3.11.3. Functional enrichment analysis using DAVID ... 162
3.3.11.4. Functional enrichment analysis using FunRich ... 168
3.3.11.5. Gene set enrichment analysis in the GOI lists using GSEA ... 180

*3.3.12. Transcription factor binding site enrichment analysis of GOI .... 200*
*3.3.13. Confirmation of gene knock-down effects on neighbour suppression ..... 206*
*3.3.14. Study of differential survival rates for levels of expression of genes identified as potential effectors ..... 220*
3.4. DISCUSSION ..... 231
*3.4.1. Design of a genome-wide shRNA screen in a homotypic co-culture system. ..... 231*
*3.4.2. Analysis pipeline ..... 233*
*3.4.3. Method for validation of the GOI selected ..... 234*
*3.4.4. Overlap in the mechanism of response to suppressive and permissive co-culture conditions. ..... 234*
*3.4.5. Validated GOI ..... 235*

**CHAPTER 4. A POOLED GENOME-WIDE SHRNA SCREEN FOR GENES INVOLVED IN B-CELL PROLIFERATION AND RESISTANCE TO DEATH CAUSED BY LIPIDS ..... 243**

4.1. INTRODUCTION ..... 243
*4.1.1 Type 2 diabetes and β-cell mass. ..... 243*
*4.1.2 Overview of genes involved in β-cells proliferation ..... 244*
4.1.2.1 PI3K/AKT pathway ..... 244
4.1.2.2 Mammalian Target of Rapamycin (mTOR) ..... 245
4.1.2.3 Wnt canonical signalling ..... 246
4.1.2.4 Other pathways. ..... 246
4.1.2.5 Proliferation of human β-cells ..... 247
*4.1.3 Apoptosis in β-cells ..... 247*
*4.1.4 Aim of this chapter ..... 249*
4.2. MATERIAL AND METHODS ..... 250
*4.2.1 Cell lines ..... 250*
*4.2.2 Bromodeoxyuridine (BrDU) incorporation assay ..... 250*
*4.2.3 Immunocytochemistry (ICC) of SV40 Large Tumour Antigen (Tag) 251*
*4.2.4 Production of retroviral particles using Phoenix cells ..... 251*
*4.2.5 Transduction of cells using non-concentrated retroviral particles . 251*

*4.2.6 Concentration of retroviral particles using Retro-X Concentrator (Clontech)* ..... *251*
*4.2.7 Concentration of retroviral particles using PEG* ..... *251*
*4.2.8 Transductions with Ecotropic receptor booster* ..... *252*
*4.2.9 Transductions with Viraductin* ..... *252*
*4.2.10 RNA extractions* ..... *252*
*4.2.11 RT PCR* ..... *252*
*4.2.12 qPCR* ..... *254*
*4.2.13 Palmitate dose response* ..... *256*
*4.2.14 PCR amplification of shRNA constructs for sequencing* ..... *258*
*4.2.15 Purification of amplified DNA using Agencourt Ampure XP beads.* *262*
*4.2.16 Analysis of sequencing data.* ..... *262*
*4.2.17 Design of the screening process* ..... *264*
4.3. RESULTS ..... 266
*4.3.1. Description of the cell line used as a β-cell in vitro model* ..... *266*
*4.3.2. Design of the screening process* ..... *273*
*4.3.3. Optimization of the delivery of shRNA constructs* ..... *275*
*4.3.4. Optimization of the screening process* ..... *279*
4.3.4.1 Choice of the reversibly immortalized β-cell line used ..... 279
4.3.4.2 Optimization of the resistance to death by lipids screen ..... 279
4.3.4.3 Optimization of cell seeding ..... 280
*4.3.5. Optimization of shRNA construct recovery from R7T1 cells* ..... *286*
*4.3.6. Reproducibility of the screening process* ..... *289*
*4.3.7. Running a genome-wide shRNA screening process* ..... *304*
*4.3.8. Sequencing output of the screening process* ..... *307*
*4.3.9. Selection of genes potentially involved in β cell proliferation* ..... *319*
*4.3.10. Selection of genes potentially involved in β cell resistance to death caused by chronic exposure to lipids* ..... *323*
*4.3.11. Genes differentially expressed upon exposure to palmitate* ..... *327*
*4.3.12. Network analysis of genes potentially involved in β-cell proliferation* ..... *329*
4.3.12.1 Network analysis of the candidate mouse genes identified, using STRING. ..... 329

4.3.12.2 Network analysis of the candidate mouse genes identified using FunCoup ..... 334
4.3.12.3 Hyperconnected mouse GOIs ..... 339
4.3.12.4 Network analysis of the candidate human genes identified using STRING 346
4.3.12.5 Network analysis of the identified human candidate genes using FunCoup ..... 350
4.3.12.6 Hyperconnected human GOI ..... 354
*4.3.13. Network analysis of genes potentially involved in β-cell resistance to death by lipids ..... 359*
4.3.13.1 Network analysis of the mouse GOI using STRING ..... 359
4.3.13.2 Network analysis of the mouse GOI using FunCoup ..... 361
4.3.13.3 Hyperconnected mouse GOI ..... 363
4.3.13.4 Network analysis of the human filtered GOI using STRING .. 363
4.3.13.5 Network analysis of the human filtered GOI using FunCoup . 363
4.3.13.6 Hyperconnected human GOI ..... 363
*4.3.14. Functional enrichment analysis of GOI ..... 366*
4.3.14.1. Functional enrichment analysis of the proliferation human GOI list using STRING ..... 366
4.3.14.2. Functional enrichment analysis of the human resistance to death by lipids GOI list using STRING ..... 371
4.3.14.3. Functional enrichment analysis of the human GOI lists using FunCoup ..... 373
4.3.14.4. Functional enrichment analysis of the human proliferation GOI lists using DAVID ..... 373
4.3.14.5. Functional enrichment analysis of the human resistance to lipotoxicity GOI list using DAVID. ..... 377
4.3.14.6. Functional enrichment analysis using FunRich ..... 379
4.3.14.7. Gene set enrichment analysis of the human GOI lists ..... 384
*4.3.15. Transcription factor binding site enrichment analysis. ..... 391*
4.4. DISCUSSION ..... 401
*4.4.1. In vitro model used for this screen. ..... 401*
*4.4.2. Screening process ..... 401*
*4.4.3. PCR amplification and sequencing ..... 402*
*4.4.4. Sequencing data analysis. ..... 403*

*4.4.5. Function of GOI targeted by two shRNA constructs. ..... 404*
*4.4.6. Hyperconnected genes..... 407*
*4.4.7. Functional and gene set enrichment..... 409*
*4.4.8. Concluding remarks and future work ..... 410*

**CHAPTER 5. FINAL DISCUSSION AND CONCLUSIONS ..... 411**

5.1 DESIGN AND DEVELOPMENT OF FUNCTIONAL GENOMIC SCREENS ..... 411
*5.1.1 Choice of an adequate in vitro model. ..... 411*
*5.1.2 Choice of knockdown technology ..... 412*
*5.1.3 Choice of library..... 416*
*5.1.4 Choice of phenotype to be studied and method of assessment. .. 417*
*5.1.5 Analysis method for quantification of effect ..... 418*
*5.1.6 False positives and the need for integrative analysis ..... 422*
*5.1.7 Validation strategies ..... 423*
5.2 OVERLAPS IN THE SCREENS. ..... 425
*5.2.1 Overlap in GOI lists. ..... 425*
*5.2.2 Overlap in the functional and gene set enrichment analysis..... 430*
5.3 LIMITATIONS OF THE SCREENS ..... 435
*5.2.1. In vitro model used ..... 435*
*5.2.2. Limitations derived from the use of RNAi and the NKI library..... 435*
*5.2.3. Limitations of the sequencing and analysis ..... 436*
*5.2.4. Validation..... 437*
5.4 FUTURE WORK ..... 438
5.5 CONCLUSIONS ..... 441

**APPENDICES ..... 442**

APPENDIX 1. REANNOTATION OF THE SHRNA LIBRARY USING BLAST ..... 442
APPENDIX 2. GREP COMMANDS FOR THE NEIGHBOUR SUPPRESSION SCREEN (CHAPTER 3) ..... 444
APPENDIX 3. PYTHON SCRIPT TO EXTRACT TARGETING SEQUENCES ..... 445
APPENDIX 4. STATA DO-FILE FOR TRANSFORMATION AND CALCULATION OF Z-SCORES 447
APPENDIX 5. RANDOM SAMPLING APPROACH FOR RECURSIVE PARTITIONING RESULTS 449
APPENDIX 6. RANDOM SAMPLING APPROACH FOR DETECTION OF "HUB" GENES IN NETWORK ANALYSIS..... 453

APPENDIX 7. STATA DO-FILE FOR TRANSFORMATION AND CALCULATION OF Z-SCORES IN THE B-CELL SCREEN. ..... 457
APPENDIX 8.GREP COMMANDS FOR THE B-CELL SCREEN (CHAPTER 4) ..... 460
APPENDIX 9. .. LIST OF FUNCTIONAL TERMS FOUND TO BE ENRICHED IN THE $3T3^{SUPP+}$, $3T3^{SUPP-}$ AND COMBINED GOI LISTS USING DAVID ..... 461
APPENDIX 10. . LIST OF FUNCTIONAL TERMS FOUND TO BE ENRICHED IN THE $3T3^{SUPP+}$ AND COMBINED GOI LISTS USING DAVID ..... 462
APPENDIX 11. ..LIST OF FUNCTIONAL TERMS FOUND TO BE ENRICHED IN THE $3T3^{SUPP-}$ AND COMBINED GOI LISTS USING DAVID ..... 465
**Bibliography……..467**

***Acknowledgments***

Many people have helped me through the years I worked on this thesis and thanking all of them would probably need around ten pages. But there are some people who really need to be in these acknowledgments.

I think it's fair to say that I wouldn't be typing this if it wasn't for Tim Frayling. He gave me the opportunity to do a PhD and didn't let me give up easily.

I'm very thankful to David Allard for his supervision and guidance. I wouldn't have dreamed of completing such a daunting task as the projects in this thesis if it wasn't for his patience and dedication. He welcomed me into his research team and helped me not only to be a better researcher but also to conserve my sanity through long days of work.

I'd like to thank everyone in the St Luke's labs for their help and enjoyable company. A special thank you goes to Ashley Nicholls for helping me through the intensive tissue culture and cloning sessions, and also for all the conversations about neighbour suppression and generally interesting science. Leighton Freeman and Francesca Prandi have also earned their places in these acknowledgments by being the most supportive and positive people I could have met. Thank you also to Joana, Sophie and Tiago for being there for me anytime, and Diana for helping with part of the neighbour suppression screen.

Outside of the lab, thanks to Tom for being a great friend and a brilliant musician. Many experiments in this thesis were done with his music keeping me company. And thank you to Cintia and Omar for their friendship which knows no distance. Going back home is always worth it just to see their smiles.

The greatest thank you must go to my family. My parents, siblings and grandparents have encouraged and supported me through so many difficulties and challenges.

I'd like to dedicate this thesis to three people who made, make or will make my life better. To my grandad Marcelino, I'm sorry you couldn't see me graduate but I know you would have been proud. To Marcos, thanks for your love and support and for always being there for me. To little Ollie, I hope one day you will want to know what your aunt was doing when you were only a tiny human.

***Abbreviations***

| | |
|---|---|
| **3T3** | **3-day transfer, inoculum 3x10^5 cells** |
| **BCIP** | 5-bromo-4-chloro-3'-indolyphosphate p-toluidine salt |
| **BLAST** | Basic Local Alignment Search Tool |
| **BRCA1** | Breast Cancer 1 |
| **BrDU** | Bromodeoxyuridine |
| **cDNA** | Complementary DNA |
| **CFTR** | Cystic Fibrosis Transmembrane conductance Regulator |
| **CRISPR** | Clustered, regularly insterspaced short palindromic repeats |
| **DMEM** | Dulbecco's Modified Eagle's Medium |
| **DMFS** | Distant metastasis-free survival |
| **DMSO** | Dimethyl sulfoxide |
| **DNA** | Deoxyribonucleic acid |
| **dNTP** | Deoxynucleotide |
| **DPBS** | Dulbecco's phosphate-buffered saline |
| **DSB** | Double strand break |
| **dsRNA** | Double-stranded RNA |
| **EDTA** | Ethylenediaminetatraacetic acid |
| **ENU** | Ethylnitrosurea |
| **FBS** | Foetal Bovine Serum |
| **FDR** | False discovery rate |
| **FTP** | File Transfer Protocol |
| **gDNA** | Genomic DNA |
| **GFP** | Green Fluorescent Protein |
| **GO** | Gene ontology |
| **GOI** | Gene of interest |
| **HS** | High sensitivity |
| **LB** | Luria Bertani |
| **miRNA** | Micro RNA |
| **MOI** | Multiplicity of infection |
| **mRNA** | Messenger RNA |
| **NBT** | Nitro-blue tetrazolium |
| **NCBI** | National Center for Biotechnology Information |
| **NCS** | Newborn Calf Serum |

| | |
|---|---|
| **NGS** | Next Generation Sequencing |
| **NHEJ** | Non-homologous end joining |
| **NKI** | Netherlands Kanker Instituut |
| **OD600** | Optical density at a wavelenght of 600 nm |
| **PCR** | Polymerase chain reaction |
| **PEG** | Polyethylene glycol |
| **PFA** | Paraformaldehyde |
| **PI** | Propidium iodide |
| **Polybrene** | Hexadimethrine bromide |
| **RISC** | RNA-induced silencing complex |
| **RNA** | Ribonucleic acid |
| **RNAi** | RNA interference |
| **RNase** | Ribonuclease |
| **Rpm** | revolutions per minute |
| **rpmp** | reads per million per pool |
| **SDS** | Sodium dodecyl sulphate |
| **shRNA** | Short hairpin RNA |
| **siRNA** | Small interfering RNA |
| **SVE** | SV40 transformed 3T3supp+ |
| **SVE-6** | GFP expressing SVE |
| **TALENs** | Transcription activator-like effector nucleases |
| **v/v** | volume in volume |
| **w/v** | weight in volume |

***List of Figures***

**Chapter 1**

Figure 1.1. Schematic depiction of the RNAi mechanism of mRNA degradation by shRNA and siRNA. ..... 44

**Chapter 2**

Figure 2.1 Diagrammatic representation of a haemocytometer chamber. Cells in grey shaded areas were counted. ..... 50
Figure 2.2 Schematic depiction of the pSuperRetroPuro plasmid containing the shRNA constructs ..... 60
Figure 2.3 Structure of the custom oligoes used to produce alternative shRNA constructs using pSuperRetroPuro as backbone. ..... 64
Figure 2.4. Schematic diagram of the random sampling approach to detect enrichment in interactors using data from FunCoup. ..... 66

**Chapter 3**

Figure 3.1. Schematic depiction of the analysis pipeline used. ..... 79
Figure 3.2. Schematic simplified overview of the screening process..... 83
Figure 3.3. Distribution of null (0, black bars), single (1, green bar) and multiple (2 or more, red bar) transduction events in the cell population for different MOI. ..... 87
Figure 3.4. Fluorescence microscopy photographs showing the SVE-6 colony size on both co-cultures and on plastic. ..... 90
Figure 3.5. Neighbour suppression effect of $3T3^{supp+}$ cells on SVE-6 cells. ..... 91
Figure 3.6. Results of a restriction digestion of genomic DNA and library plasmid with XhoI and EcoRI. ..... 96
Figure 3.7. Results of a restriction digestion of genomic DNA and library plasmid with varying amounts of XhoI and EcoRI. ..... 97
Figure 3.8. Results of a restriction digestion of sheared and un-sheared genomic DNA and library plasmid digested with 6 units of XhoI and EcoRI in a total volume of 50 µl for each reaction. ..... 98
Figure 3.9. Temperature gradient PCR results. ..... 99
Figure 3.10. Results of the optimization of PCR amplification cycles. ..... 100

Figure 3.11. PCR amplification results of un-sheared and sheared gDNA samples. ..... 101
Figure 3.12. DNA processing pipeline from extraction of DNA to amplification. ..... 102
Figure 3.13. Results of the PCR amplification of 0.1 ng of library plasmid with the available primers for targeting sequence amplification. ..... 107
Figure 3.14. Venn diagram showing the intersection of constructs present in replicate A ..... 114
Figure 3.15. Venn diagram showing the intersection of constructs present in replicate B ..... 115
Figure 3.16. Dotplots comparing replicates A and B of the pilot project. ..... 116
Figure 3.17. Primers used for PCR amplification of the samples. ..... 117
Figure 3.18. Overrepresented sequences table from FastQC report on quality control of A): duplicate A and B): duplicate B. ..... 118
Figure 3.19. Graphical demonstration of the double threshold system used to select GOI for the 3T3$^{supp+}$ co-culture sample. ..... 119
Figure 3.20. Graphical demonstration of the double threshold system used to select GOI for the 3T3$^{supp-}$ co-culture sample. ..... 120
Figure 3.21. Histograms of RPMP for pools 1 to 4. ..... 129
Figure 3.22. Gene network for 3T3$^{supp-}$ built using STRING and visualised in Cytoscape. ..... 132
Figure 3.23. Gene network for 3T3supp+ built using STRING and visualized in Cytoscape. ..... 133
Figure 3.24. Gene network for the combined GOI list built using STRING and visualized in Cytoscape. ..... 134
Figure 3.25. Gene network for 3T3$^{supp+}$ GOI list built using FunCoup. ..... 139
Figure 3.26. Gene network for 3T3$^{supp-}$ GOI list built using FunCoup. ..... 140
Figure 3.27. Gene network for combined GOI list built using FunCoup. ..... 141
Figure 3.28. Graphical depiction of the script used to produce random-sampling estimates of enrichment in connectors for each gene in the networks produced with FunCoup. ..... 145
Figure 3.29. Gene network for the 3T3$^{supp-}$ GOI list. ..... 147
Figure 3.30. Gene network for the 3T3$^{supp+}$ GOI list. ..... 149
Figure 3.31. Gene network for the combined GOI list. ..... 151

Figure 3.32. Venn diagram showing the overlaps in enriched functional categories (GO terms and KEGG pathways) in the 3T3$^{supp+}$ and 3T3$^{supp-}$ GOI lists and the combined list containing the genes from both lists. ..... 159
Figure 3.33. Venn diagram showing the overlap in functional enrichment in both GOI lists and the combined list. Enrichment analysis was performed using DAVID. ..... 167
Figure 3.34. Venn diagram showing the overlap in enriched functional terms (unadjusted p-value<0.05) amongst the GOI lists (3T3$^{supp+}$ and 3T3$^{supp-}$) and the genes targeted by at least one sequence present in the screen sequencing output (Screened). ..... 175
Figure 3.35. Venn diagram showing the overlapping results of the 3 methods used to analyse the 3T3$^{supp+}$ results with GSEA..... 188
Figure 3.36. Venn diagram showing the overlapping results of the 3 methods used to analyse the 3T3$^{supp-}$ results with GSEA. ..... 193
Figure 3.37. Venn diagram showing the overlapping results of the 3 methods used to analyse the z-score difference results with GSEA..... 198
Figure 3.38. Distribution of Fisher scores for the transcription factors analyzed by OPOSSUM and the percentage of GC in their binding sites for the 3T3$^{supp+}$ GOI list. ..... 203
Figure 3.39. Distribution of Fisher scores for the transcription factors analyzed by OPOSSUM and the percentage of GC in their binding sites for the 3T3$^{supp-}$ GOI list. ..... 205
Figure 3.40. Example of the steps in the image processing pipeline. ..... 213
Figure 3.41. Percentages of GFP positive area in the ROI in a replicate on 3T3$^{supp+}$ monolayers. ..... 214
Figure 3.42. Percentages of GFP positive area in the ROI in a replicate on 3T3$^{supp-}$ monolayers. ..... 215
Figure 3.43. Percentages of GFP positive area in the ROI in the second replicate on 3T3$^{supp+}$ monolayers. ..... 216
Figure 3.44. Percentages of GFP positive area in the ROI in the second replicate on 3T3$^{supp-}$ monolayers. ..... 217
Figure 3.45. Percentages of GFP positive area in the ROI in the third replicate on 3T3$^{supp+}$ monolayers. ..... 218
Figure 3.46 Percentages of GFP positive area in the ROI in the third replicate on 3T3$^{supp-}$ monolayers..... 219

Figure 3.47. Survival curves for expression of three genes selected from the 3T3$^{supp-}$ GOI list. ..... 225
Figure 3.48. Survival curves for expression of genes selected from the 3T3$^{supp+}$ GOI list. ..... 226
Figure 3.49. Results of the random sampling approach to evaluate the significance of the recursive partitioning results in the 3T3$^{supp+}$ GOI list.. 227
Figure 3.50. Results of the random sampling approach to evaluate the significance of the recursive partitioning results in the 3T3$^{supp-}$ GOI list. . 228
Figure 3.51. Survival curve obtained from the recursive partitioning analysis of the Loi dataset for SH3PXD2B. ..... 238
Figure 3.52. Survival curve obtained from the recursive partitioning analysis of the Loi dataset for RAC1. ..... 239
Figure 3.53. Survival curve obtained from the recursive partitioning analysis of the Taylor dataset for FLT4. ..... 240
Figure 3.54. Survival curves obtained from the recursive partitioning analysis of the Shedden (A) and Loi (B) dataset for FLT4 ..... 241
Figure 3.55. Survival curve obtained from the recursive partitioning analysis of the Taylor dataset for TTBK1. ..... 242

**Chapter 4**

Figure 4.1. Schematic depiction of the analysis pipeline used. ..... 263
Figure 4.2. Schematic simplified overview of the screening process..... 265
Figure 4.3 Levels of expression of β cell markers (A, C) and markers of other pancreatic lineages (B, D) in R7T1 in presence (+) or absence (-) of DOX (A, B). ..... 269
Figure 4.4. Levels of expression of non-β-cell pancreatic lineages in R7T1 cells. ..... 270
Figure 4.5. Immunocytochemistry detection of large T Antigen (TAg) in R7T1 before and after withdrawal of DOX. ..... 271
Figure 4.6 Immunocytochemistry detection of BrDU incorportation in R7T1 before and after withdrawal of DOX. ..... 272
Figure 4.7. Schematic timeline of the screening process. ..... 274

Figure 4.8. Fluorescence (A, C, E) and visible light (B, D, F) microscopy pictures of 3T3$^{supp+}$ (A, B), R7T1 without prior dissociation (C, D) and R7T1 with dissociation prior to transduction (E, F)..... 277
Figure 4.9. Fluorescence (A, C, E, G, I) and visible light (B, D, F, H, J) microscopy pictures of R7T1 cells transduced with PEG concentrated virus (A, B), RetroX concentrated virus (C, D); and R7T1 transduced with virus aided by the addition of ecotropic receptor booster (E, F), Viraductin and 2-fold concentration (G, H) and virus complexed with calcium phosphate (I, J).. ..... 278
Figure 4.10. Palmitate dose-response curves. ..... 284
Figure 4.11. Simplified representation of the transduction scheme. ..... 285
Figure 4.12. Results of a PCR amplification of 3 samples from the pilot project. ..... 288
Figure 4.13. Overrepresented sequences in the pilot project as identified by FastQC in the sequencing run for A) replicate A and B) replicate B..... 293
Figure 4.14. Venn diagram of the overlap in identified targeting sequences expected from the library composition of the two pools included in the pilot. ..... 302
Figure 4.15. Venn diagram of the overlap in identified targeting sequences expected from the library composition of the two pools included in the pilot. ..... 303
Figure 4.16. Tapestation quantification results. ..... 306
Figure 4.17. FastQC report for overrepresented sequences in the library-wide screen sequencing data in the T0 (A), T7 (B), T14 (C), Palmitate (D) and Control (E) samples..... 311
Figure 4.18.Number of shRNA targeting sequences in both control conditions after each analysis or thresholding step. ..... 321
Figure 4.19. Venn diagram showing the overlap of the proliferation (T7, T14) and death (palmitate) GOI lists..... 325
Figure 4.20. Protein interaction network for the GOI list of the T7 proliferation sample..... 331
Figure 4.21. Protein interaction network for the GOI list of the T14 proliferation sample..... 332
Figure 4.22. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists. ..... 333

Figure 4.23. Protein interaction network for the GOI list of the T7 proliferation sample..... 335
Figure 4.24. Protein interaction network for the GOI list of the T14 proliferation sample..... 336
Figure 4.25. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists. ..... 337
Figure 4.26. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists. ..... 338
Figure 4.27. Venn diagram depicting the partial overlap in hyperconnected genes between proliferation GOI lists and the combined proliferation GOI list (T7T14). ..... 344
Figure 4.28. Histogram showing the distribution of the number of additional connectors for each gene in a combined GOI list compared to the sum of the connectors in each list. ..... 345
Figure 4.29. Protein interaction network for the human GOI list of the T7 proliferation sample. ..... 347
Figure 4.30. Protein interaction network for the human GOI list of the T14 proliferation sample. ..... 348
Figure 4.31 Protein interaction network for the human proliferation combined GOI list. ..... 349
Figure 4.32. Protein interaction network for the human GOI list of the T7 proliferation sample. ..... 351
Figure 4.33. Protein interaction network for the human GOI list of the T14 proliferation sample. ..... 352
Figure 4.34. Protein interaction network for the human proliferation GOI list, which includes the T7 and T14 GOI lists. ..... 353
Figure 4.35. Venn diagram representing the overlap amongst the hyperconnected gene lists for both individual proliferation networks and the combined network ..... 358
Figure 4.36. Protein interaction network for the mouse GOI list obtained from the resistance to death by lipids screen. ..... 360
Figure 4.37. Protein interaction network for the mouse GOI list obtained from the resistance to death by lipids screen. ..... 362
Figure 4.38. Protein interaction network of the human filtered GOI list for the death screen. ..... 364

Figure 4.39. Protein interaction network of the human filtered GOI list for the death screen. ..... 365
Figure 4.40. Venn diagram representing the overlap among the proliferation human GOI lists (t7 and t14) and the combination of both lists (t7t14) in enriched functional categories according to STRING. ..... 370
Figure 4.41. Venn diagram representing the overlap in gene set enrichment in the lists ranked by the z-score in T7, T14 or T7+T14. ..... 389
Figure 4.42. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the T7 GOI list. ..... 394
Figure 4.43. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the T14 GOI list. ..... 396
Figure 4.44. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the combined proliferation GOI list. ..... 398
Figure 4.45. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the resistance to lipotoxicity GOI list. ..... 400

**Chapter 5**

Figure 5.1. Venn diagram depicting the overlap in the GOI lists from the three screens presented in this thesis. ..... 428
Figure 5.2 Venn diagram representing the overlap in functional enrichment between the neighbour suppression screen GOI lists (3T3supp+ and 3T3supp-) and the β-cell proliferation (Beta-cell prolif) and resistance to lipotoxicity (Beta-cell lipoto). ..... 433
Figure 5.3. Venn diagram depicting the overlap in enriched gene sets as determined by GSEA in the neighbour suppression (Neighbour suppre), and β-cell resistance to lipotoxicity (Beta-cell lipoto) and proliferation (Beta-cell prolif) screens. ..... 434

***List of Tables***

**Chapter 2**

Table 2.1. Composition of an individual PCR reaction ..... 54
Table 2.2. Composition of a restriction digestion reaction to linearize pSuperRetroPuro. ..... 62
Table 2.3. Composition of an individual ligation reaction to introduce the annealed shRNA oligonuclotides into pSuperRetroPuro. ..... 63

**Chapter 3**

Table 3.1. Composition of a single genomic DNA restriction digestion reaction for each sample in the screen. ..... 74
Table 3.2. Primers used to PCR amplify the shRNA constructs in each of the three samples..... 76
Table 3.3. PCR cycling conditions used for amplification of shRNA sequences. ..... 77
Table 3.4. Human tumour expression datasets used for the recursive partitioning analysis of effects of gene expression on metastasis-free survival of cancer patients. ..... 81
Table 3.5. Number of potential shRNA constructs found in the pilot sequencing data ..... 108
Table 3.6. Number of shRNA constructs designed to be in the library found in the pilot sequencing data. ..... 109
Table 3.7. Number of shRNA constructs designed to be in the library found in the pilot sequencing data when only those constructs previously sequenced are considered. ..... 110
Table 3.8. Correlation matrix of sequencing reads in the pilot project. ..... 111
Table 3.9. Correlation matrix of reads per million per pool (rpmp) in the pilot project ..... 112
Table 3.10. Correlation matrix of z-scores..... 113
Table 3.11. Number of potential shRNA constructs found in the sequencing data ..... 122
Table 3.12. Number of shRNA constructs designed to be in the library found in each sample in the sequencing data. ..... 123
Table 3.13. Correlation matrix of raw sequencing reads ..... 124

Table 3.14. Correlation matrix of reads per million per pool ..... 125
Table 3.15. Correlation matrix of z-scores..... 126
Table 3.16. Comparison of the data sources used by String and FunCoup to infer interactions between pairs of proteins. ..... 137
Table 3.17. Genes found to be enriched (p-value<0.05) in interactors in the 3T3$^{supp-}$ GOI list. ..... 146
Table 3.18. Genes found to be enriched (p-value<0.05) in interactors in the 3T3supp+ GOI list. ..... 148
Table 3.19. Genes found to be enriched (p-value<0.05) in interactors in the combined GOI list. ..... 150
Table 3.20. Top 15 enriched GO terms in the 3T3$^{supp+}$ GOI list ranked by their false discovery rate (FDR). ..... 154
Table 3.21. Enriched KEGG pathways in the 3T3$^{supp+}$ GOI list ranked by their false discovery rate (FDR). ..... 155
Table 3.22. Top 15 enriched GO terms in the 3T3$^{supp-}$ GOI list ranked by their false discovery rate (FDR). ..... 156
Table 3.23. Top 15 enriched GO terms in the combined GOI list ranked by their false discovery rate (FDR). ..... 157
Table 3.24. Enriched KEGG pathways in the combined GOI list ranked by their false discovery rate (FDR). ..... 158
Table 3.25. Functional categories enriched in 3T3$^{supp+}$ but not in 3T3$^{supp-}$ or the combined list according to STRING. ..... 160
Table 3.26. Functional categories enriched in 3T3$^{supp-}$ but not in 3T3$^{supp+}$ or the combined list according to STRING. ..... 161
Table 3.27. Results of functional enrichment in the 3T3$^{supp+}$ GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value. ..... 164
Table 3.28. Results of functional enrichment in the 3T3$^{supp-}$ GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value. ..... 165
Table 3.29. Results of functional enrichment in the combined GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value. ..... 166
Table 3.30. Top 10 functional terms (by Benjamini p-value) enriched in the 3T3$^{supp+}$ GOI list. ..... 171
Table 3.31. Top 10 functional terms in the 3T3$^{supp-}$ GOI list ranked by increasing Benjamini adjusted p-values. ..... 172

Table 3.32. Top 10 functional terms in the combined GOI list ranked by increasing Benjamini adjusted p-values. ..... 173
Table 3.33. Top 10 functional terms in the list of screened genes ranked by increasing Benjamini adjusted p-values. ..... 174
Table 3.34. Functional terms enriched (unadjusted p-value <0.05) in both the 3T3$^{supp+}$ GOI list and the list of genes targeted by the screened sequences. ..... 176
Table 3.35. Functional terms enriched (unadjusted p-value <0.05) in both the 3T3$^{supp-}$ GOI list and the list of genes targeted by the screened sequences. ..... 177
Table 3.36. Functional terms enriched (unadjusted p-value <0.05) exclusively in the 3T3$^{supp-}$ GOI list. ..... 178
Table 3.37. Functional terms enriched (unadjusted p-value <0.05) exclusively in the 3T3$^{supp+}$ GOI list. ..... 179
Table 3.38. Top 10 enriched gene sets in 3T3$^{supp+}$. Duplicate gene names were removed using their alphabetical order. ..... 185
Table 3.39. Top 10 enriched gene sets in 3T3$^{supp+}$. Duplicate gene names were solved by removing the construct with the lowest 3T3$^{supp+}$ z-score. ..... 186
Table 3.40. Top 10 enriched gene sets in 3T3$^{supp+}$. Duplicate gene names were solved by removing the construct with the lowest absolute 3T3$^{supp+}$ z-score. ..... 187
Table 3.41. Gene sets identified by GSEA as enriched in the 3T3$^{supp+}$ co-culture results using three different duplicate gene name removal systems. ..... 189
Table 3.42. Top 10 enriched gene sets in 3T3$^{supp-}$. Duplicate gene names were removed using their alphabetical order ..... 190
Table 3.43. Top 10 enriched gene sets in 3T3$^{supp-}$. Duplicate gene names were solved by removing the construct with the highest 3T3$^{supp-}$ z-score ..... 191
Table 3.44. Top 10 enriched gene sets in 3T3$^{supp-}$. Duplicate gene names were solved by removing the construct with the highest absolute 3T3$^{supp-}$ z-score. ..... 192
Table 3.45. Gene sets identified by GSEA as enriched in the 3T3$^{supp-}$ co-culture results using three different duplicate gene name removal systems. ..... 194
Table 3.46. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene

names were solved by removing the construct with the lowest difference in z-scores..... 195
Table 3.47. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene names were solved by removing the construct with the lowest absolute difference in z-scores ..... 196
Table 3.48. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene names were solved by removing the construct with their alphabetical order. ..... 197
Table 3.49. Gene sets identified by GSEA as enriched in the z-score difference results using three different duplicate gene name removal systems ..... 199
Table 3.50. Top 10 enriched transcription factors (ranked by Fisher scores) in the 3T3$^{supp+}$ GOI list according to OPOSSUM..... 202
Table 3.51. Top 10 enriched transcription factors (ranked by Fisher scores) in the 3T3$^{supp-}$ GOI list according to OPOSSUM ..... 204
Table 3.52. Genes selected for confirmation in both co-culture conditions. ... 211
Table 3.53. Genes selected for confirmation, vectors in which the sequences were cloned and source of the sequences or plasmids for each one. ..... 212
Table 3.54. Sub-datasets used for the study of gene expression changes associated to differential metastasis-free survival in cancer patients and parameters used to create them from the original datasets. ..... 223
Table 3.55. Number of genes whose expression was found to be associated with differential metastasis-free survival using the recursive partitioning approach in the different datasets used. ..... 224
Table 3.56. 3T3$^{supp+}$ top 10 genes with higher number of associations in the recursive partitioning datasets..... 229
Table 3.57. 3T3$^{supp-}$ top 10 genes with higher number of associations in the recursive partitioning datasets..... 230

**Chapter** 4

Table 4.1. Composition of a RT-PCR reaction ..... 253
Table 4.2. Composition of a single qPCR reaction. ..... 255

Table 4.3. Composition of the stock solutions for the palmitate dose response. ..... 257
Table 4.4. Composition of a single PCR reaction for the first step of the shRNA targeting sequence amplification. ..... 259
Table 4.5. Primers used for the first step amplification of the shRNA targeting sequences for each condition. ..... 260
Table 4.6. Composition of a single PCR reaction for the second step of the shRNA targeting sequence amplification. ..... 261
Table 4.7. Number of sequencing reads containing an exact match for the sequencing primers in each replicate of the pilot project for each screening condition ..... 294
Table 4.8. Number of different potential shRNA targeting sequences (19 bp preceding the loop sequence) found in the sequencing data for the four different conditions in the two replicates ..... 295
Table 4.9. Number of constructs designed to be in the library that had at least one read in each condition of the two replicates of the pilot project. ..... 296
Table 4.10. Correlation coefficients between all pairs of samples from the pilot project ..... 297
Table 4.11. Correlation matrix for reads per million per pool (rpmp). ..... 298
Table 4.12. Correlation matrix of the ratios of (experimental condition) to (control condition) ..... 299
Table 4.13. Correlation matrix of the ln(ratios) ..... 300
Table 4.14. Correlation matrix of the robust z-scores. ..... 301
Table 4.15. Number of sequencing reads obtained for each condition in the sequencing data for the library-wide screen. ..... 310
Table 4.16. Number of sequencing reads containing the sequence of the shRNA loop identified in each sample. ..... 312
Table 4.17. Number of shRNA targeting sequences expected to be in the shRNA library identified with at least one sequencing read in each sample. ..... 313
Table 4.18. Number of shRNA targeting sequences identified in the sequencing data and expected to be in the library pools which completed the screening process ..... 314
Table 4.19. Correlation matrix of raw reads for the pairs of conditions screened. ..... 315

Table 4.20. Correlation matrix for rpmp ..... 316
Table 4.21. Correlation coefficients for the ratios of rpmp of each condition to its control. ..... 317
Table 4.22. Correlation matrix for the z-scores..... 318
Table 4.23. List of genes found in both the T7 and T14 GOI lists and their z-scores in both datasets. ..... 322
Table 4.24. Genes present in the death screen and each one the proliferation screen samples. ..... 326
Table 4.25. Genes in the lipotoxicity screen GOI list found to be de-regulated in the presence of palmitate for 48 hours. ..... 328
Table 4.26. Genes found to be enriched (p-value<0.05) in interactors in the T7 GOI list. ..... 341
Table 4.27. Genes found to be enriched (p-value<0.05) in interactors in the T14 GOI list. ..... 342
Table 4.28. Genes found to be enriched (p-value<0.05) in interactors in the combined (T7, T14) GOI list. ..... 343
Table 4.29. List of hyperconnected genes (p-value <0.05) in the T7 GOI list as determined using a random sampling approach. ..... 355
Table 4.30. List of hyperconnected genes (p-value <0.05) in the T14 GOI list as determined using a random sampling approach. ..... 356
Table 4.31. List of hyperconnected genes in the combined GOI list..... 357
Table 4.32. Top 15 enriched GO terms in the T7 proliferation human GOI list ordered by their false discovery rate (FDR) according to STRING..... 367
Table 4.33. Enriched GO terms in the T14 proliferation human GOI list ordered by their false discovery rate (FDR) according to STRING ..... 368
Table 4.34. Top 15 enriched GO terms in the combined proliferation GOI list ordered by false discovery rate (FDR), as determined by STRING..... 369
Table 4.35. Top 15 enriched GO terms in the human resistance to death by lipids GOI list ordered by FDR, according to STRING. ..... 372
Table 4.36. Top 10 categories in the functional enrichment analysis of the T7 human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value. ..... 374
Table 4.37. Top 10 categories in the functional enrichment analysis of the T14 human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value. ..... 375

Table 4.38. Top 10 categories in the functional enrichment analysis of the combined proliferation human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value. ..... 376
Table 4.39. Top 10 categories in the functional enrichment analysis of the human resistance to lipotoxicity GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value. ..... 378
Table 4.40. Top 15 functional terms (by Benjamini p-value) enriched in the T7 human GOI list. Analysis was performed using FunRich. ..... 380
Table 4.41. Top 15 functional terms (by Benjamini p-value) enriched in the T14 human GOI list. Analysis was performed using FunRich. ..... 381
Table 4.42. Top 15 functional terms (by Benjamini p-value) enriched in the combined proliferation human GOI list. Analysis was performed using FunRich ..... 382
Table 4.43. Top 15 functional terms (by Benjamini p-value) enriched in the resistance to lipotoxicity human GOI list. Analysis was performed using FunRich ..... 383
Table 4.44. Top 15 gene sets enriched in the top of the T7 ranked list. ..... 386
Table 4.45. Top 15 gene sets enriched in the top of the T14 ranked list. ..... 387
Table 4.46. Top 15 gene sets enriched in the top of the T7+T14 ranked list. . 388
Table 4.47. Top 15 gene sets enriched in the palmitate z-score ranked list. .. 390
Table 4.48. Top 10 enriched transcription factor binding sites in the T7 human GOI list. ..... 393
Table 4.49. Top 10 transcription factor binding sites enriched in the T14 human GOI list ranked by Fisher score. ..... 395
Table 4.50. Top 10 enriched transcription factor binding sites in the T7T14 human GOI list. ..... 397
Table 4.51. Top 10 transcription factors enriched in the human lipotoxicity GOI list.7 ..... 399

**Chapter 5**

Table 5.1. Genome-wide shRNA libraries used for functional genomics screens in publications indexed on PubMed. ..... 415
Table 5.2. GOI detected in the neighbour suppression screen and the β-cell proliferation screen. ..... 429

# Chapter 1. Introduction

## 1.1. Functional genomics

Functional genomics seeks to study of the function of genes and non-genic elements in the DNA, nucleic acid and the proteins encoded by DNA in a genome-wide manner (Pevsner 2009). The study of gene function was classically performed using approach which could be broadly categorized in two main categories: forward genetics and reverse genetics.

Before genome-wide sequencing information was available, researchers were limited in prior hypotheses and therefore the most useful approaches were those which required no information on the gene but rather depended on the observation of phenotypes. This was the major advantage of forward genetics, which starts with the observation of a phenotype of interest and then investigates the genetic cause. Reverse genetics, on the other hand, requires the researcher to hypothesize on the genetic cause to then investigate whether the phenotype could be caused by manipulation of the hypothesized causing gene.

Forward genetic approaches were used in medicine with success for many years as a tool to identify the genetic causes of inherited diseases. Two examples of the success of this approach are the characterization of mutations in *CFTR* as the cause of cystic fibrosis (Kerem, Rommens et al. 1989, Riordan, Rommens et al. 1989, Rommens, Iannuzzi et al. 1989) and the role of *BRCA1* in breast carcinoma (Miki, Swensen et al. 1994).

The three articles describing the mutations in *CFTR* as the cause of cystic fibrosis were published on the same issue of Science. Previously, linkage analysis had uncovered a region in chromosome 7 which was associated with cystic fibrosis (Eiberg, Mohr et al. 1985, Knowlton, Cohen-Haguenauer et al. 1985, Tsui, Buchwald et al. 1985). A putative gene was then identified using chromosome walking and chromosome jumping (Rommens, Iannuzzi et al. 1989), its cDNA was cloned and analysed (Riordan, Rommens et al. 1989) and an extensive analysis of the locus was performed in affected individuals and their families to characterize the causative mutations (Kerem, Rommens et al. 1989).

*BRCA1* was identified following a similar strategy: linkage analysis revealed an association of familial breast (Hall, Lee et al. 1990) and ovarian (Narod, Feunteun et al. 1991) cancer to a locus on chromosome 17q. The region was then more finely mapped using a larger number of patients and a dense set of genetic markers (Easton, Bishop et al. 1993, Simard, Feunteun et al. 1993). A physical map of the region was constructed (Neuhausen, Swensen et al. 1994). Miki and colleagues built on this knowledge and identified and characterized the cDNA sequence, which allowed them to study several mutations present in the transcript of affected individuals (Miki, Swensen et al. 1994).

The completion of the first draft of the human genome in 2001 (Lander, Linton et al. 2001, Venter, Adams et al. 2001) produced a great amount of information about the structure of the human genome and greatly influenced the development of genomics research. The mouse genome was also sequenced (Waterston, Lindblad-Toh et al. 2002). Information was now available about the genome but the function of most of the genes were unknown. The possibility to make hypotheses was there, facilitating the use of reverse genetics techniques to the study of functional genomics.

## 1.2. Forward functional genomics screens

The main characteristic of forward genetics, its observational nature, has been adapted to the functional genomics era by increasing the frequency of the phenotypes of interest. Mutant cells or organisms are created using different approaches: exposure to X-rays (Russell, Montgomery et al. 1982) or mutagenic chemicals such as ethylnitrosurea or ENU (Coghill, Hugill et al. 2002), or introduction of exogenous DNA which causes mutations on insertion such as transposons (Collier, Carlson et al. 2005, Dorr, Janik et al. 2015). The resulting mutants are screened for the phenotype of interest and the mutation causing the phenotype is then identified. The development of transposons as tools to cause insertional mutagenesis in mice (Collier, Carlson et al. 2005, Dupuy, Akagi et al. 2005, Dupuy, Rogers et al. 2009) led to the discovery of many genes involved in several cancer phenotypes. For example, transposons were used to identify a role for *Csf1* in astrocytomas (Bender, Collier et al. 2010), to nominate candidate genes for histiocytic sarcoma (Been, Linden et al.

2014) and to uncover the role of *Cux1* as a tumour suppressor (Wong, Martincorena et al. 2014).

It is also possible to perform overexpression screens using forward genetics approaches. Misexpression of endogenous genes is achieved through the random insertion of transposable elements into the genome. In the screens performed by Nakayama and collaborators (Nakayama, Ishibashi et al. 2014) and Cruz and colleagues (Cruz, Glavic et al. 2009) in Drosophila, the inserted transposon includes a binding site for a yeast transcription factor. Crossing the flies with the transposon with flies with the transcription factor led to the production of flies where the transcription factor binds to the transposon and as it lacks a STOP signal, the adjoined gene is also expressed under the control of that transcription factor. This type of activation screen using transposons has also been adapted for mammalian cells (Chen, Stuart et al. 2013).

These screens have the same principal advantage as the classic forward genetics approach to the identification of genes causing a disease: no previous hypothesis or knowledge about the causative gene's structure or sequence are required. Gene expression is altered at random and if the phenotype of interest is observed the altered gene then needs to be identified.

### 1.3. Reverse functional genomics screens

The reasoning behind reverse genetic approaches to functional genomics can be over-simplistically summarized in a question: "What happens if I alter the normal expression of this gene?" In a genome-wide approach, this question may be formulated as "What happens if I alter the normal expression of each gene in the genome?" This question is not posed as generically of course, as the researcher using these techniques will be interested in a certain phenotype. This researcher may then ask "The alteration of the normal expression of which genes in the genome may result in the phenotype I seek?" Answering this question would result in the proposition of a gene or group of genes which have an effect on the phenotype of interest. Reverse functional genomics screens therefore seek to analyse the effect of the alteration of gene expression in a phenotype. Gene expression can be altered in two directions: increasing the levels of expression (gain-of-function) or decreasing them (loss-of-function).

Both strategies require a prior knowledge about the gene or genes to be studied, although a hypothesis-free approach can be implemented if all the genes in the genome are studied.

## 1.4. Gain-of-function studies

The logic behind the use of gain-of-function approaches to the study of functional genomics resides in that the overexpression of a gene with a certain function will exacerbate such function and cause an observable phenotype. The overexpression of a gene in reverse genetics may be caused by the introduction of exogenous cDNA.

cDNA sequences can be cloned into expression vectors from which they are transcribed into mRNA. A single cDNA expression construct may be used to study the effect of overexpressing one particular gene. However, a functional genomics approach is feasible with the available libraries or collections of cDNA sequences in expression vectors. One example is the “Expression Ready MGC cDNA libraries” available from Dharmacon. These genome-wide libraries are available for human, mouse and rat and are derived from the Mammalian Gene Consortium efforts to provide at least one full cDNA clone for all the human and mouse genes, and from at least 6,200 rat genes (Strausberg, Feingold et al. 2002, Gerhard, Wagner et al. 2004, Temple, Gerhard et al. 2009). Another example of expression library is the more recently constructed hORFeome v 8.1 (Yang, Boehm et al. 2011). These and other libraries have enabled high-throughput gain-of-function studies (Pritsker, Ford et al. 2006, Arnoldo, Kittanakom et al. 2014).

## 1.5. Loss-of-function studies

Loss-of-function studies comprise the reduction of gene expression followed by the observation of the resulting phenotype which indicates the potential function of the disrupted transcript. Loss-of-function can be exerted in a number of ways:

### 1.5.1. Homologous recombination

Also called “gene targeting”, homologous recombination as a functional genetics technique consists in the disruption of a gene’s sequence by inserting

a piece of DNA (defective copy of the targeted gene for example) flanked by the sequence complementary to the gene to be disrupted (Smithies, Gregg et al. 1985). Homologous recombination includes this exogenous piece of DNA into the targeted gene sequence, leading to disruption of the gene's normal function. A systematic approach to the design of gene targeting vectors enabled the production of a genome-wide resource to study gene function in a mouse model (Collier, Carlson et al. 2005).

### 1.5.2. Genome editing

Genome editing strategies rely on the mechanism of non-homologous end-joining (NHEJ) to repair double strand breaks (DSB). During NHEJ the fragments produced by DSB are repaired by re-ligation, without the need for a template sequence and it is therefore error-prone, often resulting in insertions or deletions (indels) at the break point (Moore and Haber 1996). NHEJ can be induced in a specific place in the genome using nucleases which recognize a sequence and introduce a DSB. Genome editing can be used to decrease gene function by either mutation or deletion of the gene of interest.

To introduce a mutation the DSB is induced in the coding region of the targeted gene. NHEJ will often produce a small indel causing a frameshift (Bibikova, Golic et al. 2002).

A region of the genome can be deleted by flanking the region with two DSBs (Bauer, Kamran et al. 2013).

Genome editing techniques may also be used to introduce exogenous DNA, but the DSB is repaired using homologous recombination instead of NHEJ. Homologous recombination uses a template to repair the DSB and so if the exogenous DNA is used as a template for repair it gets integrated in the genome (Rouet, Smih et al. 1994).

A requirement of all these methods is the introduction of DSBs, for which a nuclease is used. There are four main families of nucleases currently used:

#### *1.5.2.1. Zinc finger nucleases*

These nucleases are composed by a DNA cleaving domain and a DNA binding domain (Kim, Cha et al. 1996). The DNA binding domain can recognize up to 18 bp, enough to be unique in the genome (Bibikova, Carroll et al. 2001). They

have been used in different organisms (Bibikova, Golic et al. 2002, Doyon, McCammon et al. 2008, Santiago, Chan et al. 2008, Geurts, Cost et al. 2009)

#### *1.5.2.2. Meganucleases*

Endonucleases which recognize DNA sequences of 12 to 40 bp (Thierry and Dujon 1992). There are six families of meganucleases of which the LAGLIDADG family is the most common and used in genome editing. There have been efforts to engineer these nucleases into recognizing specific sequences of interest to produce targeted DSB (Arnould, Chames et al. 2006, Ashworth, Havranek et al. 2006, Rosen, Morrison et al. 2006, Jarjour, West-Foyle et al. 2009, Zaslavskiy, Bertonati et al. 2014)

#### *1.5.2.3. TALENS*

Transcription activator-like effector nucleases are the fusion of transcription activator-like effectors (TALEs) to the catalytic domain of FokI, an endonuclease (Christian, Cermak et al. 2010). TALEs are natural proteins from plant pathogenic bacteria which are translocated into their hosts' cells (Büttner and He 2009). They then enter the nucleus and bind DNA sequences, activating gene expression which results in increased pathogenicity or activation of defences (Büttner and He 2009). The DNA-binding domain is composed of 33 to 35 amino acids. There is a one-to-one code in that domain which was de-cyphered (Boch, Scholze et al. 2009, Moscou and Bogdanove 2009) and then used to produce TALENs capable of recognizing specific sequences (Christian, Cermak et al. 2010).

#### *1.5.2.4. CRISPR/Cas9*

CRISPR/Cas systems (clustered, regularly interspaced short palindromic repeats – CRISPR associated systems) mediate a mechanism of adaptable immune response in bacteria (Bolotin, Quinquis et al. 2005, Mojica, Diez-Villasenor et al. 2005, Pourcel, Salvignol et al. 2005, Barrangou, Fremaux et al. 2007). CRISPR are composed of repeats of 25 to 50 bp separated by non-repeat sequences or spacers of similar length of extrachromosomal origin (Mojica, Diez-Villasenor et al. 2005). These spacers are transcribed into RNA and guide the cleavage of exogenous DNA (Brouns, Jore et al. 2008, Marraffini and Sontheimer 2008) in conjunction with a second RNA sequence (Deltcheva, Chylinski et al. 2011). Cas endonucleases (of which Cas9 is a type) DNA cleavage is specifically produced within a sequence in the CRISPR, the

protospacer adjacent motif (PAM) (Garneau, Dupuis et al. 2010) which is common to all CRISPR sequences and drives target recognition (Bolotin, Quinquis et al. 2005).

Jiang and colleagues reported their success in harnessing the CRISPR/Cas system to edit bacterial DNA (Jiang, Bikard et al. 2013). Cong and colleagues engineered two CRISPR-cas systems to produce targeted cleavage of DNA guided by short RNA sequences in human cells (Cong, Ran et al. 2013), an advance reported simultaneously by Mali et al (Mali, Yang et al. 2013). This system has also been reported *in vivo* in a range of organisms (Hwang, Fu et al. 2013, Li, Qiu et al. 2013, Shen, Zhang et al. 2013, Yang, Xu et al. 2014).

These first reports of success in targeted genome editing led to the production of libraries or extensive collections of short RNA sequences which could act as guides for the CRISPR/Cas9 cleavage of DNA. Using these libraries numerous knockout screens have been performed in different *in vitro* models and for varying phenotypes of interest (Koike-Yusa, Li et al. 2014, Shalem, Sanjana et al. 2014, Wang, Wei et al. 2014, Bassett, Kong et al. 2015, Lu, Qi et al. 2015, Schmid-Burgk, Chauhan et al. 2016)

### 1.5.3. RNA interference

RNAi is a mechanism of post-trancriptional regulation of gene expression. Fire and Mello described the RNAi response to the introduction of exogenous long dsRNA molecules into *C. elegans* (Fire, Xu et al. 1998) although they could not describe the mechanism at play. Endogenous short dsRNA molecules inducing RNAi in the cells were later described and termed miRNA (Lau, Lim et al. 2001).

DNA sequences coding for miRNA molecules are estimated to represent as much as 5% of the human genome and are found scattered around the genome (Berezikov, Guryev et al. 2005). The expression of at least 30% of the genes in the human genome are predicted to be regulated by miRNA-mediated RNAi (Lewis, Burge et al. 2005).

miRNA coding sequences are transcribed by polII into a 100bp molecule (pri-miRNA) with a strong secondary structure (Lee, Kim et al. 2004). This secondary structure is recognized by Drosha, which cleaves the pri-miRNA into a still immature 70 bp molecule termed pre-miRNA (Lee, Ahn et al. 2003). This pre-miRNA is exported to the cytoplasm by Exportin-5 (Yi, Qin et al. 2003),

where it is recognized and cleaved by Dicer (Bernstein, Caudy et al. 2001). The resulting molecule is a 21-22 nucleotide long RNA sequence that determines the site of cleavage in the mRNA (Zamore, Tuschl et al. 2000). These short dsRNA sequences are by a multicomponent protein complex to form the cytoplasmic RNA-induced silencing complex, known as RISC (Hammond, Bernstein et al. 2000). The dsRNA molecule in RISC is then unwound in an ATP-dependant reaction (Nykänen, Haley et al. 2001). One of the strands, named the passenger strand is degraded while the other, the guide strand, remains associated with RISC (Khvorova, Reynolds et al. 2003, Schwarz, Hutvagner et al. 2003). It is unknown whether miRNA mediate cleavage or inhibition of translation of target transcripts (Baek, Villen et al. 2008, Selbach, Schwanhausser et al. 2008).

The implementation of RNAi as a tool for loss-of-function studies in mammalian cells required some further adjustments from the techniques used in other models. Long dsRNA molecules elicit an interferon response in mammalian cells which results in cell death (Lee and Esteban 1994). RNAi-mediated reduction of gene expression was later demonstrated in mammalian cells using short dsRNA molecules (small interfering RNA, siRNA) which did not cause an interferon response (Elbashir, Harborth et al. 2001), and the first study of RNAi-induced phenotypes was published shortly after (Harborth, Elbashir et al. 2001). Two main types of molecules (siRNA and shRNA) with different structures are now used to induce RNAi-mediated changes in gene expression. The RNAi pathway for each of the two types is schematically depicted in Figure 1.1.

#### *1.5.3.1. siRNA*

Artificially generated siRNA molecules can be transfected to the cell cytoplasm (Chiu, Ali et al. 2004). There, they are incorporated into RISC and guide the cleavage of mRNA (Hammond, Bernstein et al. 2000, Elbashir, Harborth et al. 2001, Elbashir, Martinez et al. 2001, Martinez, Patkaniowska et al. 2002). As a finite number of molecules are delivered to each cell, the knockdown produced by siRNA is transient (Yang, Tutton et al. 2001).

#### *1.5.3.2. shRNA*

Short hairpin RNA (shRNA) molecules are formed by two reverse complementary RNA sequences separated by a spacer or hairpin which can be

encoded in a plasmid vector (Brummelkamp, Bernards et al. 2002) and delivered by transfection or more efficiently, using viral encapsulation (Barton and Medzhitov 2002, Brummelkamp, Bernards et al. 2002, Devroe and Silver 2002, Shen, Buck et al. 2003, Stewart, Dykxhoorn et al. 2003). Delivery of shRNA coding plasmids using retroviruses facilitate the incorporation of the shRNA coding sequence into the host cells genome. The retrovirus delivers the RNA into the cytoplasm, where a viral retro transcriptase converts it into dsDNA that is integrated into the cell genomic DNA. The DNA sequence coding for the shRNA molecule has a promoter site for polIII. The shRNA DNA is translated into a pre-shRNA molecule which is recognized by Drosha. Drosha modifies the molecule, which is then exported to the cytoplasm by Exportin 5 (Ohrt, Merkle et al. 2006). The molecule is processed by DICER into an siRNA and then recognized by RISC to produce knockdown. Retroviral delivery of shRNA enables long term knockdown experiments as shRNA molecules are being produced in the cells.

#### *1.5.3.3. RNAi high throughput screens (HTS)*

The relative ease of RNAi induction in cells for loss-of-function study of functional genomics led to the implementation of large-scale screens in a multitude of phenotypes. For a few recent examples see (Breusegem and Seaman 2014, Goh, Seah et al. 2014, O'Reilly, Long et al. 2014, Burleigh, McKinney et al. 2015, Dopie, Rajakyla et al. 2015, Maia, Tanenbaum et al. 2015, Galeev, Baudet et al. 2016, Jolly, Luan et al. 2016, Zugasti, Thakur et al. 2016). Many collections or libraries of siRNA and shRNA are now available which have been used to study different phenotypes. Some examples of human libraries are the collection of “Silencer siRNA libraries” available from Thermo Fisher which in total cover 21,686 human genes with three siRNA per gene, the NKI human shRNA library (23,742 sequences targeting 7,914 genes (Berns, Hijmans et al. 2004)) and the Hannon library v2 (Silva, Li et al. 2005) composed of 79,805 shRNAs targeting 30,728 genes. Screens have been implemented in two different formats: arrayed and pooled.

#### *1.5.3.4. Arrayed screens*

Arrayed screening strategies consist of the independent assay of the effect of individual RNAi inducing sequences. This can be achieved by the distribution of

these sequences in different wells of multi-well plates. Cells are then plated onto the wells and targeting molecules are inserted into the cells, e.g. by reverse transfection. Experiments can be then carried out and the phenotype assayed. If the sequence in each well is known, effects in the phenotype are assigned to the knockdown of the gene targeted by the sequence. There are many examples of arrayed RNAi screens in the literature, such as (Zhou, Xu et al. 2008, Hu, Kim et al. 2009, Paulsen, Soni et al. 2009, Smith, White et al. 2010, Genovesio, Giardini et al. 2011, Gilot, Le Meur et al. 2011).

The arrayed screening format allows the investigation of different phenotypes although specialist equipment may be required to perform the experiments (such as automated liquid handling) and to evaluate the phenotype readout (fluorescence plate readers, automated imaging systems or others). More than one phenotype can be assayed simultaneously for example by detection of multiple antibodies or dyes (Gilbert, Erdmann et al. 2011) or multi-parametric analysis of imaging data (Collinet, Stoter et al. 2010).

#### *1.5.3.5. Pooled screens*

In pooled screens the shRNA sequences used to elicit RNAi are mixed and cells are transduced *en masse*. Then the cells undergo the experimental screening process, which typically result in a change of cell number. As shRNA sequences are integrated into the cells genome, they can be used as a “molecular barcode” for the cells. To evaluate a change in cell number, the shRNA sequence composition of a reference sample (usually from the initial time point in the experiment or a no-treatment control) is compared to the experimental endpoint. The evaluation of cell number changes was initially performed using microarrays (Berns, Hijmans et al. 2004, Brummelkamp, Fabius et al. 2006, Berns, Horlings et al. 2007). With the development of next generation sequencing (NGS) technologies, changes in cell number as indicated by changes in shRNA sequence representation started to be assayed by sequencing (Burgess, Doles et al. 2008, Silva, Marran et al. 2008, Bassik, Lebbink et al. 2009, Zuber, McJunkin et al. 2011).

The use of NGS eliminates the need to design microarrays with custom probes to detect the targeting sequences and it facilitates multiplexing as only two samples can be used per microarray while many samples may be sequenced at

the same time. NGS also has a higher sensitivity and dynamic range than microarrays (Marioni, Mason et al. 2008) and is not restricted to the identification of a set number of sequences unlike microarrays. However, the use of NGS requires careful optimisation of the steps previous to sequence (PCR amplification of the targeting sequences and sequencing library preparation) as any bias introduced in these steps will reflect in the sequencing results.

Also, data analysis of NGS output for pooled screens can be challenging. Several statistic and bioinformatic approaches have been developed to aid the correct interpretation of data derived from them (Birmingham, Selfors et al. 2009, Sims, Mendes-Pereira et al. 2011, Kumar, Goh et al. 2013, Dai, Sheridan et al. 2014, Diaz, Qin et al. 2015). There are many aspects which need considering such as the initial representation of the sequences in the library and the control cell population (Birmingham, Selfors et al. 2009). If the library is divided into sub-libraries or pools, the screening or sampling conditions may vary slightly between pools and potentially bias the results obtained. All sources of variability need assessing and if possible correcting for in the analysis steps.

## 1.6. Summary

Gene function can be studied using functional genomics approaches. In the last two decades RNA interference has been used to produce knockdown of genes and observe the phenotypes resulting from the reduction in expression. We now have several approaches available to experimentally elicit an RNAi response in mammalian cells. Importantly, genome-wide screens are now feasible and can be used to evaluate the impact of gene knockdown in phenotypes which result in cell number changes. The combination of genome-wide RNAi screens with next-generation sequencing technologies allow an unprecedented quantitative assessment of knockdown effects on cell number under various experimental conditions. However, the screening conditions need extensive optimization processes and data analysis must be performed carefully and taking into consideration the nature of the experiments and the phenotype assessed.

Pooled shRNA screens are ideally suited to identify molecular mechanisms for phenomena resulting in a change in rates of proliferation, death or both. I have chosen two biological phenomena that have important medical translation and for which limited or no molecular mechanism is currently known. One is related to the cancer microenvironment (neighbour suppression) and the other to diabetes (β-cell proliferation and resistance to death by lipids). For each of these screens I describe the optimization procedures pre-screening, the screening process and the data analysis.

The aims of the three screens introduced here were similar: to analyse the effect of the reduction of gene expression at a genome-wide scale in an *in vitro* model of each of the phenomena studied using shRNA-mediated RNAi. These aims were stated under the hypothesis that neighbour suppression, β-cell proliferation and β-cell resistance to death by lipids are influenced by the reduction of expression levels of certain genes.

To model neighbour suppression *in vitro* I used a co-culture system in which SV40-transformed 3T3 mouse fibroblasts are used as a model tumourigenic cells and co-cultured with contact inhibited non-transformed 3T3 mouse fibroblasts as a model of normal counterparts. I used two isogenic non-transformed cell lines: $3T3^{supp+}$, which suppresses the growth of the transformed cells when in co-culture; and $3T3^{supp-}$, which lacks this suppressive effect.

As a model of β-cells I have used R7T1 cells, a reversibly inmortalised mouse β-cell line.

shRNA molecules stably reduce gene expression and can therefore be used in longer experiments than siRNA. This facilitates the study of phenotypes which may take a relatively long time to manifest or to be detectable, such as subtle changes in proliferation rates in β-cells or in the suppression of tumourigenic cell colony growth. Furthermore, they do not eliminate gene expression all together, allowing the study of genes which upon complete knockout may cause cell death. The use of shRNA molecules stably integrated in the genome of cells as molecular barcodes to identify changes in cell growth rate allows the study of a large collection of shRNA targeting sequences simultaneously, in a pooled screen which can be then deconvoluted in the analysis stage. This type of pooled approach is especially suitable for high-throughput and genome-wide studies of phenomena where the molecular mechanisms are unknown or poorly characterized as it does not require a prior hypothesis to design shRNA molecules. Given that both neighbour suppression and, to a lesser extent, β-cell proliferation and death remain poorly characterized, I used a high-throughput pooled shRNA screen to identify genes with a potential effect on both phenomena.

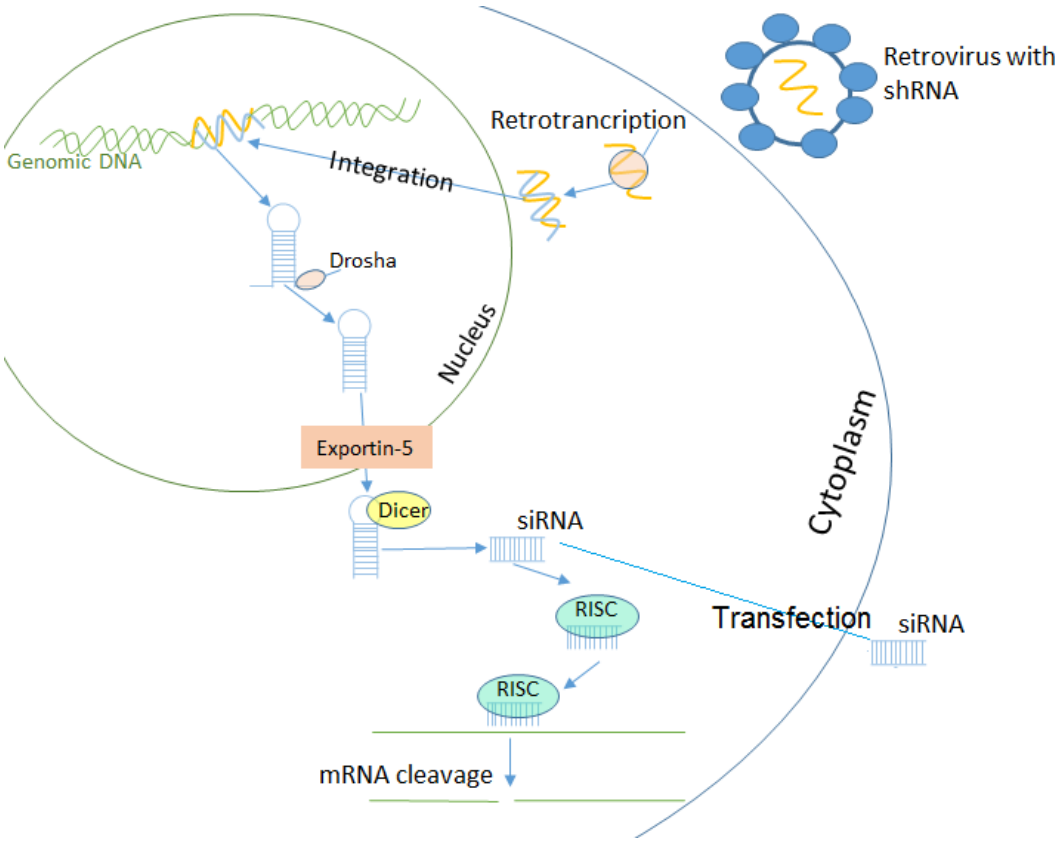



*Figure 1.1. Schematic depiction of the RNAi mechanism of mRNA degradation by shRNA and siRNA.*

# Chapter 2. Material and methods

This chapter describes the material and methods that were used during the course of the projects described in this thesis. Each subsequent chapter will introduce a detailed account of the methods specific for that chapter.

## 2.1 Cell culture

### 2.1.1 List of materials

Dulbecco's Modified Eagle's Medium (DMEM), high glucose, GlutaMAX, pyruvate (31966) (Life technologies, Paisley, UK)

DMEM, high glucose, L-glutamine (41965) (Life technologies, Paisley, UK)

Opti-MEM Reduced Serum Medium, GlutaMAX Supplement (10592) (Life Technologies, Paisley, UK)

HyClone Newborn Calf Serum (USA) (10500) (GE Healthcare Life Sciences, Little Chalfont, UK)

Tet System Approved Foetal Bovine Serum (FBS) (USA) (631106) (Takara Bio Europe/Clontech, Saint-Germain-en-Laye, France)

FBS, qualified, heat inactivated (E.U.-approved) (10500) (Life Technologies, Paisley, UK)

Penicillin Streptomycin (5,000 Units/mL Penicillin, 5,000 µg/mL Streptomycin) (15070) (Life technologies, Paisley, UK)

Dulbecco's phosphate-buffered saline (DPBS), no calcium, no magnesium (14190) (Life technologies, Paisley, UK)

0.25 % Trypsin (25200) (Life technologies, Paisley, UK)

Dimethyl sulfoxide, anhydrous, 99.7% (10103483) (Fisher Scientific, Loughborough, UK)

Disinfectant, Virkon (12358667) (Fisher Scientific, Loughborough, UK)

Puromycin dihydrochloride (10054207) (Fisher Scientific, Loughborough, UK)

Turbofect (10304780) (Fisher Scientific, Loughborough, UK)

Polybrene (H9268) (Sigma-Aldrich, Poole, UK)

ProFection® Mammalian Transfection System (E1200) (Promega, Southampton, UK)

Viraductin™ Retrovirus transduction kit (RV-201) (200 Transductions) (Cambridge Bioscience, Cambridge, UK)

Ecotropic Receptor Booster (631471) (Takara-Clontech, Saint-Germain-en-Laye, France)

Retro-X Concentrator (631455) (Takara-Clontech, Saint-Germain-en-Laye, France)

Polyethylene glycol 8000 (PEG) (P5413) (Sigma-Aldrich, Poole, UK)

Paraformaldhyde (PFA) (10131580) (Fisher Scientific, Loughborough, UK)

Ethanol absolute (≥99.8 %) (32221) (Sigma-Aldrich, Poole, UK)

Nitro-blue tetrazolium (NBT) (Thermo Scientific, Loughborough, UK)

5-bromo-4-chloro-3'-indolyphosphate p-toluidine salt (BCIP) (Thermo Scientific, Loughborough, UK)

Dimethylformamide (≥99.8 %) (Sigma-Aldrich, Poole, UK)

Mouse anti-SV40 Large T/ small t antigen (pAb419) (Monoclonal antibody service, ICRF / CR UK, London, UK)

Monoclonal mouse anti BrDU (M0744) (Dako, Ely, UK)

Polyclonal Goat Anti-Mouse Immunoglobulins (AP) (Dako, Ely, UK)

BrDU labelling reagent (Fisher Scientific, Loughborough, UK)

Propidium iodide (PI) – 1.0 g/mL solution in water (Fisher Scientific, Loughborough, UK)

Methanol (≥99.8 %) (Sigma-Aldrich, Poole, UK)

1.8 mL Cryotube vials (Thermo Scientific, Loughborough, UK)

Mr. Frosty freezing container (Thermo Scientific, Loughborough, UK)

15 mL conical centrifuge tubes (VWR, Lutterworth, UK)

50 mL conical centrifuge tubes (VWR, Lutterworth, UK)

Cell culture dish with vents, 100 x 20 mm, sterile (Greiner Bio-One, Stonehouse, UK)

Cell culture multi-well plate, 6 well, crystal-clear, with lid, sterile (Greiner Bio-One, Stonehouse, UK)

Cell culture flask, 50 ml, 25 $cm^2$, filter screw cap red, clear, sterile (Greiner Bio-One, Stonehouse, UK)

Cell culture flask, 250 ml, 75 $cm^2$, filter screw cap red, clear, sterile (Greiner Bio-One, Stonehouse, UK)

Corning® cell scrapers, blade L 30 cm, handle L 39 cm (Sigma Aldrich, Poole, UK)

### 2.1.2 Cell lines

**Phoenix - ECO**: Variant of the highly transfectable 293T cell line (human embryonic kidney cell line). This cell line expresses Moloney Murine Leukemia Virus (M-MLV) viral packaging proteins (gag-pol) and an ecotropic envelope protein. It can be used to produce retroviral particles that infect dividing mouse or rat cells (Swift, Lorens et al. 2001). They were routinely grown in DMEM (high glucose, GLUTAMAX, pyruvate) supplemented with 10% FBS and 1% Penicillin Streptomycin.

Cell lines used specifically in each chapter will be described in their relevant sections.

### 2.1.3 General tissue culture procedures

For clarity and brevity in the methods it can be assumed that unless otherwise specified, the following procedures were performed following the protocols and using the materials and equipment stated below:

- **Cell culture sterile conditions:** All cell culture procedures were carried out inside a class II microbiological safety cabinet (MSC) that was sprayed with 70% ethanol and wiped down with tissue prior to use. Any material entering the MSChad previously been sprayed with 70% ethanol.
- **Decontamination of biological material**: Consumables that had been in contact with biological samples were decontaminated with 1% (weight/volume) Virkon for at least 20 minutes before disposal.
- **Pre-warming medium**: Growth medium was placed in a 37ªC water bath for at least 30 minutes before use.

- **Pelleting cells**: A tube containing cells in suspension was centrifuged at 500 g for 5 minutes to pellet cells.
- **Aspirating**: Liquid was removed using an autoclaved unplugged glass Pasteur pipette (Fisherbrand), connected with rubber tubing to a Dry Laboratory Vacuum / Pressure Station (Free Flow 9.2L/Min 230V 50Hz Vacuum 292Mbar Pressure 2.3Bar) (Fisherbrand).
- **Growth medium**: DMEM supplemented with serum and antibiotics as stated for each cell line.
- **Cell culture conditions**: all cell lines were cultured in incubators at 37ºC and 5% $CO_2$.
- **Trypsinizing cells**: Growth medium was aspirated, cells were washed twice with DPBS and trypsin was added to the flask or dish according to its growing surface (approximately 0.02 mL per $cm^2$). The flask or dish was placed in an incubator set to standard tissue culture growing conditions for approximately 3 minutes or until cells were detached from the growth surface. Then, growth medium was added (at least 2:1 ratio to trypsin) to inactivate trypsin.
- **Freezing medium**: freezing medium was prepared as a 10% (volume/volume) dilution of DMSO in the serum used in the growth medium.

#### 2.1.4 Sub-culturing cells

Cells were not allowed to reach more than 70% confluency when being routinely cultured. Sub-confluent flasks or dishes of cells were trypsinized and plated at a ratio dependent on starting confluency (e.g. 1 in 5 of the total cells), or counted using a haemocytometer (2.1.7) to plate a specific number of cells.

#### 2.1.5 Freezing cells

Cells were considered to be apt for freezing when they had reached 50% to 70% confluency. Cells were trypsinized, the cell suspension was transferred to a 15 mL tube and cells were pelleted. The supernatant was aspirated and cells were re-suspended in freezing medium. This cell suspension was transferred to a sterile CryoTube vial, which was tightly closed and placed in a Mr. Frosty Freezing container. This was placed in a -80ºC freezer for at least 24 hours and the cryovial was transferred to a liquid nitrogen dewer for long term storage.

### 2.1.6 Thawing cells

To thaw a vial of cells, it was removed from the liquid nitrogen tank and placed in a 37ºC water bath for 3-5 minutes. When the cell solution had defrosted it was transferred to a 15 mL plastic tube containing 9 mL of pre-warmed growth medium. Cells were pelleted, supernatant was aspirated and the pellet was gently re-suspended into 2 mL of pre-warmed growth medium. The cell suspension was added to a 10 cm tissue culture dish containing 12 mL of pre-warmed growth medium and cells were grown until ready to sub-culture.

### 2.1.7 Counting cells

Cells were counted using a glass haemocytometer (Fisher Scientific). Both the haemocytometer and the coverslip were cleaned with 70% ethanol before use and the coverslip was placed on top of the haemocytometer. Trypsinized cells were pelleted and gently re-suspended in an appropriate volume of growth medium, dependant on the estimated number of cells. The cell suspension was pipetted between the coverslip and the haemocytometer until the chambers were full. Cells in the shaded areas in Figure 2.1 were counted. The sum of all cells in these areas was divided by 4 to get an average number per individual shaded area. The concentration of cells in the cell suspension was calculated as: (average number of cells per shaded area)*$10^4$=cells/mL in suspension

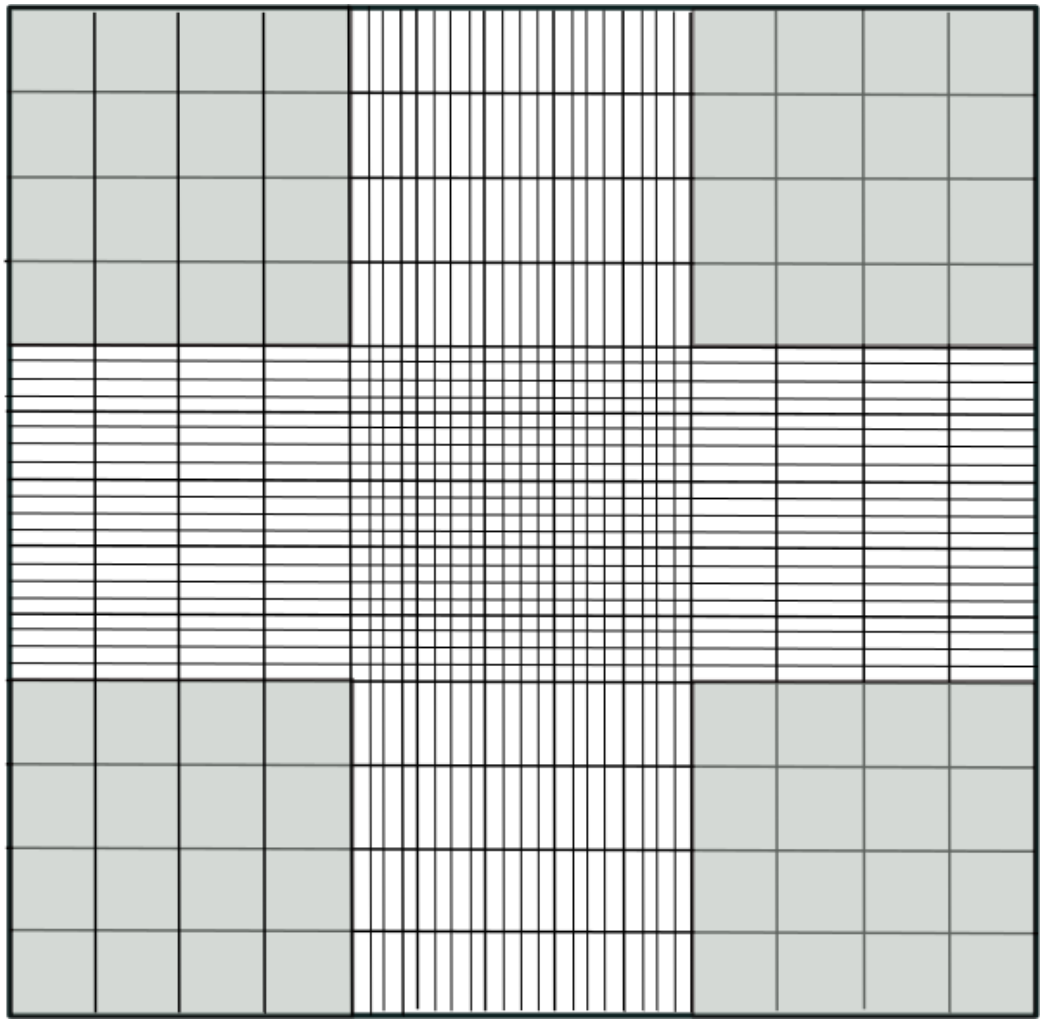


*Figure 2.1. Diagrammatic representation of a haemocytometer chamber. Cells in grey shaded areas were counted.*

### 2.1.8 Puromycin kill curve

Cells were seeded at approximately 30% confluence in a 6-well cell culture multi-well plate and grown for 24 hours. Then, medium in the wells was aspirated and replaced with growth medium containing a range of concentrations of puromycin (0 to 10 µg/mL). Cell survival was assessed after 48 hours using a Nikon Eclipse TS100-F inverted light microscope (Nikon UK Ltd). The lowest concentration at which all cells in the well had died was used as the working concentration for puromycin selection.

### 2.1.9 Puromycin selection of transduced cells.

Selection medium was prepared by transferring 12mL of growth medium to a 15mL tube and adding puromycin to reach the appropriate concentration (determined using a kill curve performed as described in section 2.1.8). The growth medium in the dish was aspirated and replaced with selection medium, and the dish was returned to the incubator. A non-transduced dish was used as a control in each batch of selections to control for potential variations in antibiotic resistance. The rate of cell death was monitored after 48 hours of selection. If the non-transduced dish had a death rate of 100%, transduced cells were considered selected and living cells were considered puromycin resistant.

### 2.1.10 Harvesting cells for DNA extraction

Growth medium was aspirated and cells were washed twice with DPBS. After the last wash, a volume of DPBS proportional to the growth surface area was added and cells were detached from the surface and into the DPBS using a cell scraper. This cell suspension was transferred to a 15 mL tube and cells were pelleted. The supernatant was aspirated and the pellet was used to extract DNA.

## 2.2 Molecular biology

### 2.2.1. List of materials

Tris HCl (T3253) (Sigma-Aldrich, Poole, UK)

Ethylenediaminetetraacetic acid (EDTA) (E9884) (Sigma-Aldrich, Poole, UK)

Sodium dodecyl sulphate (SDS) (463143) (Sigma-Aldrich, Poole, UK)

RNase A (E1008) (Cambridge Bioscience, Cambridge, UK)

Proteinase K DNase and RNase free (10407583) (Fisher Scientific, Loughborough, UK)

Phenol:chloroform (17909) (Fisher Scientific, Loughborough, UK)

Qubit dsDNA HS Assay Kit (Q32851) (Fisher Scientific, Loughborough, UK)

Q5® High-Fidelity Polymerase (M0491) (New England Biolabs, Herts, UK)

Eppendorf® PCR 0.2 mL tubes (Z316156) (Sigma-Aldrich, Poole, UK)

96-well clear plates (4483348) (Fisher Scientific, Loughborough, UK)

Agarose broad separation range for DNA/RNA, Genetic Analysis Grade (10688973) (Fisher Scientific, Loughborough, UK)

Zymoclean Gel DNA Recovery Kit (D4008) (Cambridge Bioscience, Cambridge, UK)

QIAquick Gel Extraction Kit (28706) (Qiagen, Manchester, UK)

Ethidium Bromide, 10 mg/mL (15585011) (Fisher Scientific, Loughborough, UK)

Gel loading dye (Purple) (B7024) (New England Biolabs, Herts, UK)

Scalpel disposable sterile No. 22 (11758353) (Fisher Scientific, Loughborough, UK)

EcoRI (R0101) (New England Biolabs, Herts, UK)

XhoI (R0146) (New England Biolabs, Herts, UK)

RevertAid H Minus Reverse Transcriptase (200 U/µL) (EP0451) (Fisher Scientific, Loughborough, UK)

Deoxynucleotide (dNTP) Solution Mix (R0447S) (New England Biolabs, Herts, UK)

Oligo(dT)20 Primer (18418020) (Fisher Scientific, Loughborough, UK)

26G needle (BD, 14-826-10) (Fisher Scientific, Loughborough, UK)

### 2.2.2. Phenol:chloroform extraction of DNA from cell lines

Pelleted cells obtained as per section 2.1.10 were re-suspended in an adequate volume of TE buffer (10 mM Tris HCl (pH 8.0) 10 mM EDTA) to obtain an estimated cell concentration of $10^7$ cells/mL. 10% SDS and RNase were added to a final concentration of 0.5% SDS and 2ul/mL RNase, and mixed thoroughly. This solution was incubated for at least 1 hour at 37ºC in a water bath. Proteinase K (10mg/mL) was added to reach a final concentration of 200 µg/mL, and incubated at 55ºC in a water bath for at least 2 hours. DNA was then purified according to the method by Sambrook and Russell (Sambrook and Russell 2006).

### 2.2.3. Quantification of nucleic acids

RNA and DNA (except samples that were to be sent for Illumina Next Generation Sequencing) were quantified routinely using a Nanodrop 8000 UV-Vis spectrophotometer (Thermo Scientific) according to the manufacturer instructions. Samples sent for Illumina Next Generation Sequencing were quantified with a Qubit 2.0 Fluorometer (Invitrogen) using a Qubit dsDNA HS Assay Kit (Invitrogen) and a Tapestation 2200 (Agilent), according to their respective manufacturer instructions.

### 2.2.4. PCR amplification of DNA

The shRNA sequences inserted in the pSuperRetroPuro plasmid were amplified by PCR using Q5 high fidelity DNA polymerase. PCR reaction mixtures were set up in 0.2 mL flat cap thin wall tubes or 96-well plates if amplifying more than 24 samples simultaneously. The final composition of an individual PCR reaction is shown in Table 2.1. Primers used and cycling conditions are experiment-specific and will be indicated in their relevant sections.

| **Component** | **Stock concentration** | **Volume** | **Final concentration** |
|---|---|---|---|
| Q5 Reaction Buffer | 5X | 5 µl | 1X |
| dNTP mix | 10mM | 0.5 µl | 0.2 mM |
| Forward Primer | 10 µM | 1.25 µl | 0.5 µM |
| Reverse Primer | 10 µM | 1.25 µl | 0.5 µM |
| Q5 High Fidelity DNA Polymerase | 2 U/ µl | 0.25 µl | 0.02 U/µl |
| DNA | Variable | Variable | <40 ng/ul |
| PCR grade $H_2O$ | NA | To 25 µl | NA |

*Table 2.1. Composition of an individual PCR reaction*

### 2.2.5. Preparation of agarose gels

A volume of 1X gel running buffer (TBE: 0.1 M Tris base, 0.1 M boric acid, 2mM EDTA or TAE: 40 mM Tris base, 20 mM acetic acid, 1mM EDTA), dependent on the size of gel to be run, was measured using a measuring cylinder and transferred to a clean Pyrex glass bottle. The appropriate weight of agarose to be added was calculated as agarose(g) = (percentage of agarose in gel) * volume (mL) / 100.

Agarose was dissolved in the running buffer using a microwave to heat up the solution until no agarose particles were visible. When the solution was homogenous it was left to cool down at room temperature. When the bottle was cold enough to be handled, ethidium bromide was added to a final concentration of 1µg/mL. The solution was swirled to evenly distribute the ethidium bromide and then poured in the appropriate tray to cast the agarose gel.

### 2.2.6. Agarose gel electrophoresis of DNA

Agarose gels were prepared as per 2.2.5. DNA Gel loading dye was added to the DNA samples and DNA size ladders to a final concentration of 1X. DNA samples and DNA size ladders were loaded into the gels using micropipettes. Gels were run using a Bio-Rad Model 200/2.0 power supply (Bio-Rad, Hemel Hempstead, UK)

### 2.2.7. DNA extraction from agarose gel

After the electrophoresis had completed, DNA in the gels was visualized under UV light using a Corning Labnet Enduro UV Transiluminator. Pictures of the gels were taken using a BioRad ChemiDoc XRS. Fragments of the gel containing the DNA of interest were cut out from the gel using a disposable sterile scalpel (Fisher Scientific).

DNA was extracted from the agarose gel fragments using either a QIAquick Gel Extraction Kit (Qiagen) or a Zymoclean Gel DNA Recovery Kit (Zymo Research, California, USA), according to the manufacturer's instructions.

### 2.2.8. Sequencing

Purified DNA samples were sent to be sequenced in the University of Exeter Sequencing Service, using the standard TruSeq v.3 protocol on an Illumina HiSeq2500.

## 2.3 Cloning

### 2.3.1 List of materials

LB Broth High Salt (L1704) (Melford, Ipswich, UK)

Agar-agar, pure, granular powder (10548030) (Fisher Scientific, Loughborough, UK)

Petri dish 150x20mm Triple Vented (82.1184) (Sarstedt, Leicester, UK)

Spreader inoculating, L-shape (11836191) (Fisher Scientific, Loughborough, UK)

$CaCl_2$ (C1016) (Sigma-Aldrich, Poole, UK)

BRAND® microcentrifuge tube, 0.5 mL with lid (Z334006) (Sigma-Aldrich, Poole, UK)

T4 DNA Ligase (M0202) (New England Biolabs, Herts, UK)

EcoRI (R0101) (New England Biolabs, Herts, UK)

XhoI (R0146) (New England Biolabs, Herts, UK)

HindIII (R0104) (New England Biolabs, Herts, UK)

BglII (R0144) (New England Biolabs, Herts, UK)

### 2.3.2 Bacterial growing medium

Bacteria were grown using Luria Bertani (LB) medium. LB medium was prepared using a commercial mix of components, dissolving the mix in an appropriate volume of distilled water to produce the following composition: 0.5 % (w/v) Yeast Extract, 1 % (w/v) NaCl and 1 % (w/v) Tryptone. PH was adjusted to 7.2 using a S220 SevenCompact pH/Ion meter (Mettler Toledo), and the medium was then autoclaved.

### 2.3.3 LB agar plates

Agar plates for bacterial growth were prepared by addition of 1.5% (w/v) agar to LB medium in glass bottles. This mix was autoclaved and then left to cool down until temperature was lower than 55ºC. Then, selection antibiotic was added to the solution, which was mixed and poured into a Petri dish and allowed to set at room temperature. Plates were stored inverted at 4ºC.

### 2.3.4 Preparation of competent bacteria

Competent bacteria for plasmid DNA propagation (*Escherichia coli* (*E. coli*), DH5α strain) were prepared by spreading 10µl of DH5α on a LB-agar plate

using a disposable sterile plastic cell spreader. The plate was left for 5 minutes and then inverted and incubated at 37ºC for approximately 16 hours. A single colony was picked from the plate using a sterile wooden toothpick and inoculated into 100 mL of sterile LB medium. Bacteria were incubated for approximately 12 hours in a Minitron AB70 bacterial shaker (INFORS HT) at 37°C shaking at 250 rpm. Then, 1 mL was inoculated into 100 mL of sterile LB medium and the culture was grown in the same conditions as above until the optical density ($OD_{600}$) was 0.4-0.5. The culture was then incubated on ice for 20 minutes, transferred to pre-cooled 50 ml centrifuge tubes and centrifuged at 4000g for 10 minutes at 4ºC. The supernatant was discarded and the pellet in each tube was re-suspended in 10 mL of ice-cold 0.1M $CaCl_2$ before being centrifuged again using the same conditions. Supernatant was discarded and the pellet was re-suspended in 1mL of ice-cold 0.1M $CaCl_2$ with 10 % (w/v) glycerol and placed on ice. 0.5mL microcentrifuge tubes were pre-cooled on dry ice for 5 minutes. The cell suspensions were then pipetted into the tubes in 100 µL aliquots and stored at -80°C.

### 2.3.5 Heat-shock transformations

An aliquot of competent DH5α cells (2.3.4) was thawed on ice before adding 10 ng of the plasmid to be transformed. The tube was gently mixed to distribute the DNA evenly and incubated on ice for 30 minutes. Then, bacteria were heat-shocked by placing the tube at 42ºC for 45 seconds and back on ice for 2 minutes. The mix was added to 3ml of LB medium in a 15 ml tube and incubated at 37ºC in a shaking incubator at 250rpm. The tube was then centrifuged at 3200xg for 5 minutes in a cooled centrifuge at 4ºC. The supernatant was discarded and the pellet re-suspended in 200 µl of LB medium. An appropriate volume of this solution was pipetted onto an LB-agar plate and evenly spread using a disposable sterile plastic cell spreader. The plate was left at room temperature for 5 minutes and then inverted and placed into an incubator set to 37ºC for approximately 16 hours.

### 2.3.6 Plasmid isolations

Individual colonies from transformed DH5α cells were picked using a sterile wooden toothpick and placed into tubes containing LB medium with the appropriate selection antibiotic for expansion of the clonal bacteria. The tubes were placed in a shaking incubator at 190 rpm and bacteria were left to grow for

approximately 12 hours. Plasmids were then extracted using the Plasmid Midi or Mini Kit (Qiagen) according to the manufacturer instructions.

## 2.4 NKI mouse shRNA Library

The NKI (Netherlands Kanker Instituut) mouse shRNA library is composed of approximately 28,000 different shRNA sequences that target approximately 14,000 genes with a two-fold redundancy. These shRNA constructs are inserted into the pSUPER.retro.puro plasmid, a retroviral expression vector derived from the Murine Embryonic Stem Cell virus (pMSCV). The structure of pSUPER.retro .puro is depicted in Figure 2.2. Important elements of this plasmid are:

- Ampicillin resistance gene (AMP). Allows propagation of the plasmid in bacteria using ampicillin to select transformed bacterial colonies.

- Puromycin resistance gene (puro). Allows selection of mammalian cells containing the plasmid.

- Viral packaging signal. Necessary for the production of retroviral particles by the packaging cell line.

- EcoRI and XhoI restriction sites. They flank the shRNA sequence.

- Hairpin sequence. Composed of a sense 19 nucleotide (19-mer) sequence, a hairpin loop and an antisense 19-mer, reverse complement of the sense 19-mer. Its transcript is processed into a 19 nucleotide siRNA that ultimately produces the RNAi response.

The NKI library was a gift from Dr Roderick Beijersbergen and amplified in-house as described in (Nicholls 2015). The in-house amplified library was used throughout the projects described in this thesis. The plasmids in this library are distributed into 73 different tubes which constitute the different “pools” of the library. Each pool contains approximately 384 different constructs.

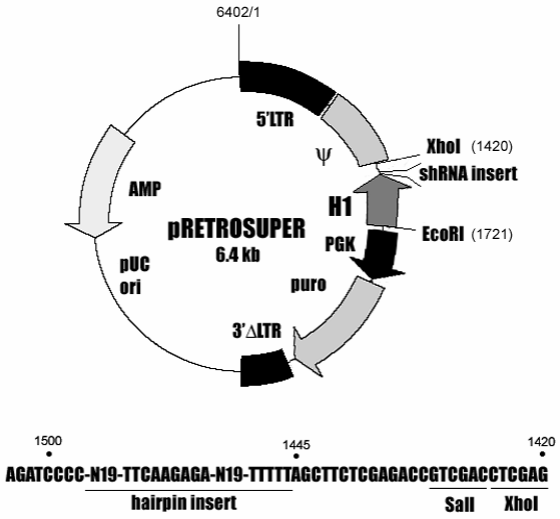



*Figure 2.2. Schematic depiction of the pSuperRetroPuro plasmid containing the shRNA constructs*

## 2.5 Re-annotation of the NKI shRNA library

A standalone version of BLAST+ version 2.2.30 was downloaded from the NCBI BLAST FTP server (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) on 17th March 2015 as a WINDOWS executable file. This was installed on a PC with an Intel Core i7 processor and 8 GB of RAM. The RefSeq mRNA database for *Mus musculus* was downloaded from NCBI nucleotide (http://www.ncbi.nlm.nih.gov/nuccore) as a FASTA file. A database was constructed from that FASTA file using BLAST+.

A text file containing all shRNA constructs designed to be in the library and constructs present in previous sequencing runs was built. These sequences were aligned to the mouse transcriptome using the standalone BLAST+. A detailed account of the commands and options used for building the RefSeq mRNA database and for the re-annotation process can be found on Appendix 1.

## 2.6 Preparation of plasmids with custom shRNA constructs

Pairs of oligonucleotides with the structure depicted in (Figure 2.3) were purchased from Sigma-Aldrich. The targeting sequence will for each oligo used will be specified in their relevant sections. Oligonucleotides were dissolved in DNA-grade $H_2O$ to a concentration of 3 mg/mL. An annealing buffer was prepared (100 mM NaCl and 50 mM HEPES pH 7.4) and the annealing reaction was set up by mixing 2 µL of each oligo in the pair with 46 µL of annealing buffer. The mix was incubated at 90ºC for 5 minutes on a heater block and then removed and left to cool down at room temperature. Annealed oligo pairs were stored at 20ºC until needed. Meanwhile, pSuper.retro.puro was linearized by digestion with XhoI and BglII as described in Table 2.2. Annealed oligo pairs were ligated into linearized plasmid as described in Table 2.3, and the reactions were treated with BglII prior to transformation by adding 1 µL of BglII to the reaction and incubating at 37ºC for 30 minutes. This step reduces background transformation of unsuccessful ligations, as the BglII site is destroyed upon ligation of the oligo pairs. Ligation products were used to transform competent bacteria as described in 2.3.5. At least three different colonies were picked and plasmids were grown and isolated as described in 2.3.6.

| | **Stock concentration** | **Volume** | **Final concentration** |
|---|---|---|---|
| **pSuper.retro.puro** | Variable | Variable | 4 µg |
| **XhoI/HindIII** | 20 U/ µL | 4 µL | 0.8 U/ µL |
| **BglII** | 10 U/ µL | 4 µL | 0.4 U/ µL |
| **Buffer** | 10 X | 10 | 1X |
| **$H_2O$** | Not applicable | Up to 100µL | Not applicable |

*Table 2.2. Composition of a restriction digestion reaction to linearize pSuperRetroPuro.*

| | **Stock concentration** | **Volume** | **Final concentration** |
|---|---|---|---|
| **Annealed oligonucleotides** | 240 µg/ mL | 2 µL | 48 µg/ mL |
| **Linearized plasmid** | 500 µg/mL | 1 µL | 50 µg/ mL |
| **T4 ligase buffer** | 10X | 1 µL | 1X |
| **T4 DNA ligase** | 400 U/ µL | 1 µL | 40 U/ µL |
| **$H_2O$** | Not applicable | 5 µL | Not applicable |

*Table 2.3. Composition of an individual ligation reaction to introduce the annealed shRNA oligonucleotides into pSuperRetroPuro.*

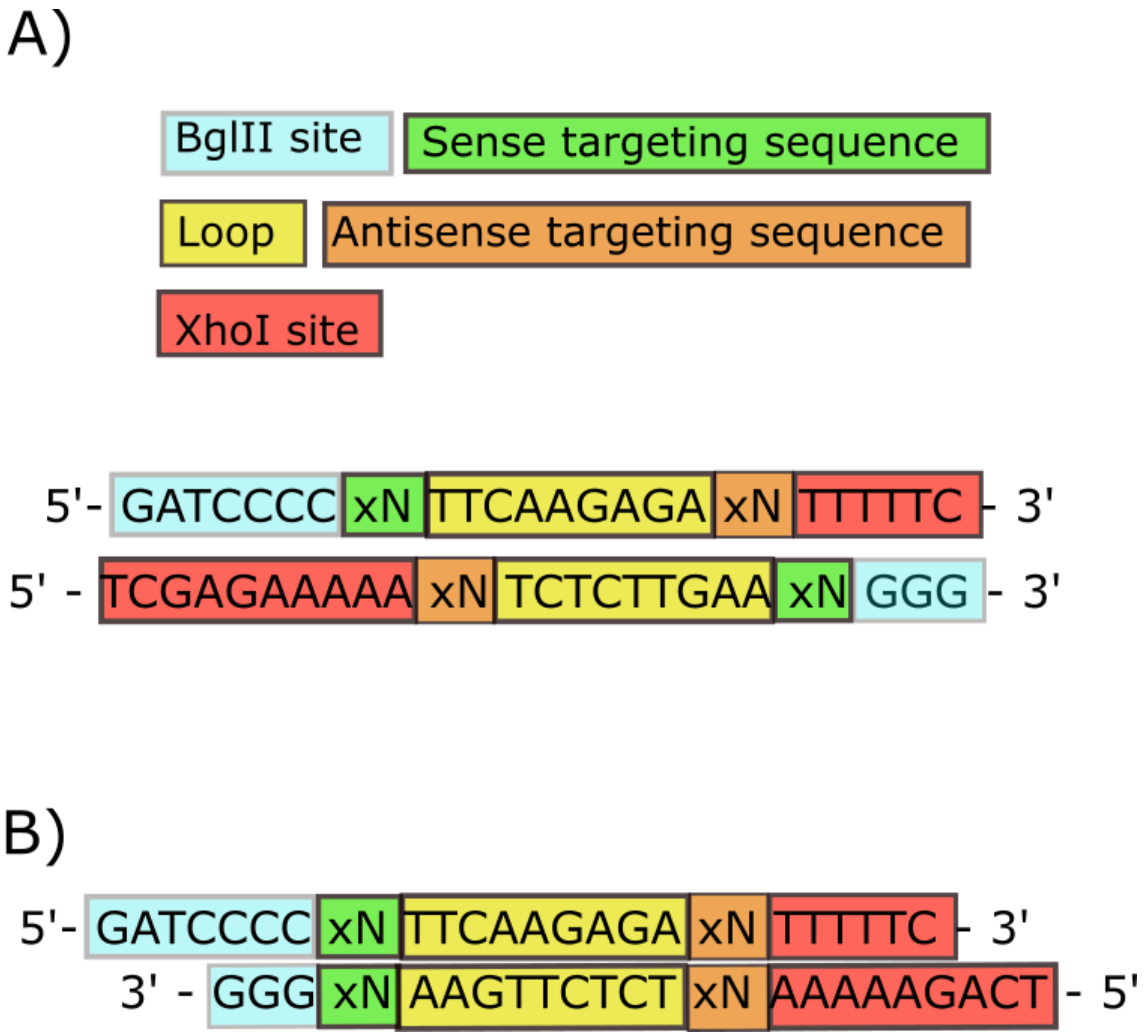



*Figure 2.3 Structure of the custom oligonucleotides used to produce alternative shRNA constructs using pSuperRetroPuro as backbone.*
*A: Structure of the oligonucleotides as designed before annealing, and legend. B: Oligonucleotides after annealing.*

## 2.7 Analysis

The initial processing steps of sequencing data were performed using pipelines specific for each project and therefore they will be explained in their respective chapters. Later steps of the analysis were common to both projects and are explained below. This analysis process is graphically summarized in Figure 2.4.

### 2.7.1 Building gene networks

Gene networks were built for the genes selected as genes of interest (GOI) using FunCoup (http://funcoup.sbc.su.se/) (Schmitt, Ogris et al. 2014). Default parameters were used except for the expansion depth, under “Advanced search”. This was set to 0 so FunCoup would produce networks using only the GOI list. Network files were downloaded and imported into Cytoscape 3.3.0 for visualization and analysis of gene interactions.

### 2.7.2 Analysis of gene interactions

Gene networks were analysed using Cytoscape (Tools>>NetworkAnalyzer>>Network Analysis>>Analyze Network). The analysed files were exported and opened using Microsoft Excel. A text file was exported from Excel containing gene names and number of interactors in the queried genes.

### 2.7.3 Enrichment in gene interactors

The FunCoup compact database of interactions in *Mus musculus* was downloaded from the FunCoup website (http://funcoup.sbc.su.se/downloads/, downloaded file: FC3.0_M.musculus_compact.gz). The database file was extracted and opened using TextPad. Interactions were filtered by confidence threshold, producing a smaller file with interactions above 0.8 confidence score (referred to as “interaction file” from now). A Python script (Appendix 6) was written to produce 1000 random samples of constructs that included the gene to be analysed for enrichment of interactors. The script then used the interaction file to search for interactors in each sample produced and outputted a file with each gene name and the number of random samples where the number of interactors was the same or above the number of interactors for the same gene in the GOI list. This number divided by 10 to reflect percentages and was used as a p-value for enrichment in interactors.

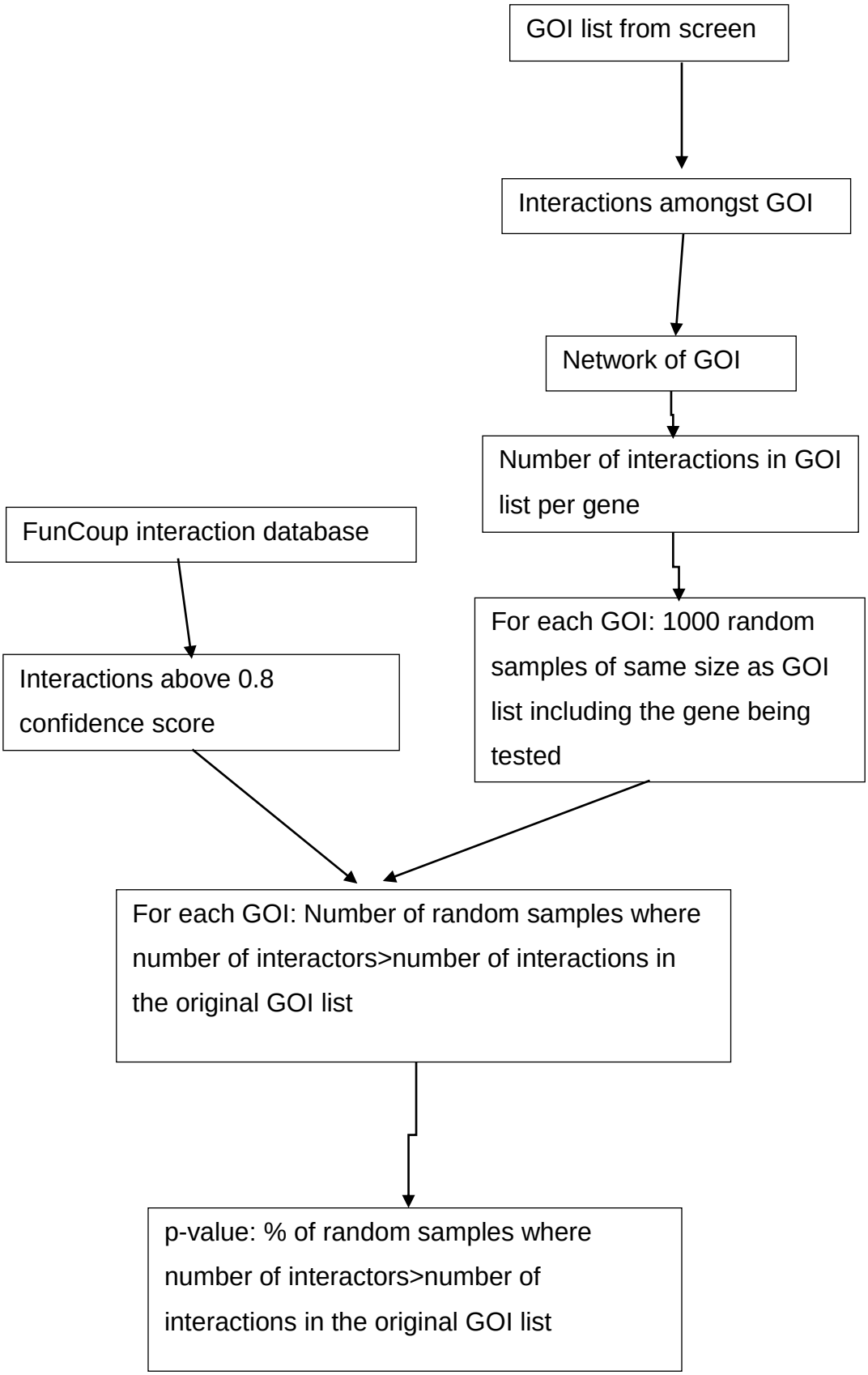



*Figure 2.4. Schematic diagram of the random sampling approach to detect enrichment in interactors using data from FunCoup.*

# Chapter 3. A pooled genome-wide shRNA screen of neighbour suppression effectors in tumorigenic cells

## 3.1. Introduction

### 3.1.1. Cancer

Multicellular organisms can be viewed as ecosystems where a balance in cell number is required to maintain the body's structure and order. This is achieved through cell proliferation, death and survival (Pienta, McGregor et al. 2008). The initiation of cancer may be defined in base of this view as a perturbation of the balance by some cells. These tumorigenic cells acquire mutations which can confer them with certain advantages which result in a higher rate of proliferation or lower rate of cell death (Pienta, McGregor et al. 2008). This over-simplistic view does not fully explain the nature of cancer but outlines the concept of tumorigenesis as a process that starts with a cell surrounded by other cells of the same genetic background. This genetic background changes in the tumorigenic cell while it is still in a homotypic environment.

Although the incidence of cancer may seem high given the number of new cases diagnosed each year (352,197 in the UK in 2013 according to Cancer Research UK (Cancer Research UK, http://www.cancerresearchuk.org/health-professional/cancer-statistics, Accessed June 2016)), the prevalence in the worldwide population of all types of cancer combined is 0.2% (Ferlay, I. et al. 2012). Considering the exposure to mutagenic stimuli, the prevalence of spontaneous mutations and the number of cells in the average human body, cancer is not as prevalent as expected (Seshadri, Kutlaca et al. 1987, Araten, Golde et al. 2005). Furthermore it has been observed that the occurrence of cancer in animals is not directly predicted by the number of cells they are formed by and their lifespans, which points to discrepancies in the resistance to carcinogenesis (Cairns 1975, Peto, Roe et al. 1975).

### 3.1.2. Mechanisms of resistance to carcinogenesis

Many intra- and inter-cellular mechanisms have been described that can result in a lower incidence of carcinogenesis than expected. The intra-cellular mechanisms of resistance to carcinogenesis include genetic and epi-genetic aspects and regulation of tumorigenic cell death. These intra-cellular

mechanisms will not be discussed in this introduction as the focus of this thesis lies in a type of inter-cellular carcinogenesis suppression.

Of the inter-cellular mechanisms of resistance against carcinogenesis, inmmunosurveillance is perhaps the most commonly studied. The focus of this chapter, however, is on neighbour suppression, a mechanism of growth suppression of tumorigenic cells by normal cells which form part of the same tissue.

### 3.1.3. Neighbour suppression

Berenblum noted in 1941 that the development of tumours in mice skin following the application of a cancer promoter substance was uncommon unless a pre-treatment with a cancer initiator was applied (Berenblum 1941). A hypothesis was formed that carcinogenesis was a two-stage process. This characteristic may explain the presence of genotypically abnormal and even premalignant epithelial mammary cells in half of American healthy women with no elevated risk of developing breast cancer (Holst, Nuovo et al. 2003). The prevalence of breast cancer in American women is 12%, which indicates that those abnormal cells do not always acquire a carcinogenic phenotype. Numerous *in vivo* studies have been performed which suggest this phenomenon is stronger in early development (Dolberg, Hollingsworth et al. 1985) but is present in a range of tissues (Medina, Shepherd et al. 1978, Terzaghi-Howe 1987, Strickland, Ueda et al. 1992, Widschwendter, Berger et al. 1997, Javaherian, Vaccariello et al. 1998, Roberts, Min et al. 2002, Holst, Nuovo et al. 2003).

Despite the apparent involvement of ’normal’ cells in halting the growth of precancerous cells, the molecular mechanism of the phenomenon remains elusive. Early *in vitro* work performed by Rubin showed that a high proportion of cultured cells needed to be infected with an oncogenic virus (Rous Sarcoma Virus or RSV) for the infected cells to display the phenotype associated with RSV infection (Rubin 1960). A similar phenomenon was described by Stoker, Todaro and Green also in the 1960’s. Using polyoma virus-transformed mouse fibroblast as a model for carcinogenic cells, they showed that their growth could be suppressed by a monolayer of density-dependent growth inhibited mouse fibroblasts by a mechanism dependent on direct contact between both types of cells (Stoker 1964, Stoker, Shearer et al. 1966, Todaro and Green 1966).

Fibroblasts transformed by Myc but not Src and Ras have also been shown to be suppressed by non-transformed counterparts (Bignami, Rosa et al. 1988) while suppression is seen in myoblasts transformed with Myc and Src (La Rocca, Grossi et al. 1989).

The role of neighbouring cells in controlling cell growth has been studied under different terms such as “contact inhibition” or “postconfluence inhibition of cell division” (Martz and Steinberg 1972). This inhibition of cell growth by surrounding cells was experimentally observed as a plateau in growth rate and a reversible cell cycle arrest in G1 phase in monolayers of fibroblasts (Eagle and Levine 1967, Rubin 1971, Deleu, Fuks et al. 1998). The loss of contact inhibition in tumour cells is considered a hallmark of cancer (Hanahan and Weinberg 2011) and has been used to produce a prognosis in cancer development and as a tool to identify oncogenes (Bargmann, Hung et al. 1986) (Fuse, Tanikawa et al. 2000). Contact inhibition has been observed and described in the context of homologous cells, i.e. cells of the same type or genetic background. However, in the context of cancer initiation tumourigenic cells are not yet surrounded by other tumourigenic cells but rather by “normal” cells. Therefore, the mechanism of growth suppression of these tumourigenic cells by their surrounding cells may be considered different from the mechanism of contact inhibition. The distinction of these two forms of growth inhibition will be possible once the molecular mechanism for both phenomena are identified.

This phenomenon of carcinogenic cell growth inhibition has been demonstrated using human cancer cells as opposed to virus-transformed cell lines. For example, Flaberg and collaborators reported the inhibitory effect of primary human fibroblasts on the growth of human cancer cell lines (Flaberg, Markasz et al. 2011). They found that variation in the extent of this suppressive behaviour was at least partially explained by the age of the fibroblast donor as well as the site of origin of the fibroblast explant.

A few *in vivo* studies have also highlighted a mechanism of tumour growth suppression by surrounding cells. An early example is the work of Illmensee (Illmensee and Mintz 1976), who found that when mouse teratocarcinoma cells were injected in mouse blastocysts these cells contributed to the development of various tissues in tumour-free mice.

- Cell-cell contact and gap junctions in neighbour suppression.

Neighbour suppression has been demonstrated to be dependent on contact or close proximity with growth inhibited cells (Stoker 1964) and not a result of nutrient depletion in the medium (Stoker 1967). This requirement for close proximity lead to the study of cell communication between the suppressive and the suppressed cells. Gap junctions were shown to be essential for this type of growth suppression (Mehta, Bertram et al. 1986). It was hypothesized that the suppressive cells may send growth inhibitory signals to the suppressed cells using gap junctions, and that conversely the suppressed cells may release growth-promoting signals to the neighbouring not transformed cells. However, almost 20 years after this study, Alexander and collaborators showed using cells derived from connexin knockout mice that, although direct cell contact between transformed and non-transformed cells was required for neighbour suppression, this phenomenon was not dependent on the formation of gap junctions (Alexander, Ichikawa et al. 2004).

The aim of this chapter has been to design, optimize, execute and analyse a non-biased RNAi screen to find molecular effectors of neighbour suppression in the tumorigenic cells. It was hoped that molecules could be identified in the tumourigenic cells that were required for the growth suppression by non-tumourigenic neighbours.

## 3.2. Material and methods

This section introduces methods that are specific to this chapter. All materials used are listed in Chapter 2. Generic methods are referenced back to the relevant section in this thesis when needed.

### 3.2.1. Cell lines

**3T3$^{supp-}$ and 3T3$^{supp+}$**. Isogenic derivatives of early passage Swiss 3T3 mouse embryonic fibroblasts (Allard, Stoker et al. 2003), (Todaro and Green 1963).

**SVE**. Derivative of the 3T3$^{supp+}$ cell line, transformed using Simian virus 40 (SV40) (Eddy, Borman et al. 1962).

**SVE-6**. Derivative of the SVE cell line (Nicholls 2015). Expresses high levels of cytoplasmic green fluorescent protein (eGFP) from a lentivirus expression construct (Marr, Guan et al. 2004).

These 4 cell lines were routinely grown in DMEM (high glucose, GlutaMAX, pyruvate) supplemented with 10% (v/v) NCS and 1% (v/v) Penicillin Streptomycin.

### 3.2.2. Production of retroviral particles using Phoenix cells

Phoenix cells were trypsinized and seeded 24 hours before transfection in a 100 X 20 mm sterile cell culture dish so that they were 70% confluent on the day of transfection. Cells were transfected using Turbofect as follows: 10 µg of the plasmid to be transfected were diluted into 1 mL of Opti-MEM Reduced Serum Medium in a sterile 1.5 mL microcentrifuge tube and mixed by vortexing. A vial of Turbofect was vortexed briefly and 40 µl were added to the DNA solution. The mixture was incubated for 20 minutes at room temperature. The growth medium from the Phoenix cell dish was aspirated and replaced with 3 mL of fresh growth medium. The Turbofect mix was then added dropwise into the dish. To ensure even distribution of the mix, the dish was gently rocked and returned to the incubator. After 24 hours of incubation the transfection medium was aspirated and cells were very gently washed once with DPBS. 12 mL of growth medium were added to the dish before returning it to the incubator. After 24 hours the medium was removed and transferred to a 15 mL tube. Phoenix

cells and cell debris in the tube were pelleted, and the supernatant containing viral particles was transferred to another 15 mL tube. Polybrene was added to a final concentration of 8 µg/ml.

### 3.2.3. Transduction of SVE-6 cells

SVE-6 cells were seeded 24 hours before transduction in a 100 X 20 mm sterile cell culture dish at a density that would give approximately 50% confluence. To perform the transduction, the growth medium in the dish was removed and replaced with 10 mL of viral medium containing polybrene. The dish was returned to the incubator for 24 hours, after which the viral medium was aspirated and replaced with 10 mL of viral growth medium. Cells were returned to the incubator and grown for 24 hours before puromycin selection.

### 3.2.4. Seeding monolayers

$3T3^{supp+}$, $3T3^{supp-}$ and SVE cells were trypsinized and seeded at approximately 25% confluency. Cells were grown under standard conditions and monitored daily. When they had reached confluency (i.e., reached 100% density) they were cultured for another 24 hours before considered to have formed a monolayer of cells and used for further experiments.

### 3.2.5. Plating SVE-6 cells on monolayers and plastic

Monolayers of $3T3^{supp+}$, $3T3^{supp-}$ or SVE cells were prepared as per section 3.2.4. To plate SVE-6 on plastic, the appropriate volume of growth medium was added to the same type of tissue culture dish, flask or multi-well plate used to plate the monolayers. SVE-6 cells were trypsinized and counted as described in section 2.1.7. Growth medium was added to the cell suspension to adjust the concentration of cells so that the number of cells to be plated was contained in at least 20µl. The appropriate volume of cell suspension was pipetted on the monolayer and plastic dishes and they were gently swirled to distribute the cells as homogenously as possible.

### 3.2.6. Restriction enzyme digestion of genomic DNA

Genomic DNA was sheared with a 26G needle. The DNA solution was passed through the needle 20 times or until the viscosity of the solution had been reduced enough to allow accurate pipetting. Genomic DNA was then fragmented using EcoRI and XhoI. Digestions were set up as in Table 3.1 and

incubated overnight. A higher amount of DNA was used on the 3T3$^{supp+}$ digestions due to a lower ratio of target sequences to genomic DNA in these co-cultures. Loading buffer was added to each digestion reaction for a final concentration of 1X. The digested DNA was separated by electrophoresis in 1.5% agarose gels prepared as explained in section 2.2.5. A band around the size of the shRNA construct was extracted as in section 2.2.7

| | **Plastic** | **3T3$^{supp-}$** | **3T3E$^{supp+}$** |
|---|---|---|---|
| **Needle-sheared DNA** | 100 µl | 100 µl | 500 µl |
| **EcoRI** | 2 µl | 3 µl | 5 µl |
| **XhoI** | 2 µl | 3 µl | 5 µl |
| **Buffer** | 22 µl | 22 µl | 110 µl |
| **$H_2O$** | 94 µl | 92 µl | 480 µl |

*Table 3.1. Composition of a single genomic DNA restriction digestion reaction for each sample in the screen.*

### 3.2.7. PCR amplification and purification of shRNA constructs

ShRNA sequences were PCR amplified using Q5 High fidelity polymerase as explained in section 2.2.4 using the primers in Table 3.2 and conditions in Table 3.3. PCR reactions were mixed with loading buffer at a final 1X concentration and separated in a 1.5% agarose gel prepared as in section 2.2.5. Electrophoresis was run and bands extracted as explained in section 2.2.7.

| Primer name | Sequence | Sample |
|---|---|---|
| **HiSeq 1 forward** | TCGAGGTCGACGGTATCGATA | 3T3$^{supp-}$ |
| **HiSeq 2 forward** | ATCTCTCGAGGTCGACGGTA | 3T3$^{supp+}$ |
| **HiSeq 4 forward** | GAGGTCGACGGTATCGAT | Plastic |
| **HiSeq reverse** | TGTCTTTGGATTTGGGAATC | Common to all samples |

*Table 3.2. Primers used to PCR amplify the shRNA constructs in each of the three samples*

| Step | Temperature | Time |
|---|---|---|
| **Initial denaturation** | 98ºC | 30 seconds |
| **28 cycles** | 98ºC<br>52ºC<br>72ºC | 5 seconds<br>15 seconds<br>15 seconds |
| **Final extension** | 72ºC | 2 minutes |
| **Hold** | 4ºC | |

*Table 3.3. PCR cycling conditions used for amplification of shRNA sequences.*

### 3.2.8. Analysis of sequencing data

Figure 3.1 depicts the definitive analysis pipeline. The sequencing output was received as .fastq files. These files were moved to a UNIX server. Sequencing reads containing the primer sequence used for the amplification of each sample were extracted from the .fastq files using the command “grep” in the server (see format of the command in Appendix 1). Custom Python scripts (in Appendix 2) were written to extract the shRNA constructs from the files and build a plain text file which contained the shRNA constructs present in the sequencing output and the number of times that each shRNA construct was present. Briefly, the shRNA sequences were detected by locating the shRNA loop sequence in the sequencing reads. If the loop was present in a read, the 19 letters before the start of the loop were considered to be a potential shRNA construct. Then, the script counted how many times each construct appeared and outputted the shRNA sequence and the number of reads found in the sequencing. These text files were merged using Stata with a plain text file containing the shRNA constructs that were designed to be in the shRNA library. Constructs with less than 50 reads in the control sample (Plastic) were counted as absent to avoid random fluctuations in subsequent analysis steps.

Reads were converted to reads per million per pool (rpmp), and two ratios were calculated: (rpmp 3T3$^{supp+}$)/(rpmp Plastic) and (rpmp 3T3$^{supp-}$)/(rpmp Plastic). These ratios were used to calculate a robust z-score per pool (see Stata do-file in Appendix 3 for particulars of calculations).

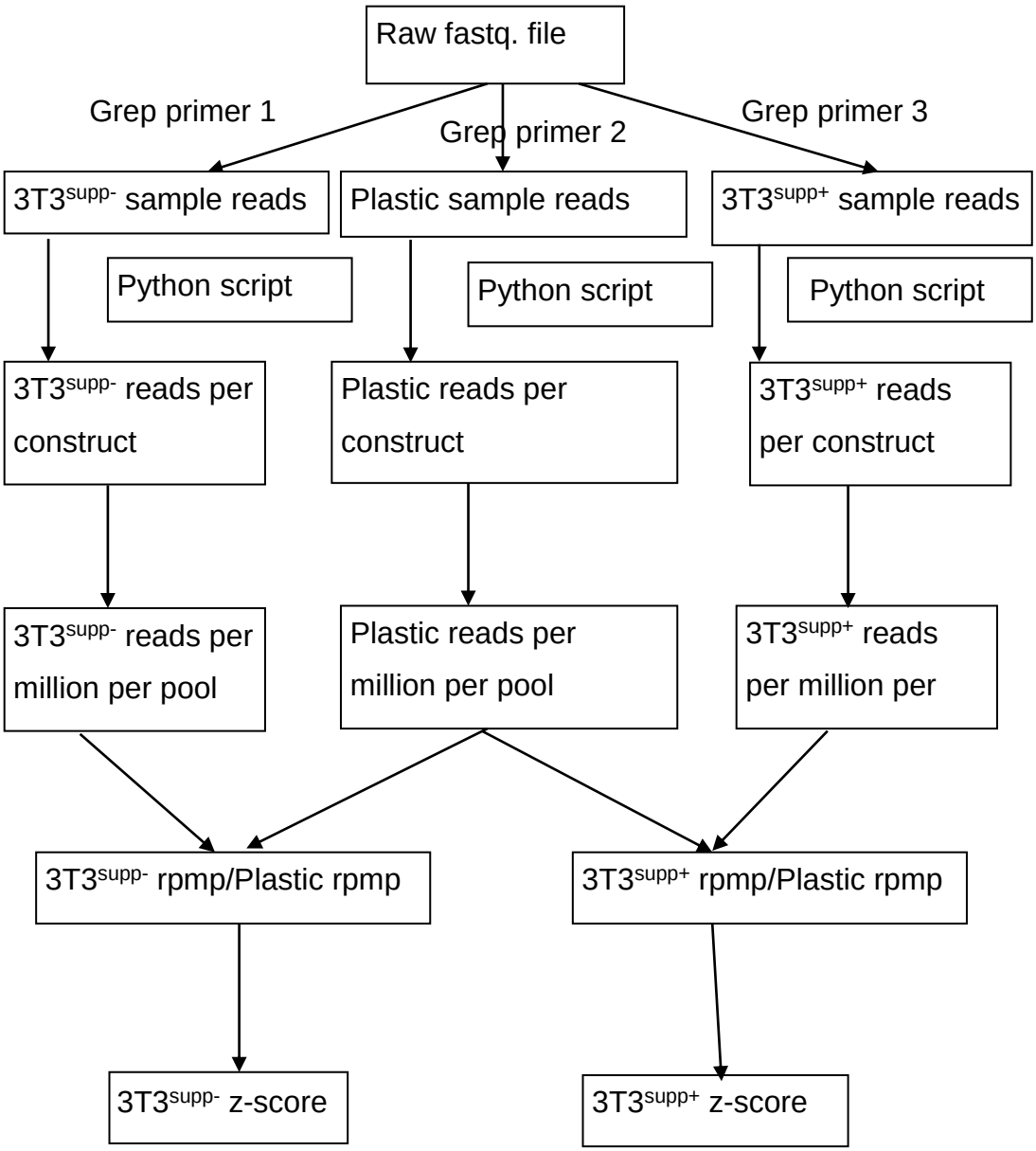



*Figure 3.1. Schematic depiction of the analysis pipeline used.*
*Sequencing reads were extracted from the .fastq files in a UNIX server using the command "grep". Reads were aligned to the shRNA library and counted to produce a file containing the number of reads per shRNA construct. These raw number of reads were converted to reads per million per library pool. A ratio of each experimental condition to the growth control (Plastic, "P") was calculated and used to build robust z-scores as estimates of shRNA construct effect sizes.*

### 3.2.9. Recursive partitioning analysis of tumour expression data

Microarray expression data from various types of tumour were obtained from Dr Sarah Vowler, who also provided the R scripts to perform the recursive partitioning analysis. This analysis finds genes whose expression are correlated with survival rates of patients and is based on a method described first by Breiman (Breiman, Friedman et al. 1984). Distant metastasis free survival was used as the dependent variable as neighbour suppression is thought to be relevant for the initial phase of tumour growth. Although neighbour suppression is not amenable to study *in vivo* when it is expected to be taken place (i.e., when tumorigenic cells are still not forming tumours but either as single or few cells), the closest stage is represented by the establishment of metastasis. Metastasised cells need to settle in a new environment and grow. However, the surrounding tissue may exert a neighbour suppression effect on them, delaying or impeding the appearance of distant metastasis.

The different datasets used are explained in Table 3.4. The detection of a high number of genes whose expression correlates with distant metastasis in the datasets used cannot be directly interpreted without evaluating whether this number of genes is unexpected. For this, an approach was designed to calculate the frequency at which the same or higher number of genes in a random group of genes were detected as being correlated with distant metastasis.

Enrichment for significant effects in the list of GOI was tested using a random sampling approach (see Appendix 4 for the Python script to perform this analysis). Briefly, the entire shRNA library was fed to the recursive partitioning R script and then random samples of genes of the same size as the GOI list were picked. The output consisted on the number of significant effects found in each sample for each dataset.

| Dataset name | Type of tumour | Sample size | Publication |
|---|---|---|---|
| **Taylor** | Prostate cancer | 218 | (Taylor, Schultz et al. 2010) |
| **Shedden** | Lung adenocarcinoma | 442 | (Shedden, Taylor et al. 2008) |
| **Loi** | Breast cancer | 335 | (Loi, Haibe-Kains et al. 2007) |
| **Desmedt** | Breast cancer | 198 | (Desmedt, Piette et al. 2007) |

*Table 3.4. Human tumour expression datasets used for the recursive partitioning analysis of effects of gene expression on metastasis-free survival of cancer patients.*

### 3.2.10. Gene set enrichment analysis

Data were analysed for enrichment in gene set members using GSEA PreRanked. shRNA constructs were ranked using different metrics: $3T3^{supp+}$ z-score, $3T3^{supp-}$ z-score or difference of z-scores (($3T3^{supp-}$ z-score)-($3T3^{supp+}$ z-score)). Duplicate gene names were resolved by removing duplicates in Excel. This was done after sorting the constructs on z-score (high to low or low to high depending on the sample, further details are explained in 3.3.11.5), absolute z-score or alphabetically. Each combination of ranking and duplicate removal approaches was tried and analysed separately. The following parameters were used in all analyses:

- Gene set database: “c2.all.v5.1.symbols.gmt”.
- Collapse probes into gene names: False.
- Scoring scheme: Classic.
- Minimum size of sets to analyse: 15.
- Maximum size of sets to analyse: 500.
- Number of permutations: 1000.
- Seed for randomisation: timestamp.

### 3.2.11. Design of the screening process

The screening process was designed to uncover molecular effectors of neighbour suppression in tumorigenic cells. More details of the screening will be given in the Results section of this chapter (section 3.3.1), but a simplified schematic depiction of the screening process is shown below (Figure 3.2).

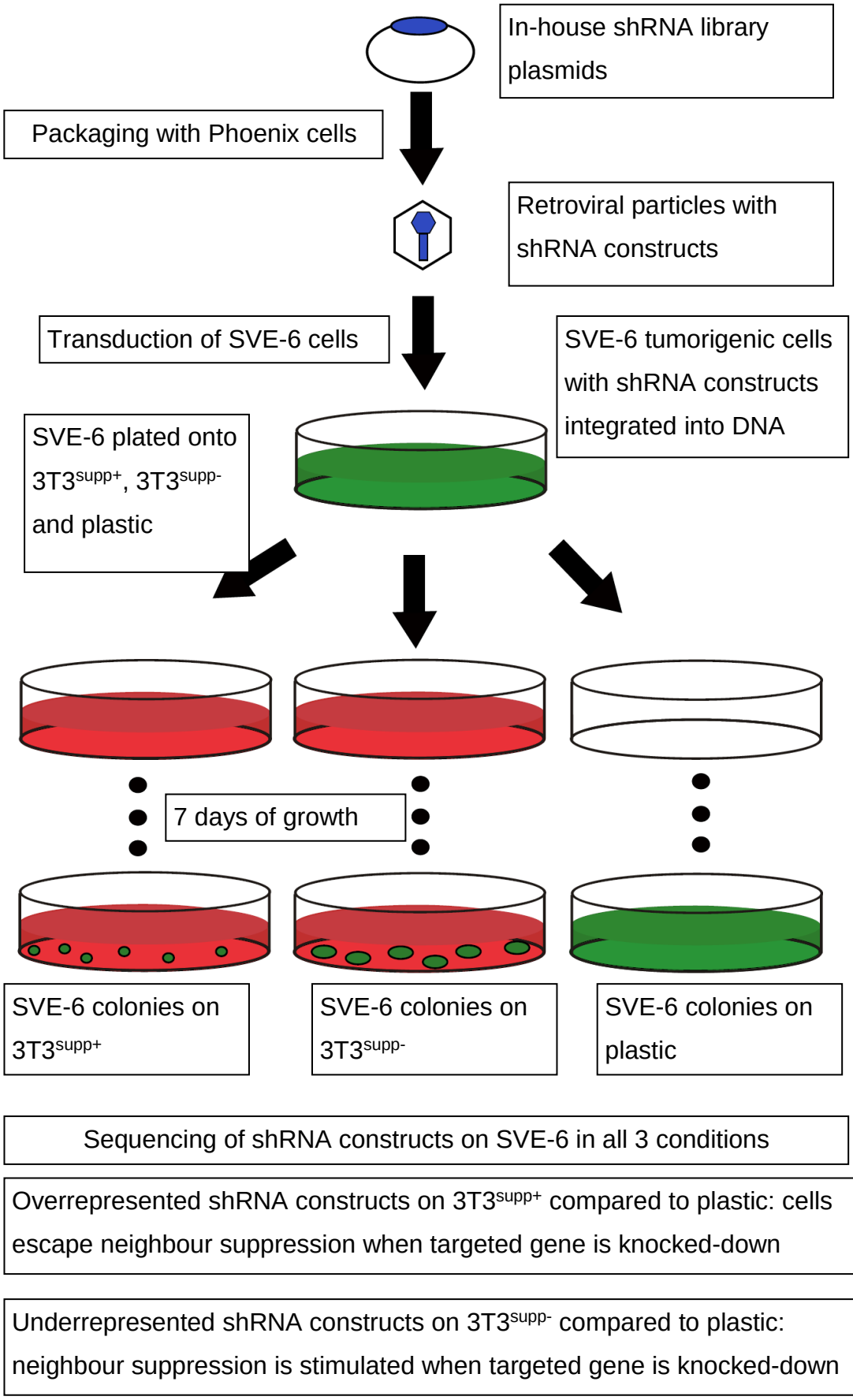



*Figure 3.2. Schematic simplified overview of the screening process*

## 3.3. Results

### 3.3.1. Description of the screening process

$3T3^{supp-}$ and $3T3^{supp+}$ are two isogenic cell lines derived from mouse embryonic fibroblasts that differ in their ability to suppress the growth of oncogene-transformed derivatives in co-culture. Such effect is depicted in Figure 3.3, where SV40-transformed $3T3^{supp+}$ expressing GFP (SVE-6 cells) were added on top of confluent monolayers of either $3T3^{supp-}$ or $3T3^{supp+}$ cells and cultured for seven days. The aim of the screening process was to detect genes which upon reduction of expression in the tumorigenic cells would either allow the growth of these cells in suppressive co-cultures (co-culture with $3T3^{supp+}$ cells) or suppress growth in non-suppressing co-cultures (co-culture with $3T3^{supp-}$).

A schematic depiction of the screening process is represented in Figure 3.2. The in-house NKI mouse shRNA library was transfected into a packaging cell line (Phoenix ECO), which produced ecotropic (murine-specific) retroviral particles containing the shRNA targeting constructs. One 10 cm dish of Phoenix cells was used per library pool. Tumorigenic (SVE-6) cells were transduced with the retroviral particles and selected for stable integration using puromycin. One 10 cm dish of SVE-6 cells was used per library pool. Monolayers of $3T3^{supp+}$ and $3T3^{supp-}$ were established in T175 and T75 flasks as explained in section 3.2.4. An average of 100 cells per shRNA construct (38400 cells per library pool) were screened. Puromycin selected SVE-6 cells were added on top of confluent $3T3^{supp+}$ and $3T3^{supp-}$ monolayers and to the same area of tissue culture plastic. Co-cultures were grown for seven days, with growth medium being renewed every three days. After seven days, cells were harvested and genomic DNA extracted. As cells were harvested together, genomic DNA from both the monolayer and the SVE-6 cells was extracted.

### 3.3.2. Overview of the NKI shRNA library used

The NKI shRNA library, described in section 2.4 was amplified in-house as previously described (Nicholls 2015). Using the data obtained by Nicholls, a file was constructed that contained the shRNA constructs present in the in-house shRNA library, and the pool in which they were expected to be as indicated by Dr Roderick Beijersbergen (NKI). This file also contained the name of the

transcripts that were expected to be targeted by the sequences in the library. However, this annotation was not been performed with the latest release of the mouse transcriptome reference. To update the annotation provided, the shRNA targeting sequences in this library file were re-aligned to the mouse transcriptome using a standalone version of BLAST (section 2.5).

Not all sequences were found to have only one match: 3,367 sequences were found to be duplicated. This does not mean that 3,367 targeting sequences aligned to two RefSeq mRNA sequences as some of those might be the same sequence which has matched to more than two mRNAs. Of the matches produced, 2,185 (of 33,547) were not present in the original library design and 25,163 had an e-value as calculated by BLAST of 0.004. The e-value is a measure of the probability of finding another mRNA sequence with the same degree of match to the sequence blasted. E-values of less than 0.06 were assigned to 25,655 sequences in total, with the rest of them (7,892) presenting an e-value ranging from 0.23 to 866. Of these 7,892 sequences with a high e-value 1,991 were found to be duplicated and 2,056 were not in the original library design.

### 3.3.3. Optimization of multiplicity of infection (MOI)

It was important for this screen that the resolving power was as high as possible to detect moderate effects of knock-downs. If the tumorigenic cells were transduced with more than one plasmid, the effects of a functional plasmid could be dampened by the effects of the others present in the same cell. Also, it is possible that the knock-down of more than one gene simultaneously would produce an effect that the knock-down of each of them separately would not exert. This would result in false positives as the effect would be detected for all the genes knocked-down equally and it would not be possible to resolve their individual effects. As viral infections are binomial phenomena (a viral particle either infects or not a cell), their behaviour is modelled by the Poisson distribution (Figure 3.3). The probability of a multiple transduction decreases with the MOI, but if the MOI is too low the efficiency of the experiment also decreases as it produces a very low number of transduced cells. To satisfy both conditions a MOI of 0.3 was chosen. As seen in Figure 3.3, when cells are transduced at a 0.3 MOI, most (22% of total cells or 73% of the infected cells) are transduced with a single construct. This produced enough infected cells for

a library pool in a 10 cm dish without too many cells containing multiple sequences. This multiple sequence integration issue is a common limitation in pooled RNAi screens and requires a balance between optimal conditions for the experiment (i.e. it would be optimal to eliminate the possibility of multiple sequences inside the cells) and efficiency (i.e. to eliminate this possibility one would need to perform the transfections with a lower MOI, which increases the number of cells needed).

To optimize the conditions in order to obtain the target 0.3 MOI, PHOENIX cells were transfected with a Turbofect/DNA mixture (as in section 3.2.2) but with varying ratios of Turbofect/DNA. The retroviral supernatant harvested from these various conditions was used to transduce SVE-6 cells as in 3.2.3. Cells were puromycin selected and the survival rate examined. The conditions which produced a cell survival of approximately 30% were chosen as the ideal conditions for the screen and maintained throughout the screening process. As pools were transfected and transduced separately they were monitored for consistency. If any pool showed a survival rate substantially higher or lower than approximately 30%, those SVE-6 cells were discarded and not used in the screen.

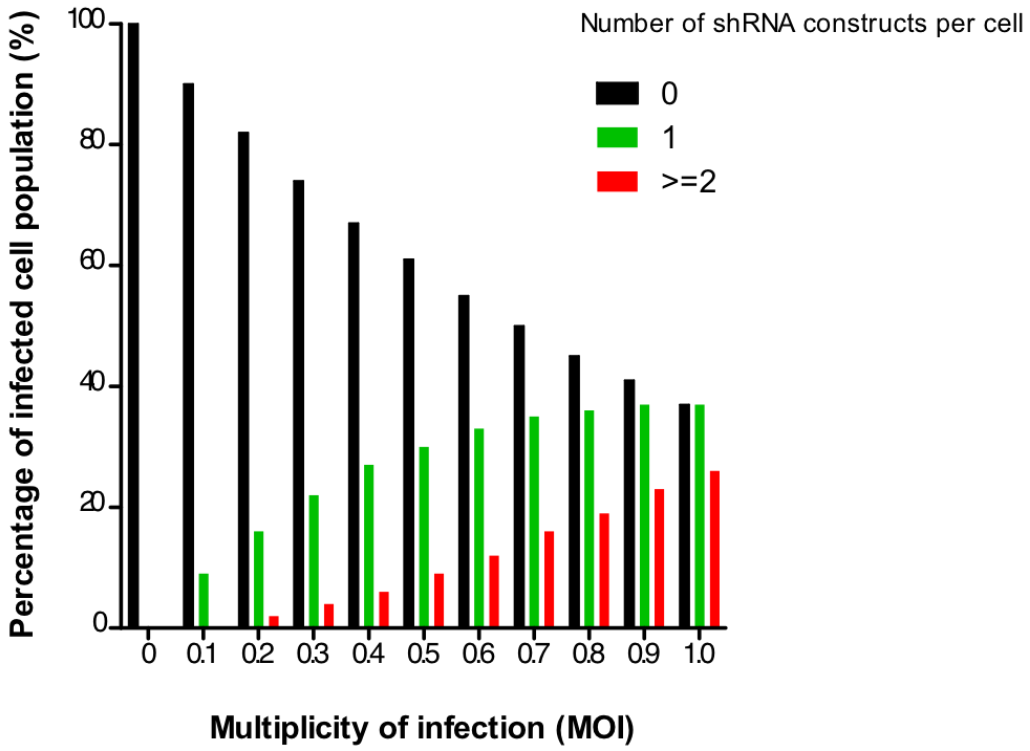



*Figure 3.3. Distribution of null (0, black bars), single (1, green bar) and multiple (2 or more, red bar) transduction events in the cell population for different MOI.*

### 3.3.4. Optimization of co-cultures

To maximize the resolution of the screen it was critical to establish the co-culture conditions that would produce a maximum difference in growth of transformed cells in suppressive and non-suppressive conditions. Two parameters were considered of most importance: seeding density of transformed cells and length of co-culture. Seeding transformed cells in too close proximity could result in their growth being impeded by space limitations, which in co-cultures with $3T3^{supp-}$ would cause a decrease in SVE-6 colony size. Modest decreases in colony size would not be detected as all colonies would be smaller than that achieved by unhindered growth. Seeding cells too sparsely could affect the plating efficiency but of more concern was the practical limitation from increasing the scale of the screen.

To optimize the co-culturing conditions, monolayers of $3T3^{supp+}$ and $3T3^{supp-}$ were seeded (section 3.2.4). SVE-6 cells were plated at different densities on top of these monolayers and directly on plastic (section 3.2.5). Growth of SVE-6 cells was monitored using a fluorescence microscope every 24 hours for up to ten days. Although the difference in SVE-6 growth was maximal after ten days of co-culture, it was noted that monolayers started to detach from the tissue culture plastic surface, which could confound the results as SVE-6 could start growing on plastic instead of on monolayer. Seven days of co-culture provided the maximum resolution without causing the underlying monolayer to detach. It was determined by examination of the cell colonies that 166 SVE-6 cells per $cm^2$ could be plated without their growth being limited by the proximity of other colonies (Figure 3.4).

Having chosen the appropriate seeding density, it was important to calculate how many cells per construct would be necessary to minimize a potential bias due to sampling effects. As the effect on reduction of expression for each targeting sequence was unknown and therefore power calculations could not be performed, this parameter was researched in the available literature. Most RNAi screens have been performed using a dropout protocol, where knockdown of genes are expected to increase the survival of cells to a certain stimulus. For this type of dropout screen, literature suggests a 1,000 fold representation. The screen being described here however was based on cellular proliferation and

therefore the number of cells per hairpin was expected to increase. In a dropout screen, the number of cells required per hairpin to maintain representation of the original library is higher due to the number of cells per hairpin decreasing upon exposure to the stimulus causing the dropout effect. Chapter 4 will present a dropout screen where 1,000 cells per hairpin were used.

The effect of shRNA hairpin fold representation in pooled screens was examined by Strezoska (Strezoska, Licon et al. 2012) who concluded that although using 500 cells per hairpin sequence resulted in a higher level of reproducibility between biological replicates, using a 100 fold representation still resulted in highly reproducible results: 65% of GOI overlap using 100 cells per hairpin compared to 73% using 500 cells per hairpin. Considering the trade-off between reproducibility and feasibility in this type of screen, 100 cells per hairpin were used.

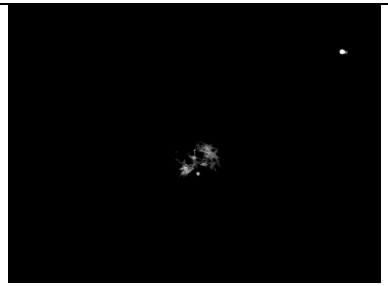
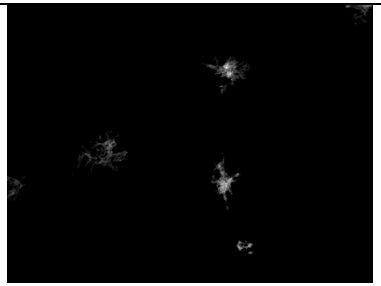
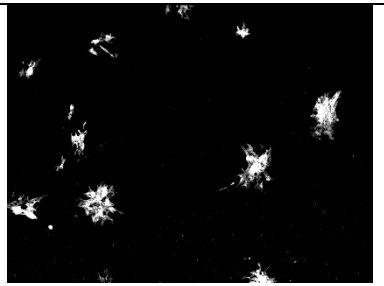
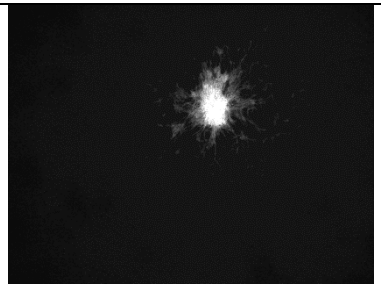
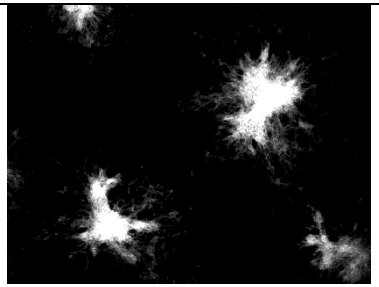
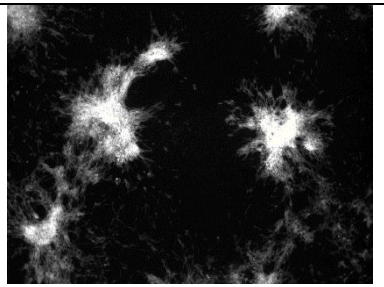
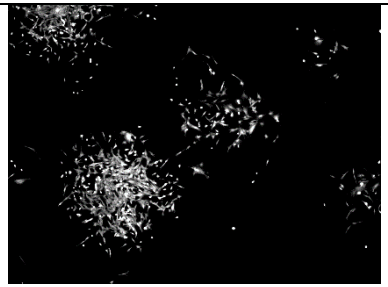
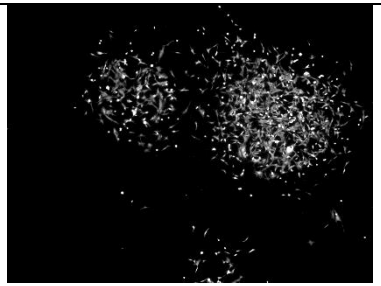
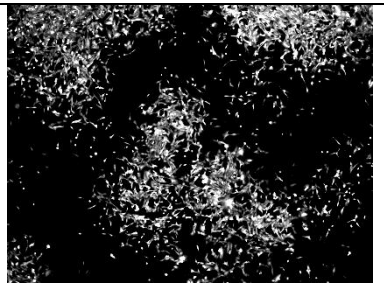
| | 41.5 cells/cm$^2$ | 83 cells/cm$^2$ | 166 cells/cm$^2$ |
|---|---|---|---|
| 3T3$^{supp+}$ |  |  |  |
| 3T3$^{supp-}$ |  |  |  |
| Plastic |  |  |  |

*Figure 3.4. Fluorescence microscopy photographs showing the SVE-6 colony size on both co-cultures and on plastic.*
*The number of SVE-6 cells per surface unit plated in each experiment is indicated on the top row of the table. SVE-6 cells were plated on monolayers or plastic and cultured for 7 days. Cells were fixed and a fluorescence microscope was used with a 4X objective and a FITC filter to take the photographs.*

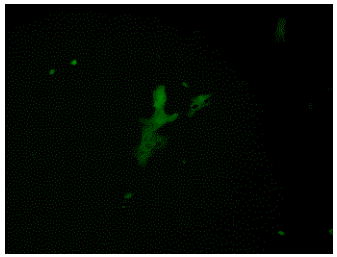
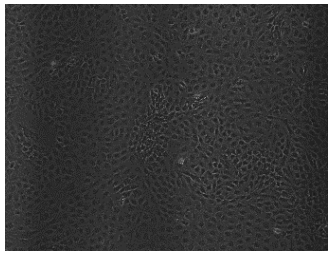
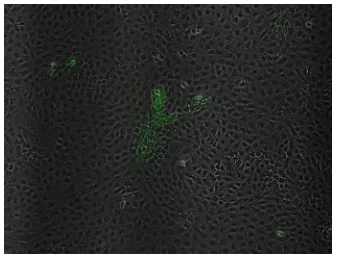
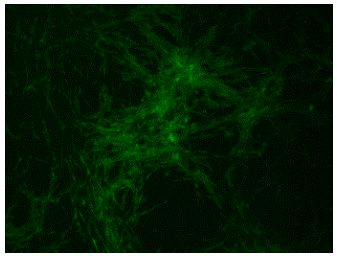
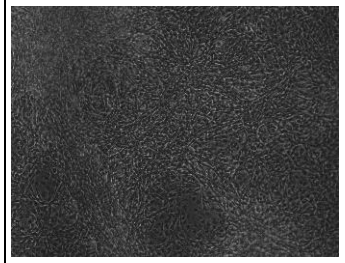
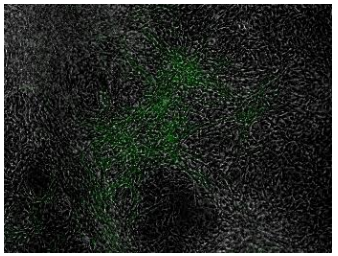
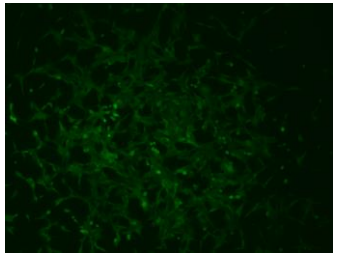
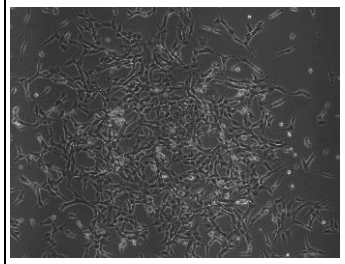
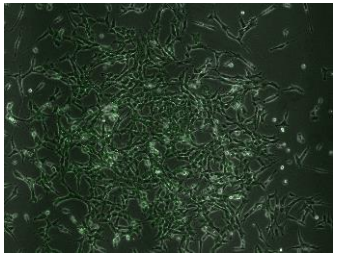
| Condition | GFP | Visible | Merge |
|---|---|---|---|
| 3T3$^{supp+}$ |  |  |  |
| 3T3$^{supp-}$ |  |  |  |
| Plastic |  |  |  |

*Figure 3.5. Neighbour suppression effect of 3T3$^{supp+}$ cells on SVE-6 cells.*
*Monolayers of 3T3$^{supp+}$ and 3T3$^{supp-}$ were set up and SVE-6 cells were plated on top of both monolayers and alone on plastic. Colonies of SVE-6 cells were imaged using fluorescence microscopy with a 10X objective after 7 days of growth. Monolayers were imaged using p hase contrast on the same microscope. GFP mages were artificially coloured and merged with monolayer images using ImageJ.*

### 3.3.5. Optimization of shRNA construct recovery from SVE-6 cells

The readout of the screen used second generation Illumina sequencing to assess the change on cell number as estimated by changes on representation of the shRNA constructs integrated in the tumorigenic cells genome. As cells were antibiotic selected, only those cells with a stably integrated shRNA sequence were used in the screening process. This provided in effect a molecular barcode such that sequencing through the targeting sequence integrated in the genomic DNA would be a proxy of the number of cells containing that targeting sequence. This barcoding strategy was first described in the context of RNAi pooled screens by Berns and collaborators (Berns, Hijmans et al. 2004). To use this barcoding approach, targeting sequences needed to be recovered from the cells, PCR amplified and sequenced. Two steps were optimized in the recovery of the shRNA constructs from the tumorigenic cells: the preparation of DNA for amplification and the amplification of the shRNA constructs itself.

As the growth of tumorigenic cells is significantly different on $3T3^{supp+}$ and $3T3^{supp-}$ monolayers and on plastic, the proportion of shRNA constructs to genomic DNA varied considerably between the three culture conditions. While on plastic for each copy of genomic DNA there was approximately one shRNA construct, DNA samples from the co-cultures included genomic DNA from the underlying monolayers of $3T3^{supp+}$ and $3T3^{supp-}$, which increased the proportion of genomic DNA to shRNA construct. This was particularly prominent in the suppressive co-culture as the growth of tumorigenic cells is very limited. Therefore, to achieve the same amount of template DNA (shRNA targeting sequences) a variable amount of genomic DNA would need to be used in PCR amplification for each screening condition.

PCR amplification is known to be inhibited by the presence of high DNA concentration in the PCR reaction, which limited the amount of genomic DNA that could be used in each PCR reaction. A possible solution for this limitation is to simply reduce the DNA to be amplified. This would reduce the template DNA proportionally, which could lead to sampling effects that could alter the representation of targeting sequences. The literature suggests that to limit sampling effects an average of 1,000 copies of each shRNA construct should be PCR amplified (Sims, Mendes-Pereira et al. 2011, Strezoska, Licon et al.

2012), which would not be practical if the amount of DNA template per tube was to be reduced. An alternative solution was then explored: to maintain the amount of DNA to be amplified, but increase the number of PCR reactions performed such that the concentration of DNA in each reaction was in effect reduced. To achieve the amplification of the equivalent to 1,000 copies of each targeting sequenced it was estimated that it would be necessary to perform a minimum of 25 PCR reactions per pool for the $3T3^{supp-}$ and 100 per pool for the $3T3^{supp+}$ samples.

To circumvent the impracticality of reducing the input DNA using these approaches, solutions were explored to increase the template DNA to genomic DNA ratio. Targeting sequences in the library had been cloned into the pSuperRetroPuro vector using the restriction enzymes EcoRI and XhoI and their recognition sequence is maintained when they are integrated in the genomic DNA. It was therefore reasoned that it would be possible to excise a region of the viral DNA containing the targeting sequences from the genomic DNA. It was necessary that the recognition sequence for both enzymes was not included into any of the targeting sequences such they would not be cleaved by the enzymes. If this was the case, the cleaved targeting sequences would not be PCR amplified or sequenced. It was expected that the targeting sequences would not contain the recognition sequence for any of the two enzymes as the original library was designed to be cloned into the vector plasmid with these enzymes, but this was still investigated to be certain this step would not alter the composition of the samples. A file with the original sequences provided by the library providers was searched for the recognition sequence and no matches were found. Therefore, this digestion step was incorporated into the protocol.

The composition of the restriction digestion reactions were then optimized with respect to the amount of DNA that could be digested in a reaction and the units of each enzyme to be used. Concentration of DNA in the reaction seemed to limit the performance of the enzymes, leading to undigested DNA (Figure 3.6). To increase the efficiency of the reaction, overnight digestions with varying amounts of enzymes were performed (Figure 3.7). Digestions appeared to be incomplete, with very prominent high weight bands which could be a result of genomic DNA not being digested.

Genomic DNA samples were highly concentrated, at an estimated concentration of 5 µg/µl. Viscous DNA solutions reduce enzyme diffusion and can limit the efficiency of the digestion reactions. To reduce the viscosity of the solution, DNA samples were sheared before enzymatic digestion using a 26G needle. To investigate the effect that shearing DNA could have on restriction digestion and subsequent PCR reactions, two DNA samples were processed in parallel. One of them was sheared, and both samples (sheared and un-sheared) were digested using EcoRI and XhoI overnight. The digestion reactions were analysed on an agarose gel (Figure 3.8). These sheared samples show a more complete digestion of DNA with the restriction enzymes such that a soft smear is visible instead of the sharp band visualized in the un-sheared sample. A sample of sheared un-digested genomic DNA was also visualized and no DNA fragmentation is evident.

A band around the expected size for the EcoRI-XhoI region of the shRNA targeting vector was extracted to be used as template for the PCR amplification of the targeting sequences. Before amplification, two parameters of the PCR required optimization. The optimal annealing temperature was investigated using a gradient PCR which results in the same cycling conditions in several wells of the thermocycler except for the annealing temperature, which varies. A range of temperatures from 50ºC to 72ºC were used to determine the optimal annealing temperature of the primers. Figure 3.9 shows the results of an electrophoresis on an agarose gel of the products of these reactions. Reactions with annealing temperature between 50ºC and 59ºC resulted in successful strong amplification. 52ºC was chosen as the optimal annealing temperature, as it resulted in successful PCR amplification with no unexpected bands that could be result of unspecific primer binding and the band obtained appeared to be the brightest by eye.

Once the annealing temperature had been optimized, the number of amplification cycles was investigated. It was critical that PCR amplification did not alter the relative representation of targeting sequences so that the sequencing output could be used for quantitative analysis of the library composition after co-culture. To establish the optimal number of cycles that would yield products still in the exponential phase of the PCR, reactions were amplified for different number of cycles and compared on an agarose gel

(Figure 3.10). The brightness of the amplified band increased with each increase in cycle number up to 38 cycles. As the band produced after 26 cycles seemed to be slightly faint, and to accommodate variation in DNA concentration in the digested sample DNA templates, 28 cycles were performed in the PCR amplifications.

Using the optimal annealing temperature and number of cycles, PCR amplification of the band extracted from the sheared and un-sheared DNA digestions was examined. The effect of DNA concentration was also investigated.

A band of the expected size was observed in both sets of reactions, sheared and un-sheared DNA, showing that shearing did not affect PCR amplification (Figure 3.11). Increasing amounts of template DNA produced molecules of higher molecular weight. These molecules have been described before as annealed shRNA molecules and named “X-molecules” (Du, Ge et al. 2006). To minimize the proportion of X-molecules in the samples, 1 ng of template DNA was used. The shearing step was included into the DNA processing protocol. The optimal DNA processing pipeline is depicted in Figure 3.12. This process was used with all the samples in the final screen.

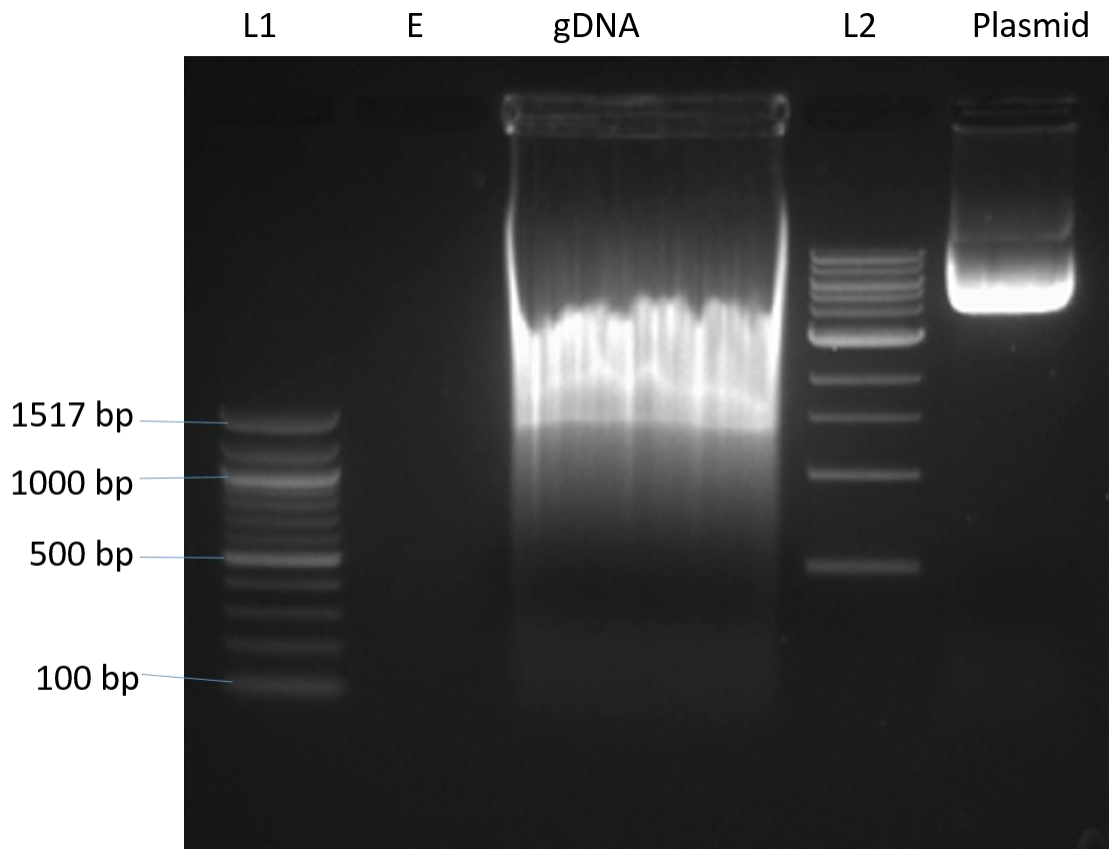



*Figure 3.6. Results of a restriction digestion of genomic DNA and library plasmid with XhoI and EcoRI.*
*10 µg of DNA were digested with 10 units of each enzyme at 37ºC for 3 hours in a total volume of 50 µl and run on a 1.2% agarose gel. L1: NEB 100 bp DNA ladder. E: empty. gDNA: Genomic DNA extracted from a co-culture experiment. L2: NEB 1 kb DNA ladder. Plasmid: library plasmid*

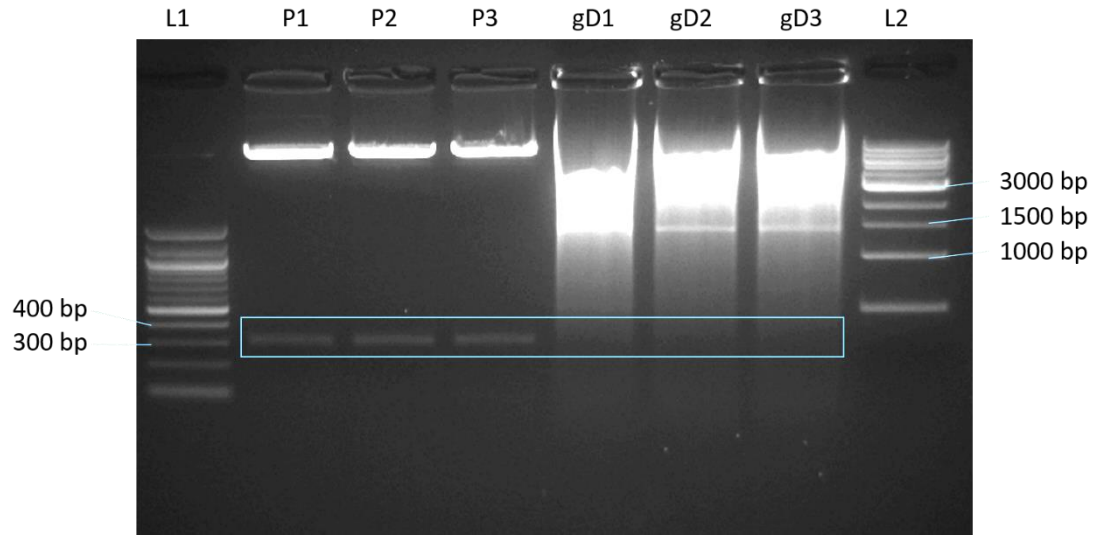



*Figure 3.7. Results of a restriction digestion of genomic DNA and library plasmid with varying amounts of XhoI and EcoRI.*
*10 µg of DNA was digested in a total volume of 50 µl for each reaction at 37ºC overnight and run on a 1.2% agarose gel. L1: NEB 100 bp DNA ladder. P1: library plasmid digested with 12 units of each enzyme. P2: library plasmid digested with 4 units of each enzyme. P3: library plasmid digested with 20 units of each enzyme. gD1: genomic DNA digested with 12 units of each enzyme. gD2: genomic DNA digested with 4 units of each enzyme. gD3: genomic DNA digested with 20 units of each enzyme. L2: NEB 1kb DNA ladder. The blue rectangle indicates the portion of gel that was extracted from each lane.*

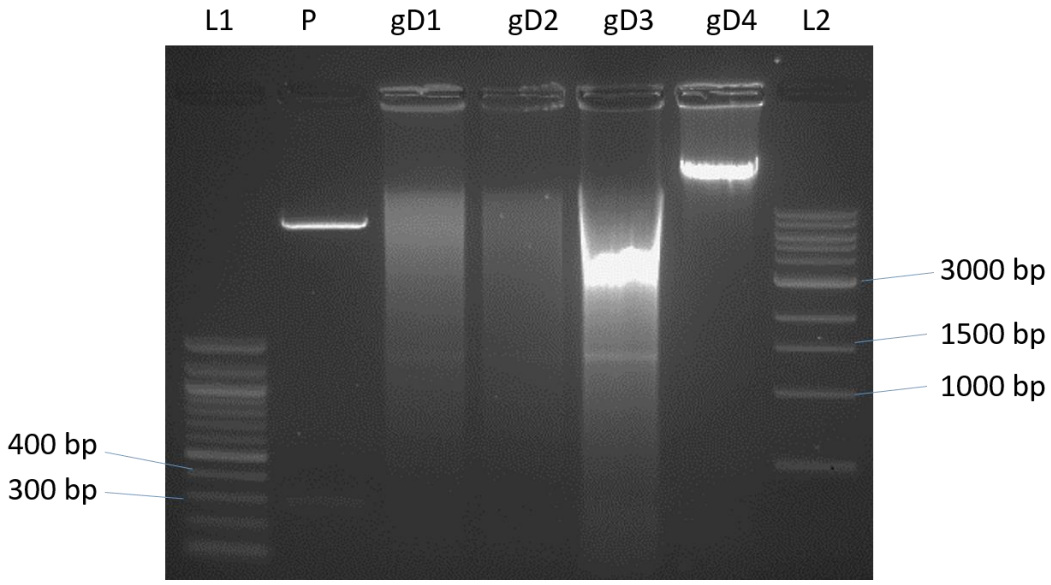



*Figure 3.8. Results of a restriction digestion of sheared and un-sheared genomic DNA and library plasmid digested with 6 units of XhoI and EcoRI in a total volume of 50 µl for each reaction.*
*Samples were incubated at 37ºC overnight and run on a 1.2% agarose gel. L1: NEB 100 bp DNA ladder. P: library plasmid digested with 12 units of each enzyme. gD1: 5 µg of sheared genomic DNA. gD2: 2 µg of sheared genomic DNA. gD3: 5 µg of un-sheared genomic DNA. gD4: undigested sheared DNA. L2: NEB 1kb DNA ladder.*

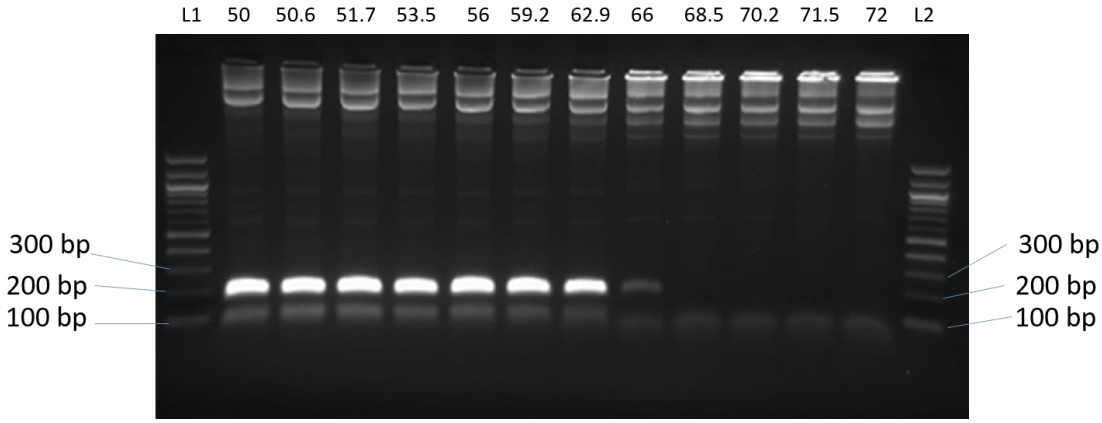



*Figure 3.9. Temperature gradient PCR results.*
*0.2 µg of library plasmid were amplified using a pair of HiSeq primers (3.2.7) with the annealing temperature set to the temperature above the lanes. L1 and L2: NEB 100 bp DNA ladder.*

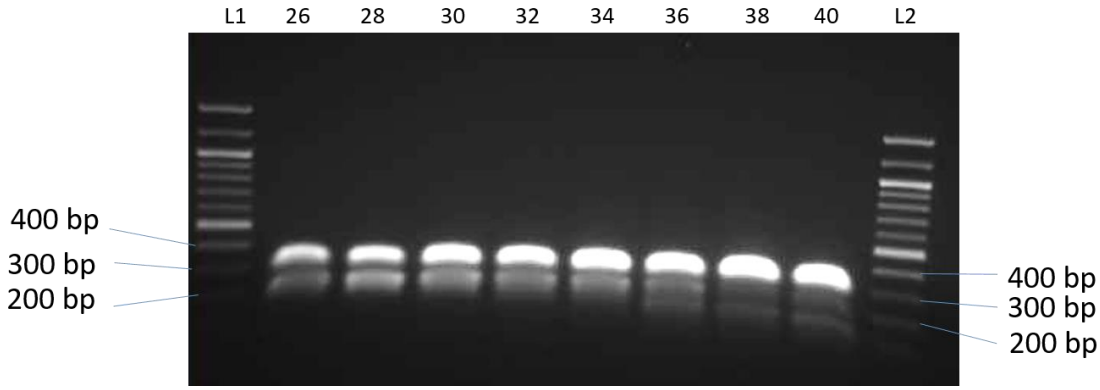



*Figure 3.10. Results of the optimization of PCR amplification cycles.*
*0.01 µg of library plasmid were amplified using a pair of HiSeq primers (3.2.7) for the number of cycles specified on the lane names. L1 and L2: NEB 100bp DNA ladder.*

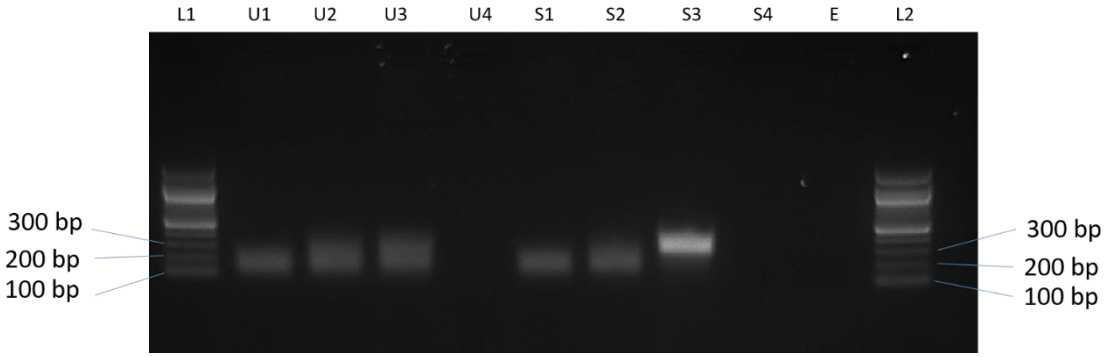



*Figure 3.11. PCR amplification results of un-sheared and sheared gDNA samples. L1 and L2: NEB 100 bp DNA ladder. U1: 1 ng of un-sheared DNA. U2: 2 ng of un-sheared DNA. U3: 4 ng of un-sheared DNA. U4: 4 ng of un-sheared DNA, not amplified. S1: 1 ng of sheared DNA. S2: 2 ng of sheared DNA. S3: 4 ng of sheared DNA. S4: 4 ng of sheared DNA, not amplified. E: empty reaction, no DNA negative control..*

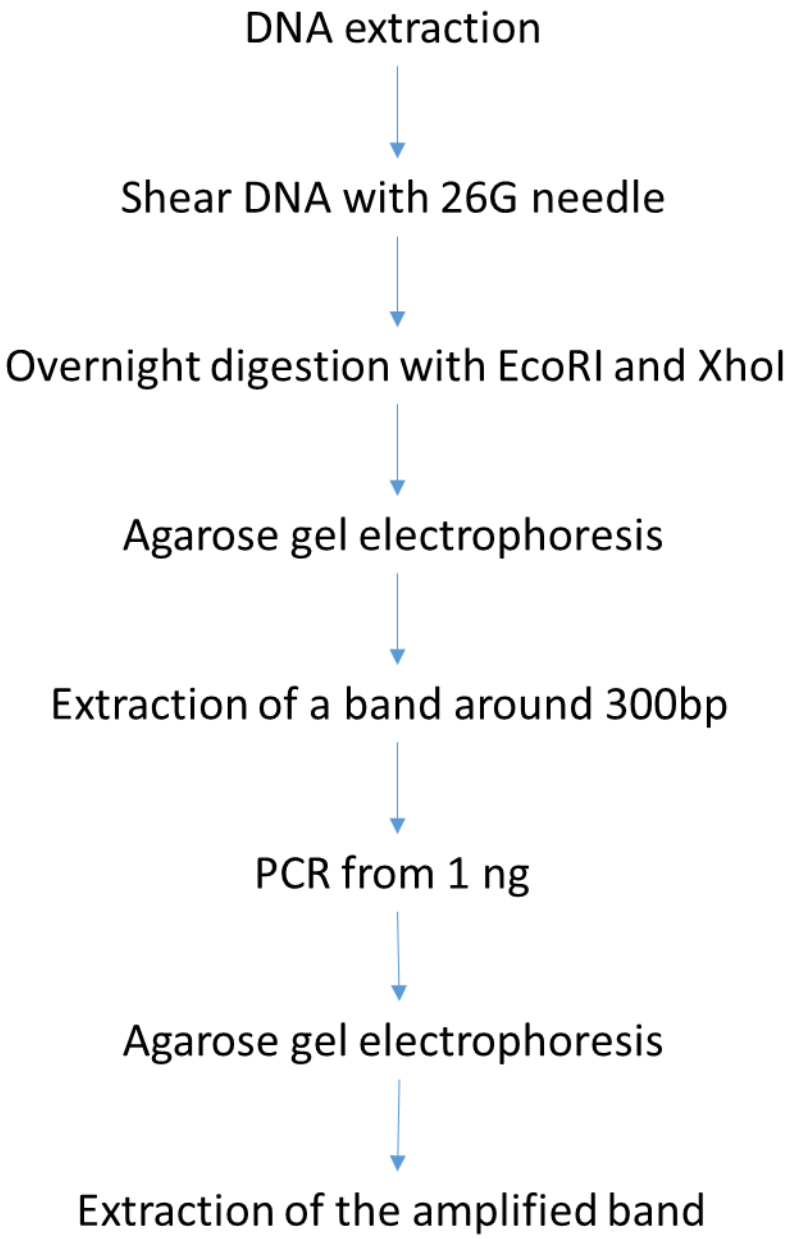



*Figure 3.12. DNA processing pipeline from extraction of DNA to amplification. The amplified band obtained from this pipeline was then sequenced by the University of Exeter Sequencing Service.*

### 3.3.6. Reproducibility of the screening process

To investigate the biological and technical variability of the screening process, three library pools were processed in duplicate. Both replicates were carried out in parallel but using independent cell cultures and reactions. Monolayers were set up using the same cell stocks but in separate flasks, and viral particles were derived from separate transfections of the packaging cell line. All further processes were carried out in parallel. Each condition (co-cultures and plastic culture) was PCR amplified using a different forward primer. PCR amplified shRNA constructs were sequenced using a different Illumina adaptor for each duplicate but on the same sequencing lane. As an indicator of the efficiency of the primers used for PCR amplification, PCR reactions were performed with the same DNA template using the four available forward primers.

Figure 3.13 shows the result of the amplification reactions. Although this is not quantitative there seems to be a slightly higher amount of DNA in the reaction using primer 2. Even if this is the case, an increase in efficiency would not alter the results because relative abundance of targeting sequences within a sample is used in the data analysis rather than absolute number of targeting sequences. It would not alter reproducibility either as long as this efficiency was maintained because the same condition was amplified using the same primer for both duplicates.

A total of 112,995,003 reads were obtained for duplicate A across all conditions, and 131,096,006 for duplicate B. Figure 3.17 shows the sequence of the PCR primers for each condition and how their differences are used to separate the sequencing reads that belong to each of the samples. All primers have a common sequence, which was used to extract the reads to a file containing the reads for all 3 samples (referred to as “general” below). Part of the sequence was common to both co-culture samples and it was used to extract the reads to a file containing the reads for both samples (referred to as “co-culture” below). Another part of the sequence was specific to the $3T3^{supp+}$ monolayer samples and it was used to produce a file containing only these sequences (referred to as “supp+” below). To detect potential shRNA constructs (sequencing reads with the expected structure of a shRNA) in these files, a Python script was written (Appendix 3) which identifies reads containing the exact sequence for the shRNA loop and extracts the 19bp before the start of the loop as a targeting

sequence. The script then counts how many times that targeting sequence is found in the sequencing reads and outputs a file with each unique targeting sequence and the number of times it appears in the sample.

These files were imported into Stata such that each targeting sequence was an observation and each sample a variable. To resolve the incomplete separation of the samples due to partial overlapping of the amplification primers, the numbers of reads in the samples were subtracted as follows: “general”-“co-culture” produced the “Plastic” sample; “co-culture”-“supp+” produced the “supp-” sample. After these operations, the file contained variables with the number of times each construct was detected in each of the three samples.

The number of potential shRNA constructs obtained per sample is shown in Table 3.5. Potential shRNA construct sequences were aligned against a file with the expected shRNA targeting sequences of the library. Only sequences that align 100% to the targeting sequences designed to be present in the library were used in subsequent analysis. Between 92.1% and 93.8% of the constructs designed to be in the pools screened were detected in the different samples (Table 3.6). This incomplete representation can be explained by the composition of the library. The in-house library was sequenced to analyse the shRNA targeting sequences contained in it, revealing that not all expected targeting sequences were detectable at the read depth used (Nicholls 2015). When only those targeting sequences detected in the library sequencing (1,024 out of 1,152) are considered to be in the library, between 92.2% and 99.1% of the sequences are found (Table 3.7). Therefore, not all of the sequences present in the library as determined through sequencing were detected. Three main reasons can explain this finding. The sequences might have not been sequenced in the screen due to insufficient read depth, their effects might be deleterious for the cells and therefore these are not present in the screen, or they might not have been transduced into the cells due to their low representation in the library. The average of reads for those targeting sequences in the in-house library sequencing dataset is approximately ten times lower than the average sequence in the same dataset, suggesting that at least part of the sequences are not present in the screening process due to their low representation in the library.

z-scores were produced for both replicates as explained in section 3.2.8. shRNA construct reads, reads per million per pool (rpmp) and z-scores were plotted for both replicates using Stata. Dotplots comparing both replicates at each state of the analysis are depicted in Figure 3.16. Correlation of z-scores as calculated by Stata between both replicates was 0.27 for $3T3^{supp-}$ and 0.54 for $3T3^{supp+}$. Complete tables of correlations for the various steps of the analysis are depicted in Table 3.8, Table 3.9 and Table 3.10. Correlations coefficients were in general higher between the different conditions of the same replicate than between both replicates of the same condition. This reveals an effect on either the transfection and transduction steps or during amplification and sequencing, or both. More than 100,000 packaging cells per targeting sequence were transfected, and that representation was kept during the transduction of the sequences into SVE-6. This fold representation is high and therefore unlikely to be the source of variation. A lower representation (100 cells per targeting sequence) was used when seeding SVE-6 and that may be a source of error, but it has been previously explained why increasing the number of SVE-6 cells seeded on monolayers would not be viable.

Therefore, PCR amplification and sequencing were targeted for improvement before the full-scale screen was carried out. The following modifications were included in the final pipeline. The amount of genomic DNA used as template for shRNA sequence amplification was doubled for all conditions. A higher sequencing depth was also requested when performing the full-scale screen (average of 2,000 reads instead of 1,000 reads per construct) as many of the sequencing reads obtained did not correspond to the samples but rather to Illumina adaptor dimers (Figure 3.18). Other sequences were detected to be overrepresented in the sequencing output but could not be assigned to an Illumina adaptor. These sequences could not be mapped to the mouse genome or transcriptome using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) nor to the plasmid vector sequence.

In general, most constructs detected in this pilot screen were common to all conditions, as shown in Figure 3.14 (replicate A) and Figure 3.15 (replicate B). This indicates that even though the PCR amplification and the sequencing was not optimal, the pipeline was robust enough to avoid loss of targeting sequences from some conditions and not others.

Applying the analysis pipeline that was to be used with the final output of the screen, one targeting sequence was detected to have an effect on $3T3^{supp+}$ on both replicates, while 36 were detected only in one replicate. These were selected using a double threshold method, shown in Figure 3.19 and Figure 3.20. The objective of this method was to avoid those targeting sequences that had the same effect on both co-cultures. This requirement is a consequence of the *in vitro* model used. Tumorigenic cell growth on $3T3^{supp+}$ is strongly suppressed and therefore an increase in suppression in this co-culture is highly unlikely to represent an effect on suppression but rather a less specific cell growth effect. Therefore, sequences that exert this effect on both co-cultures are most likely having an effect on tumorigenic cell growth and are outside the scope of this work. The same reasoning can be applied for $3T3^{supp-}$ co-cultures, where a targeting sequence is unlikely to trigger a reduction of suppression given that this is almost non-existent. Therefore, for a construct to have an effect in the $3T3^{supp+}$ co-cultures, the z-score must be in the top 5% of the z-score distribution in the $3T3^{supp+}$ sample but at the same time it must not be in the top 5% of the z-score distribution in the $3T3^{supp-}$. Complementarily, for a construct to be a considered a GOI on $3T3^{supp-}$ its z-score must be in the bottom 5% of the z-score distribution in the $3T3^{supp-}$ sample while its z-score in $3T3^{supp+}$ must not be in the bottom 5% of the z-score distribution in the $3T3^{supp+}$.

Two GOI (*Pax2* and *Tyk2*) were found to be targeted by two constructs in replicate A in $3T3^{supp+}$, meaning that both targeting sequences for those two genes were detected to have had an effect. In replicate B, no GOI were targeted by two constructs. *Taf6l* was identified as a GOI in both replicates in $3T3^{supp+}$.

No constructs were detected to have an effect on $3T3^{supp-}$ on both replicates, with 12 being detected in either one of the replicates but not in the other.

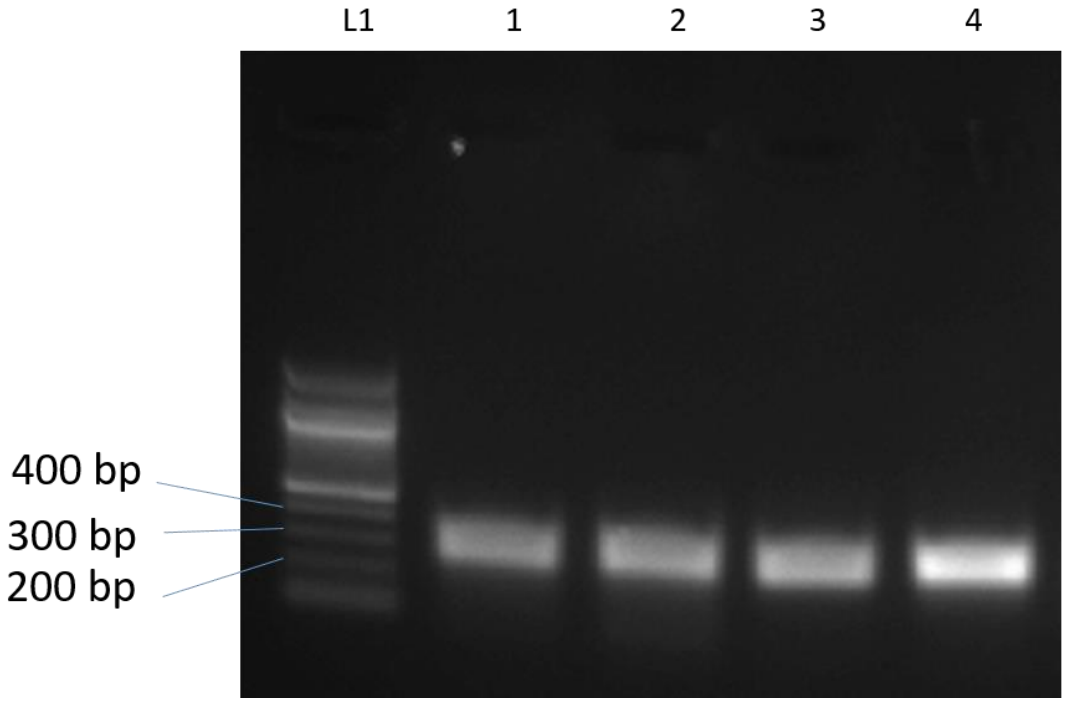



*Figure 3.13. Results of the PCR amplification of 0.1 ng of library plasmid with the available primers for targeting sequence amplification.*
*L1: NEB 100bp DNA ladder. 1: PCR amplification using primer HiSeq 4. 2: PCR amplification using primer HiSeq 8. 3: PCR amplification using HiSeq 1. 4: PCR amplification using HiSeq 2.*

| | Replicate A | Replicate B |
|---|---|---|
| **Plastic** | 76,844 | 108,342 |
| **3T3$^{supp+}$** | 306,836 | 392,206 |
| **3T3$^{supp-}$** | 90,859 | 107,758 |

*Table 3.5. Number of potential shRNA constructs found in the pilot sequencing data*

| | Replicate A | Replicate B |
|---|---|---|
| **Plastic** | 1,061 (92.1%) | 1,070 (92.9%) |
| **3T3$^{supp+}$** | 1,071 (93.0%) | 1,081 (93.8%) |
| **3T3$^{supp-}$** | 1,061 (92.1%) | 1,070 (92.9%) |

*Table 3.6. Number of shRNA constructs designed to be in the library found in the pilot sequencing data.*
*Numbers inside brackets are percentages of expected shRNA constructs according to the theoretical composition of the library.*

| | **Replicate A** | **Replicate B** |
|---|---|---|
| **Plastic** | 944 (92.2%) | 960 (93.8%) |
| **3T3$^{supp+}$** | 1,008 (98.4%) | 1,015 (99.1%) |
| **3T3$^{supp-}$** | 1,003 (97.9%) | 1,011 (98.7%) |

*Table 3.7. Number of shRNA constructs designed to be in the library found in the pilot sequencing data when only those constructs previously sequenced are considered. Numbers inside brackets are percentages of expected shRNA constructs according to the composition of the library determined by sequencing.*

| | **3T3$^{supp+}$A** | **3T3$^{supp-}$A** | **Plastic A** | **3T3$^{supp+}$B** | **3T3$^{supp-}$B** | **Plastic B** |
|---|---|---|---|---|---|---|
| **3T3$^{supp+}$A** | 1.00 | | | | | |
| **3T3$^{supp-}$A** | 0.82 | 1.00 | | | | |
| **Plastic A** | 0.67 | 0.94 | 1.00 | | | |
| **3T3$^{supp+}$B** | 0.53 | 0.39 | 0.28 | 1.00 | | |
| **3T3$^{supp-}$B** | 0.68 | 0.57 | 0.47 | 0.86 | 1.00 | |
| **Plastic B** | 0.71 | 0.92 | 0.94 | 0.44 | 0.60 | 1.00 |

*Table 3.8. Correlation matrix of sequencing reads in the pilot project.*

| | **3T3$^{supp+}$ A** | **3T3$^{supp-}$ A** | **Plastic A** | **3T3$^{supp+}$ B** | **3T3$^{supp-}$ B** | **Plastic B** |
|---|---|---|---|---|---|---|
| **3T3$^{supp+}$ A** | 1.00 | | | | | |
| **3T3$^{supp-}$ A** | 0.83 | 1.00 | | | | |
| **Plastic A** | 0.75 | 0.88 | 1.00 | | | |
| **3T3$^{supp+}$ B** | 0.80 | 0.92 | 0.75 | 1.00 | | |
| **3T3$^{supp-}$ B** | 0.80 | 0.88 | 0.80 | 0.85 | 1.00 | |
| **Plastic B** | 0.70 | 0.92 | 0.84 | 0.88 | 0.83 | 1.00 |

*Table 3.9. Correlation matrix of reads per million per pool (rpmp) in the pilot project*

| | **3T3$^{supp+}$A** | **3T3$^{supp-}$A** | **3T3$^{supp+}$B** | **3T3$^{supp-}$B** |
|---|---|---|---|---|
| **3T3$^{supp+}$A** | 1.00 | | | |
| **3T3$^{supp-}$A** | 0.73 | 1.00 | | |
| **3T3$^{supp+}$B** | 0.54 | 0.41 | 1.00 | |
| **3T3$^{supp-}$B** | 0.36 | 0.27 | 0.70 | 1.00 |

*Table 3.10. Correlation matrix of z-scores*

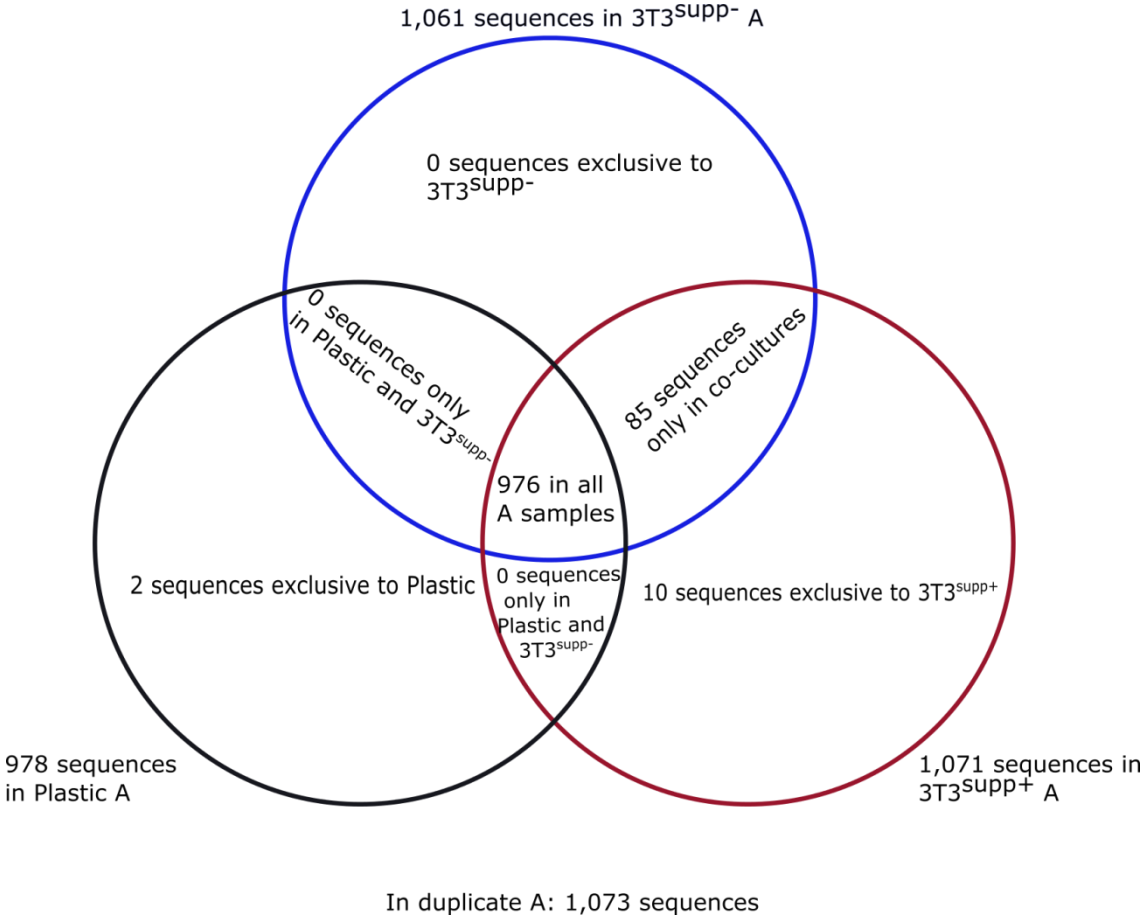



*Figure 3.14. Venn diagram showing the intersection of constructs present in replicate A*

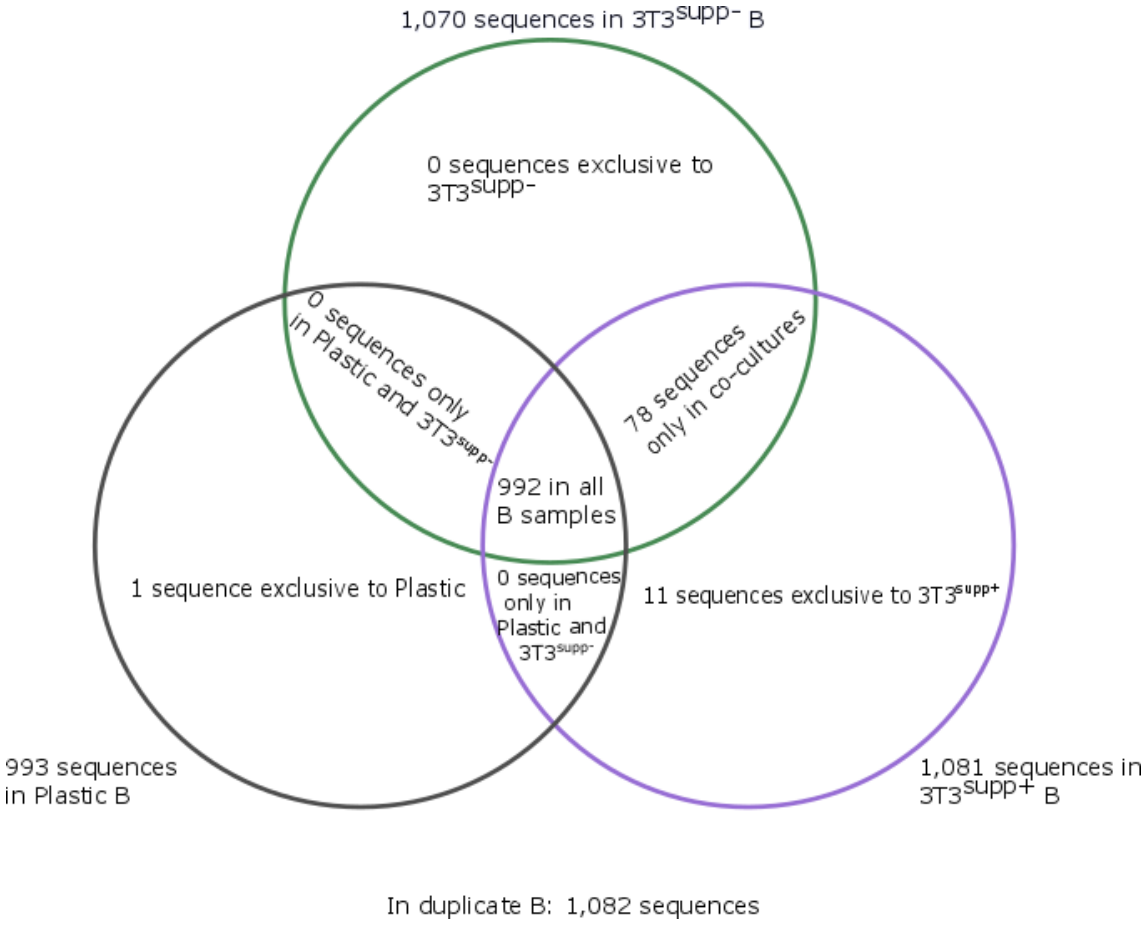



*Figure 3.15. Venn diagram showing the intersection of constructs present in replicate B*

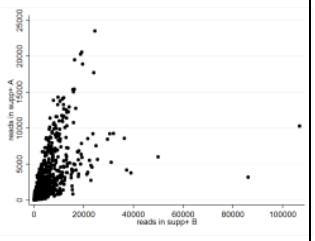
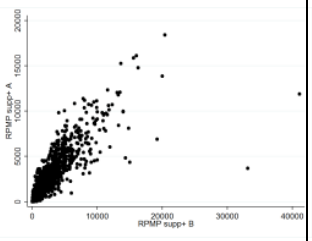
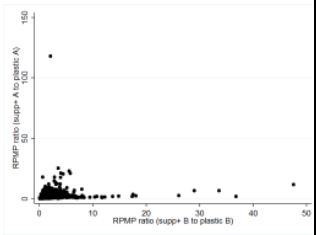
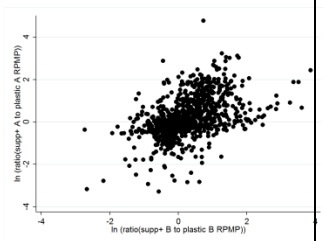
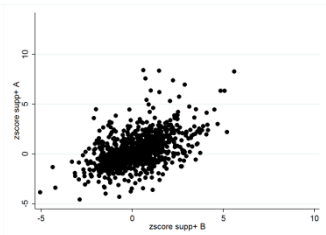
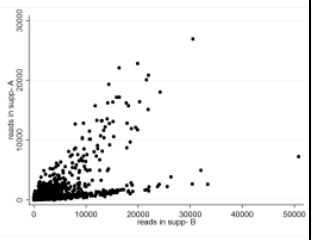
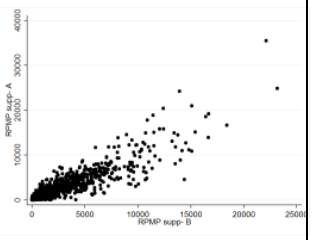
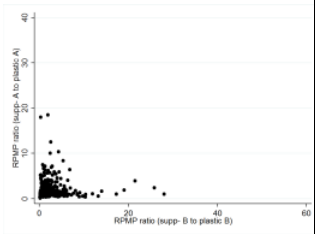
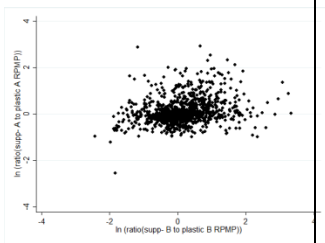
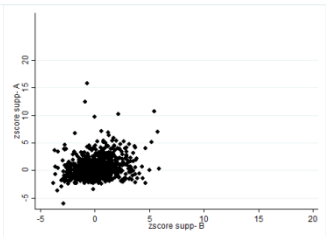
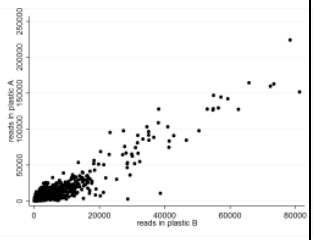
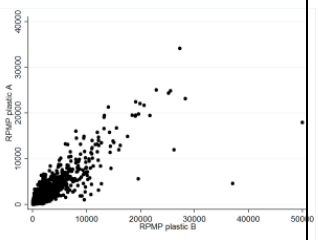
| | Raw read number | Rpmp | rpmp ratio to P | ln (rpmp to P) | zscore |
|---|---|---|---|---|---|
| 3T3$^{supp+}$ |  |  |  |  |  |
| 3T3$^{supp-}$ |  |  |  |  |  |
| P |  |  | Not applicable | Not applicable | Not applicable |

*Figure 3.16. Dotplots comparing replicates A and B of the pilot project.*

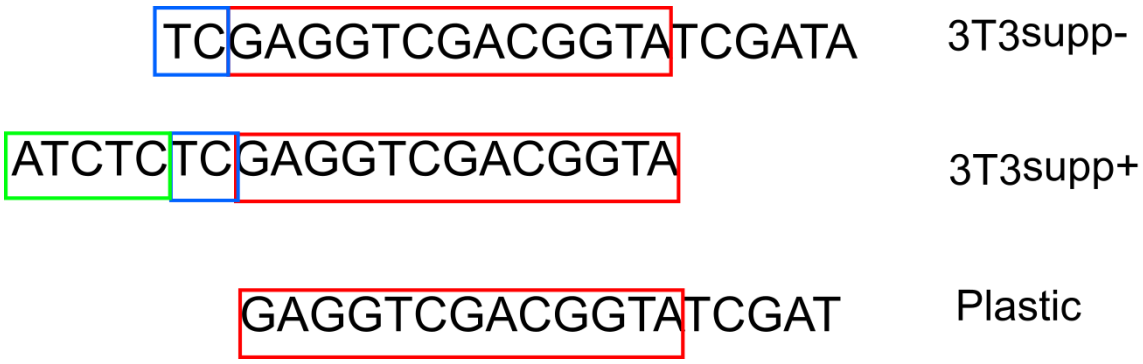



*Figure 3.17. Primers used for PCR amplification of the samples.*
*The sequence marked with the red rectangle was common to all samples and used to extract the reads to a file. Reads containing the red+blue marked sequence were removed from that file (which now contained the "plastic" reads) and placed into a "co-culture reads" file. From this file, reads containing the sequence marked in green were removed (generating a "3T3supp-" file) and placed into a "3T3supp+" file*
.

A)

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGAAGATCTCGTATGC | 13796004 | 12.209393011830798 | TruSeq Adapter, Index 10 (97% over 36bp) |
| GTCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTCC | 126373 | 0.11183945895377337 | No Hit |
| TGTCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTC | 123850 | 0.10960661685189743 | No Hit |

B)

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGGCAATCTCGTATGC | 23185778 | 17.68610555534392 | TruSeq Adapter, Index 10 (97% over 36bp) |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGGCAATATCGTATGC | 286812 | 0.2187801205781967 | TruSeq Adapter, Index 10 (97% over 36bp) |
| TGTCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTC | 146495 | 0.11174634870264469 | No Hit |
| GTCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTCC | 143558 | 0.10950600585039943 | No Hit |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGGCAATCTCGTATTC | 140947 | 0.1075143357151552 | TruSeq Adapter, Index 10 (97% over 36bp) |

*Figure 3.18. Overrepresented sequences table from FastQC report on quality control of A): duplicate A and B): duplicate B.*

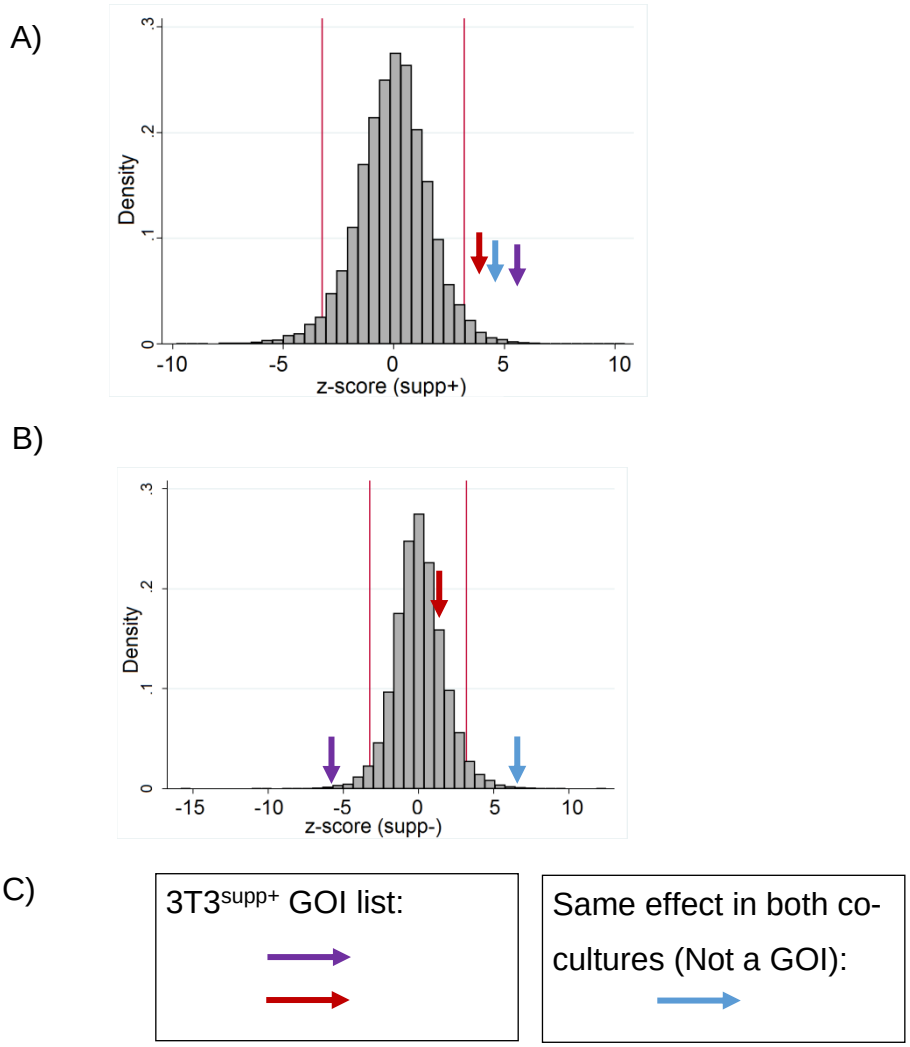



*Figure 3.19. Graphical demonstration of the double threshold system used to select GOI for the 3T3$^{supp+}$ co-culture sample.*
*A) Targeting sequences in the higher 5% of the 3T3supp+ z-score distribution (to the right of the red vertical line) are selected as potential GOIs. For illustration purposes, three arrows have been drawn to represent three random targeting sequences. B) Histogram for the 3T3supp- z-score distribution. Arrows represent the same three targeting sequences as in A. The purple arrow is situated in the bottom 5% of the distribution for 3T3supp-, the red arrow is in the 90% around the average z-score and the blue arrow is situated in the top 5% of the distribution for 3T3supp-.C) The purple and red arrows are included in the 3T3supp+ GOI list. The blue arrow has been determined to have the same effect in both co-cultures and therefore is not included in any GOI list.*

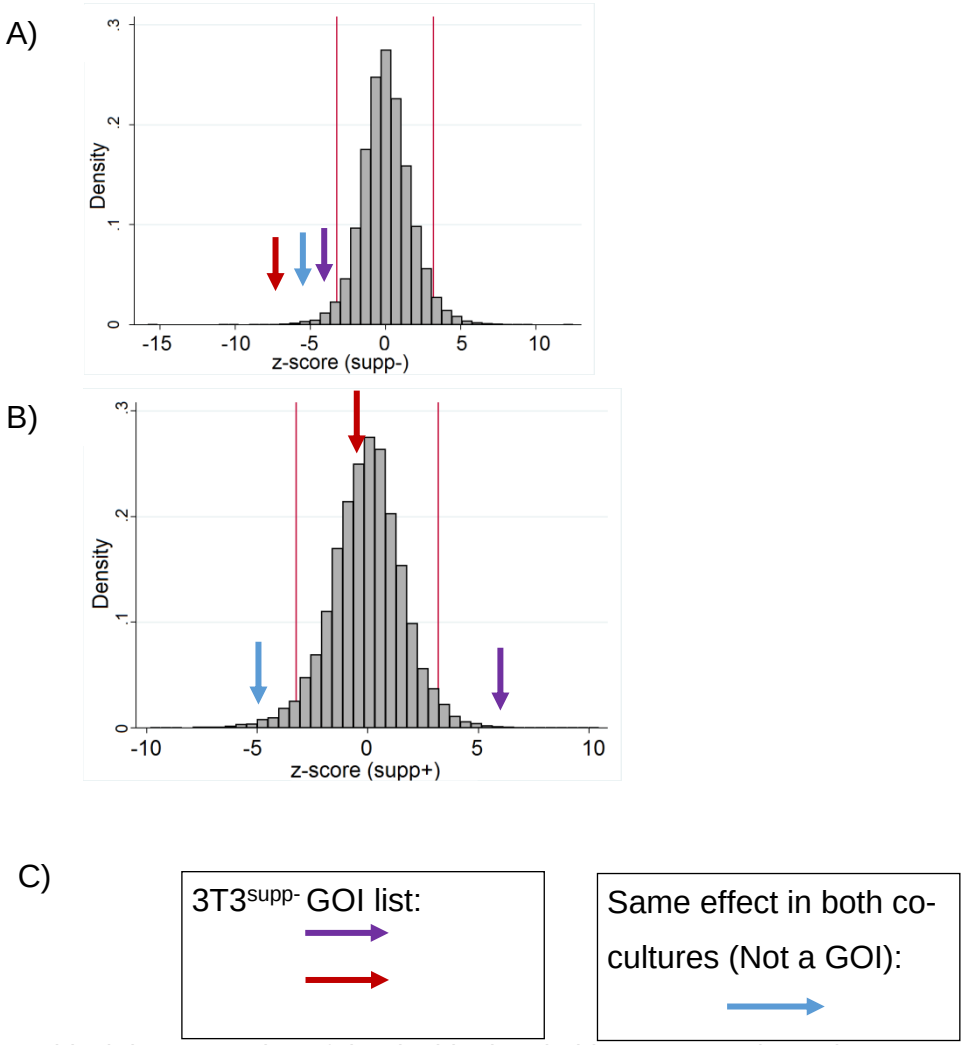



*Figure 3.20. Graphical demonstration of the double threshold system used to select GOI for the 3T3$^{supp-}$ co-culture sample.*
*A) Targeting sequences in the lower 5% of the 3T3$^{supp-}$ z-score distribution (to the left of the red vertical line) are selected as potential GOIs. For illustration purposes, three arrows have been drawn to represent three random targeting sequences. B) Histogram for the 3T3$^{supp+}$ z-score distribution. Arrows represent the same three targeting sequences as in A. The purple arrow is situated in the top 5% of the distribution for 3T3$^{supp+}$, the red arrow is in the 90% around the average z-score and the blue arrow is situated in the bottom 5% of the distribution for 3T3$^{supp+}$.C) The purple and red arrows are included in the 3T3$^{supp+}$ GOI list. The blue arrow has been determined to have the same effect in both co-cultures and therefore is not included in any GOI list.*

### 3.3.7. Running the genome-wide shRNA screening process

The full screen was performed with all 73 pools following the revised, optimal protocol and analysis pipeline identified in section 3.3.6. Only one replicate of the genome-wide screen was performed. Pools were run in batches of 10, with each pool being physically separate as indicated for the pilot pools. DNA samples were kept separate until combining them for loading into the HiSeq flow-cell. At that point, DNA samples were quantified using a Qubit fluorimeter and mixed in equal proportions. The mixture was then processed and sequenced by the University of Exeter Sequencing Service.

### 3.3.8. Sequencing output of the screening process.

All 73 library pools were screened until the point of extracting genomic DNA. However, 3 library pools (9, 45 and 48) failed to PCR amplify in at least one condition. Those pools were therefore not included in the samples that were sequenced and are absent from any further analysis.

Sequencing results were received as a .fastq file which contained the reads for all the conditions pooled together. A total of 84,652,417 sequencing reads were obtained from the sequencing run. This file was managed as explained in section 3.3.6, producing a file with shRNA construct sequences and the corresponding reads containing that sequence in each of the samples. The number of potential shRNA targeting sequences (19bp preceding the shRNA loop sequence) detected in each sample is specified in Table 3.11, and the number of shRNA constructs designed to be in the library that were found in each sample is in Table 3.12. Robust z-scores were calculated for each targeting sequence as explained in section 3.3.6. Correlation matrices for the different data transformations are shown in Table 3.13 (raw read numbers), Table 3.14 (reads per million per pool) and Table 3.15 (z-scores).

| | |
|---|---|
| **Plastic** | 163,731 |
| **3T3$^{supp+}$** | 266,787 |
| **3T3$^{supp-}$** | 155,801 |

*Table 3.11. Number of potential shRNA constructs found in the sequencing data*

| | |
|---|---|
| **Plastic** | 25,246 (93.9%) |
| **3T3$^{supp+}$** | 25,610 (95.3%) |
| **3T3$^{supp-}$** | 25,357 (94.3%) |

*Table 3.12. Number of shRNA constructs designed to be in the library found in each sample in the sequencing data.*
*In brackets, percentage of the number of shRNA constructs screened according to the library design.*

| | **3T3$^{supp+}$** | **3T3$^{supp-}$** | **Plastic** |
|---|---|---|---|
| **3T3$^{supp+}$** | 1 | | |
| **3T3$^{supp-}$** | 0.8593 | 1 | |
| **Plastic** | 0.9574 | 0.8910 | 1 |

*Table 3.13. Correlation matrix of raw sequencing reads*

| | **3T3$^{supp+}$** | **3T3$^{supp-}$** | **Plastic** |
|---|---|---|---|
| **3T3$^{supp+}$** | 1 | | |
| **3T3$^{supp-}$** | 0.9673 | 1 | |
| **Plastic** | 0.4381 | 0.4882 | 1 |

*Table 3.14. Correlation matrix of reads per million per pool*

| | **3T3$^{supp+}$** | **3T3$^{supp-}$** |
|---|---|---|
| **3T3$^{supp+}$** | 1 | |
| **3T3$^{supp-}$** | 0.6014 | 1 |

*Table 3.15. Correlation matrix of z-scores*

### 3.3.9. Selection of potential effectors of neighbour suppression in the sequencing data

Several methods of analysis of the sequencing data were considered. The objective of the analysis pipeline was to detect constructs that had a specific effect on each of the co-culture samples as explained in section 3.3.6.

First it was considered to use the reads per million per pool (RPMP) as a proxy for cell number. However, the variation in the different pools in the three conditions studied was not homogenous and therefore this direct comparison was considered inefficient and biased. Histograms of RPMP for four of the pools in the three conditions studied are depicted in Figure 3.21 to illustrate this point. The density distribution for RPMP in these four pools is different and also varies between conditions. If a threshold was to be set using one of the pools and consider any constructs above this threshold in any of the other pools as having had an effect, the threshold would be very different depending on the pools used to set it.

To account for this, an analysis method using z-scores calculated from reads per million per pool was investigated. Although this method accounted for differences in the variability levels amongst pools it was difficult to detect effects due to the nature of the screen. For a construct to be considered of interest, it needs to have an effect in one of the co-cultures but not in the other co-culture and also not on plastic monoculture. To achieve this, a system of three thresholds was studied. For each pool, a threshold would need to be implemented for each of the co-cultures and the monocultures. Targeting sequences would have needed to pass a threshold in a co-culture, a different one in the other co-culture to ensure that the effects were not the same in both co-cultures and then a third one to discard any sequences with an effect on growth as a monoculture. Given the number of pools screened, this approach although logical was considered inefficient and an alternative solution was sought. To reduce this system of multiple thresholds, a ratio of co-culture z-scores to plastic z-scores was tried next. Due to the distributions of z-scores, which ranged from negative to positive numbers some artefacts were detected for negative z-scores. Finally, an analysis pipeline was devised where reads per million per pool for each construct in $3T3^{supp+}$ and $3T3^{supp-}$ were divided by the

reads per million per pool for the same construct in plastic. The distribution of these ratios was very variable amongst pools. To achieve a more homogenous range of distributions, the ratios were converted to logarithms. A robust z-score was built from these log(ratios). To select potential effectors in each co-culture sample, a double threshold system was used. Constructs on the top or bottom 5% of one of the conditions but not on the other were considered to have an effect.

A total of 593 constructs were selected using this method, 281 from $3T3^{supp+}$ and 312 from $3T3^{supp-}$. Eight genes were targeted by two constructs: *Ttbk1*, *Cd5l*, *Snap23* and *Coa3* in the $3T3^{supp+}$ GOI list, and *Rrp7a*, *Nr2c2ap*, *Actrt3* and *Gapdh* in the $3T3^{supp-}$. As the library is composed of two constructs per gene, having detected an effect with both constructs may be an indicator of robustness of the effect. It may also however be due to chance and those GOI would still need replication.

In total, 277 genes were selected as GOI for $3T3^{supp+}$ and 308 for $3T3^{supp-}$.

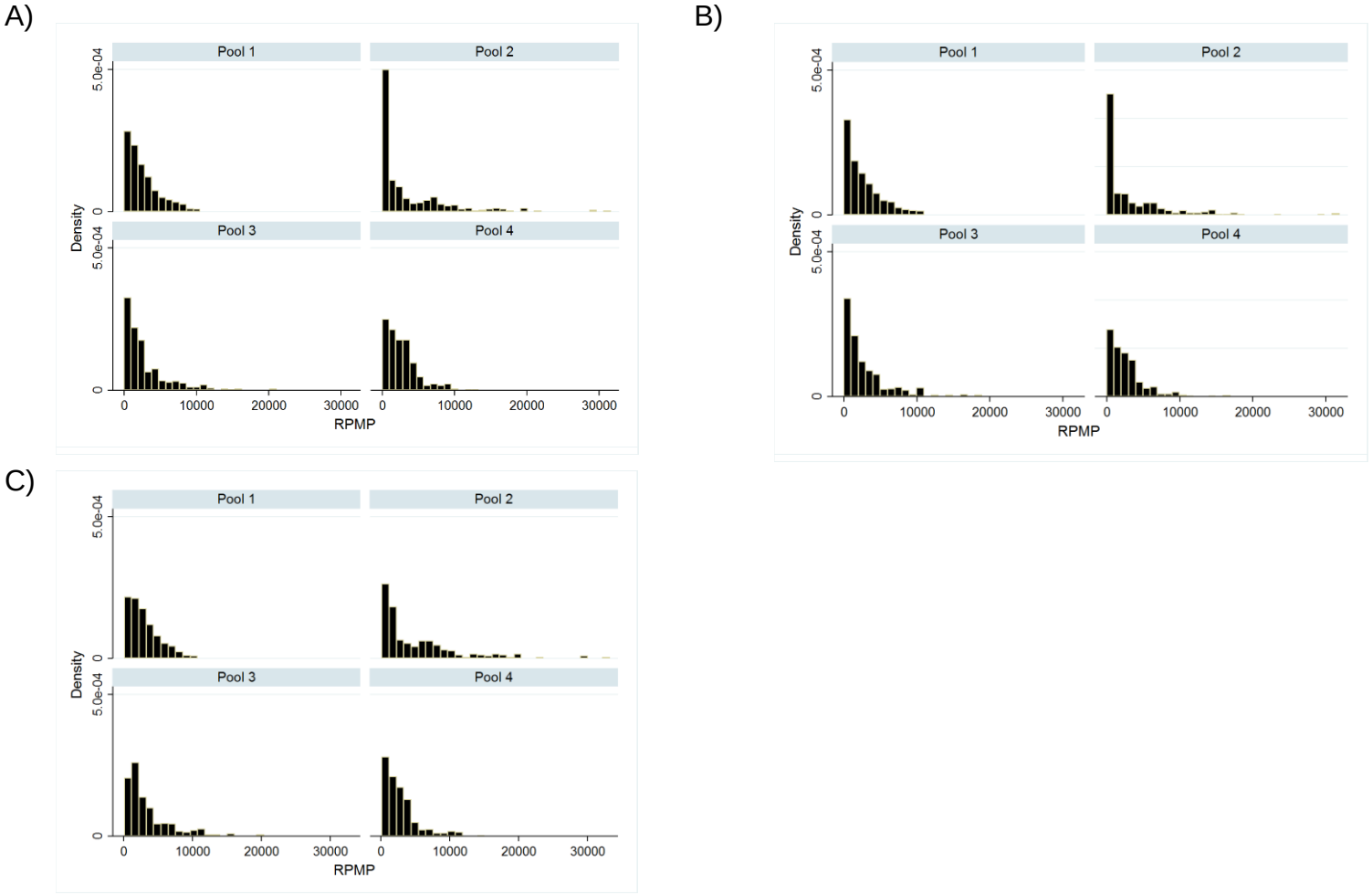



*Figure 3.21. Histograms of RPMP for pools 1 to 4.*
*A) 3T3$^{supp-}$ co-culture sample. B) 3T3$^{supp+}$ co-culture sample. C) Plastic sample.*

### 3.3.10. Network analysis of the sequencing output

As this screen used a comprehensive genome-wide, non-biased approach it was possible that the genes targeted by some of the constructs selected as having an effect could be part of protein complexes or have a function in the same pathway. Therefore gene interactions were investigated using network analysis. It was hoped that this analysis would result in the production of a list of potential "essential" genes or a pathway essential to the phenotype – those connected with many other GOI. With this aim of producing a second list of GOI to follow-up based on their connections with other GOI, network analysis was produced using two different approaches.

#### *3.3.10.1. Network analysis using STRING*

STRING (available at http://string-db.org/) was initially used to build gene networks. The list of GOI for each condition was uploaded to STRING and networks were built using the "High confidence" setting (confidence of the interaction higher than or equal to 0.7). "Textmining" was not used to predict interactions as this method links any proteins that are mentioned in the same abstract of published paper, which was thought to result in non-specific or non-validated connections.

Networks for both co-culture samples GOI are shown in Figure 3.22 ($3T3^{supp-}$) and Figure 3.23 ($3T3^{supp+}$). A network was also produced for both lists combined (Figure 3.24). STRING provides some basic network analysis. For the $3T3^{supp+}$ network, STRING indicates that there are 65 interactions and 23 were expected under the null hypothesis that the network is not enriched for interactions. It produced a p-value for enrichment in interactions of $3.95*10^{-13}$. The same analysis of the $3T3^{supp-}$ revealed that although 32 interactions were expected the sample contains 63, which is a significant enrichment with a p-value of $1.38*10^{-6}$. When the network produced for the combined list containing both $3T3^{supp+}$ and $3T3^{supp-}$, STRING indicated that 110 interactions were expected but 261 were found (p-value for enrichment was reported as 0, which indicates it is under the minimum computed by STRING). If the connections were only found between pairs of gene inside the individual co-culture networks, 128 connections were to be expected and that is not the case, which suggests the genes from both GOI lists are interconnected. To visualize this, the genes in

Figure 3.24 have been colour-coded according to whether they pertain to either the $3T3^{supp+}$ or the $3T3^{supp-}$. There is no apparent differential distribution of GOI according to which sample they were identified in. $3T3^{supp+}$ GOI are often connected to $3T3^{supp-}$ GOI and not only to $3T3^{supp+}$ GOI. This level of inter-sample connection may suggest the involvement of a common mechanism in the tumorigenic cells response to suppressive cues from neighbouring cells and their growth in absence of those suppressive cues.

To analyse the networks and identify “hubs” or genes with more connections than expected by chance (section 3.3.11), it was first necessary to be able identify all the possible interactors of each gene in the screened genes. Unfortunately, the interaction database used by STRING to produce networks is not readily available for download. No further analyses were therefore produced using these networks and an alternative database that could be fully analysed was used instead.

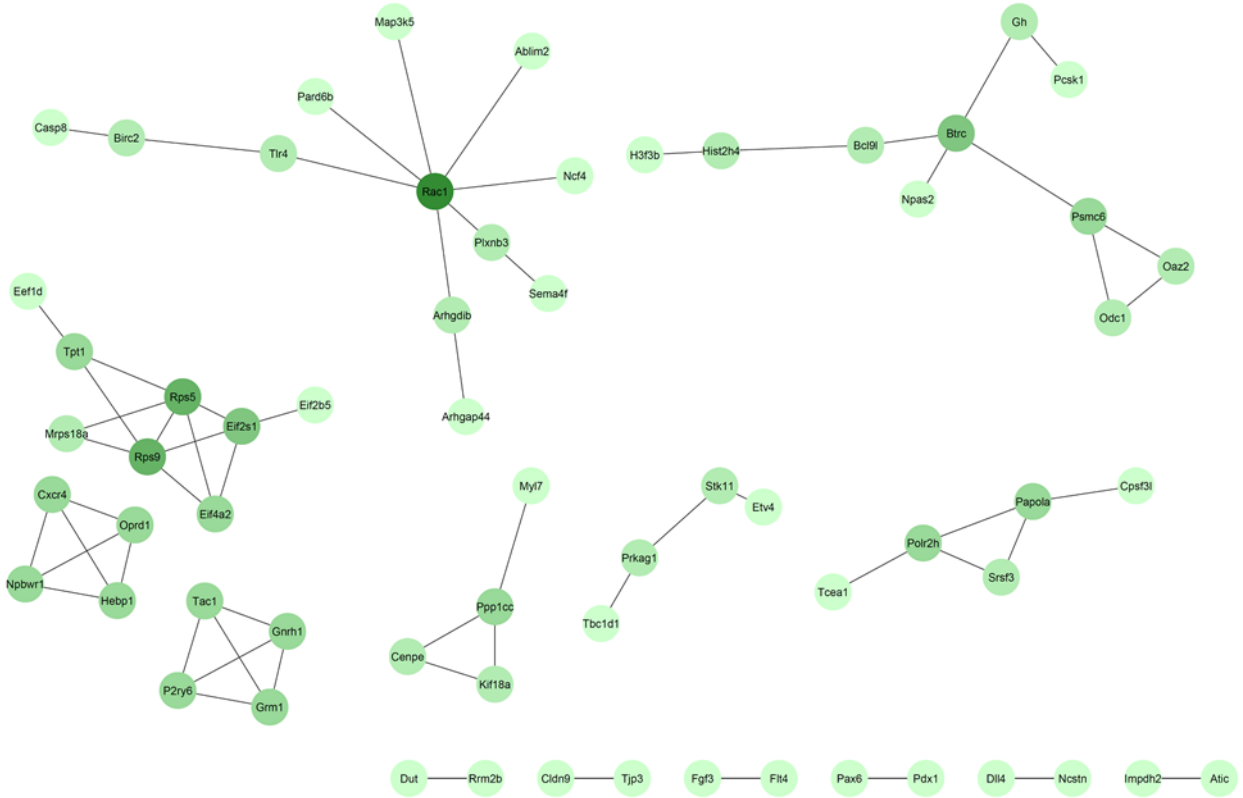



*Figure 3.22. Gene network for 3T3$^{supp-}$ built using STRING and visualised in Cytoscape.*
*Circles represent genes, colour coded by the number of interactions in the sample. Lines between two genes represent the interaction between them. Genes with no interactors are not represented in this network.*

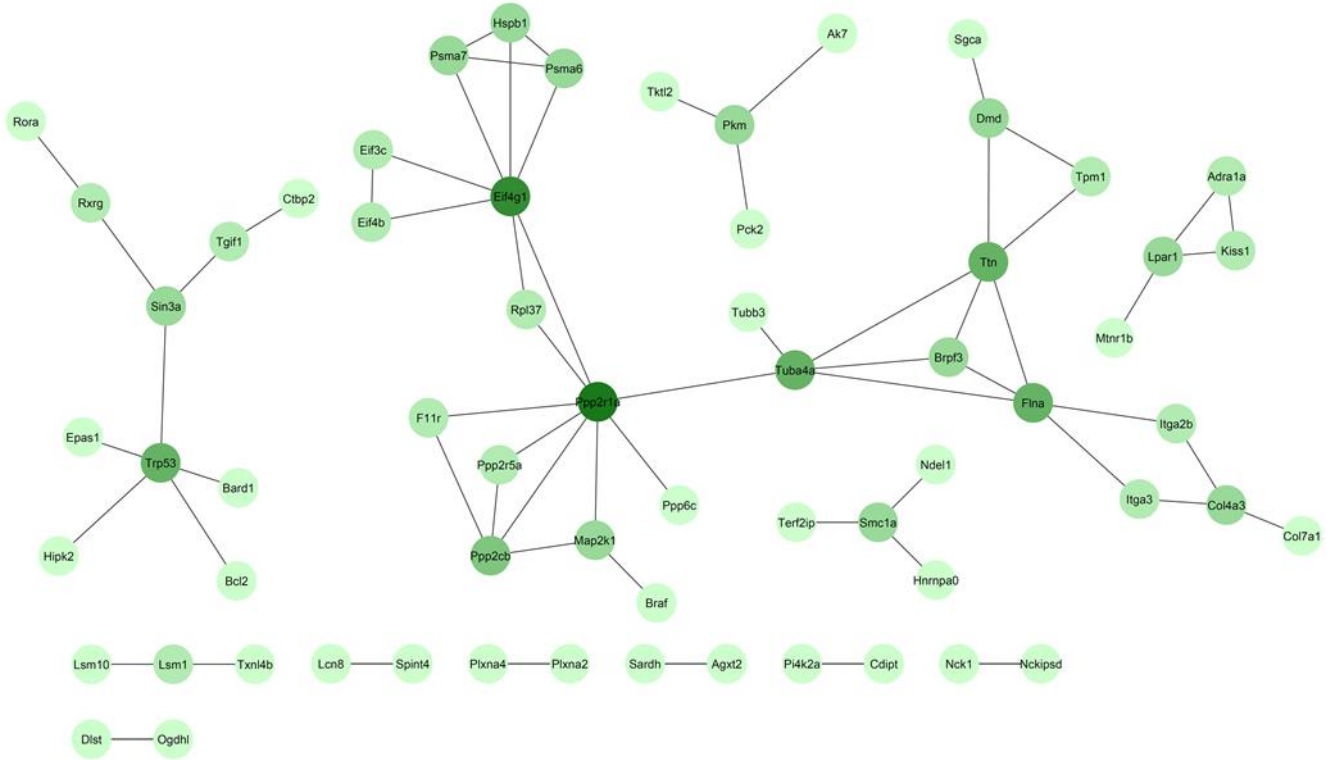



*Figure 3.23. Gene network for 3T3supp+ built using STRING and visualized in Cytoscape.*
*Circles represent genes, colour coded by the number of interactions in the sample. Lines between two genes represent the interaction between them. Genes with no interactors are not represented in this network*

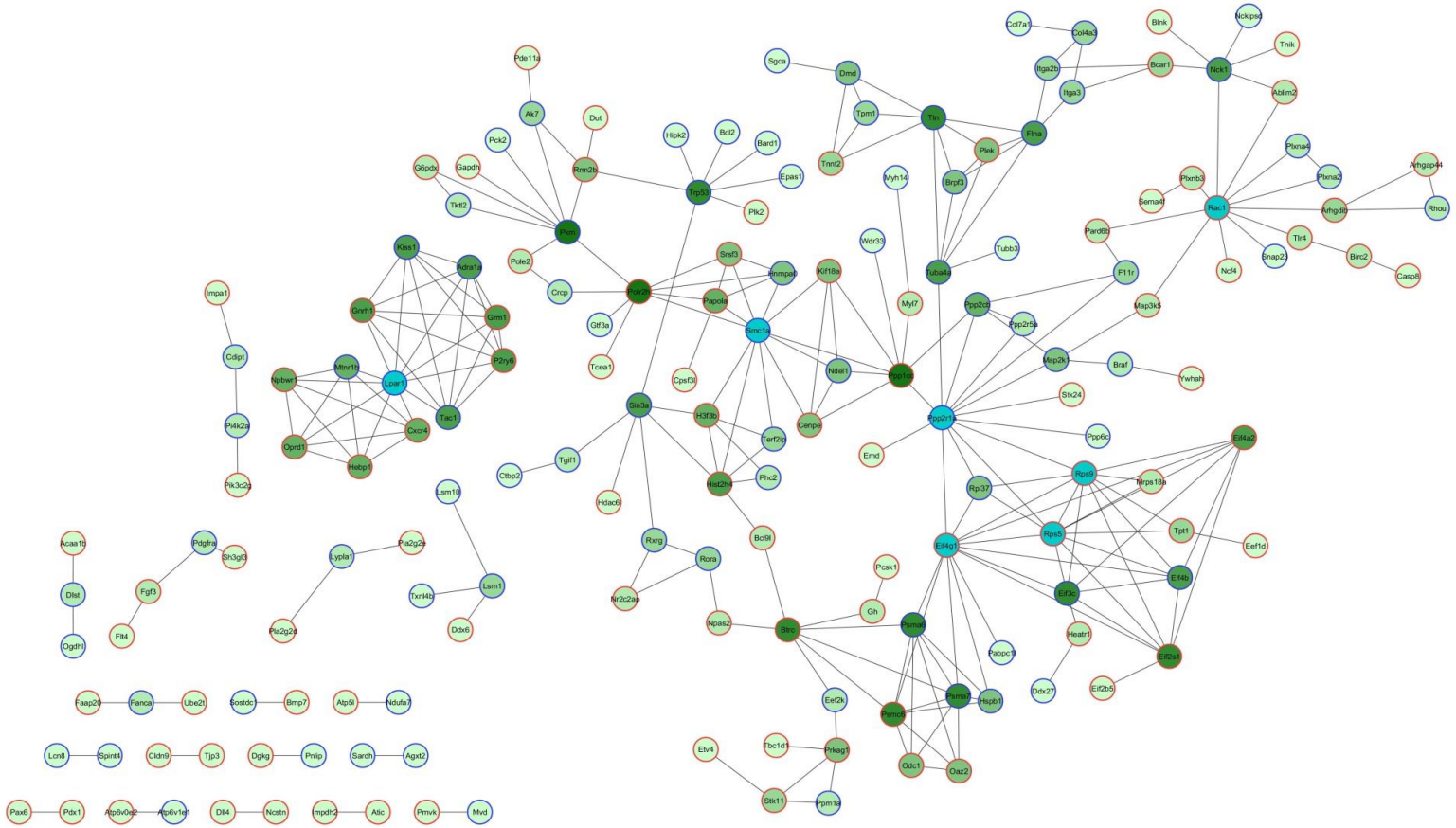



*Figure 3.24. Gene network for the combined GOI list built using STRING and visualized in Cytoscape.*
*Circles represent genes, with filling colour representing the number of interactions in the sample. Blue marks genes that have more than 9 connections in the sample. Lines between two genes represent the interaction between them. Circles with dark blue border are 3T3$^{supp+}$ GOI and red border marks 3T3$^{supp-}$ GOI. Genes with no interactors are not represented in this network*

### *3.3.10.2. Network analysis using FunCoup*

FunCoup v 3.0 (Schmitt, Ogris et al. 2014) is available online at http://funcoup.sbc.su.se/search/. It produces networks of genes using data from a number of sources, with some but not all of them overlapping with the sources used by STRING. Table 3.16 compares the data sources used by STRING and FunCoup. The protein interaction database used by FunCoup is available to download from the same website.

Gene interaction networks were built using FunCoup. Only queried genes were used to build the networks by setting the “Expansion depth” to 0. A confidence threshold of 0.8 was used to filter interactions between pairs of genes. This threshold retains only those connections with a high level of confidence as calculated by FunCoup from the evidence the database uses. Connections among genes were exported as text files. These files consist on a list of connections with the names of the two genes and the score for the diverse type of connections evaluated. The type of connection with the higher score is reported as the most likely connection, and the score of that connection is reported as the connection strength. The files for $3T3^{supp+}$ and $3T3^{supp-}$ were imported into Cytoscape 3.3.0 for visualization and initial analysis.

General views of the networks are represented in Figure 3.25 ($3T3^{supp+}$) and Figure 3.26 ($3T3^{supp-}$). A network was also produced for the combined GOI list (Figure 3.27) obtained in section 3.3.10.1. FunCoup does not produce any analysis of the connections in the network and therefore no enrichment in connections is available for these networks. However, Cytoscape does indicate the number of genes and connections amongst them when a network is displayed. The network produced by FunCoup from the $3T3^{supp+}$ GOI list has 78 genes connected by 192 interactions. The $3T3^{supp-}$ network is composed of 92 genes and 441 interactions, which indicates the average number of interactions per gene is higher in the $3T3^{supp-}$ network than in the $3T3^{supp+}$ network. In a similar manner to the results obtained with STRING, the network produced from the combination of both GOI lists has more genes with at least one connection (200) than the sum of both individual networks (170). A total of 1247 interactions connect these 200 genes. When the combined network (Figure 3.27) is examined it is apparent that genes from both GOI lists interact and

there is not a clear separation between the individual GOI lists. This interaction between pairs of genes from both GOI lists was also observed in the network obtained using STRING in the previous section.

| | **STRING** | **FunCoup** |
|---|---|---|
| **Conserved neighbourhood** | Genes that appear closely together (less than 300bp of separation between genes) in prokaryotic genomes | Not considered |
| **Co-occurrence** | Linked proteins are either present or absent across species | Gene conservation across species obtained from InParanoid |
| **Co-expression** | The expression levels of pairs of genes correlate in the same or other species (by homology) | mRNA (from GEO datasets) and protein co-expression (from Human Protein Atlas) |
| **Protein-protein interactions (PPI)** | PPI has been demonstrated experimentally and published | Combined PPIs from iRefIndex. The score depends on the number of publications confirming it and the scale of the experiments (larger scale experiments have less weight) |
| **Text-mining** | Data extracted from paper abstracts that mention both proteins | Not considered |
| **Genetic interactions** | Linked genes are fused in some species | Considers interactions between pathways under the assumption that a gene in pathway A will probably interact with a gene in pathway B if any gene in pathway A interacts with any gene in pathway B |
| **Shared transcription factor binding sites** | Not considered | Presence of binding sites for the same transcription factor near two genes is considered as evidence of functional coupling |
| **Co-regulation by miRNA** | Not considered | Regulation of expression by a common miRNA is considered as evidence of functional coupling |
| **Sub-cellular localization** | Not considered | From GO terms weighted on specificity of location |
| **Domain interaction** | Not considered | From UniDomInt, interactions between protein domains are used. |

*Table 3.16. Comparison of the data sources used by String and FunCoup to infer interactions between pairs of proteins.*

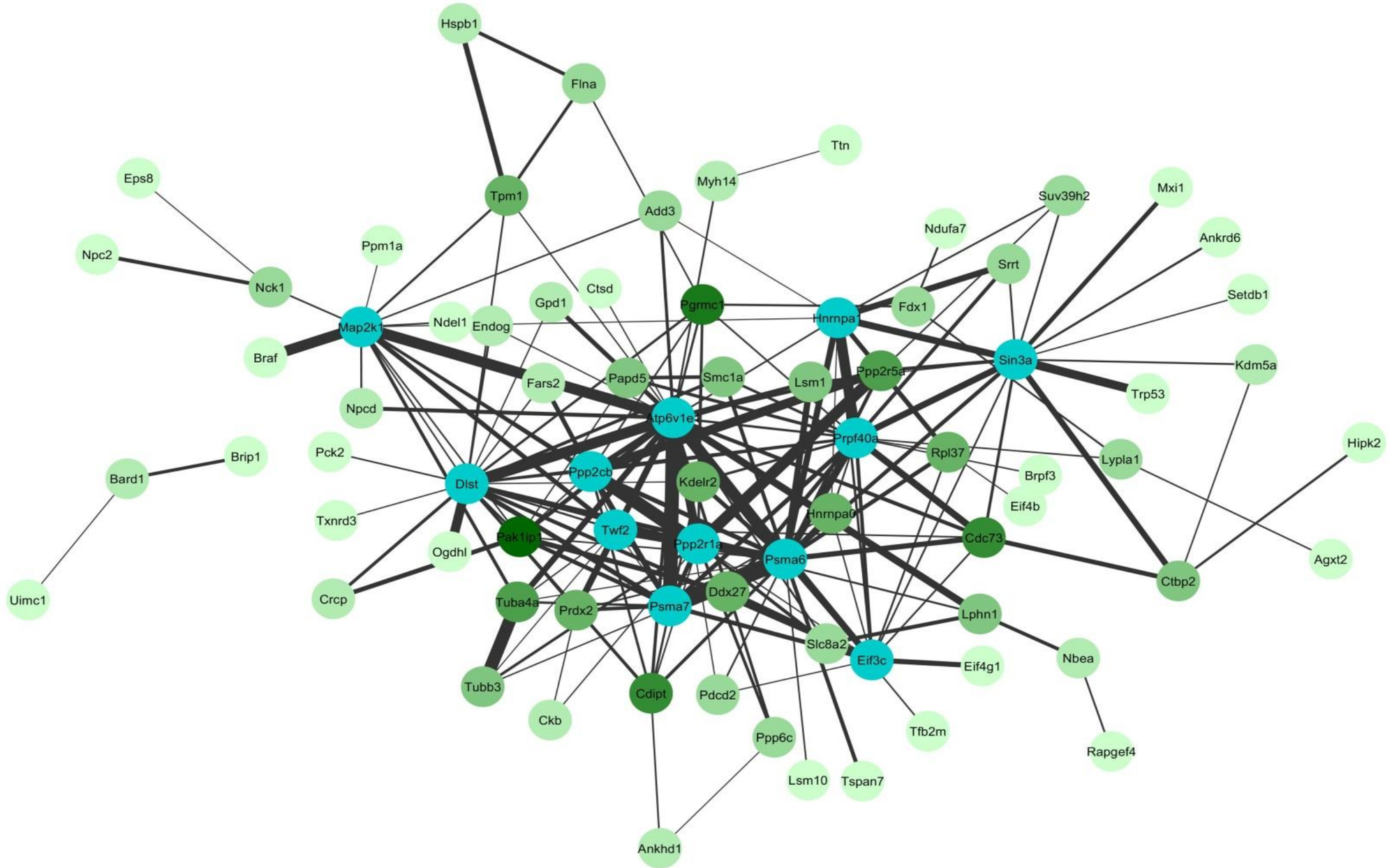



*Figure 3.25. Gene network for 3T3$^{supp+}$ GOI list built using FunCoup.*
*Circles represent genes, colour coded for number of interactors in the GOI list. Blue: 9 interactors or more, and continuous spectrum of green from more interactors (dark green) to fewer (light green). Black lines represent the connections between pairs of genes, with thickness representing the strength of the interaction as per FunCoup.*

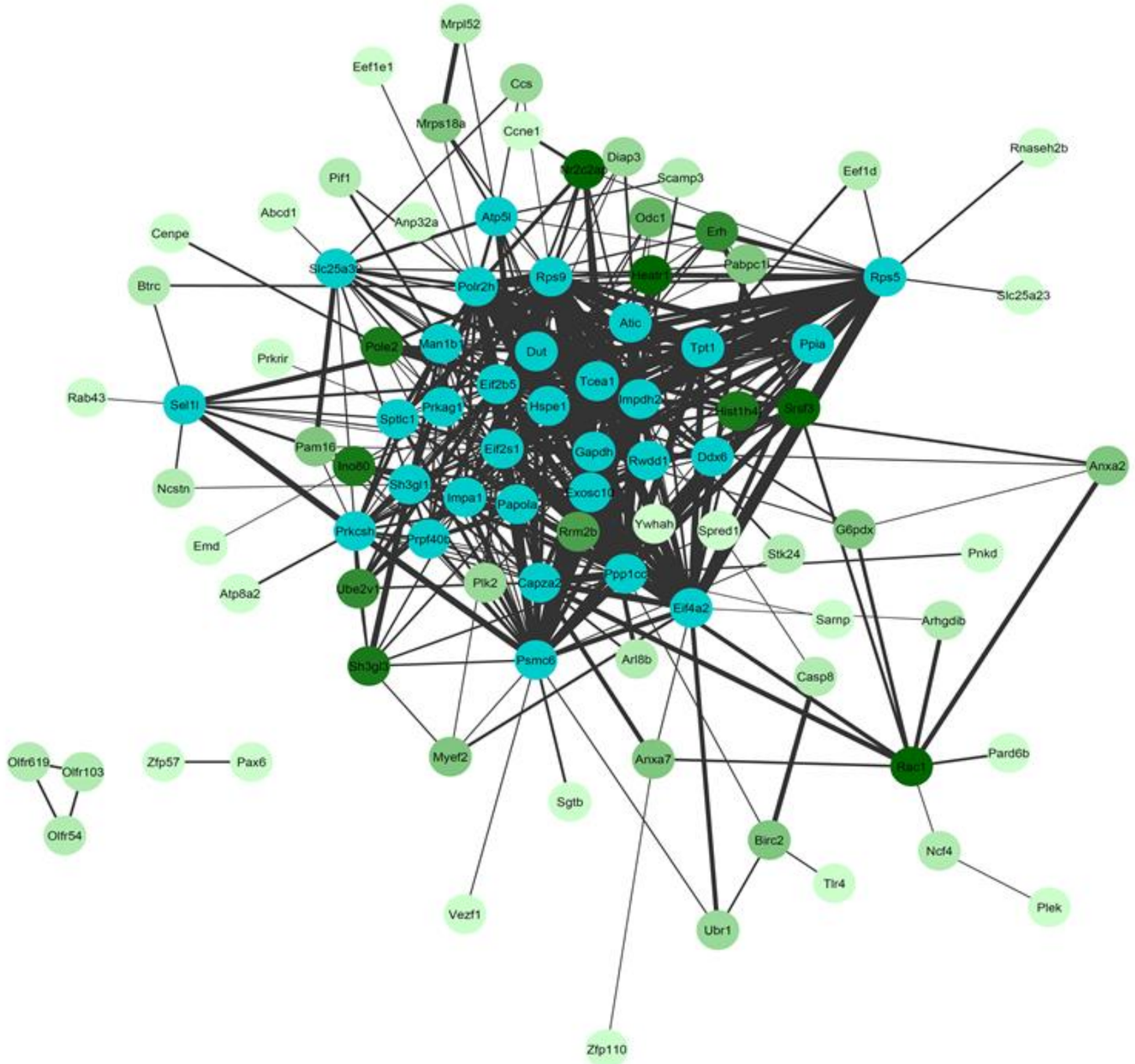



*Figure 3.26. Gene network for 3T3$^{supp-}$ GOI list built using FunCoup.*
*Circles represent genes, colour coded for number of interactors in the GOI list. Blue: 9 interactors or more, and continuous spectrum of green from more interactors (dark green) to fewer (light green). Black lines represent the connections between pairs of genes, with thickness representing the strength of the interaction as per FunCoup.*

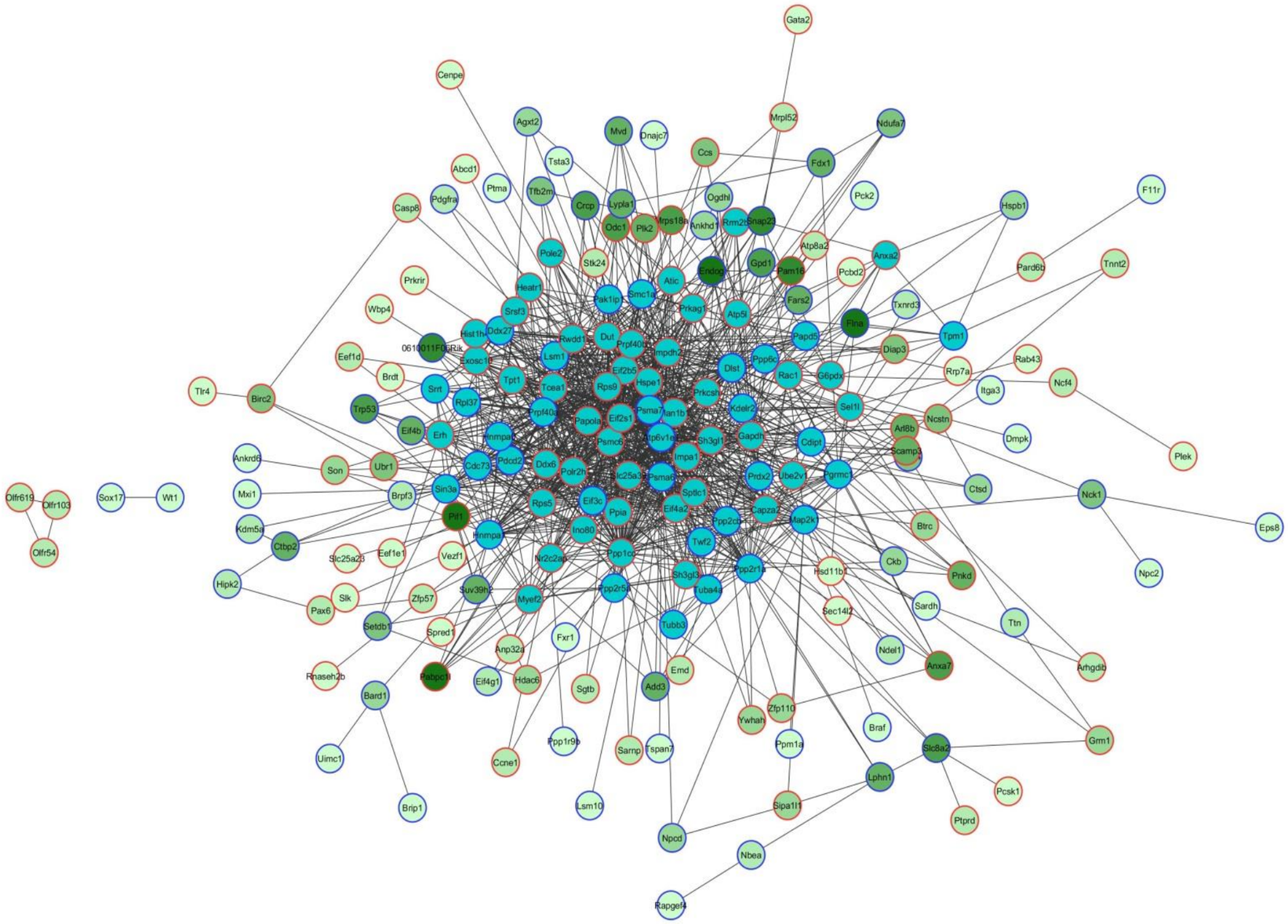



*Figure 3.27. Gene network for combined GOI list built using FunCoup.*
*Circles represent genes, colour coded for number of interactors in the GOI list. Blue: 9 interactors or more, and continuous spectrum of green from more interactors (dark green) to fewer (light green). Black lines represent the connections between pairs of genes. Circles with dark blue border are 3T3$^{supp+}$ GOI and red border marks 3T3$^{supp-}$ GOI. Genes with no interactors are not represented in this network*

#### *3.3.10.3. Prioritization of potential effectors of neighbour suppression in the network analysis data*

The network analysis data was used to determine which genes to prioritise in the further confirmation experiments

The networks produced by FunCoup were analysed for the presence of genes with a higher number of connections than expected given the number of interacting genes present in the screened genes.

Genes were first ranked by the number of connections they had with other genes in the GOI list, obtained from the “NetworkAnalyzer” tool in Cytoscape (section 2.7.2). A script was written in Python and used to test the significance of the number of connections for each gene using a random sampling approach (Appendix 6).

The database of gene interactions in *Mus musculus* used by FunCoup was downloaded and filtered for connections with more than 0.8 confidence score (same threshold used for queried genes). The logic behind the script structure is summarised in Figure 3.28. Basically, the script iterated through the list of GOI producing 1,000 random samples of the same size of the GOI list, all including the gene currently being tested. Random samples were drawn without replacement from the list of constructs in the library. The script then iterated through the genes in the random sample and checked whether they interacted with the gene being tested according to the FunCoup database. For each GOI a count of how many samples had the same or higher number of interactions than in the GOI list was outputted. The proportion of the samples with same or higher number of connections was used as a p-value for significance of enrichment of gene interactors in the GOI list. A total of 25 genes were found to have a significantly higher number of interactors in the GOI list with a 0.05 level of significance. Nine of these genes belonged to the $3T3^{supp+}$ co-culture (Table 3.18) and sixteen to the $3T3^{supp-}$ co-culture (Table 3.17). Networks for these hyper-connected genes in the GOI lists are depicted in Figure 3.29 ($3T3^{supp-}$) and Figure 3.30 ($3T3^{supp+}$).

The same analysis was performed using the combined GOI list produced in section 3.3.10.2 (network in Figure 3.31). In this case 29 genes (Table 3.19) were statistically enriched (p-value<0.05) in connections. Of those, fourteen

genes had been determined as statistically enriched for connections in their respective GOI list – seven in each of the conditions. These fourteen genes have been highlighted in Table 3.19. Fifteen genes had not been previously determined to be enriched for connections. Some of them did not have any interactors at all in their original GOI list whilst others did but gained more interactors when the GOI lists were combined.

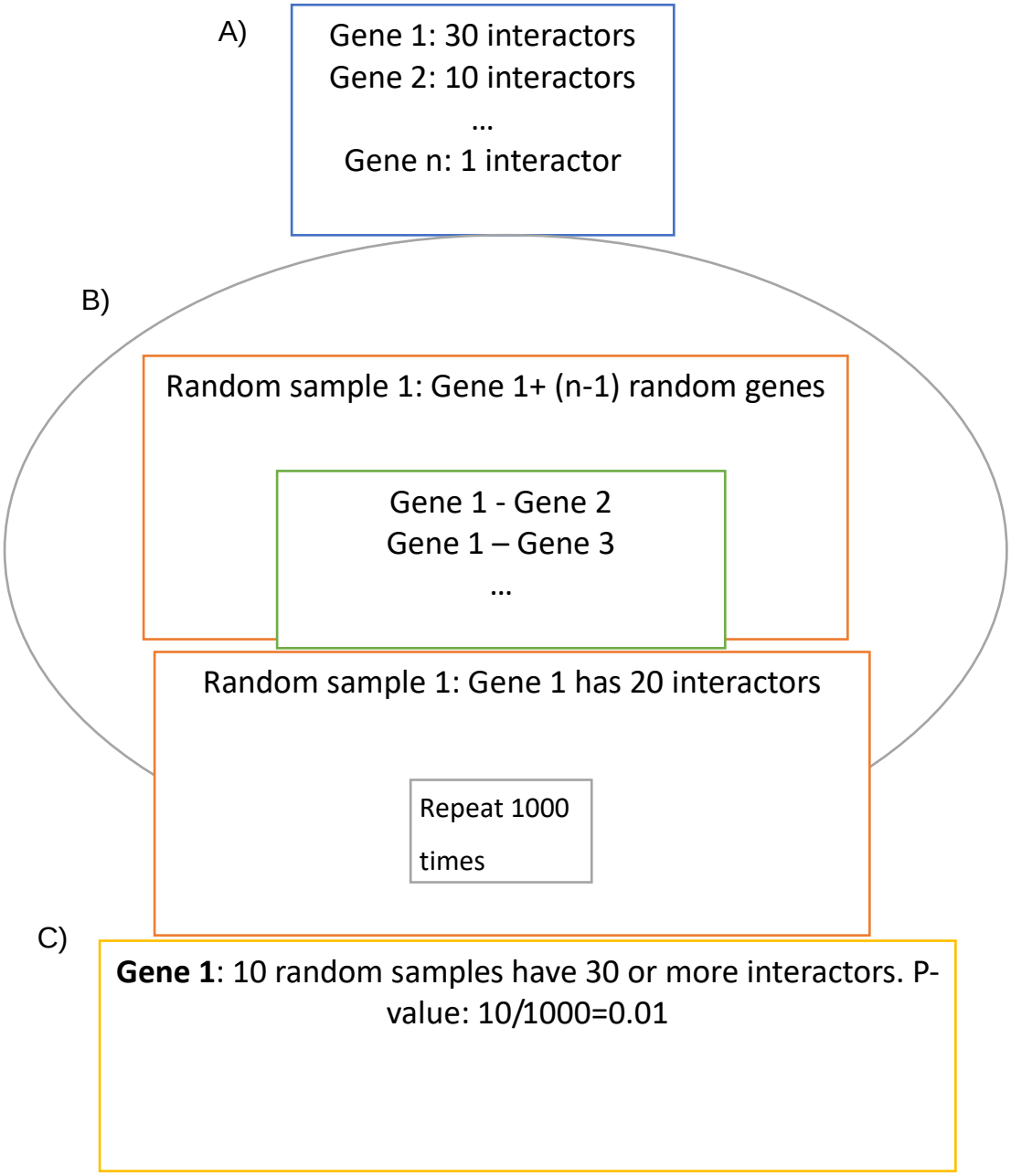



*Figure 3.28. Graphical depiction of the script used to produce random-sampling estimates of enrichment in connectors for each gene in the networks produced with FunCoup. A) List of GOI with 1 or more interacting genes in the network, containing the number of interacting genes. B) Random sample process. Briefly, a random sample of the same size as the GOI list is drawn from a list that contains all the genes targeted by at least one sequence identified in the sequencing output of the screen. One of the genes in this random sample is always the gene being assessed for enrichment. The script searches for interactions between the gene being assessed and the rest of the genes in the random sample. The interactors are counted. This process is repeated 1,000 times. C) Of the 1,000 samples the script counts how many have at least the same number of genes interacting with the gene under assessment as interactors were found for that gene in the GOI list. P-value is then calculated as the proportion of random samples with at least the same number of interactors. B and C are repeated for each gene in the GOI list.*

| Gene | Potential interactors screened | Interactors in GOI list | p-value |
|---|---|---|---|
| *Prkrir* | 1 | 1 | <0.001 |
| *Spred1* | 3 | 1 | 0.002 |
| *Zfp57* | 6 | 1 | 0.008 |
| *Dut* | 705 | 29 | 0.009 |
| *Tlr4* | 6 | 1 | 0.012 |
| *Vezf1* | 9 | 1 | 0.015 |
| *Plek* | 7 | 1 | 0.016 |
| *Pole2* | 140 | 8 | 0.017 |
| *Rab43* | 7 | 1 | 0.017 |
| *Hspe1* | 480 | 20 | 0.027 |
| *Odc1* | 86 | 5 | 0.029 |
| *Sel1l* | 228 | 10 | 0.032 |
| *Prkag1* | 253 | 11 | 0.036 |
| *Ncf4* | 35 | 2 | 0.041 |
| *Ube2v1* | 157 | 7 | 0.045 |
| *Rps5* | 853 | 31 | 0.049 |

*Table 3.17. Genes found to be enriched (p-value<0.05) in interactors in the 3T3$^{supp-}$ GOI list.*

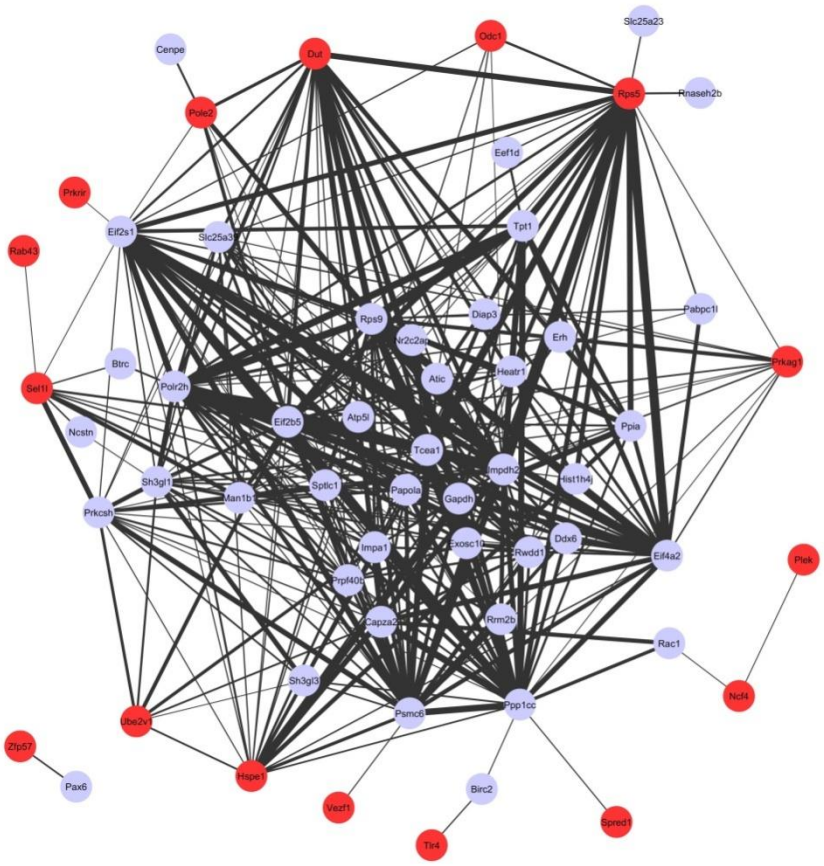



*Figure 3.29. Gene network for the 3T3$^{supp-}$ GOI list.*
*Only those genes determined to be statistically enriched in connections (red circles) and the genes interacting with them are depicted.*

| Gene | Potential interactors screened | Interactors in GOI list | p-value |
|---|---|---|---|
| *Mxi1* | 2 | 1 | 0.001 |
| *Npcd* | 10 | 2 | 0.001 |
| *Tspan7* | 4 | 1 | 0.001 |
| *Ankrd6* | 5 | 1 | 0.002 |
| *Twf2* | 209 | 10 | 0.008 |
| *Txnrd3* | 12 | 1 | 0.028 |
| *Agxt2* | 13 | 1 | 0.029 |
| *Pck2* | 14 | 1 | 0.03 |
| *Hipk2* | 13 | 1 | 0.035 |

*Table 3.18. Genes found to be enriched (p-value<0.05) in interactors in the 3T3supp+ GOI list.*

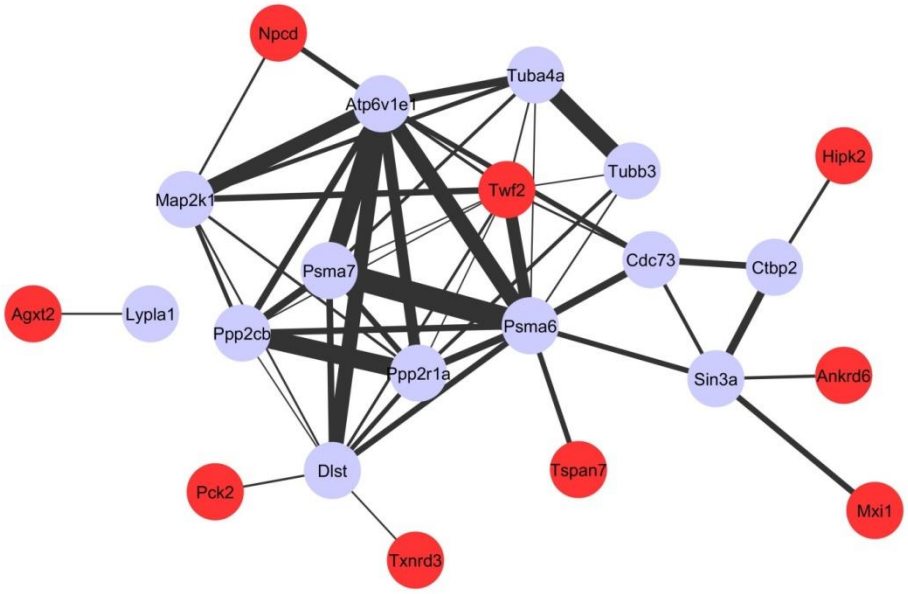



*Figure 3.30. Gene network for the 3T3$^{supp+}$ GOI list.*
*Only those genes determined to be statistically enriched in connections (red circles) and the genes interacting with them are depicted.*

| Gene | Potential interactors screened | Interactors in GOI list | P-value |
|---|---|---|---|
| *Dmpk* | 1 | 1 | <0.001 |
| *Prkrir* | 1 | 1 | <0.001 |
| *Ptma* | 1 | 1 | <0.001 |
| *Zfp57* | 6 | 2 | <0.001 |
| *Agxt2* | 13 | 3 | 0.001 |
| *Npcd* | 10 | 3 | 0.001 |
| *Mxi1* | 2 | 1 | 0.002 |
| *Spred1* | 3 | 1 | 0.004 |
| *Tspan7* | 4 | 1 | 0.004 |
| *Hsd11b1* | 3 | 1 | 0.006 |
| *Papd5* | 171 | 15 | 0.006 |
| *Wt1* | 3 | 1 | 0.009 |
| *Snap23* | 62 | 7 | 0.012 |
| *F11r* | 3 | 1 | 0.013 |
| *Pnkd* | 42 | 5 | 0.013 |
| *Tnnt2* | 10 | 2 | 0.015 |
| *Cdipt* | 198 | 16 | 0.022 |
| *Hipk2* | 13 | 2 | 0.023 |
| *Hnrnpa0* | 249 | 20 | 0.024 |
| *Tsta3* | 6 | 1 | 0.031 |
| *Gata2* | 5 | 1 | 0.032 |
| *Txnrd3* | 12 | 2 | 0.032 |
| *Tlr4* | 6 | 1 | 0.034 |
| *Sel1l* | 228 | 16 | 0.039 |
| *Zfp110* | 30 | 3 | 0.039 |
| *Pcsk1* | 6 | 1 | 0.04 |
| *Twf2* | 209 | 16 | 0.041 |
| *Dut* | 705 | 46 | 0.044 |
| *Plek* | 7 | 1 | 0.046 |

*Table 3.19. Genes found to be enriched (p-value<0.05) in interactors in the combined GOI list.*
*Shaded in blue: genes statistically enriched in connections in the 3T3$^{supp+}$ GOI list. In green: genes statistically enriched in connections in the 3T3$^{supp-}$ GOI list. In white: genes not found to be statistically enriched in their original GOI list.*

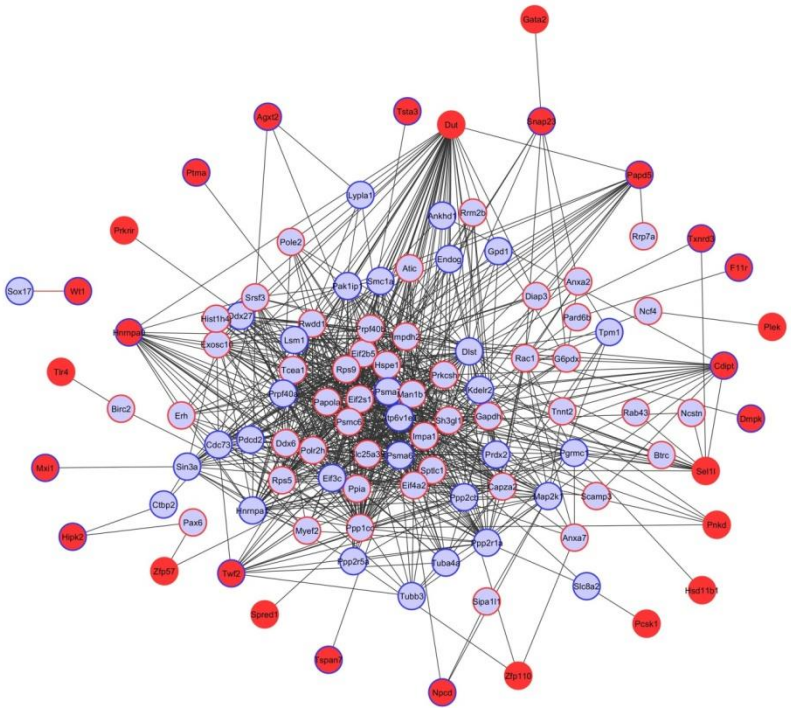



*Figure 3.31. Gene network for the combined GOI list.*
*Only those genes determined to be statistically enriched in connections (red circles) and the genes interacting with them are depicted. Circles with dark blue border are 3T3$^{supp+}$ GOI and red border marks 3T3$^{supp-}$ GOI.*

### 3.3.11. Functional enrichment analysis of GOI

Functional enrichment analysis provides information about the overrepresentation of genes involved in certain biological pathways, expressed in specific cellular components or with certain molecular functions. This type of analysis can highlight putative mechanisms at work in the phenotype studied. As the molecular mechanism of neighbour suppression is unknown, functional enrichment was used to investigate whether the data obtained from the screen could highlight any biological pathway or molecular function of importance to it. For the same reason this functional enrichment analysis was performed using different methods followed by combination of these results. The following subsections describe five functional enrichment analysis approaches (three web-based and two standalone applications) and their outcomes.

#### *3.3.11.1. Functional enrichment analysis using STRING*

STRING was used for building protein interaction networks as described in section 3.3.10. However, it also produces a functional enrichment analysis of the gene list inputted. It uses three gene ontology (GO) categories: biological process, molecular function and cellular component; the pathway database KEGG; and the protein domain databases PFAM and INTERPRO. STRING does not allow the use of a custom background for the enrichment analysis and performs it using the whole genome as the background. As such, this method may include false positive enrichments due to the targeting sequences screened providing an incomplete coverage of the mouse genome.

Functional enrichment analysis of the $3T3^{supp+}$ GOI list using STRING resulted in 64 biological process, ten molecular function, 26 cellular component and six KEGG pathway terms being significantly enriched. This enrichment was defined as a false discovery rate (FDR) under 0.05. The fifteen enriched terms with the lowest FDR from the GO categories (biological process, molecular function and cellular component) are displayed in Table 3.20. This analysis yielded some interesting enriched categories such as “homotypic cell-cell adhesion” and several categories related to the extracellular matrix and membrane-bound vesicles. Neighbour suppression has been shown to depend upon cell-cell contact between the tumorigenic cells and the cells exerting the suppression

and so these results were encouraging. Table 3.21 shows the enriched KEGG pathways for the same dataset and method of analysis.

This analysis was performed for the $3T3^{supp-}$ GOI list. In this case, only GO terms were found to be enriched in the list with no KEGG pathways passing the FDR threshold. Table 3.22 contains the fifteen enriched GO terms with the lowest FDR.

The combined GOI list generated for the network analysis in section 3.3.10.1 was also analysed for functional enrichment using STRING. Three KEGG pathways were identified as significantly enriched, with 368 biological process terms, 37 molecular function terms and 47 cellular component terms passing the threshold for enrichment in the GO terms analysis. The fifteen enriched GO terms with the lowest FDR are indicated in Table 3.23 and the enriched KEGG pathways in Table 3.24.

The functional enrichment results were analysed for overlapping and specific annotations in both co-cultures and the combined list. A Venn diagram representing the overlapping categories is depicted in Figure 3.32. All the terms in common between both co-culture samples are also enriched in the combined list. In $3T3^{supp+}$ enrichment was detected for 11 categories (Table 3.25) which were not enriched in the $3T3^{supp-}$ or the combined GOI list. These include three categories related to cancer which are encouraging as neighbour suppression is hypothesized to be relevant in the early stages of carcinogenesis. Four categories were detected as enriched exclusively in the $3T3^{supp-}$ GOI list (Table 3.26).

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0008150 | biological_process | 183 | $4.06 \times 10^{-07}$ |
| GO.0005488 | Binding | 150 | $2.20 \times 10^{-06}$ |
| GO.0097159 | organic cyclic compound binding | 96 | $2.20 \times 10^{-06}$ |
| GO.1901363 | heterocyclic compound binding | 95 | $2.20 \times 10^{-06}$ |
| GO.0065007 | biological regulation | 141 | $1.52 \times 10^{-05}$ |
| GO.0009987 | cellular process | 167 | $1.63 \times 10^{-05}$ |
| GO.0050789 | regulation of biological process | 134 | $5.54 \times 10^{-05}$ |
| GO.0005575 | cellular_component | 199 | $1.06 \times 10^{-04}$ |
| GO.0003674 | molecular_function | 162 | $2.29 \times 10^{-04}$ |
| GO.0043167 | ion binding | 93 | $5.91 \times 10^{-04}$ |
| GO.0005737 | Cytoplasm | 140 | $6.80 \times 10^{-04}$ |
| GO.0044424 | intracellular part | 163 | $6.80 \times 10^{-04}$ |
| GO.0005622 | Intracellular | 165 | $7.09 \times 10^{-04}$ |
| GO.0005623 | Cell | 178 | $7.09 \times 10^{-04}$ |
| GO.0044763 | single-organism cellular process | 138 | $8.89 \times 10^{-04}$ |

*Table 3.20. Top 15 enriched GO terms in the 3T3$^{supp+}$ GOI list ranked by their false discovery rate (FDR).*
*The enrichment analysis was performed with STRING using the whole mouse genome as the background.*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| 4151 | PI3K-Akt signaling pathway | 13 | 0.029 |
| 5200 | Pathways in cancer | 12 | 0.029 |
| 5216 | Thyroid cancer | 4 | 0.029 |
| 5222 | Small cell lung cancer | 6 | 0.029 |
| 5410 | Hypertrophic cardiomyopathy (HCM) | 6 | 0.029 |
| 5414 | Dilated cardiomyopathy | 6 | 0.030 |

*Table 3.21. Enriched KEGG pathways in the 3T3$^{supp+}$ GOI list ranked by their false discovery rate (FDR).*
*The enrichment analysis was performed with STRING using the whole mouse genome as the background.*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0003674 | molecular_function | 189 | $5.98 \times 10^{-06}$ |
| GO.0050789 | regulation of biological process | 154 | $8.00 \times 10^{-06}$ |
| GO.0050794 | regulation of cellular process | 147 | $1.04 \times 10^{-05}$ |
| GO.0019222 | regulation of metabolic process | 113 | $1.66 \times 10^{-05}$ |
| GO.0065007 | biological regulation | 155 | $2.96 \times 10^{-05}$ |
| GO.0044763 | single-organism cellular process | 159 | $4.01 \times 10^{-05}$ |
| GO.0032501 | multicellular organismal process | 105 | $5.51 \times 10^{-05}$ |
| GO.0048519 | negative regulation of biological process | 87 | $5.51 \times 10^{-05}$ |
| GO.0005623 | cell | 206 | $6.05 \times 10^{-05}$ |
| GO.0044464 | cell part | 205 | $6.05 \times 10^{-05}$ |
| GO.0044707 | single-multicellular organism process | 102 | $7.42 \times 10^{-05}$ |
| GO.0050896 | response to stimulus | 108 | $1.43 \times 10^{-04}$ |
| GO.0005488 | binding | 160 | $2.26 \times 10^{-04}$ |
| GO.0048518 | positive regulation of biological process | 96 | $2.47 \times 10^{-04}$ |
| GO.0050767 | regulation of neurogenesis | 25 | $2.47 \times 10^{-04}$ |

*Table 3.22. Top 15 enriched GO terms in the 3T3$^{supp-}$ GOI list ranked by their false discovery rate (FDR).*
*The enrichment analysis was performed with STRING using the whole mouse genome as the background.*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0050789 | regulation of biological process | 286 | $1.09 \times 10^{-12}$ |
| GO.0065007 | biological regulation | 294 | $1.16 \times 10^{-12}$ |
| GO.0005488 | binding | 309 | $2.35 \times 10^{-12}$ |
| GO.0003674 | molecular_function | 348 | $9.19 \times 10^{-12}$ |
| GO.0009987 | cellular process | 344 | $5.61 \times 10^{-11}$ |
| GO.0050794 | regulation of cellular process | 268 | $5.61 \times 10^{-11}$ |
| GO.0044763 | single-organism cellular process | 295 | $1.02 \times 10^{-10}$ |
| GO.0097159 | organic cyclic compound binding | 187 | $5.07 \times 10^{-10}$ |
| GO.0019222 | regulation of metabolic process | 203 | $5.08 \times 10^{-10}$ |
| GO.0005623 | cell | 383 | $5.80 \times 10^{-10}$ |
| GO.1901363 | heterocyclic compound binding | 184 | $9.82 \times 10^{-10}$ |
| GO.0044464 | cell part | 380 | $1.13 \times 10^{-09}$ |
| GO.0005575 | cellular_component | 413 | $1.54 \times 10^{-09}$ |
| GO.0044424 | intracellular part | 345 | $1.54 \times 10^{-09}$ |
| GO.0005634 | nucleus | 208 | $3.61 \times 10^{-09}$ |

*Table 3.23. Top 15 enriched GO terms in the combined GOI list ranked by their false discovery rate (FDR).*
*The enrichment analysis was performed with STRING using the whole mouse genome as the background.*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| 4114 | Oocyte meiosis | 10 | 0.028 |
| 4151 | PI3K-Akt signaling pathway | 21 | 0.028 |
| 5410 | Hypertrophic cardiomyopathy (HCM) | 9 | 0.028 |

*Table 3.24. Enriched KEGG pathways in the combined GOI list ranked by their false discovery rate (FDR).*
*The enrichment analysis was performed with STRING using the whole mouse genome as the background*

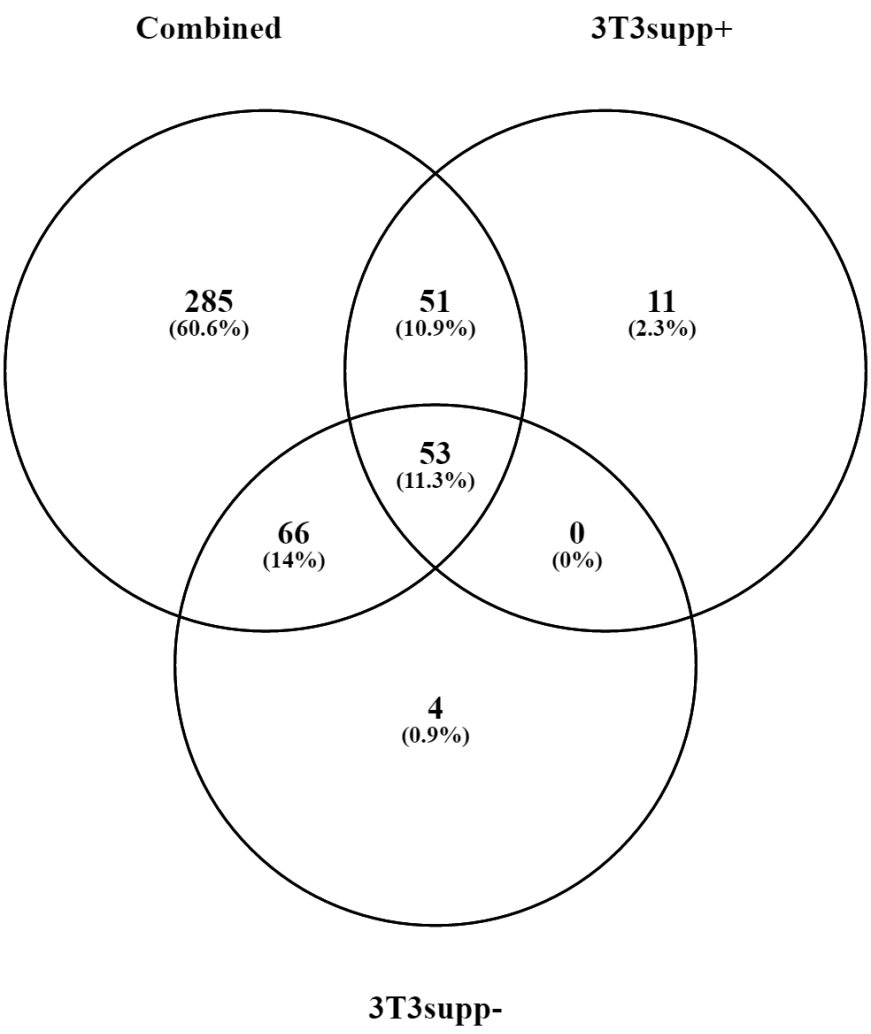



*Figure 3.32. Venn diagram showing the overlaps in enriched functional categories (GO terms and KEGG pathways) in the 3T3$^{supp+}$ and 3T3$^{supp-}$ GOI lists and the combined list containing the genes from both lists.*

| |
|---|
| nephron development |
| Pathways in cancer |
| Thyroid cancer |
| Small cell lung cancer |
| Dilated cardiomyopathy |
| extracellular matrix |
| purine nucleobase transmembrane transporter activity |
| pyrimidine nucleobase transmembrane transporter activity |
| nucleobase transmembrane transporter activity |
| positive regulation of cell-substrate adhesion |
| contractile fiber part |

*Table 3.25. Functional categories enriched in 3T3$^{supp+}$ but not in 3T3$^{supp-}$ or the combined list according to STRING.*

| |
|---|
| regulation of dendrite development |
| response to prostaglandin E |
| regulation of dendrite morphogenesis |
| semicircular canal development |

*Table 3.26. Functional categories enriched in 3T3$^{supp-}$ but not in 3T3$^{supp+}$ or the combined list according to STRING.*

#### *3.3.11.2. Functional enrichment analysis using FunCoup*

FunCoup provides some basic results on GO enrichment on the list of genes used when networks are computed. These include “Cellular component”, “Biological process” and “Molecular function”. No terms were found to be enriched in any of the GOI lists analysed. No information is available on how this analysis is performed.

#### *3.3.11.3. Functional enrichment analysis using DAVID*

DAVID (https://david.ncifcrf.gov/) is an on-line resource commonly used to perform functional annotation and enrichment analysis. However, the performance of this system is limited when long lists of genes are analysed. As the shRNA library used might already be enrichment for some functional categories, this analysis would require the use of the list of genes screened (i.e. those genes that have more than 50 reads in the sequencing) as a background. Despite various attempts to set the list of screened genes as a background in DAVID this analysis could not be performed as the server could not accept a list of genes of that size as a background. Therefore a pre-existent background had to be used. The results of the functional enrichment analysis (Table 3.27, Table 3.28, and Table 3.29) thus indicate enrichment with respect to the entire *Mus musculus* genome and may include a bias caused by the potential enrichment of the screened genes list in the same functional categories.

Only two categories were considered statistically enriched (Benjamini adjusted p-value below 0.05) in the $3T3^{supp+}$ GOI list (Table 3.27): “phosphoproteins” and “isopeptide bonds”. Using the unadjusted p-value, 122 categories were enriched in the $3T3^{supp+}$ GOI list.

No categories were found to be statistically enriched using the Benjamini adjusted p-value in the $3T3^{supp-}$ GOI list (Table 3.28). Using the unadjusted p-value to judge enrichment, 71 terms were enriched in this list.

It was reasoned that the same pathways might be of relevance for both suppressive and permissive signal integration, therefore both GOI lists were combined and analysed for functional enrichment using DAVID (Table 3.29). Three terms were found to be enriched above the statistical significance threshold using the Benjamini p-value adjustment: “phosphoproteins”,

“acetylation” and “nucleus”. Using the unadjusted p-values, 183 terms were enriched in the combined list.

The overlap of the functional enrichment analyses (unadjusted p-values below 0.05) for the different GOI lists was visualised using a Venn diagram (Figure 3.33). Only 10 terms were found to be common to all GOI lists, with 61 being shared between $3T3^{supp+}$ and the combined list and 28 between $3T3^{supp-}$ and the combined list. The terms that form part of each of the overlaps between the different GOI lists are available in Appendices 9, 10 and 11.

| Category | Term | % of GOI in the category | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 44.664 | 6.28E-05 | 1.355 | 0.018 |
| SP_PIR_KEYWORDS | isopeptide bond | 5.138 | 3.20E-04 | 3.550 | 0.046 |
| GOTERM_CC_FAT | GO:0008287~protein serine/threonine phosphatase complex | 1.976 | 9.96E-04 | 11.077 | 0.110 |
| GOTERM_CC_FAT | GO:0000159~protein phosphatase type 2A complex | 1.581 | 5.86E-04 | 23.177 | 0.128 |
| KEGG_PATHWAY | mmu05414:Dilated cardiomyopathy | 2.372 | 0.008 | 4.737 | 0.239 |
| KEGG_PATHWAY | mmu05200:Pathways in cancer | 4.348 | 0.012 | 2.474 | 0.255 |
| KEGG_PATHWAY | mmu05216:Thyroid cancer | 1.581 | 0.007 | 10.018 | 0.296 |
| KEGG_PATHWAY | mmu00310:Lysine degradation | 1.581 | 0.018 | 7.086 | 0.307 |
| SP_PIR_KEYWORDS | nucleus | 26.877 | 0.005 | 1.351 | 0.326 |
| SP_PIR_KEYWORDS | ubl conjugation | 6.324 | 0.004 | 2.310 | 0.331 |

*Table 3.27. Results of functional enrichment in the 3T3$^{supp+}$ GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value.*

| Category | Term | % of GOI in the category | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| GOTERM_CC_FAT | GO:0042995~cell projection | 5.841924 | 0.005 | 2.149 | 0.516 |
| SP_PIR_KEYWORDS | acetylation | 17.86942 | 0.003 | 1.501 | 0.527 |
| GOTERM_CC_FAT | GO:0005829~cytosol | 5.841924 | 0.004 | 2.251 | 0.605 |
| SP_PIR_KEYWORDS | dna-binding | 10.99656 | 0.017 | 1.530 | 0.716 |
| SP_PIR_KEYWORDS | cytoplasm | 20.27491 | 0.022 | 1.307 | 0.727 |
| SP_PIR_KEYWORDS | nucleus | 24.39863 | 0.027 | 1.251 | 0.739 |
| SP_PIR_KEYWORDS | CBS domain | 1.030928 | 0.032 | 10.598 | 0.742 |
| SP_PIR_KEYWORDS | nucleotide-binding | 11.68385 | 0.042 | 1.399 | 0.753 |
| SP_PIR_KEYWORDS | DNA binding | 3.092784 | 0.039 | 2.341 | 0.770 |
| SP_PIR_KEYWORDS | P-loop | 1.37457 | 0.072 | 4.130 | 0.793 |

*Table 3.28. Results of functional enrichment in the 3T3$^{supp-}$ GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value.*

| Category | Term | % of GOI in the category | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 41.58965 | 4.40E-06 | 1.276 | 0.002 |
| SP_PIR_KEYWORDS | acetylation | 17.92976 | 4.09E-05 | 1.493 | 0.008 |
| SP_PIR_KEYWORDS | nucleus | 25.69316 | 3.66E-04 | 1.306 | 0.046 |
| GOTERM_CC_FAT | GO:0005829~cytosol | 5.730129 | 1.79E-04 | 2.095 | 0.057 |
| SP_PIR_KEYWORDS | isopeptide bond | 3.512015 | 8.17E-04 | 2.454 | 0.075 |
| SP_PIR_KEYWORDS | cytoplasm | 20.7024 | 0.001 | 1.323 | 0.083 |
| SP_PIR_KEYWORDS | metal-binding | 18.48429 | 0.002 | 1.334 | 0.093 |
| SP_PIR_KEYWORDS | ubl conjugation | 5.175601 | 0.002 | 1.912 | 0.104 |
| SP_PIR_KEYWORDS | dna-binding | 10.72089 | 0.003 | 1.478 | 0.108 |
| SP_PIR_KEYWORDS | nucleotide-binding | 12.01479 | 0.003 | 1.426 | 0.117 |

*Table 3.29. Results of functional enrichment in the combined GOI list performed using the online tool DAVID, ranked by Benjamini adjusted p-value.*

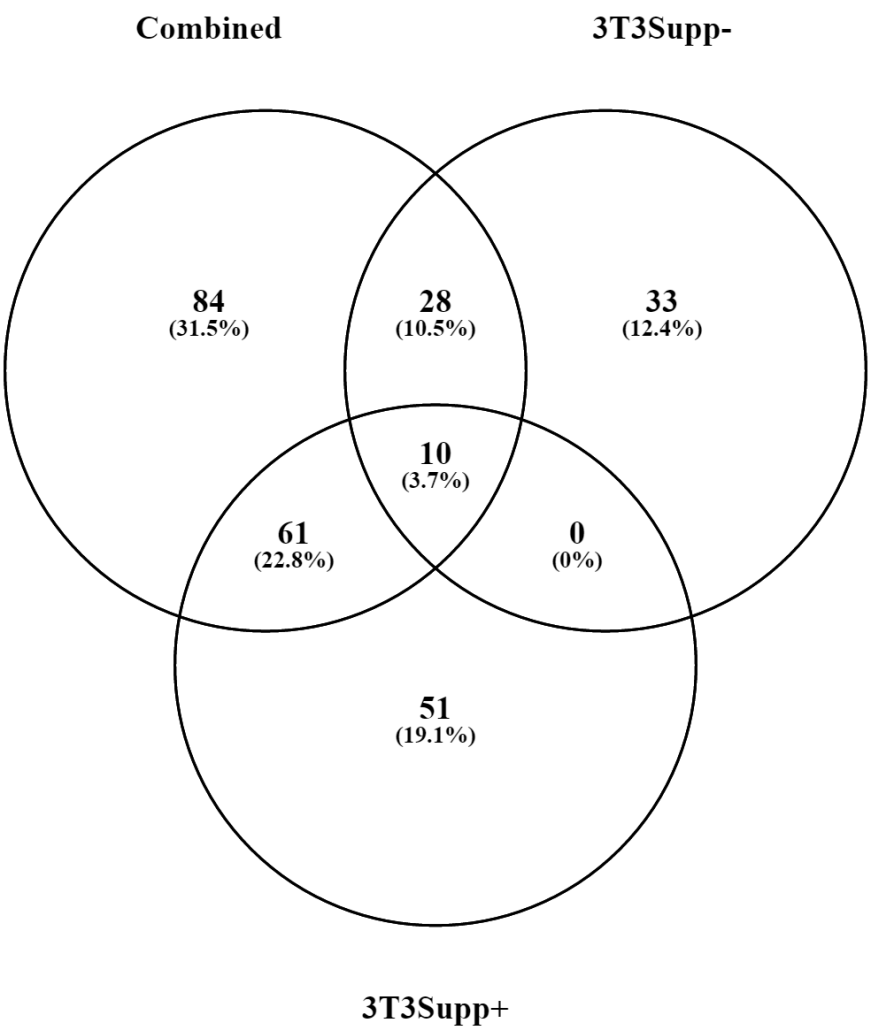



*Figure 3.33. Venn diagram showing the overlap in functional enrichment in both GOI lists and the combined list. Enrichment analysis was performed using DAVID.*

#### *3.3.11.4. Functional enrichment analysis using FunRich*

FunRich (http://funrich.org/) is a freely available functional enrichment analysis tool developed by Pathan and colleagues (Pathan, Keerthikumar et al. 2015). FunRich 2.1.2 was obtained from the FunRich website and installed locally. The GOI lists for both co-culture samples were used as input and analysed for enrichment in “Cellular component”, “Molecular function”, “Biological process”, and “Transcription factor”. This last category is equivalent to the analysis performed using Opossum in section 3.3.14.

FunRich does not accept a custom list of genes as the background for enrichment. Furthermore, its custom database integrating evidence from different publicly available sources is not available for mouse and thus the mouse gene names were converted to their homologs in human. To do this, a human-mouse homology database was downloaded from HomoloGene (http://www.ncbi.nlm.nih.gov/homologene/) and gene names were converted by merging this database with the gene lists. HomoloGene uses sequence similarity to determine homology between species. Not all genes in the shRNA library could be assigned a human homologue. Out of 13,342 genes identified in the library composition file 12,283 were assigned a human gene name.

Analysis of the $3T3^{supp+}$ GOI list highlighted one significantly enriched (Benjamini adjusted p-value < 0.05) functional term: “Protein serine/threonine phosphatase activity”. An alternative q-value is reported by FunRich, the Storey and Tibshirani q-value (Storey and Tibshirani 2003). This type of adjustment is a false discovery rate (FDR) based measure of significance which indicates the proportion of false positives expected if the functional category is regarded as significant. Using this q-value as a measurement of enrichment, 24 functional terms were found to be enriched (q-value<0.05) in the $3T3^{supp+}$ GOI list. The 10 categories with the lowest q-value are reported in Table 3.30.

The analysis of the $3T3^{supp-}$ produced no significantly enriched (Benjamini adjusted p-value < 0.05) functional categories. Four functional categories have a q-value under 0.05: “cytoplasm” and “cytoplasmic part”, “cell leading edge” and “eukaryotic translation initiation factor 2B complex”.

A combined GOI list composed of the GOI from both co-culture conditions was also analysed for enrichment using FunRich (Table 3.25). Three functional

categories have Benjamini adjusted p-value under 0.05: “cytoplasm”, “TFAP4” (indicating enrichment in binding sites for that transcription factor) and “nucleus”. Enrichment for transcription factor binding sites is the biggest contributor to the list of functional terms with q-value below 0.05, which is composed of 141 terms – 135 are transcription factors. The other six terms are “cytoplasm”, “signal transduction”, “protein serine/threonine phosphatase activity”, “cell communication” and “muscle contraction”.

Given that FunRich does not accept custom background gene lists I investigated which terms were enriched in the screened genes to determine whether the above functional enrichment resulted from the selected genes that were screened. Enrichment analysis was therefore performed on the list of all the genes that were targeted by the sequences screened (Table 3.33). A total of 156 functional terms were detected as significantly enriched (Benjamini adjusted p-value under 0.05).

To non-formally assess the differences in enriched terms between the screened genes and the co-culture GOI lists, all functional categories with an unadjusted p-value below 0.05 were examined for overlap amongst the lists (Figure 3.34). Three functional categories were found to be enriched in all lists of genes using that threshold: “cytoplasm”, “signal transduction” and “TFAP4”. The number of functional terms found to be enriched in either of the screening co-culture conditions and the screened gene list are quite similar: 17 in common between the screened genes and the $3T3^{supp+}$ GOI list (Table 3.34) and 19 between the screened genes and the $3T3^{supp-}$ GOI list (Table 3.35). This overlap analysis indicates that half of the functional categories enriched in the $3T3^{supp-}$ GOI list and almost half in the $3T3^{supp+}$ GOI list are enriched in the list of screened genes. Those categories might therefore be enriched in the GOI list due to the enrichment in the background. Furthermore, all the categories enriched in both GOI lists are also enriched in the background, suggesting there might not be an overlap in the mechanism behind both phenotypes.

The list of functional categories enriched exclusively in the $3T3^{supp-}$ (Table 3.36) is composed amongst others of transcription factors and cytoskeleton, membrane and vesicle-related terms. This may suggest a role of cell membrane and motility in the formation of tumorigenic cell colonies in a permissive co-culture.

The categories enriched exclusively in the $3T3^{supp+}$ GOI list (Table 3.37) however are mostly transcription factors, extracellular matrix component and chromosome-related terms, suggesting an involvement of transcription events and importance of the extracellular matrix in the response of tumorigenic cells to a suppressive cue by neighbouring cells.

| Term | % of GOI in the term | Fold enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| Protein serine/threonine phosphatase activity | 2.119 | 10.666 | <0.001 | 0.016 | 0.007 |
| Protein phosphatase type 2A complex | 1.508 | 21.444 | 0.001 | 0.079 | 0.021 |
| BRCA1-A complex | 1.005 | 28.325 | 0.004 | 0.089 | 0.024 |
| Cytoplasm | 48.744 | 1.254 | 0.003 | 0.101 | 0.024 |
| Basement membrane | 1.508 | 13.734 | 0.003 | 0.112 | 0.024 |
| DNA polymerase III complex | 0.503 | 71.901 | 0.014 | 0.124 | 0.033 |
| Eukaryotic translation initiation factor 4F complex | 1.005 | 22.798 | 0.006 | 0.125 | 0.033 |
| Chromosome | 2.010 | 6.265 | 0.009 | 0.130 | 0.033 |
| Dystrophin-associated glycoprotein complex | 1.005 | 20.772 | 0.008 | 0.133 | 0.033 |
| Cell communication | 26.695 | 1.316 | 0.014 | 0.133 | 0.032 |

*Table 3.30. Top 10 functional terms (by Benjamini p-value) enriched in the 3T3$^{supp+}$ GOI list.*
*The analysis was performed with FunRich*

| Term | % of GOI in the term | Fold enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| Cytoplasm | 49.351 | 1.268 | 0.001 | 0.108 | 0.023 |
| Cell leading edge | 0.866 | 27.791 | 0.004 | 0.144 | 0.030 |
| Cytoplasmic part | 0.866 | 21.782 | 0.007 | 0.197 | 0.041 |
| Ion channel activity | 1.866 | 6.277 | 0.003 | 0.198 | 0.110 |
| Eukaryotic translation initiation factor 2B complex | 0.866 | 27.791 | 0.004 | 0.216 | 0.030 |
| Muscle contraction | 0.373 | 23.955 | 0.073 | 0.254 | 0.082 |
| Transcription factor activity | 8.209 | 1.848 | 0.007 | 0.258 | 0.143 |
| Electron transport | 0.373 | 20.651 | 0.086 | 0.269 | 0.087 |
| Vesicle docking | 0.373 | 28.518 | 0.059 | 0.273 | 0.082 |
| Regulation of immune response | 0.373 | 16.186 | 0.114 | 0.289 | 0.093 |

*Table 3.31. Top 10 functional terms in the 3T3$^{supp-}$ GOI list ranked by increasing Benjamini adjusted p-values. The analysis was performed using FunRich.*

| Term | % of GOI in the term | Fold enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| Cytoplasm | 49.065 | 1.258 | $1.42x10^{-05}$ | 0.003 | 0.001 |
| TFAP4 | 13.304 | 1.608 | <0.001 | 0.048 | 0.003 |
| Signal transduction | 27.545 | 1.271 | 0.001 | 0.049 | 0.016 |
| Transcription factor activity | 7.784 | 1.715 | 0.001 | 0.060 | 0.037 |
| Muscle contraction | 0.399 | 18.549 | 0.007 | 0.098 | 0.031 |
| Protein serine/threonine phosphatase activity | 1.198 | 5.845 | 0.001 | 0.110 | 0.037 |
| Receptor signaling complex scaffold activity | 3.593 | 2.138 | 0.004 | 0.128 | 0.079 |
| Cell communication | 25.150 | 1.231 | 0.007 | 0.139 | 0.031 |
| Eukaryotic translation initiation factor 4F complex | 0.701 | 13.908 | 0.002 | 0.170 | 0.054 |
| Translation regulator activity | 1.597 | 3.227 | 0.007 | 0.174 | 0.108 |

*Table 3.32. Top 10 functional terms in the combined GOI list ranked by increasing Benjamini adjusted p-values. The analysis was performed using FunRich*

| Term | % of GOI in the term | Fold enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| Cytoplasm | 43.833 | 1.121 | $4.15x10^{-37}$ | $2.83x10^{-34}$ | $9.8x10^{-35}$ |
| Transport | 8.767 | 1.298 | $8.77x10^{-27}$ | $1.34x10^{-24}$ | $9.82x10^{-25}$ |
| Signal transduction | 25.073 | 1.145 | $4.72x10^{-26}$ | $3.61x10^{-24}$ | $2.64x10^{-24}$ |
| Energy pathways | 11.265 | 1.245 | $6.23x10^{-25}$ | $3.18x10^{-23}$ | $2.33x10^{-23}$ |
| Metabolism | 11.558 | 1.238 | $1.89x10^{-24}$ | $7.22x10^{-23}$ | $5.28x10^{-23}$ |
| Plasma membrane | 27.362 | 1.138 | $1.06x10^{-24}$ | $3.61x10^{-22}$ | $1.25x10^{-22}$ |
| Cell communication | 23.599 | 1.147 | $1.96x10^{-23}$ | $5.99x10^{-22}$ | $4.38x10^{-22}$ |
| SP1 | 50.697 | 1.081 | $1.03x10^{-22}$ | $2.11x10^{-22}$ | $3.17x10^{-21}$ |
| Transcription factor activity | 5.976 | 1.282 | $7.82x10^{-17}$ | $1.62x10^{-14}$ | $1.09x10^{-14}$ |
| EGR1 | 23.670 | 1.123 | $2.08x10^{-16}$ | $2.13x10^{-14}$ | $3.2x10^{-14}$ |

*Table 3.33. Top 10 functional terms in the list of screened genes ranked by increasing Benjamini adjusted p-values. The analysis was performed using FunRich*

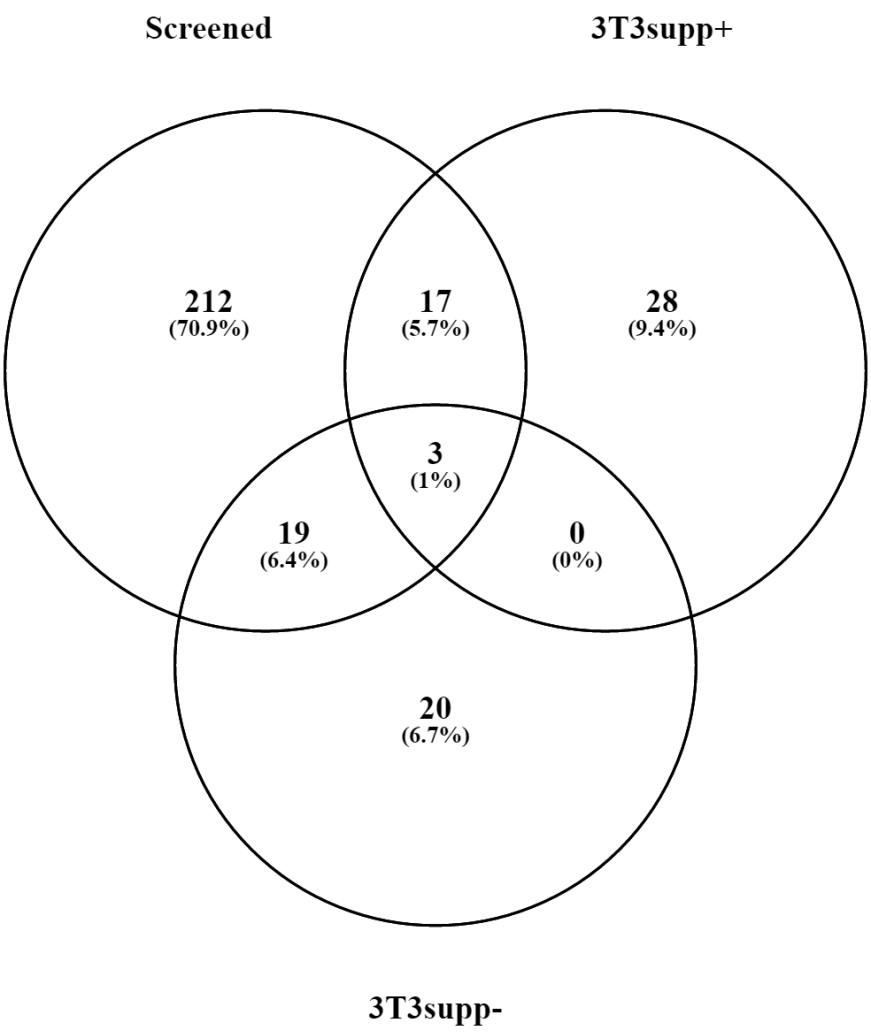



*Figure 3.34. Venn diagram showing the overlap in enriched functional terms (unadjusted p-value<0.05) amongst the GOI lists (3T3$^{supp+}$ and 3T3$^{supp-}$) and the genes targeted by at least one sequence present in the screen sequencing output (Screened).*

| |
|---|
| Cell communication |
| RREB1 |
| Receptor signaling complex scaffold activity |
| Nucleus |
| IRF1 |
| RXRA |
| Cell surface |
| ELF3 |
| HOXA3 |
| REST |
| Chromosome |
| Voltage-gated potassium channel complex |
| FEV |
| Dystrophin-associated glycoprotein complex |
| Protein phosphatase type 2A complex |
| VSX2 |
| ARID3A |

*Table 3.34. Functional terms enriched (unadjusted p-value <0.05) in both the 3T3$^{supp+}$ GOI list and the list of genes targeted by the screened sequences. Enrichment analysis performed using FunRich.*

| |
|---|
| Transcription factor activity |
| Exosomes |
| Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism |
| JUN |
| JUNB |
| JUND |
| FOS |
| FOSB |
| ONECUT1 |
| Soluble fraction |
| Translation regulator activity |
| FOXA1 |
| NKX2-1 |
| Cytoplasmic part |
| Centrosome |
| PDX1 |
| E2F1 |
| GTPase activity |
| HOXD8 |

*Table 3.35. Functional terms enriched (unadjusted p-value <0.05) in both the 3T3$^{supp-}$ GOI list and the list of genes targeted by the screened sequences.*
*Enrichment analysis performed using FunRich.*

| |
|---|
| Ion channel activity |
| Cell leading edge |
| Eukaryotic translation initiation factor 2B complex |
| FOXJ1 |
| Ruffle membrane |
| HMX1 |
| PLAG1 |
| Dense-core vesicle (None) |
| Dynein complex |
| SPOTS complex |
| Tubulin complex |
| Cytoplasmic membrane-bounded vesicle |
| Cytoskeleton |
| Binding |
| Biological_process |
| Humoral immune response |
| Ion transporter activity |
| Eukaryotic translation elongation factor 1 complex |
| Vesicular fraction |
| Condensed chromosome, centromeric region |

*Table 3.36. Functional terms enriched (unadjusted p-value <0.05) exclusively in the 3T3$^{supp-}$ GOI list.*
*Enrichment analysis performed using FunRich*

| |
|---|
| Basement membrane |
| BRCA1-A complex |
| Eukaryotic translation initiation factor 4F complex |
| Condensed nuclear chromosome |
| Nuclear body |
| Cohesin core heterodimer |
| Collagen type VII |
| DNA polymerase III complex |
| Protein phosphatase type 1 complex |
| M band |
| BRCA1-BARD1 complex |
| Ribosome |
| Protein serine/threonine phosphatase activity |
| Catalase activity |
| Kinase activity |
| Water channel activity |
| Cytoskeletal anchoring activity |
| Extracellular matrix structural constituent |
| Protein kinase activity |
| Helicase activity |
| Glycoprotein metabolism |
| Bone remodeling |
| ZNF143 |
| HOXB3 |
| POU6F1 |
| FOXJ2 |
| HOXC10 |
| HOXB4 |

*Table 3.37. Functional terms enriched (unadjusted p-value <0.05) exclusively in the 3T3$^{supp+}$ GOI list.*
*Enrichment analysis performed using FunRich*

#### *3.3.11.5. Gene set enrichment analysis in the GOI lists using GSEA*

Functional enrichment analysis provides insight into the functional categories which may be of importance for a mechanism. However, a plethora of studies have been published in the last years highlighting lists of genes which are thought to be involved in a high number of phenotypes. These studies are not contemplated as functional categories and therefore not included in functional enrichment analysis approaches. Gene set enrichment analysis on the other hand does not rely on the same functional categories but includes gene sets defined by the literature. Given that a molecular mechanism has not as yet been proposed for neighbour suppression it was reasoned that a gene set enrichment analysis approach may identify phenotypes with mechanistic similarities to neighbour suppression and highlight pathways with an alternative approach to those used in the last sections.

To study the potential enrichment of gene set members amongst the GOI, GSEA software (Broad Institute, available for download at http://software.broadinstitute.org/gsea) was used.

Two factors were of most importance when using this software for analysis of the GOI list. Firstly, GSEA was originally designed to work with expression data. Secondly, the software does not analyse repeated gene names. This last issue is usually solved inside the software by collapsing the observations with the same gene name into one consensus observation which is assigned either the mean or the median value of the observations. This was originally used for expression array data where a single gene may be analysed by more than one probe but is applicable to shRNA screening where a single gene is targeted by more than one sequence. However, the first problem requires more consideration as a two-fold increase in the metric inputted would be interpreted as a two-fold increase in expression. z-scores, as used in the screen, cannot be directly assimilated to changes in expression, and the differences in the distributions of z-scores in both co-culture samples suggest that they cannot approximate a change of phenotype as used by GSEA.

To avoid using z-scores as expression data, the GSEA Preranked tool was used. This tool uses a list of user-ranked genes and analyses the enrichment of members of gene sets in both tails of the rank, but it does not use a numeric

value for anything apart from sorting the observations. To circumvent the problems arising from having several constructs targeting the same gene, constructs were ordered by their z-score and then duplicates were removed retaining the first construct to appear in the list and remove any repeats further down the list. This approach was deemed as the least stringent as it would favour extreme z-scores, however it is expected that extreme z-scores are produced by shRNA sequences with functional effects on gene expression. This approach could generate a bias by favouring positive z-scores when more extremely negative z-scores are present for the same gene, and so the absolute value of z-score was explored as an alternative metric to use to remove duplicates. As an alternative to these approaches using a metric to decide which constructs would remain in the dataset, alphabetic sorting by Excel was also considered. Lists were ordered alphabetically using the function implemented in Excel and duplicates were then removed as above.

Two approaches were used to rank the constructs after removing duplicates. Initially, two lists were produced where genes were ranked on either $3T3^{supp+}$ or $3T3^{supp-}$ z-scores, and then analysed separately. To rank genes in an order that was consistent with the analysis method used, a new metric named “z-score distance” was calculated as the difference in z-score between $3T3^{supp+}$ and $3T3^{supp-}$.

GSEA Preranked outputs a report indicating the p-values for enrichment of the different gene sets analysed in the top or bottom of the GOI list. However, this is not equivalent to finding enrichment in the GOI list with respect to the genes screened but rather to enrichment of gene sets in high z-scores in a pre-defined list of GOI. With this in mind, the entire list of constructs screened was used as the GOI list, ordered as indicated above. The different analyses performed were:

1) Genes screened, ranked by $3T3^{supp+}$ z-scores, duplicates removed using $3T3^{supp+}$ z-scores.

2) Genes screened, ranked by $3T3^{supp+}$ z-scores, duplicates removed using the absolute value of $3T3^{supp+}$ z-scores.

3) Genes screened, ranked by $3T3^{supp+}$ z-scores, duplicates removed using the Excel alphabetical sorting.

4) Genes screened, ranked by $3T3^{supp-}$ z-scores, duplicates removed using $3T3^{supp-}$ z-scores.

5) Genes screened, ranked by $3T3^{supp-}$ z-scores, duplicates removed using the absolute value of $3T3^{supp-}$ z-scores.

6) Genes screened, ranked by $3T3^{supp-}$ z-scores, duplicates removed using the Excel alphabetical sorting.

7) Genes screened, ranked by ($3T3^{supp+}$-$3T3^{supp-}$) z-scores, duplicates removed using ($3T3^{supp+}$-$3T3^{supp-}$) z-scores.

8) Genes screened, ranked by ($3T3^{supp+}$-$3T3^{supp-}$) z-scores, duplicates removed using the absolute value of ($3T3^{supp+}$-$3T3^{supp-}$) z-scores.

9) Genes screened, ranked by ($3T3^{supp+}$-$3T3^{supp-}$) z-scores, duplicates removed using the Excel alphabetical sorting.

GSEA Preranked orders the input files by descending order, meaning that the constructs would be ordered from higher to lower z-score. Thus, gene sets were considered enriched in the $3T3^{supp+}$ GOI list in cases 1, 2 and 3 above if they were enriched on the top of the appropriate list (higher z-scores). Equivalently, gene sets were considered enriched in the $3T3^{supp-}$ GOI list in cases 4, 5 and 6 if they were enriched on the bottom of the appropriate list (lower z-scores). When analysis was performed using the differences in z-scores (cases 7, 8 and 9 above), gene set enrichment was considered only at the top of the list as the bottom of the list would be composed of constructs with no effects. This ranking prioritizes gene sets enriched in constructs with an effect on both conditions.

The enrichment analysis in $3T3^{supp+}$ identified 98 gene sets enriched in case 1 (Table 3.38), 45 in case 2 (Table 3.39) and 101 in case 3 (Table 3.40). The overlap of the gene sets detected as being enriched with a p-value lower than 0.05 is shown in Figure 3.35. The absolute (z-score) method is the most conservative, producing the lowest overlaps with the two other methods. Of all the gene sets that this method shares with the other two, only one gene set is shared between the absolute (z-score) results and only one of the other methods (z-score). This gene set is “KEGG_THYROID_CANCER” and its p-value in the alphabetical method is 0.051. The gene sets identified using all three approaches (Table 3.41) include a mix of cancer-related terms and

various gene sets with no apparent relationship to each other. Although the apparent robustness of cancer gene sets may be encouraging it could also be a result of a high number of studies being focused in the identification of de-regulated pathways in carcinogenesis.

The enrichment analysis in $3T3^{supp-}$ identified 130 gene sets enriched in case 4 (Table 3.42), 86 in case 5 (Table 3.43) and 122 in case 6 (Table 3.44). Figure 3.36 depicts the overlap of the gene sets determined as enriched in $3T3^{supp-}$. The absolute (z-score) method is again the most conservative with a lower number of gene sets detected as enriched. The overlap between the different combinations of methods is lower than for $3T3^{supp+}$, although the number of gene sets identified using all approaches (Table 3.45) is the same. There are less cancer-related gene sets, which may indicate that the gene set enrichment in $3T3^{supp+}$ is not totally derived from a high number of studies in cancer as suggested above. The list of enriched gene sets identified with all approaches in this case is more heterogeneous, with a higher number of "classical" pathways identified (i.e. pathways defined by KEGG, Reactome and Biocarta, susceptible of analysis with functional enrichment). These pathways include the Rac1 and TNF pathways amongst others.

Using the difference in z-scores to rank the screened genes detected 147 gene sets enriched in the higher ranks of the list when duplicate gene names were resolved by removing the constructs with lower difference in z-scores (Table 3.46, case 7). Using the absolute difference in z-score to remove duplicates (case 8), 156 gene sets were detected to be enriched in the top of the list (Table 3.47). In case 9 above, 181 gene sets were detected as enriched in constructs with high difference in z-scores (Table 3.48). The degree of overlap in enriched gene sets detected with these three methods (Figure 3.37) follows the same tendencies as the individual z-score methods: using the absolute difference in z-scores to remove duplicate gene names produces a list of enriched gene sets which overlap the least with the other two. Table 3.49 contains the list of gene sets detected using all three methods. The list is again a mix of gene sets most of which do not have an apparent relationship. Two of the gene sets are cancer-related, one refers to cell death and two to signalling by GPCR. The gene sets found to be enriched using this difference of z-scores approach may represent an unspecific response mechanism to both

phenomena (response to suppressive cues and growth in absence of suppressive cues).

| **Name of gene set** | **p-value** |
|---|---|
| REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES | <0.001 |
| KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_UP | <0.001 |
| MIKKELSEN_MCV6_ICP_WITH_H3K27ME3 | <0.001 |
| KIM_MYCN_AMPLIFICATION_TARGETS_DN | <0.001 |
| GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_DN | <0.001 |
| LIN_MELANOMA_COPY_NUMBER_DN | 0.002 |
| WANG_BARRETTS_ESOPHAGUS_UP | 0.002 |
| PID_PI3KCI_PATHWAY | 0.004 |
| AKL_HTLV1_INFECTION_UP | 0.006 |
| ASTON_MAJOR_DEPRESSIVE_DISORDER_DN | 0.006 |

*Table 3.38. Top 10 enriched gene sets in $3T3^{supp+}$. Duplicate gene names were removed using their alphabetical order.*

| **Name of gene set** | **p-value** |
|---|---|
| REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES | <0.001 |
| KIM_MYCN_AMPLIFICATION_TARGETS_DN | <0.001 |
| MIKKELSEN_MCV6_ICP_WITH_H3K27ME3 | <0.001 |
| LIN_MELANOMA_COPY_NUMBER_DN | <0.001 |
| KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_UP | 0.002 |
| GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_DN | 0.002 |
| REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S | 0.004 |
| KEGG_VIBRIO_CHOLERAE_INFECTION | 0.004 |
| JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_DN | 0.006 |
| PENG_LEUCINE_DEPRIVATION_UP | 0.006 |

*Table 3.39. Top 10 enriched gene sets in 3T3$^{supp+}$. Duplicate gene names were solved by removing the construct with the lowest 3T3$^{supp+}$ z-score.*

| **Name of gene set** | **p-value** |
|---|---|
| MIKKELSEN_MCV6_ICP_WITH_H3K27ME3 | <0.001 |
| RICKMAN_TUMOR_DIFFERENTIATED_WELL_VS_MODERATELY_DN | 0.002 |
| REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS | 0.002 |
| LIN_MELANOMA_COPY_NUMBER_DN | 0.004 |
| KYNG_ENVIRONMENTAL_STRESS_RESPONSE_NOT_BY_GAMMA_IN_WS | 0.004 |
| KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_UP | 0.004 |
| HU_GENOTOXIN_ACTION_DIRECT_VS_INDIRECT_4HR | 0.004 |
| PENG_LEUCINE_DEPRIVATION_UP | 0.004 |
| ELLWOOD_MYC_TARGETS_DN | 0.006 |
| REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S | 0.006 |

*Table 3.40. Top 10 enriched gene sets in 3T3$^{supp+}$. Duplicate gene names were solved by removing the construct with the lowest absolute 3T3$^{supp+}$ z-score.*

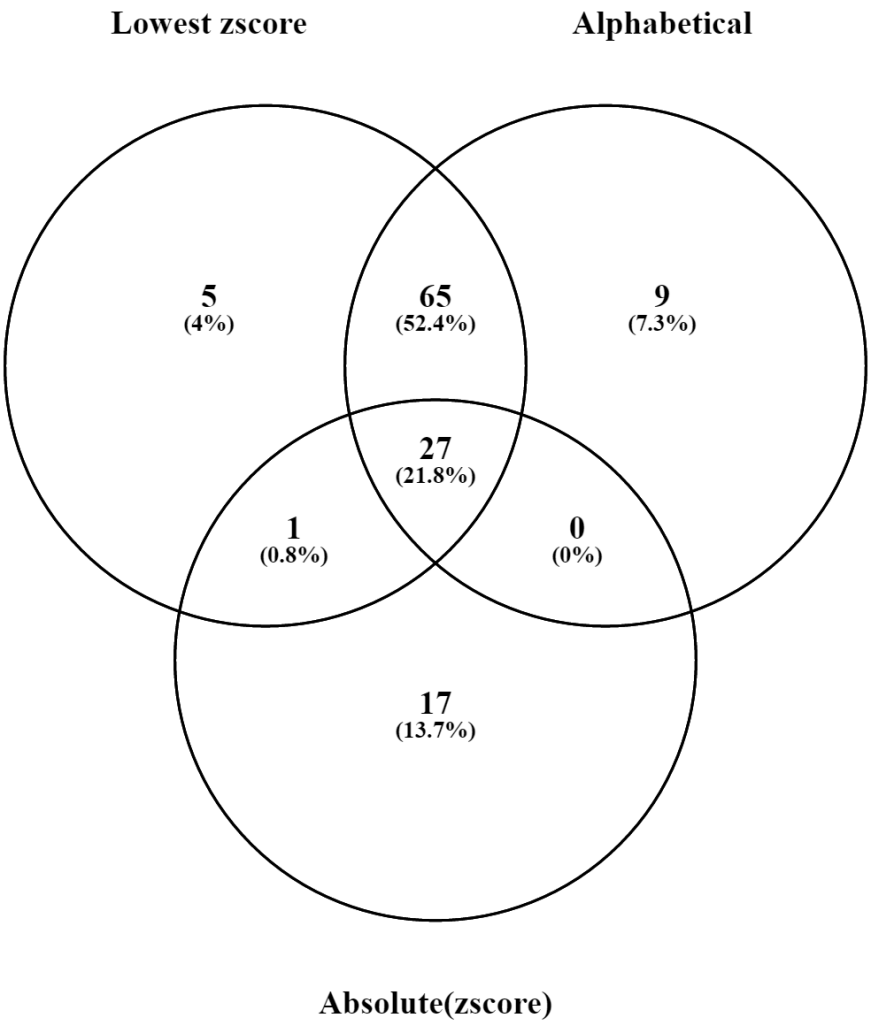



*Figure 3.35. Venn diagram showing the overlapping results of the 3 methods used to analyse the 3T3$^{supp+}$ results with GSEA.*

| |
|---|
| KIM_MYCN_AMPLIFICATION_TARGETS_DN |
| MIKKELSEN_MCV6_ICP_WITH_H3K27ME3 |
| LIN_MELANOMA_COPY_NUMBER_DN |
| KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_UP |
| REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S |
| PENG_LEUCINE_DEPRIVATION_UP |
| BREDEMEYER_RAG_SIGNALING_VIA_ATM_NOT_VIA_NFKB_UP |
| REACTOME_INWARDLY_RECTIFYING_K_CHANNELS |
| ELLWOOD_MYC_TARGETS_DN |
| KAYO_CALORIE_RESTRICTION_MUSCLE_UP |
| SERVITJA_LIVER_HNF1A_TARGETS_DN |
| MAGRANGEAS_MULTIPLE_MYELOMA_IGG_VS_IGA_DN |
| KAMIKUBO_MYELOID_CEBPA_NETWORK |
| RORIE_TARGETS_OF_EWSR1_FLI1_FUSION_DN |
| ROZANOV_MMP14_TARGETS_DN |
| RICKMAN_TUMOR_DIFFERENTIATED_WELL_VS_MODERATELY_DN |
| SATO_SILENCED_BY_METHYLATION_IN_PANCREATIC_CANCER_1 |
| KYNG_ENVIRONMENTAL_STRESS_RESPONSE_NOT_BY_GAMMA_IN_WS |
| KUMAR_AUTOPHAGY_NETWORK |
| BIOCARTA_IGF1MTOR_PATHWAY |
| SASSON_RESPONSE_TO_FORSKOLIN_UP |
| REACTOME_G_ALPHA_S_SIGNALLING_EVENTS |
| REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TRIC |
| ZHANG_TLX_TARGETS_36HR_UP |
| LINDSTEDT_DENDRITIC_CELL_MATURATION_D |
| MODY_HIPPOCAMPUS_PRENATAL |
| HU_GENOTOXIN_ACTION_DIRECT_VS_INDIRECT_4HR |

*Table 3.41. Gene sets identified by GSEA as enriched in the 3T3$^{supp+}$ co-culture results using three different duplicate gene name removal systems.*

| **Name of gene set** | **p-value** |
|---|---|
| BIOCARTA_NFKB_PATHWAY | <0.001 |
| PID_RAC1_PATHWAY | <0.001 |
| DAIRKEE_CANCER_PRONE_RESPONSE_BPA_E2 | <0.001 |
| REACTOME_INNATE_IMMUNE_SYSTEM | <0.001 |
| PID_BCR_5PATHWAY | 0.002 |
| FORTSCHEGGER_PHF8_TARGETS_UP | 0.002 |
| VERHAAK_GLIOBLASTOMA_NEURAL | 0.002 |
| SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6 | 0.002 |
| BILANGES_SERUM_RESPONSE_TRANSLATION | 0.002 |
| RAO_BOUND_BY_SALL4_ISOFORM_B | 0.002 |

*Table 3.42. Top 10 enriched gene sets in 3T3$^{supp-}$. Duplicate gene names were removed using their alphabetical order*

| **Name of gene set** | **p-value** |
|---|---|
| BIOCARTA_PPARA_PATHWAY | <0.001 |
| REACTOME_INNATE_IMMUNE_SYSTEM | <0.001 |
| BIOCARTA_KERATINOCYTE_PATHWAY | <0.001 |
| VANHARANTA_UTERINE_FIBROID_DN | <0.001 |
| BIOCARTA_RACCYCD_PATHWAY | <0.001 |
| PID_TNF_PATHWAY | 0.002 |
| RAO_BOUND_BY_SALL4_ISOFORM_B | 0.002 |
| PID_AVB3_OPN_PATHWAY | 0.002 |
| SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6 | 0.002 |
| VERHAAK_GLIOBLASTOMA_NEURAL | 0.004 |

*Table 3.43. Top 10 enriched gene sets in $3T3^{supp-}$. Duplicate gene names were solved by removing the construct with the highest $3T3^{supp-}$ z-score*

| **Name of gene set** | **p-value** |
|---|---|
| PID_TNF_PATHWAY | <0.001 |
| PID_CDC42_PATHWAY | <0.001 |
| BIOCARTA_NFKB_PATHWAY | <0.001 |
| KUROKAWA_LIVER_CANCER_CHEMOTHERAPY_DN | <0.001 |
| ST_TUMOR_NECROSIS_FACTOR_PATHWAY | <0.001 |
| PID_PRL_SIGNALING_EVENTS_PATHWAY | <0.001 |
| ACEVEDO_FGFR1_TARGETS_IN_PROSTATE_CANCER_MODEL_UP | <0.001 |
| REACTOME_INNATE_IMMUNE_SYSTEM | 0.002 |
| KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY | 0.002 |
| LIEN_BREAST_CARCINOMA_METAPLASTIC | 0.002 |

*Table 3.44. Top 10 enriched gene sets in $3T3^{supp-}$. Duplicate gene names were solved by removing the construct with the highest absolute $3T3^{supp-}$ z-score.*

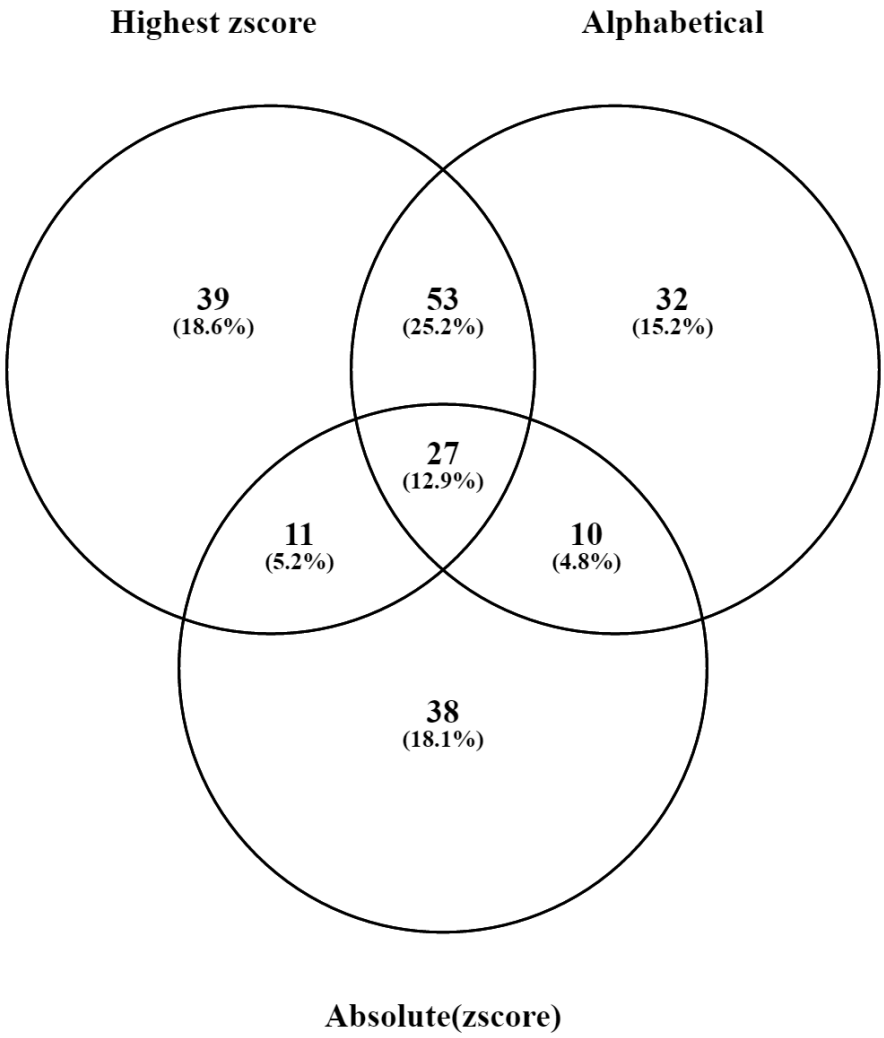



*Figure 3.36. Venn diagram showing the overlapping results of the 3 methods used to analyse the 3T3$^{supp-}$ results with GSEA.*

| |
|---|
| BIOCARTA_PPARA_PATHWAY |
| REACTOME_INNATE_IMMUNE_SYSTEM |
| BIOCARTA_KERATINOCYTE_PATHWAY |
| VANHARANTA_UTERINE_FIBROID_DN |
| BIOCARTA_RACCYCD_PATHWAY |
| PID_TNF_PATHWAY |
| PID_AVB3_OPN_PATHWAY |
| SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6 |
| VERHAAK_GLIOBLASTOMA_NEURAL |
| KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY |
| PID_RAC1_PATHWAY |
| PID_CDC42_PATHWAY |
| YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_1 |
| ST_TUMOR_NECROSIS_FACTOR_PATHWAY |
| SENESE_HDAC3_TARGETS_UP |
| REACTOME_P75_NTR_RECEPTOR_MEDIATED_SIGNALLING |
| LIAN_LIPA_TARGETS_3M |
| WOOD_EBV_EBNA1_TARGETS_DN |
| PID_S1P_S1P1_PATHWAY |
| COULOUARN_TEMPORAL_TGFB1_SIGNATURE_UP |
| ZHANG_TLX_TARGETS_60HR_DN |
| SIG_CD40PATHWAYMAP |
| PLASARI_TGFB1_SIGNALING_VIA_NFIC_1HR_DN |
| TING_SILENCED_BY_DICER |
| BIOCARTA_TOLL_PATHWAY |
| REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR 4_SIGNALING_REPERTOIRE |
| BILANGES_SERUM_RESPONSE_TRANSLATION |

*Table 3.45. Gene sets identified by GSEA as enriched in the 3T3$^{supp-}$ co-culture results using three different duplicate gene name removal systems.*

| NAME | P-value |
|---|---|
| REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE | <0.001 |
| COLIN_PILOCYTIC_ASTROCYTOMA_VS_GLIOBLASTOMA_DN | <0.001 |
| REACTOME_CELL_DEATH_SIGNALLING_VIA_NRAGE_NRIF_AND_NADE | <0.001 |
| KEGG_MAPK_SIGNALING_PATHWAY | <0.001 |
| KEGG_CELL_CYCLE | <0.001 |
| BIOCARTA_MAPK_PATHWAY | <0.001 |
| REACTOME_TRIF_MEDIATED_TLR3_SIGNALING | 0.002 |
| BIOCARTA_TOLL_PATHWAY | 0.002 |
| BURTON_ADIPOGENESIS_10 | 0.002 |
| ST_DIFFERENTIATION_PATHWAY_IN_PC12_CELLS | 0.002 |

*Table 3.46. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene names were solved by removing the construct with the lowest difference in z-scores.*

| NAME | P-value |
|---|---|
| KEGG_PURINE_METABOLISM | <0.001 |
| REACTOME_NEURONAL_SYSTEM | <0.001 |
| LIAN_LIPA_TARGETS_6M | <0.001 |
| LIAN_LIPA_TARGETS_3M | <0.001 |
| REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION | <0.001 |
| REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL | <0.001 |
| REACTOME_GAP_JUNCTION_TRAFFICKING | <0.001 |
| REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES | <0.001 |
| KEGG_LYSOSOME | <0.001 |
| REACTOME_POTASSIUM_CHANNELS | <0.001 |

*Table 3.47. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene names were solved by removing the construct with the lowest absolute difference in z-scores*

| NAME | P-value |
|---|---|
| REACTOME_CELL_DEATH_SIGNALLING_VIA_NRAGE_NRIF_AND_NADE | <0.001 |
| REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE | <0.001 |
| REACTOME_NRAGE_SIGNALS_DEATH_THROUGH_JNK | <0.001 |
| BIOCARTA_RAS_PATHWAY | <0.001 |
| COLIN_PILOCYTIC_ASTROCYTOMA_VS_GLIOBLASTOMA_DN | <0.001 |
| ST_DIFFERENTIATION_PATHWAY_IN_PC12_CELLS | <0.001 |
| REACTOME_TRIF_MEDIATED_TLR3_SIGNALING | <0.001 |
| REACTOME_ABC_FAMILY_PROTEINS_MEDIATED_TRANSPORT | <0.001 |
| SA_B_CELL_RECEPTOR_COMPLEXES | <0.001 |
| KEGG_MAPK_SIGNALING_PATHWAY | <0.001 |

*Table 3.48. Top 10 enriched gene sets enriched in the shRNA constructs with higher z-score differences between co-culture samples. Duplicate gene names were solved by removing the construct with their alphabetical order.*

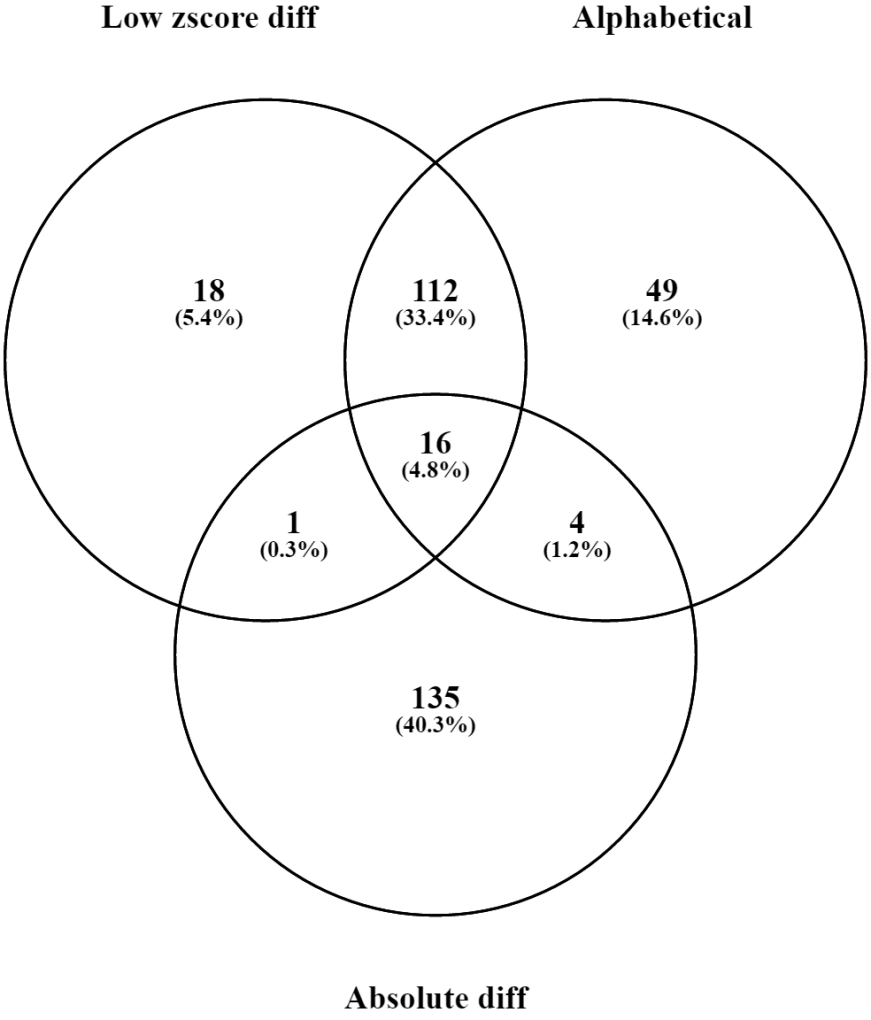



*Figure 3.37. Venn diagram showing the overlapping results of the 3 methods used to analyse the z-score difference results with GSEA*

| |
|---|
| REACTOME_CELL_DEATH_SIGNALLING_VIA_NRAGE_NRIF_AND_NADE |
| KEGG_MAPK_SIGNALING_PATHWAY |
| REACTOME_NRAGE_SIGNALS_DEATH_THROUGH_JNK |
| RODRIGUES_THYROID_CARCINOMA_DN |
| REACTOME_SIGNALING_BY_GPCR |
| REACTOME_GPCR_DOWNSTREAM_SIGNALING |
| KEGG_GNRH_SIGNALING_PATHWAY |
| REACTOME_ION_CHANNEL_TRANSPORT |
| REACTOME_P75_NTR_RECEPTOR_MEDIATED_SIGNALLING |
| BIOCARTA_CREB_PATHWAY |
| REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOPMENT_AND_PLATELET_PRODUCTION |
| LE_EGR2_TARGETS_UP |
| LIN_MELANOMA_COPY_NUMBER_DN |
| REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL |
| LIAN_LIPA_TARGETS_3M |
| KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM |

*Table 3.49. Gene sets identified by GSEA as enriched in the z-score difference results using three different duplicate gene name removal systems*

### 3.3.12. Transcription factor binding site enrichment analysis of GOI

Transcription factors are DNA binding proteins that regulate gene expression. They recognize and bind to certain DNA sequences in the vicinity of the genes whose expression they regulate. A transcription factor may regulate the expression of many genes, some of which might be involved in a common mechanism. Therefore, transcription factors were investigated as possible regulators of the response to neighbour suppression in tumorigenic cells.

OPOSSUM 3.0 (http://opossum.cisreg.ca/oPOSSUM3/) was used to investigate whether there were any transcription factors regulating more genes in the GOI list than what was expected by chance. The "Mouse Single Site Analysis" option was used. The GOI list for $3T3^{supp+}$ and $3T3^{supp-}$ were used as the target genes, and the background was set to the list of genes for which at least one targeting sequence was analysed. All parameters were kept to their default options.

OPOSSUM calculates a z-score and a Fisher score as enrichment statistics. The z-score for a transcription factor indicates how far the number of binding sites in the GOI list for that transcription factor is from the expected number of binding sites given the number of binding sites in the background list. In essence, a highly positive z-score indicates that the GOI list is enriched in binding sites for a transcription factor. On the contrary, a negative z-score indicates that there are less binding sites than expected. The Fisher score is calculated similarly but it does not take into account that genes might have more than one binding site for a transcription factor. Rather, it considers the presence of binding sites for a transcription factor as a binary event: a gene either has no binding sites for the transcription factor under consideration or it has at least one. Fisher scores were used for assessment of enrichment as the aim of this analysis was to find transcription factors that could regulate the expression of a higher number of genes than expected without taking into consideration the existence of multiple binding sites in the GOI. There is no formal statistical method of assessment of enrichment in OPPOSUM so this needs to be considered for each of the analyses performed.

Results were ranked by Fisher scores. Table 3.50 contains the 10 transcription factors with the highest Fisher scores in the $3T3^{supp+}$ GOI list. OPPOSSUM recommends visualizing the distribution of the transcription factors z-scores

against the GC composition of their binding sites as a clustering on any of the extremes (high or low GC content) may be an indicative of a bias in analysis produced by a differential composition in binding sites in the background used (the screened genes in this case). The website provides this graphical representation. For the $3T3^{supp+}$ (Figure 3.38) there is no evident bias and the graphical representation indicates that the transcription factor with the highest Fisher score (Zfp423) is separate from the rest of the transcription factors detected as potentially enriched.

The 10 transcription factors with the highest Fisher score in the $3T3^{supp-}$ GOI list are listed in Table 3.51. The graphical representation as explained above is available on Figure 3.39. In this case, the two transcription factors with the highest Fisher score (ELK4 and EBF1) apparently cluster separate from the rest transcription factors again without an apparent bias in GC composition. Zfp423 has a low Fisher score (0.256), which may suggest a difference in transcriptional regulation of the $3T3^{supp-}$ and $3T3^{supp+}$ GOIs and points to different mechanisms being involved as also suggested by the functional and gene set enrichment data. The Fisher scores for ELK4 and EBF1 are not as low in $3T3^{supp+}$ however: EBF1 has a Fisher score of 3.688 and is included in the top 10 and ELK4 has a Fisher score of 1.572.

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Z-score | Fisher score |
|---|---|---|---|---|---|---|---|---|
| Zfp423 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 102 | 153 | 3015 | 6840 | 2.960 | 6.877 |
| FOXD1 | Winged Helix-Turn-Helix | Forkhead | 199 | 56 | 7023 | 2832 | 5.837 | 4.646 |
| ELK1 | Winged Helix-Turn-Helix | Ets | 196 | 59 | 6942 | 2913 | -3.223 | 4.251 |
| FOXA1 | Winged Helix-Turn-Helix | Forkhead | 199 | 56 | 7132 | 2723 | 2.598 | 3.695 |
| EBF1 | Zipper-Type | Helix-Loop-Helix | 170 | 85 | 5956 | 3899 | -1.422 | 3.688 |
| Egr1 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 98 | 157 | 3201 | 6654 | 0.363 | 3.585 |
| Gata1 | Zinc-coordinating | GATA | 207 | 48 | 7515 | 2340 | 4.613 | 3.285 |
| GABPA | Winged Helix-Turn-Helix | Ets | 140 | 115 | 4862 | 4993 | -4.405 | 3.095 |
| TEAD1 | Helix-Turn-Helix | Homeo | 83 | 172 | 2739 | 7116 | 2.795 | 2.878 |
| Esrrb | Zinc-coordinating | Hormone-nuclear Receptor | 158 | 97 | 5606 | 4249 | -0.967 | 2.818 |

*Table 3.50. Top 10 enriched transcription factors (ranked by Fisher scores) in the 3T3$^{supp+}$ GOI list according to OPOSSUM*

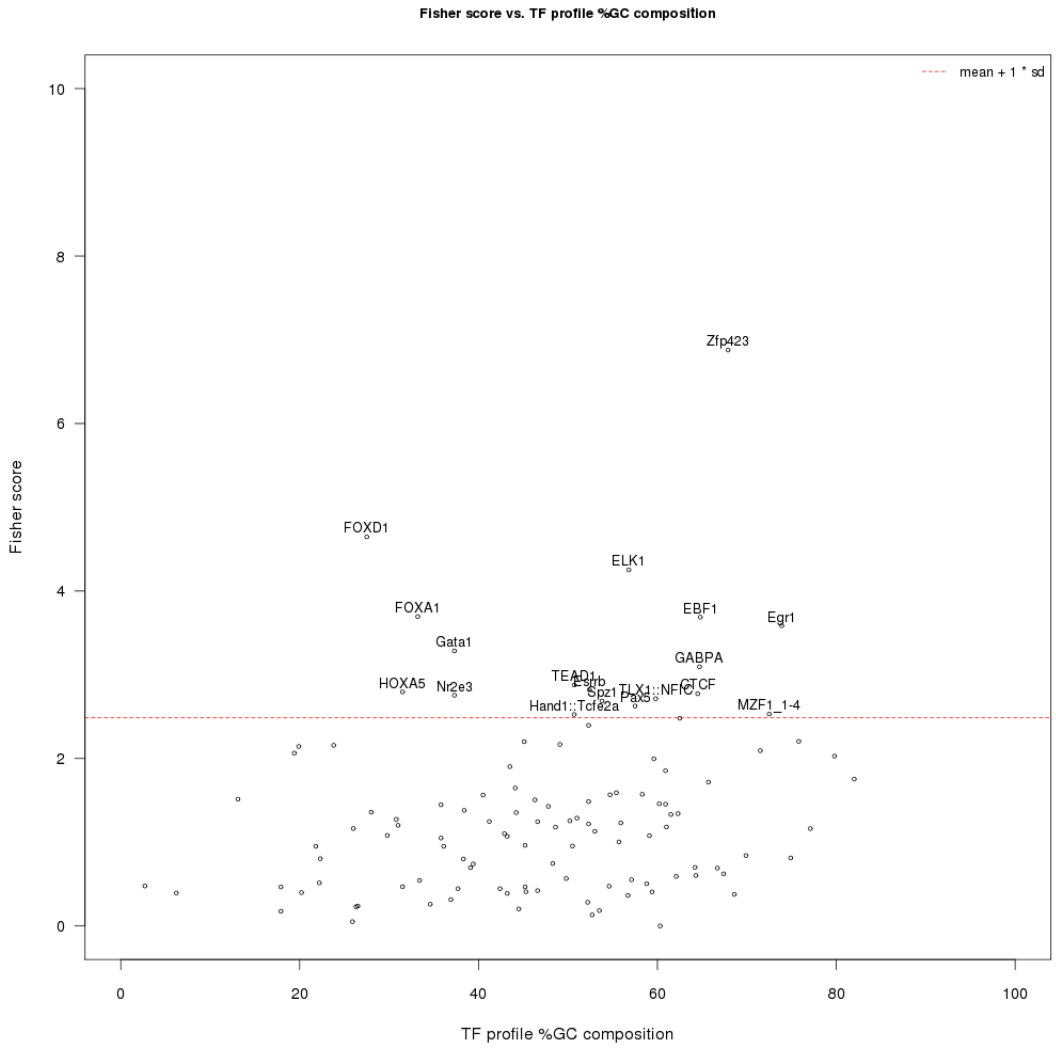



*Figure 3.38. Distribution of Fisher scores for the transcription factors analyzed by OPOSSUM and the percentage of GC in their binding sites for the 3T3$^{supp+}$ GOI list. The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors.*

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Z-score | Fisher score |
|---|---|---|---|---|---|---|---|---|
| ELK4 | Winged Helix-Turn-Helix | Ets | 116 | 180 | 3258 | 6597 | 4.707 | 4.08 |
| EBF1 | Zipper-Type | Helix-Loop-Helix | 197 | 99 | 5956 | 3899 | -2.261 | 3.97 |
| NFKB1 | Ig-fold | Rel | 68 | 228 | 1880 | 7975 | 5.589 | 2.868 |
| GABPA | Winged Helix-Turn-Helix | Ets | 159 | 137 | 4862 | 4993 | 2.863 | 2.566 |
| USF1 | Zipper-Type | Helix-Loop-Helix | 162 | 134 | 4964 | 4891 | 0.071 | 2.553 |
| PPARG::RXRA | Zinc-coordinating | Hormone-nuclear Receptor | 112 | 184 | 3335 | 6520 | 2.089 | 2.449 |
| Foxa2 | Winged Helix-Turn-Helix | Forkhead | 198 | 98 | 6297 | 3558 | 8.814 | 1.835 |
| RUNX1 | Ig-fold | Runt | 229 | 67 | 7360 | 2495 | 4.156 | 1.808 |
| Zfx | Zinc-coordinating | BetaBetaAlpha-zinc finger | 158 | 138 | 4960 | 4895 | 1.891 | 1.803 |
| ESR1 | Zinc-coordinating | Hormone-nuclear Receptor | 8 | 288 | 175 | 9680 | 4.636 | 1.796 |

*Table 3.51. Top 10 enriched transcription factors (ranked by Fisher scores) in the 3T3$^{supp-}$ GOI list according to OPOSSUM*

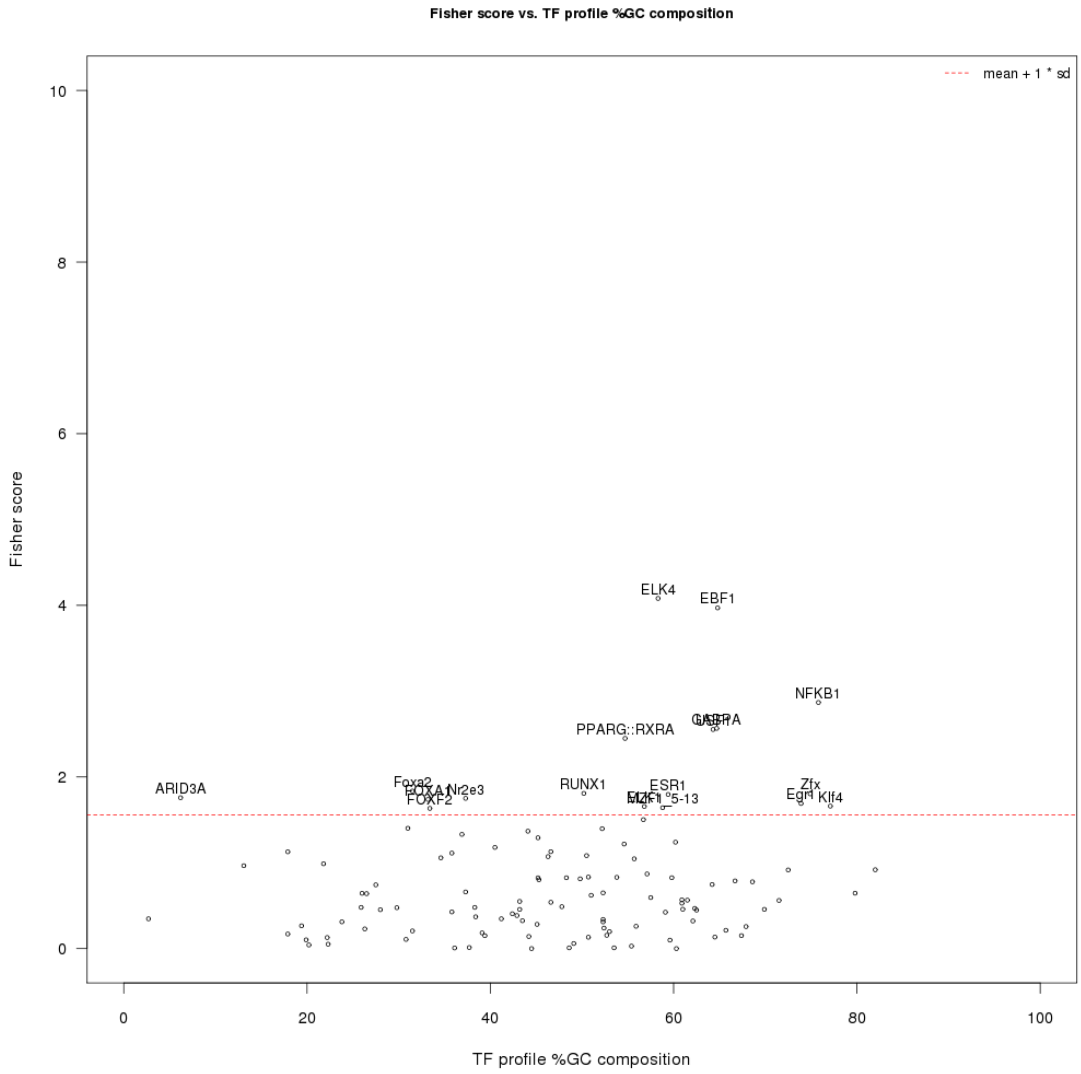



*Figure 3.39. Distribution of Fisher scores for the transcription factors analyzed by OPOSSUM and the percentage of GC in their binding sites for the 3T3$^{supp-}$ GOI list. The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors*

### 3.3.13. Confirmation of gene knock-down effects on neighbour suppression

RNAi screens are known to produce a high number of false positives. It was therefore imperative to re-test the effect of GOI gene knockdown on neighbour suppression. A list of genes to be confirmed was built based on different criteria. Genes selected for confirmation are listed in Table 3.52.

Firstly, genes from the $3T3^{supp+}$ and the $3T3^{supp-}$ GOI lists were selected if two targeting sequences had been included in the relevant GOI list and at least one of the targeting sequences was in the top 10 targeting sequences with the biggest effects as approximated by their z-score. This was the case for *Rrp7a* in $3T3^{supp-}$ and *Ttbk1, Cd5l, Snap23* and *Coa3* in $3T3^{supp+}$.

Next, genes were included in the confirmation list based on their z-scores. Ten genes were selected from the $3T3^{supp-}$ GOI list and eight from the $3T3^{supp+}$ GOI list using this approach.

Upon network analysis using STRING, three additional genes were added to the list of genes to be confirmed based on the apparent high number of genes in the GOI list they interacted with. Two of these genes (*Tp53* and *Wt1*) were part of the $3T3^{supp+}$ GOI list and one (*Rac1*) belonged to the $3T3^{supp-}$ GOI list.

Three genes were added to the list of confirmations after network analysis using FunCoup.

When possible, shRNA constructs were designed using targeting sequences already published with available data about their knockdown efficiency. This was not the case for most of the genes, for which no knockdown literature existed. In that case, the RNAi Consortium website was used to search for targeting sequences. If no sequences were available they were designed using the same website. The plasmid containing a *Tp53* shRNA targeting sequence and a control plasmid with a shRNA construct targeting Firefly Luciferase (referred to as FF below) were obtained from Addgene (Tp53 plasmid #19751 and FF plasmid #19744). The plasmid containing the targeting sequence for Wt1 and its control (targeting LacZ) were a gift from Dr Abdelkader Essafi (University of Bristol). For *Rac1*, a plasmid expressing a human dominant negative form of RAC1, Rac1-T17N (commonly named $Rac1^{N17}$) and another

expressing a constitutively active form of the protein, Rac1-G12V ($Rac1^{V12}$) were obtained from Dr Fabio Luis Forti (University of Sao Paulo). Table 3.53 has information about the different plasmids used. For all plasmids cloned into the pSuperRetroRuro vector a scrambled negative control was available. Empty cells (not transduced) were also used as a control for transduction effects.

Plasmids were delivered to the tumorigenic cells using viral particles. Wt1 and LacZ targeting sequences were cloned into a lentiviral rather than a retroviral vector. As Phoenix cells can only produce retroviral particles, 293T cells were used for lentiviral packaging of *Wt1* and LacZ plasmids. The lentiviral packaging plasmids 8.91 and DMD were obtained from Dr Giles Cory (University of Exeter). 293T cells were co-transduced with both packaging plasmids and either *Wt1* or LacZ. Delivery of the viral particles to SVE-6 was performed following the optimized protocols for retroviral transductions.

Stable cell lines were derived by antibiotic selection of clones containing each plasmids. All vectors except for pCM, containing the $Rac1^{N17}$ or $Rac1^{V12}$ provided resistance to puromycin to cells upon integration. The selection marker for pCM in mammalian cells is gentamycin, so SVE-6 cells were selected using G418.

To assess the effect that the reduction of expression of the GOI had on neighbour suppression, the derived cell lines were plated on top of monolayers of $3T3^{supp+}$ and $3T3^{supp-}$. Cells were cultured for 7 days and then fixed using 4% paraformaldehyde. Selecting which colonies to take pictures of was identified as a potential source of bias. Therefore, a systematic approach was designed. Each well of the 6-well plates was scanned on a Leica Widefield microscope using an automated scanning protocol in the Leica LasX software which produced a tile-scan composed of 100 pictures with a 5x objective. The pictures were merged together using ImageJ to produce an individual picture per well. All pictures were processed in ImageJ before colony quantification using the same approach:

“Unsharp mask” with radius=1 and mask=0.60

“Subtract Background” with rolling=50

“Auto Threshold” with method=Default white

“Make Binary”

An example of this processing pipeline is shown in Figure 3.40.

The binary pictures obtained at the end of the pre-processing pipeline were then used to quantify the size of the colonies. A method of quantification was devised that could produce comparable results between technical and biological replicates. A degree of variability on degree of suppression had been observed before between monolayers formed from the same or different cell stocks of the same cell line. It was therefore reasoned that an absolute quantification of colony size would not be adequate. A normalization method was needed to allow a comparison between replicates and thus a control well, with non-transduced cells, was present in all experiments.

The shape of the SVE-6 colonies is also variable, which limits the viability of an accurate surface area calculation. It was reasoned that a method was needed where a constant surface was interrogated for GFP presence, producing a percentage of surface containing GFP positive cells (SVE-6). As it had been previously observed that a colony on a certain stock of $3T3^{supp-}$ had a size approximating the surface area of a 20x objective picture, this was used as the measurement area. A square of that size was drawn in ImageJ using the command “Specify...”. The square was placed on the centre of each colony present in the picture to obtain the regions of interest (ROI), and the percentage of GFP positive area was then calculated for each ROI using the “Measure” command.

The percentage of GFP positive area for all the ROI in each well were imported into GraphPad for graphical interpretation.

Two replicates were performed for the first set of clones designed from the initial z-score list and STRING analysis. The results of the first replicate are depicted in Figure 3.41 for the $3T3^{supp+}$ co-cultures and Figure 3.42 for the $3T3^{supp-}$ co-cultures. It can be seen that the control conditions (SVE-6 cells either un-transduced or transduced with a scrambled targeting sequence) produce smaller colonies when plated on $3T3^{supp-}$ , which makes interpretation difficult. However, it is still possible to see a difference in the colony size of some clones (*Flt4, Rrp7a, Cd5l* and *Ttbk1*) compared with the rest of the clones.

On the $3T3^{supp+}$ co-culture, the same tendency of the controls to be smaller than the clones is still maintained. Within the clones selected for confirmation from

the GOI list, *Sh3px2bd*, *Psma7*, *Ttll1*, *Gab2*, *Plxnb3* and *Serpinb3* seem to have a higher median colony size than the rest.

A second replicate was performed on both co-culture conditions. This replicate included the genes under examination in the first replicate and three new genes: *Rac1*, *p53* and *Wt1*.

Results of the second replicate on the $3T3^{supp-}$ are represented in Figure 3.44. In this case control wells were not available due to problems in the culture of the control cell lines and therefore a comparison amongst the different clones is necessary to evaluate the effect of the targeting sequences. Similarly to the first replicate in Figure 3.41, *Flt4*, *Rrp7a* and *Ttbk1* appear to have a lower median than other clones, although that is not the case with *Cd5l*, which had been identified in the first replicate. *Rac1* was tested with an expression vector containing a dominant negative form of RAC1 ($RAC1^{N17}$) and a constitutively active form ($RAC1^{V12}$). Both forms seem to produce smaller colonies than the targeting sequences used in pSuperRetroPuro. However, no specific negative control was available for comparison. Similarly, both *p53* and *Wt*1 appear to reduce the colony size slightly compared to their own controls (firefly luciferase, labelled “FF” and LacZ respectively).

In the case of the $3T3^{supp+}$ co-culture, the second replicate (Figure 3.43) agrees only partially with the first replicate. *Sh3px2bd* still appears to have increased the colony size, but the rest of the effects seen in the first replicated are not apparent. The targeting sequence against *Snap23* appears to have increased the colony size greatly. However, the same effect is present in the $3T3^{supp-}$ co-culture (Figure 3.44) and upon manual inspection of the well it is evident that a very high density of SVE-6 cells transduced with that targeting sequence were plated in the well which might have confounded the analysis. Strikingly, LacZ appears to have increased the colony size greatly as well, and given that the number of colonies identified in that well is also higher than the number of colonies in most of the other wells it is possible that the same “over-plating” situation has confounded the results of this control targeting sequence.

A third experiment was performed to include the GOI selected from the FunCoup network analysis and including some of the genes already included in the first replicates. Figure 3.45 represents the results for this third experiment in $3T3^{supp+}$. *Nptxr* knockdown appears to have an effect on colony size and so

does the constitutively active form of RAC1 ($RAC1^{V12}$). In this last case a high number of colonies were detected which may indicate a contribution of high plating density to the detected effect of this form of RAC1

A third replicate of the $3T3^{supp-}$ (Figure 3.46) suggests that the knockdown of *Twf2* in SVE6 cells reduces the size of colonies, an effect also produced by the expression of a dominant negative form of RAC1 ($RAC1^{N17}$).

| Gene | Condition | Selection method |
|---|---|---|
| *Abcd1* | 3T3supp- | Z-score |
| *Atp8a2* | 3T3supp- | Z-score |
| *Btbd6* | 3T3supp- | Z-score |
| *Flt4* | 3T3supp- | Z-score |
| *Gab2* | 3T3supp- | Z-score |
| *Impa1* | 3T3supp- | Z-score |
| *Plxnb3* | 3T3supp- | Z-score |
| *Pmvk* | 3T3supp- | Z-score |
| *Rrp7a* | 3T3supp- | Z-score |
| *Serpinb3a* | 3T3supp- | Z-score |
| *Rac1* | 3T3supp- | STRING network analysis |
| *Rps5* | 3T3supp- | FunCoup network analysis |
| *Cd5l* | 3T3supp+ | Z-score |
| *Cdc73* | 3T3supp+ | Z-score |
| *Coa3* | 3T3supp+ | Z-score |
| *Myocd* | 3T3supp+ | Z-score |
| *Plxna4* | 3T3supp+ | Z-score |
| *Psma7* | 3T3supp+ | Z-score |
| *Sh3pxd2b* | 3T3supp+ | Z-score |
| *Snap23* | 3T3supp+ | Z-score |
| *Ttbk1* | 3T3supp+ | Z-score |
| *Ttll1* | 3T3supp+ | Z-score |
| *Wisp1* | 3T3supp+ | Z-score |
| *Zeb2* | 3T3supp+ | Z-score |
| *Wt1* | 3T3supp+ | STRING network analysis |
| *Tp53* | 3T3supp+ | STRING network analysis |
| *Twf2* | 3T3supp+ | FunCoup network analysis |
| *Nptxr* | 3T3supp+ | FunCoup network analysis |

*Table 3.52. Genes selected for confirmation in both co-culture conditions.*

| Gene | Vector | Source |
|---|---|---|
| *Abcd1* | pSuperRetroPuro | The RNAi Consortium |
| *Atp8a2* | pSuperRetroPuro | (Xu, Yang et al. 2012) |
| *Btbd6* | pSuperRetroPuro | The RNAi Consortium |
| *Cd5l* | pSuperRetroPuro | The RNAi Consortium |
| *Cdc73* | pSuperRetroPuro | The RNAi Consortium |
| *Coa3* | pSuperRetroPuro | The RNAi Consortium |
| *Flt4* | pSuperRetroPuro | The RNAi Consortium |
| *Gab2* | pSuperRetroPuro | (Huang, Tong et al. 2008) |
| *Impa1* | pSuperRetroPuro | The RNAi Consortium |
| *Myocd* | pSuperRetroPuro | The RNAi Consortium |
| *Plxna4* | pSuperRetroPuro | (Okada and Tomooka 2013) |
| *Plxnb3* | pSuperRetroPuro | The RNAi Consortium |
| *Pmvk* | pSuperRetroPuro | The RNAi Consortium |
| *Psma7* | pSuperRetroPuro | The RNAi Consortium |
| *Rrp7a* | pSuperRetroPuro | (Maserati, Dai et al. 2014) |
| *Serpinb3a* | pSuperRetroPuro | The RNAi Consortium |
| *Sh3pxd2b* | pSuperRetroPuro | The RNAi Consortium |
| *Snap23* | pSuperRetroPuro | (Yatsu, Ohbayashi et al. 2013) |
| *Ttbk1* | pSuperRetroPuro | The RNAi Consortium |
| *Ttll1* | pSuperRetroPuro | The RNAi Consortium |
| *Wisp1* | pSuperRetroPuro | The RNAi Consortium |
| *Zeb2* | pSuperRetroPuro | The RNAi Consortium |
| $RAC1^{N17}$ | pCM | (Espinha, Osaki et al. 2015) |
| $RAC1^{V12}$ | pCM | (Espinha, Osaki et al. 2015) |
| *Wt1* | pLenti | (Essafi, Webb et al. 2011) |
| *LacZ* | pLenti | (Essafi, Webb et al. 2011) |
| *p53* | MSCV P2Gm | (Stern, Astrof et al. 2008) |
| *FF* | MSCV P2Gm | (Stern, Astrof et al. 2008) |
| *Twf2* | pSuperRetroPuro | The RNAi Consortium |
| *Nptxr* | pSuperRetroPuro | The RNAi Consortium |
| *Rps5* | pSuperRetroPuro | (O'Donohue, Choesmel et al. 2010) |

*Table 3.53. Genes selected for confirmation, vectors in which the sequences were cloned and source of the sequences or plasmids for each one.*

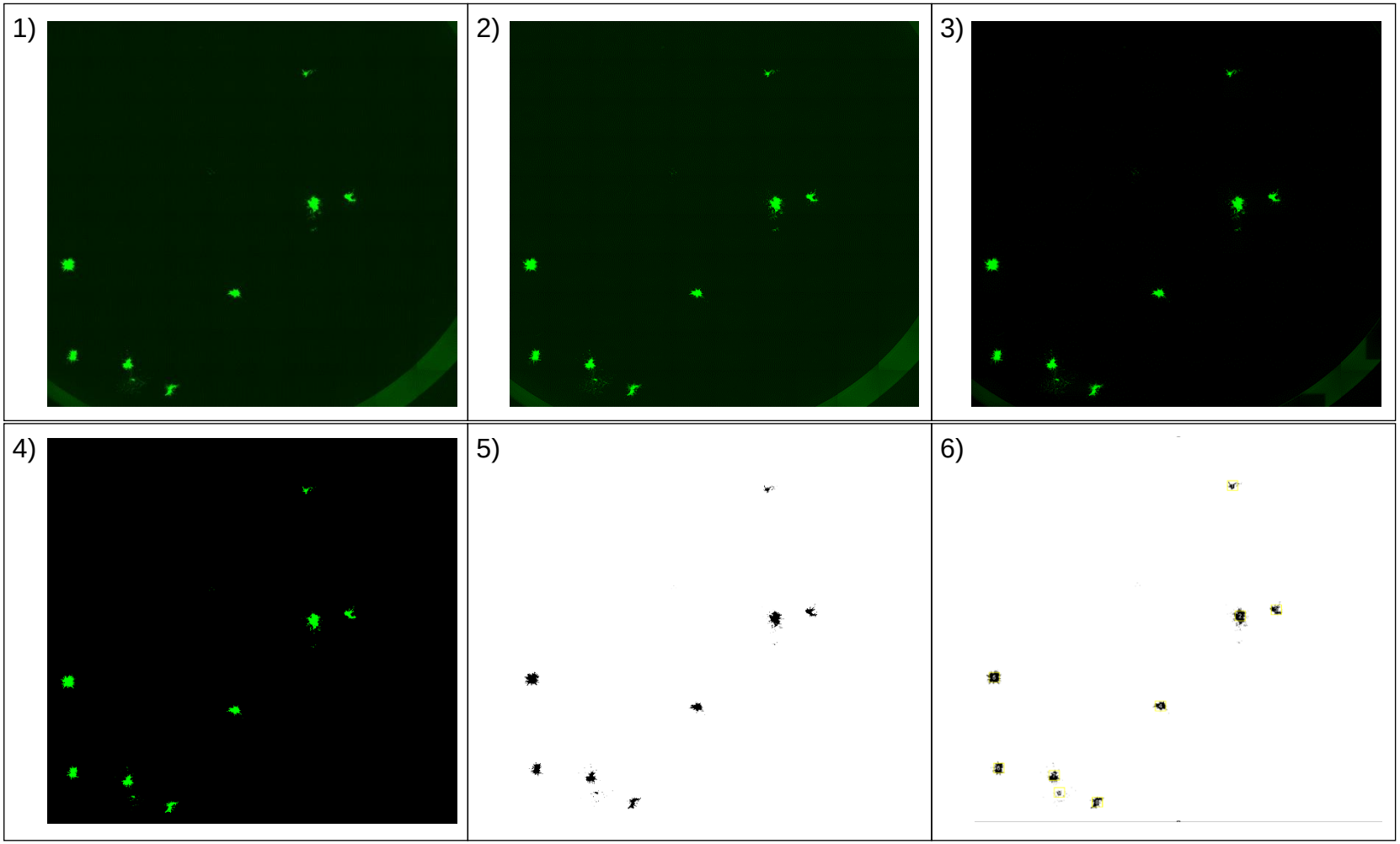



*Figure 3.40. Example of the steps in the image processing pipeline.*
*1) Original merged tilescan. 2) Image in 1 after "Unsharp mask" was applied. 3) Image in 2 after "Subtract background". 4) Image in 3 after "Auto threshold". 5) Image in 4 after "Make binary". 6) Image in 6 with the colonies marked for quantification (yellow squares with colony number in white).*

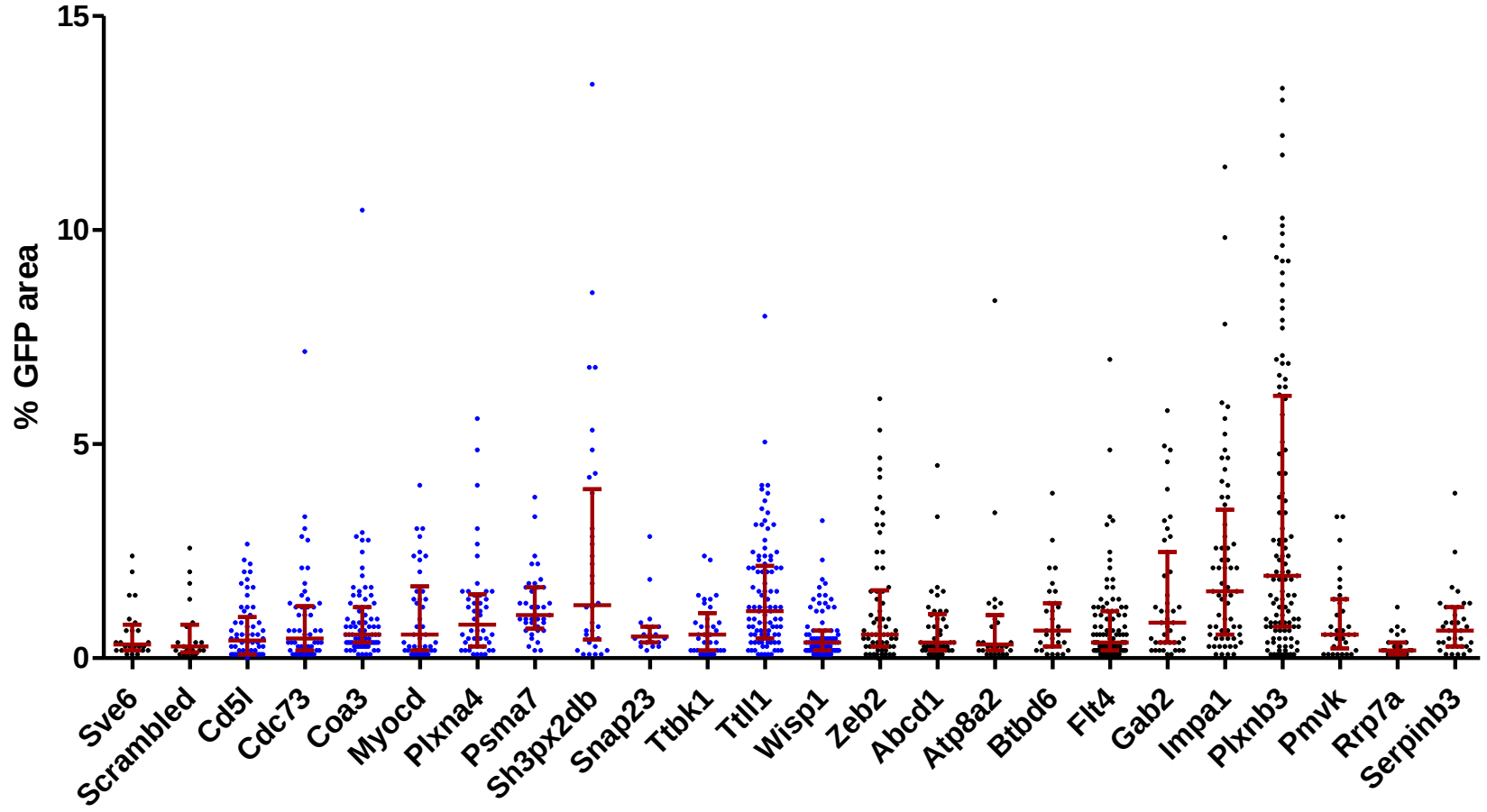



*Figure 3.41. Percentages of GFP positive area in the ROI in a replicate on 3T3$^{supp+}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Sve6 are untransduced cells and Scrambled, cells transduced with a pSuperRetroPuro plasmid containing a scrambled targeting sequencing. Datasets plotted in blue are those genes which were identified in the 3T3$^{supp+}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp-}$ GOI list. Red bars represent the median and interquartile range.*

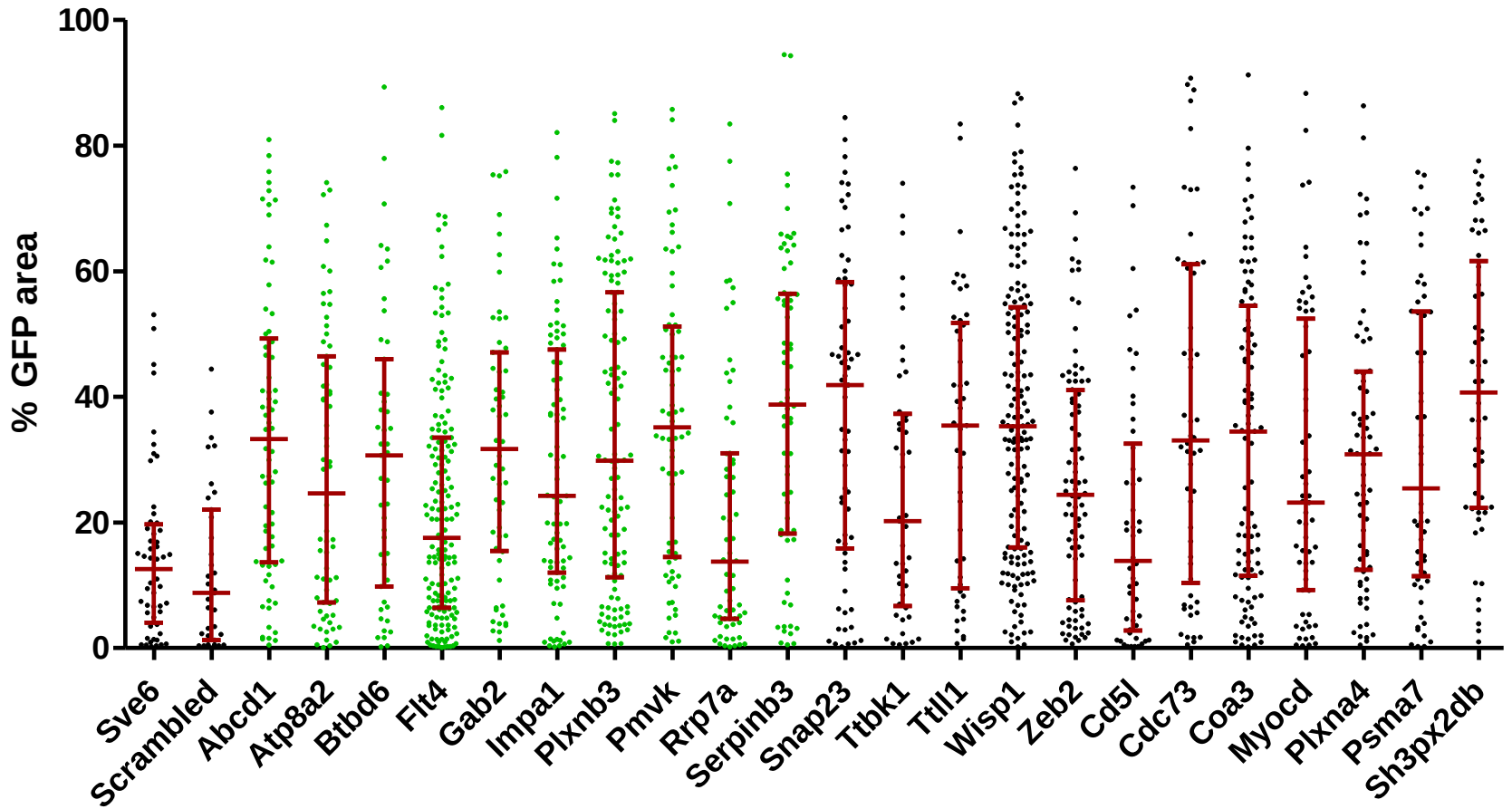



*Figure 3.42. Percentages of GFP positive area in the ROI in a replicate on 3T3$^{supp-}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Sve6 are untransduced cells and Scrambled, cells transduced with a pSuperRetroPuro plasmid containing a scrambled targeting sequencing. Datasets plotted in green are those genes which were identified in the 3T3$^{supp-}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp+}$ GOI list. Red bars represent the median and interquartile range.*

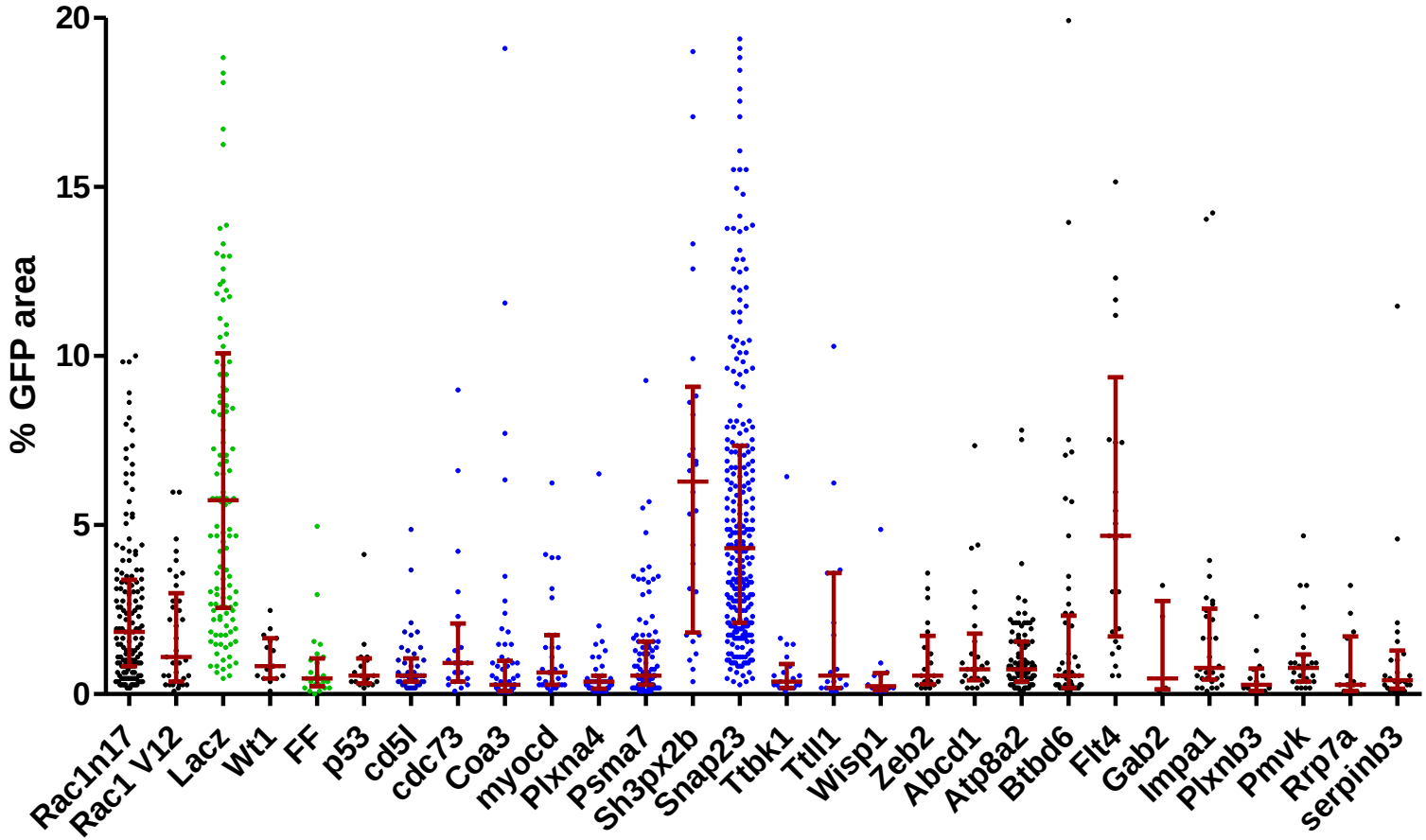



*Figure 3.43. Percentages of GFP positive area in the ROI in the second replicate on 3T3$^{supp+}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Datasets in green are the control for the dataset in black on their right (Lacz is the control plasmid for Wt1 and FF is the control plasmid for p53). Datasets plotted in blue are those genes which were identified in the 3T3$^{sup+-}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp-T}$ GOI list. Red bars represent the median and interquartile range.*

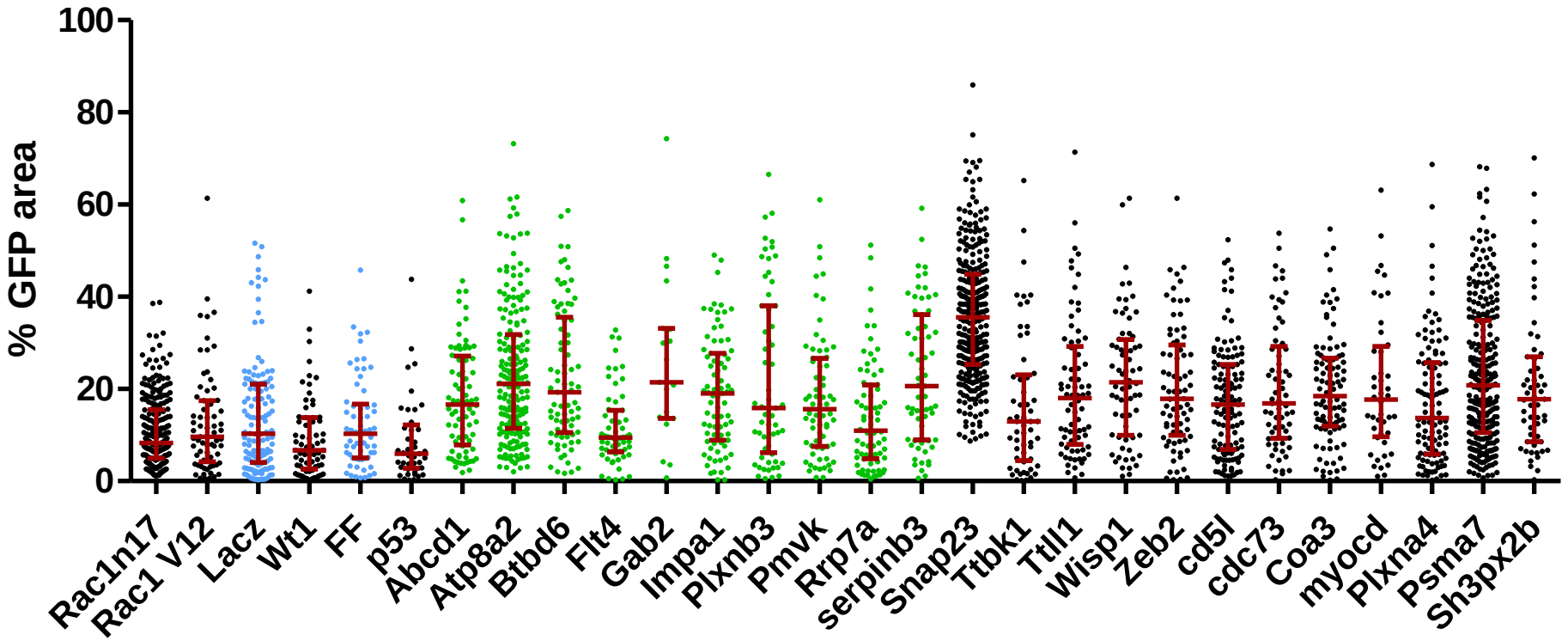



*Figure 3.44. Percentages of GFP positive area in the ROI in the second replicate on 3T3$^{supp-}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Datasets in blue are the control for the dataset in black on their right (Lacz is the control plasmid for Wt1 and FF is the control plasmid for p53). Datasets plotted in green are those genes which were identified in the 3T3$^{supp-}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp+}$ GOI list. Red bars represent the median and interquartile range.*

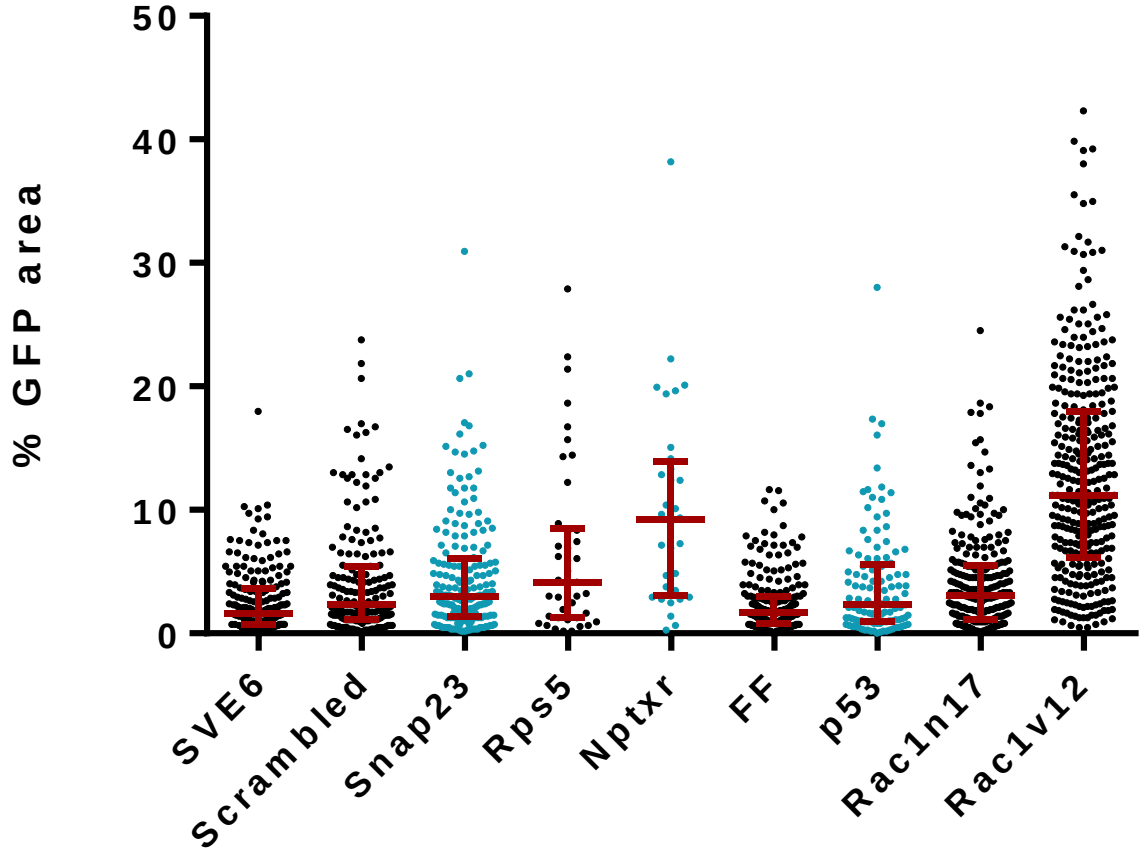



*Figure 3.45. Percentages of GFP positive area in the ROI in the third replicate on 3T3$^{supp+}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Sve6 are un-transduced cells and Scrambled, cells transduced with a pSuperRetroPuro plasmid containing a scrambled targeting sequencing. Datasets plotted in blue are those genes which were identified in the 3T3$^{supp+}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp-}$ GOI list or are controls. Red bars represent the median and interquartile range.*

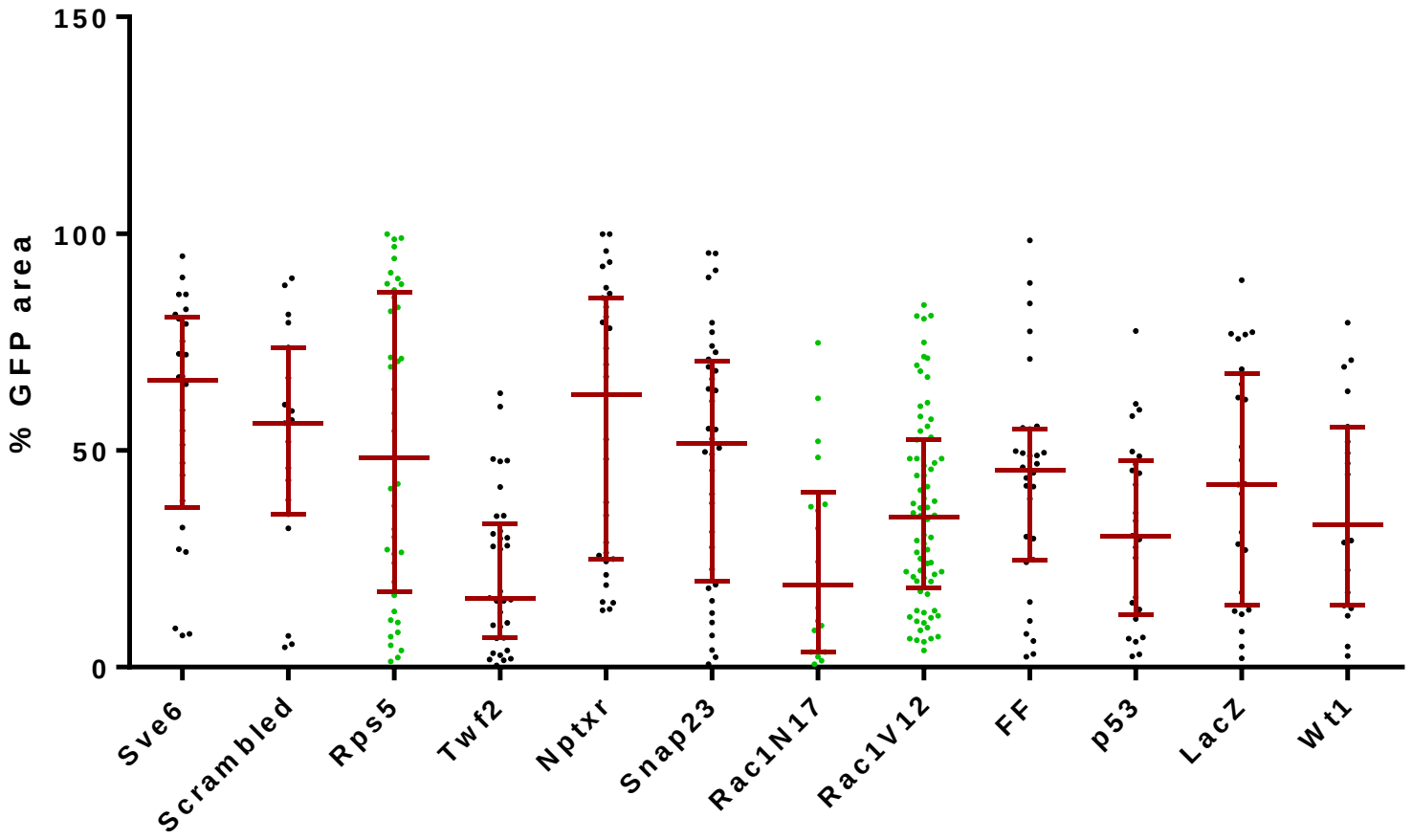



*Figure 3.46 Percentages of GFP positive area in the ROI in the third replicate on 3T3$^{supp-}$ monolayers.*
*The labels in the X axis correspond to the genes against which the targeting sequences were designed. Sve6 are un-transduced cells and Scrambled, cells transduced with a pSuperRetroPuro plasmid containing a scrambled targeting sequencing. Datasets plotted in green are those genes which were identified in the 3T3$^{supp-}$ GOI list. Genes with data represented by black dots were identified in the 3T3$^{supp+}$ GOI list or are controls. Red bars represent the median and interquartile range.*

### 3.3.14. Study of differential survival rates for levels of expression of genes identified as potential effectors

To perform an independent validation of the GOI lists and produce an initial assessment of the potential of the screen to be translated into human cancer biology, an approach was sought to evaluate the effect of changes in expression of the GOI in human tumours.

Most cancers are epithelial in nature and therefore the use of a fibroblast *in vitro* model adds a potential complication to the extrapolation of the results to human biology. However, the study of gene expression of the GOI in different tumours may provide an insight into subsets of genes which can be altering tumour development or growth using a mechanism similar to the observed in this *in vitro* model. Additionally, if neighbour suppression is, as hypothesized, present in various cell types then the mechanisms will be observed in both epithelial and fibroblast systems. Furthermore, fibroblasts have been used to obtain molecular mechanisms in other aspects of cancer biology such as the signalling downstream of various oncogenes, studied in mouse embryonic fibroblasts.

Numerous studies have attempted to use tumour expression data to predict prognosis of cancer patients. A few of these studies (see section 3.2.9) were available for analysis in-house using an R script. The files and the script were provided by Dr Sara Vowler (University of Cambridge). The script investigates a gene list provided by the user and correlates the expression level of that gene in the expression dataset specified with the survival rate of the patients. If a significant difference in terms of survival is found between high and low expression levels, the script then looks for the point of inflexion where that difference starts. The script outputs a number of files containing the results of the test for each gene, and for each gene that has come up as having a significant effect it produces a survival curve. Although net survival can be used as an output when dealing with clinical data, the aim of this screen was to identify genes potentially involved in early steps of tumorigenesis. Upon evaluation of the available outcomes, the script was therefore configured to look for effects on distant metastasis free survival. It was reasoned that the establishment of tumorigenic cells in a new niche after metastasis would most closely resemble the situation studied here, where a tumorigenic cell is found surrounded by non-tumorigenic neighbours. Although both situations differ,

most basically due to the homotypic interactions in neighbour suppression compared to heterotypic interactions after metastasis, the mechanisms of neighbour suppression may still be applicable.

For the same reason and when possible, only primary or low stage tumour samples were used to increase relevance to early cancer development. In some datasets, this cut-off produced small subsamples without enough power to detect differences and therefore the entire dataset was used. Eight sub-datasets were produced from the original datasets. The sub-datasets and the parameters used to create them are specified in Table 3.54.

To search for correlation of the expression level of the GOI with metastasis free survival of cancer patients in these sub-datasets, the GOI lists had to be used as input files for the script. However the expression data is for human genes. The gene symbols were converted into their human sequence homologs as explained in 3.3.11.4.

Expression data was not available for all genes in the GOI lists in all datasets: 244 (of 277 GOI) genes in the $3T3^{supp+}$ GOI list were assigned a human homologue. Of those, 229 had at least 1 probe in Taylor, 173 in Shedden and Desmedt and 215 in Loi. From the $3T3^{supp-}$ GOI list, 297 (of 308 GOI) genes were assigned a human gene name, of which 262 had at least 1 probe in Taylor, 195 in Shedden and Desmedt and 249 in Loi.

The numbers of genes found to be statistically associated with metastasis-free survival in each dataset for each screening condition and for the library are specified in Table 3.55. Genes were considered associated only if the direction of effect was consistent with the direction predicted by the screen. As the knockdown of the GOI in $3T3^{supp+}$ produced an increment on cell number, higher expression of the genes was expected to be beneficial in patient metastasis-free survival. Conversely, only genes where a higher expression was detrimental to metastasis-free survival were considered to be associated in the $3T3^{supp-}$ GOI list. The survival curves for some of these genes are shown in Figure 3.47 ($3T3^{supp-}$) and Figure 3.48 ($3T3^{supp+}$).

To investigate the functional relevance of the GOI according to this expression approach, a score was given to genes in the GOI lists according to how many datasets they were found to be significant in. The top 10 genes are listed in

Table 3.56 ($3T3^{supp+}$) and Table 3.57 ($3T3^{supp-}$). Genes were ordered first by the number of datasets in which they were found to be associated with disease progression. In both GOI lists, the maximum number of datasets where a single gene was found to be statistically associated with disease progression was 3 (5 genes in $3T3^{supp+}$ and 7 in $3T3^{supp-}$). From the $3T3^{supp+}$ GOI list, 19 genes were associated with disease progression in 2 datasets. These genes were ordered by increasing sum of p-values for the relevant datasets and the 5 genes with the lowest sum were included in the top 10. The same process was applied for the $3T3^{supp-}$ GOI list.

The relevance of the number of GOI for which expression levels correspond with survival in human patients is difficult to assess without comparing them to the number of screened genes which are also correlated. To investigate the probability of such a number of genes to be associated in these datasets if the GOI lists were random subsets of screened genes rather than genes with functional importance, a random sampling approach similar to that used in section 3.3.10.3 was implemented.

First, the list of screened genes was used as the input gene list for the recursive partitioning script. The list of human homologs available from 3.3.11.4 was used as input. Similarly to the GOI, not all genes had expression data in all datasets. 6,962 genes had probes in Taylor; 8,561 in Shedden and Desmedt (100% overlap); and 10,608 in Loi.

A Python script was then written to produce random samples of constructs from the shRNA library of the same size as the GOI. The script then looked those genes up in the different expression dataset results for the library and counted the number of genes that were found to have a significant effect on distant metastasis free survival. These numbers were then compared to the results found using the GOI lists.

Results from this random sampling approach are shown in Figure 3.49 ($3T3^{supp+}$) and Figure 3.50 ($3T3^{supp-}$). The $3T3^{supp+}$ GOI list was found to be significantly enriched in genes whose expression level affected distant metastasis free survival in expression datasets Shedden, Loi_tamoxifen and Desmed_allgrades. The GOI list for $3T3^{supp-}$ was deemed significantly enriched in genes with expression level mediated effects on distant metastasis free survival in both Loi expression datasets.

| Name | Dataset | Survival measure | Tumour type | Number of samples | Tumour grade | Treatment |
|---|---|---|---|---|---|---|
| Taylor_all | Taylor | DMFS | Prostate cancer | 140 | All | N/A |
| Taylor_primary | Taylor | DMFS | Prostate cancer | 61 | Primary tumours, low Gleason grade (4-6), t-stage 2 | N/A |
| Shedden_all | Shedden | DMFS | Non-small cell lung cancer | 442 | N/A | N/A |
| Loi_tamoxifen | Loi | DMFS | Breast cancer | 255 | N/A | Tamoxifen |
| Loi_notamox | Loi | DMFS | Breast cancer | 125 | N/A | No treatment |
| Desmedt_all | Desmedt | DMFS | Breast cancer (lymph node negative) | 198 | All | N/A |
| Desmedt_3 | Desmedt | DMFS | Breast cancer (lymph node negative) | 83 | Grade 3 | N/A |
| Desmedt_1 | Desmedt | DMFS | Breast cancer(lymph node negative) | 30 | Grade 1 | N/A |

*Table 3.54. Sub-datasets used for the study of gene expression changes associated to differential metastasis-free survival in cancer patients and parameters used to create them from the original datasets.*
*DMFS: distant metastasis free survival.*

| | Taylor_all | Taylor_primary | Shedden | Loi_tamoxifen | Loi_notamox | Desmedt_all | Desmedt_1 | Desmedt_3 |
|---|---|---|---|---|---|---|---|---|
| Library | 3,474 | 170 | 1,413 | 2,034 | 480 | 543 | 221 | 464 |
| $3T3^{supp+}$ | 33 | 3 | 22 | 38 | 6 | 8 | 2 | 3 |
| $3T3^{supp-}$ | 52 | 1 | 24 | 46 | 14 | 8 | 4 | 7 |

*Table 3.55. Number of genes whose expression was found to be associated with differential metastasis-free survival using the recursive partitioning approach in the different datasets used.*
*Note that the number of genes in "$3T3^{supp+}$" and "$3T3^{supp-}$" corresponds to the number of genes found to have an effect on the direction predicted by the screen; i.e. $3T3^{supp+}$: lower expression is detrimental, $3T3^{supp-}$: lower expression is beneficial. On the contrary, numbers for the "library" dataset correspond to both directions of effect.*

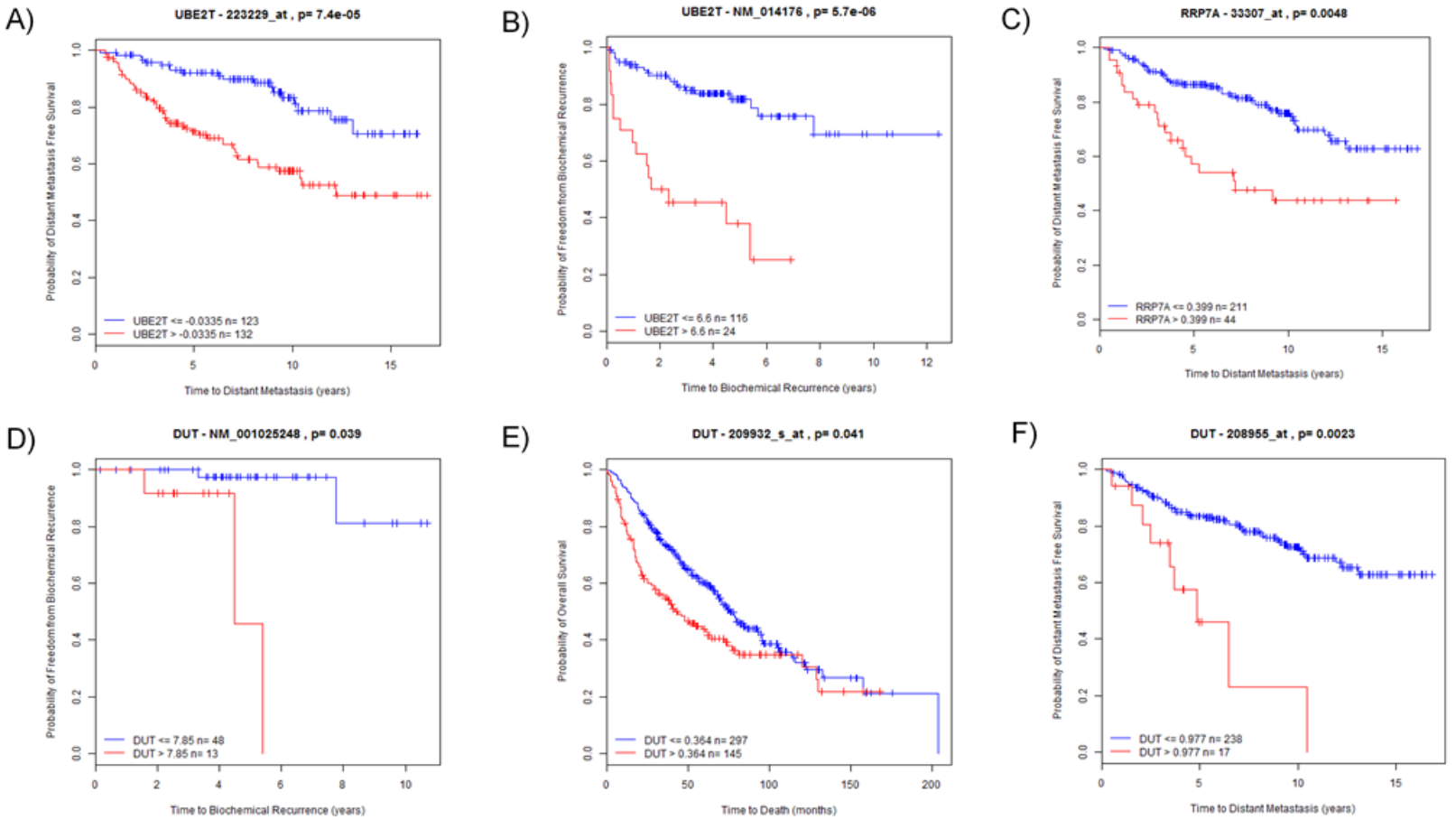



*Figure 3.47. Survival curves for expression of three genes selected from the 3T3$^{supp-}$ GOI list.*
*A) Survival curve for UBE2T from dataset Loi_tamoxifen. B) Survival curve for UBE2T from dataset Taylor_all. C) Survival curve for RRP7A from dataset Loi_tamoxifen. D) Survival curve for DUT from dataset Taylor_primary. E) Survival curve for DUT from dataset Shedden. F) Survival curve for DUT from dataset Loi_tamoxifen*

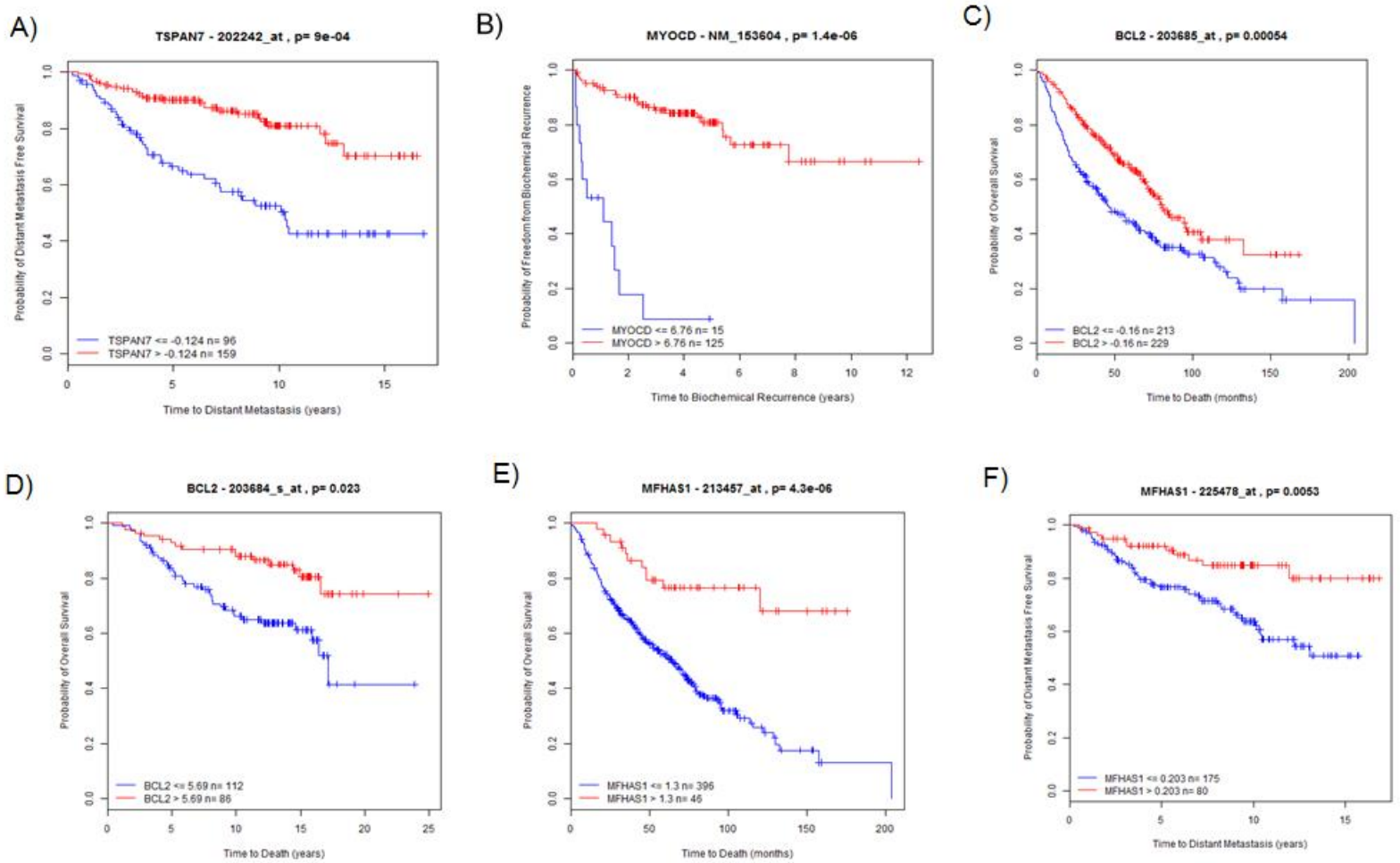



*Figure 3.48. Survival curves for expression of genes selected from the 3T3$^{supp+}$ GOI list.*
*A) Survival curve for TSPAN7 from dataset Loi_tamoxifen. B) Survival curve for MYOCD from dataset Taylor_All.. C) Survival curve for BCL2 from dataset Shedden. D) Survival curve for BCL2 from dataset Desmedt_all. E) Survival curve for MFHAS1 from dataset Shedden. F) Survival curve for MFHAS1 from dataset Loi_tamoxifen.*

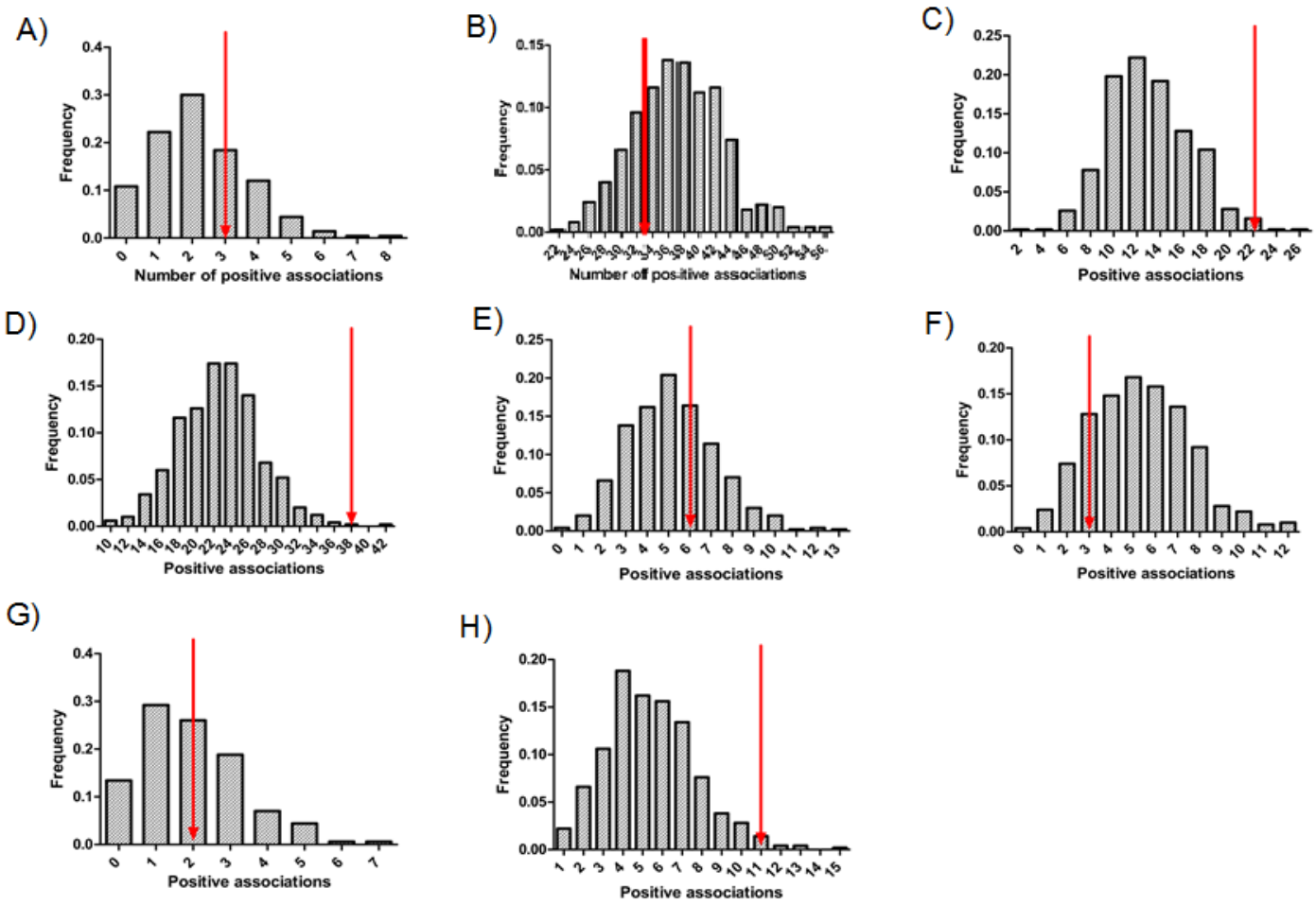



*Figure 3.49. Results of the random sampling approach to evaluate the significance of the recursive partitioning results in the 3T3$^{supp+}$ GOI list. Red arrows mark the number of genes with changes in expression associated with differential metastasis rates. A) Supp+ Taylor primary. B) Supp+ Taylor All. C) Shedden. D) Loi Tamoxifen. E) Loi no treatment. F) Desmedt grade 3. G) Desmedt grade 1. H) Desmedt all grades.*

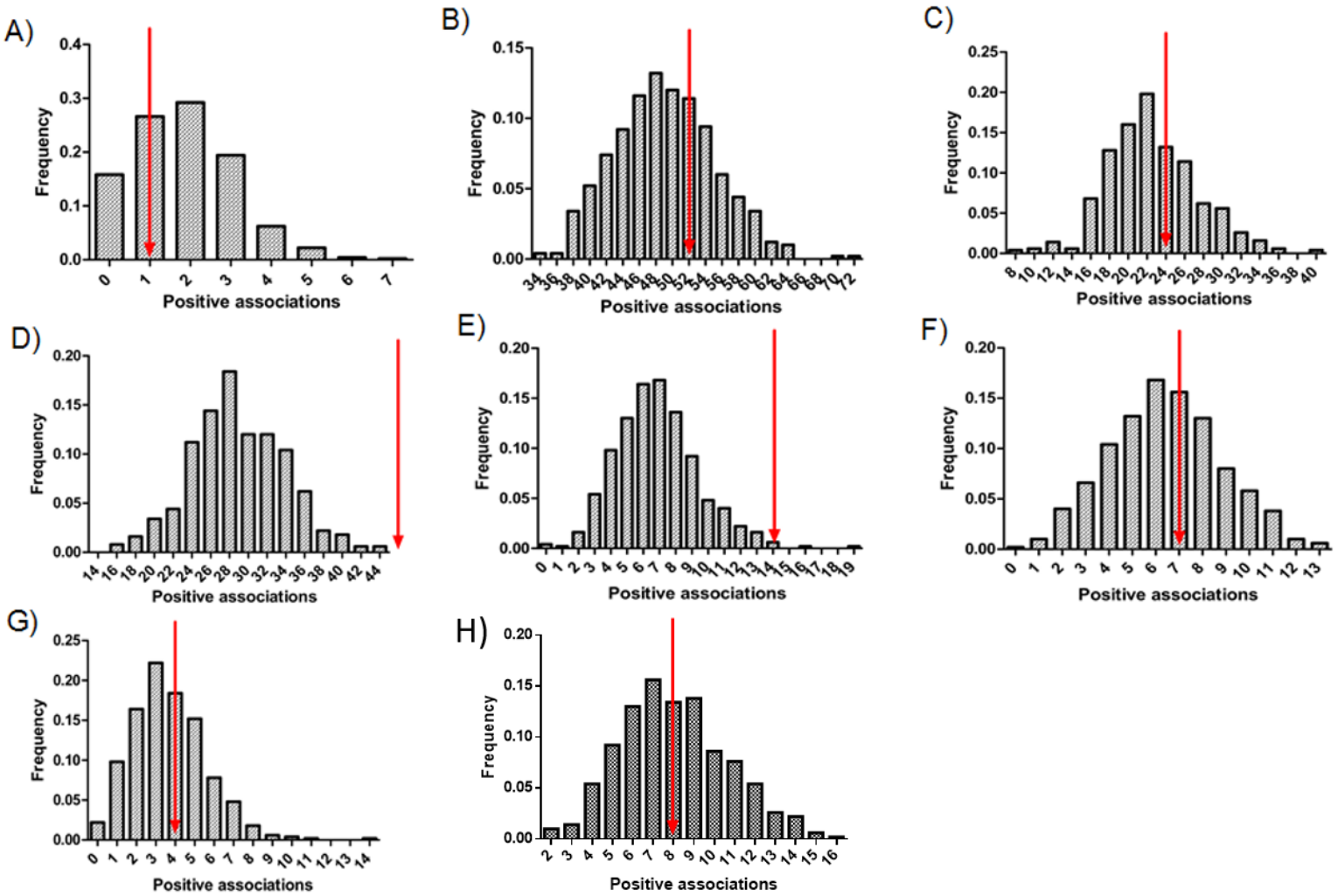



*Figure 3.50. Results of the random sampling approach to evaluate the significance of the recursive partitioning results in the 3T3$^{supp-}$ GOI list. Red arrows mark the number of genes with changes in expression associated with differential metastasis rates. A) Supp- Taylor primary. B) Supp- Taylor All. C) Shedden. D) Loi Tamoxifen. E) Loi no treatment. F) Desmedt grade 3. G) Desmedt grade 1. H) Desmedt all grades.*

| Gene name | Number of datasets | Dataset names (p-values) |
|---|---|---|
| *DMD* | 3 | Taylor_all (0.035), Shedden (0.044), desmedt_grade3 (0.04) |
| *MFHAS1* | 3 | Taylor_all (0.048), Shedden ($4.3*10^{-6}$), Loi_tamoxifen (0.0053) |
| *NDEL1* | 3 | Taylor_primary (0.013), Taylor_all (0.00057), Loi_tamoxifen (0.0086) |
| *FBXO30* | 3 | Taylor_all (0.0058), Shedden (0.039), Loi_notreatment (0.026) |
| *RPL37* | 3 | Loi_tamoxifen (0.041), Desmedt_grade3 (0.0032), Desmedt_all (0.0075) |
| *TGIF1* | 2 | Taylor_all (0.00027), Taylor_primary (0.021) |
| *PPP2CB* | 2 | Taylor_all (0.0014), Loi_tamoxifen (0.00065) |
| *EIF4B* | 2 | Loi_tamoxifen (0.0057), Desmedt_all (0.0039) |
| *NPC2* | 2 | Shedden (0.0064), Loi_tamoxifen (0.008) |
| *COL4A3* | 2 | Shedden (0.0067), Desmedt_grade1 (0.0085) |
| *PDGFRA* | 2 | Taylor_all (0.015), Loi_tamoxifen (0.0016) |
| *MS4A1* | 2 | Shedden (0.0054), Loi_notreatment (0.051) |

*Table 3.56. $3T3^{supp+}$ top 10 genes with higher number of associations in the recursive partitioning datasets.*

| Gene name | Number of datasets | Dataset names (p-values) |
|---|---|---|
| *ABCD1* | 3 | Loi_tamoxifen (0.047), Desmedt_grade3 (0.00091), Desmedt_all (0.0013) |
| *CCNE1* | 3 | Taylor_all (0.0014), Shedden (0.013), Loi_tamoxifen (0.0082) |
| *CENPE* | 3 | Taylor_all (0.00048), Shedden (0.0085), Loi_tamoxifen (0.00018) |
| *KIF18A* | 3 | Taylor_all (0.002), Shedden (0.031), Loi_tamoxifen (0.015) |
| *NR2C2AP* | 3 | Taylor_all (0.027), Loi_tamoxifen (0.0031), Loi_notreatment (0.0037) |
| *PLXNB3* | 3 | Taylor_all (0.0051), Loi_tamoxifen (0.012), Loi_notreatment(0.016) |
| *TBX1* | 3 | Taylor_all (0.045), Shedden (0.029), Desmedt_all (0.013) |
| *DUT* | 3 | Taylor_primary (0.039), Shedden(0.041), Loi_tamoxifen (0.0099) |
| *EIF2S1* | 3 | Shedden (0.015), Desmedt_grade3 (0.025), Desmedt_all (0.017) |
| *GATA2* | 2 | Taylor_all (0.00048), Desmedt_grade3 (0.041) |
| *UBE2T* | 2 | Taylor_all ($5.7 \times 10^{-6}$), Loi_tamoxifen ($7.4 \times 10^{-5}$) |
| *DIAPH3* | 2 | Taylor_all (0.00033), Loi_tamoxifen (0.016) |

*Table 3.57. $3T3^{supp-}$ top 10 genes with higher number of associations in the recursive partitioning datasets.*

## 3.4. Discussion

### 3.4.1. Design of a genome-wide shRNA screen in a homotypic co-culture system.

Since the advent of the RNAi approach for the study of functional genomics, RNAi screens have gained popularity year by year. Pooled screens such as the one presented here have been performed in a number of *in vitro* models of cancer in the last few years (Zhou, Patel et al. 2014, Burleigh, McKinney et al. 2015, Kushwaha, O'Leary et al. 2015) and are amenable especially for phenotypes that result in an increase or decrease in cell number. An RNAi screen has even been performed in a co-culture model of the tumour microenvironment (Burleigh, McKinney et al. 2015). However the expansion of our knowledge about the initial phases of tumorigenesis, when tumour cells are surrounded by cells of the same type (that is, in a homotypic context) has been more conservative. To gain some insight into the molecular effectors at play in the initial steps of tumorigenesis and the tumour growth suppression effect exerted by homotypic cell interactions, a genome-wide shRNA screen was designed.

Normal cells surrounding tumorigenic cells have been previously shown to suppress the growth of the latter by a still under-characterized mechanism. A genome-wide shRNA screen was designed and performed to identify the molecular effectors of neighbour suppression in the normal cells (Nicholls 2015). No mechanism has as of yet been proposed to explain the response of the tumorigenic cells to the suppression cues they receive from their neighbouring non-tumorigenic cells. Therefore this unexplored signal integration process occurring in the tumorigenic cells was targeted.

To investigate this process an appropriate *in vitro* model was first sought. To mimic a homotypic interaction context of tumour initiation it was necessary to use a non-tumorigenic cell line and a transformed counterpart. It is common that this type of *in vitro* model of homotypic interactions result in the growth suppression of the transformed cells. However, the $3T3^{supp+}$/$3T3^{supp-}$ cell lines provide a model of suppressive and permissive homotypic interactions when co-cultured with a SV-40 large Tag transformed cell line derived from $3T3^{supp+}$. This system composed of two different co-culture conditions has the advantage of

allowing the study of the phenotype of interest in two ways. Molecular effectors of suppression can be studied through the reversal of tumorigenic cell growth suppression in the co-culture with $3T3^{supp+}$; but also through the induction of suppression in the $3T3^{supp-}$ permissive co-culture condition.

An embryonic fibroblast *in vitro* model for the study of the initial phases of tumorigenesis might not be the most relevant to an *in vivo* situation. However it can be argued that the use of any cell line would yield results that are only specific to tumorigenesis in the tissue from which the cell line was derived as cancers in different tissues differ in their mechanisms. Also, monolayers of cells suppress the growth of transformed cells and therefore a permissive co-culture condition would be difficult to translate into any other cell type.

The decision of performing only one replicate of the entire screen after a pilot in duplicate is a limitation of this study. However, the implementation of this screen in duplicate would became a significantly larger endeavour and only approximately 65% of the GOI would be expected to overlap between duplicates (Strezoska, Licon et al. 2012). The NKI library has a twofold redundancy, which in the case of the GOI for which the two targeting sequences were detected as having an effect provides a reassurance in the confidence of their selection as GOI. However this is not the case for most of the GOI. This may be a result of different factors stemming from the RNAi technology itself or from the process by which the library was designed. The NKI library was designed using algorithms which predict the efficiency and limit the off-target effects of the targeting sequences. This is a limitation as far as the algorithm may be flawed due to either the information available to predict these parameters and to the design of the algorithm itself. The level of expression knockdown for each targeting sequence is unknown and may vary for both constructs that target the same gene and amongst the different genes targeted. The same can be said about the potential off-target effects of the targeting sequences.

These limitations were considered in the design of the screening process, the analysis pipeline and the validation experiments.

### 3.4.2. Analysis pipeline

Genome-wide RNAi screens provide a considerable amount of data that need to be carefully analysed to reduce the number of false positives as much as possible. Methods of analysis for this type of screening have been discussed at length and new applications and statistical pipelines are continuously being proposed (Diaz, Qin et al. 2015, Dutta, Azhir et al. 2016, List, Schmidt et al. 2016), highlighting the lack of a strong consensus on an optimal analysis pipeline.

Furthermore, most RNAi screens have been performed using a “control vs. treatment” approach, where only two conditions are compared. In contrast the screening process described here was composed of a “control” condition – growth on plastic; and two “treatment” conditions – co-culture with either a suppressive or a permissive underlying monolayer. These two conditions add an extra element for analysis, as effects exerted in the same direction (i.e. increasing or decreasing the suppressive effect in both conditions) would most likely represent non-specific effects in tumour cell growth or plating which are outside the scope of this screen.

This pooled RNAi screen was performed in pools or groups of constructs which required separate normalization due to the difference in variability across pools. For this reason an approach based on z-scores was selected. To include both levels of control conditions (growth in plastic and growth in the other co-culture condition), the growth in plastic was used as normalization for both co-cultures by calculating a ratio and the z-score was then built. Given the non-normality of the data, robust z-scores were calculated using the median and median absolute deviation. A method of two thresholds, one for each co-culture conditions was used to select genes of interest (GOI). To my knowledge this is the first report of this method in RNAi screens.

The use of integrated analysis pipelines including network biology and functional enrichment approaches has been advocated for as a method to circumvent some of the limitations of the RNAi technology (Dutta, Azhir et al. 2016, List, Schmidt et al. 2016, Wilson, Dalin et al. 2016). A pipeline of this type was adopted for the analysis of this screen with the aim of exploring the molecular mechanisms behind the phenomenon of neighbour suppression and also to refine the list of GOI produced by the screen. An important limitation of

the use of these approaches with a pooled RNAi screen lies in the choice of a background from which to draw conclusions about enrichments. Although the ideal situation in this study has been to use the list of screened genes as the background, this list may still be biased. It is assumed that all screened constructs could be functional, an assumption that it is possibly incorrect. In relation to the technical aspects of the screen, the constructs may not produce knockdown of the targeted gene, the level of knockdown may not be sufficient to produce an effect in the phenotype being investigated or off-target effects of the constructs may result in a phenotype due to the knockdown of a non-intended gene. Biologically, the knockdown of certain genes may not have the predicted effect if there is a compensation mechanism in place.

### 3.4.3. Method for validation of the GOI selected

The validation of the GOI was performed using additional targeting sequences or other gene disruption techniques. Other technologies (e.g. CRISPR, TALENS) may have been used for these validation studies, although this would have required optimization of additional procedures and the value added by these alternative approaches was not deemed to be sufficient to justify the implementation efforts.

The choice of an additional targeting sequence instead of re-validating the targeting sequences identified as effectors in the screen was based on the reasoning that using the same sequence, an approach taken in some screens (Cooper and Brockdorff 2013), could repeat the potential off-target effects of those sequences. However, this decision results in a limitation. The level of knockdown for the additional targeting sequence chosen may not be the same as the targeting sequence detected in the screen, resulting in the non-validation of genuine GOI. However, any GOI validated with an alternative targeting sequence would have greater confidence in a role in the phenotype.

### 3.4.4. Overlap in the mechanism of response to suppressive and permissive co-culture conditions.

The network analysis of the GOI lists from the $3T3^{supp+}$ and the $3T3^{supp-}$ co-cultures and the list combining the GOI from both co-cultures indicates the existence of interaction between genes with an effect in both co-cultures. This may indicate that there is a shared mechanism in the response to suppressive

cues and the growth in permissive conditions. However, this hypothesis is not supported by the functional enrichment approaches. There is very little overlap in the functional categories enriched in both GOI lists. The functional categories and gene sets enriched in the $3T3^{supp+}$ GOI list include many terms related to cancer, which suggests a potential role for neighbour suppression evasion in carcinogenesis. Other functional categories and gene sets in this co-culture condition include cell-cell communication, junctions and membrane proteins. Neighbour suppression has been described as a phenomenon which needs cell-cell contact and therefore the enrichment in these terms supports the choice of the $3T3^{supp+}$/SVE-6 co-culture system as a model of neighbour suppression.

The functional categories and gene sets found to be enriched in the $3T3^{supp-}$ include a wide range of terms.

### 3.4.5. Validated GOI

The results of the experimental validation of the GOI lists are very variable for most genes. Considering that 100 cells were used per shRNA sequence in the initial screening, it was expected that approximately 65% of the GOI would be validated if the same sequence was used (Strezoska, Licon et al. 2012). However, validation was performed in this case using different sequences from those in the primary screen as off-target effects are not expected to be repeated with this approach.

In view of the experiments performed to date in this study, *Sh3pxd2b* and *Rac1* would be the genes proposed to have an effect on neighbour suppression in the $3T3^{supp+}$/SVE-6 co-culture model. The effect of *Nptxr* in this model has been replicated once.

*Sh3pxd2b* encodes the protein Tks4, which is hypothesized to be an important component of actin-rich membrane structures called podosomes which are important for invasiveness (Buschman, Bromann et al. 2009). In the same study, Buschman and collaborators showed that knockdown of *Sh3pxd2b* reduced the ability of Src-transformed fibroblast to degrade the extracellular matrix and form podosomes and decreased proliferation. The increased colony size of SVE-6 upon introduction of a shRNA construct targeting this gene is therefore contradictory to their conclusions. Data from the recursive partitioning approach

to human tumour expression analysis agrees with the effect detected in this RNAi screen. In the Loi dataset, decreased expression is correlated with lower distant metastasis free survival (Figure 3.51).

*Rac1* was detected as a GOI in the $3T3^{supp-}$ co-culture sample. Rac1 is a Rho-GTPase signal transducer implicated in the regulation of a variety of cellular functions (Van Aelst and D'Souza-Schorey 1997) including organization of the actin cytoskeleton and recruitment of integrins (Hall 1998). It has also been shown to be involved in cell growth and regulation of the cell cycle (Saci, Cantley et al. 2011), formation of cell-cell adhesions (Ehrlich, Hansen et al. 2002) and contact inhibition of cell proliferation (Bosco, Nakai et al. 2010). *Rac1* involvement in neighbour suppression was validated using a dominant negative and a constitutively active form of the human protein. Introduction of both cDNA constructs seem to affect the colony size of SVE-6 in co-culture with $3T3^{supp+}$, although the magnitude of the effect for each of the constructs varied between replicates. Higher expression of RAC1 in human tumour samples correlates with lower probability of distant metastasis free survival (Figure 3.52).

The effect of *Nptxr* on the SVE-6 colony size in a suppressive co-culture was replicated once and therefore more validation experiments will be necessary to decide on its involvement. The protein encoded by the *Nptxr* (Neuronal Pentraxin Receptor) gene was first described as a synapse-associated protein (Hsu and Perin 1995). No expression data is available for this gene in the human tumour datasets used by the recursive partitioning script.

In the $3T3^{supp-}$ co-culture validation experiments, the effect of four genes (*Rac1*, *Flt4*, *Rrp7a* and *Ttbk1*) were replicated.

The dominant-negative form of *Rac1* had the largest effect of the two forms used for validation, although the constitutively-active form also decreased the size of the SVE-6 colonies. The expression data agrees with the dominant-negative form reducing the size of the colonies as lower expression correlates with higher probability of distant metastasis free survival.

*Flt4* encodes a tyrosine kinase receptor for vascular endothelial growth factors C and D. Inhibition of this signalling pathway was recently shown to decrease mammary tumour cell growth and survival (Varney and Singh 2015) and identified as a marker of metastasis in colorectal cancer (Xiao, Liu et al. 2015). Recursive partitioning analysis of the human tumour dataset Taylor (Figure 3.53)

indicates that higher expression of FLT4 correlates with lower probability of distant metastasis free survival, agreeing with those two studies and with the findings in this screen. However, the effect detected in the expression datasets may be mediated by an increase in angiogenesis and reflect the number of endothelial cells rather than indicating the level of expression in tumour cells.

The available literature on *Rrp7a* is limited. It encodes a ribosomal RNA processing factor, but the single publication within Pubmed only indicates that expression is required for the transition between morula and blastocyst in mouse embryo development (Maserati, Dai et al. 2014). Expression data in datasets Shedden and Loi indicate that high expression in human tumours is correlated with low probability of distant metastasis free survival (Figure 3.54).

*Ttbk1* encodes a kinase which phosphorylates the protein TAU and has been suggested to be implicated in the pathology of Alzheimer's disease (Lund, Cowburn et al. 2013). No publications are available on a possible role in carcinogenesis, but a high level of expression is correlated in Taylor with lower probabilities of distant metastasis free survival (Figure 3.55).

The effect of *Twf2* in the colony size of SVE-6 cells in co-culture with $3T3^{supp-}$ was validated in one replicate. This gene was selected as a GOI in $3T3^{supp+}$ from the FunCoup network analysis data. The protein encoded by *Twf2* binds actin and inhibits its polymerization (Vartiainen, Sarkkinen et al. 2003). No correlation between expression levels and distant metastasis free survival was found in the recursive partitioning datasets.

The proportion of experimentally validated GOI was low. However, most of the genes which were validated by a second targeting sequence were found to have their expression levels associated with distant metastasis free survival. This suggests that an integrated approach to the analysis of pooled shRNA screens may aid the identification of genes involved in the phenotype studied.

Two of the validated GOIs bind actin, which may indicate a role of the cytoskeleton in neighbour suppression.

Further experiments should be performed to confirm the role of the validated GOI in neighbour suppression and their importance in carcinogenesis.

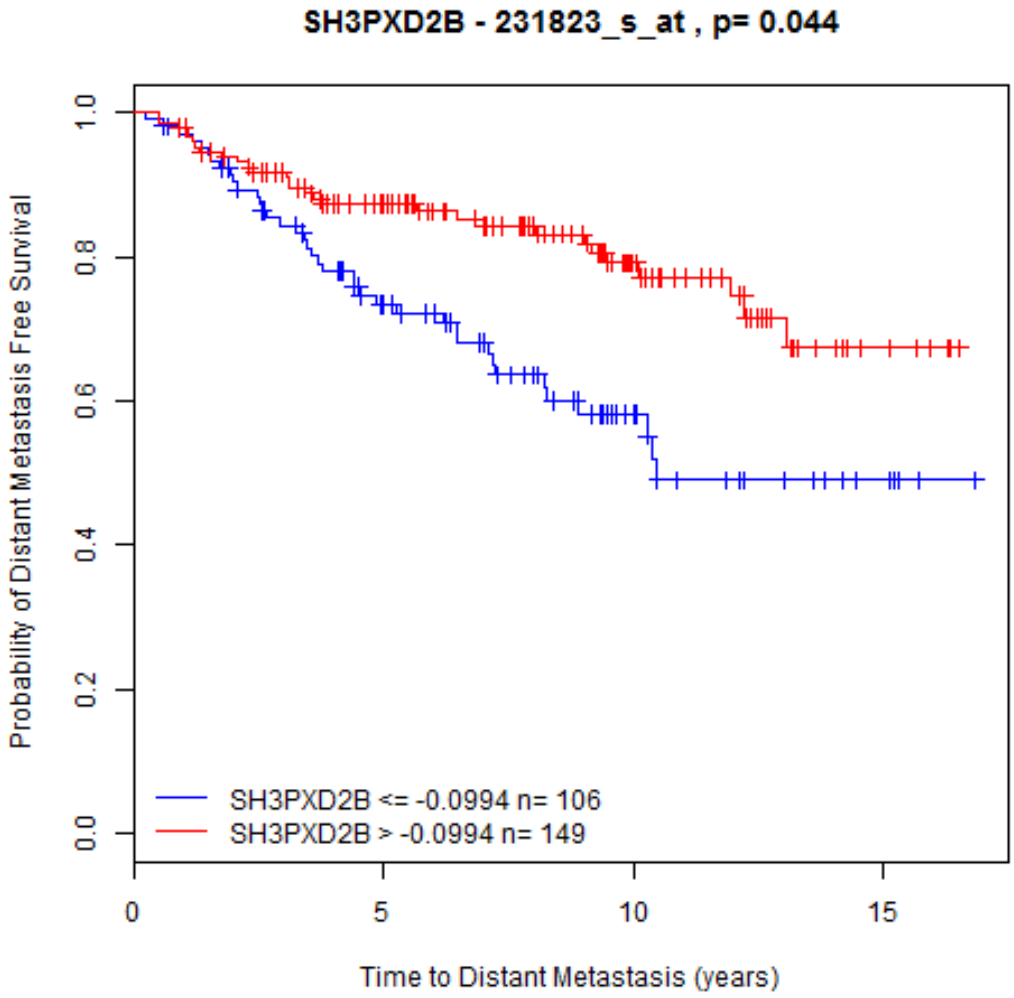



*Figure 3.51. Survival curve obtained from the recursive partitioning analysis of the Loi dataset for SH3PXD2B.*

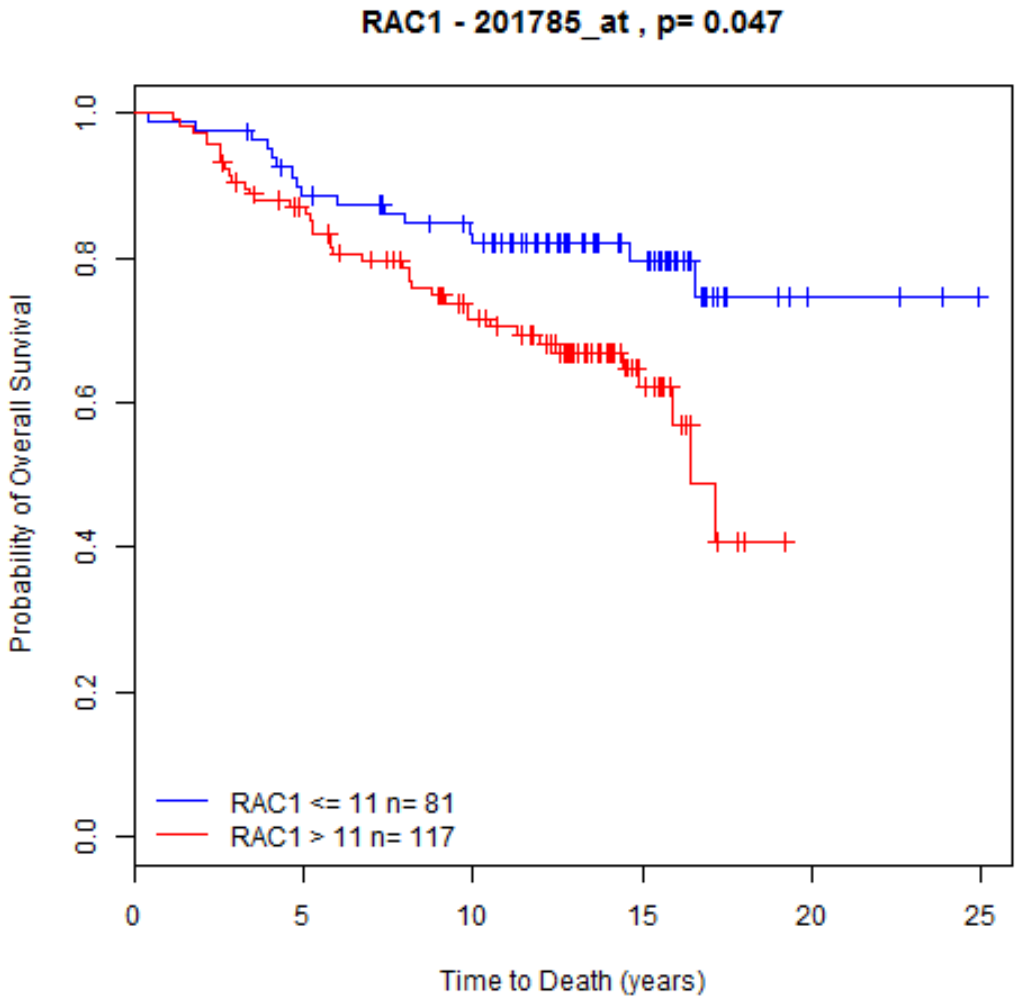



*Figure 3.52. Survival curve obtained from the recursive partitioning analysis of the Loi dataset for RAC1.*

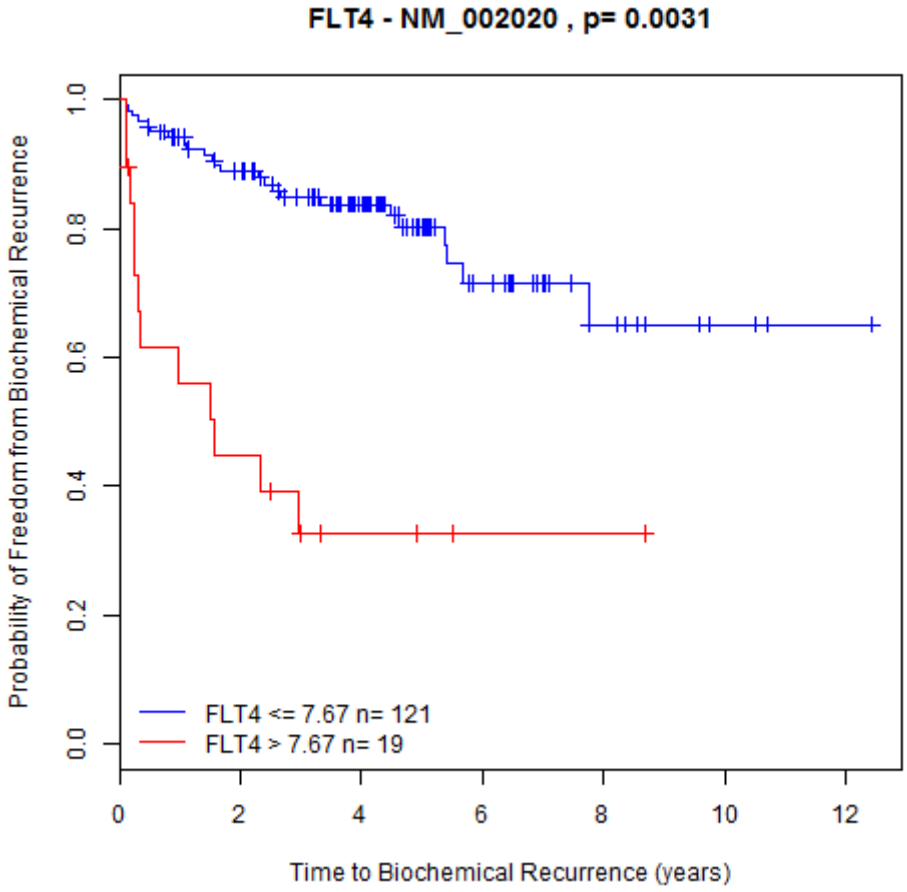



*Figure 3.53. Survival curve obtained from the recursive partitioning analysis of the Taylor dataset for FLT4.*

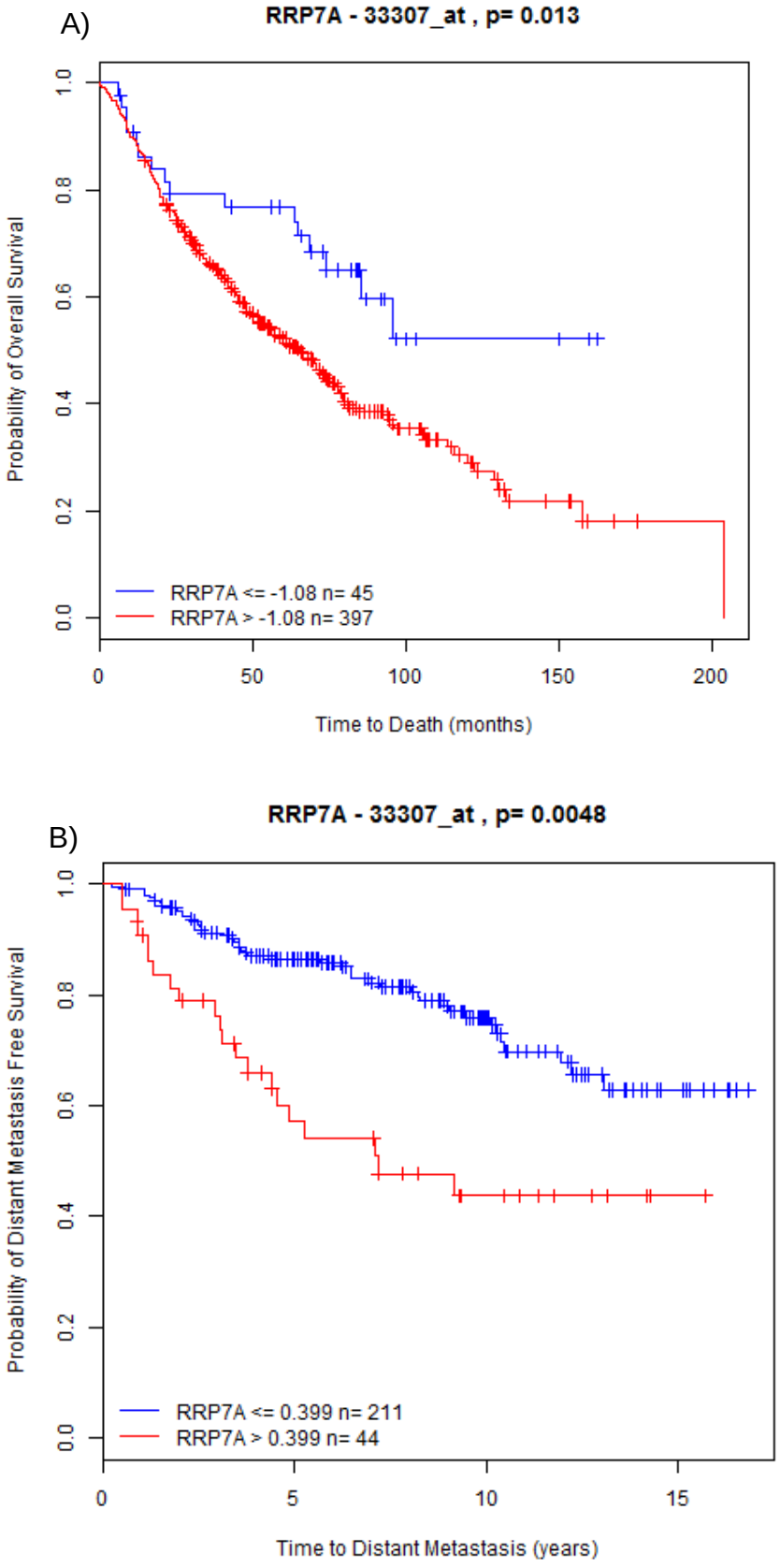



*Figure 3.54. Survival curves obtained from the recursive partitioning analysis of the Shedden (A) and Loi (B) dataset for FLT4*

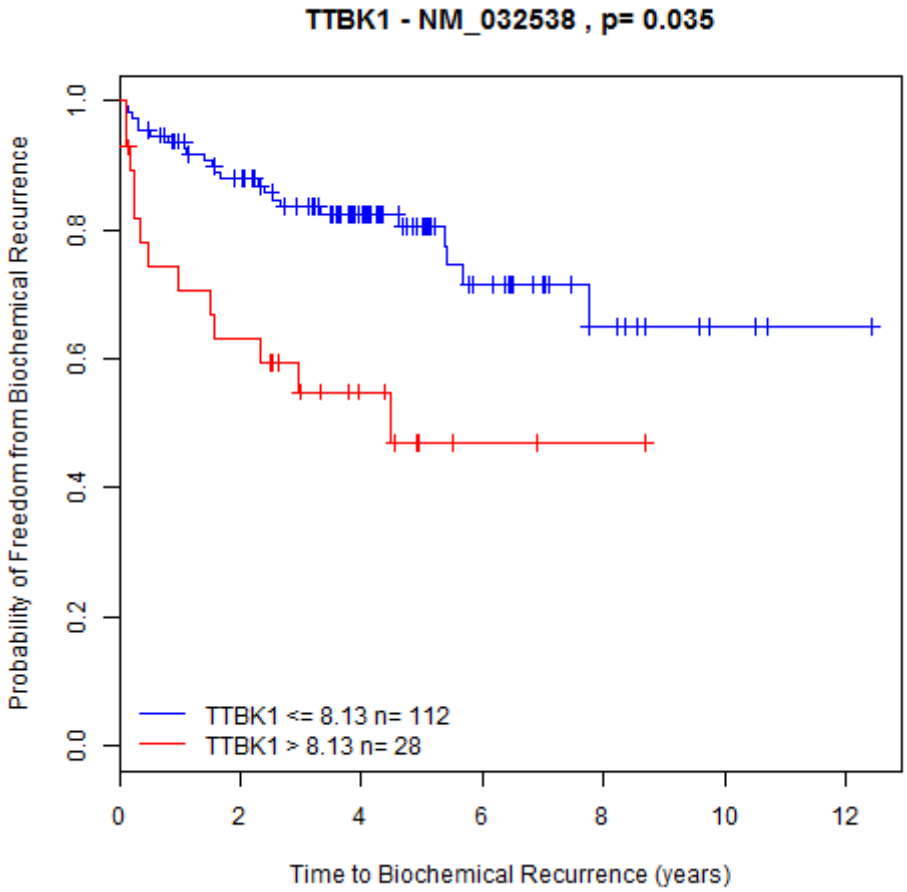



*Figure 3.55. Survival curve obtained from the recursive partitioning analysis of the Taylor dataset for TTBK1.*

# Chapter 4. A pooled genome-wide shRNA screen for genes involved in β-cell proliferation and resistance to death caused by lipids.

## 4.1. Introduction

### 4.1.1 Type 2 diabetes and β-cell mass.

Type 2 diabetes (T2D) is a chronic disease characterized by a disequilibrium in blood glucose homeostasis as a result of an increase in insulin resistance and a degree of β-cell dysfunction. The relative importance of both factors and the order in which they arise has been and still is largely discussed (Martin, Warram et al. 1992, Prentki, Joly et al. 2002, Weir and Bonner-Weir 2004, Kahn, Zraika et al. 2009).

β-cell mass is highly regulated and normally increases by β-cell proliferation during development and childhood, with a marked decrease in adulthood (Dor, Brown et al. 2004, Georgia and Bhushan 2004, Meier, Butler et al. 2008). A β-cell mass increase can also be demonstrated in response to pathological or physiological changes in insulin requirements such as obesity (Terauchi, Takamoto et al. , Kloppel, Lohr et al. 1985, Pick, Clark et al. 1998, Butler, Janson et al. 2003, Hull, Kodama et al. 2005, Gonzalez, Merino et al. 2013, Mao, Dillon et al. 2014, Irles, Neco et al. 2015) or pregnancy (Green and Taylor 1972, Van Assche, Aerts et al. 1978, Sorenson and Brelje 1997). The increase in β-cell mass in obesity in humans is calculated to be 50% compared to non-obese individuals (Butler, Janson et al. 2003) while in mouse this increase has been shown to reach nine-fold (Butler, Janson et al. 2003).

β-cell mass is lower in diabetic patients than weight-matched controls (Maclean and Ogilvie 1955, Butler, Janson et al. 2003, Rahier, Guiot et al. 2008). It is difficult to assess in humans whether this decrease in β-cell mass is a cause or consequence of diabetes, but recent studies show a reduction in β-cell function in pre-diabetic states (Do, Gunton et al. 2016) which may indicate that β-cell mass loss precedes overt diabetes.

Variations in β-cell mass can be due to an increase in proliferation, a decrease in apoptosis or both. An overview of the knowledge on the mechanisms of β-cell proliferation and apoptosis is provided in the rest of this introduction.

### 4.1.2 Overview of genes involved in β-cells proliferation

Proliferation of β-cells has been predominantly studied in *in vitro* or *in vivo* rodent models due to the difficulty to obtain and culture human β-cells. Numerous studies performed in mouse β-cells have provided insight into the mechanism of proliferation in response to nutrients and growth factors.

#### *4.1.2.1 PI3K/AKT pathway*

The PI3K/AKT pathway is a highly conserved signalling pathway involved in the regulation of cell proliferation and apoptosis amongst other cellular functions. In mouse β-cells, insulin and insulin-like growth factor 1 (IGF) (Shepherd, Navé et al. 1995) and GLP-1 (Park, Dong et al. 2006) activate the PI3K/AKT pathway through insulin receptor substrates (IRS). Signalling via the insulin receptor is critical for adult β-cell mass regulation in insulin resistant mice but not for the early development of β-cells (Okada, Liew et al. 2007). The role of Igf1 as an enhancer of β-cell proliferation in mice has however been questioned (Lu, Herrera et al. 2004, Okada, Liew et al. 2007). Indeed, mice lacking the receptor for Igf1 (Igf1r) do not show a decrease in β-cell mass although they develop impaired glucose tolerance (Kulkarni, Holzenberger et al. 2002).

Two IRS isoforms are ubiquitously expressed: *Irs1* (Sun, Rothenberg et al. 1991) and *Irs2 (*Sun, Wang et al. 1995), but their role in β-cell proliferation signalling differs. *Irs2* knockout mice have a significantly lower β-cell mass than *Irs1* knockouts (Withers, Gutierrez et al. 1998, Kubota, Tobe et al. 2000). Choudhury and colleagues provided further evidence of the requirement of Irs2 for β-cell mass increase by showing that β-cells expressing *Irs2* repopulated the islets in partial knockouts (Choudhury, Heffron et al. 2005).

The requirement of components of the PI3K/AKT pathway for β-cell proliferation has been demonstrated in studies where deletion of either *Pi3k* (Kaneko, Ueki et al. 2010) or *Akt1* (Bernal-Mizrachi, Fatrai et al. 2004) caused a decrease in β-cell mass. Deletion of *Akt2* in mice causes diabetes due to increased insulin resistance despite β-cell mass increase (Cho, Mu et al. 2001, Garofalo, Orena et al. 2003). Therefore, *Pi3k* and *Akt1* but not *Akt2* are required for β-cell proliferation.

Pi3k modulates the activity of Akt through Pdk1. Mice lacking *Pdk1* in β-cells develop overt uncontrolled diabetes as a result of their significantly decreased β-cell mass (Hashimoto, Kido et al. 2006).

Control of cell proliferation by *Akt1* in β-cells can be achieved through different pathways. Overexpression of constitutively active *Akt1* in mice has been shown to increase phosphorylation of GSK3β and levels of cdk4, cyclin D1, cyclin D2, and p21 (Fatrai, Elghazi et al. 2006). The same study found that inactivating phosphorylation of Foxo1 was increased in response to the overexpression of active *Akt1* (Fatrai, Elghazi et al. 2006).

The PI3K/AKT pathway is modulated by Pten, which inhibits Pi3k. Mice with specific deletion of *Pten* in β-cells have increased β-cell mass and are protected against streptozotocin-induced diabetes (Stiles, Kuralwalla-Martinez et al. 2006). This role of Pten is not confined to β-cell proliferation in early development as deletion of *Pten* in adult β-cells also results in the same phenotypes (Yang, Bayan et al. 2014).

#### *4.1.2.2 Mammalian Target of Rapamycin (mTOR)*

Mammalian Target of Rapamycin (mTOR) is involved in the regulation of cell growth, proliferation and survival through the formation of two different complexes, mTORC1 and mTORC2. Growth factors induce mTORC1-mediated cell growth and proliferation through inactivation of TSC (tuberous sclerosis complex), which is inhibited by Akt (Inoki, Li et al. 2002, Manning, Tee et al. 2002).

mTORC1 activation through deletion of *Tsc2* in mouse β-cells leads to increased β-cell size and number (Rachdi, Balcazar et al. 2008). This effect is reverted by the inhibition of mTORC1 using rapamycin (Rachdi, Balcazar et al. 2008), which demonstrates the involvement of mTORC1 in the control of β-cell mass.

mTORC1 also mediates the effect of PKCζ in β-cell proliferation (Velazquez-Garcia, Valle et al. 2011). PKCζ has been shown to be involved in the stimulation of β-cell proliferation by GLP-1 using the rat insulinoma cell line INS1 (Buteau, Foisy et al. 2001). Hepatocyte growth factor (HGF) and

parathyroid hormone-related protein (PTHrP) also stimulate β-cell proliferation through PKCζ according to *in vitro* studies using INS1, MIN6 (a mouse insulinoma cell line) and primary mouse and human islets (Miele, Raciti et al. 2007, Vasavada, Wang et al. 2007). Mice expressing constitutively active PKCζ in β-cells display increased β-cell proliferation but also an increase in inactivating phosphorylation of GSK3β (Velazquez-Garcia, Valle et al. 2011), suggesting that the effect of PKCζ on β-cell proliferation may also be mediated through GSK3β.

#### *4.1.2.3 Wnt canonical signalling*

In the canonical Wnt pathway, Wnt ligands bind the Frizzled receptor and initiate a cascade by which β-catenin localizes to the nucleus, where it interacts with transcription factors from the T-cell-specific factor/lymphoid enhancer-binding factor family (TCF/LEF) and stimulates the transcription of Wnt regulated genes (reviewed in (Welters and Kulkarni 2008)). Activation of the Wnt pathway in INS1 cells results in higher proliferation rate (Welters, Oknianska et al. 2008). In mouse primary islets, the activation of Wnt signalling increases the expression of cyclin D2 and increased β-cell proliferation (Rulifson, Karnik et al. 2007).

GLP-1 has been shown to stimulate β-cell proliferation in isolated primary islets and INS1 cells at least in part through activation of Wnt signalling (Liu, Remedi et al. 2009).

#### *4.1.2.4 Other pathways.*

Prolactin and growth hormone induce β-cell proliferation through Stat5, and it has been shown that in INS1 cells activation of Stat5 drives proliferation via cyclin D2 (Friedrichsen, Richter et al. 2003). Stat5 upregulates *Bcl6* expression, which inhibits menin and promotes β-cell proliferation (Karnik, Chen et al. 2007).

Glucose can upregulate the expression of cyclins and increase β-cell proliferation via ChREBP (carbohydrate response element-binding protein) (Metukuri, Zhang et al. 2012). It can also act through Lkb1 and calcineurin (Heit, Apelqvist et al. 2006, Fu, Ng et al. 2009, Fu, Ng et al. 2009, Granot, Swisa et al. 2009, Metukuri, Zhang et al. 2012).

Other growth factors and pathways have been demonstrated to be involved in the regulation of β-cell proliferation such as the epidermal growth factor (EGF) (Hakonen, Ustinov et al. 2011), cAMP (Schisler, Fueger et al. 2008, Song, Schreiber et al. 2008), leptin (Morioka, Asilmaz et al. 2007), transforming growth factor- β (TGFβ) (Brown and Schneyer 2010), or platelet derived growth factor (PDGF) (Chen, Gu et al. 2011).

#### *4.1.2.5 Proliferation of human β-cells*

Much less is known about the control of β-cell mass through proliferation in humans. Insulin is known to affect human β-cell proliferation through a mechanism impaired in T2D patients (Lopez, Cypess et al. 2011, Mari, Tura et al. 2011, Halperin, Lopez et al. 2012). A role for mTORC has also been proposed (Liu, Remedi et al. 2009) and so have PKCζ (Vasavada, Wang et al. 2007) and PDGF (Chen, Gu et al. 2011). Glucose seems to act through similar pathways than in mice (Metukuri, Zhang et al. 2012, Poungvarin, Lee et al. 2012).

Variants in one of the TCF/LEF transcription factors, *TCF7L2*, have been associated with increased risk of T2D (Damcott, Pollin et al. 2006, Florez , Jablonski et al. 2006, Grant, Thorleifsson et al. 2006, Groves, Zeggini et al. 2006, Saxena, Gianniny et al. 2006, Scott, Bonnycastle et al. 2006, Zhang, Qi et al. 2006), which may indicate a role of Wnt canonical signalling in human β-cell proliferation.

### 4.1.3 Apoptosis in β-cells

The decrease in β-cell mass observed in T2D patients may be due at least in part to an increased level of apoptosis (Butler, Janson et al. 2003). Conditions associated with T2D, such as hyperglycaemia have shown to induce apoptosis of human β-cells (Efanova, Zaitsev et al. 1998, Maedler, Spinas et al. 2001). Elevated glucose leads to overexpression of pro-apoptotic genes such as *Bad*, *Bid* and *Bik* (Federici, Hribal et al. 2001). Furthermore, β-cells release interleukin (IL)-1β in hyperglycaemic conditions (Maedler, Sergeev et al. 2002), which leads to apoptosis via NF-κB. These IL-1β producing cells are found in pancreata from T2D patients but not in non-diabetic controls (Maedler, Sergeev

et al. 2002). Glucotoxicity may also be a result of increased mitochondrial generation of reactive oxygen species (ROS) due to excessive glucose metabolism. Anti-oxidant enzymes are not highly expressed in β-cells (Atkinson, Bluestone et al. 2011) and markers of oxidative stress are higher in pancreata of T2D affected individuals (Del Guerra, Lupi et al. 2005).

High concentrations of saturated fatty acids such as palmitate are also known to be toxic for β-cells (Maedler, Spinas et al. 2001, Piro, Anello et al. 2002) by a process not completely understood but which seems to implicate ceramide synthesis and serine protease activity (Shimabukuro, Zhou et al. 1998, Lupi, Dotta et al. 2002). Lipotoxicity seems to be associated with the accumulation of saturated fatty acids but not with exposure to unsaturated fatty acids (Cnop, Hannaert et al. 2001, Maedler, Spinas et al. 2001).

Amyloid deposits have also been investigated as pro-apoptotic stimuli in the pathogenesis of T2D. β-cells co-secrete islet amyloid polypeptide (IAPP) with insulin, although its role is still debated. It was first thought to play a role in insulin resistance as it inhibits insulin-stimulated glucose uptake in rat (Cooper, Leighton et al. 1988) and human (Zierath, Galuska et al. 1992) skeletal muscle *in vitro*. However, the concentrations used in these studies were higher than the physiological ranges. IAPP has been reported to affect insulin secretion *in vivo* in perfused rats (Dégano, Silvestre et al. 1993) and *in vitro* in cultured β-cells (Sandler and Stridsberg 1994) while other studies did not find an effect in insulin secretion (O'Brien, Westermark et al. 1990, Pettersson and Ahren 1990, Broderick, Brooke et al. 1991). IAPP inhibits gastric emptying (Reidelberger, Arnelo et al. 2001) and osteoclastic activity (Zaidi, Datta et al. 1990).

IAPP deposits have are frequently found in β-cells from T2D affected individuals (Westermark and Wilander 1978, Clark, Wells et al. 1988, O'Brien, Butler et al. 1994), and human islets cultured in the presence of high levels of glucose show an increase in IAPP deposits (Lorenzo, Razzaboni et al. 1994, Konarkowska, Aitken et al. 2006)

### 4.1.4 Aim of this chapter

In this chapter I present two genome-wide RNAi screens to identify genes involved in β-cell proliferation and death by lipids in a reversibly immortalized mouse β-cell line. To my knowledge, no genome-wide RNAi screens have yet been performed in β-cell lines for any of these phenotypes. As mentioned above, only a few molecular mechanisms are known for these beta cell phenotypes which were identified predominantly through targeted investigations.

This chapter presents the design, optimization and analysis required for these two screens and their potential implications for this research area. The main objective of both screens was to uncover genes which upon knockdown increased proliferation or resistance to death by lipids, two processes which could be of importance in the understanding of β-cell dysfunction in T2D.

## 4.2. Material and methods

### 4.2.1 Cell lines

R7T1 cells were a gift from Dr Shimon Efrat. They are a mouse insulinoma cell line that conditionally express an oncogene (SV40 large T antigen) under the tet-on system (Milo-Landesman, Surana et al. 2001). R7T1 cells were derived from mice resulting from the crossing of transgenic mice expressing the RIP7-rtTA (rat insulin promoter-reverse tetracycline-controlled transactivator) construct and transgenic mice expressing a tet-TAG construct, which consists of the SV40 large T antigen gene (TAG) flanked by repeats of tet operator sequences and a minimal promoter (Milo-Landesman, Surana et al. 2001). Double transgenic mice remained euglycaemic and had normal islet morphology unless they were treated with a tetracycline analogue, in which case they developed insulinomas and manifested hypoglycaemia (Milo-Landesman, Surana et al. 2001). Milo-Landesman and collaborators characterized R7T1 and found that the rate of BrDU incorporation in these cells was 50 times higher following treatment with doxycycline, a tetracycline analogue (Milo-Landesman, Surana et al. 2001). R7T1 cells were routinely grown in DMEM (high glucose, no pyruvate) supplemented with 10% FBS (tetracycline-free certified) and doxycycline (DOX) at (5 µg/mL). Medium was renewed every 48 hours.

### 4.2.2 Bromodeoxyuridine (BrDU) incorporation assay

Bromodeoxyuridine (BrDU) stock was prepared as a 50 mM BrDU solution in double distilled $H_2O$. This solution was filter-sterilised and stored at 4ºC.

Cells were incubated overnight at standard tissue culture conditions with growth medium supplemented with 50 µM BrdU. Then, growth medium was aspirated and cells were very gently washed with DPBS. Ice cold methanol was added to the cells, the dish was rocked and methanol was aspirated immediately. Fresh ice cold methanol was added to cover the dish and it was left to incubate for four minutes. Cells were then washed five times with DPBS. After the last wash, cells were covered with 1.5 M HCl and incubated for 20 minutes at 37ºC. HCl was then aspirated and cells were washed five times with DPBS. Incorporation of BrDU was detected with an anti-BrDU mouse antibody diluted 1:100 in 10% FCS in DPBS, and incubated overnight at 4ºC. Cells were then washed three

times with DPBS before being incubated with AP conjugated goat anti-mouse IgG diluted 1:100 in 10% FCS in DPBS for 45 minutes. After the incubation, cells were washed five times with DPBS. Cells were covered with fresh AP substrate solution (70 ml of alkaline phosphatase buffer (100 mM NaCl, 5 mM MgCl2, 100 mM Tris pH 9.5) with 462 µl of NBT (Nitro-blue tetrazolium chloride, dissolved at 50 mg/ml in 70% dimethylformamide) and 231 µl of BCIP (5-bromo-4-chloro-3'-indolyphosphate p-toluidine salt, dissolved at 50 mg/ml in 100% dimethylformamide)). SVE cells (section 3.2.1) were used as a positive control for incorporation of BrDU. They were cultured for 30 minutes in the presence of BrDU instead of overnight as they divide rapidly. Development of staining in the positive control was monitored under a bright field microscope until the positive nuclei were clearly visible. Then, the reaction was stopped by adding DPBS and cells were washed five times. DPBS was added to the wells and the stained plate was stored at 4ºC until imaged.

#### 4.2.3 Immunocytochemistry (ICC) of SV40 Large Tumour Antigen (Tag)

Growth medium was aspirated and cells washed with DPBS. Cells were washed with ice cold methanol. Cells were incubated at room temperature with fresh ice cold methanol for ten minutes and then washed five times with DPBS. Cells were covered with a 1:1000 dilution of anti-Tag antibody in 10% FCS in DPBS and incubated overnight at 4ºC. Cells were washed five times with DPBS before being incubated for 45 minutes with AP conjugated goat anti-mouse IgG (1:100 in 10% FCS in DPBS). Staining was performed as in 4.2.2.

#### 4.2.4 Production of retroviral particles using Phoenix cells.

Ecotropic Phoenix cells were seeded 24 hours before transfection at a density that would give approximately 70% confluence on the day of transfection. Cells were transfected using the ProFection Mammalian Transfection System (Promega) using 15 µg of DNA per 10 cm dish according to the manufacturer instructions. After 24 hours growth medium was replaced with 12 mL of fresh growth medium.

#### 4.2.5 Transduction of cells using non-concentrated retroviral particles

Retroviral medium was collected 48 hours after transfection, transferred to 15 mL tubes and centrifuged at 500 g for five minutes to pellet any Phoenix cells in

suspension. The supernatant was carefully transferred to a new 15 mL tube and polybrene was added to a final concentration of 8 µg/ml. Growth medium was aspirated from the dishes to be transduced and replaced with the retroviral medium.

#### 4.2.6 Concentration of retroviral particles using Retro-X Concentrator (Clontech).

Retroviral medium was collected and centrifuged as in 4.2.5. Retroviral particles were concentrated according to the manufacturer instructions and re-suspended into the appropriate volume of growth medium to reach the concentration factor stated in the results section for each figure.

#### 4.2.7 Concentration of retroviral particles using PEG

A solution of 40% polyethylene glycol (PEG) 8000 was prepared in sterile glass bottles by dissolution of the appropriate amount of PEG in autoclaved ddH2O and stored at 4ºC. Retroviral medium was harvested as in 4.2.5. Three volumes of retroviral medium were mixed with one volume of PEG solution and incubated overnight at 4ºC on a shaker. The solution was centrifuged at 4ºC for 45 minutes at 1000 g. The supernatant was decanted and the pellet gently re-suspended in complete growth medium using a volume that would give the concentration factor stated in the results section for each figure.

#### 4.2.8 Transductions with Ecotropic receptor booster

Cells were pre-treated with the Ecotropic Receptor Booster following the manufacturer instructions. Retroviral medium was collected and centrifuged at 500 g for five minutes to pellet any Phoenix cells in suspension. The supernatant was transferred to a new 15 mL tube and polybrene was added to a final concentration of 8 µg/mL.

#### 4.2.9 Transductions with Viraductin

Retroviral medium was collected and centrifuged as in 4.2.5. Cells were transduced following the manufacturer instructions. Briefly, reagent A was added to the retroviral medium and mixed. This was repeated with reagent B. The retroviral medium was then aliquoted to 2 mL microcentrifuge tubes and incubated at 37ºC for 30 minutes. The tubes were then centrifuged at 12,000 g for 10 minutes. The supernatant was carefully aspirated using a micropipette

and the pellet was re-suspended in a volume of growth medium determined by the concentration factor stated in the results section for each figure.

#### 4.2.10 RNA extractions

Total RNA was extracted from cells in 10 cm dishes using the RNeasy Mini Kit (Qiagen) according to the manufacturer instructions. RNA was eluted in 50 µl of nuclease-free water.

#### 4.2.11 RT PCR

cDNA was produced by RT-PCR using RevertAid™ H Minus Reverse Transcriptase (Fermentas). The final composition of a RT-PCR is shown in Table 4.1.

| Component | Volume |
|---|---|
| Template RNA | (1 µg) |
| Oligo $(dT)_{20}$ primer | 1 µL |
| 5X reaction buffer | 4 µL |
| RNase inhibitor (20 U/µL) | 1 µL |
| 10 mM dNTP mix | 2 µL |
| RevertAid RT (200 U/µL) | 1 µL |
| RNase free water | To 20 µL |

*Table 4.1. Composition of a RT-PCR reaction*

4.2.12 qPCR

cDNA obtained in 4.2.11 was diluted 1:20 using RNase-free water and qPCR reactions were performed on an ABI 7900HT real-time PCR system using SYBR Green dye. Reactions were performed in triplicates and analysed using the ΔΔCT method. The composition of a PCR reaction is depicted in Table 4.2.

| Component | Volume |
|---|---|
| 2X Sybr Green Master Mix | 10 µL |
| Forward primer (10 mM) | 0.9 µL |
| Reverse primer (10 mM) | 0.9 µL |
| cDNA | 4 µL |
| Water | 4.2 µL |

*Table 4.2. Composition of a single qPCR reaction.*

### 4.2.13 Palmitate dose response

A 50 mM stock of palmitate in ethanol was prepared and stored as 200 µl aliquots at -20ºC. An aliquot was thawed and heated to 80ºC prior to use. Working solutions of different concentrations were prepared (Table 4.3) and incubated for at least 1 hour at 37ºC. “Dose” medium was prepared as a 1:10 dilution of each working solutions. Growth medium was aspirated and replaced by palmitate-containing medium, which was refreshed every 24 hours. At the end of the experimental time period, cell death was assayed by PI exclusion.

| Dose (mM) | Stock (mM) | BSA (10%) | Ethanol (50%) | Palmitate (50 mM) |
|---|---|---|---|---|
| Control | 0 | 630 µL | 70 µL | 0 µL |
| 0.025 | 0.25 | 630 µL | 66.5 µL | 3.5 µL |
| 0.05 | 0.5 | 630 µL | 63 µL | 7 µL |
| 0.1 | 1 | 630 µL | 56 µL | 14 µL |
| 0.25 | 2.5 | 630 µL | 35 µL | 35 µL |
| 0.5 | 5 | 630 µL | 0 µL | 70 µL |

*Table 4.3. Composition of the stock solutions for the palmitate dose response.*

### 4.2.14 PCR amplification of shRNA constructs for sequencing

The targeting regions of the shRNA constructs were PCR amplified in two steps. The first reaction was setup as in Table 4.4 with primers specified in Table 4.5. PCR products were separated by electrophoresis on a 1.5% agarose gel and purified as in section 2.2.7. PCR products of the correct size were extracted from the gel, quantified using Qubit and all the pools mixed in equal proportions. This mix was then used as template for a second round of amplification set up as in Table 4.6. This second round introduced sample-specific adaptor sequences for the Illumina sequencer using a different combination of Nextera index primers for each sample (T0: N711 and S517, T7: N711 and S502, T14: N711 and S503, P: N711 and S504, Control: N711 and S505).

| Component | Volume |
|---|---|
| Q5 buffer 5X | 20 µL |
| dNTP mix (10 mM) | 2 µL |
| Forward primer (10 mM) | 5 µL |
| Reverse primer (10 mM) | 5 µL |
| Betaine | 20 µL |
| Q5 polymerase | 1 µL |
| DNA | 3 µL |
| Water | 44 µL |

*Table 4.4. Composition of a single PCR reaction for the first step of the shRNA targeting sequence amplification.*

| Primer | Sequence | Condition amplified |
|---|---|---|
| Twostep_hiseq1 | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGAGGTCGACGGTCGACGGTAT | T14 and Control |
| Twostep_hiseq2 | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTCTGAAGGTCGACGGTA | T0 |
| Twostep_hiseq8 | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATCTCTCGAGGTCGAC | T7 |
| Twostep_hiseq4 | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGGTCGACGGTATCGAT | Palmitate |
| Twostep_hiseq6 | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTCTTTGGATTTGGGAATC | Reverse, common to all |

*Table 4.5. Primers used for the first step amplification of the shRNA targeting sequences for each condition.*

| Component | Volume |
|---|---|
| Q5 buffer 5X | 20 µL |
| dNTP mix (10 mM) | 2 µL |
| Forward primer (10 mM) | 5 µL |
| Reverse primer (10 mM) | 5 µL |
| Betaine | 20 µL |
| Q5 polymerase | 1 µL |
| DNA | 50 ng |
| Water | Up to 50 µL |

*Table 4.6. Composition of a single PCR reaction for the second step of the shRNA targeting sequence amplification.*

4.2.15 Purification of amplified DNA using Agencourt Ampure XP beads.

PCR products containing the shRNA targeting sequences amplified as in 4.2.14 were purified using a bead system (Agencourt Ampure XP) following the manufacturer instructions and eluted in 40 µl of TE buffer.

4.2.16 Analysis of sequencing data.

The sequencing output was received as one .fastq file per sample, which were transferred to a UNIX server. Sequencing reads were extracted to plain text files using the command “grep” on the command line. As the samples were already in different files, reads containing shRNA targeting sequences were directly extracted from the corresponding file using the shRNA loop sequence instead of the PCR primer as in section 3.2.8. The Python scripts generated for the Chapter 3 were used to produce a file with the targeting sequences found in the reads and the number of reads containing each sequence in the sequencing data. These files were merged using Stata to produce one file with the number of reads containing each sequence in each sample. Similarly to the procedures in chapter 3, sequences with less than 50 reads in any of the control samples were coded as “missing data” in that sample. Then, raw number of reads were converted to reads per million per pool, and the ratio of (experimental condition) to (control condition) was calculated. From this ratio, a robust z-score was determined for assessment of construct effects. Figure 4.1 depicts this analysis pipeline, and the Stata do-files corresponding to this analysis steps can be found in Appendix 7.

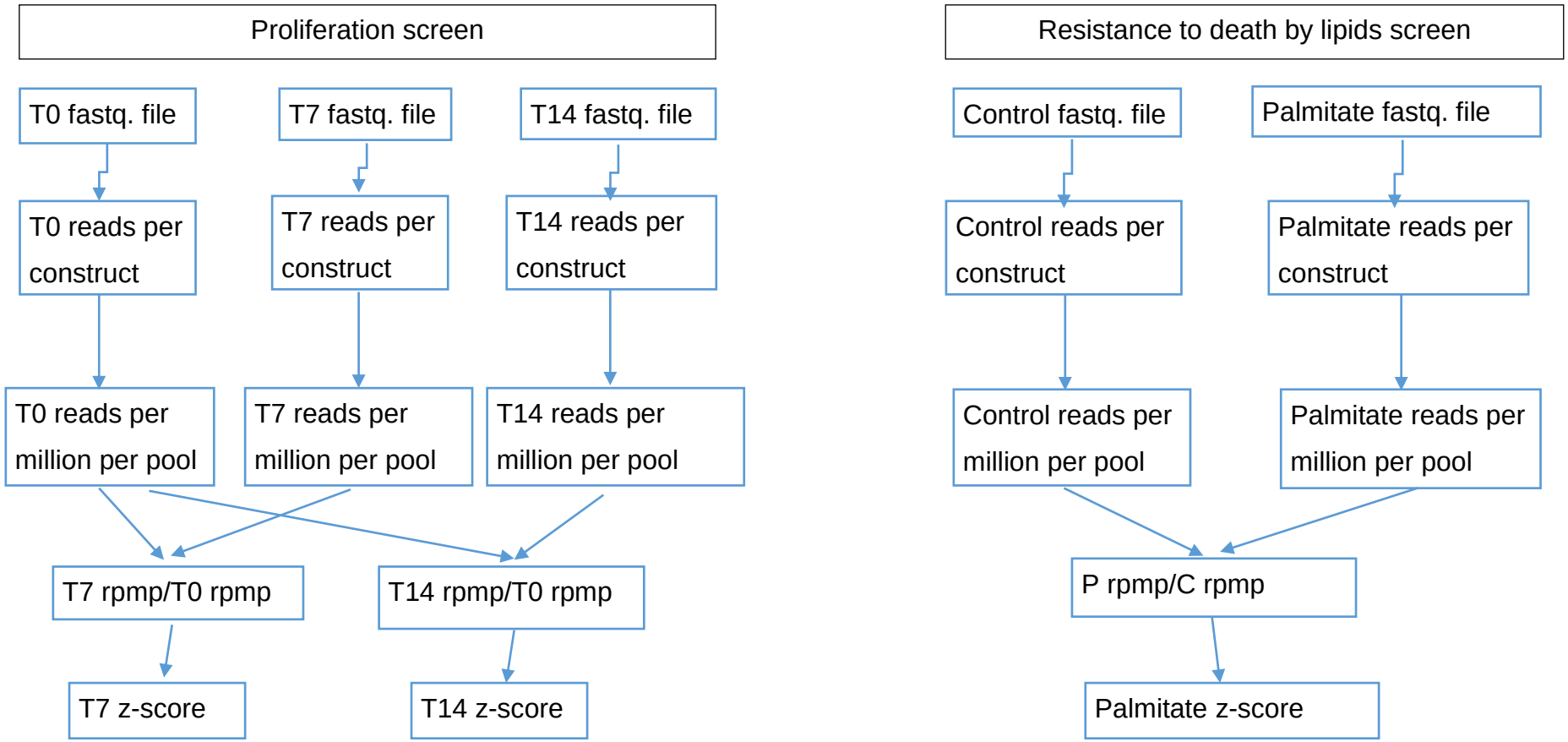



*Figure 4.1. Schematic depiction of the analysis pipeline used.*
*Sequencing reads were extracted from the .fastq files in a UNIX server using the command "grep". Reads were aligned to the shRNA library and counted to produce a file containing the number of reads per shRNA construct. These raw number of reads were converted to reads per million per library pool. A ratio of each experimental condition to the appropriate control (T0 for the proliferation screen samples, C for the palmitate sample) was calculated and used to build robust z-scores as estimates of shRNA construct effect sizes.*

### 4.2.17 Design of the screening process

A genome wide RNAi screen using a shRNA library was designed to identify genes with functions on β-cell proliferation and resistance to death induced by lipids. The screening process will be explained in more detail in the sections to follow, but a simplified schematic depiction of the screening strategy is shown in Figure 4.2

Briefly, the library of shRNA targeting sequences described in section 2.4 was packaged into retroviral particles using an ecotropic packaging cell line (Phoenix). The retroviral particles were then used to transduce a β-cell line (R7T1) which conditionally expresses an oncogene in the presence of doxycycline. Cells containing the library sequences were selected using puromycin and cultured in absence of doxycycline for seven days to remove the expression of the oncogene. A control sample ("T0") was then harvested and the rest of the cells entered one of the screening processes.

For the first screen, cells in the 'proliferation' screen were cultured in absence of doxycycline for seven or fourteen days and then harvested, constituting the "T7" and "T14" proliferation samples respectively.

For the second screen, cells in the 'death by lipids' screen were cultured in absence of doxycycline and exposed to palmitate or vehicle-only treatments for 96 hours. Cells were then harvested to constitute the "P" and "C" samples respectively.

DNA was extracted from all the samples and the shRNA targeting sequences were PCR amplified and sequenced to be used as a proxy for cell number. Increases in representation of the shRNA targeting sequences in the experimental samples with respect to their controls were assessed using robust z-scores. These z-scores were used to identify and prioritize those shRNA targeting sequences with a stronger effect on cell number, yielding a list of genes of interest (GOI) for each phenotype screened – increase of proliferation of a β-cell line in the absence of an oncogene and increase in the resistance to death induced by palmitate of a β-cell line in the absence of an oncogene.

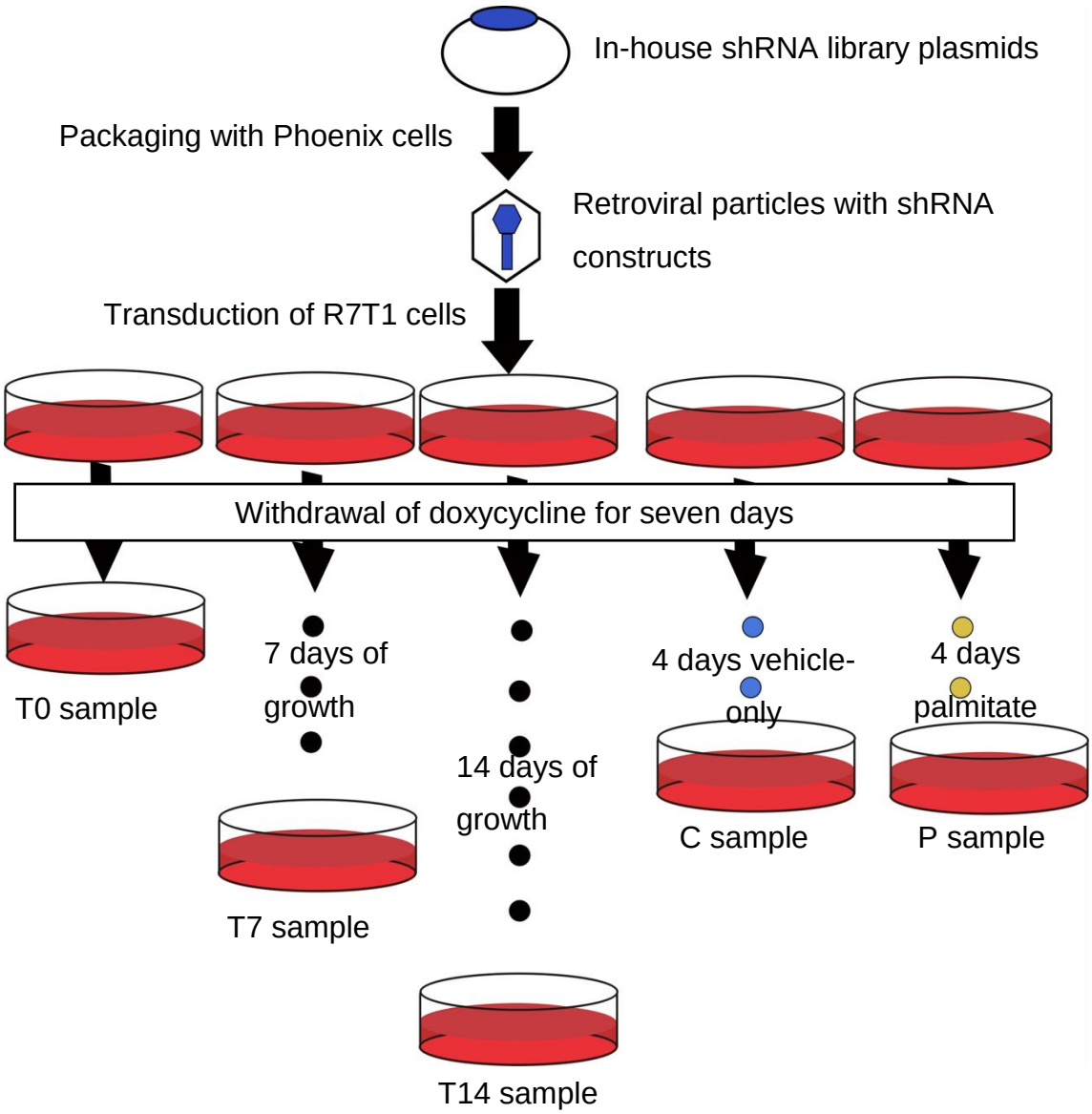



*Figure 4.2. Schematic simplified overview of the screening process*

## 4.3. Results

### 4.3.1. Description of the cell line used as a β-cell *in vitro* model

R7T1 is a mouse insulinoma cell line reversibly immortalized with the SV-40 large T oncogene expressed under a tet-on system. The oncogene expression is activated in the presence of tetracycline or one of its analogues. As a consequence, this cell line is reversibly immortalized. The cell line was maintained in culture with expression of the oncogene by addition of a tetracycline analogue (doxycycline, DOX) to the growth medium.

R7T1 has not been extensively used as an *in vitro* model of β-cells and it was therefore necessary to assess its suitability. Expression of widely used markers of β-cell and non-β pancreatic cells was assessed using QRT-PCR. It was also of interest to investigate whether a higher level of expression of the oncogene affected the expression of these markers. RNA was extracted from cells cultured in presence of DOX to constitute an “oncogene positive” sample. An “oncogene negative” sample was obtained from cells cultured in the absence of DOX for seven days. These RNA samples were used to obtain cDNA and then perform qPCR. MIN-6 cells, a widely used mouse insulinoma cell line that expresses the SV40 large T antigen, was used as the positive control for β-cell markers and as a negative control for makers of other pancreatic cell lineages. Reactions were performed in duplicate using SYBR green dye and two biological replicates were analysed. *Tbp* was used as the housekeeping gene to normalise expression levels in all cell lines. Relative expression levels in R7T1 compared to MIN6, represented as ΔΔCt, are depicted in Figure 4.3. These results suggest that expression of the oncogene may alter the level of expression of several of these markers, including insulin. Expression of insulin is highly downregulated when expression of the oncogene is induced in the presence of DOX. In the absence of oncogene expression induction, R7T1 cells express similar levels of β-cell markers (*Ins*, *Pdx1*, *Kcnj11*, *Nkx6.1* and *NeuroD*) to MIN6, a widely used *in vitro* model of β-cells. Relative expression of markers for other pancreatic lineages is highly variable. This is due to the variation on expression of these markers in MIN6 rather than to a high level of expression in R7T1 cells. Expression of *Cftr*, a marker of duct cells, is very low in R7T1 but almost undetected in MIN6, which yields a high fold level of expression. This is

true also for *Sst*, a marker for δ-cells and *Cpa1*, expressed in exocrine pancreatic cells. The opposite situation is seen in the fold of expression of *Gcg* (glucagon, a marker of alpha cells). Expression of *Gcg* is detected in R7T1 but it is much higher in MIN6, which causes the fold expression to appear undetected. The expression of *Gcg* in MIN6 has been previously reported (Lima, Docherty et al. 2012).

To obtain a more meaningful comparison the expression of the pancreatic markers was studied using a pancreatic cell line (DT508M) as a control instead of MIN6 (Figure 4.4). These cells are ductal cells derived from a $Kras^{LSL\text{-}G12D/+}$ mouse (Allard et al., under review) and hence are essentially normal apart from hemizygous expression of *Kras*. This comparison indicates that the expression of other pancreatic cell markers in R7T1 is not very different from a non-β-cell pancreatic cell line. Expression levels of *Gcg* appear to be higher in R7T1 than in DT508M while *Sst* and *Cpa*1 tend to be less expressed in R7T1.

Therefore, R7T1 cells in the absence of oncogene expression stimulation express similar levels of β-cell markers to MIN6 and higher levels of *Cftr*, *Sst* and *Cpa1*, and lower levels of *Gcg*. The expression of these last four markers of non-β-cell pancreatic lineages is similar to that of a non-β-cell cell line.

The QRT-PCR results also highlighted that the presence of DOX in the growth medium produced a change in expression on some of the markers under study. Importantly, the expression of insulin was highly downregulated in the presence of DOX. As the presence of DOX stimulates the expression of the oncogene, this change in expression indicates an effect of the oncogene expression and provides additional value to the use of a reversibly immortalized cell-line.

To evaluate the presence of the oncoprotein in R7T1 cells in the presence and absence of DOX, cells were cultured for different time lengths in the absence of DOX and then fixed. The oncoprotein was then detected by immunocytochemistry (ICC). The microscopy pictures in Figure 4.5 confirm that the levels of the SV40 large T antigen oncoprotein diminish in absence of DOX.

Downregulation of oncogene expression was predicted to have a direct effect on cell proliferation rate. As β-cells *in vivo* present a very low proliferation rate (Butler, Janson et al. 2003, Meier, Butler et al. 2008, Reers, Erbel et al. 2009) it was important to determine how different the proliferation rate of R7T1 was in absence of DOX. BrDU incorporation was assessed in R7T1 cells cultured in

absence of DOX for the same length of time used for the oncoprotein. Cells were incubated with BrDU overnight and then fixed. BrDU incorporation was detected as explained in section 4.2.2. Microscopy pictures of this are shown in Figure 4.6. BrDU incorporation, and therefore number of cells that have synthesized DNA, diminishes with in the absence of DOX. This decrease depends on the time that the cells are cultured without DOX.

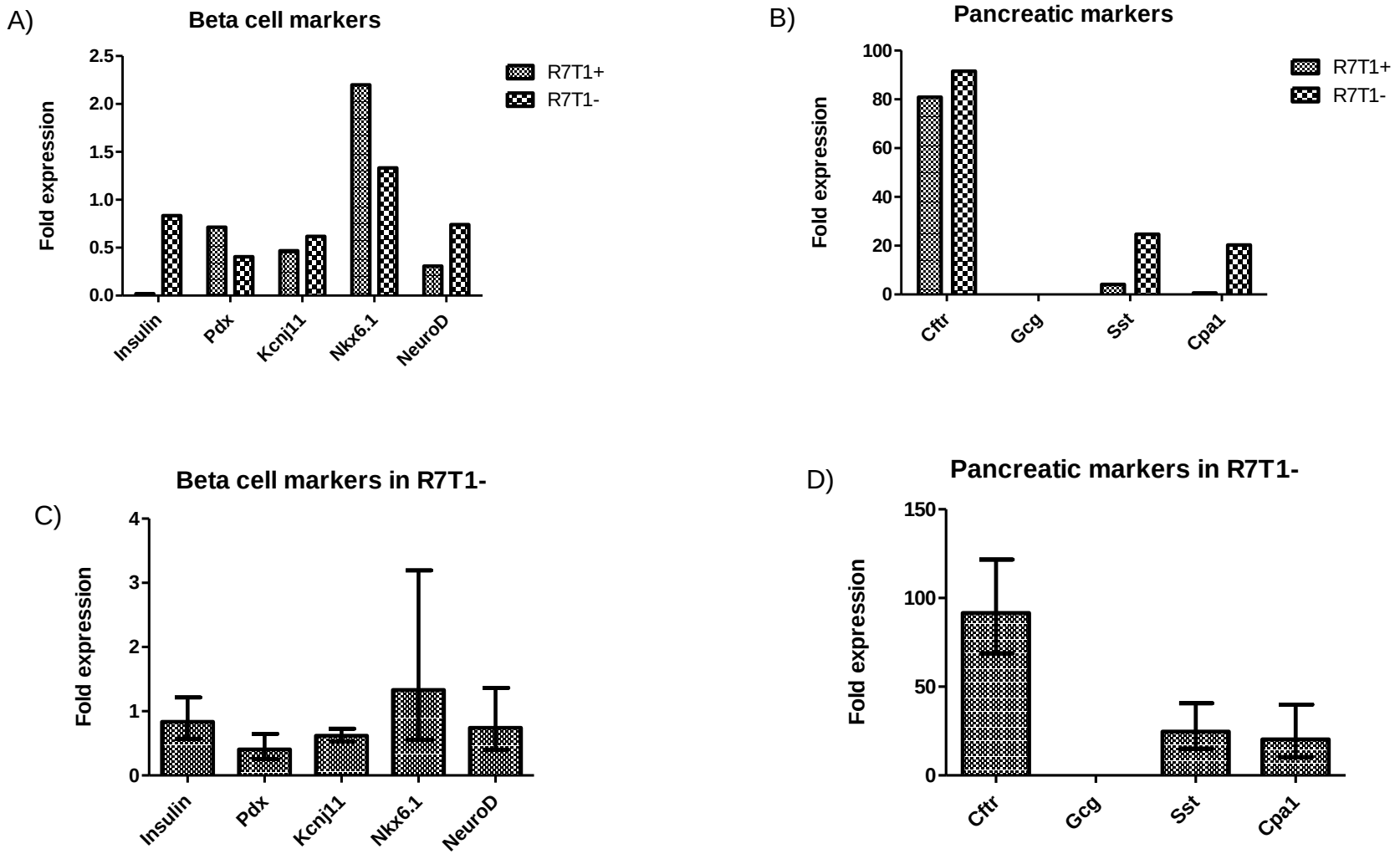



*Figure 4.3 Levels of expression of β cell markers (A, C) and markers of other pancreatic lineages (B, D) in R7T1 in presence (+) or absence (-) of DOX (A, B).*
*C and D: expression levels in absence of DOX. All expressed as ΔΔCt normalized to expression in MIN6 cells.*

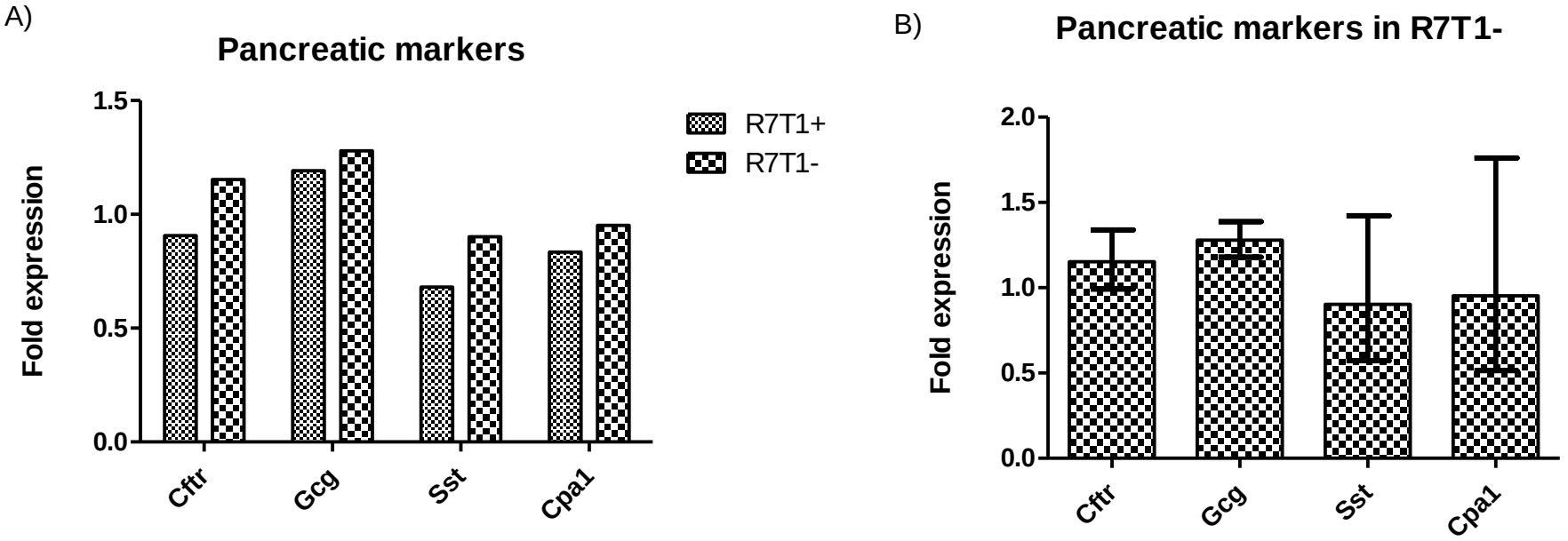



*Figure 4.4. Levels of expression of non-β-cell pancreatic lineages in R7T1 cells.*
*A) R7T1 in presence (+) or absence (-) of DOX. B) R7T1 in absence of DOX (bars indicate the 95% CI) All expressed as ΔΔCt normalized to expression in DT508M pancreatic cells.*

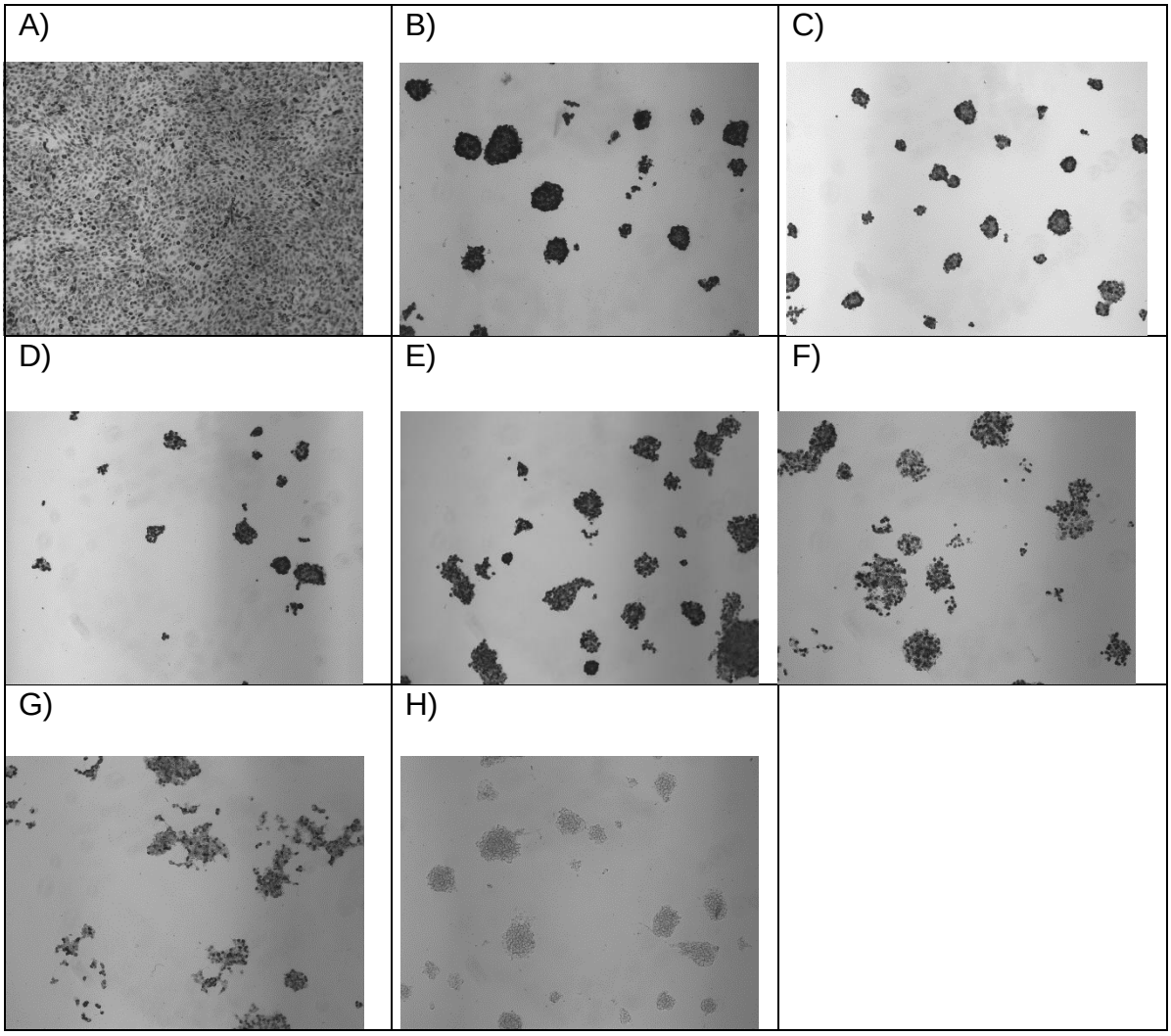



*Figure 4.5. Immunocytochemistry detection of large T Antigen (TAg) in R7T1 before and after withdrawal of DOX.*
*A) SVE cells, positive control for TAg. B) R7T1 cells cultured in the presence of DOX. C) R7T1 cells cultured in absence of DOX for 1 day. D) R7T1 cells cultured in absence of DOX for 2 days. E) R7T1 cells cultured in absence of DOX for 3 days. F) R7T1 cells cultured in absence of DOX for 5 days. G) R7T1 cells cultured in absence of DOX. H) R7T1 cells cultured in the presence of DOX, no primary antibody negative control.*

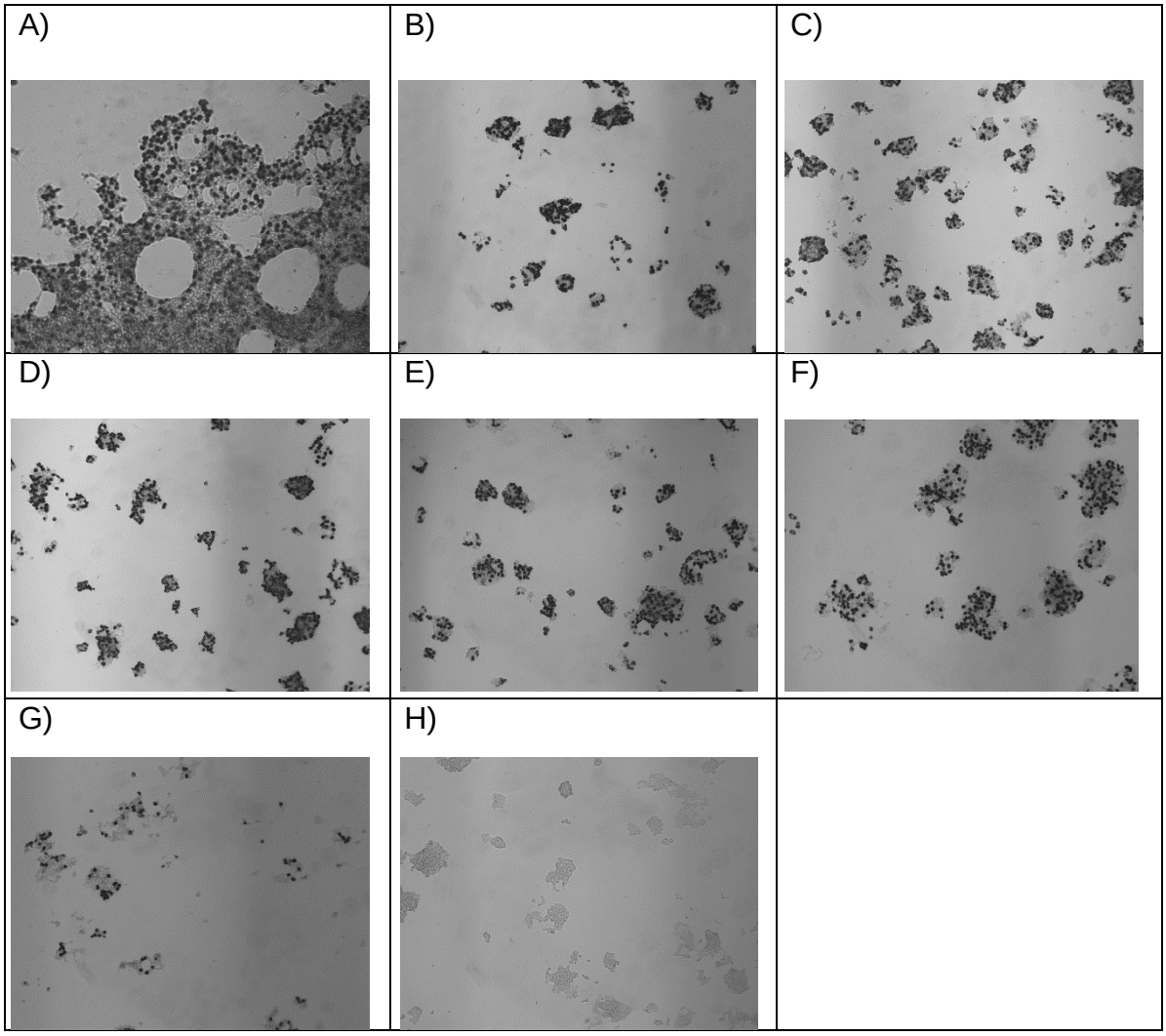



*Figure 4.6 Immunocytochemistry detection of BrDU incorportation in R7T1 before and after withdrawal of DOX.*
*A) MIN6 cells, positive control for BrDU incorporation. B) R7T1 cells cultured in the presence of DOX. C) R7T1 cells cultured in absence of DOX for 1 day. D) R7T1 cells cultured in absence of DOX for 2 days. E) R7T1 cells cultured in absence of DOX for 3 days. F) R7T1 cells cultured in absence of DOX for 5 days. G) R7T1 cells cultured in absence of DOX. H) R7T1 cells cultured in the presence of DOX, no primary antibody negative control.*

### 4.3.2. Design of the screening process

A genome wide RNAi screen using shRNA constructs was designed to search for genes involved in β-cell proliferation and resistance of β-cells to death stimulated by chronically high levels of lipids.

A schematic representation of the screening process is depicted in Figure 4.1. The in-house amplified NKI library was transfected into Phoenix cells, which packaged the shRNA constructs into retroviral particles as described in section 4.2.4. These retroviral particles were then used to transduce R7T1 cells.

A schematic timeline of the different steps in the screen is depicted in Figure 4.7. Briefly, 72 hours after transduction cells which had stably integrated the viral construct were selected with puromycin. Puromycin resistant cells were cultured in absence of DOX for seven days before starting the experimental conditions. A sample was collected before any further experiments. This sample constituted the proliferation control and was named T0. Two screening processes took place simultaneously: a proliferation screen, which consisted on culturing the cells in absence of DOX for 7 or 14 days; and a resistance to death screen, which consisted in culturing the cells for 96 hours in absence of DOX and in presence of either a solution of palmitate in ethanol (death sample) or just ethanol (vehicle-only sample). Cells were collected at the end of their respective experiments and genomic DNA was extracted. shRNA targeting sequences were then amplified and sequenced.

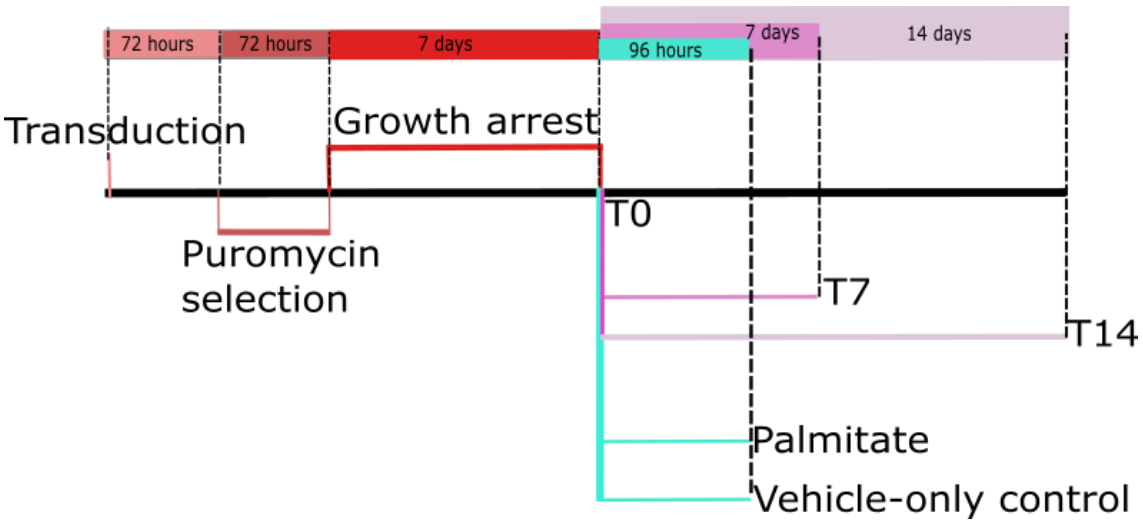



*Figure 4.7. Schematic timeline of the screening process.*
*In tones of red, steps to prepare the cells. In blue, the resistance to death screen. In tones of purple, the proliferation screen. T0, T7, T14, Palmitate and Vehicle-only control are the four samples produced by the screen and indicate the points in the timeline when those samples are collected.*

### 4.3.3. Optimization of the delivery of shRNA constructs

Initial attempts at delivering the library using the protocols already optimized for the screening process described in Chapter 3 showed a much lower success rate, with no more than 1% of the cells surviving puromycin selection. To assess whether this was problem with the transduction efficiency or with the puromycin used, $3T3^{supp+}$ fibroblasts were transduced with a GFP expressing retroviral construct as a positive control. Transduction levels of approximately 30% were detected when using the transduction protocol as in Chapter 3. The reasons for this low transduction rate in R7T1 were investigated in order to produce an optimized workflow.

The transfection of the Phoenix cells was optimised to increase the number of retroviral particles obtained. Transfection levels were first investigated using a GFP expression plasmid. Transfections using Turbofect as in Chapter 3 showed that approximately 70% of the Phoenix packaging cells expressed the fluorescent protein. Although this transfection efficiency is not low it was still optimized to maximize the number of cells producing retroviral particles. An attempt to optimize the rate of transfection using Turbofect did not yield any higher transfection rates. However, calcium-phosphate transfection using a commercial kit (Promega) resulted in a higher transfection efficiency of around 95%.

The retroviral medium produced using the newly optimized transfection protocol was used to transduce R7T1 cells. However, the transduction levels were not changed, indicating a limiting step in the transduction. It has previously been reported that β-cells are not as readily transduced as other types of cells, and that dispersion of cells could increase the transduction rate (Dupraz, Rinsch et al. 1999). Dispersion of R7T1 was therefore tried as a method to increase transduction rates Figure 4.8). Dissociation of R7T1 did increase the transduction rate to approximately 5%, but that increase was not enough for an efficient screening process. It was then investigated whether the limiting step was uptake of the viral particles by the cells.

The virus produced by the ecotropic Phoenix cells is internalized into cells after binding a membrane receptor (Scl7a1). This step is necessary for the virus to deliver the plasmid into the transduced cells. The levels of expression of the

ecotropic viral receptor were measured using QRT-PCR. The expression level of the viral receptor was 10 times lower than in the SV40 transformed 3T3 cell line that was successfully transduced in Chapter 3. This difference could explain the lower transduction rate obtained in R7T1. Several options were then studied: to increase the concentration of viral particles, to increase the levels of the ecotropic receptor on the R7T1 cell membrane and to deliver the virus using receptor-independent methods.

Two methods were evaluated to increase the concentration of viral particles: a commercial retrovirus concentrator solution (Retro-X, Clontech) and a solution of polyethylene glycol (PEG). Both methods rely on the same principle: viral particles get included into heavy complexes during an incubation step. These viral complexes are then pelleted by centrifugation. Supernatant can then be discarded and the pellet re-suspended in a lower volume of growth medium, producing a concentrated retroviral medium.

To increase the levels of the ecotropic receptor on R7T1 cell membrane, Clontech's Ecotropic Receptor Booster was tested. This commercial reagent consists on lipid vesicles containing molecules of the receptor which are incorporated into the cell membrane and transiently raise the receptor levels. Retroviral particles were complexed with calcium phosphate and with a commercial solution (Viraductin) to attempt a receptor-independent delivery. Delivery with Viraductin can be coupled with a concentration step as the virus is complexed and pelleted.

These methods were tried using the same retroviral supernatant and R7T1 cells grown in parallel to keep these variables constant across this comparison. Representative images of cells expressing GFP are shown in fluorescent microscopy pictures in Figure 4.9. Concentration of the virus using PEG and RetroX produced transduction rates of approximately 10%, similar to the rate obtained using the Ecotropic receptor Booster to increase virus binding to the membrane. Viraductin-aided delivery of 2-fold concentrated viral supernatant produced a transduction rate of approximately 25% while the use of virus in complex with calcium phosphate did not increase the transduction rate above 1%. Thus, Viraductin-aided delivery following concentration of the viral supernatant 2-fold was used in the screening process.

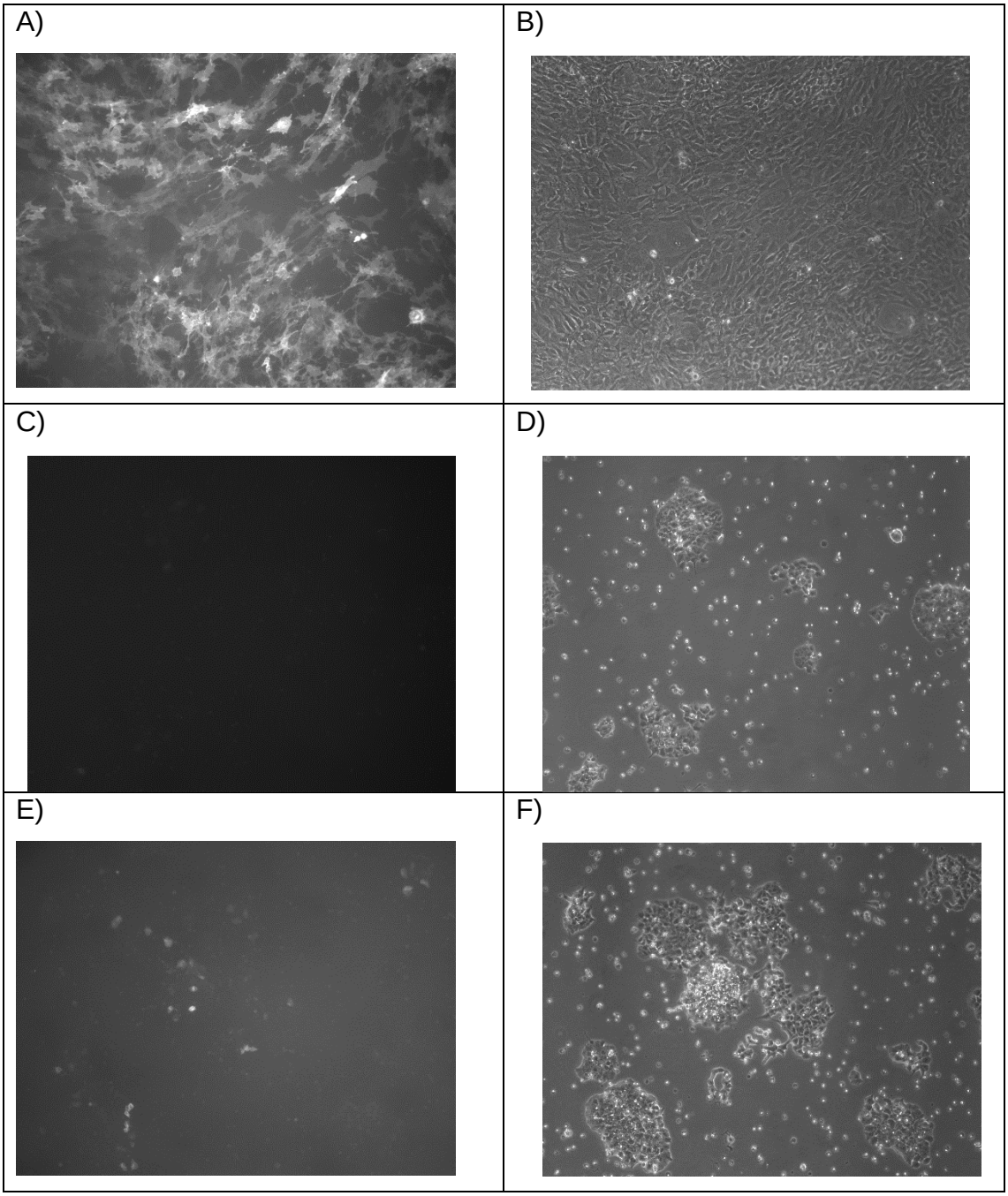



*Figure 4.8. Fluorescence (A, C, E) and visible light (B, D, F) microscopy pictures of 3T3$^{supp+}$ (A, B), R7T1 without prior dissociation (C, D) and R7T1 with dissociation prior to transduction (E, F).*

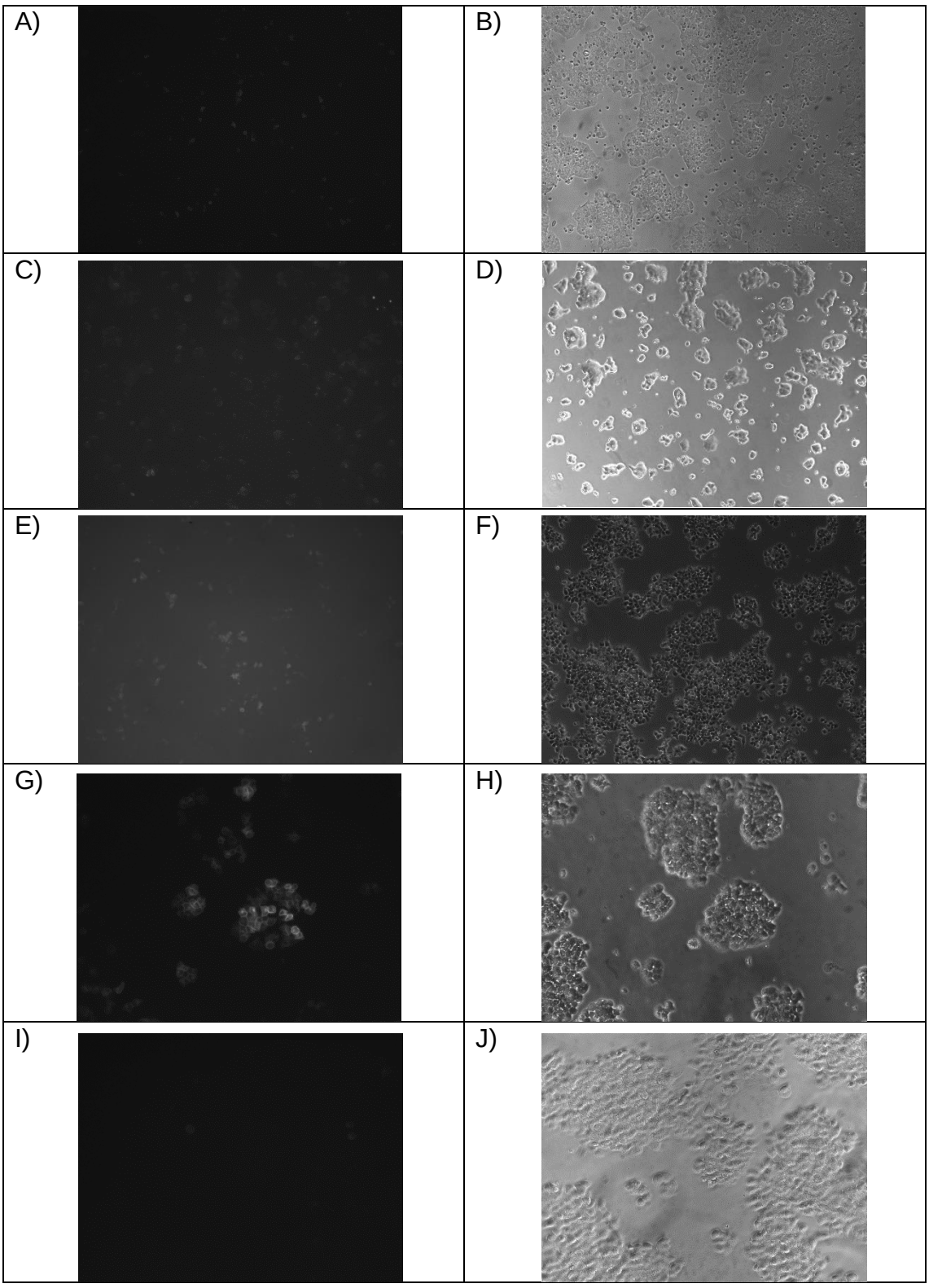



*Figure 4.9. Fluorescence (A, C, E, G, I) and visible light (B, D, F, H, J) microscopy pictures of R7T1 cells transduced with PEG concentrated virus (A, B), RetroX concentrated virus (C, D); and R7T1 transduced with virus aided by the addition of ecotropic receptor booster (E, F), Viraductin and 2-fold concentration (G, H) and virus complexed with calcium phosphate (I, J)..*

### 4.3.4. Optimization of the screening process

#### *4.3.4.1 Choice of the reversibly immortalized β-cell line used*

The aim of the screen was to find genes which had an effect on any of two phenotypes: β-cell proliferation rate and resistance of β-cells to death caused by excess of lipids in the microenvironment. The proliferation rate of β-cells *in vivo* is known to be very low (Dor, Brown et al. 2004, Meier, Butler et al. 2008) compared to the proliferation rate of immortalised, oncogene-expressing β-cell lines routinely used in the literature. Therefore, studying changes in proliferation rate using these cell lines is not likely to yield results that are easily transferred to an *in vivo* situation. Also, the expression of an oncogene used to immortalize a cell line could alter the expression of other genes and lead to the detection of changes in the phenotype that would not happen *in vivo*. To better approximate a physiologically relevant level of cell proliferation, a reversibly immortalized β-cell line was used.

#### *4.3.4.2 Optimization of the resistance to death by lipids screen*

Fatty acids are chronically elevated in type 2 diabetes (Paolisso, Tataranni et al. 1995, Charles, Eschwe`ge et al. 1997) and elevated levels of circulating free fatty acids are associate with risk of T2D (Pankow, Duncan et al. 2004). Abnormally high levels of fatty acids have been shown to be toxic to β-cells (Shimabukuro, Zhou et al. 1998, Maedler, Spinas et al. 2001, Lupi, Dotta et al. 2002, Piro, Anello et al. 2002). Although palmitate is one of the most abundant fatty acid in circulation and elevated in type 2 diabetes the mechanism by which it might cause β-cell dysfunction remains under-characterized. Therefore an RNAi screen was designed to investigate mechanisms of protection of β-cells from palmitate-induced toxicity.

Two conditions were optimized for this screening process: the concentration of palmitate that cells would be exposed to and the duration of the stimulus. As type 2 diabetes is a chronic condition, longer exposure to a lower level of stimulus was thought to be more physiologically relevant.

A palmitate dose-response curve was performed as stated in section 4.2.13 after reversal of immortalization (Figure 4.10). Palmitate treatment is commonly

performed with serum-free media. However, when cells were cultured in absence of palmitate and without serum (Figure 4.10 A), the level of cell death was still high. The difference in survival between the control condition (0 mM palmitate) and the treatment condition limits the resolution of the screening process through the setting of the maximum effect size. This maximum effect size corresponds to the highest protection level that the downregulation of a certain gene may have. A targeting sequence with a full protective effect would approximate the number of cells surviving the palmitate treatment to the number of cells surviving the control condition. Therefore, if the difference between these two numbers is small, the maximum effect size is also small. For the reason explained above, the exposure to palmitate needed to be of a chronic nature. However, long periods of culture in the control condition produced high rates of cell death. One of the reasons for this high rate of cell death might have been a toxic effect of the vehicle. Another reason might be a starvation effect by which cells in complete absence of serum were dying after 96 hours. To assess whether this might be the case, an additional dose-response curve was performed in the presence of 1% FBS for 72 and 96 hours (Figure 4.10 B). The supplementation with 1% FBS had a protective effect on the cells. The level of cell death after 96 hours in culture with 0.5 mM palmitate in the presence of 1% FBS approximated the level of cell death after 72 hours with no serum supplementation. Importantly, the level of cell death in the control condition (0 mM palmitate) was also reduced. However a lower death rate in the control condition with no serum was found in this second experiment. This may be due to a different batch of BSA being used.

The difference in cell death rate between the control and the 0.5 mM palmitate treatment conditions was highest in the presence of 1% FBS for 96 hours. These conditions (0.5 mM palmitate in 1% BSA, 1% FBS DMEM) were selected for the screening process.

#### *4.3.4.3 Optimization of cell seeding.*

In pooled RNAi screens the targeting sequences are delivered to a stock of cells that are then subdivided and seeded onto the plates or dishes and exposed to the conditions being studied. This allows a first sample of cells being collected at the time of seeding that will inform of the representation of the targeting sequences used at the time of plating. However, several aspects had

to be considered in the screen being described due to the nature of the *in vitro* model used.

Firstly, the shRNA constructs had to be delivered prior to the oncogene expression reversal as retroviral particles will only infect dividing cells. As the screen was to be performed after reversion of the oncogene expression, this added seven days of growth after transduction. During this time the representation of the targeting sequences might have changed. Hence, the control sample which is usually collected before the screening would have to be collected after the reversion time.

Secondly, cells had to be puromycin selected before growth arrest. After seven days of culturing R7T1 cells in the absence of DOX cell death was evident. This led to a lower cell density and seemed to impair puromycin resistance even in those cells that expressed the resistance gene as determined by the co-expression of a fluorescent protein. Therefore, puromycin selection was implemented after transduction and before the start of the growth arrest period.

Lastly, R7T1 cells do not plate efficiently after detachment from a culture surface. It was observed during the culture of R7T1 that when they were trypsinized and plated onto a new surface less than 30% of cells seeded and grew. This aspect of the culture was managed during routine subculture by plating an excess of cells onto the new culture surface. During the screening process however this would require the transduction of more than three times the amount of cells required for the screen which in turn increased the viral supernatant required. Implementing these changes would reduce the efficiency of the screen and therefore an alternative solution was investigated. A non-enzymatic dissociation reagent was assessed as an alternative to trypsin but the re-plating efficiency was not increased. The re-plating step was then avoided by seeded R7T1 onto the screening dishes before transduction. Cells were then puromycin selected and growth arrested in the plates that underwent the screening process, with no re-seeding steps.

It has already been explained in Chapter 3 the importance of representation in RNAi screens. A literature search indicated that 1,000 cells per sequence would provide adequate levels of representation for a drop-out screen (Sims, Mendes-Pereira et al. 2011). Library pools formed by 384 different shRNA constructs

were kept separate, so for each experiment 384,000 cells were required for each condition after puromycin selection and immortalization reversal, processes which also had an effect on cell survival. To avoid lack of representation due to stochastic drop out during the oncogene expression reversal process, it was estimated that an average of 500,000 cells should remain after puromycin selection and prior to reversal of oncogene expression.

As described in section 4.3.3, the transduction protocol that I optimised resulted in approximately a MOI of 0.3. To obtain 500,000 cells using a 0.3 MOI, 1.5 million cells needed to be exposed to the retroviral medium. A well in a 6-well multiwell plate contains approximately 1.6 million R7T1 cells at 70% confluency (determined experimentally). Therefore, one well was used for each screening condition: time 0 (T0), proliferation for seven days (T7), proliferation for fourteen days (T14), cells exposed to palmitate for 96 hours (P) and cells exposed to vehicle only for 96 hours (C).

To facilitate the treatment of cells, each condition was performed in separate 6-well plates, so that each 6-well plate would contain cells for six different library pools that underwent the same conditions during screening, one well per pool. This approach has been summarised in Figure 4.11. Phoenix cells were plated in 10 cm dishes and transfected. Five wells of R7T1 in five different plates were infected with the retroviral medium for each pool as explained in section 4.2.9. The cells were puromycin selected for 48 hours and then DOX was withdrawn from the growth medium. Cells were grown in the absence of DOX for seven days. Then, cells in the T0 6-well plates were harvested and stored at -20ºC. The rest of the cells were used in either the proliferation screen or the resistance to death by lipids screen.
For the proliferation screen, cells in two wells per pool were kept growing in growth medium without DOX, which was renewed every three days. After seven days of growth, cells in one of the wells were harvested and stored at -20ºC, constituting a T7 sample. Cells in the other well were kept growing for seven more days and then harvested and stored at -20ºC (T14 sample).

The resistance to death by lipids screen was carried out at the same time as the proliferation screen. Two wells of cells per library pool were used for this screen. One well was treated with palmitate (P sample) and another with ethanol,

vehicle-only control termed “C sample”. Wells were exposed to the appropriate treatment for 96 hours, with the treatment being refreshed every 24 hours. At the end of the treatment period cells were harvested and pelleted. Pellets were stored at -20ºC until DNA extraction.

DNA extractions were performed in parallel for all the pools and all the conditions to avoid a potential source of technical variability that could arise from extracting the DNA in different days.

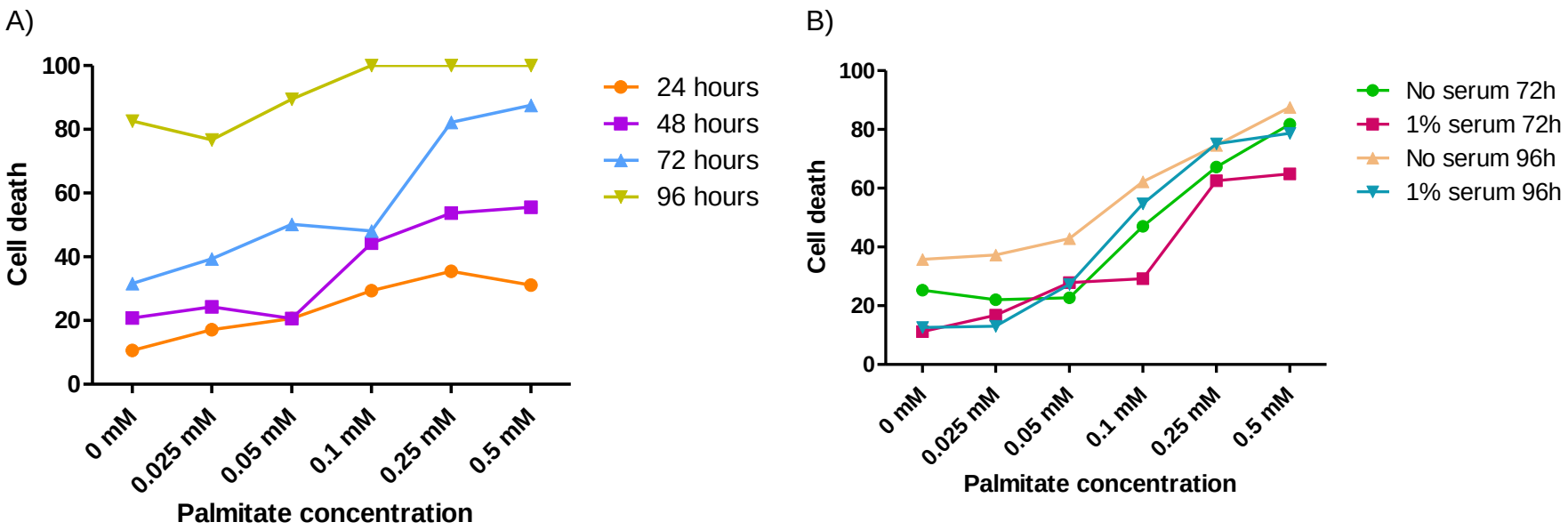



*Figure 4.10. Palmitate dose-response curves.*
*A) Dose-response curve for 4 time-points (24 hours, 48 hours, 72 hours and 96 hours) in absence of serum. B) Dose-response curve for two time-points (72 hours and 96 hours) in absence of serum and in presence of 1% FBS.*

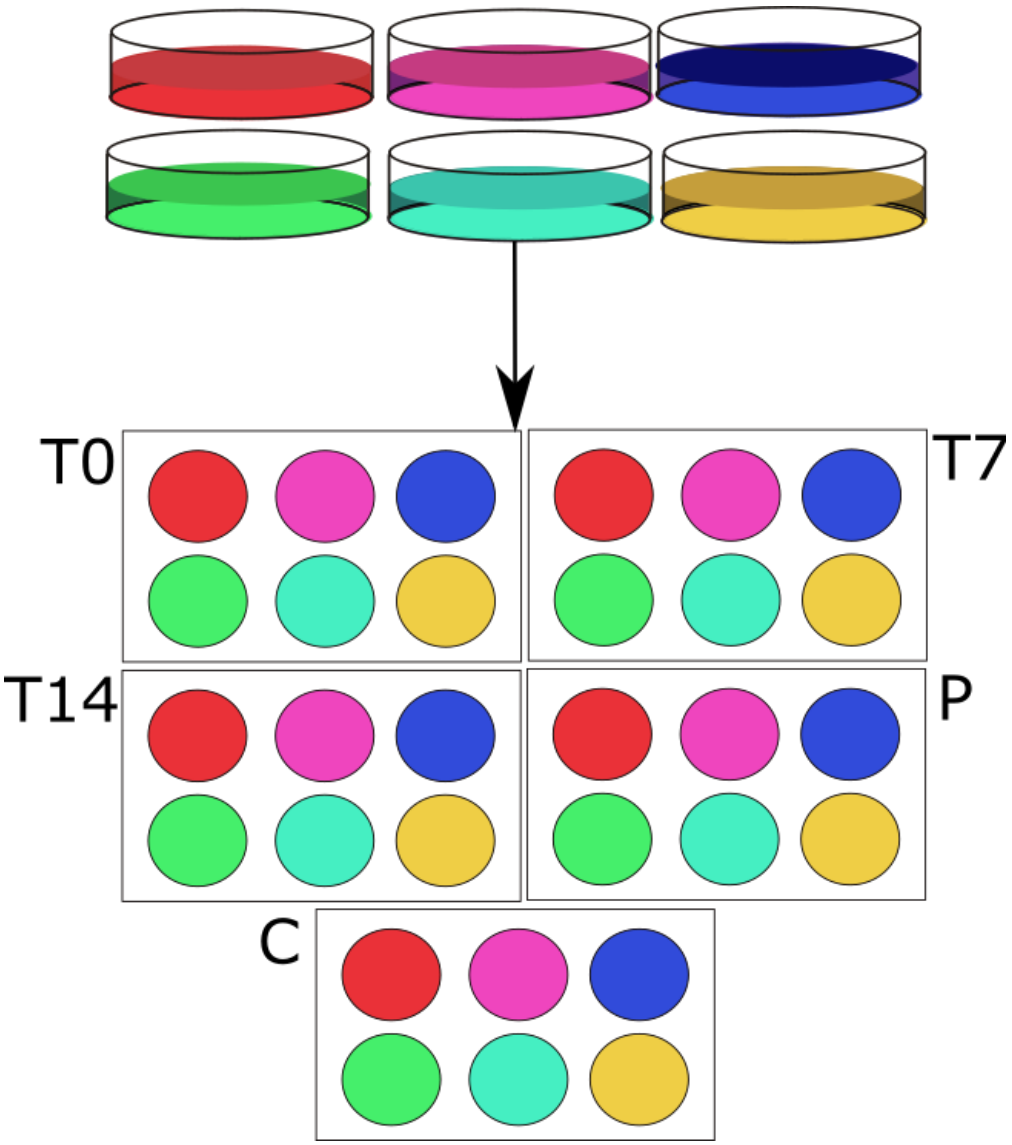



*Figure 4.11. Simplified representation of the transduction scheme.*
*Each colour represents a different library pool. Retroviral medium for each of the six pools represented was used to transduce five wells in five different plates, one per screening condition. Each 6-well plate contained cells transduced for six different pools.*

### 4.3.5. Optimization of shRNA construct recovery from R7T1 cells

This screening process used sequencing and counting of the shRNA targeting sequences as a proxy for the number of cells containing each shRNA construct as each cell contained a single shRNA targeting sequence inserted into the genome. Cells containing a targeting sequence with a positive effect on proliferation will divide more rapidly than cells containing targeting sequences with no effect on this phenotype and hence increase the number of a particular shRNA targeting sequence. As the proliferation rate of R7T1 after growth arrest is very low, only increases in proliferation rate were considered. Similarly if the effect of a targeting sequence resulted in the protection from cell death caused by palmitate, cells containing that sequence would be more abundant in the palmitate sample than cells containing other sequences with no effect on protection.

Targeting sequences therefore had to be extracted from the cells, amplified in a linear manner and sequenced to quantify their representation in the screening samples.

Cells were stored as pellets at -20ºC after the screen. Cell pellets were thawed and genomic DNA was extracted as explained in section 2.2.2 for all samples simultaneously to reduce variation (Sims, Mendes-Pereira et al. 2011). DNA was then used as a template for PCR amplification of shRNA constructs. Genomic DNA extractions and PCR amplifications were performed in parallel for all samples again to reduce any potential variation. The expected amount of genomic DNA extracted from the screening samples is lower than in the screen in Chapter 3. This difference was also increased due to the absence of a co-culture condition in this screen. However, the entirety of the genomic DNA extracted belonged to R7T1 cells containing the shRNA targeting sequences and therefore one copy of the targeting sequence was expected for each copy of genomic DNA. Therefore the extracted DNA was not further processed before PCR amplification as it was not necessary to enrich the DNA samples in targeting sequences in contrast with the screening process in Chapter 3.

A pilot project was run to investigate the level of reproducibility of the screen (section 4.3.6). The shRNA targeting sequences were amplified as optimized in

Chapter 3 and sent to be sequenced. However, analysis of the sequencing data of this pilot project together with observation of agarose gel pictures revealed that the PCR amplification yielded multiple bands, possibly due to the formation of secondary structures, which could affect the sequencing output. Figure 4.12 depicts an agarose gel picture showing both a band for the expected shRNA PCR product and a slower migrating band, probably the result of a polymerization and tertiary structure of the amplification products.

Inspection of the sequencing data revealed that a high amount of adaptor dimers was seen in the sequencing output, pointing to a possible decrease in ligation efficiency (Figure 4.17). The presence of these adaptor primers limited the number of sequencing reads obtained from each HiSeq lane. To avoid any decrease in sequencing efficiency that these abnormal configurations could cause a two-step PCR amplification strategy was adopted that eliminated any adaptor ligation steps (section 4.2.14). A first round of amplification was performed with sample-specific primers that incorporated a short sequence complementary to a second primer pair. This second pair of primers was used in a second round of amplification and introduced the Illumina barcodes and primer binding sites necessary for the sequencing process.

PCR amplification has the potential to alter sequence representation as has been discussed in Chapter 3. To reduce any bias introduced by the PCR reactions, amplification cycles were minimized using the optimization strategy described in section 3.3.5.

The shRNA secondary structure of different DNA templates can affect PCR amplification. This issue has been discussed (Strezoska, Licon et al. 2012) and solved by the supplementation of the PCR reaction with a relaxing agent such as DMSO or betaine. Addition of betaine to the PCR amplification reactions resulted in a reduction of the slower migrating bands in the electrophoresis gels. Betaine was therefore used in the two rounds of PCR reactions to relax any secondary structures and facilitate the amplification. PCR products thus produced were quantified using a Qubit fluorometer and mixed in equal proportions. They were then purified (section 4.2.15) and quantified again using Qubit and Tapestation. Amplified DNA samples were then submitted for sequencing to the University of Exeter Sequencing Service.

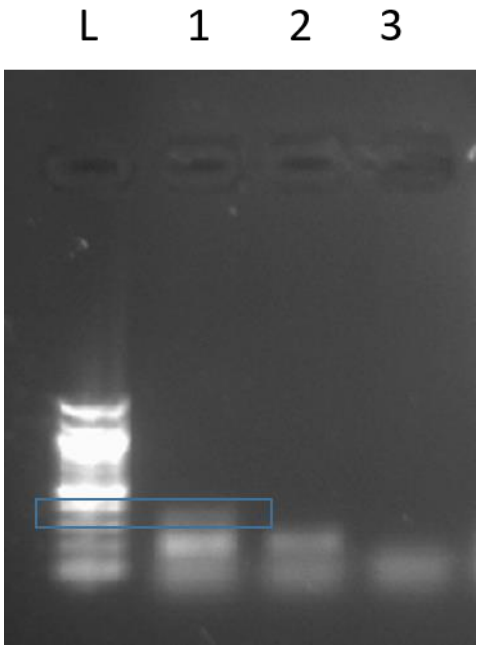



*Figure 4.12. Results of a PCR amplification of 3 samples from the pilot project.*
*L: 100 bp NEB DNA ladder. 1: sample with an extra band (marked by blue rectangle), and a band of the expected size. 2: sample with a band of the expected size. 3: failed amplification.*

### 4.3.6. Reproducibility of the screening process

A pilot project (section 4.3.6) was performed to determine the reproducibility of the screen design before the final screening process. Although a reproducibility pilot had already been performed for the screening process in Chapter 3, it was reasoned that the screening processes were different enough both biologically and technically that the reproducibility needed to be assessed in this case as well.

Biologically there are many differences that could alter the reproducibility when compared to the screening process in Chapter 3. Firstly, R7T1 cells undergo a period of oncogene expression reversal that results in a high rate of cell death. If this cell death is random and not affected by any targeting sequence it would be expected that the number of cells containing each sequence would be reduced in a similar way across all wells of any library pool as cells are transduced with the same stock virus. However, if the virus containing the targeting sequences was not distributed equally amongst the wells some of the targeting sequences could be represented more often in some wells than others. This can happen as a result of a sampling effect due to the low representation of certain sequences in the library, which leads to their under-representation in the virus and therefore in the cells. Cell death could then result in the elimination of those under-represented sequences from some of the wells.

It was also considered that the proliferation rate of R7T1 cells is very low when the oncogene expression has been reduced. Where transformed cells in the screening process, described in Chapter 3, will proliferate quickly and therefore the screen samples will contain abundant cells with each targeting sequence, the number of R7T1 cells would not increase as much. This could result in a sampling effect if the PCR amplifications were not performed with enough template DNA.

To assess the performance of the experimental design and the biological and technical variation introduced before the sequencing, a pilot project was run. Two pools were run in duplicate in parallel. Transfections, transductions and all tissue culture steps were performed at the same time in parallel independent cell cultures. DNA was extracted, amplified, purified and mixed separately. PCR amplification of the T14 sample in replicate A was not successful and as the

purpose of this pilot project was to compare both replicates the T14 sample in replicate B was not sequenced.

Replicates were sequenced in the same Illumina HiSeq 2500 lane but with different Illumina adaptors. The sequencing output was received as two separate .fastq files, each one containing all conditions for one of the replicates as the sequencing service separated them using the Illumina adaptor sequences. The sequencing run produced 2,738,774 reads for replicate A and 2,722,427 for replicate B. Of these, 172,242 reads were identified by FastQC as adaptor primers in sample A and 311,160 in sample B (Figure 4.13). In the latter, FastQC also detected two 50bp overrepresented sequences that correspond to the vector sequence. These sequences were removed from all subsequent analysis and quantification.

The “grep” command in UNIX was used to extract the reads containing the correct sequencing primer sequencing for each of the conditions screened. This process has been explained in Chapter 3 (section 3.2.8). Table 4.7 contains the numbers of sequencing reads containing an exact match for the sequencing primer used to identify each screening condition. As the pilot project was composed of two library pools of 384 sequences each, 768 targeting sequences were expected. If all of the sequences were present in the sequencing data, an average of 523 reads would be obtained for each targeting sequence. This is one half of the target depth and therefore a higher number of sequencing reads was requested for the sequencing on the final screen.

Once the reads containing the primers had been de-multiplexed into a file for the condition that they pertained to, the targeting sequences were extracted. This was performed as described for the screen in Chapter 3. Custom Python scripts (described in section 3.2.8 and in full in Appendix 3) were used to extract the shRNA targeting sequences and the number of sequencing reads which contained each targeting sequence to text files. Briefly, the script identifies the sequence of the shRNA loop when it is present in a sequencing read. Then, it records the 19bp sequence just before the start of the loop sequence as a “potential shRNA targeting sequence”. Once all the potential sequences have been identified, the script records the occurrence of each individual sequence. The potential shRNA targeting sequences identified and their occurrence are written to a text file. The number of potential shRNA targeting sequences found

at least once in each condition and replicate are summarized in Table 4.8. These files were imported into Stata and aligned to the expected sequences according to the theoretical composition of the pools screened. Table 4.9 shows the number of constructs that were present in the output for each condition in both replicates. The degree of overlap of the expected targeting sequences amongst the four conditions (T0, T7, P and C) was investigated for each replicate. It was expected that most sequences would be common for all conditions and that the controls might have some sequences that were not identified in the experimental conditions due to the cell death experienced during the treatment with palmitate and the additional week of culture in absence of DOX between T0 and T7. Venn diagrams representing these overlaps were drawn for replicate A (Figure 4.14) and replicate B (Figure 4.15) and confirm that this was the case.

The number of reads in replicate B was lower than in replicate A and in turn the number of potential shRNA targeting sequences and the number of expected targeting sequences detected in replicate B were also lower. The palmitate sample in replicate A has the highest number of potential targeting sequences identified but not the highest number of reads or expected targeting sequences. The palmitate sample is exposed to a toxic stimulus that greatly reduces the number of cells. This results in less template DNA which should lead to less targeting sequences being detected. Upon investigation of the potential targeting sequences it was evident that a large proportion of the potential targeting sequences (6,188 sequences) had only one read in the sequencing data. This could indicate that those sequences might be derived from other sequences by a polymerase-introduced error at either the PCR amplification or the sequencing processes.

Equivalently to the approach implemented for the analysis of the data in Chapter 3, a threshold was established for the number of reads that a targeting sequence had to have been assigned for it to be considered present in the dataset. This threshold was used only in the control samples (T0 and C). Targeting sequences with lower amounts of reads were set to “missing” in the Stata dataset so that they were not used in any further calculations.

Variability between replicates was assessed using correlations. Table 4.10 shows the correlation matrix for the raw reads amongst all conditions and

replicates. These correlations are all above 0.500 indicating an acceptable degree of correlation between replicates and between conditions in each replicate.

Raw reads were converted to reads per million per pool (rpmp) as explained in section 3.2.8. A correlation matrix for rpmp is shown in Table 4.11. All correlations coefficients are higher than 0.600 again indicating a high degree of correlation.

For each experimental condition (T7, T14 and Palmitate) a ratio of experimental condition to control was calculated. For the proliferation experiments (T7 and T14) T0 was used as a control. For Palmitate, vehicle only was used as the control to account for potential changes in cell representation from T0 due to proliferation during the 96 hours of exposure or to effects of the vehicle (ethanol) rather than treatment (palmitate) on cells. A correlation matrix for these ratios is available in Table 4.12.

The logarithm of these ratios (correlation matrix in Table 4.13) was used to build a robust z-score as explained in section 4.2.16. The correlation matrix for this metric is shown Table 4.14.

The degree of correlation of the replicates decreases with the transformation steps performed as was observed in Chapter 3. The palmitate sample in replicate B has a consistently low correlation with the rest of the samples in both replicates. It was expected that the palmitate samples would have the lowest correlation with other samples as the treatment with palmitate would reduce the representation of the shRNA targeting sequences, potentially causing a sampling effect in the PCR amplification and sequencing effects. The highest correlation coefficient after data transformation and calculation of the z-scores corresponds to the T7 samples between both replicates. This indicates that the reproducibility of the proliferation screen is high even taking into account that the sequencing depth was lower than expected. The low reproducibility of the death by lipids screen highlighted the need for a higher representation of the shRNA targeting sequences and an improvement in the PCR amplification and sequencing steps. Both optimizations were included in the final library-wide screen.

A)

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGC | 172242 | 6.289018370993736 | TruSeq Adapter, Index 3 (100% over 50bp) |

B)

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC | 302885 | 11.125550841216311 | TruSeq Adapter, Index 5 (100% over 50bp) |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATATCGTATGC | 8275 | 0.30395672684703756 | TruSeq Adapter, Index 5 (98% over 50bp) |
| TGTCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTC | 3950 | 0.1450911264103684 | No Hit |
| TCTTTGGATTTGGGAATCTTATAAGTTCTGTATGAGACCACAGATCTCCC | 2892 | 0.10622874369083175 | No Hit |

*Figure 4.13. Overrepresented sequences in the pilot project as identified by FastQC in the sequencing run for A) replicate A and B) replicate B.*

| Screening condition | Reads in replicate A | Reads in replicate B |
|---|---|---|
| **T0** | 401,685 | 313,819 |
| **T7** | 266,097 | 267,653 |
| **P** | 262,059 | 207,584 |
| **C** | 183,300 | 220,750 |

*Table 4.7. Number of sequencing reads containing an exact match for the sequencing primers in each replicate of the pilot project for each screening condition*

| Screening condition | Replicate A | Replicate B |
|---|---|---|
| **T0** | 9,984 | 1,819 |
| **T7** | 6,441 | 1,013 |
| **P** | 12,734 | 1,137 |
| **C** | 7,787 | 1,454 |

*Table 4.8. Number of different potential shRNA targeting sequences (19 bp preceding the loop sequence) found in the sequencing data for the four different conditions in the two replicates.*

| Screening condition | Replicate A | Replicate B |
|---|---|---|
| **T0** | 602 (78.39%) | 560 (72.92%) |
| **T7** | 557 (72.53%) | 406 (52.86%) |
| **P** | 565 (73.57%) | 411 (53.52%) |
| **C** | 451 (58.72%) | 472 (61.46%) |

*Table 4.9. Number of constructs designed to be in the library that had at least one read in each condition of the two replicates of the pilot project.*
*In brackets, percentage of the expected constructs identified in the sequencing data.*

| | T0A | T7A | PA | CA | T0B | T7B | PB | CB |
|---|---|---|---|---|---|---|---|---|
| **T0A** | 1 | | | | | | | |
| **T7A** | 0.710 | 1 | | | | | | |
| **PA** | 0.629 | 0.800 | 1 | | | | | |
| **CA** | 0.787 | 0.673 | 0.589 | 1 | | | | |
| **T0B** | 0.848 | 0.848 | 0.750 | 0.769 | 1 | | | |
| **T7B** | 0.640 | 0.862 | 0.777 | 0.573 | 0.867 | 1 | | |
| **PB** | 0.635 | 0.635 | 0.757 | 0.572 | 0.775 | 0.843 | 1 | |
| **CB** | 0.649 | 0.721 | 0.665 | 0.639 | 0.801 | 0.680 | 0.680 | 1 |

*Table 4.10. Correlation coefficients between all pairs of samples from the pilot project*

| | T0A | T7A | PA | CA | T0B | T7B | PB | CB |
|---|---|---|---|---|---|---|---|---|
| **T0A** | 1 | | | | | | | |
| **T7A** | 0.797 | 1 | | | | | | |
| **PA** | 0.697 | 0.831 | 1 | | | | | |
| **CA** | 0.803 | 0.807 | 0.690 | 1 | | | | |
| **T0B** | 0.874 | 0.887 | 0.803 | 0.793 | 1 | | | |
| **T7B** | 0.751 | 0.895 | 0.825 | 0.744 | 0.933 | 1 | | |
| **PB** | 0.775 | 0.877 | 0.810 | 0.768 | 0.886 | 0.880 | 1 | |
| **CB** | 0.703 | 0.785 | 0.734 | 0.689 | 0.830 | 0.831 | 0.779 | 1 |

*Table 4.11. Correlation matrix for reads per million per pool (rpmp).*

| | T7A | T7B | PA | PB |
|---|---|---|---|---|
| **T7A** | 1 | | | |
| **T7B** | 0.467 | 1 | | |
| **PA** | 0.319 | 0.164 | 1 | |
| **PB** | 0.037 | 0.006 | 0.010 | 1 |

*Table 4.12. Correlation matrix of the ratios of (experimental condition) to (control condition)*

| | T7A | T7B | PA | PB |
|---|---|---|---|---|
| **T7A** | 1 | | | |
| **T7B** | 0.458 | 1 | | |
| **PA** | 0.385 | 0.219 | 1 | |
| **PB** | 0.047 | 0.099 | 0.072 | 1 |

*Table 4.13. Correlation matrix of the ln(ratios)*

| | T7A | T7B | PA | PB |
|---|---|---|---|---|
| **T7A** | 1 | | | |
| **T7B** | 0.467 | 1 | | |
| **PA** | 0.381 | 0.221 | 1 | |
| **PB** | 0.053 | 0.100 | 0.084 | 1 |

*Table 4.14. Correlation matrix of the robust z-scores.*

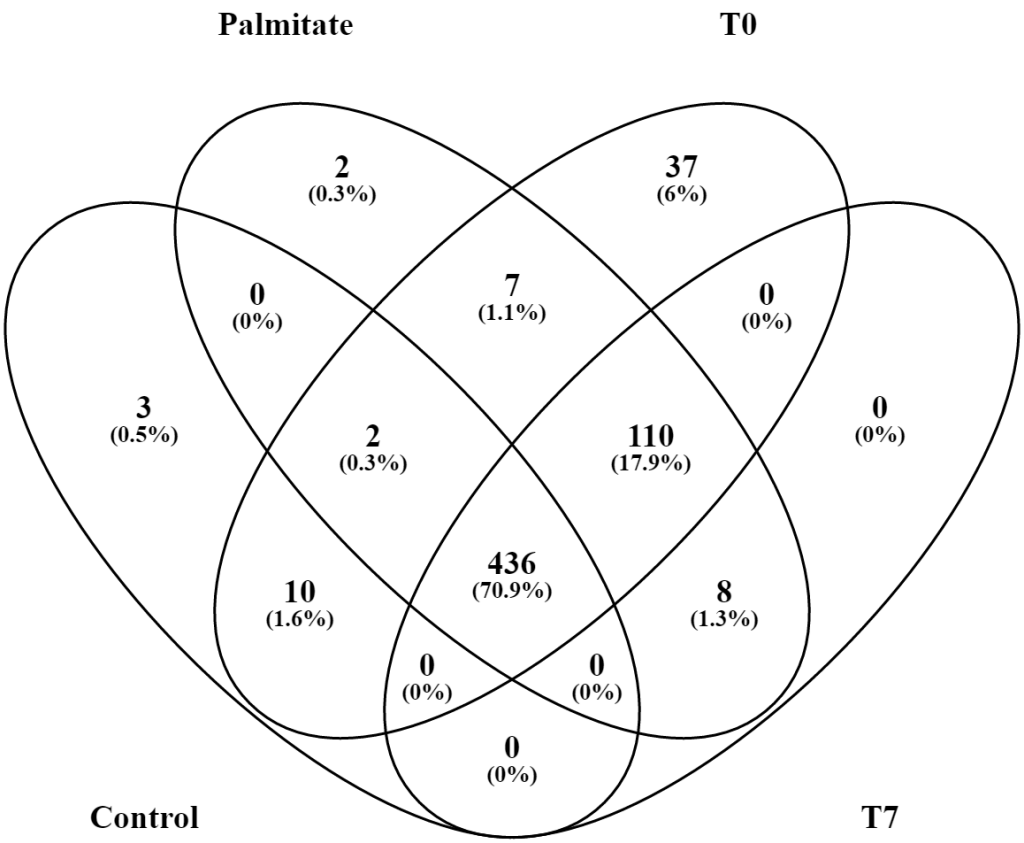



*Figure 4.14. Venn diagram of the overlap in identified targeting sequences expected from the library composition of the two pools included in the pilot.*
*Only replicate A is represented.*

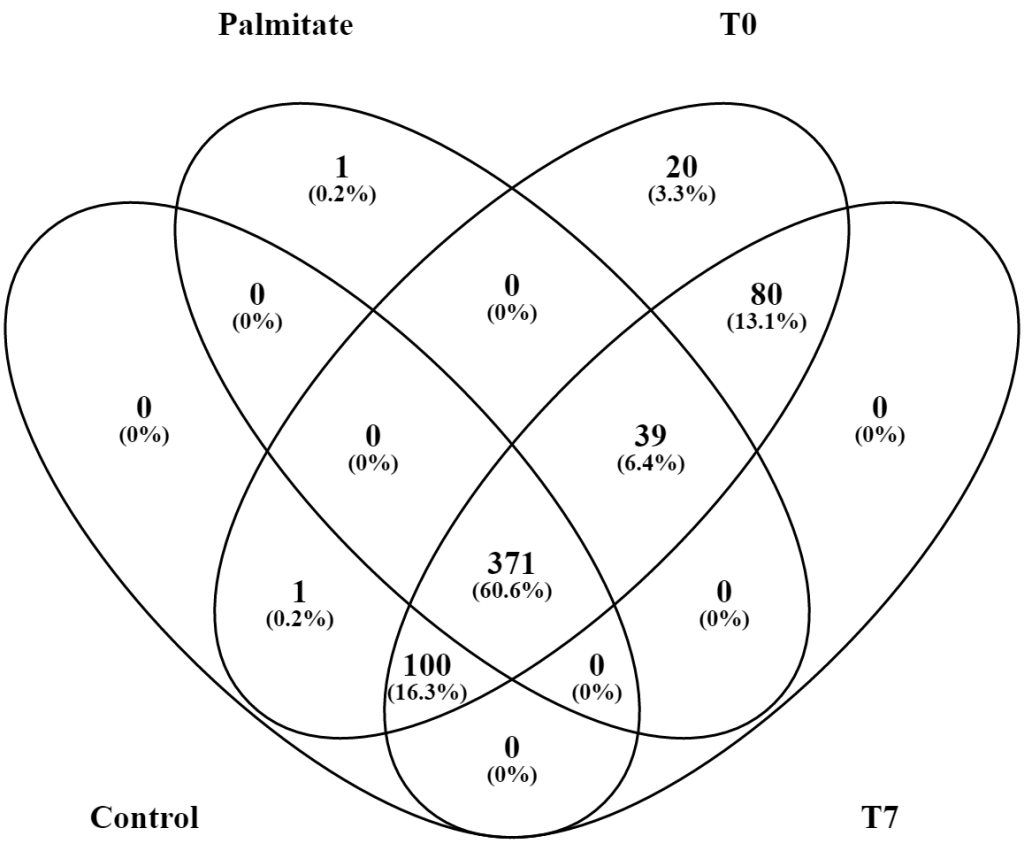



*Figure 4.15. Venn diagram of the overlap in identified targeting sequences expected from the library composition of the two pools included in the pilot.*
*Only replicate B is represented.*

### 4.3.7. Running a genome-wide shRNA screening process

After checking that the variability in the pilot project was acceptable, the entire shRNA library was used on a screening process carried out as explained in 4.3.2.

The reproducibility pilot project highlighted some potential sources of variability that could be addressed to increase the quality of the screening process. R7T1 cells were plated at a higher density before being transduced to increase the representation of the shRNA targeting sequences.

To improve the sequencing data, the sequencing libraries for the different samples were prepared in-house using a PCR based preparation method. This removes the ligation step on the library preparation protocol and should limit the presence of Illumina adaptor dimers in the sequencing data. Primers for PCR amplification were therefore designed with a region that would anneal with the primers used for the sequencing library preparation. A first round of amplification was performed with these primers. Each condition was amplified with a sample specific primer so that the samples could be de-multiplexed after sequencing. This first PCR amplification was performed in parallel for each of the pools. Then, PCR products were purified and quantified using a Qubit. Equal amounts of each pool were mixed for each condition.

The mixtures were then used as templates for the second round of amplification using the primers containing the Illumina adaptor sequences. This second round of amplification introduced a different combination of Illumina indexes for each sample so an additional level of de-multiplexing was possible. This removed the possible uncertainty in assigning sequencing reads to the correct sample that was present in the pilot project. The new de-multiplexing strategy will be explained in section 4.3.8.

Products of the second round of PCR amplification were purified using the Agencourt AMPure beads and quantified using a Qubit and a Tapestation. The results obtained using the Tapestation are depicted in Figure 4.16. A peak of the expected size after the two rounds of amplification (approximately 270 bp) is visible in both the digital electrophoresis separation gel and the peak profile. Three other peaks are visible that might be derived from the first and/or second rounds of amplification. A small peak of approximately 150 bp may correspond

to the products of the first round of PCR amplification. These products would not contain the Illumina adaptor sequence and subsequently would not be sequenced. Two heavier peaks approximately 660 and 1,200 bp may be the result of secondary structures formed by more than one shRNA construct. Such structures have been described in the literature (Du, Ge et al. 2006). As it was unknown whether the presence of that peak would decrease the number of reads obtained, the sample was sent to the University of Exeter Sequencing Service and three times the number of reads estimated to produce good quality quantitative data was requested.

Not all pools completed the screening process in all conditions, due to either contamination of the cultures during the screening or to complete cell death in some of the pools. Out of the 73 pools the library is divided into, 67 pools completed the screening process in all conditions. Two pools did not complete the screening in T14 and four were not screened in the death by lipids screen.

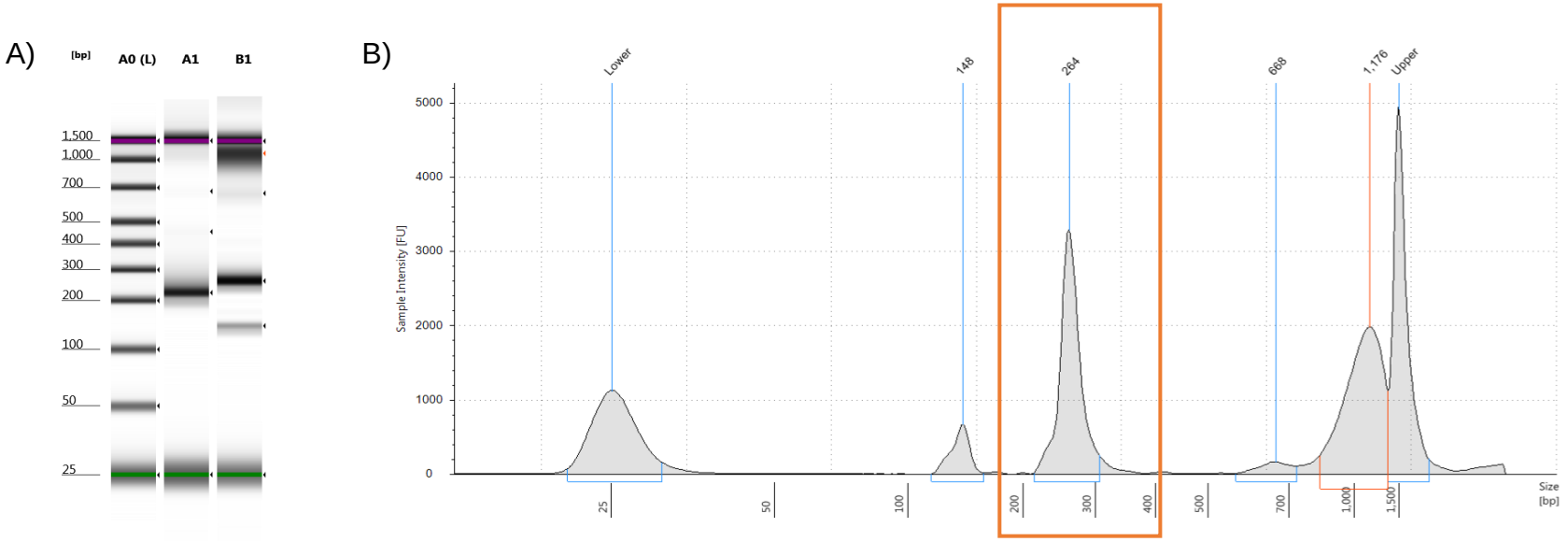



*Figure 4.16. Tapestation quantification results.*
A) *Digital gel. Lane A0(L): ladder. A1: sample not analysed in this thesis. B1: β cell PCR amplified sample. The expected band for the amplification (approximately 270 bp) is highlighted in orange.* B) *Peak profile for the β cell sample. "Lower" and "Upper" are size markers introduced by the software. The expected peak is highlighted in orange.*

### 4.3.8. Sequencing output of the screening process

The introduction of sample specific combinations of Illumina indexes resulted in the sequencing output being different from the output received for the screen in Chapter 3. De-multiplexing of the samples using the Illumina indexes was performed by the University of Exeter Sequencing Server. As a result, one .fastq file was received per condition. Reads in each file corresponded to the sample for which the Illumina indexes were used and therefore the sample-specific primers used for the first round of PCR amplification were not used for de-multiplexing as they had been used in the pilot project and in Chapter 3. The number of sequencing reads obtained for each screening condition is shown in Table 4.15. The vehicle-only control for the death by lipids screen, named C, had the lowest number of reads. If every sequence in the 62 pools that completed the screen had been sequenced the average representation of the sequences in this condition would have been 395 reads.

Quality control analysis performed by the Sequencing Service using the software package FastQC indicated that no Illumina indexes were overrepresented in the samples as expected. However, overrepresented sequences were detected in all the samples. The FastQC output for this diagnostic test for all samples is available in Figure 4.17. The frequency of these overrepresented sequences is lower than the frequency of the Illumina indexes detected in the pilot sequencing data and the data from the screen in Chapter 3. All the overrepresented sequences start with the sequence of the primer used in the first round of PCR amplification indicating they are not external contamination from other samples with the same Illumina index. These sequences could be overrepresented due to their higher representation in the samples themselves or they could be a result of contamination of the PCR reactions. A no-template control was amplified in parallel with the samples and no bands were visible in the agarose gel used to purify the amplified samples suggesting that the reactions were not contaminated. The overrepresented sequences are common to all samples which indicates that if they are due to contamination this contamination was present in a similar degree in all samples, and if they are just highly represented sequences they are so in all samples. In both cases the data analysis pipeline is expected to correct the overrepresentation as the experimental samples are normalized to their

corresponding controls where the sequences are equally overrepresented. Therefore I proceeded with the analysis process and noted these sequences.

As the sequencing reads were already separated into sample specific .fastq files, the sequence of the shRNA loop that separates both reverse-complement targeting sequences was used to extract the potential shRNA targeting sequences in each sample. The commands used for this are indicated in Appendix 8. Briefly, the “grep” command was used to identify sequencing reads containing an exact match to the shRNA construct loop. Reads where the loop was identified were placed in a file containing reads with potential shRNA constructs. The numbers of reads containing the shRNA loop for each condition are shown in Table 4.16.

The shRNA targeting sequences and the number of sequencing reads for each targeting sequence identified in each of the screening conditions were extracted using a custom Python script as explained in section 3.2.8 (script in Appendix 3). The files were imported into Stata and processed as discussed in section 3.2.8.

Table 4.17 shows the number of shRNA sequences and the proportion of those designed to be in the library detected in each condition. Most of the sequences expected to be in the library are identified in at least one sequencing read for all the conditions. However, sequences with very low representation can lead to artefacts in the analysis and therefore a threshold was used to exclude targeting sequences with low representation in the control conditions as in (Sims, Mendes-Pereira et al. 2011). Similarly to the approach taken in Chapter 3, targeting sequences with less than 50 sequencing reads in the control conditions (T0 and C) were coded as “missing” data in Stata. Ratios and therefore z-scores cannot be calculated with “missing” data so those targeting sequences were in effect eliminated from any further analysis. All targeting sequences that were expected to be a part of one of the library pools that did not complete the screening process were also discarded as they were predicted to be residual contamination. These sequences had very low numbers of sequencing reads. The numbers of the targeting sequences that were screened, hence, were identified to have more than 50 sequencing reads in the T0 and C samples, are shown in Table 4.18. The percentage of the library screened is still over 70% for T0 and T7, for which more library pools were screened. As less

pools completed the death by lipids screen it is not unexpected to have identified less targeting sequences in the P and C samples.

The different measures of targeting sequence abundance (raw reads, rpmp, ratios of rpmp and z-scores) were expected to be correlated specially amongst the proliferation screening samples.

Table 4.19 shows the correlation matrix for raw number of reads. Correlation coefficients are notably low for any pair of conditions that include T14 indicating a possible bias in T14. These correlation coefficients are however much higher when reads per million per pool (rpmp) are used instead of raw sequencing reads (Table 4.20). Correlation coefficients are much lower when the ratios of the experimental conditions to their controls are used (Table 4.21). This is not unexpected for the correlations between the proliferation conditions and the death condition but it is for the correlation between the two proliferation conditions. This correlation increases when the variation in each pool for each condition is taken into account by the calculation of a robust z-score (Table 4.22). It has been previously discussed that the additional seven days of culture of R7T1 cells in absence of DOX between T7 and T14 introduces a high level of cell death. The low correlation between T14 and other conditions can be a consequence of this cell death. If the increase in proliferation rate due to the effect of any particular targeting sequence is not strong enough to overcome the high rate of cell death, the cells containing that targeting sequence will not be present when the T14 sample is obtained. This results in the equivalent to a sampling effect in T14, where the representation of targeting sequences is highly variable and different from T7.

| Condition | Number of reads |
|---|---|
| *T0* | 14,638,600 |
| *T7* | 16,457,499 |
| *T14* | 15,629,863 |
| *P* | 12,926,085 |
| *C* | 9,404,944 |

*Table 4.15. Number of sequencing reads obtained for each condition in the sequencing data for the library-wide screen.*

A) Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| ATCTCTCGAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCGCT | 164198 | 1.1216783025699177 | No Hit |
| ATCTCTCGAGGTCGACGGTATCGATCTGTGGTCTCATACAGAACTTATAA | 94305 | 0.6442214419411693 | No Hit |
| ATCTCTCGAGGTCGACGGTATCGATTCCCAAATCCAAAGACACTGTCTCT | 30583 | 0.2089202519366606 | No Hit |
| ATCTCTCGAGGTCGACGGTATCGATAAGCTTTTCCAAAAAGCGCGCTTTG | 21075 | 0.14396868553003703 | No Hit |
| CTCTCTCGAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCGCT | 19252 | 0.13151530884100937 | No Hit |

B) Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| CGATCTCTCGAGGTCGACGGTATCGATCTGTGGTCTCATACAGAACTTAT | 91362 | 0.5551390281111365 | No Hit |
| CGATCTCTCGAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCG | 32025 | 0.1945921430710705 | No Hit |
| CGATCTCCGAGCCCACGAGACAAGAGGCAATCTCGTATGCCGTCTTCTGC | 20864 | 0.12677503428680142 | RNA PCR Primer, Index 33 (96% over 26bp) |
| CGATCTCTCGAGGTCGACGGTATCGATTCCCAAATCCAAAGACACTGTCT | 20547 | 0.12484886069262406 | No Hit |

C) Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| TCGAGGTCGACGGTATCGATCTGTGGTCTCATACAGAACTTATAAGATTC | 231501 | 1.4811454201485963 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCGCTTAGAT | 104587 | 0.6691485395617351 | No Hit |
| TCGAGGTCGACGGTATCGATTCCCAAATCCAAAGACACTGTCTCTTATAC | 77845 | 0.4980529899718251 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTTAAAAACCCAACTGGTCATCATAGA | 35856 | 0.22940700120020244 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTTAAAAACCTAACATTGTGGTTCATA | 33579 | 0.21483873531073175 | No Hit |

D) Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GAGGTCGACGGTATCGATTCCCAAATCCAAAGACACTGTCTCTTATACAC | 238780 | 1.8472723953153642 | No Hit |
| GAGGTCGACGGTATCGATCTGTGGTCTCATACAGAACTTATAAGATTCCC | 235763 | 1.8239319948770256 | No Hit |
| GAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCGCTTAGATCT | 35739 | 0.27648742832806683 | No Hit |
| GAGGTCGACGGTATCGATAAGCTTAAAAACCAAGAAGTTGGTATGGGTTC | 25677 | 0.19864483329639251 | No Hit |
| GAGGTCGACGGTATCGATAAGCTTAAAAACCACCTACATCAACCACAATC | 21887 | 0.16932427722701807 | No Hit |

E) Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| TCGAGGTCGACGGTATCGATTCCCAAATCCAAAGACACTGTCTCTTATAC | 234898 | 2.497601261634306 | No Hit |
| TCGAGGTCGACGGTATCGATCTGTGGTCTCATACAGAACTTATAAGATTC | 162704 | 1.7299837192013052 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTTGTGAATTCATGCGGCCGCTTAGAT | 123445 | 1.3125543331252159 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTTTTCCAAAAAGCGCGCTTTGTAGGA | 28575 | 0.3038295602823366 | No Hit |
| TCGAGGTCGACGGTATCGATAAGCTGTGGTCTCATACAGAACTTATAAGA | 14043 | 0.1493150836411147 | No Hit |
| TCGAGGTCGACGGTATCGATGATCTGTGGTCTCATACAGAACTTATAAGA | 10447 | 0.11107987458511184 | No Hit |

*Figure 4.17. FastQC report for overrepresented sequences in the library-wide screen sequencing data in the T0 (A), T7 (B), T14 (C), Palmitate (D) and Control (E) samples.*

| Sample | Number of reads containing the loop |
|---|---|
| *T0* | 13,114,383 (89.59%) |
| *T7* | 15,053,328 (91.47%) |
| *T14* | 14,168,888 (90.65%) |
| *P* | 11,533,202 (89.22%) |
| *C* | 8,142,763 (86.58%) |

*Table 4.16. Number of sequencing reads containing the sequence of the shRNA loop identified in each sample.*
*In brackets, percentage of the total sequencing reads that contain the loop.*

| Condition | Number of targeting sequences |
|---|---|
| *T0* | 25,692 (91.67%) |
| *T7* | 25,896 (92.38%) |
| *T14* | 22,766 (81.23%) |
| *P* | 23,355 (83.31%) |
| *C* | 23,990 (85.58%) |

*Table 4.17. Number of shRNA targeting sequences expected to be in the shRNA library identified with at least one sequencing read in each sample.*
*In brackets, percentage of the shRNA library identified in each sample. All 73 pools were considered for this calculation.*

| Condition | Number of targeting sequences |
|---|---|
| *T0* | 20,508 (79.71%) |
| *T7* | 20,953 (81.44%) |
| *T14* | 19,020 (73.93%) |
| *P* | 18,155 (75.05%) |
| *C* | 17,862 (73.83%) |

*Table 4.18. Number of shRNA targeting sequences identified in the sequencing data and expected to be in the library pools which completed the screening process.*
*For the T0 and C conditions only sequences with more than 50 sequencing reads were considered as identified.*

| | T0 | T7 | T14 | P | C |
|---|---|---|---|---|---|
| **T0** | 1 | | | | |
| **T7** | 0.783 | 1 | | | |
| **T14** | 0.280 | 0.164 | 1 | | |
| **P** | 0.886 | 0.791 | 0.254 | 1 | |
| **C** | 0.855 | 0.715 | 0.174 | 0.773 | 1 |

*Table 4.19. Correlation matrix of raw reads for the pairs of conditions screened.*

| | T0 | T7 | T14 | P | C |
|---|---|---|---|---|---|
| **T0** | 1 | | | | |
| **T7** | 0.988 | 1 | | | |
| **T14** | 0.705 | 0.651 | 1 | | |
| **P** | 0.824 | 0.781 | 0.786 | 1 | |
| **C** | 0.836 | 0.834 | 0.578 | 0.837 | 1 |

*Table 4.20. Correlation matrix for rpmp*

| | T7 | T14 | P |
|---|---|---|---|
| **T7** | 1 | | |
| **T14** | 0.020 | 1 | |
| **P** | -0.038 | 0.012 | 1 |

*Table 4.21. Correlation coefficients for the ratios of rpmp of each condition to its control.*

| | T7 | T14 | P |
|---|---|---|---|
| **T7** | 1 | | |
| **T14** | 0.204 | 1 | |
| **P** | -0.043 | 0.027 | 1 |

*Table 4.22. Correlation matrix for the z-scores*

### 4.3.9. Selection of genes potentially involved in β cell proliferation

Analysis of the sequencing results and prioritization of any candidate genes from the proliferation screen were performed similarly to the approaches used for the screen in Chapter 3. However, differences in the experimental screen and the biology studied resulted in the implementation of a few changes.

The proliferation screen performed in R7T1 was composed of two conditions (T7 and T14) and a control (T0) in a similar manner to the two co-cultures and plastic approach in the neighbour suppression screen. However, T7 and T14 are expected to have a similar distribution of targeting sequences except for the potential 'drop outs' resulting from the high level of cell death when cells are cultured for an extended period of time in the absence of DOX. Therefore there is no equivalent in this screening process to the two threshold system used in Chapter 3 to prioritize targeting sequences. On the contrary, targeting sequences were prioritized using the changes in their representation in the experimental conditions compared to their control samples.

Genes of interest (GOI) for β cell proliferation were defined as those with targeting sequences whose z-score was in the top 5% of the z-score distribution of the whole dataset in either T7 or T14. In the T7 dataset, 492 constructs were selected as targeting GOIs. In total 477 genes were prioritized as GOI with 15 genes being targeted by both sequences. In the T14 data 317 targeting sequences were prioritized. The T14 GOI list was composed of 313 genes, with four genes targeted by both constructs. Of the targeting sequences prioritized, 38 were selected in both T7 and T14.

To further refine the GOI list and prioritize genes that may be translatable to human β-cell biology an RNA sequencing (RNA-seq) dataset was used to identify genes that are expressed in human β-cells. The dataset used was made publicly available by Nica and collaborators (Nica, Ongen et al. 2013). It contains expression data normalized as reads per kilobase per million (RPKM), a commonly used measure of expression in RNA-seq datasets. The file contains data for eleven non-diabetic individuals, with expression data for whole islets, FACS separated preparations of β-cells and non-β-cells of the islets. Expression data from β-cells in Individual_1 was used for analysis.

Gene names in the screening output were first assigned a human homologue gene name using the approach explained in section 3.3.11.4 which identifies homologues based on sequence homology. The RNASeq data was then aligned to the output of the screens. Expression data was available for 14,461 of the 20,508 targeting sequences in the control condition of the proliferation screen. A flow diagram with the number of targeting sequences in each step of the analysis is presented in Figure 4.18. To filter for genes expressed in β-cells, a threshold was set up that corresponded to the expression level in individual 1 of *CFTR*, a gene known to be expressed in pancreas ductal epithelia but not in β-cells. Any gene expressed above that threshold (0.095815 RPKM) was considered to be expressed in human β-cells. After applying this filter, 336 constructs remained in the T7 data that corresponded to 329 GOI, with seven genes targeted by two sequences. These seven genes are *ZMYND11*, *HAX1*, *ANAPC10*, *BCR*, *ARHGEF1*, *PPIA* and *N4BP2*.

In the T14 dataset 194 targeting sequences remained after filtering for expression in human β-cells. No genes were found to be targeted by two targeting sequences and therefore the GOI list for T14 is composed of 194 genes. Of these 194 GOI, 29 are also part of the T7 GOI list. A list of genes included in both GOI lists and their z-scores in both samples of the proliferation screen is available in Table 4.23.

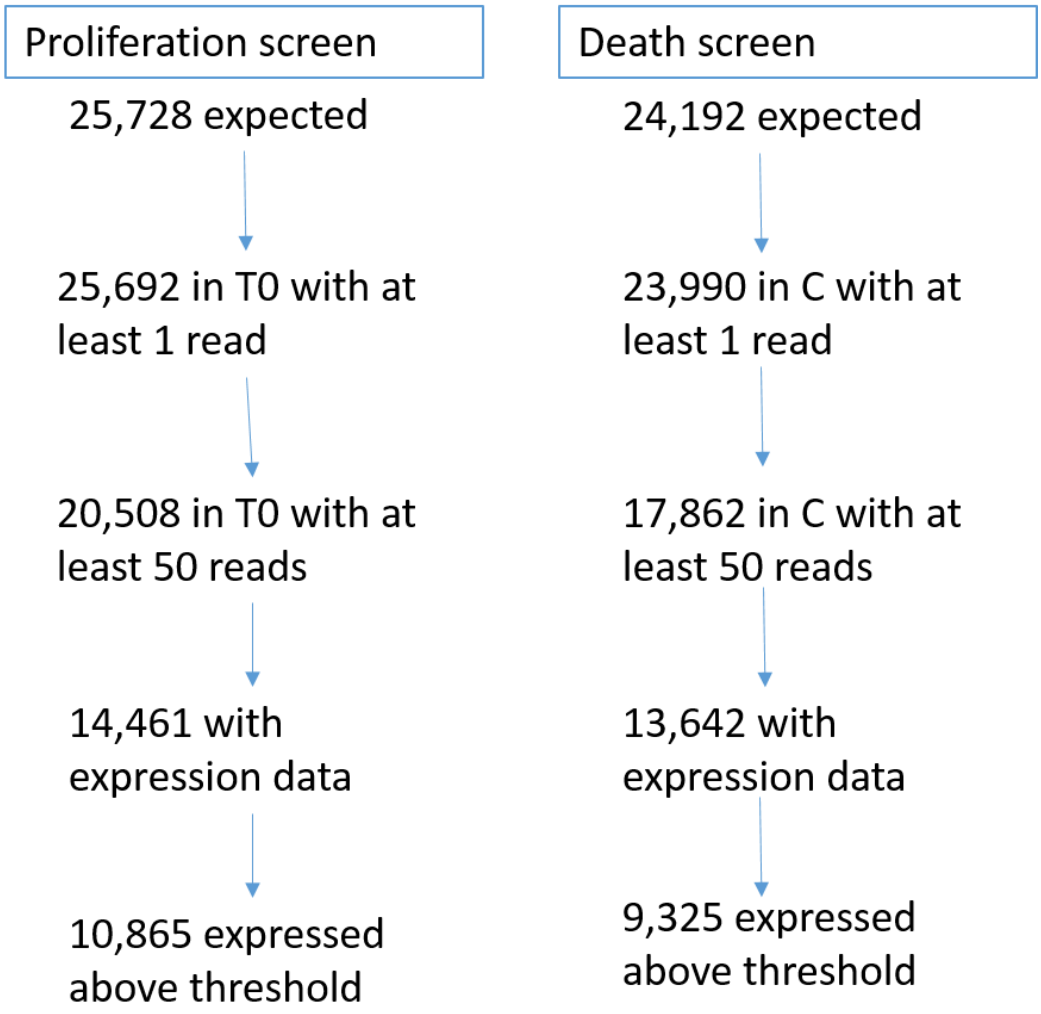



*Figure 4.18.Number of shRNA targeting sequences in both control conditions after each analysis or thresholding step.*

| Human gene name | Z-score (T7) | Z-score (T14) |
|---|---|---|
| *PWP2* | 6.927 | 4.910 |
| *TMEM165* | 6.760 | 5.271 |
| *SNRNP27* | 6.509 | 6.123 |
| *ZMYND11* | 6.344 | 5.354 |
| *MKRN2* | 5.538 | 3.793 |
| *KCNA5* | 5.465 | 3.883 |
| *ARHGAP26* | 5.423 | 4.479 |
| *FARSB* | 5.045 | 3.605 |
| *NAP1L1* | 4.921 | 3.558 |
| *ZNF8* | 4.527 | 3.708 |
| *NAA20* | 4.520 | 6.535 |
| *EPHA4* | 4.315 | 3.983 |
| *ATP6V0E1* | 4.268 | 3.859 |
| *GOT2* | 4.136 | 5.332 |
| *CDH20* | 4.040 | 4.068 |
| *PABPC1L* | 4.016 | 5.443 |
| *ATG5* | 3.921 | 3.647 |
| *WDR25* | 3.848 | 3.588 |
| *GEMIN6* | 3.796 | 5.263 |
| *NAT6* | 3.757 | 4.229 |
| *PDZK1* | 3.730 | 3.643 |
| *PRICKLE2* | 3.703 | 4.364 |
| *SPTB* | 3.664 | 3.728 |
| *CACNB2* | 3.576 | 6.042 |
| *PLCD4* | 3.519 | 4.256 |
| *SRSF7* | 3.486 | 3.588 |
| *SNX7* | 3.334 | 3.653 |
| *GPATCH2* | 3.311 | 4.550 |
| *UROS* | 3.270 | 3.582 |

*Table 4.23. List of genes found in both the T7 and T14 GOI lists and their z-scores in both datasets.*
*Genes are ordered by their z-score in T7.*

### 4.3.10. Selection of genes potentially involved in β cell resistance to death caused by chronic exposure to lipids

The same approach used to prioritize targeting sequences in the proliferation screens was used in the death screen. Targeting sequences were considered as having increased resistance to chronic exposure to lipids if their z-score was on the top 5% of the z-score distribution for sample P. Of the constructs screened, 320 identified as such. The GOI list was composed of 308 genes, 12 genes were targeted by two sequences.

A step that filtered for genes that are expressed in human β-cells was included using the same dataset and approach as in section 4.3.9. After this filtering step 204 targeting sequences remained in the list of prioritized sequences. The GOI list was composed of 195 genes, of which nine were targeted by two different targeting sequences: LRIT3, EPRS, MGA, CHRNB1, TBC1D10C, ZBED3, CP, SPON2 and HIC2.

- Overlap with the proliferation screen

Genes identified in this screen could result in resistance to death caused by the chronic exposure to palmitate. However, it was hypothesized that the effect of some of the targeting sequences could be due to an increase in proliferation that would overcome the cell death rate. As this is independent of the death-inducing stimulus, these genes would also be more highly represented in the vehicle-only control, but the high cell death rate could drastically reduce the proportion of other constructs and generate high z-scores in the palmitate sample. Therefore to identify genes that might exert their effect through an increase in proliferation rate, the overlap between this GOI list and the proliferation screen GOI lists was investigated. A Venn diagram has been used in Figure 4.19 to represent the degree of overlap between the two screens. The GOI lists used were those which had been filtered for expression in human β-cells. Only one gene is present in all GOI lists: MKRN2. As expected given the different biological mechanisms of both screens, more genes are common between both samples of the proliferation screen (T7 and T14) than between the death screen and any of the two proliferation samples. Table 4.24 contains

the names of the genes overlapping the death screen and each of the proliferation screen samples.

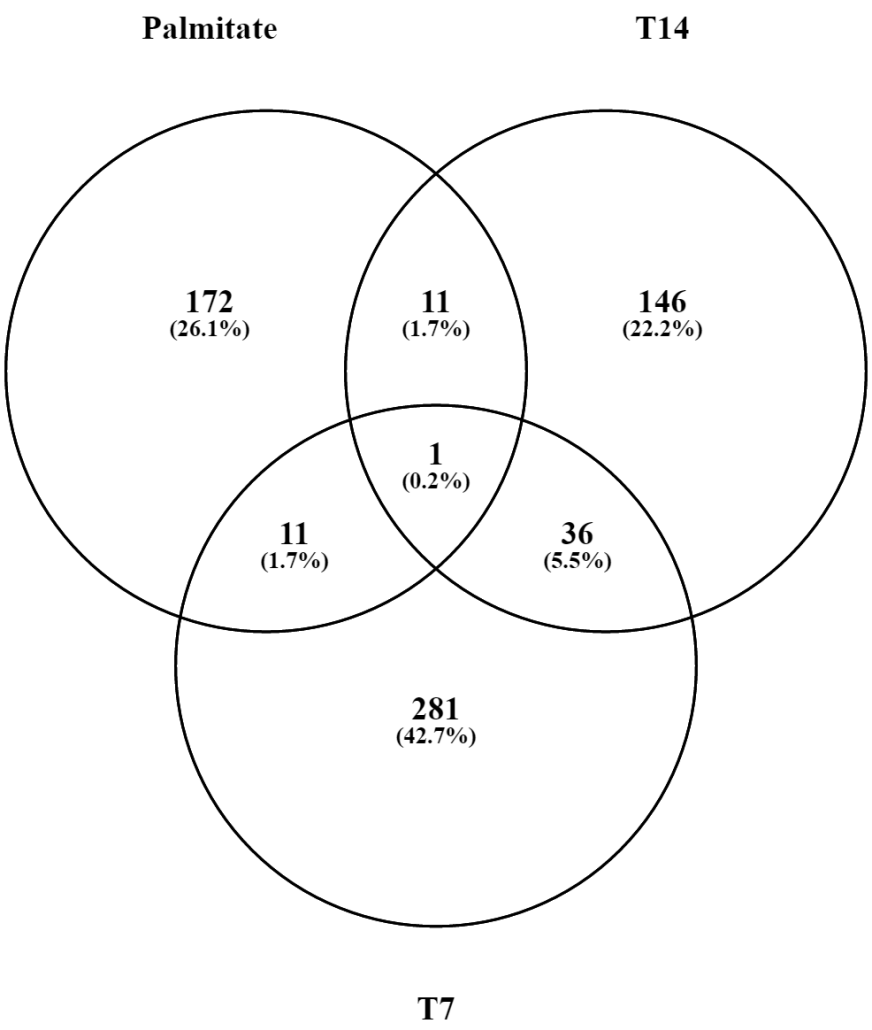



*Figure 4.19. Venn diagram showing the overlap of the proliferation (T7, T14) and death (palmitate) GOI lists.*

| Genes in P and T7 GOI lists | Genes in P and T14 GOI lists |
|---|---|
| *BEX4* | *LRIT3* |
| *FAM189B* | *CHRNE* |
| *CHD6* | *IMPA1* |
| *NRG3* | *TSPAN32* |
| *ARNTL* | *LTBR* |
| *LIG3* | *ZMYND8* |
| *RALGPS2* | *PPP1R14C* |
| *DPP4* | *VAV3* |
| *AIP* | *REEP6* |
| *PLCL2* | *SLC4A5* |
| *AGBL5* | *SUCLA2* |

*Table 4.24. Genes present in the death screen and each one the proliferation screen samples.*

### 4.3.11. Genes differentially expressed upon exposure to palmitate

The mechanism of action of palmitate as a lipotoxic agent is not well understood, but efforts have been made to study the effect that palmitate exposure has on β-cells. Specifically, gene expression changes upon exposure to palmitate have been studied to attempt to elucidate its effects.

It was hypothesized that as the death screen was designed to detect genes which upon knockdown would protect the β-cells from the toxic effects of exposure to palmitate, the GOIs might be enriched in palmitate-regulated genes.

A dataset of genes differentially expressed in human islets after 48h exposure to palmitate was obtained. This dataset was produced by Hall and collaborators (Hall, Volkov et al. 2014) and made publicly available. It contains only those genes with a q-value <0.05 for differential expression upon exposure to palmitate compared to a vehicle-only control. The authors stated that no cell death was observed in the course of this experiment, suggesting that palmitate exerts a lower cell toxicity to human beta cells. This might be of importance since a high level of cell death was observed in the course of the lipotoxicity screen.

The GOI list obtained from the screen was merged with the expression dataset. Of the 195 genes in the GOI list, 22 were found to be differentially expressed in the dataset. A random sampling approach was used to evaluate the statistical significance of this finding. Using an adapted version to the script prepared for the random sampling performed in section 3.3.10.3, 1000 iterations of 195 gene names were randomly drawn from a list containing the names of the screened genes. No statistically significant enrichment for genes differentially expressed after 48h exposure to palmitate was detected.

The list of genes found to be differentially expressed in the dataset and the direction of change detected by the authors are available in Table 4.25. Although a high proportion of the genes identified in my screen that are also differentially expressed in response to palmitate are down-regulated by palmitate, the majority of genes differentially regulated by palmitate are downregulated. As such, no enrichment for down-regulated genes was found using a random sampling approach.

| Gene | Direction of effect |
|---|---|
| *SPON2* | Down-regulated |
| *SEZ6L* | Up-regulated |
| *TMEM50A* | Down-regulated |
| *PRKAA1* | Down-regulated |
| *PLEKHB1* | Down-regulated |
| *HNRNPK* | Down-regulated |
| *PTPRG* | Down-regulated |
| *WWTR1* | Down-regulated |
| *METTL7A* | Down-regulated |
| *SMAD3* | Down-regulated |
| *ACADM* | Down-regulated |
| *MLEC* | Down-regulated |
| *BEX4* | Up-regulated |
| *REEP6* | Down-regulated |
| *PLSCR4* | Down-regulated |
| *MAPKAP1* | Down-regulated |
| *SPTLC3* | Down-regulated |
| *PRIM1* | Down-regulated |
| *ODC1* | Up-regulated |
| *AIP* | Up-regulated |
| *ACSS1* | Down-regulated |
| *FOS* | Up-regulated |

*Table 4.25. Genes in the lipotoxicity screen GOI list found to be de-regulated in the presence of palmitate for 48 hours.*

### 4.3.12. Network analysis of genes potentially involved in β-cell proliferation

The advantages of a network analysis approach to prioritize candidate genes from an RNAi screen have been described in section 3.3.10. A similar pipeline was used in the analysis of the screen described in this chapter. One important difference was implemented however: as the screen was performed in mouse β-cells, the analysis was first performed using the mouse gene names. As the expression data in human β-cells was used to filter the GOI lists these reduced GOI lists composed of human gene names were then used to perform the same analyses.

#### *4.3.12.1 Network analysis of the candidate mouse genes identified, using STRING.*

STRING was used as described in section 3.3.10.1. Briefly, a confidence threshold of 0.7 was used and "textmining" was not used as a source of evidence for protein interactions to enrich for functional interactions. The resulting interactions were downloaded and inputted into Cytoscape for visualization.

The GOI lists for the proliferation screen samples, T7 and T14 were first inputted separately and then merged to constitute a "proliferation GOI list". Mouse gene names were used and no filtering was performed.

Upon inputting of the T7 GOI list, STRING mapped 464 gene names of the 477 GOIs. A view of the network is shown in Figure 4.20. The interaction enrichment analysis produced by STRING indicates that there 117 interactions in the GOI list inputted when 62 were expected. The p-value for the enrichment in interactions is $2.27 \times 10^{-10}$ which indicates that the GOI list is enriched for interacting genes or gene products.

Of the 319 gene names in the T14 GOI list, STRING mapped 304. The network produced (Figure 4.21) is composed of 100 interactions when 38 were expected by STRING, which indicates that the T14 GOI list is also enriched for interactions (p-value for enrichment calculated by STRING is $1.11 \times 10^{-16}$).

Inputting the proliferation GOI list (T7+T14) produced a network with 348 interactions when the expected number calculated by STRING was 174. The p-

value for this apparent enrichment in connections is reported as 0, indicating the network is statistically enriched in interacting genes. This proliferation network is represented in Figure 4.22.

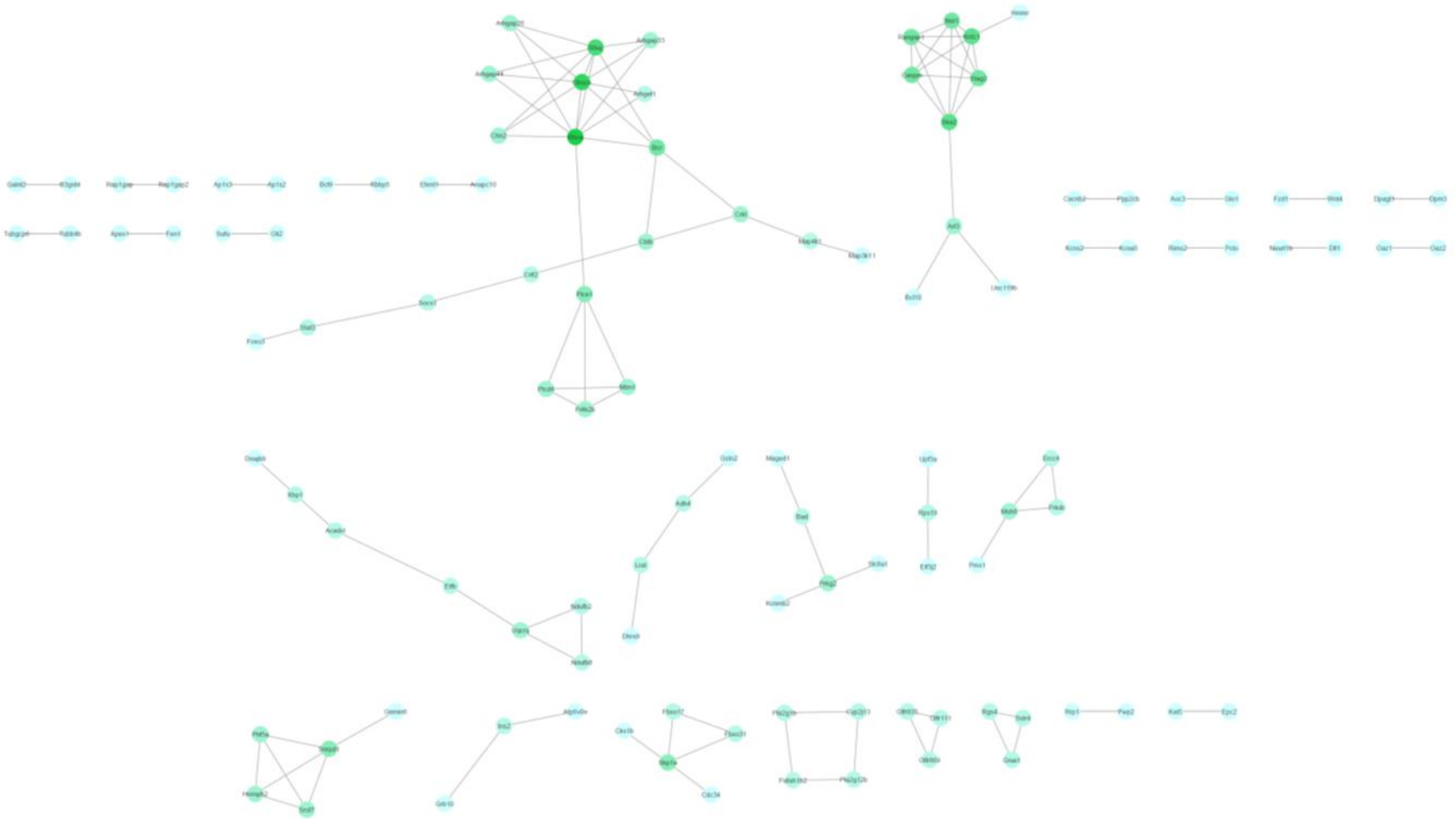


*Figure 4.20. Protein interaction network for the GOI list of the T7 proliferation sample.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue.*

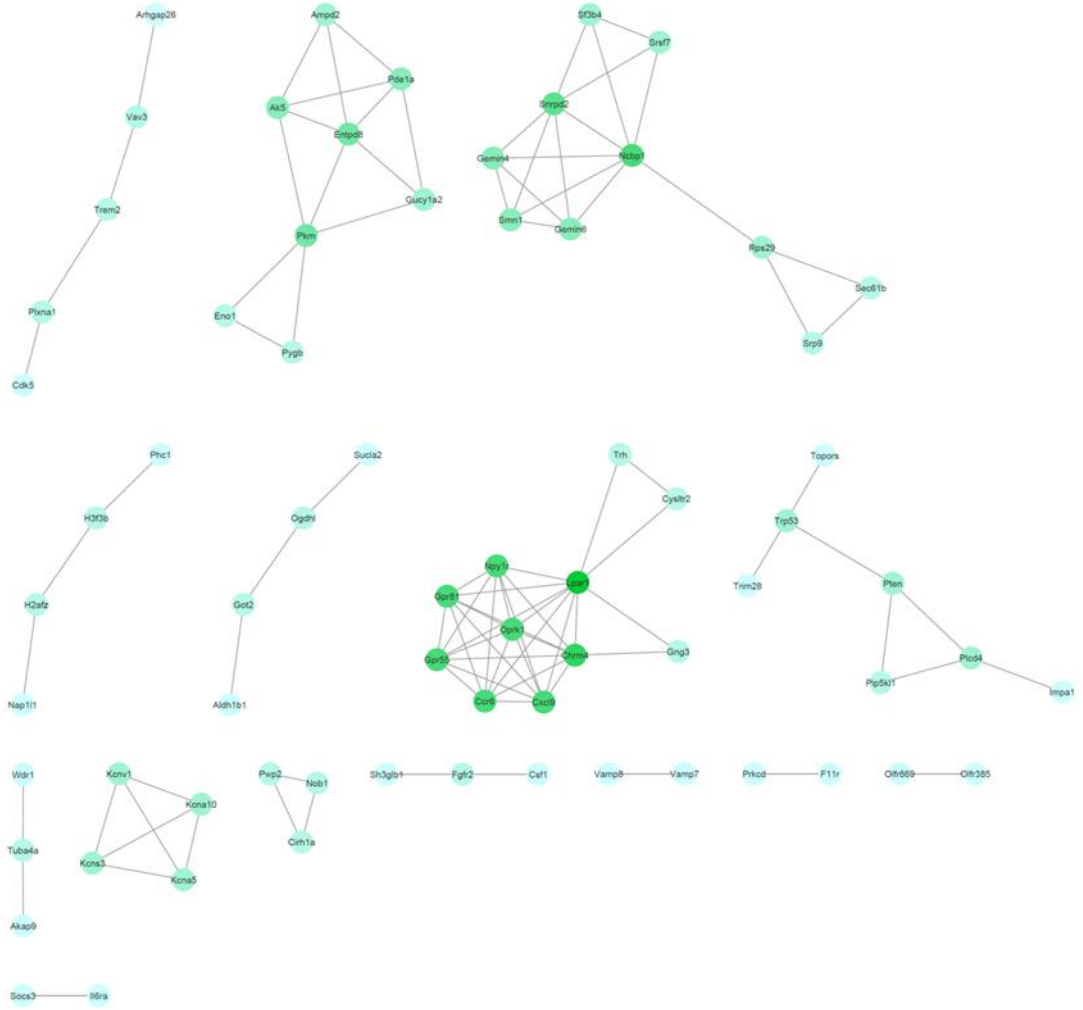



*Figure 4.21. Protein interaction network for the GOI list of the T14 proliferation sample.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue.*

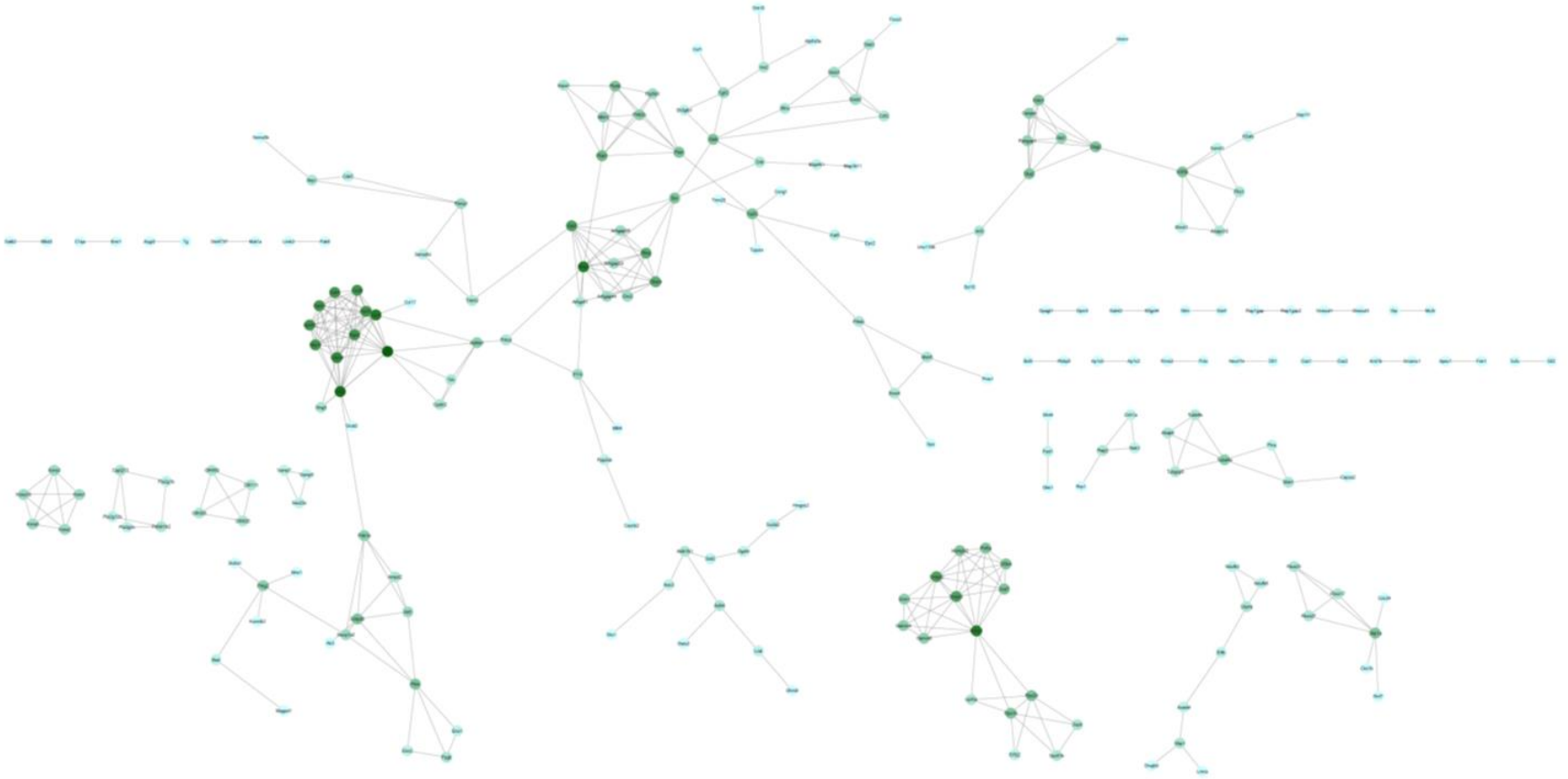


*Figure 4.22. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue.*

#### *4.3.12.2 Network analysis of the candidate mouse genes identified using FunCoup*

FunCoup was used to construct gene interaction networks for the reasons explained in section 3.3.10.2. The same parameters used to construct the interaction networks for the screen results in Chapter 3 (section 3.3.10.2) were used. Briefly, the *Mus musculus* database was used, the confidence threshold was set to 0.8 and the expansion depth used was 0, which means that the networks were built using only the gene names inputted and no additional genes were added by FunCoup. Therefore the whole network was comprised of genes identified in my screen.

The networks produced with FunCoup were imported into Cytoscape for visualization. The T7 network (Figure 4.23) is composed of 159 genes and 598 connections. The T14 network (Figure 4.24) is smaller, with 90 genes connected by 253 edges. Similarly to the results found in section 3.3.10.2 when the GOI lists for both co-culture samples were merged, the union of the GOI lists for both proliferation samples produced a network that includes more nodes and edges than the sum of both if they were completely independent. However in this case it is not unexpected that the GOI from both lists interact, as both T7 and T14 are samples derived from the same phenotype: an increase in proliferation. The network built from the combination of both GOI lists is depicted in Figure 4.26. The outside of the circles representing the genes are colour coded to distinguish which GOI list they were identified in.

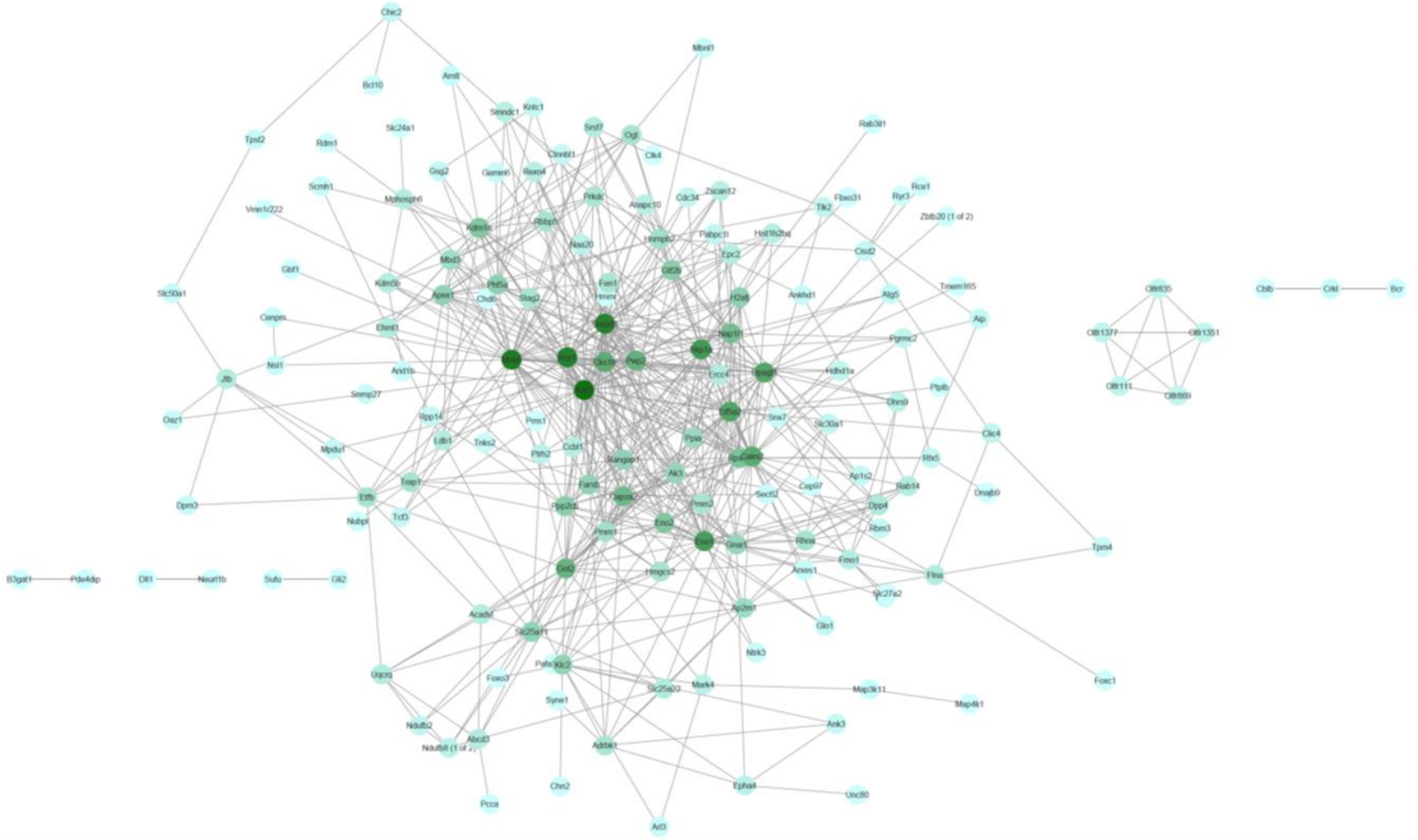



*Figure 4.23. Protein interaction network for the GOI list of the T7 proliferation sample.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue.*

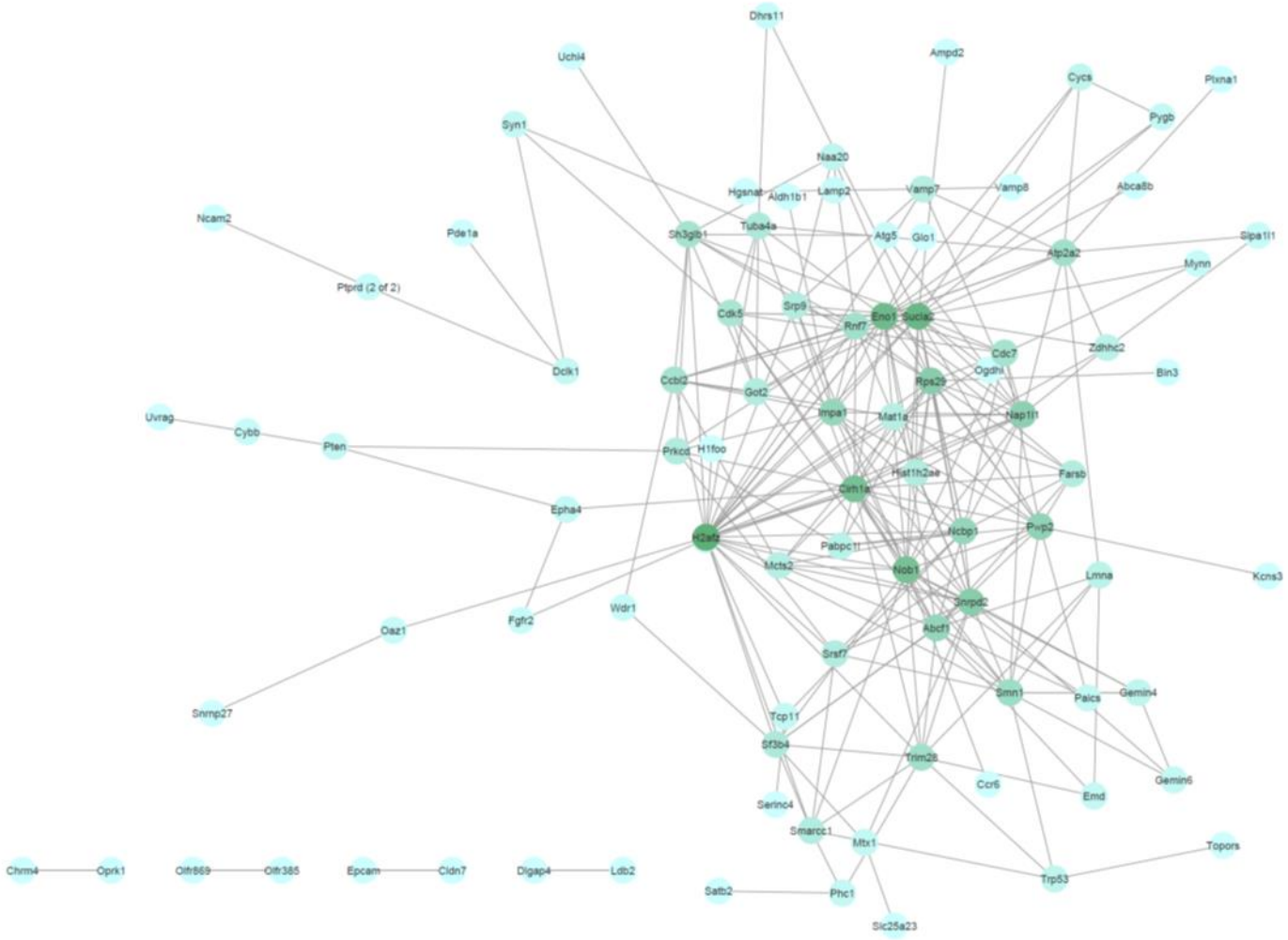



*Figure 4.24. Protein interaction network for the GOI list of the T14 proliferation sample.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue.*

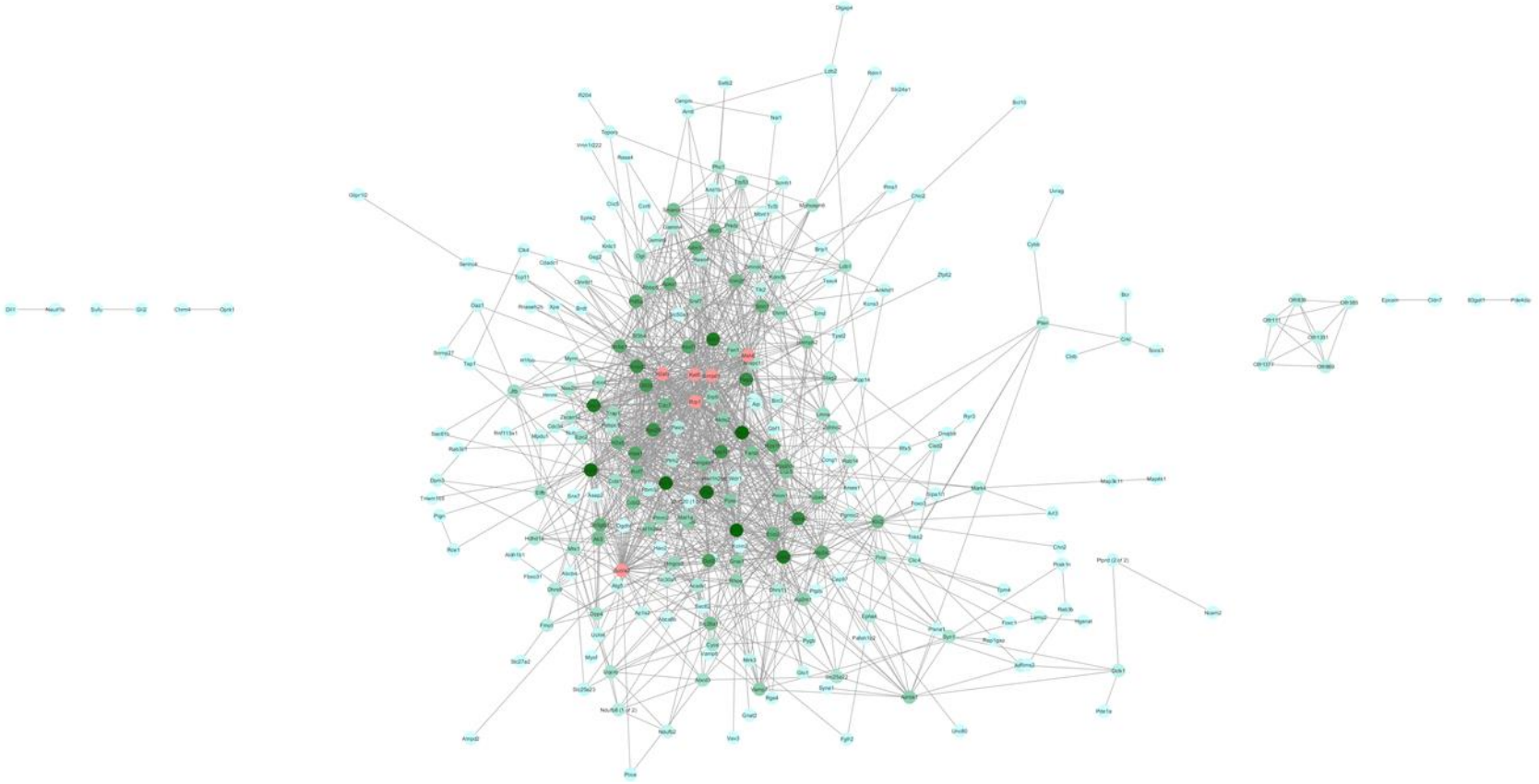


*Figure 4.25. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue. Pink circles represent genes with more than 50 interactors in the sample.*

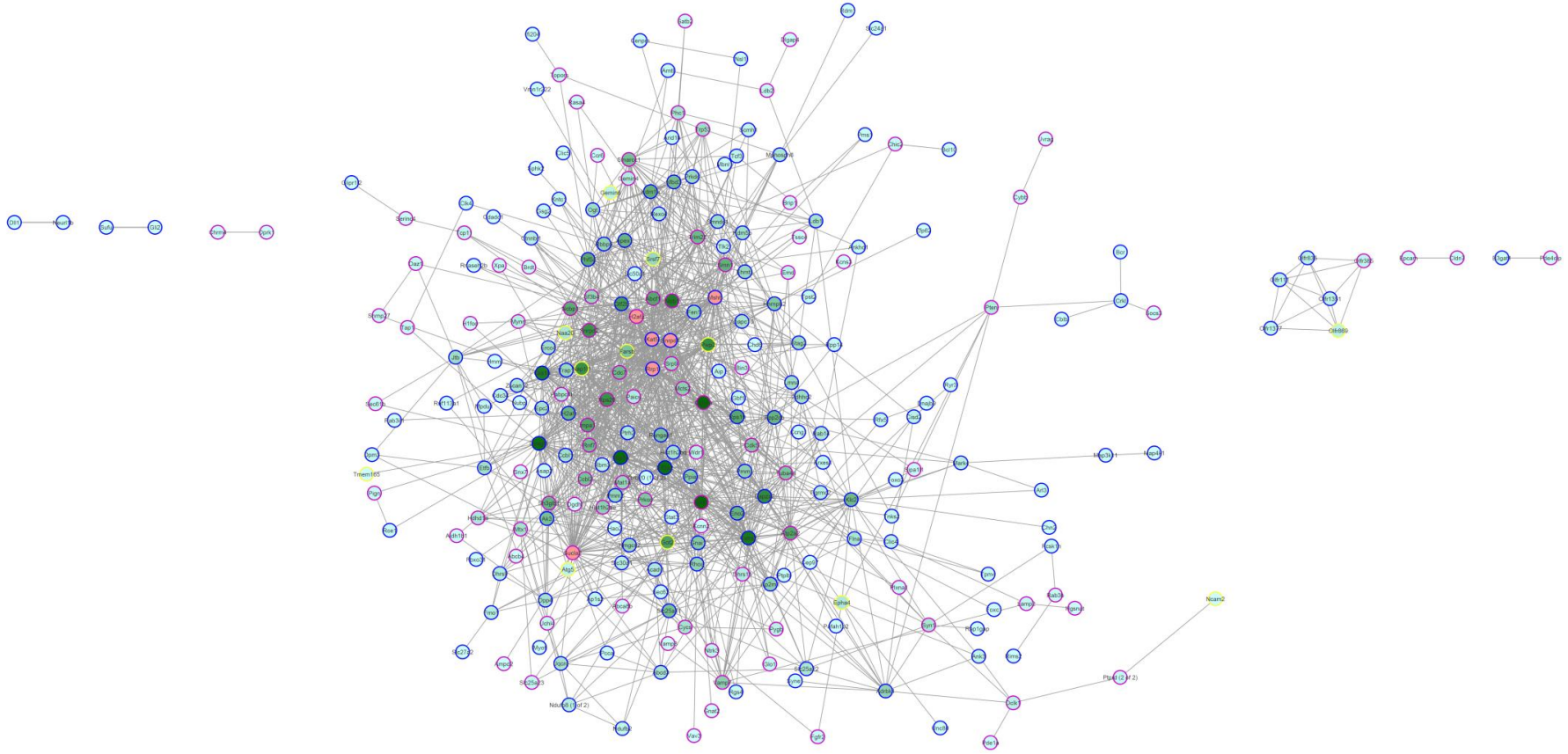


*Figure 4.26. Protein interaction network for the proliferation GOI list, which includes the T7 and T14 GOI lists.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue. Pink circles represent genes with more than 50 interactors in the sample. The border colour corresponds to the GOI list they were identified in. Blue: T7, pink: T14. Yellow border indicates that the gene was identified in both GOI lists.*

#### *4.3.12.3 Hyperconnected mouse GOIs*

“Hyperconnected” genes or hubs were identified using the same approach used in section 3.3.10.3. Briefly, a custom script was written in Python to produce 1,000 random samples of the size of the GOI lists for each gene in the network. Each of those 1,000 random samples for one gene includes the gene being studied. The script counts how many connectors the GOI has in each of the 1,000 samples and reports the number of random samples with at least as many genes that interact with the gene being investigated as that found in the GOI list. The proportion of samples with at least as many connectors as the GOI list is used as a p-value.

This random sampling approach provides a measure of the enrichment in connectors of each gene in the GOI list. If less than 5% of the random samples contain at least the same number of connectors for a given gene than the GOI list, this gene is determined as being “hyperconnected” with a p-value<0.05.

In the T7 GOI list, 19 genes were determined as being hyperconnected (Table 4.26). In the T14 GOI list, 13 genes were considered hyperconnected (Table 4.27). When the GOI lists were combined, 44 genes were found to be hyperconnected (Table 4.28). From these 44 genes it was expected that some might be hyperconnected in the individual GOI lists. The overlap among the different GOI lists is represented by a Venn diagram in Figure 4.27.

Although most of the genes in the individual GOI lists are included in the combined GOI list, three hyperconnected genes in each of the individual lists are found to be hyperconnected in the combined GOI list. This is not unexpected in the case of genes which have a low number of potential connectors amongst the screened genes. These potential connectors might not have been included in many of the 1,000 random samples due to a small sample size when assessing the individual GOI list they were identified in. However, combining both GOI lists results in a larger sample size which can more easily lead to the sampling of those few potential connectors. In the T7 GOI list, the three genes that are not hyperconnected in the combined list are *Glo1*, *Slc27a2* and *Clic4*. Although the number of connectors in the screened gene list is not very low (18, 7 and 38 respectively), the fraction of the connectors identified in the GOI list is not high (1/6) compared to other of the

hyperconnected genes which remained as hyperconnected in the combined GOI list.

The three genes which were detected as hyperconnected in T14 but not in the combined list are *Cldn7*, *Ntrk3* and *Fgfr2* (5, 26 and 24 connectors in the screened genes).

Another factor contributing to genes not being in the combined list might be that while some genes may have connectors in the other individual GOI list, some may not. This could result from either the connectors not having an effect strong enough to be in the GOI list or from the connectors not having been screened in the other sample. This is a possibility for the hyperconnected genes in T7 because some of the pools that were screened in T7 were not screened in T14.

The effect of combining the GOI lists in the number of connectors found for each gene in each of the individual networks was calculated as (number of connectors in the combined list)-(number of connectors in T7)-(number of connectors in T14). The distribution for the number of connectors gained when a network is built from the combined GOI list is depicted in Figure 4.28. Most of the genes either gained no connectors or a small number of them when the GOI lists were combined to draw a proliferation network. The highest number of connectors gained was 28 (*Cirh1a* and *H2afz*). Some genes have a negative number of additional connectors, meaning that the connectors for those genes overlap between both lists and their combination produces less connectors than expected if the number of connectors in both lists is simply summed.

The six genes that went from hyperconnected in their respective individual GOI lists to not hyperconnected in the combined GOI list network have either not gained any additional connectors or their connectors overlapped with the other GOI list. When the random sample size was increased to model the combined list it resulted in more random samples having at least the same number of connectors than the combined list and therefore they were not identified as hyperconnected.

| Gene | Potential interactors screened | Interactors in GOI list | p-value |
|---|---|---|---|
| *B3gat1* | 1 | 1 | <0.001 |
| *Gli2* | 1 | 1 | <0.001 |
| *Neurl1b* | 1 | 1 | <0.001 |
| *Rgs4* | 4 | 2 | <0.001 |
| *Ryr3* | 1 | 1 | <0.001 |
| *Slc24a1* | 1 | 1 | <0.001 |
| *Foxc1* | 2 | 1 | 0.001 |
| *Mbnl1* | 5 | 2 | 0.001 |
| *Rce1* | 2 | 1 | 0.002 |
| *Arxes1* | 7 | 2 | 0.003 |
| *Map3k11* | 9 | 2 | 0.004 |
| *Slc50a1* | 7 | 2 | 0.005 |
| *Rfx5* | 14 | 3 | 0.008 |
| *Cep97* | 5 | 1 | 0.013 |
| *Mark4* | 33 | 5 | 0.018 |
| *Glo1* | 18 | 3 | 0.025 |
| *Slc27a2* | 7 | 1 | 0.034 |
| *Scmh1* | 11 | 2 | 0.038 |
| *Clic4* | 38 | 4 | 0.043 |

*Table 4.26. Genes found to be enriched (p-value<0.05) in interactors in the T7 GOI list.*

| Gene | Potential interactors screened | Interactors in GOI list | p-value |
|---|---|---|---|
| *Chrm4* | 2 | 1 | 0.003 |
| *Kcns3* | 2 | 1 | 0.003 |
| *Epcam* | 3 | 1 | 0.006 |
| *Gemin4* | 40 | 4 | 0.011 |
| *Pde1a* | 4 | 1 | 0.011 |
| *Oprk1* | 5 | 1 | 0.012 |
| *Cldn7* | 5 | 1 | 0.018 |
| *Ntrk3* | 26 | 3 | 0.018 |
| *Ccr6* | 6 | 1 | 0.019 |
| *H1foo* | 8 | 1 | 0.02 |
| *Serinc4* | 7 | 1 | 0.023 |
| *Plxna1* | 6 | 1 | 0.035 |
| *Fgfr2* | 24 | 2 | 0.048 |

*Table 4.27. Genes found to be enriched (p-value<0.05) in interactors in the T14 GOI list.*

| Gene | Potential interactors screened | Interactors in GOI list | p-value |
|---|---|---|---|
| *B3gat1* | 1 | 1 | <0.001 |
| *Gli2* | 1 | 1 | <0.001 |
| *Neurl1b* | 1 | 1 | <0.001 |
| *Pcsk1n* | 2 | 2 | <0.001 |
| *Rce1* | 2 | 2 | <0.001 |
| *Ryr3* | 1 | 1 | <0.001 |
| *Scmh1* | 11 | 4 | <0.001 |
| *Slc24a1* | 1 | 1 | <0.001 |
| *Arxes1* | 7 | 3 | 0.001 |
| *Rab3b* | 4 | 2 | 0.001 |
| *Rgs4* | 4 | 2 | 0.001 |
| *Socs3* | 2 | 1 | 0.004 |
| *Rasa4* | 6 | 2 | 0.005 |
| *Brdt* | 6 | 2 | 0.006 |
| *Ccr6* | 6 | 2 | 0.006 |
| *Ifi204* | 2 | 1 | 0.009 |
| *Kcns3* | 2 | 1 | 0.009 |
| *Chrm4* | 2 | 1 | 0.01 |
| *Foxc1* | 2 | 1 | 0.01 |
| *H1foo* | 8 | 2 | 0.011 |
| *Mbnl1* | 5 | 2 | 0.011 |
| *Slc50a1* | 7 | 2 | 0.011 |
| *Rap1gap* | 3 | 1 | 0.012 |
| *Serinc4* | 7 | 2 | 0.012 |
| *Gemin4* | 40 | 7 | 0.013 |
| *Plxna1* | 6 | 2 | 0.017 |
| *Map3k11* | 9 | 2 | 0.024 |
| *Rfx5* | 14 | 3 | 0.024 |
| *Rims2* | 4 | 1 | 0.025 |
| *Epha4* | 51 | 8 | 0.026 |
| *Phc1* | 57 | 9 | 0.026 |
| *Mark4* | 33 | 6 | 0.028 |
| Crkl | 18 | 4 | 0.03 |
| Xpa | 9 | 2 | 0.032 |
| Oprk1 | 5 | 1 | 0.035 |
| Ehmt1 | 95 | 13 | 0.036 |
| Epcam | 3 | 1 | 0.036 |
| Jtb | 95 | 12 | 0.038 |
| Arntl | 16 | 3 | 0.039 |
| Hdhd1a | 48 | 8 | 0.039 |
| Pde1a | 4 | 1 | 0.039 |
| Cep97 | 5 | 1 | 0.04 |
| Rnf113a1 | 11 | 2 | 0.042 |
| Ldb1 | 68 | 10 | 0.048 |

*Table 4.28. Genes found to be enriched (p-value<0.05) in interactors in the combined (T7, T14) GOI list.*

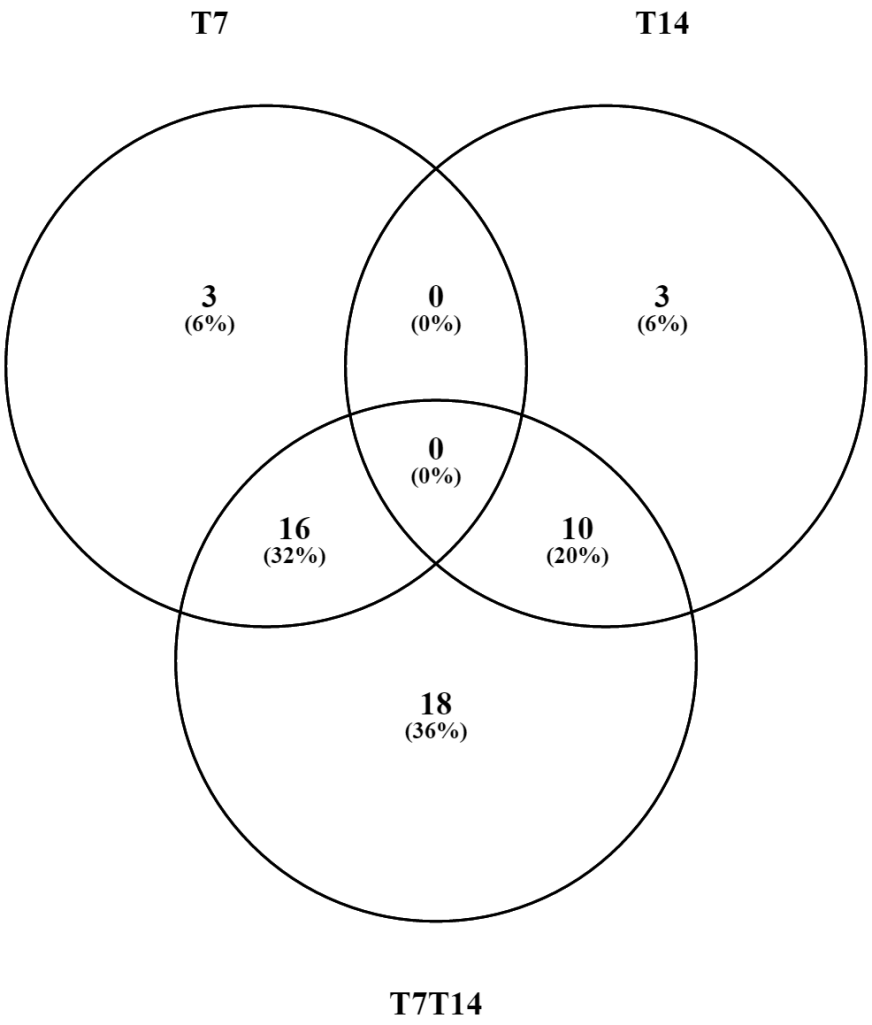



*Figure 4.27. Venn diagram depicting the partial overlap in hyperconnected genes between proliferation GOI lists and the combined proliferation GOI list (T7T14).*

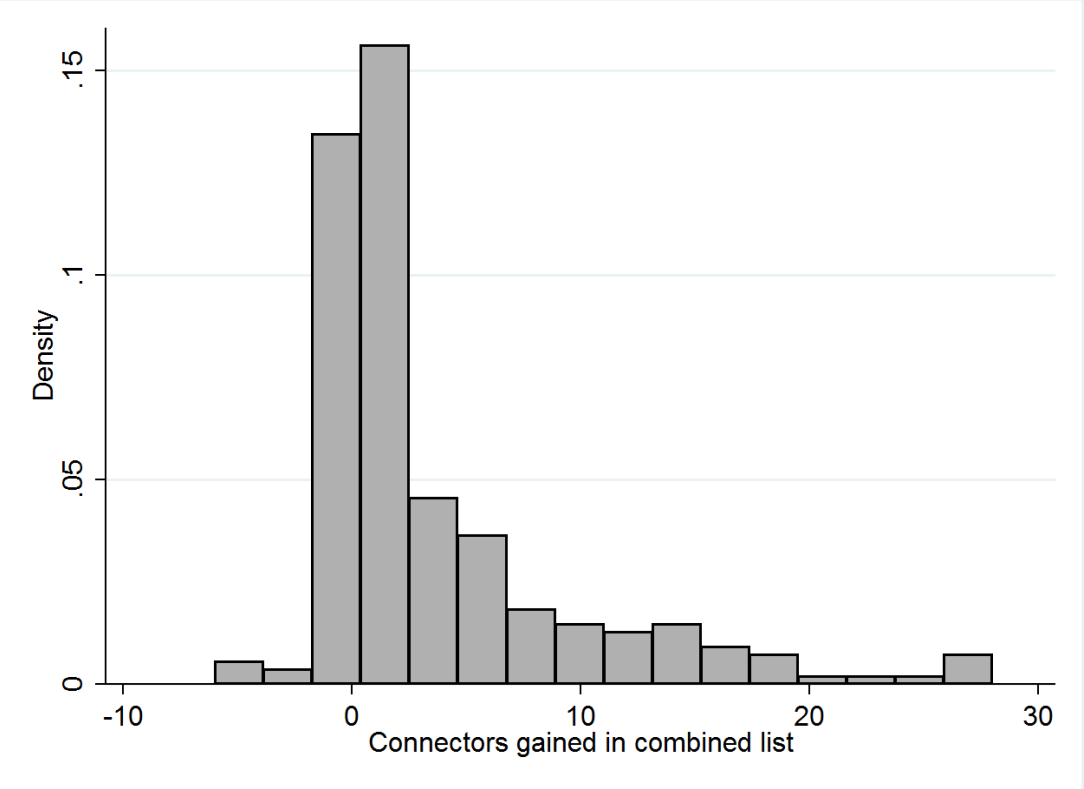



*Figure 4.28. Histogram showing the distribution of the number of additional connectors for each gene in a combined GOI list compared to the sum of the connectors in each list.*

#### *4.3.12.4 Network analysis of the candidate human genes identified using STRING*

The GOI lists filtered for expression in human β-cells as explained in section 4.3.9 were used to build gene networks in STRING following the approach in 4.3.12.1 but using the human database.

STRING mapped all the 329 genes in the T7 GOI list and produced a network (Figure 4.29) where 93 genes are connected by 91 edges. This number of edges is higher than the number expected by STRING (68), which produced an enrichment in connections with a p-value of $4x10^{-5}$ as calculated by STRING, indicating that the network is more highly connected than expected in a random gene list of the same size.

The network produced by STRING (Figure 4.30) for the T14 human GOI list is composed of 65 genes connected by 69 edges. According to the enrichment calculations performed by STRING, this network is not enriched in connections (p-value = 0.15).

The combination of both GOI lists to produce a “proliferation” GOI list, an approach already explained in section 4.3.12.1, produced a network formed of 171 genes connected by 246 edges. STRING indicates that the network is significantly enriched in connections with a p-value of $6.66x10^{-16}$. As the T14 network had been determined as not being enriched in connections it might seem that this enrichment in the combined list is a result of the enrichment in T7. However, although the number of connected genes (171) in the combined GOI list is not much higher than the sum of both lists (158), the high number of connections between pairs of genes indicates that genes from both GOI lists are connected. This is not unexpected as both lists are comprised of genes involved in the same phenotype, i.e. an increase in proliferation. These connections across GOI lists are visible in Figure 4.31, where the borders of the circles representing the genes have been colour coded.

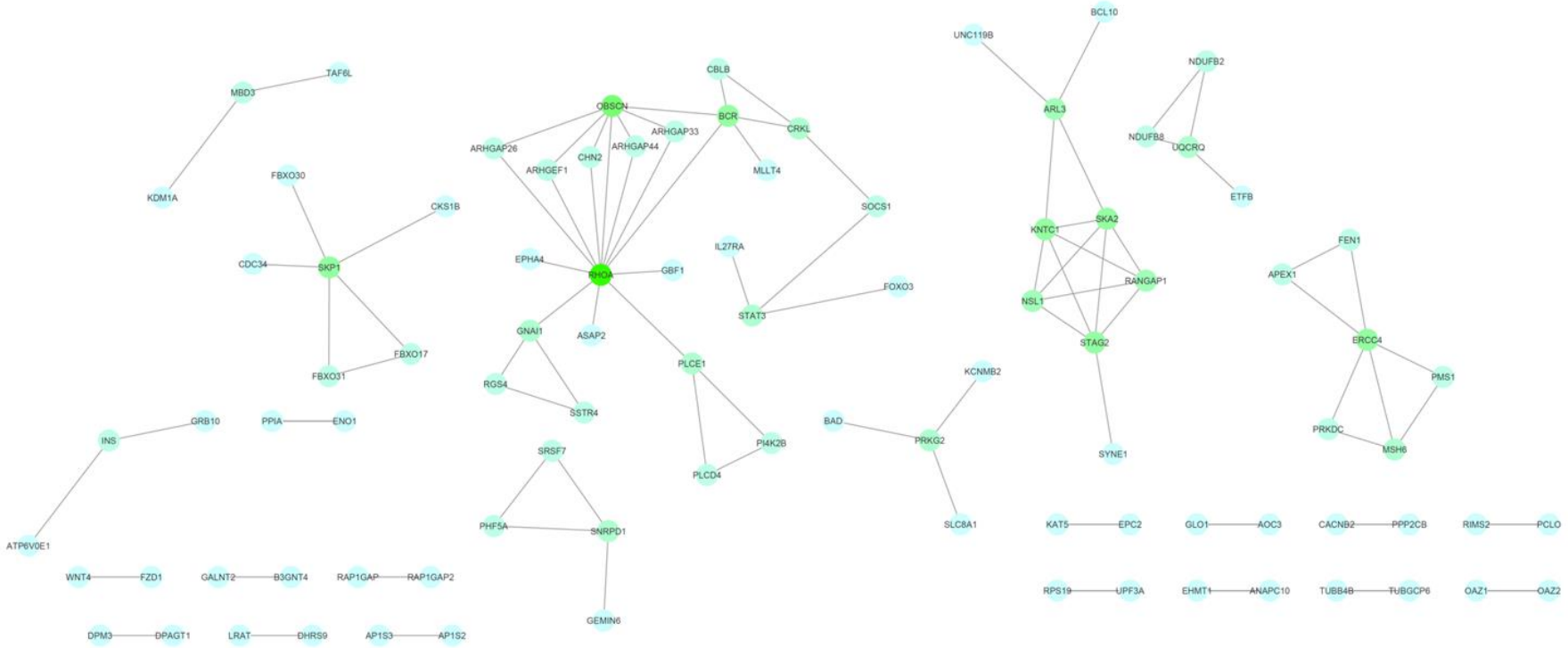



*Figure 4.29. Protein interaction network for the human GOI list of the T7 proliferation sample.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue*

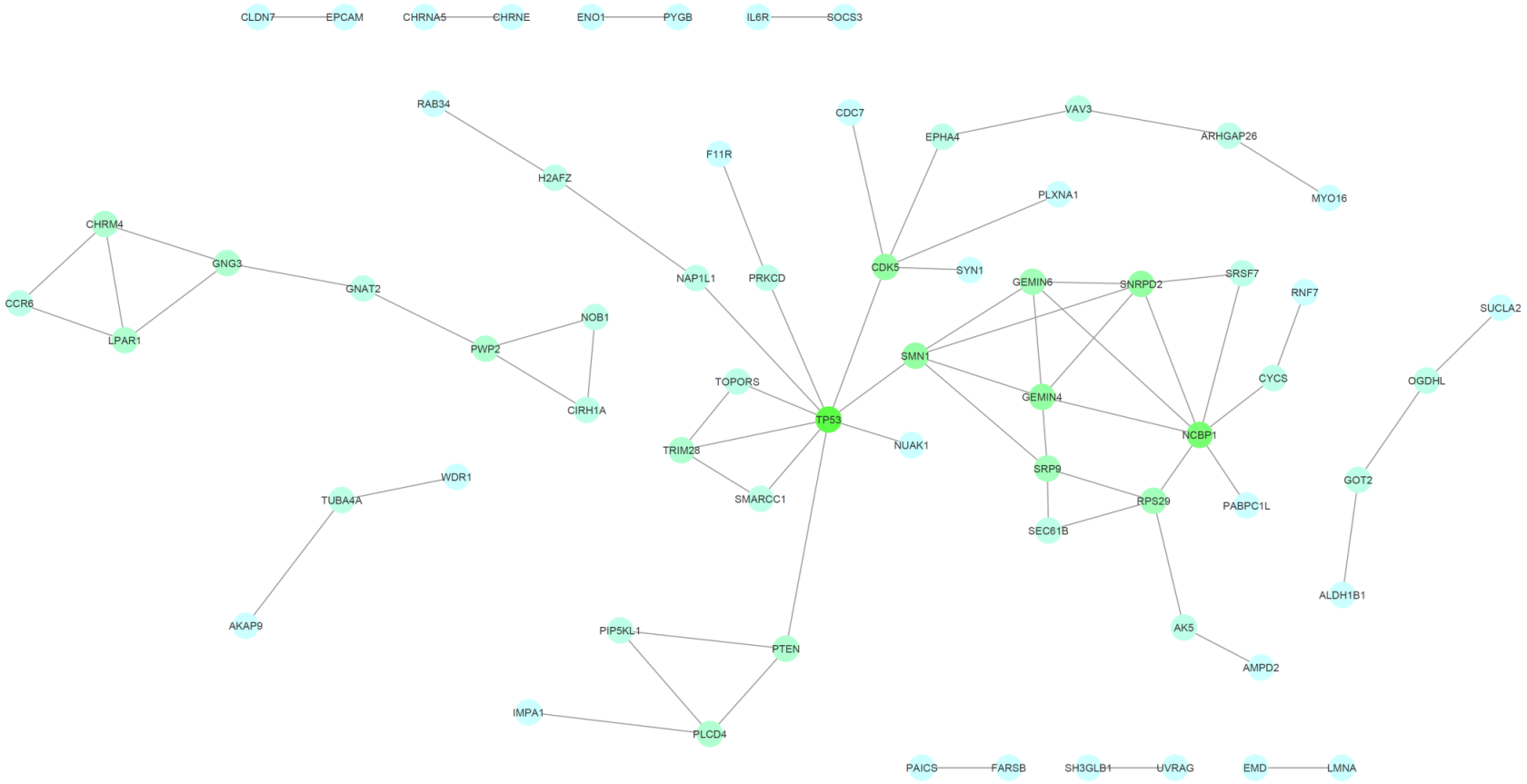



*Figure 4.30. Protein interaction network for the human GOI list of the T14 proliferation sample.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue*

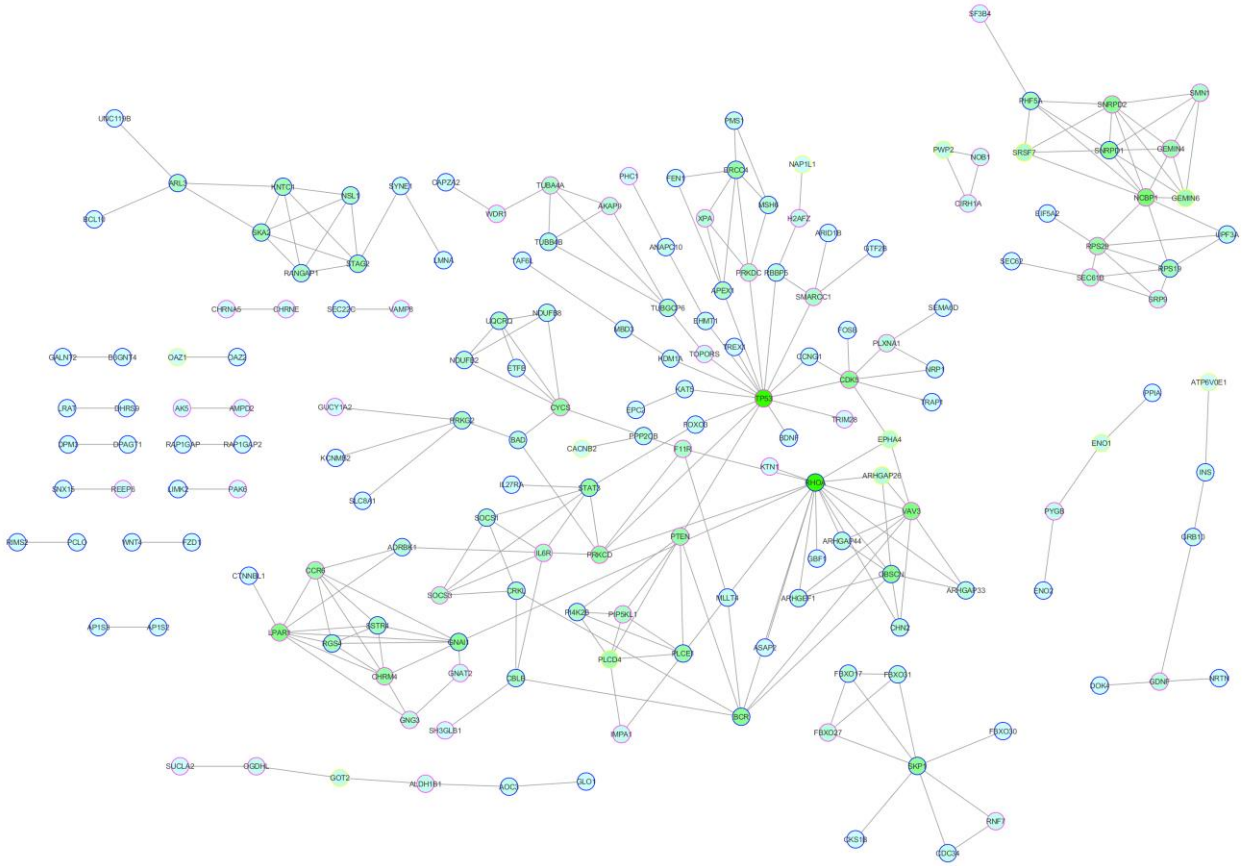


*Figure 4.31 Protein interaction network for the human proliferation combined GOI list.*
*Network was calculated using STRING and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue. The border colour corresponds to the GOI list they were identified in. Blue: T7, pink: T14. Yellow border indicates that the gene was identified in both GOI lists.*

#### *4.3.12.5 Network analysis of the identified human candidate genes using FunCoup*

Gene interaction networks were produced using FunCoup as described in section 4.3.12.2 but using the lists filtered for expression in human β-cells.

The network produced using the T7 GOI list is depicted in Figure 4.32. In this network, 121 genes were found to be connected by 489 edges. The network for the T14 GOI list has the same number of genes as the network produced for the same list using STRING in section 4.3.12.4 , 65 genes, but in contrast with the 69 edges connecting the STRING network FunCoup reports 222 connections in the network. The combined GOI list produced a network (Figure 4.34) with not many more genes than expected by the simple combination of both networks (199 in the combined GOI list versus 186 expected). However, the number of edges in this combined network is higher than the sum of the connections in both individual networks with 1,112 edges versus 711 expected edges. This agrees with the results from STRING indicating that both networks are highly interconnected.

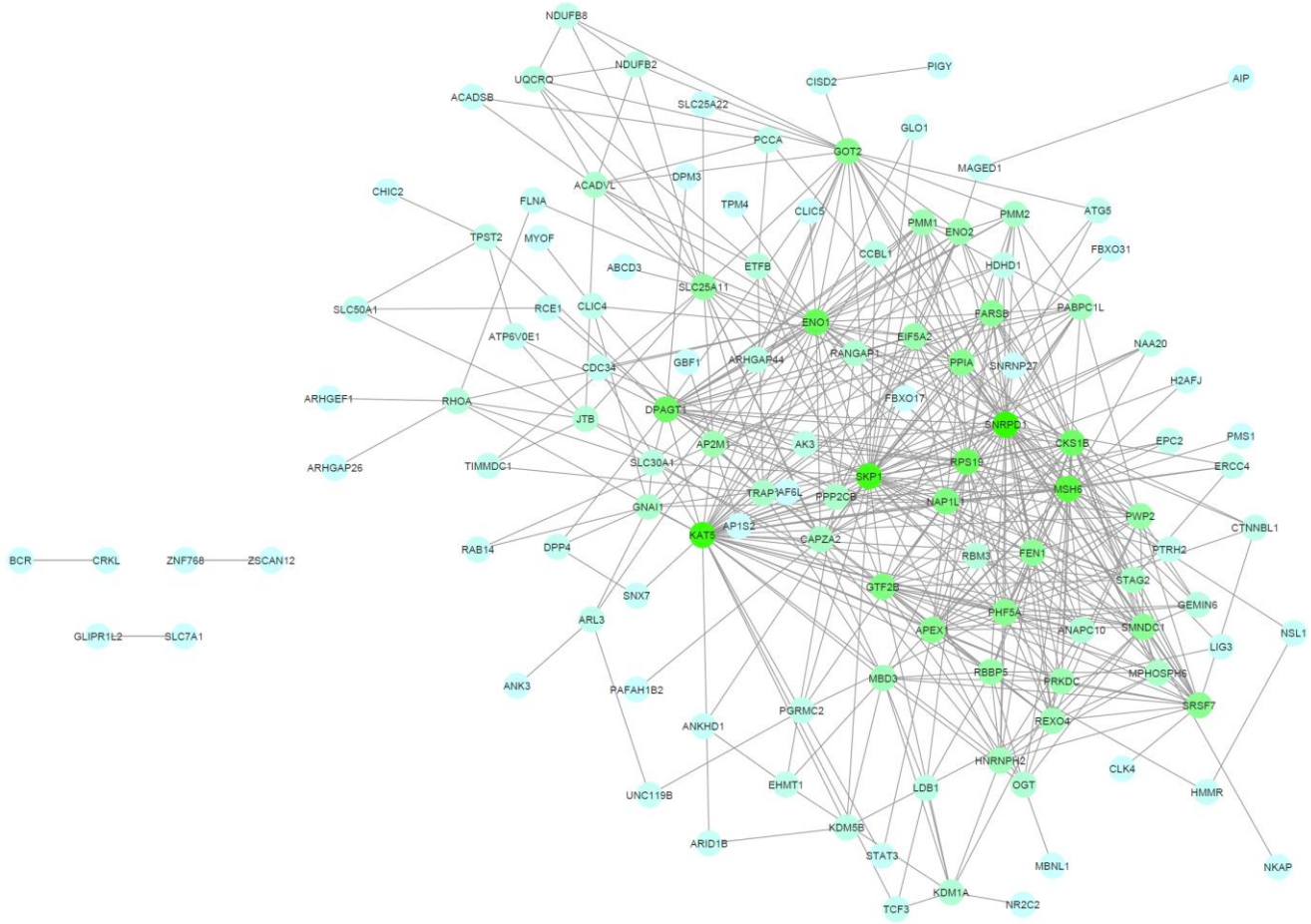



*Figure 4.32. Protein interaction network for the human GOI list of the T7 proliferation sample.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue*

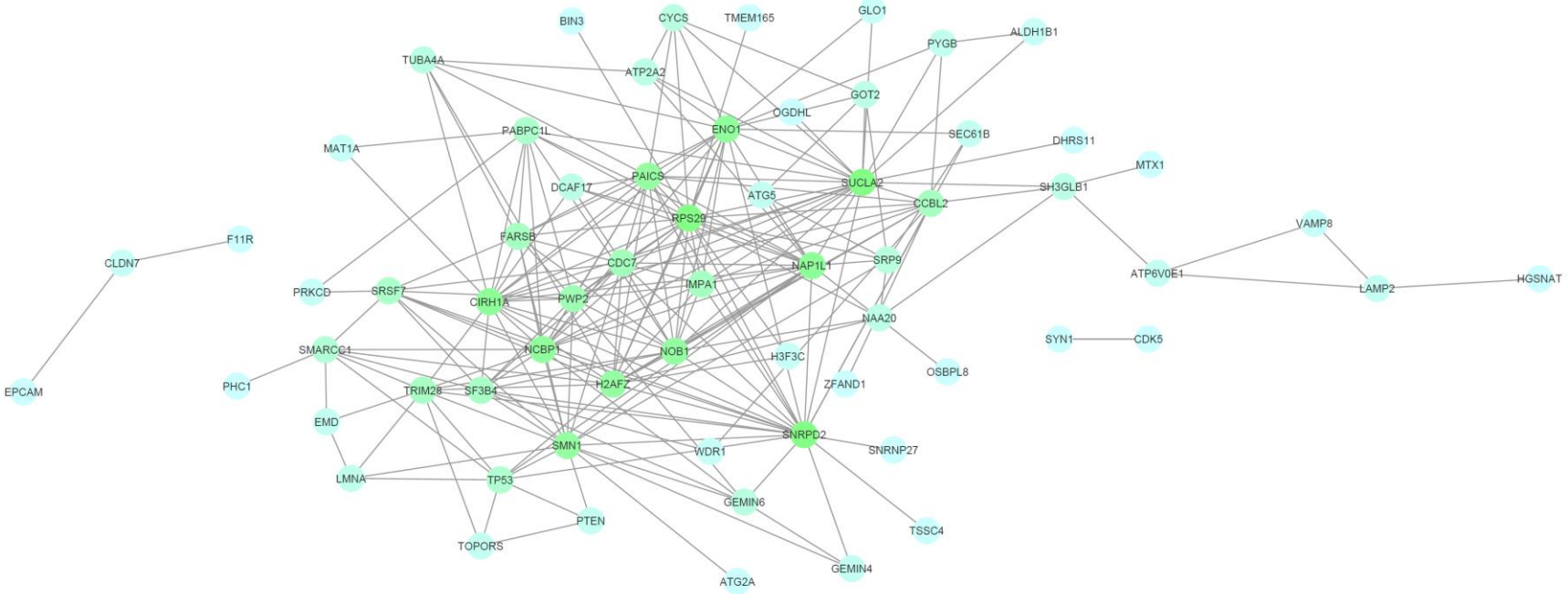



*Figure 4.33. Protein interaction network for the human GOI list of the T14 proliferation sample.*
*Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue*

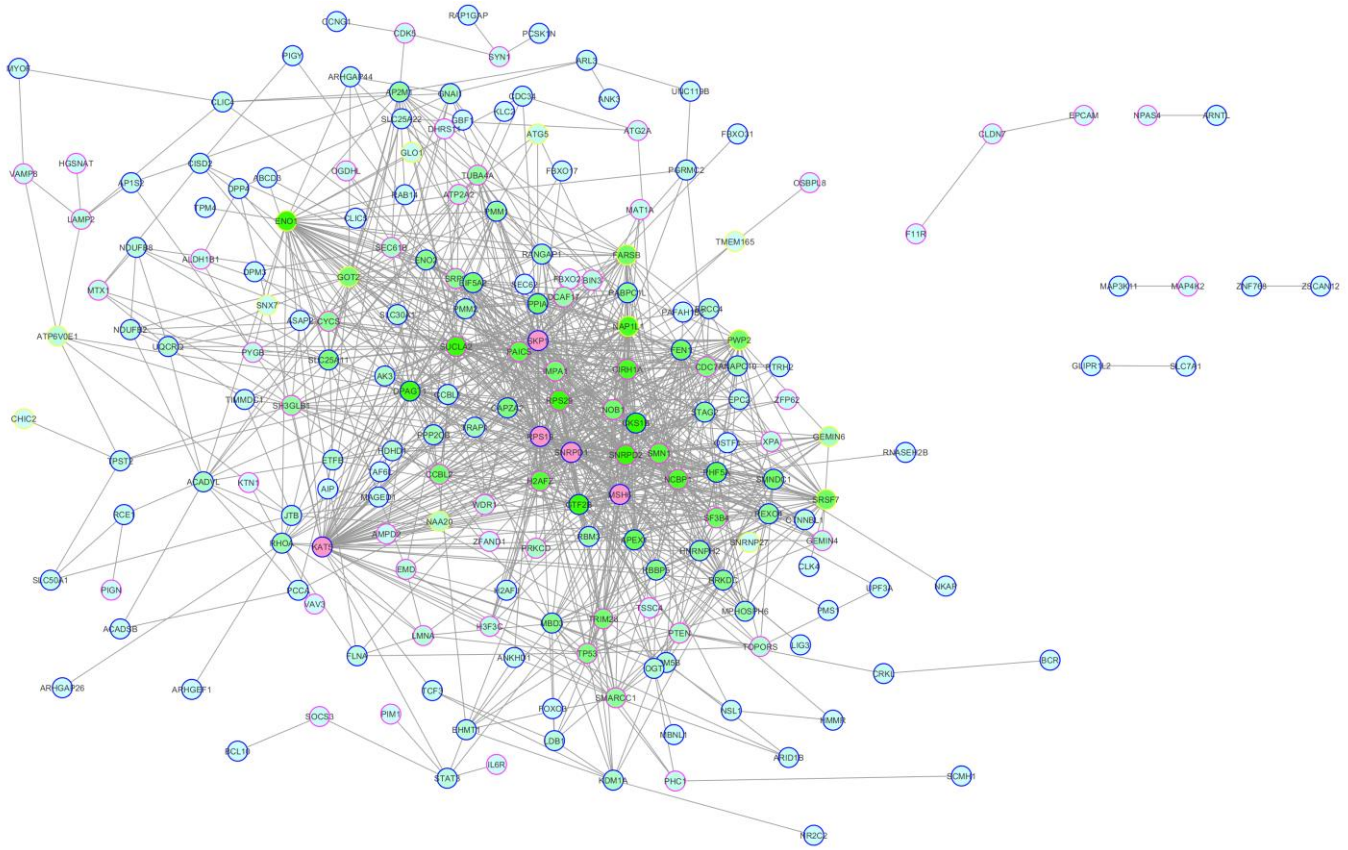


*Figure 4.34. Protein interaction network for the human proliferation GOI list, which includes the T7 and T14 GOI lists. Network was calculated using FunCoup and visualized with Cytoscape. Nodes represent genes and are colour-coded for the number of genes they interact with in the GOI list. Green nodes correspond to genes with higher number of interactions than those in blue. Pink circles represent genes with more than 50 interactors in the sample.*

#### *4.3.12.6 Hyperconnected human GOI*

To complement the network analysis performed with the human GOI, the FunCoup results were used to look for GOI which had more connections in the GOI list than expected by chance. The approach implemented in section 3.3.10.3 and used with the mouse GOI list in this screen in section 4.3.12.3 was adapted to the data.

Although the principle of the analysis and therefore the script used remained unchanged, the GOI list in this case is formed of human gene names. Therefore, I built an interaction database for pairs of human genes similar to the mouse database in section 3.3.10.3. Briefly, the FunCoup database of interactions for human genes was downloaded from http://funcoup.sbc.su.se/downloads/. A more compact file was produced by the removal of any interactions with a confidence score lower than 0.8, the threshold used for building the networks in FunCoup. The gene names in this file are expressed as Ensembl IDs. As the RNA-seq data used to filter the screen results for expression in human β-cells already contains the Ensembl ID, no further conversion was necessary.

The results of this random sampling approach indicate that in the T7 GOI list there are 14 genes which are hyperconnected in the network produced using FunCoup (Table 4.29). The same analysis of the T14 network produced a list of seven hyperconnected genes, available in Table 4.30. Following the analysis performed on the mouse GOI lists, the combined T7T14 network was also searched for hyperconnected nodes. In this case 28 genes were found to have a p-value for enrichment in connections below 0.05 and were therefore considered enriched or hyperconnected (Table 4.31). The same phenomenon of previously not-hyperconnected genes now appearing as hyperconnected is seen using the human β-cell expressed genes. Only half of the hyperconnected genes in the combined list were found to be hyperconnected in their respective individual lists (Figure 4.35). Three genes from the T7 hyperconnected list lost their hyperconnection status when transferred to the combined list: *Arl3*, *Pcca* and *Nr2c2*. From the T14 GOI list network, four hyperconnected genes were not found to be enriched in the combined list: F11R, PYGB, GEMIN6 and HGSNAT as above.

| Gene | Screened interactors | Interactors in GOI | P-value |
|---|---|---|---|
| *Fbxo31* | 2 | 1 | 0.002 |
| *Slc50a1* | 16 | 3 | 0.002 |
| *Ank3* | 3 | 1 | 0.003 |
| *Zscan12* | 2 | 1 | 0.004 |
| *Mbnl1* | 3 | 1 | 0.007 |
| *Acadsb* | 11 | 2 | 0.009 |
| *Maged1* | 11 | 2 | 0.010 |
| *Hmmr* | 10 | 2 | 0.011 |
| *Znf768* | 4 | 1 | 0.013 |
| *Rce1* | 12 | 2 | 0.015 |
| *Arid1b* | 13 | 2 | 0.027 |
| *Arl3* | 34 | 4 | 0.035 |
| *Pcca* | 38 | 4 | 0.043 |
| *Nr2c2* | 8 | 1 | 0.044 |

*Table 4.29. List of hyperconnected genes (p-value <0.05) in the T7 GOI list as determined using a random sampling approach.*

| Gene | Screened interactors | Interactors in GOI | P-value |
|---|---|---|---|
| *Cldn7* | 7 | 2 | <0.001 |
| *Dhrs11* | 4 | 1 | 0.004 |
| *Epcam* | 4 | 1 | 0.004 |
| *F11r* | 6 | 1 | 0.014 |
| *Pygb* | 60 | 4 | 0.028 |
| *Gemin6* | 105 | 6 | 0.035 |
| *Hgsnat* | 13 | 1 | 0.047 |

*Table 4.30. List of hyperconnected genes (p-value <0.05) in the T14 GOI list as determined using a random sampling approach.*

| Gene | Screened interactors | Interactors in GOI | P-value |
|---|---|---|---|
| *Fbxo27* | 1 | 1 | <0.001 |
| *Il6r* | 1 | 1 | <0.001 |
| *Npas4* | 1 | 1 | <0.001 |
| *Socs3* | 5 | 2 | 0.002 |
| *Fbxo31* | 2 | 1 | 0.003 |
| *Rce1* | 12 | 3 | 0.003 |
| *Arid1b* | 13 | 3 | 0.005 |
| *Maged1* | 11 | 3 | 0.005 |
| *Pcsk1n* | 2 | 1 | 0.005 |
| *Zscan12* | 2 | 1 | 0.006 |
| *Mbnl1* | 3 | 1 | 0.011 |
| *Slc50a1* | 16 | 3 | 0.011 |
| *Map4k2* | 3 | 1 | 0.012 |
| *Cldn7* | 7 | 2 | 0.013 |
| *Ank3* | 3 | 1 | 0.014 |
| *Dhrs11* | 4 | 1 | 0.015 |
| *Rap1gap* | 3 | 1 | 0.016 |
| *Pim1* | 3 | 1 | 0.018 |
| *Znf768* | 4 | 1 | 0.02 |
| *Hmmr* | 10 | 2 | 0.024 |
| *Pten* | 78 | 10 | 0.028 |
| *Acadsb* | 11 | 2 | 0.029 |
| *Vav3* | 5 | 1 | 0.035 |
| *Ankhd1* | 18 | 3 | 0.041 |
| *Epcam* | 4 | 1 | 0.044 |
| Myof | 15 | 2 | 0.044 |
| *Scmh1* | 5 | 1 | 0.045 |
| *Tssc4* | 12 | 2 | 0.048 |

*Table 4.31. List of hyperconnected genes in the combined GOI list.*

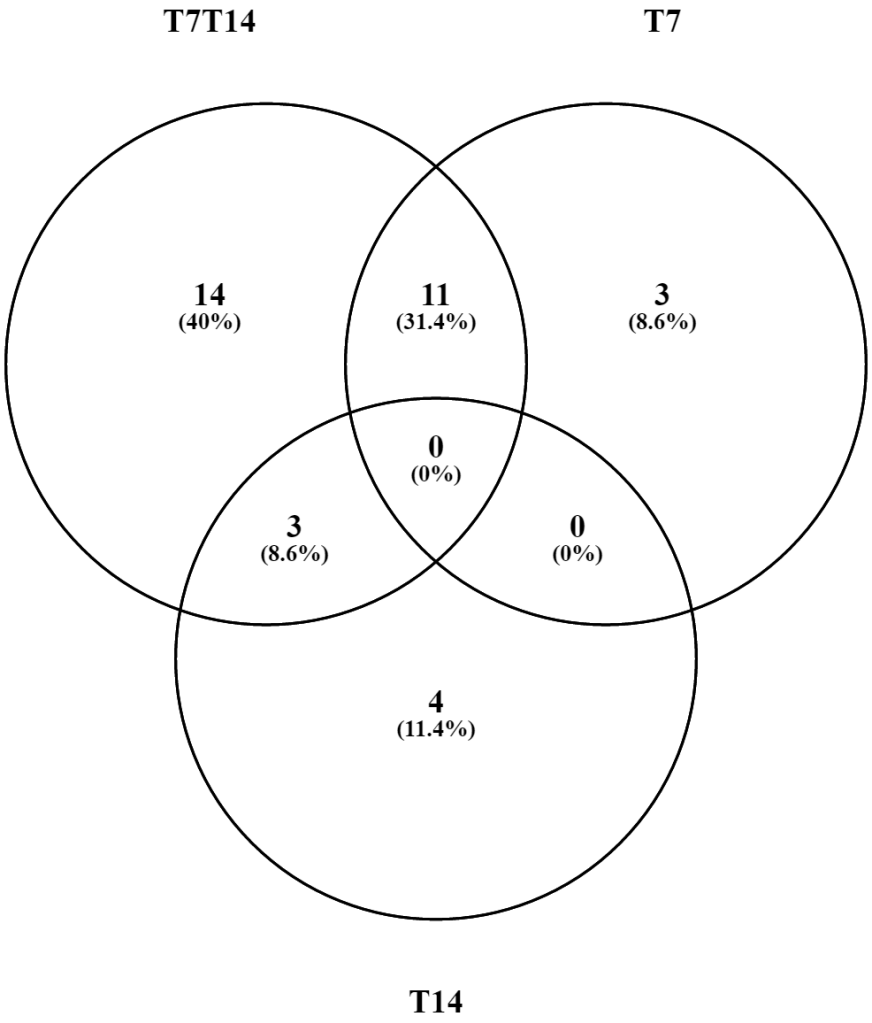



*Figure 4.35. Venn diagram representing the overlap amongst the hyperconnected gene lists for both individual proliferation networks and the combined network*

### 4.3.13. Network analysis of genes potentially involved in β-cell resistance to death by lipids

The approach used in 4.3.12 was applied to the resistance to death by lipids screen to identify essential genes or pathways in the phenotype studied. STRING and FunCoup were used to produce gene networks using the mouse dataset and the data after filtering for expression in human β-cells.

#### *4.3.13.1 Network analysis of the mouse GOI using STRING*

The GOI list obtained in 4.3.10, genes involved in protection from palmitate induced death, was used to produce a protein network using STRING. The resulting network is composed of 56 genes connected by 57 genes. According to STRING, this network (Figure 4.36) is enriched in interactions with a p-value of 0.00336. The network is simpler than the networks produced for the proliferation samples in 4.3.12.1, which could be an effect of fewer GOIs being identified in the death screen. However, the T14 GOI list is composed of only five genes more than the death screen GOI list and its network is formed of 70 genes connected by 100 edges.

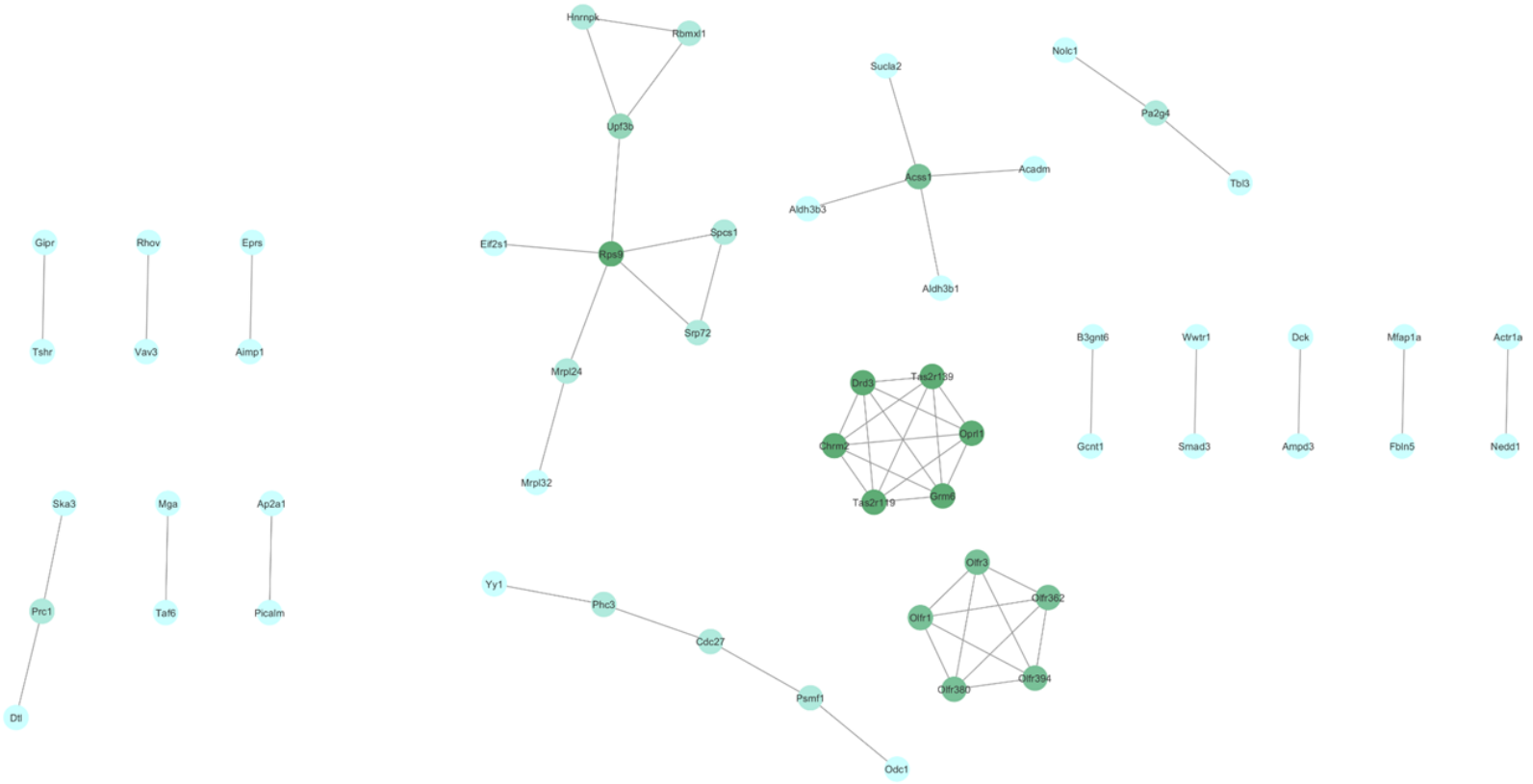



*Figure 4.36. Protein interaction network for the mouse GOI list obtained from the resistance to death by lipids screen.*
*The network was produced using STRING and visualized using Cytoscape. Genes are colour-coded according to the number of interactors they have in the GOI list. Genes in green tones have more interactors than genes in blue.*

#### *4.3.13.2 Network analysis of the mouse GOI using FunCoup*

The mouse GOI list produced for the resistance to death by lipids was used to produce a network using FunCoup (Figure 4.37). This network is formed of 79 genes connected by 288 interactions. Networks produced by FunCoup in this screen and the screen in Chapter 3 all contain more genes and connections than the same networks drawn by STRING. However, the death screen network is especially enriched in edges in comparison. This network in STRING was compared in 4.3.13.1 with the network for T14 as the number of GOI is similar to the death screen. The same comparison using the FunCoup network indicates that although the number of genes is higher in the T14 network (90), the genes in the death screen network are more highly connected with 288 edges compared to 253 in T14.

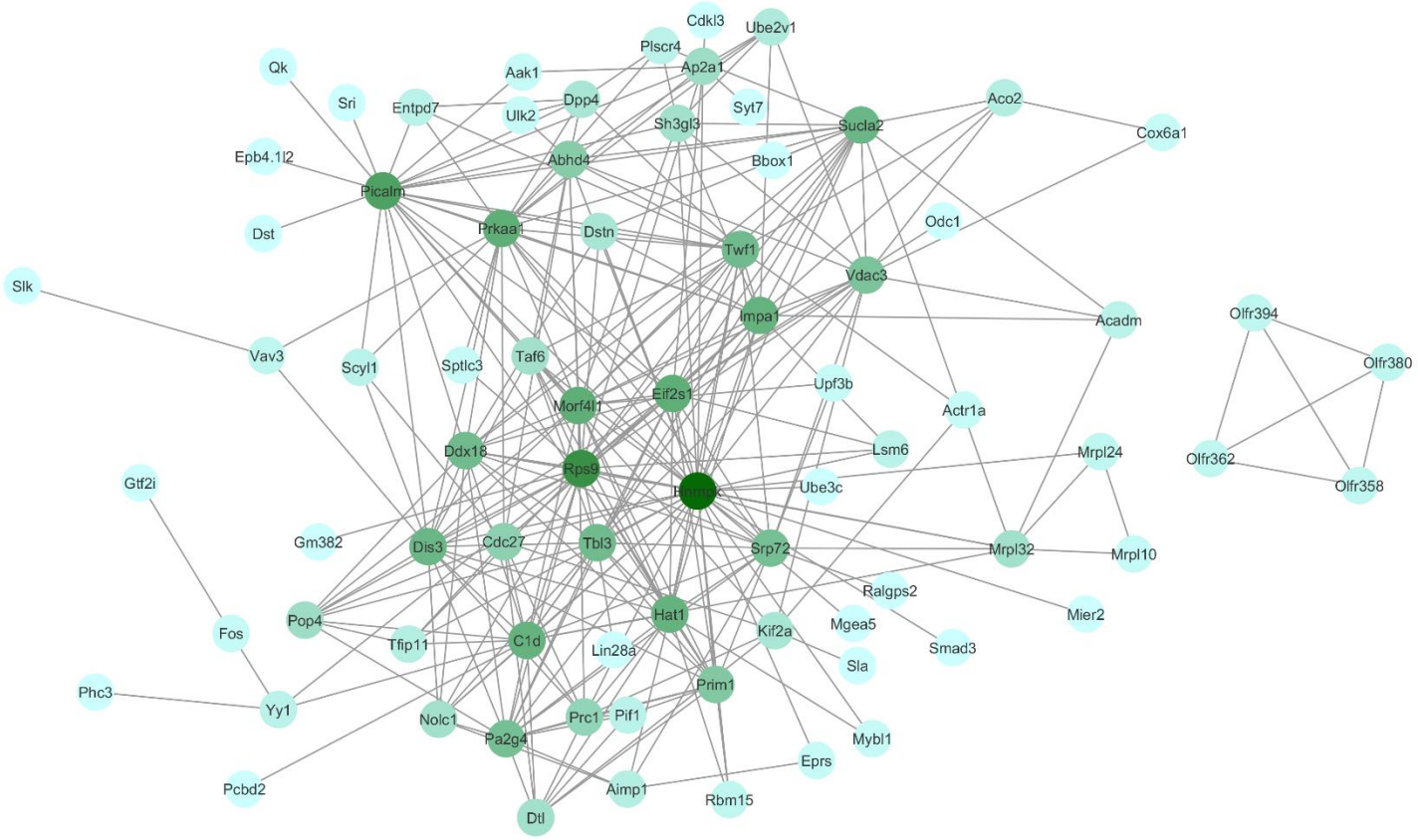



*Figure 4.37. Protein interaction network for the mouse GOI list obtained from the resistance to death by lipids screen.*
*The network was produced using FunCoup and visualized using Cytoscape. Genes are colour-coded according to the number of interactors they have in the GOI list. Genes in green tones have more interactors than genes in blue.*

#### *4.3.13.3 Hyperconnected mouse GOI*

The same approach explained in 4.3.12.3 was used to produce a list of hyperconnected genes in the mouse network. Only one gene was determined as being hyperconnected with a p-value below 0.05: Vav3.

#### *4.3.13.4 Network analysis of the human filtered GOI using STRING*

The death screen GOI list was filtered for expression in human β-cells as described in 4.3.10. This filtered GOI list was used to construct a gene interaction network in STRING as previously described. The resulting network is formed of 46 nodes connected by 28 edge, which STRING determined as not enriched in gene interactions. The network as visualized using Cytoscape is available in Figure 4.38.

#### *4.3.13.5 Network analysis of the human filtered GOI using FunCoup*

The filtered GOI list used with STRING in section 4.3.13.4 was used to construct a network in FunCoup. The network was visualized with Cytoscape and is available in Figure 4.39.

#### *4.3.13.6 Hyperconnected human GOI*

The human FunCoup network was examined for hyperconnected genes using the random sampling approach as described in previous sections. Only three genes were found to be significantly enriched in connections in this network: *AAK1*, *CHRM3* and *PCBD2.*

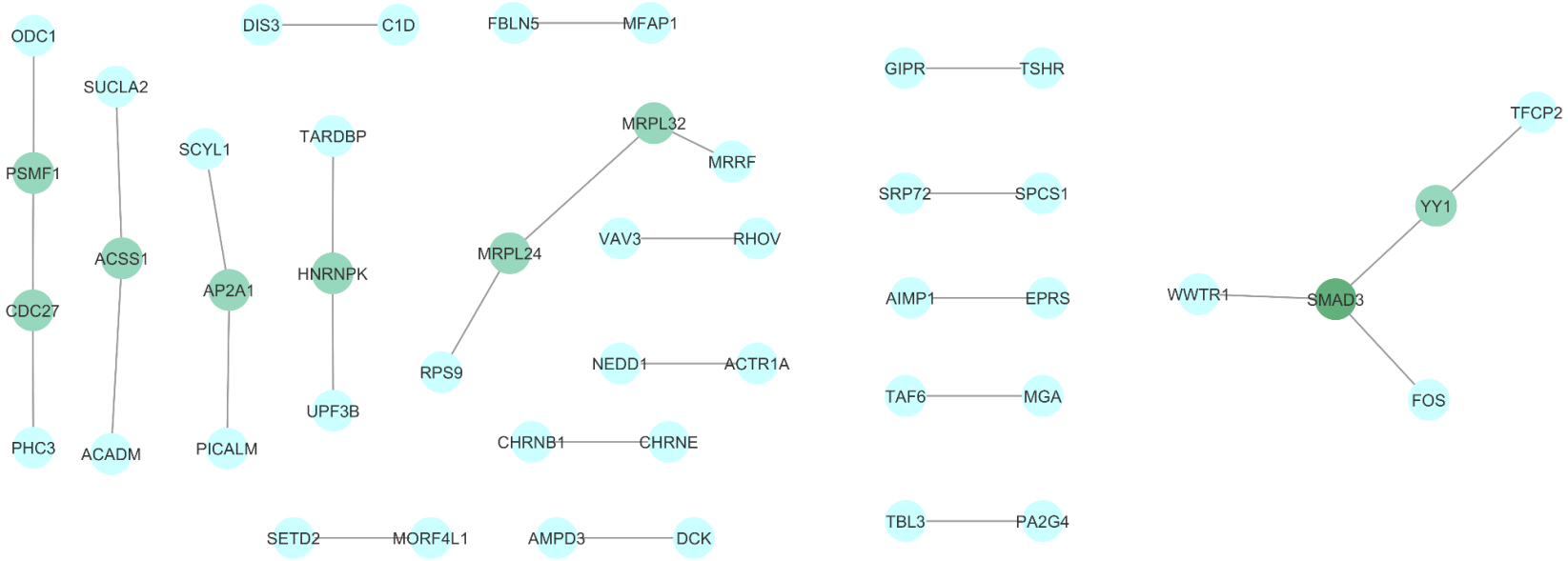



*Figure 4.38. Protein interaction network of the human filtered GOI list for the death screen.*
*The network was produced using STRING and visualized in Cytoscape. Genes are colour-coded according to the number of interactors in the GOI list. Genes in green have more interactors in the list than genes in blue*

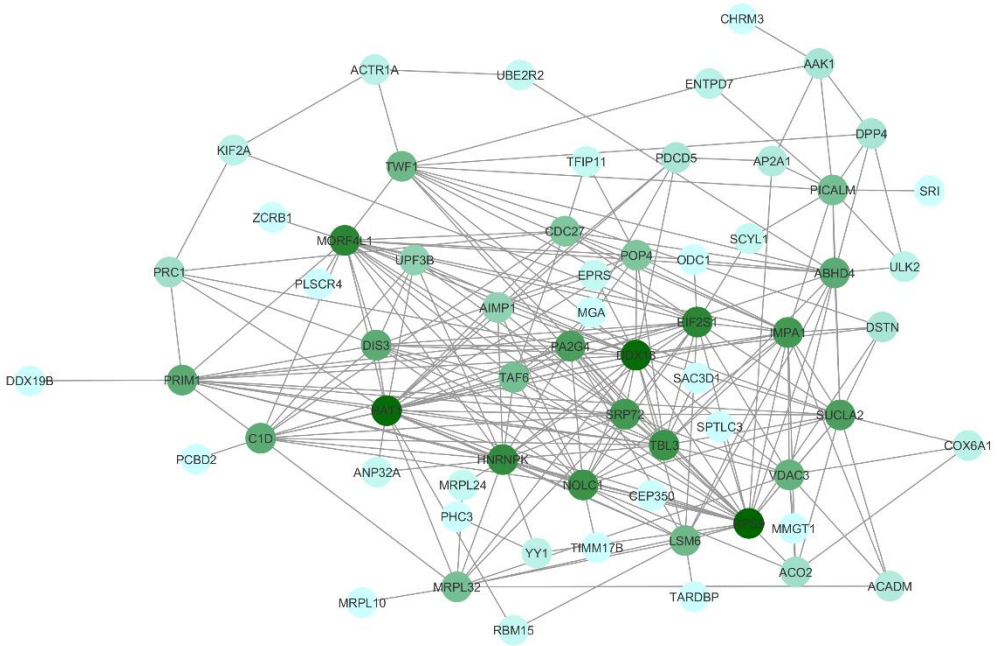



*Figure 4.39. Protein interaction network of the human filtered GOI list for the death screen.*
*The network was produced using FunCoup and visualized in Cytoscape. Genes are colour-coded according to the number of interactors in the GOI list. Genes in green have more interactors in the list than genes in blue.*

### 4.3.14. Functional enrichment analysis of GOI

The reasons to implement a functional enrichment analysis approach into the analysis of the screens have been described in section 3.3.11. In the case of the screening process in this chapter, it was hypothesized that the functional enrichment analysis of the GOI lists could highlight novel pathways or mechanisms involved in β-cell proliferation or in the toxicity of palmitate.

The approach used for the analysis of the GOI lists in the screen in Chapter 3 was followed for the human GOI lists. Each of the analyses was performed in both the proliferation and the death screens. The results for both screens are indicated in the following sub-sections.

#### *4.3.14.1. Functional enrichment analysis of the proliferation human GOI list using STRING*

The GOI lists for the T7 and T14 samples were filtered for genes expressed in human β-cells as previously explained. The enrichment analysis produced by STRING when the networks in section 4.3.12.4 were constructed revealed a total of 94 functional categories significantly enriched in the T7 human GOI list. From those, 40 are biological processes, 11 molecular functions and 42 cellular component terms.

In the T14 human GOI list no enrichment was found in the biological process GO category and only 13 terms (ten cellular components and three molecular functions) were determined as being enriched.

The combined list was determined to be enriched in 119 biological processes, 30 molecular functions and 43 cellular components. The top 15 enriched GO terms are available in Table 4.34.

The overlap among the different human proliferation lists is available in Figure 4.40.

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0044422 | organelle part | 179 | $2.86x10^{-10}$ |
| GO.0044446 | intracellular organelle part | 174 | $8.51x10^{-10}$ |
| GO.0044444 | cytoplasmic part | 167 | $4.20x10^{-09}$ |
| GO.0005622 | intracellular | 245 | $1.46x10^{-08}$ |
| GO.0044424 | intracellular part | 240 | $1.83x10^{-08}$ |
| GO.0005737 | cytoplasm | 203 | $6.57x10^{-08}$ |
| GO.0043229 | intracellular organelle | 217 | $3.01x10^{-07}$ |
| GO.0043231 | intracellular membrane-bounded organelle | 204 | $9.49x10^{-07}$ |
| GO.0005829 | cytosol | 88 | $1.09x10^{-06}$ |
| GO.0043227 | membrane-bounded organelle | 219 | $1.09x10^{-06}$ |
| GO.0043226 | organelle | 226 | $2.00x10^{06}$ |
| GO.0005488 | binding | 208 | $2.22x10^{-0}$ |
| GO.0005623 | cell | 250 | $2.23x10^{-05}$ |
| GO.0044464 | cell part | 248 | $5.34x10^{05}$ |
| GO.0044428 | nuclear part | 90 | 0.000106 |

*Table 4.32. Top 15 enriched GO terms in the T7 proliferation human GOI list ordered by their false discovery rate (FDR) according to STRING*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0043167 | ion binding | 87 | 0.000966 |
| GO.0005488 | binding | 122 | 0.00393 |
| GO.0003674 | molecular_function | 138 | 0.0107 |
| GO.0034719 | SMN-Sm protein complex | 4 | 0.00997 |
| GO.0043234 | protein complex | 58 | 0.00997 |
| GO.0098805 | whole membrane | 35 | 0.00997 |
| GO.0030670 | phagocytic vesicle membrane | 5 | 0.0302 |
| GO.0032797 | SMN complex | 3 | 0.0302 |
| GO.0045335 | phagocytic vesicle | 6 | 0.0302 |
| GO.0097504 | Gemini of coiled bodies | 3 | 0.0302 |
| GO.0030666 | endocytic vesicle membrane | 7 | 0.0328 |
| GO.0005737 | cytoplasm | 111 | 0.0357 |
| GO.0016604 | nuclear body | 11 | 0.0415 |

*Table 4.33. Enriched GO terms in the T14 proliferation human GOI list ordered by their false discovery rate (FDR) according to STRING*

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0005488 | binding | 308 | $7.01 \times 10^{-11}$ |
| GO.0044422 | organelle part | 246 | $3.85 \times 10^{-10}$ |
| GO.0005737 | cytoplasm | 294 | $4.35 \times 10^{-10}$ |
| GO.0044444 | cytoplasmic part | 233 | $6.26 \times 10^{-10}$ |
| GO.0044446 | intracellular organelle part | 239 | $6.66 \times 10^{-10}$ |
| GO.0005622 | intracellular | 350 | $1.09 \times 10^{-09}$ |
| GO.0044424 | intracellular part | 343 | $1.15 \times 10^{-09}$ |
| GO.0043227 | membrane-bounded organelle | 318 | $9.99 \times 10^{-09}$ |
| GO.0003674 | molecular_function | 344 | $2.08 \times 10^{-08}$ |
| GO.0005829 | cytosol | 121 | $2.36 \times 10^{-07}$ |
| GO.0043226 | organelle | 323 | $4.97 \times 10^{-07}$ |
| GO.0043229 | intracellular organelle | 302 | $1.69 \times 10^{-06}$ |
| GO.0043231 | intracellular membrane-bounded organelle | 285 | $1.95 \times 10^{-06}$ |
| GO.0005515 | protein binding | 157 | $2.31 \times 10^{-06}$ |
| GO.0043234 | protein complex | 135 | $2.40 \times 10^{-06}$ |

*Table 4.34. Top 15 enriched GO terms in the combined proliferation GOI list ordered by false discovery rate (FDR), as determined by STRING.*

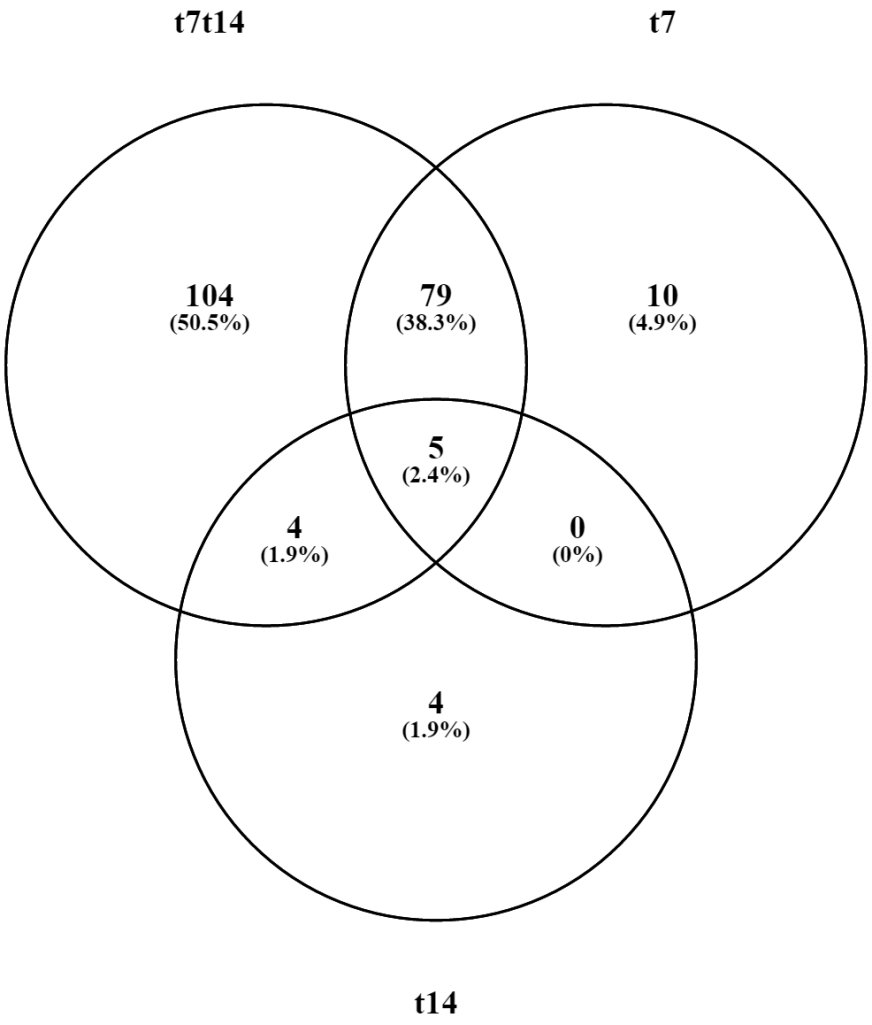



*Figure 4.40. Venn diagram representing the overlap among the proliferation human GOI lists (t7 and t14) and the combination of both lists (t7t14) in enriched functional categories according to STRING.*

#### *4.3.14.2. Functional enrichment analysis of the human resistance to death by lipids GOI list using STRING*

The functional enrichment analysis performed by STRING on the resistance to death by lipids GOI list after filtering for genes expressed in human β-cells indicates that only one biological process and 17 cellular components were enriched (Table 4.35). The lack of enriched functional categories may be a result of the lack of knowledge of the specific mechanism of toxicity induced by palmitate in β-cells. If a known apoptotic process was used this should have been identified, particularly effector pathways. This absence of enrichment may therefore indicate that the apoptotic process involved in palmitate toxicity is not shared with other stimuli for which the effector pathways have been characterised.

| Pathway ID | Pathway description | Observed gene count | FDR |
|---|---|---|---|
| GO.0043227 | membrane-bounded organelle | 136 | $4.48x10^{-05}$ |
| GO.0005622 | intracellular | 146 | $4.75x10^{-05}$ |
| GO.0043231 | intracellular membrane-bounded organelle | 124 | 0.000166 |
| GO.0043226 | organelle | 136 | 0.000367 |
| GO.0044424 | intracellular part | 140 | 0.000367 |
| GO.0005634 | nucleus | 90 | 0.000536 |
| GO.0005737 | cytoplasm | 118 | 0.000588 |
| GO.0043229 | intracellular organelle | 127 | 0.000794 |
| GO.0070013 | intracellular organelle lumen | 60 | 0.00273 |
| GO.0031981 | nuclear lumen | 52 | 0.00399 |
| GO.0005654 | nucleoplasm | 44 | 0.0162 |
| GO.0044428 | nuclear part | 53 | 0.0162 |
| GO.0044446 | intracellular organelle part | 90 | 0.0193 |
| GO.0005623 | cell | 144 | 0.0299 |
| GO.0044422 | organelle part | 90 | 0.0398 |

*Table 4.35. Top 15 enriched GO terms in the human resistance to death by lipids GOI list ordered by FDR, according to STRING.*

#### *4.3.14.3. Functional enrichment analysis of the human GOI lists using FunCoup*

FunCoup performs a basic functional enrichment analysis using GO categories when it calculates a protein network. Results of the functional enrichment analysis were obtained when the different networks were produced. However, similarly to the analysis in Chapter 3, no categories were found to be enriched with a q-value below 0.05 in any of the human GOI lists.

#### *4.3.14.4. Functional enrichment analysis of the human proliferation GOI lists using DAVID*

The use of the on-line resource DAVID to perform functional enrichment analyses has been previously discussed in section 3.3.11.3. Functional enrichment analysis using DAVID was performed on the proliferation GOI lists (T7, T14 and combined T7T14) after filtering for genes expressed in human β-cells.

The whole genome of Homo sapiens was used as the background for the enrichment analysis resulting in the same limitations for this analysis method discussed in Chapter 3.

Analysis of the T7 human GOI list resulted in one functional term being determined as enriched with an FDR under 0.05: “phosphoprotein”. Using a p-value threshold of 0.05 to determine enrichment, 136 categories are enriched. The top 10 categories with lower Benjamini-adjusted p-value are listed in Table 4.36.

In the T14 GOI list no categories were found to be enriched with an FDR lower than 0.05 and 144 categories were enriched with a p-value under 0.05. Table 4.37 contains the top 10 categories as ranked by their Benjamini-adjusted p-value.

Analysis of the combined proliferation GOI list indicated that four categories were enriched with an FDR below 0.05 and 245 had a p-value under 0.05. The 10 categories with lower Benjamini-adjusted p-value are available on Table 4.38.

| Category | Term | % of GOI in the term | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 56.42633 | $7.79x10^{-12}$ | 1.499069 | $2.60x10^{-09}$ |
| SP_PIR_KEYWORDS | cytoplasm | 25.07837 | $3.63x10^{-04}$ | 1.452279 | 0.029866 |
| SP_PIR_KEYWORDS | nucleus | 31.03448 | $2.02x10^{-04}$ | 1.398145 | 0.033191 |
| SP_PIR_KEYWORDS | GTPase activation | 3.448276 | $3.14x10^{-04}$ | 4.184664 | 0.034386 |
| SP_PIR_KEYWORDS | transferase | 12.22571 | 0.001553 | 1.692259 | 0.098608 |
| SMART | SM00233:PH | 4.702194 | 0.001019 | 2.762037 | 0.146121 |
| SP_PIR_KEYWORDS | disease mutation | 13.16614 | 0.002837 | 1.596777 | 0.146257 |
| SP_PIR_KEYWORDS | transport | 13.47962 | 0.003933 | 1.557461 | 0.171423 |
| SP_PIR_KEYWORDS | DNA damage | 3.134796 | 0.007101 | 2.950606 | 0.232371 |
| SP_PIR_KEYWORDS | magnesium | 5.015674 | 0.006774 | 2.18959 | 0.247073 |

*Table 4.36. Top 10 categories in the functional enrichment analysis of the T7 human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value.*

| Category | Term | % of GOI in the term | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 51.5625 | $8.83x10^{-05}$ | 1.365558 | 0.025381 |
| SP_PIR_KEYWORDS | metal-binding | 25 | $6.28x10^{-04}$ | 1.618018 | 0.059114 |
| SP_PIR_KEYWORDS | disease mutation | 16.14583 | $4.87x10^{-04}$ | 1.952012 | 0.068371 |
| SP_PIR_KEYWORDS | transferase | 14.0625 | 0.001396 | 1.940403 | 0.096642 |
| SP_PIR_KEYWORDS | acetylation | 21.875 | 0.00207 | 1.596833 | 0.113622 |
| SP_PIR_KEYWORDS | serine/threonine-protein kinase | 5.729167 | 0.004889 | 2.892402 | 0.184322 |
| SP_PIR_KEYWORDS | lyase | 3.125 | 0.006492 | 5.094015 | 0.189886 |
| SP_PIR_KEYWORDS | zinc-finger | 15.10417 | 0.006232 | 1.691086 | 0.203396 |
| SP_PIR_KEYWORDS | prenylation | 3.645833 | 0.004695 | 4.495359 | 0.20407 |
| SP_PIR_KEYWORDS | zinc | 17.70833 | 0.009498 | 1.556052 | 0.22312 |

*Table 4.37. Top 10 categories in the functional enrichment analysis of the T14 human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value.*

| Category | Term | % of GOI in the term | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 5.01258 | $4.89x10^{-14}$ | 1.447097 | $2.02x10^{-11}$ |
| SP_PIR_KEYWORDS | disease mutation | 1.316044 | $9.67x10^{-06}$ | 1.734413 | 0.00199 |
| SP_PIR_KEYWORDS | transferase | 1.161215 | $3.02x10^{-05}$ | 1.746636 | 0.00414 |
| SP_PIR_KEYWORDS | GTPase activation | 0.27095 | $1.58x10^{-04}$ | 3.573097 | 0.010791 |
| SP_PIR_KEYWORDS | acetylation | 1.838591 | $1.06x10^{-04}$ | 1.463042 | 0.010908 |
| SP_PIR_KEYWORDS | nucleus | 2.690149 | $2.33x10^{-04}$ | 1.316984 | 0.011949 |
| SP_PIR_KEYWORDS | serine/threonine-protein kinase | 0.445133 | $2.05x10^{-04}$ | 2.449721 | 0.011994 |
| SP_PIR_KEYWORDS | cytoplasm | 2.206309 | 1.48E-04 | 1.388397 | 0.012109 |
| SP_PIR_KEYWORDS | nucleotide-binding | 1.257983 | 3.18E-04 | 1.564479 | 0.01444 |
| SMART | SM00233:PH | 0.406425 | 1.61E-04 | 2.597357 | 0.029749 |

*Table 4.38. Top 10 categories in the functional enrichment analysis of the combined proliferation human GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value.*

*4.3.14.5. Functional enrichment analysis of the human resistance to lipotoxicity GOI list using DAVID.*

The human GOI list for resistance to lipotoxicity was found to be enriched in two functional categories with an FDR under 0.05 and 48 with a p-value below 0.05. The list of the ten functional terms with the lowest Benjamini-adjusted p-values is available on Table 4.39.

| Category | Term | % of GOI in the term | P-Value | Fold Enrichment | Benjamini |
|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | phosphoprotein | 53.36787565 | $7.32x10^{-06}$ | 1.420731935 | 0.00201734 |
| SP_PIR_KEYWORDS | acetylation | 24.87046632 | $3.56x10^{-05}$ | 1.824952562 | 0.004895105 |
| SP_PIR_KEYWORDS | nucleus | 34.19689119 | $1.22x10^{-04}$ | 1.543785022 | 0.011182769 |
| SP_PIR_KEYWORDS | rna-binding | 8.29015544 | $3.36x10^{-04}$ | 2.968364198 | 0.022903168 |
| SP_PIR_KEYWORDS | transcription regulation | 17.61658031 | 0.002972563 | 1.681242802 | 0.151537876 |
| SP_PIR_KEYWORDS | Transcription | 17.61658031 | 0.004182524 | 1.644711693 | 0.175352749 |
| SP_PIR_KEYWORDS | alternative splicing | 47.66839378 | 0.008480025 | 1.230872173 | 0.285220373 |
| SP_PIR_KEYWORDS | cytoplasm | 24.35233161 | 0.010602688 | 1.413135567 | 0.307706813 |
| GOTERM_BP_FAT | GO:0016072~rRNA metabolic process | 3.626943005 | 9.01E-04 | 6.203878407 | 0.393508302 |
| SP_PIR_KEYWORDS | rrna processing | 2.07253886 | 0.022769176 | 6.569330601 | 0.470432235 |

*Table 4.39. Top 10 categories in the functional enrichment analysis of the human resistance to lipotoxicity GOI list performed using DAVID, ranked by their Benjamini-adjusted p-value.*

#### *4.3.14.6. Functional enrichment analysis using FunRich*

The use of FunRich as a tool to analyse a list of genes for functional enrichment has been discussed in this thesis in section 3.3.11.4.

Only the human GOI lists (human gene names filtered for expression in human β-cells) were used to perform this analysis. Functional enrichment was considered in “cellular component”, “molecular function”, “biological process” and “transcription factor”.

In the T7 human GOI list 24 terms were found to be enriched with a Benjamini-adjusted p-value under 0.05. The 15 terms with the lowest Benjamini p-value are listed in Table 4.40. Only two of them are not transcription factors: “GTPase activator activity” and “Cytoplasm”. The other 13 indicate enrichment for genes with transcription factor binding sites for each of the transcription factors in the table.

Analysis of the T14 human GOI list found six functional terms enriched with a Benjamini adjusted p-value lower than 0.05, all of them transcription factor binding sites. Table 4.41 contains the 15 functional terms with the lowest Benjamini adjusted p-value.

In the combined list, 40 terms were found to be enriched with a Benjamini-adjusted p-value under 0.05, the majority are transcription factors. A list of the 15 functional terms with the lowest Benjamini adjusted p-value is available in Table 4.42.

The analysis of the human GOI list from the resistance to death by chronic exposure to lipids did not find any functional terms enriched with a Benjamini-adjusted p-value under 0.05. The 15 functional terms with the lowest Benjamini-adjusted p-values are listed in Table 4.43. As opposed to the lists for the proliferation GOI lists, transcription factors are not the main category of the enriched terms.

| Term | % of GOI in the term | Fold Enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| SP1 | 59.5959596 | 1.273943 | $6.32x10^{-06}$ | 0.001213 | 0.000118 |
| KLF7 | 39.73063973 | 1.36854 | $5.68x10^{-05}$ | 0.003636 | 0.000354 |
| SP4 | 39.05723906 | 1.377023 | $5.25x10^{-05}$ | 0.00504 | 0.000354 |
| NKX2-1 | 9.764309764 | 2.174731 | 0.0001381 | 0.006603 | 0.000643 |
| RAB40B | 10.43771044 | 2.030947 | 0.0002611 | 0.008333 | 0.000812 |
| RARA | 10.43771044 | 2.030947 | 0.00026 | 0.01 | 0.000812 |
| POU4F3 | 9.090909091 | 2.090137 | 0.000459 | 0.012598 | 0.001227 |
| ZNF238 | 9.764309764 | 1.928922 | 0.000976 | 0.017039 | 0.001659 |
| GTPase activator activity | 3.095975232 | 4.339353 | 0.000257 | 0.017986 | 0.004163 |
| ZNF513 | 9.764309764 | 1.928922 | 0.000976 | 0.018743 | 0.001659 |
| PLAU | 11.78451178 | 1.800485 | 0.000893 | 0.019053 | 0.001659 |
| Cytoplasm | 49.82578397 | 1.279078 | 0.000129 | 0.019399 | 0.003722 |
| ATF1 | 11.78451178 | 1.800485 | 0.000893 | 0.021434 | 0.001659 |
| NFYA | 23.23232323 | 1.423751 | 0.001647 | 0.026344 | 0.002566 |
| POU3F4 | 5.387205387 | 2.391388 | 0.002355 | 0.02826 | 0.002752 |

*Table 4.40. Top 15 functional terms (by Benjamini p-value) enriched in the T7 human GOI list. Analysis was performed using FunRich.*

| Term | % of GOI in the term | Fold Enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| TCF12 | 15.88235294 | 2.280834 | 8.64 x$10^{-05}$ | 0.015721 | 0.003287 |
| RARA | 12.35294118 | 2.440331 | 0.000272 | 0.016526 | 0.003455 |
| SP1 | 59.41176471 | 1.272722 | 0.000669 | 0.024363 | 0.005093 |
| RAB40B | 12.35294118 | 2.440331 | 0.000272 | 0.024789 | 0.003455 |
| NR4A2 | 14.11764706 | 2.152615 | 0.00056 | 0.025483 | 0.005093 |
| CTCF | 26.47058824 | 1.574659 | 0.001411 | 0.042793 | 0.008946 |
| Energy pathways | 15.34391534 | 1.754941 | 0.003399 | 0.067971 | 0.021331 |
| Transaminase activity | 1.587301587 | 16.5044 | 0.001752 | 0.09285 | 0.054785 |
| Metabolism | 14.81481481 | 1.645235 | 0.009657 | 0.096575 | 0.030307 |
| Carbohydrate mediated signalling | 0.529100529 | 65.1961 | 0.020989 | 0.104945 | 0.032933 |
| Cell communication | 26.45502646 | 1.308567 | 0.030231 | 0.120923 | 0.037948 |
| JUNB | 17.64705882 | 1.584272 | 0.011197 | 0.135863 | 0.028404 |
| Hemopoiesis | 0.529100529 | 40.35949 | 0.04154 | 0.138465 | 0.043453 |
| Reproductive behaviour | 0.529100529 | 65.1961 | 0.020989 | 0.139926 | 0.032933 |
| BARHL1 | 7.058823529 | 2.296262 | 0.012374 | 0.14076 | 0.029427 |

*Table 4.41. Top 15 functional terms (by Benjamini p-value) enriched in the T14 human GOI list. Analysis was performed using FunRich.*

| Term | % of GOI in the term | Fold Enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| SP1 | 60.32110092 | 1.288208125 | $8.12 \times 10^{-09}$ | $1.58 \times 10^{-06}$ | $2.63 \times 10^{-07}$ |
| KLF7 | 39.44954128 | 1.356197245 | $2.15 \times 10^{-06}$ | 0.000209 | $3.49 \times 10^{-05}$ |
| RAB40B | 10.55045872 | 2.029650963 | $7 \times 10^{-06}$ | 0.000339 | $5.67 \times 10^{-05}$ |
| RARA | 10.55045872 | 2.029650963 | $7 \times 10^{-06}$ | 0.000453 | $5.67 \times 10^{-05}$ |
| NKX2-1 | 9.174311927 | 2.022808482 | $3.36 \times 10^{-05}$ | 0.001088 | 0.000182 |
| SP4 | 37.3853211 | 1.315783094 | $3.19 \times 10^{-05}$ | 0.001238 | 0.000182 |
| ZNF238 | 9.633027523 | 1.881160888 | 0.000108 | 0.002615 | 0.000437 |
| ZNF513 | 9.633027523 | 1.881160888 | 0.000108 | 0.002989 | 0.000437 |
| PITX1 | 9.633027523 | 1.776311126 | 0.000367 | 0.007121 | 0.00119 |
| NR4A2 | 11.23853211 | 1.677806593 | 0.000407 | 0.007181 | 0.0012 |
| OTX1 | 9.633027523 | 1.776311126 | 0.000367 | 0.007912 | 0.00119 |
| GTPase activator activity | 2.725366876 | 3.726596475 | 0.00011 | 0.008925 | 0.003687 |
| Signal transduction | 28.1512605 | 1.299140477 | 0.000574 | 0.010043 | 0.002159 |
| Auxiliary transport protein activity | 4.402515723 | 2.493737491 | 0.000249 | 0.010078 | 0.004163 |
| Cell communication | 27.10084034 | 1.325740604 | 0.000332 | 0.011603 | 0.002159 |

*Table 4.42. Top 15 functional terms (by Benjamini p-value) enriched in the combined proliferation human GOI list. Analysis was performed using FunRich*

| Term | % of GOI in the term | Fold Enrichment | Unadjusted p-value | Benjamini adjusted p-value | Storey and Tibshirani q-value |
|---|---|---|---|---|---|
| Cell-cell signaling | 1.041666667 | 17.46714288 | 0.01242225 | 0.068322377 | 0.028557109 |
| Regulation of circadian rhythm | 0.520833333 | 92.71021992 | 0.010717276 | 0.078593358 | 0.028557109 |
| LHX4 | 7.514450867 | 3.165241267 | 0.000548949 | 0.104300322 | 0.044951126 |
| NRF1 | 21.96531792 | 1.684252985 | 0.001213177 | 0.115251795 | 0.049670968 |
| Secretory pathway | 0.520833333 | 92.71021992 | 0.010717276 | 0.117890036 | 0.028557109 |
| Spindle microtubule | 1.724137931 | 14.39888524 | 0.002572104 | 0.119602849 | 0.058177902 |
| Signal complex formation | 0.520833333 | 49.08188114 | 0.031810236 | 0.139965037 | 0.058502016 |
| Regulation of endocytosis | 0.520833333 | 39.7329514 | 0.04218833 | 0.154690545 | 0.064656924 |
| Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism | 22.39583333 | 1.477686531 | 0.007044393 | 0.154976652 | 0.028557109 |
| Aminoacyl-tRNA synthetase multienzyme complex | 0.574712644 | 82.15834522 | 0.012085851 | 0.160569166 | 0.078104972 |
| LHX3 | 14.45086705 | 1.88807312 | 0.002652585 | 0.167997041 | 0.072402999 |
| Invadopodium membrane | 0.574712644 | 82.15834522 | 0.012085851 | 0.187330694 | 0.078104972 |
| Nucleus | 48.27586207 | 1.21444344 | 0.016347063 | 0.190034613 | 0.092437723 |
| AP-2 adaptor complex | 1.149425287 | 36.82960303 | 0.002110082 | 0.196237662 | 0.058177902 |
| Signal peptidase complex | 0.574712644 | 82.15834522 | 0.012085851 | 0.224796833 | 0.078104972 |

*Table 4.43. Top 15 functional terms (by Benjamini p-value) enriched in the resistance to lipotoxicity human GOI list. Analysis was performed using FunRich*

#### *4.3.14.7. Gene set enrichment analysis of the human GOI lists*

The GSEA software has been previously described in this thesis in section 3.3.11.5. Briefly, the “Preranked” tool implemented in GSEA was used to analyse the enrichment in genes belonging to certain gene sets in the top positions of the list of screened genes. The genes were ranked by the phenotype of interest in each case. Similarly to the analyses produced in Chapter 3, several approaches were taken to produce gene set enrichment reports. As the method of removing duplicate gene names in the lists did not seem to make a big difference in the results, duplications were resolved by keeping the targeting sequence with the highest z-score. Therefore, the lists analysed were as follows:

1) Genes screened in the T7 proliferation sample, ranked by z-score in T7.

2) Genes screened in the T14 proliferation sample, ranked by z-score in T14.

3) Genes screened in the T14 and T7 proliferation samples, ranked by the sum of z-scores in both samples.

4) Genes screened in the resistance to lipotoxicity sample, ranked by z-score in P.

Enrichment was considered only in the top of the list, which corresponds to the highest z-scores in each list.

Analysis of the T7 ranked list using the Preranked tool identified 133 gene sets enriched in the top of the list. The 15 gene sets with the lowest p-values are listed in Table 4.44.

Using the T14 z-scores to rank the list of genes, 73 gene sets resulted enriched in the top of the list. Table 4.45 contains the 15 gene sets with the lowest p-values for enrichment according to GSEA.

In the list ranked by the sum of T7 and T14 z-scores, 89 gene sets were identified as being enriched in the top of the list. The 15 gene sets with the lowest p-values are available in Table 4.46.

A degree of overlap of the T7+T14 ranked list with the analysis of the lists ordered by the T7 or T14 z-scores was expected. An overlap is indeed

observed as depicted in Figure 4.41 although this is for a few gene sets. The relatively low percentage of gene sets that overlap between conditions may be due to the smaller sample size in T14, which leads to the list ranked by T7+T14 being shorter than the T7 ranked lists. Genes in the enriched gene sets might have not been included in the T7+T14 analysis due to not having a z-score available in the T14 sample.

In the list ranked by the resistance to lipotoxicity z-score, 93 gene sets were determined as enriched at the top of the list with a p-value under 0.05. The 15 gene sets with the lowest p-values are listed in Table 4.47.

| Name of the gene set | P-value |
|---|---|
| LI_WILMS_TUMOR_VS_FETAL_KIDNEY_2_UP | <0.001 |
| HONMA_DOCETAXEL_RESISTANCE | <0.001 |
| SAKAI_CHRONIC_HEPATITIS_VS_LIVER_CANCER_DN | <0.001 |
| AZARE_STAT3_TARGETS | 0.001931 |
| REACTOME_LATENT_INFECTION_OF_HOMO_SAPIENS_WITH_MYCOBACTERIUM_TUBERCULOSIS | 0.001969 |
| KEGG_PARKINSONS_DISEASE | 0.00198 |
| PUJANA_CHEK2_PCC_NETWORK | 0.00198 |
| YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_17 | 0.002024 |
| BIOCARTA_AKT_PATHWAY | 0.002058 |
| TARTE_PLASMA_CELL_VS_PLASMABLAST_DN | 0.004049 |
| SHEPARD_CRUSH_AND_BURN_MUTANT_DN | 0.005682 |
| VANHARANTA_UTERINE_FIBROID_WITH_7Q_DELETION_DN | 0.005941 |
| RIZ_ERYTHROID_DIFFERENTIATION | 0.005976 |
| YOSHIMURA_MAPK8_TARGETS_DN | 0.005976 |
| ALONSO_METASTASIS_EMT_UP | 0.006122 |

*Table 4.44. Top 15 gene sets enriched in the top of the T7 ranked list.*

| Name of gene set | P-value |
|---|---|
| GUTIERREZ_CHRONIC_LYMPHOCYTIC_LEUKEMIA_DN | <0.001 |
| MIKKELSEN_ES_LCP_WITH_H3K4ME3 | <0.001 |
| MIKKELSEN_IPS_LCP_WITH_H3K4ME3 | 0.001961 |
| PARK_HSC_MARKERS | 0.001969 |
| KEGG_VIBRIO_CHOLERAE_INFECTION | 0.001972 |
| LEE_BMP2_TARGETS_DN | 0.001996 |
| SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN | 0.003795 |
| BRACHAT_RESPONSE_TO_CAMPTOTHECIN_DN | 0.003802 |
| PID_P53_REGULATION_PATHWAY | 0.003854 |
| YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_17 | 0.003914 |
| BLUM_RESPONSE_TO_SALIRASIB_DN | 0.003953 |
| FEVR_CTNNB1_TARGETS_DN | 0.00396 |
| BIOCARTA_WNT_PATHWAY | 0.004049 |
| BIOCARTA_CHREBP2_PATHWAY | 0.004228 |
| FERRANDO_T_ALL_WITH_MLL_ENL_FUSION_DN | 0.005747 |

*Table 4.45. Top 15 gene sets enriched in the top of the T14 ranked list.*

| Name of the gene set | P-value |
|---|---|
| REACTOME_LATENT_INFECTION_OF_HOMO_SAPIENS_WITH_MYCOBACTERIUM_TUBERCULOSIS | <0.001 |
| SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6 | <0.001 |
| MODY_HIPPOCAMPUS_POSTNATAL | <0.001 |
| MIKKELSEN_ES_LCP_WITH_H3K4ME3 | <0.001 |
| PID_PTP1B_PATHWAY | <0.001 |
| REACTOME_IRON_UPTAKE_AND_TRANSPORT | 0.001942 |
| GROSS_HYPOXIA_VIA_HIF1A_UP | 0.001996 |
| CHIANG_LIVER_CANCER_SUBCLASS_INTERFERON_DN | 0.00202 |
| YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_17 | 0.002049 |
| RODWELL_AGING_KIDNEY_DN | 0.003992 |
| MIYAGAWA_TARGETS_OF_EWSR1_ETS_FUSIONS_DN | 0.006012 |
| CREIGHTON_AKT1_SIGNALING_VIA_MTOR_UP | 0.006073 |
| REACTOME_MITOCHONDRIAL_PROTEIN_IMPORT | 0.006135 |
| BRUINS_UVC_RESPONSE_VIA_TP53_GROUP_D | 0.007921 |
| PID_RAC1_PATHWAY | 0.008282 |

*Table 4.46. Top 15 gene sets enriched in the top of the T7+T14 ranked list.*

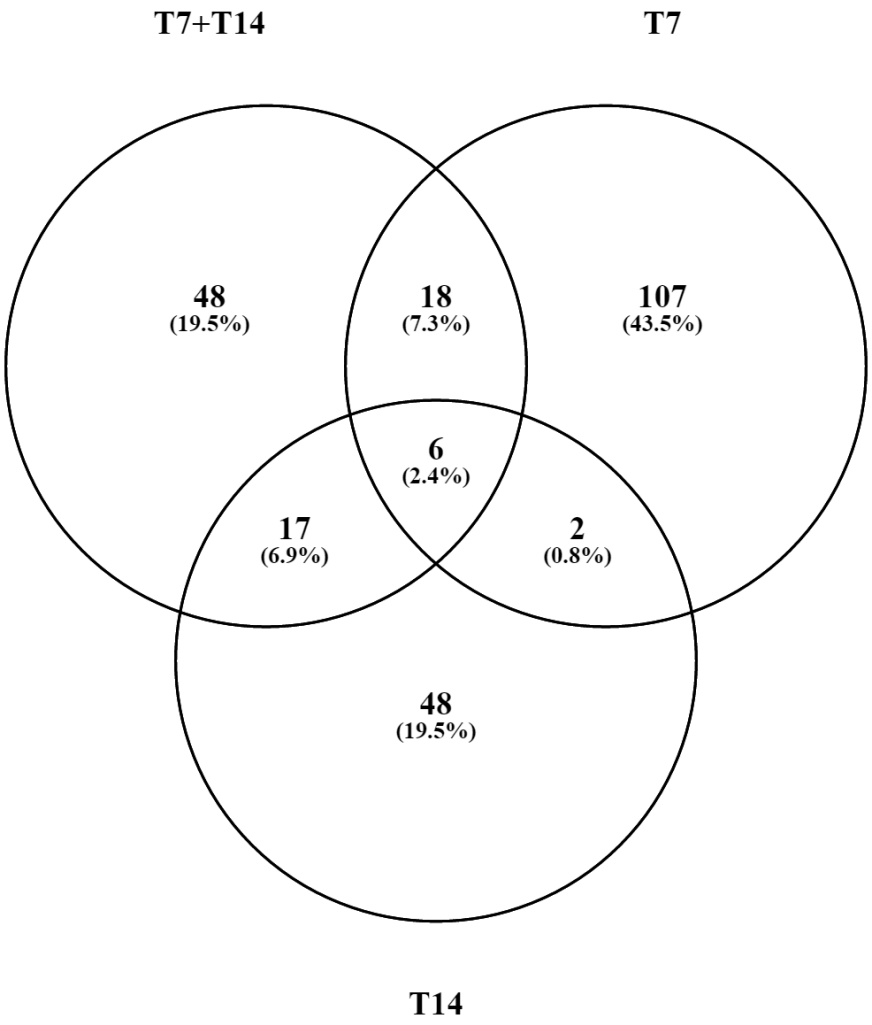



*Figure 4.41. Venn diagram representing the overlap in gene set enrichment in the lists ranked by the z-score in T7, T14 or T7+T14.*

| Name of gene set | P-value |
|---|---|
| BHATI_G2M_ARREST_BY_2METHOXYESTRADIOL_UP | <0.001 |
| DURCHDEWALD_SKIN_CARCINOGENESIS_DN | <0.001 |
| ZHANG_TLX_TARGETS_36HR_DN | <0.001 |
| RAO_BOUND_BY_SALL4_ISOFORM_B | <0.001 |
| BIDUS_METASTASIS_UP | <0.001 |
| HOFFMANN_LARGE_TO_SMALL_PRE_BII_LYMPHOCYTE_UP | 0.002 |
| IGARASHI_ATF4_TARGETS_DN | 0.002012 |
| GRADE_METASTASIS_DN | 0.002028 |
| HOFMANN_CELL_LYMPHOMA_UP | 0.002066 |
| BIOCARTA_ALK_PATHWAY | 0.002079 |
| BORCZUK_MALIGNANT_MESOTHELIOMA_UP | 0.002141 |
| MARCHINI_TRABECTEDIN_RESISTANCE_DN | 0.003883 |
| FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_OK_VS_DONOR_UP | 0.003914 |
| DACOSTA_UV_RESPONSE_VIA_ERCC3_DN | 0.004082 |
| HAMAI_APOPTOSIS_VIA_TRAIL_UP | 0.004246 |

*Table 4.47. Top 15 gene sets enriched in the palmitate z-score ranked list.*

### 4.3.15. Transcription factor binding site enrichment analysis.

The approach taken in section 3.3.12 was used to analyse the potential enrichment in transcription factor binding sites in the human GOI lists. The human single site analysis was used for this analyses, leaving all options were used as predetermined

The human T7 GOI list was analysed using as a background a list of the human homologues for the genes screened in T7 and expressed in human β-cells. The 10 transcription factor binding sites with the highest Fisher scores are listed in Table 4.48. As explained in section 3.3.12, OPPOSSUM recommends visualizing the distribution of the transcription factors z-scores against the GC composition of their binding sites as a clustering on any of the extremes (high or low GC content) may be an indicative of a bias in analysis produced by a differential composition in binding sites in the background used (the screened genes in this case). The website provides this graphical representation. For the T7 GOI list (Figure 4.42) there is no evident bias and the graphical representation indicates that the transcription factor with the highest Fisher score (INSM1) is separate from the rest of the transcription factors detected as potentially enriched.

Equivalently to the analysis performed in the T7 human GOI list, the human T14 GOI list was searched for enrichment using as a background the list of human homologues for genes screened in T14 which are expressed in human β-cells. Table 4.49 contains the ten transcription factors with the highest Fisher scores. The graphical representation obtained from OPOSSUM indicates no apparent bias in %GC content in the background of the GOI list, and RORA_1 is separate from the rest of the transcription factors.

The GOI list composed of GOI in both proliferation samples which are expressed in human β-cells was analysed using as a background the genes screened in any of the samples which were determined as being expressed in human β-cells. The results of this analysis are in Table 4.50. Fisher scores were plotted against %GC Figure 4.44 and no bias was apparent. RORA_1 is maintained separate from other transcription factors as in the T14 GOI list, indicating that it might be important for the control of proliferation-related genes.

The list of genes from the lipotoxicity GOI list which are expressed in human β-cells were examined for enrichment in transcription factor binding sites using the list of genes expressed in human β-cells which were screened. The top 10 transcription factors with the highest Fisher scores as determined by OPOSSUM are listed in Table 4.51. The graphical representation obtained from OPOSSUM indicates that there is no apparent bias in %GC in the background used, although the transcription factors with higher Fisher score are asymmetrically distributed towards the low end of the %GC distribution. HNF1B and Foxq1 are apparently separated from the main cluster of transcription factors, indicating they may potentially be regulating more genes than expected by chance.

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Fisher score | Z-score |
|---|---|---|---|---|---|---|---|---|
| INSM1 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 166 | 161 | 2929 | 3895 | 5.749 | -0.131 |
| NFYA | Other Alpha-Helix | NFY CCAAT-binding | 135 | 192 | 2370 | 4454 | 4.658 | 3.574 |
| HNF4A | Zinc-coordinating | Hormone-nuclear Receptor | 139 | 188 | 2453 | 4371 | 4.623 | 2.37 |
| RORA_1 | Zinc-coordinating | Hormone-nuclear Receptor | 175 | 152 | 3209 | 3615 | 4.374 | 5.029 |
| Zfp423 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 123 | 204 | 2192 | 4632 | 3.773 | 4.518 |
| Zfx | Zinc-coordinating | BetaBetaAlpha-zinc finger | 204 | 123 | 3872 | 2952 | 3.701 | -13.547 |
| Tcfcp2l1 | Other | CP2 | 205 | 122 | 3899 | 2925 | 3.631 | -8.037 |
| NFE2L2 | Zipper-Type | Leucine Zipper | 120 | 207 | 2151 | 4673 | 3.522 | 3.77 |
| CTCF | Zinc-coordinating | BetaBetaAlpha-zinc finger | 75 | 252 | 1276 | 5548 | 3.349 | 5.114 |
| USF1 | Zipper-Type | Helix-Loop-Helix | 187 | 140 | 3547 | 3277 | 3.299 | -2.825 |

*Table 4.48. Top 10 enriched transcription factor binding sites in the T7 human GOI list.*

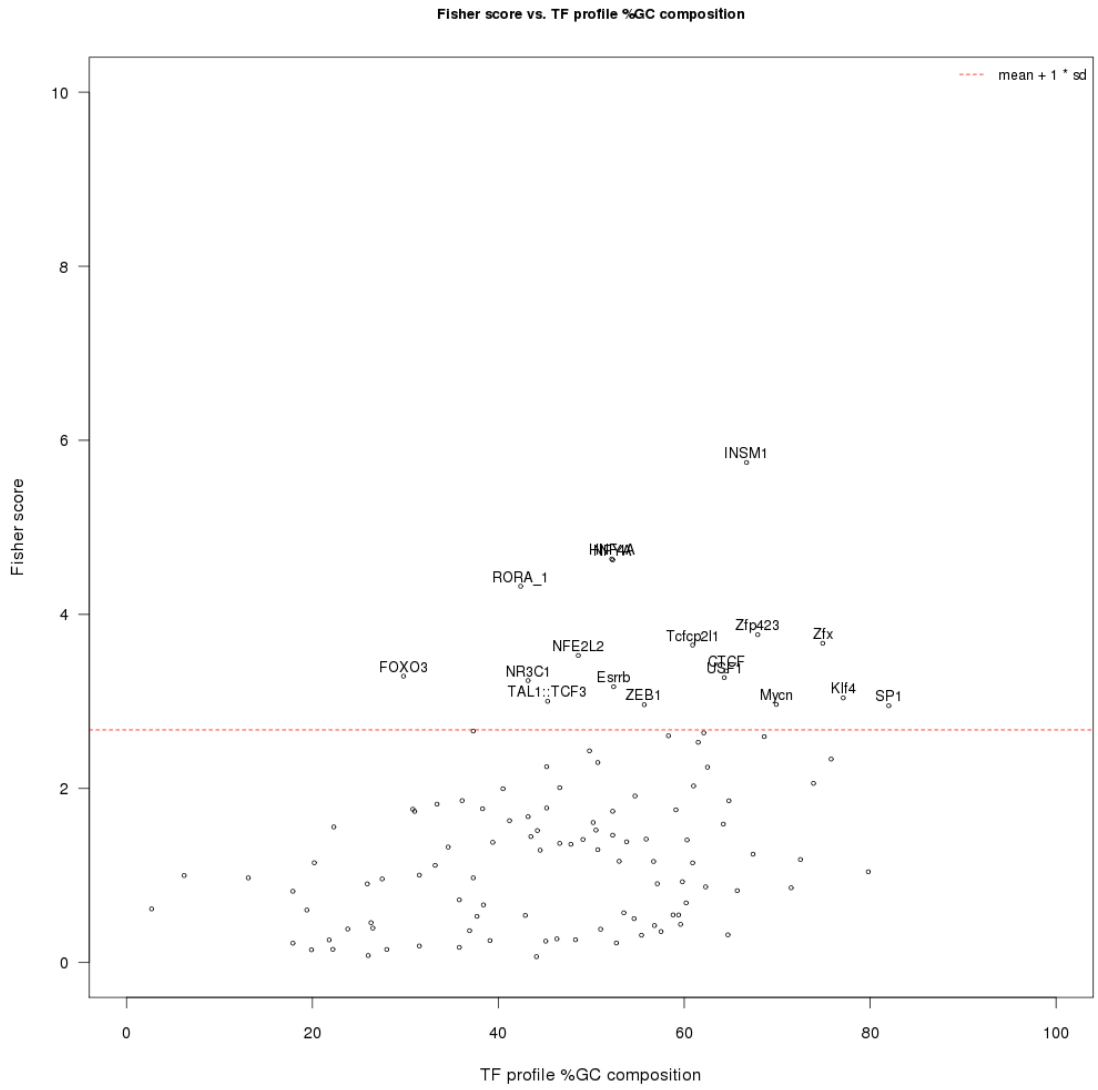



*Figure 4.42. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the T7 GOI list. The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors.*

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Fisher score | Z-score |
|---|---|---|---|---|---|---|---|---|
| RORA_1 | Zinc-coordinating | Hormone-nuclear Receptor | 113 | 80 | 3063 | 3478 | 7.082 | 10.464 |
| HNF1A | Helix-Turn-Helix | Homeo | 46 | 147 | 1124 | 5417 | 4.364 | 4.916 |
| STAT1 | Ig-fold | Stat | 84 | 109 | 2315 | 4226 | 4.34 | 7.734 |
| RORA_2 | Zinc-coordinating | Hormone-nuclear Receptor | 50 | 143 | 1315 | 5226 | 3.432 | 8.966 |
| HLF | Zipper-Type | Leucine Zipper | 72 | 121 | 2014 | 4527 | 3.393 | 3.773 |
| NFIL3 | Zipper-Type | Leucine Zipper | 83 | 110 | 2374 | 4167 | 3.37 | 0.901 |
| ELK4 | Winged Helix-Turn-Helix | Ets | 80 | 113 | 2292 | 4249 | 3.213 | -1.383 |
| NFYA | Other Alpha-Helix | NFY CCAAT-binding | 79 | 114 | 2270 | 4271 | 3.112 | -0.586 |
| CTCF | Zinc-coordinating | BetaBetaAlpha-zinc finger | 46 | 147 | 1227 | 5314 | 3.012 | 5.033 |
| TEAD1 | Helix-Turn-Helix | Homeo | 67 | 126 | 1906 | 4635 | 2.869 | 2.322 |

*Table 4.49. Top 10 transcription factor binding sites enriched in the T14 human GOI list ranked by Fisher score.*

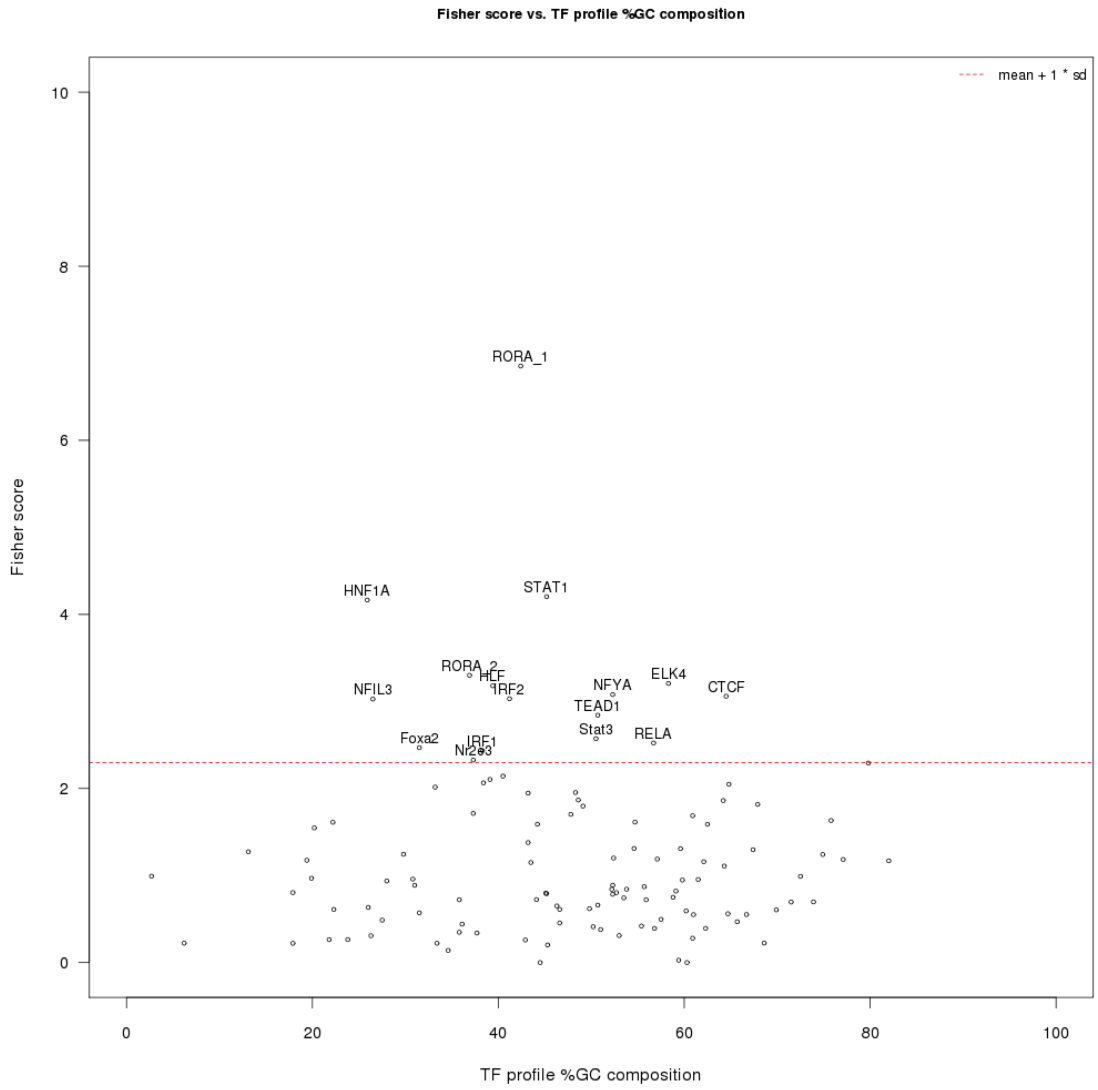



*Figure 4.43. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the T14 GOI list. The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors.*

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Fisher score | Z-score |
|---|---|---|---|---|---|---|---|---|
| RORA_1 | Zinc-coordinating | Hormone-nuclear Receptor | 266 | 217 | 3209 | 3616 | 7.91 | 9.256 |
| NFYA | Other Alpha-Helix | NFY CCAAT-binding | 199 | 284 | 2370 | 4455 | 5.986 | 2.102 |
| CTCF | Zinc-coordinating | BetaBetaAlpha-zinc finger | 115 | 368 | 1276 | 5549 | 5.504 | 8.066 |
| INSM1 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 234 | 249 | 2929 | 3896 | 4.58 | -1.497 |
| Zfp423 | Zinc-coordinating | BetaBetaAlpha-zinc finger | 180 | 303 | 2192 | 4633 | 4.447 | 5.012 |
| STAT1 | Ig-fold | Stat | 197 | 286 | 2429 | 4396 | 4.37 | 7.689 |
| ELK4 | Winged Helix-Turn-Helix | Ets | 194 | 289 | 2393 | 4432 | 4.283 | -1.38 |
| Tcfcp2l1 | Other | CP2 | 301 | 182 | 3899 | 2926 | 4.254 | -8.07 |
| Zfx | Zinc-coordinating | BetaBetaAlpha-zinc finger | 299 | 184 | 3872 | 2953 | 4.226 | -11.295 |
| NFE2L2 | Zipper-Type | Leucine Zipper | 174 | 309 | 2151 | 4674 | 3.761 | 4.611 |

*Table 4.50. Top 10 enriched transcription factor binding sites in the T7T14 human GOI list.*

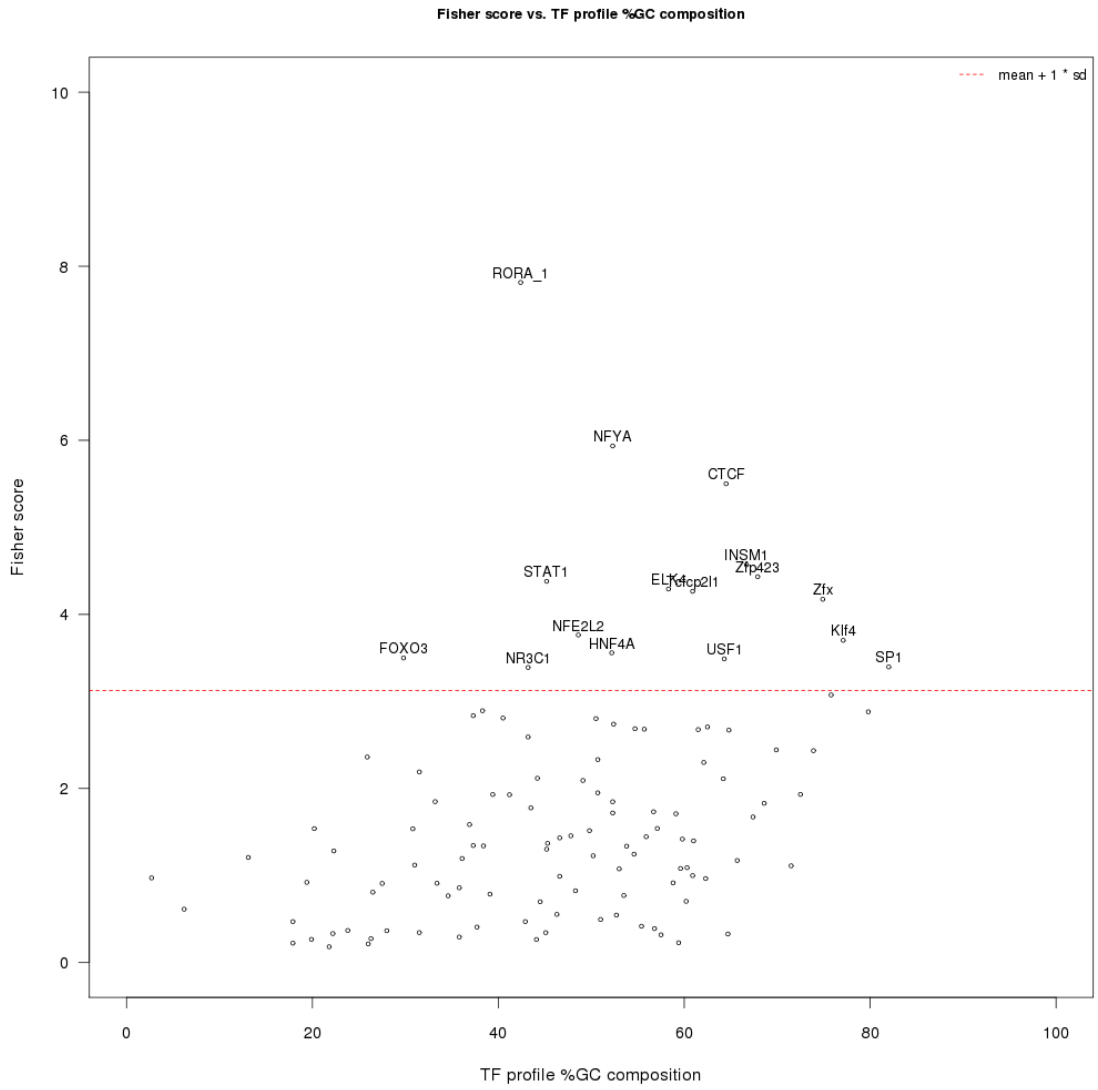



*Figure 4.44. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the combined proliferation GOI list.*
*The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors.*

| Name | Class | Family | Target gene hits | Target gene non-hits | Background gene hits | Background gene non-hits | Fisher score | Z-score |
|---|---|---|---|---|---|---|---|---|
| HNF1B | Helix-Turn-Helix | Homeo | 59 | 134 | 1417 | 4753 | 4.588 | 11.707 |
| Foxq1 | Winged Helix-Turn-Helix | Forkhead | 98 | 95 | 2611 | 3559 | 4.421 | 12.034 |
| CEBPA | Zipper-Type | Leucine Zipper | 139 | 54 | 4034 | 2136 | 3.45 | 2.489 |
| Pou5f1 | Helix-Turn-Helix | Homeo | 49 | 144 | 1211 | 4959 | 3.431 | 1.965 |
| Foxd3 | Winged Helix-Turn-Helix | Forkhead | 126 | 67 | 3625 | 2545 | 3.223 | 5.311 |
| FOXF2 | Winged Helix-Turn-Helix | Forkhead | 54 | 139 | 1376 | 4794 | 3.206 | 4.469 |
| TEAD1 | Helix-Turn-Helix | Homeo | 68 | 125 | 1798 | 4372 | 3.178 | 2.126 |
| Ddit3::Cebpa | Zipper-Type | Leucine Zipper | 71 | 122 | 1906 | 4264 | 3.005 | -2.272 |
| MEF2A | Other Alpha-Helix | MADS | 85 | 108 | 2363 | 3807 | 2.771 | -1.513 |
| HLF | Zipper-Type | Leucine Zipper | 71 | 122 | 1934 | 4236 | 2.732 | 0.755 |

*Table 4.51. Top 10 transcription factors enriched in the human lipotoxicity GOI list.*

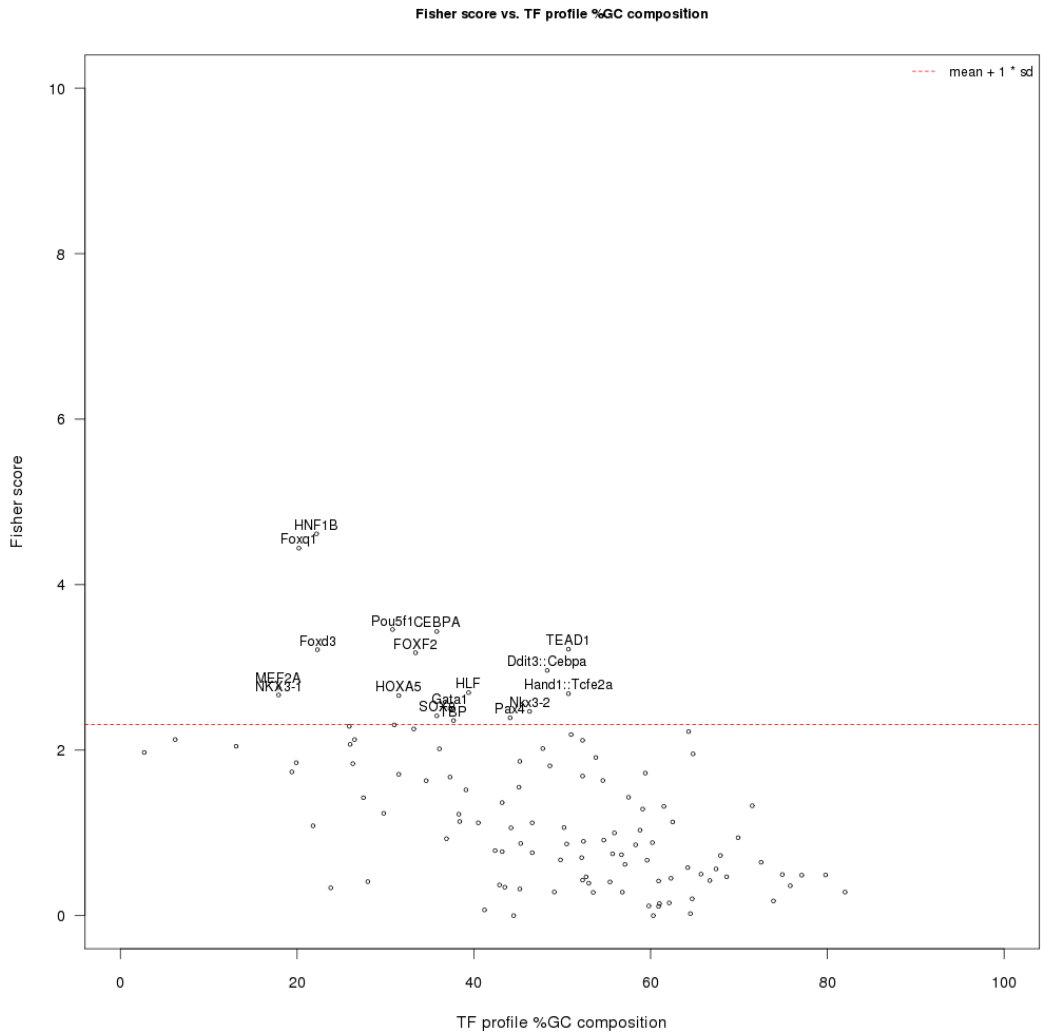



*Figure 4.45. Distribution of Fisher scores for the transcription factors analysed by OPOSSUM and the percentage of GC in their binding sites for the resistance to lipotoxicity GOI list.*
*The red line indicates a threshold set by OPOSSUM for graphical representation which may vary between samples. In this case, the threshold chosen is 1 standard deviation from the mean Fisher score for the transcription factors.*

## 4.4. Discussion

### *4.4.1. In vitro model used for this screen.*

These shRNA screens used R7T1, a reversibly immortalised mouse β-cell line, as an *in vitro* model of β-cells. The use of a cell line was necessary due to the scale of the screening process, which required a large number of β-cells with a stable phenotype. The choice of a reversibly immortalised cell line responded principally to the objective of screening for genes involved in proliferation control. Immortalisation of cells using SV40 large T antigen increases the proliferation rate of the cells via at least the pRb and p53 pathways (Ali and DeCaprio 2001). Cells therefore are not proliferating in a way which may be physiologically relevant and therefore any results obtained on mechanisms of proliferation may not be translated to an *in vivo* situation. Also, detection of an increase in proliferation in a rapidly proliferating cell line may have been difficult as the baseline of proliferation could be too high to discern any further effects.

I observed a decline in proliferation rate after withdrawal of doxycycline in a similar pattern to the decline in oncoprotein content in the cells. The cells did not stop proliferating completely as indicated by the incorporation of BrDU. However, the rate of BrDU incorporation had greatly reduced and β-cells still proliferate, albeit slowly, *in vivo* (Dor, Brown et al. 2004, Georgia and Bhushan 2004, Meier, Butler et al. 2008).

### *4.4.2. Screening process*

This genome-wide RNAi screen was performed in parallel for two phenotypes with their own independent controls, constituting two different screens. The design of the screen has many potential limitations.

Firstly, cells were transduced in parallel tissue culture vessels for the different samples of the same screen. In a genome-wide RNAi screen cells are usually transduced together as a pool and then separated into the screening conditions by dissociation from the tissue culture surface and seeding into the screening vessels (Luo, Cheung et al. 2008, Silva, Marran et al. 2008, Sims, Mendes-Pereira et al. 2011). This process however induced a high level of cell death in R7T1 and so cells were transduced in parallel in the wells where the screening was performed.

Second, the period of doxycycline withdrawal necessary to reduce the expression of the immortalizing oncogene may introduce a bias in shRNA targeting sequence representation if any targeting sequence affects growth in the presence of the oncoprotein or as its expression is reduced. However, the control condition for the proliferation screen, T0, is harvested after the withdrawal period and therefore it is expected that such effects would already be present in the control. As the proportion of targeting sequences in the screening time points (T7 and T14) is normalised to the proportion in T0, such effects would be discarded.

The use of palmitate as an apoptotic stimulus relevant to T2D may also be discussed. Although free fatty acid levels are elevated in T2D patients (Reaven, Hollenbeck et al. 1988, Boden and Shulman 2002), the levels of free fatty acids are measured in blood samples and the concentration to which β-cells are exposed in the islet are unknown.

Another limitation to these screens is that only one replicate was performed. However, the scale of the screens limited this possibility. The NKI library has a two-fold redundancy, which is not a solution for the lack of additional biological replicates but increases the confidence on those GOI targeted by two shRNA constructs.

### *4.4.3. PCR amplification and sequencing*

The two-step PCR amplification and Illumina library preparation used in this screen replaces the PCR amplification followed by library preparation using a PCR-free approach in Chapter 3. Performing two PCR amplifications may have reduced the linearity of amplification, but the number of PCR cycles was kept to a minimum. This resulted in a higher proportion of shRNA targeting sequences being identified in the sequencing data than in the screen in Chapter 3 or in the pilot project in this screen.

PCR amplification also needed the addition of betaine to relax the potential tertiary structures detected in the first PCR amplification reactions. The addition of betaine produced PCR amplification without tertiary structures. It cannot be excluded that secondary structure resulted in some sequences not being analysed. However, this would occur equally in the experimental and control

samples so it would not affect the results obtained but some biologically active genes may not have been assayed. This would have affected the network analysis as this was done in the absence of first neighbours.

This two-step PCR and library preparation in the presence of betaine resulted in the sequencing dataset with the highest quality as judged by the absence of primer dimers in the sequencing lane and the proportion of sequencing reads which contained a shRNA construct with a perfect match to the loop between complementary strands of the targeting sequence.

The improvement in shRNA construct identification may have also be a result of the samples being identified by their Illumina indexes instead of the PCR primer used to amplify them. Using the PCR primers for deconvolution of the sequencing output as in Chapter 3 could only identify those sequencing reads with a perfect match to the primer sequence, which may have been modified in the sequencing reads by a PCR or sequencing polymerase error.

#### *4.4.4. Sequencing data analysis.*

Identification of potential genes of interest (GOIs) was performed using the same strategy as in Chapter 3. The use of a z-score of the ratio between experimental condition and control has already been discussed in section 3.4.2.

A step of filtering for genes expressed in human β-cells was implemented to reduce the number of GOI to those which may be of interest for human biology. Using only one RNA sequencing dataset and only one individual is a limitation in this process. Ideally, several datasets including multiple individuals should be integrated to produce a better threshold of expression. However, this analysis was performed using only the individual with the median level of expression for *CFTR* for simplicity in these screens. This aspect may be re-considered in the future in order to produce a better dataset. Nevertheless, highly expressed or essential genes are not likely to not be expressed in one individual alone and therefore a more convoluted approach could result in more lowly expressed genes or genes expressed heterogeneously in the population being included.

*4.4.5. Function of GOI targeted by two shRNA constructs.*

Unlike the screen in Chapter 3, no confirmation experiments with a second targeting sequence were performed. As such, the closest to a confirmation experiment available at the moment are those GOI for which two targeting sequence have been considered to have had an effect, or those selected in T7 and T14.

In T7, seven genes were targeted by two sequences with an effect: *Zmynd11*, *Hax1*, *Anapc10*, *Bcr*, *Arhgef*1, *Ppia* and *N4bp2*. No gene was targeted by two sequences in T14.

*Zmynd11* was also selected as a GOI in T14. It is also known as *Bs69* and encodes a protein which binds adenovirus and inhibits Adenovirus serotype 5 E1A 289R (13S) (Hateboer, Gennissen et al. 1995), a nuclear phosphoprotein which transactivates cellular genes and other adenovirus early genes. *Bs69* also binds and inhibits the transcriptional activity of c-Myb (Ladendorff, Wu et al. 2001) and acts as a gene expression co-repressor upon binding with Ets-2 (Wei, Schaffner et al. 2003). This protein regulates genes involved in growth, differentiation and survival of normal and tumour cells and has been identified as a potential tumour suppressor (Ansieau and Sergeant 2003). *Zmynd11* has also been identified as a mediator of transcription elongation by histone H3.3K36me3, and binding of this histone seems to confer Zmynd11 its tumour suppressor activity (Wen, Li et al. 2014). Based on this information, knockdown of this GOI may induce cell proliferation in R7T1 cells.

*Hax1* encodes an anti-apoptotic protein related to Bcl-2 (Chao, Parganas et al. 2008) and therefore its knockdown would be expected to have the opposite effect of that detected in this screen. Bcl-2 reduces glucose signalling in β-cells (Luciani, White et al. 2013) and it can protect β-cells from apoptosis *in vitro* (Allison, Thomas et al. 2000).

*Anapc10*, also known as *Apc10* is part of the anaphase-promoting complex or cyclosome which is essential for cell cycle progression (Chang, Zhang et al. 2014). Its knockdown should produce the opposite effect to the observed.

Not much is known about *Bcr* in its complete form outside of the fusion protein BCR/ABL. *Bcr* is a tyrosine kinase that inactivates *Rac1* (Chuang, Xu et al.

1995) and a negative modulator of the *Wnt* pathway (Ress and Moelling 2005). Thus, its knockdown may upregulate cell proliferation.

*Arhgef1* encodes a Rho-guanine exchange factor which stimulates RhoA activation (Croisé, Brunaud et al. 2016). *RhoA* has been associated with cell proliferation and malignancy (Fritz, Brachetti et al. 2002, Horiuchi, Imai et al. 2003, Kamai, Tsujii et al. 2003) and therefore knockdown of *Arhgef1* may result in cell proliferation. Inhibition of RhoA has been shown to increase the level of human β-cell proliferation stimulation *in vitro* when used as a complement to Wnt signalling activation (Aly, Rohatgi et al. 2013).

*Ppia* encodes a protein also known as Cyclophilin A. Its downregulation has been shown to reduce cell proliferation (Obchoei, Weakley et al. 2011), which contradicts the results from the screen.

*N4bp2* binds Bcl-3 and its expression is upregulated in tumours (Zheng, Qin et al. 2007). Based on this information, knockdown of *N4bp2* should decrease cell proliferation.

From the genes found in both T7 and T14, only those with a sum of z-scores higher than 10 will be discussed.

The function of *Snrnp27* is not very well known. It is also known as *Ry1* and was discovered by the alteration of its expression following the integration of human T lymphotropic virus type 1 (Nakamura, Moriuchi et al. 1994). I therefore cannot comment on the possible effects of its knockdown.

*Tmem165* seems to have a role in glycosylation (Foulquier, Amyere et al. 2012).

*Pwp2* encodes the Periodic Tryptophan Protein Homolog in humans. Not much information is available on the human protein, but the yeast protein is necessary for ribosome assembly (Dosil and Bustelo 2004).

Znf8 binds Smad1 (Jiao, Zhou et al. 2002), which mediates the effects of transforming growth factor (TGF-β) and bone morphogenetic proteins (BMPs) in determining cell fate (Kretzschmar, Doody et al. 1997)

In the resistance to lipotoxicity screen, nine GOI were targeted by two shRNA sequences: *Mkrn2*, *Lrit2*, *Eprs*, *Mga*, *Chrnb1*, *Tbc1d10c*, *Zbed3*, *Cp*, *Spon2* and *Hic2*.

*Mkrn2* was detected in all GOI lists. It is a highly conserved gene antisense of *Raf1* (Gray, Azama et al. 2001), although the expression of both genes does not seem to be correlated (Lee, Chan et al. 2014). Overexpression of *Mkrn2* induces cell proliferation (Lee, Chan et al. 2014), which contradicts the results of the screens. However, if the antisense sequence of *Raf1* in *Mkrn2* was regulating the expression of *Raf1*, a reduction of expression in *Mkrn2* could increase the expression of *Raf1* thereby causing an increase in proliferation. *Raf1* activation in β-cells in mice through the deletion of a gene coding an inhibitory kinase increases β-cell mass (Pardo, Altirriba et al. 2012).

*Lrit3* has been identified as a regulator of fibroblast growth factor receptors (Fgfr) (Kim, Liu et al. 2012) which signal through Ras/MAPK pathways (Kouhara, Hadari et al. 1997), so its effect might be on proliferation. *Lrit3* was also found to be a GOI in the T14 sample.

*Eprs* is part of the gamma-interferon-activated inhibitor of translation (GAIT) which regulates gene expression in response to inflammation (Arif, Jia et al. 2009).

*Mga* encodes a protein which participates in c-Myc mediated signalling (Hurlin, Steingrìmsson et al. 1999). It also binds *Bs69/Zmynd11* (Ansieau and Leutz 2002). Thus, its effect in survival to lipotoxicity may be mediated through an effect on proliferation.

*Chrnb1* encodes a subunit of the muscle acetylcholine receptor (Le Novère, Corringer et al. 2002).

*Tbcdc10c* encodes a protein named Carabin, which can bind Calcineurin and inhibit Ras signalling (Pan, Sun et al. 2007). Knockdown of this protein may therefore deregulate cell proliferation or apoptosis in a non-palmitate specific manner. Suppression of calcineurin activity in β-cells in mice causes a decrease in β-cell mass (Heit, Apelqvist et al. 2006), therefore activation of proliferation may be independent of calcineurin.

The product of *Zbed3* interacts with Axin, a modulator of Wnt/β-catenin pathway (Chen, Li et al. 2009). A single nucleotide polymorphism (SNP) has been proposed to be associated with T2D risk (Ohshige, Iwata et al. 2011). More recently, secretion of ZBED3 protein was detected and higher levels were identified in serum of T2D patients versus non-diabetic controls (Jia, Yuan et al. 2014). *ZBED3* knockdown inhibits cancer cell proliferation (Fan, Jiang et al. 2015), which indicates that the effect detected in the screen may not be due to over-proliferation but to protection from cell death.

*Cp* encodes ceruloplasmin, a metalloprotein which binds copper (Holmberg and Laurell 1948) and oxidises $Fe^{2+}$ to $Fe^{3+}$ (Roeser, Lee et al. 1970) which is then transported by transferrin (Eid, Hémadi et al. 2014). Its expression is regulated by the GAIT complex (Sampath, Mazumder et al. 2003), providing a potential link with *Eprs*.

*Spon2* was found to be targeted by two constructs and it was also downregulated upon palmitate treatment in the dataset used in section 4.3.11. Spon2, also known as Mindin, interacts with PPARα (Zhu, Wang et al. 2014). PPARα is a transcription factor which controls lipid metabolism (Ziouzenkova, Perrey et al. 2003) and has been shown to protect human islets from lipotoxicity (Lalloyer, Vandewalle et al. 2006). Downregulation of *Spon2* should then reduce the protection against lipotoxicity contrary to my findings.

Finally, *Hic2* was identified as a homologue of the putative tumour suppressor *Hic1* (Deltour, Pinte et al. 2001). It has been proposed to be involved in cardiac development (Dykes, van Bueren et al. 2014).

### *4.4.6. Hyperconnected genes*

Although far from being a confirmation assay, evidence of a higher number of genes interacting with certain GOI as detected in 4.3.12.6 support the indication that those genes are involved in the phenotype as downregulation of their interacting genes has the same effect more often than expected.

Most of the genes detected as hyperconnected only have one or two connecting genes in the shRNA library targeted genes, which makes the interpretation of hyperconnection difficult. However, some of the genes have a high number of potential interacting proteins in the library and still were detected as

hyperconnected. I will discuss some of the genes showing that pattern in the combined proliferation and lipotoxicity GOI list.

In the proliferation GOI list, *Pten* has the highest number of potential connectors in the hyperconnected gene list. Of the 78 connectors in the library, 10 were included in the GOI list. PTEN is a tumour suppressor mutated in numerous cancers (Li, Yen et al. 1997) and negatively regulates the PI3K/AKT pathway (Maehama and Dixon 1998, Stambolic, Suzuki et al. 1998). Deletion of *Pten* in β-cells increases proliferation *in vivo* (Stiles, Kuralwalla-Martinez et al. 2006, Yang, Bayan et al. 2014). The effect of its knockdown in R7T1 cells is therefore not unexpected.

In the resistance to lipotoxicity screen only three genes were hyperconnected: *Aak1, Chrm3* and *Pcbd2*, while in the proliferation list 28 genes were detected as hyperconnected. This low number of hyperconnected genes may possibly be the result of two conditions: the lipotoxicity GOI list is too short for enrichment in connections to be detected, and the mechanism of lipotoxicity may be less studied and described than proliferation. Gene networks would therefore be less described for these genes, resulting in smaller networks. The gene network for the lipotoxicity screen was found to be smaller than the T14 network despite the GOI list being only five genes shorter (section 4.3.13.1).

*Aak1* is implicated in endocytosis (Conner and Schmid 2002) and has been identified as a positive regulator of the Notch1 signalling pathway (Gupta-Rossi, Ortica et al. 2011).

*Chrm3* is a muscarinic acetylcholine receptor which can stimulate phospholipase C (PLC) (Evellin, Nolte et al. 2002). PLC is involved in the palmitate-induced stress response in podocytes (Xu, Nam et al. 2015). Therefore, *Chrm3* knockdown may protect from palmitate-induced lipotoxicity by downregulating PLC activity.

*Pcbd2* is also known as *Dcoh2*. It is a co-activator of *Hnf1a* (Rose, Pullen et al. 2004). Heterozygous mutations in HNF1A cause MODY3 (Yamagata, Oda et al. 1996) and mice expressing a dominant negative form of Hnf1a have reduced β-cell mass (Yang, Yamagata et al. 2002). HNF1A has been shown to be de-regulated in rat β-cells in the presence of palmitate (Johnstone, Diakogiannaki

et al. 2011), although only in one (BRIN-BD11) of two cell lines tested (INS1 and BRIN-BD11). *Hnf1a* was not screened in the lipotoxicity and therefore no data is available on its knockdown effect in this study. However, transcriptional regulation of *Hnf1a* may be a mechanism for the effect of *Pcbd2* in protecting R7T1 cells from lipotoxicity.

### *4.4.7. Functional and gene set enrichment*

Functional enrichment analysis was performed with the objective of potentially finding functional categories which may be implicated in the phenotypes studied and to reduce the GOI list. The major limitation of the enrichment analysis using FunCoup, STRING, DAVID and FunRich is that the entire human genome was used as the background to determine enrichment.

The functional enrichment analysis of the human GOI lists using FunCoup, STRING and DAVID produced an extensive overlap of enriched categories for all of them. However, the enriched categories seem to be of a broad nature and not very specific, such as “regulation of cellular macromolecule biosynthetic process”. This may be a result of the source of functional category definition in these websites.

Analysis using FunRich indicated that all the proliferation GOI lists were enriched in binding sites for certain transcription factors. However, the same transcription factors were not detected in the analysis using OPOSSUM. This discordance may be due to the analysis being performed using the screened genes in OPOSSUM and the human genome in FunRich.

In contrast with the proliferation GOI lists, the enriched categories detected using FunRich in the resistance to lipotoxicity GOI list were not mainly transcription factors but functional categories related to endocitosis, molecule transport using channels and secretion process. This may indicate that uptake and release of molecules is important for the lipotoxicity of palmitate.

The gene set enrichment performed using GSEA PreRanked indicates that several gene sets defined by cancer datasets are enriched in the proliferation GOI lists, as well as proliferation-related pathways such as AKT. The resistance to lipotoxicity GOI list is also enriched in cancer-related gene sets, some of them overlapping the proliferation GOI lists, which may indicate an effect of

higher proliferation rate in the overrepresentation of targeting sequences in the screen dataset. It may also represent a bias in the GSEA database of gene sets towards cancer-related gene sets, which may be a result of the extensive research performed in cancer biology. An apoptosis-related gene set is also present in the list of enriched gene sets in the lipotoxicity GOI list, which is encouraging. However, the GSEA results would need to be analysed in more detail to provide a good interpretation of the data.

#### *4.4.8. Concluding remarks and future work*

The main limitation of this study in relation to interpreting and drawing conclusions from the GOI lists is the absence of replication studies, especially in the lipotoxicity screen as not much information is available about the mechanisms at play. Extensive validation will be needed to assess the role of the GOI in the mechanism of lipotoxicity. To reduce the list of GOI to be validated it may be necessary to produce new analysis methods which can mine and interpret available functional data.

Validation may be performed similarly to the Chapter 3, with additional targeting sequences used to evaluate the effects of the knockdown of some of the GOI. Chemical inhibitors are available for some of the GOIs or pathways detected in the screen and could also be used for validation of the effect of gene activity reduction.

Replication of the GOIs in other β-cell lines, and especially in human β-cell lines will be necessary to extrapolate these results to human biology. Confirmation of gene expression in human β-cells and comparison of expression data between T2D patients and non-diabetic controls could be also used as a validation approach.

# Chapter 5. Final discussion and conclusions

## 5.1 Design and development of functional genomic screens

I have presented in this thesis the development and analysis of three genome-wide RNAi screens in two *in vitro* models of complex diseases: suppression of tumorigenesis, β-cell proliferation and resistance of β-cells to lipotoxicity. I performed these three sets of experiments in parallel, using a similar set of techniques. Several general aspects of this work can be applied to any functional genomics screening approach. I will discuss these aspects and their implementation in the screens I performed for this thesis.

### 5.1.1 Choice of an adequate *in vitro* model.

The *in vitro* model chosen needs to reflect the biology being studied. Cell lines are known to differ at the transcriptomic and proteomic level from primary cells (Sandberg and Ernberg 2005, Olsavsky, Page et al. 2007, Pan, Kumar et al. 2009). These differences make difficult the translation of results from screens performed in cell lines to primary cells or to the *in vivo* situation.

Researchers have proposed and developed methods to perform RNAi screens using primary cells (Ovcharenko, Jarvis et al. 2005). High-throughput RNAi screening may be feasible in certain types of primary cells, especially those which are more easily obtained. For example, Wermke and colleagues performed a large-scale shRNA screen in primary leukaemia cells (Wermke, Camgoz et al. 2015) and Galeev and collaborators used a genome-wide shRNA library to find genes involved in renewal and differentiation of hematopoietic stem cells (Galeev, Baudet et al. 2016).

The screens presented in this thesis required prolonged culturing of cells and primary cells can only be cultured for a limited number of passages before they become senescent (Hayflick 1965) or progress through crisis and spontaneously immortalise. A large number of cells was required for both screens, which I would have needed to obtain from multiple donors. Variability amongst donors would then have introduced a new source of variability in the screening. Primary β-cells are extremely scarce and difficult to source, more so to achieve the numbers needed for the screens. Additionally, isolating primary cells for the neighbour suppression screen would require me to identify cells

within a tissue that have acquired an oncogenic mutation but have not formed a tumour. These cells would need to be cultured with normal cells from the tissue, which are more frequent but less likely to grow in cell culture conditions.

### 5.1.2 Choice of knockdown technology

Many RNAi and genome editing technologies are available for disrupting gene expression in mammalian cells, as presented in the introduction of this thesis. The choice of one needs to be evaluated in relation to the screen design and biology. If the screen is designed to evaluate a phenotype for a short period of time, siRNA may be adequate. siRNA molecules are delivered to the cytoplasm (Chiu, Ali et al. 2004) and downregulate the expression of the targeted gene whilst they are present (Yang, Tutton et al. 2001). However, screens observing a longer period of time may need the use of shRNA, usually coupled with virus delivery to the cell. shRNA molecules can be inserted into the genome of the cell. Once inserted, the shRNA molecules are transcribed by the cell's machinery and the reduction of gene expression is maintained. As the shRNA sequence is inserted in the genome it will be inherited by the daughter cells, maintaining the reduction of gene expression in the next generations as well.

The main criticism of RNAi approaches has been the existence of off-target effects (Scacheri, Rozenblatt-Rosen et al. 2004). It was first hypothesized that close sequence similarity between the targeting sequence and the unintended target mRNAs was the reason behind off-target effects. However, the first studies specifically looking at off-target effects of siRNA disagreed. Measuring genome-wide gene expression, Jackson and collaborators found that siRNA designed to target *MAPK14* and *IGFR1* affected expression levels of many other genes with varying degrees of similarity (Jackson, Bartz et al. 2003). The authors concluded that although a certain degree of similarity was needed, not all similar mRNA transcripts were downregulated. They could not find patterns to predict off-target effects and suggested that increasing the number of siRNA constructs against each gene of interest may help discern the true phenotype associated with its decreased expression (Jackson, Bartz et al. 2003).

It was later discovered that complementarity of siRNA nucleotides 2-7 or 2-8 (known as "seed region") to the 3' UTR contributes to the off-target effects (Birmingham, Anderson et al. 2006, Jackson, Burchard et al. 2006) and that the

magnitude of the off-target effect was influenced by the number of times the complementary region was present in the unintended target (Lin, Ruan et al. 2005). Partial mismatch at positions 9 and 10 of the sequences seem to produce changes in protein levels rather than mRNA levels (Alemán, Doench et al. 2007). Furthermore, the RNA-induced silencing complex (RISC) can load either of the sense and antisense strands of the shRNA molecule (Gu, Jin et al. 2011, Kwak and Tomari 2012, Schürmann, Trabuco et al. 2013)

Integration of this knowledge about the mechanisms of siRNA-induced RNAi led to building databases to predict off-target effects of targeting sequences such as SeedSeq (Das, Ghosal et al. 2013) and to the design and generation of new libraries. For example, Kampmann and colleagues (Kampmann, Horlbeck et al. 2015) examined many of the factors in library construction and used machine learning approaches to design an algorithm which could predict shRNA efficiency. Using that algorithm they have designed libraries targeting the human and mouse genome with 25 targeting sequences per gene (Kampmann, Horlbeck et al. 2015). This approach can be compared with the relatively simple design process for the NKI library, which can be divided into ten steps (Berns, Hijmans et al. 2004):

1) Does not contain stretches of more than three T or A.
2) Its GC content is between 30 and 70%.
3) Lies within coding sequence of the target gene.
4) The first nucleotide is a G or C.
5) Begins after an AA dimer in the 5’ flanking sequence.
6) Ends just before a TT, TG or GT in the 3’ flanking sequence.
7) Does not contain XhoI or EcoRI restriction-enzyme sites.
8) Minimize similarity with other genes.
9) Targets all the RefSeq mRNAs for that gene.
10) Does not overlap with the other targeting sequence in the library for the same gene.

Apart from the design of the targeting sequences, off-target effects may be limited with other approaches to promote the loading of only the antisense strand by RISC. The introduction of “decoy” sequences which sequester the sense strand of the shRNA leaving the antisense free to be loaded by RISC can

increase specificity of the targeting sequences (Mockenhaupt, Grosse et al. 2015)

Increasing the redundancy of the libraries to discern off-target effects has been the tendency in the last few years. In Table 5.1. I have compiled information about the shRNA libraries used in published functional genomics screens available on PubMed. Several of the libraries were used in more than one screen and only the earliest publication is shown. Two of the libraries (van de Weijer, Bassik et al. 2014, Shukla, Somwar et al. 2016) stand out for their high-coverage, although the library used by Shukla and colleagues has a very variable redundancy ranging from 1 to 58 shRNA targeting sequences per targeted transcript, with a median of 5. Most of the shRNA libraries used in publications in the last five years have a five-fold or lower redundancy.

CRISPR/Cas9 can also be used in screening processes. CRISPR/Cas9 systems are available not only for reducing gene expression but also to knockout and overexpress a gene. Similarly to RNAi, researchers have reported off-target effects using CRISPR/Cas9 (Cradick, Fine et al. 2013, Fu, Foden et al. 2013), although methods have been proposed to reduce off-target effects (Pattanayak, Lin et al. 2013) and to predict and analyse these (Hsu, Scott et al. 2013). Another limitation of the CRISPR/Cas9 system is that knockdown is produced at the DNA level in contrast to mature mRNA in RNAi. Therefore, knockdown of specific transcripts is not plausible.

The screens presented in this thesis used retroviral-delivered shRNA to produce stable knockdowns. The first genome-scale CRISPR screens performed were published in 2014 (Shalem, Sanjana et al. 2014, Wang, Wei et al. 2014), when the screens were already being performed and therefore the use of CRISPR was not considered.

| Library | Species | Transcripts targeted | Targeting sequences per gene | Year of publication | Reference |
|---|---|---|---|---|---|
| shRNAmir 2.16 (Open Biosystems) | *Mus musculus* | Approx. 28,000 | 2-3 | 2007 | (Gazin, Wajapeyee et al. 2007) |
| Hannon-Elledge second generation | *Homo sapiens* | 32,293 | 2-3 | 2010 | (O'Connell, Adamson et al. 2010) |
| Systems Bioscience | *Mus musculus* | Approx. 39,000 | 3-5 | 2011 | (Wang, Studach et al. 2011) |
| TRC | *Mus musculus* | Approx. 13,000 | 5 | 2011 | (Colombi, Molle et al. 2011) |
| TRC "80k" | *Homo sapiens* | 16,056 | 4-5 | 2011 | (Marcotte, Brown et al. 2012) |
| GeneNET Human 50k | *Homo sapiens* | 38,500 | 5 | 2012 | (Yoshino, Hara et al. 2012) |
| Unknown | *Homo sapiens* | Approx. 20,000 | 30 | 2012 | (Jiang, Pan et al. 2012) |
| GIPZ shRNAmir human genome-wide library | *Homo sapiens* | 16,487 | 3-4 | 2012 | (Mendes-Pereira, Sims et al. 2012) |
| pSM2c shRNAmir library (version 1.15) | *Homo sapiens* | Approx. 20,000 | 2-3 | 2012 | (Angus and Nevins 2012) |
| Expression Arrest™ GIPZ lentiviral shRNA | *Homo sapiens* | Approx. 20,000 | 2-3 | 2012 | (Iorns, Ward et al. 2012) |
| Unknown | *Homo sapiens* | 11,255 | 4-5 | 2012 | (Simpson, Hurren et al. 2012) |
| Decode RNAi | *Homo sapiens* | 11,954 | 2-3 | 2013 | (Yang, Fan et al. 2013) |
| Unknown | *Homo sapiens* | Approx. 18,000 | 30 | 2014 | (van de Weijer, Bassik et al. 2014) |
| MISSION™ human genome-wide shRNA | *Homo sapiens* | 15,000 | 5-6 | 2016 | (Gari, Gearheart et al. 2016) |
| pGIPZ lentiviral shRNAmir | *Homo sapiens* | 21,416 | 3-4 | 2016 | (Kurahara, Bohl et al. 2016) |
| Sigma-Aldrich shRNA Human Genome Library | *Homo sapiens* | 11,748 | 1-58 (median of 5) | 2016 | (Shukla, Somwar et al. 2016) |

*Table 5.1. Genome-wide shRNA libraries used for functional genomics screens in publications indexed on PubMed.*

### 5.1.3 Choice of library

Several libraries and collections of RNAi inducing molecules are now available, for example: (Berns, Hijmans et al. 2004, Silva, Li et al. 2005, Moffat, Grueneberg et al. 2006). Table 5.1. has information on some genome-wide shRNA libraries used in published screens.

Depending on the phenotype studied and the hypothesis being tested one needs to consider whether an unbiased library would yield any more information than a focused library targeting only part of the genome. For example, if the phenomenon under study is known to involve kinases and the aim of the screen is to define which kinases are involved a kinome library would suffice. For the same experimental effort, there is a trade-off between the number of genes targeted and the redundancy of the library. For example, with 6,000 plasmids a research study may benefit more from targeting 3,000 genes with two plasmids each while another study may find it more beneficial to target 1,000 genes with six plasmids each.

For the screens reported in this thesis, the aim was to uncover new molecular effectors for phenotypes where little or no molecular mechanism is known. Therefore I used an unbiased approach with a two-fold redundancy.

The choice of an RNAi library may also depend on the delivery method to be used, as libraries are available in plasmids suitable for direct transfection and for retroviral (Berns, Hijmans et al. 2004), lentiviral (Root, Hacohen et al. 2006) and adenoviral (Oehmig, Klotzbücher et al. 2008) delivery.

Apart from the targeting sequence design discussed above in section 5.1.3, shRNA libraries also differ in how the targeting sequences are designed and how the hairpins will be processed in the cell. The first libraries constructed used RNA pol III promoters to drive the expression of hairpins mimicking the pre-miRNA structures in the RNAi pathway. These include the NKI library used in this thesis and the first TRC (Moffat, Grueneberg et al. 2006) and Hannon-Elledge (Paddison, Silva et al. 2004) libraries. A second generation of shRNA libraries replaced the simple hairpins with structures similar to pri-miRNA which are transduced by RNA pol II, the enzyme responsible for transcription of pri-miRNA in human cells (Lee, Kim et al. 2004). The second generation Hannon-

Elledge library uses this structure, which they termed shRNA-mir (Silva, Li et al. 2005).

### 5.1.4 Choice of phenotype to be studied and method of assessment.

Pooled shRNA screens are evaluated in terms of changes in cell number as indicated by the change in representation of the targeting sequences in the library (Strezoska, Licon et al. 2012). As such, two types of phenotypes are readily assayed: changes in proliferation rate (enrichment screens) and protection from a depleting stimulus (depletion screens).

The targeting sequence may be used as a molecular barcode as they are unique in the library (Berns, Hijmans et al. 2004). Some libraries have a barcode outside of the targeting sequence (Silva, Li et al. 2005) which can be used to identify the targeting sequence without amplification of the hairpin structure. However, the use of a barcode separated from the hairpin may be a source of error in screen evaluation (Sack, Davoli et al. 2016). Retroviruses carry two viral genomes (Bender and Davidson 1976) and recombination may occur between both copies and produce recombinant sequences (Hu and Temin 1990). Sack and colleagues found that integration of the targeting sequences in the genomic DNA produced recombinant sequences where the barcode was no longer associated with the original targeting sequence (Sack, Davoli et al. 2016). Although reducing the distance between the barcode and the targeting sequence decrease the recombination rate, Sack and colleagues estimated a recombination rate of approximately 6% when the distance between both sequences is 96 bp (Sack, Davoli et al. 2016).

Changes in targeting sequence or accompanying barcode representation have been detected using two methods: hybridization to microarrays and next-generation sequencing (NGS) (Sims, Mendes-Pereira et al. 2011). The use of microarrays requires the design of custom microarrays with probes hybridizing to the barcodes used in the library, and only one sample can be used per microarray. NGS approaches do not require a customized platform and are not restricted to the detection of a predefined set of sequences as microarrays are. NGS has higher sensitivity and dynamic range than microarrays (Marioni, Mason et al. 2008) and can be easily multiplexed for the analysis of several samples simultaneously. However, the use of NGS to evaluate screens requires

extensive optimisation of the PCR amplification prior to sequencing, and sequencing libraries need to be prepared (Sims, Mendes-Pereira et al. 2011, Strezoska, Licon et al. 2012). For the screen in Chapter 3 sequencing libraries were prepared using a PCR-free approach which resulted in the overrepresentation of Illumina adapter dimers in the sequencing output, resulting in a loss of sequencing reads. This limitation was solved in Chapter 4 by preparation of libraries using a two-step PCR amplification approach. Although this two-step PCR protocol for the preparation of sequencing libraries required an extra step of optimization for the inclusion of the second PCR amplification, the sequencing data received did not include adapter primers and therefore the sequencing depth obtained was higher. Sequencing depth needs considering if NGS is used to detect changes in number. The higher the depth, the more resolution is obtained and smaller the changes which can be detected (Strezoska, Licon et al. 2012).

The results of the screens presented in this thesis were evaluated using next generation sequencing (NGS) and a molecular barcode system whereby the representation of each unique targeting sequence is used as a proxy for the number of cells containing that sequence.

### 5.1.5 Analysis method for quantification of effect

Analysis of the screening output needs considerable optimisation and consideration. Researchers have made available their methods implemented in different programming languages such as R, Perl or Python (Kumar, Goh et al. 2013, Dai, Sheridan et al. 2014). However, the methods used need to be assessed for each screen as they may not be adequate for all. Analysis methods for arrayed and pooled screens will have different requirements, and assessing pooled screens using microarrays or NGS will also dictate the type of analysis used. I will only discuss analysis methods for screens which use NGS for detecting changes in representation as it was the method used in this thesis.

The first step in the analysis of NGS data from pooled shRNA screen is to obtain the sequencing reads for each of the samples. If the samples were mixed before library construction as in Chapter 3 of this thesis, sequencing reads will need separating on the basis of the sequence of the primer used to amplify each sequence. This separation can be performed in the UNIX command line

using the “grep” function, which looks for lines in a file containing an exact match to the sequence of characters provided. The sequences containing that sequence can be copied into a new file. This process is easy to use and does not require the installation of any software, but relies on exact matches to the expected PCR primer. Polymerase errors during PCR amplification or sequence may have introduced errors in the primer sequence, and molecules carrying those errors will not be considered as sequencing reads for the samples using “grep”. There are utilities such as OBITools (Boyer, Mercier et al. 2016), a meta-barcoding package for UNIX which include functions to separate samples amplified with different PCR. In OBITools this utility is called “ngsfilter” and provides an option to allow for partial matches. OBITools requires installation and files need to be prepared in a certain format.

If the samples are not mixed and libraries are prepared separately as in Chapter 4, this step is not necessary. The Illumina adapter sequences are different and therefore sequencing reads for each sample are already separated. This eliminates the need for PCR primer identification in the sample and the associated loss of sequencing reads due to errors in amplification, but requires an extra step in the sample preparation protocol.

After sequencing reads have been assigned to the correct sample, it is necessary to provide a count of how many sequencing reads contain each targeting sequence in each sample. I performed this step using a custom Python screen which identifies the targeting sequence as a 19bp long sequence before the sequence of the shRNA loop. This method has the advantage of allowing the identification of not pre-defined sequences but requires the sequencing reads to have a perfect match to the shRNA loop sequence. Any PCR errors in the loop sequence would result in the targeting sequence not being identified.

Other methods have been developed which do not rely on perfect matches. For example, the shALIGN program developed by Sims and colleagues (Sims, Mendes-Pereira et al. 2011) trims the sequencing reads to obtain the 19bp targeting sequence and then aligns each of those 19mers to the expected composition of the library allowing for mismatches. This approach has two main disadvantages: it requires the user to provide the position where the 19mer is

expected and restricts the analysis to those sequences contained in the library. In the sequencing data obtained from the screens in this thesis, the 19mers were not found in a constant position but rather it could be shifted a few nucleotides. This shift appears to be due to polymerase errors resulting in the inclusion or deletion of nucleotides in the first bases of the sequencing reads. Therefore I decided to use the loop sequence as an indication of the targeting sequence position.

Another example of software package which includes an alignment utility is edgeR (Dai, Sheridan et al. 2014), which can be used as a part of R. The alignment strategy in this case also requires the user's input on the expected position of the targeting sequence but it allows a shift in the location of the sequence. It requires the user to provide a file with the expected composition of the library in a pre-defined format and it also requires the use of the R software. Although allowing for shifts in the position of the targeting sequence is a strong advantage of this method, I wanted to study the presence of sequences not present in the original library design and therefore did not consider using this method.

Once a count of the sequencing reads obtained for each targeting sequence has been obtained, a method needs to be used to detect those sequences with an effect in the phenotype studied. Many aspects of the screening process will influence the choice of an analysis method. The goal of the analysis should be to inform on the effect of knockdown in the phenotype as accurately as possible.

The biological basis of the phenotype needs to be considered carefully to select the appropriate control conditions. In the screen in Chapter 3, I used a control for growth of tumorigenic cells in the absence of interactions with normal cells (termed "plastic" condition) and two experimental conditions: growth in contact with a suppressive monolayer (3T3$^{supp+}$ cells) and a non-suppressive monolayer (3T3$^{supp-}$). At the same time, these experimental conditions were used as controls for each other to control for non-specific changes in growth due to cell-cell interactions. This experimental design was reflected in the analysis method via a normalisation of each co-culture sample to the "plastic" control and the implementation of a double threshold method for detection of effects which accounted for effects common to both co-cultures. In the case of the β-cell

screens in Chapter 4, different controls were used for each of the screens. In the proliferation screen, a control sample was established after reversal of oncogene expression to account for possible effects of gene knockdowns in the reversal process. For the resistance to lipotoxicity screen, the control sample was treated with the vehicle used to deliver the toxic stimulus such that effects due to unspecific toxicity of the vehicle could be controlled for.

Different aspects of the analysis will depend on the design of the library used. If the library is subdivided in smaller pools of sequences and each pool is used individually, variation may be introduced in the screening process. The analysis method chosen then needs to introduce a normalisation step in which this variation is eliminated or at least mitigated. In the case of both screens in this thesis, I calculated the relative abundance of each sequence inside the pool where it was predicted to originate from. This step also normalised the abundance for the differences in read depth which were detected between pools derived from an imperfect proportion of PCR products sequenced from each pool.

The subdivision of the library can also produce differences in the level of variability in representation inside each pool. This variability may be present in the plasmid pool already if each plasmid is not represented equally. It may also be introduced after transfection or transduction if any sequences have an effect on cell viability, or during the screening process if the culture conditions are changed and growth is affected by this change. To reduce the impact of this variability the analysis method should normalise the change in representation observed for each sequence by the change in representation observed in other sequences in the same pool. A robust z-score was used in this thesis, but other statistical methods are available. Many methods have been adopted from high throughput screening of chemical compounds and as such are more valid for arrayed screens than pooled screens (Birmingham, Selfors et al. 2009). One example is the “Percentage of control”, where each value is divided by the mean of several negative of positive controls. This is easily implemented in arrayed screens where a number of plates with negative or positive controls can be exposed to same screening conditions as the screen plates. In some simple pooled screens it may also be possible to introduce shRNA constructs which have been demonstrated to produce an effect in the same *in vitro* model and

conditions used in the screen (for example, a targeting sequence against a tumour suppressor in the study of tumorigenesis potential). Such controls were not available for the screens presented in this thesis and therefore this analysis method was not adequate.

A threshold needs to be chosen to discern targeting sequences with an effect in the phenotype studied from those with no effects. This has usually been done arbitrarily or using a probabilistic method (Birmingham, Selfors et al. 2009). In the screens presented in this thesis, the threshold for determination of effects was chosen based on the distribution of the targeting sequence z-scores. It is possible that this method may have failed to identify targeting sequences with small effects and included false positives.

### 5.1.6 False positives and the need for integrative analysis

RNAi screens are prone to the detection of false positives, in part due to off-target effects (discussed above) and in part to deficient analysis methods. The analysis of RNAi screens has changed since the first genome-wide screens and multiple approaches are now used to reduce the proportion of false positives. Recently, researchers have implemented integrative analysis pipelines and used several sources of data to interpret RNAi screens more fully. Expression data can be used to estimate whether the knockdown of a gene is likely to affect the phenotype. Network biology can be used to detect “essential” genes or hubs (Jiang, Wang et al. 2015), and functional or gene set enrichment can provide information on pathways involved in a phenotype (Huang, Wang et al. 2008). Analysis packages are available now which integrate all these steps (Diaz, Qin et al. 2015, Dutta, Azhir et al. 2016).

No expression data was available for the screen in Chapter 3, but network analysis and functional and gene set enrichment analysis approaches were implemented. I also used these methods for the analysis of the screen in Chapter 4. Expression data was included in this last screen to study only those genes which are expressed in human β-cells as targeting sequences designed against genes which are not expressed will most probably be acting through an off-target effect.

### 5.1.7 Validation strategies

Validation is critical in RNAi screens due to the high proportion of false positives, an issue already raised when reports of genome-wide screens started to be published (Echeverri, Beachy et al. 2006). The method of validation varies between screens. It is common to design additional targeting sequences binding different sites in the mRNA (Westbrook, Martin et al. 2005, Li, Ding et al. 2013, Landi, Vermeire et al. 2014), although this approach may yield false negative results if the level of knockdown is very different from the discovery sequence. However, a different sequence is unlikely to result in the same off-target effects as the discovery sequence. Some studies have also used the same targeting sequence to validate the results e.g. (Cooper and Brockdorff 2013). In this case, the level of knockdown should be equivalent to the original experiment and therefore false negatives are unlikely. However, if the effect detected in the first screen was due to an off-target effect this will be replicated, resulting in the perpetuation of a false positive.

Some studies have validated the effect of the knockdown by rescuing the phenotype by introducing an expression plasmid containing a transcript which cannot be downregulated by the targeting sequence used to produce the knockdown (Zhou, Ding et al. 2015). This approach would not replicate an effect due to off-target reduction of gene expression as the rescue would not be possible unless the targeted gene was expressed. However, phenotype rescue may not always be possible, which implies a risk of reporting false negatives (Datler and Grimm 2013). Phenotype rescue is not feasible on a large scale screen.

Chemical inhibitors are available for some proteins and they could be used to validate the effect produced by an shRNA construct. However, the level of inhibition may also be different to the screen and produce false negatives. Inhibitors are also rarely specific. Furthermore, although activity of the protein is inhibited, some of its activities may be conserved, such as serving as a scaffold for other proteins and this activity will be eliminated by the knockdown (Weiss, Taylor et al. 2007). The same limitation can be extrapolated to the knockdown of the gene using other technologies (e.g. siRNA or CRISPR/Cas9).

Other validation strategies may be implemented that use publicly available data to analyse the correlation of a reduction in expression of a gene *in vivo*. This approach was implemented in Chapter 3 with the study of gene expression levels in tumour samples and their correlation with distant metastasis free survival in patients. Although it is not a validation study *per se* in that it does not replicate and observe the effect of a reduction in expression, it does provide information about the association of that reduction with a phenotype of interest. This method depends on the availability of data which is relevant to the phenotype of interest.

## 5.2 Overlaps in the screens.

The screens in Chapter 3 and 4 in this thesis have been performed with the same shRNA library and analysed using similar approaches. Therefore analysing the overlap between the screens may inform us about false positive effects. It is important to consider that certain true positive effects may cause an overlap. This is expected especially in the case of genes with an effect in proliferation which could have been detected as increasing colony size in the co-cultures in Chapter 3 and any of the screens in Chapter 4.

### 5.2.1 Overlap in GOI lists.

There is a partial overlap in the GOI lists amongst the screens as depicted in Figure 5.1. The overlap was calculated using the human GOI lists from the β-cell screens and with the human homologues of the neighbour suppression screen. For this last screen, GOI lists from the $3T3^{supp+}$ and $3T3^{supp-}$ were pooled together.

Only one gene was found in all screens: *Impa1*. This gene encodes the protein inositol monophosphatase 1, a cytoplasmic enzyme which hydrolyses inositol monophosphate to produce inositol. This enzyme is required for embryo development (Cryns, Shamir et al. 2008). Due to its function in inositol signalling it may have a role in cellular proliferation.

The overlap between the neighbour suppression GOI list and the β-cell lipotoxicity screen GOI list is the smallest as expected from their distinct biology. Eleven genes are common between these two GOI lists: *Myocd, Fopnl, Odc1, Ube2v1, Pcbd2, Slk, Olig3, Eif2s1, Sh3gl3, Thsd7a* and *Rps9*.

I performed a literature search for functions of the overlapping GOI which may explain why they were selected as GOI in both lists.

*Myocd* inhibits cell proliferation (Tang, Zheng et al. 2008) and may be increasing the cell number in the lipotoxicity screen and in the neighbour suppression screen as it was detected as a GOI in the $3T3^{supp+}$ sample.

*Pcbd2* has already been discussed in section 4.4.6 as a hyperconnected GOI in the lipotoxicity screen. It was reasoned that it may protect β-cells through regulation of HNF1A. Being a GOI in the neighbour suppression list does not invalidate its possible importance in the resistance to lipotoxicity but it does

make it more probable that it may be a false positive. It was included in the $3T3^{supp-}$ GOI list.

*Olig3* is regulated by canonical Wnt signalling but not required for proliferation of spinal cord neurons in mice (Zechner, Muller et al. 2007), however I could not find published data on other cell types and may be involved in cell proliferation. It was included in the $3T3^{supp-}$ GOI list and therefore an effect in cell proliferation is not logical for both screens.

*Eif2s1* inhibition protects from ER apoptosis (Boyce, Bryant et al. 2005) and hence it may protect from lipotoxicity, although it was detected as a GOI in the $3T3^{supp-}$ sample and protection from apoptosis is not a logical mechanism to decrease colony size.

*Sh3gl3* is associated with invasion of glioma and decreasing its expression *in vitro* in glioma cells inhibits invasion (Delic, Lottmann et al. 2012). It was also included in the $3T3^{supp-}$ GOI list which may be logical with the effect in glioma cells.

Decreased expression of *Thsd7a* enhances migration of human umbilical vein endothelial cells *in vitro* (Wang, Su et al. 2010). It was included in the $3T3^{supp-}$ GOI list.

Depletion of For20, the protein encoded by *Fopnl*, inhibits cell cycle progression (Shen, Cai et al. 2013) and therefore it is more likely a false positive as it would decrease proliferation in the screens as it was included in the $3T3^{supp+}$ screen.

Similarly, downregulation of *Odc1* expression in mouse myoblasts decreases cell number (Lee, Skinner et al. 2011), although in the neighbour suppression screen it was included in the $3T3^{supp-}$ GOI list where a decrease in cell number is a plausible mechanism for decrease in colony size.

*Ube2v1* is also known as *Uev1a* and it has been linked to tumorigenesis through regulation of metalloproteinases (Wu, Shen et al. 2014) but its inhibition decreases proliferation and survival in *in vitro* models of diffuse large B-cell lymphoma (Pulvino, Liang et al. 2012). This effect of the targeting sequence is logical with its inclusion in the $3T3^{supp-}$ GOI list but not with an increase in cell number.

Inhibition of *Slk* should decrease proliferation according to experiments in mouse fibroblasts (O'Reilly, Wagner et al. 2005) and it was indeed included in the $3T3^{supp-}$ GOI list.

Decrease of *Rps9* expression inhibits proliferation of glioma and osteosarcoma in vitro models (Lindstrom and Nister 2010), which agrees with its inclusion as a GOI in the $3T3^{supp-}$ list.

Of the eleven overlapping genes between the resistance to lipotoxicity and the neighbour suppression screen, nine were detected as GOI for $3T3^{supp-}$. Although the function of most of these genes may provide a potential mechanism for this effect in tumour cell colony size it does not readily explain their effect in resistance to lipotoxicity. They may be false positives or be acting through a mechanism not considered here.

A larger overlap is observed between the neighbour suppression and the β-cell proliferation GOI lists. The list of these overlapping GOI is in Table 5.2 along with the neighbour suppression GOI list each of the GOI was included in. A larger proportion of GOI (13 of 30) originated from the $3T3^{supp+}$ list than in the overlap with the lipotoxicity screen (2 of 11). This may be a consequence of the proliferation and $3T3^{supp+}$ higher similarity in terms of mechanisms, as a higher proliferation rate may result in larger tumour cell colony size. The same logic cannot be applied to 17 GOI overlapping between the $3T3^{supp-}$ and β-cell proliferation lists as a higher proliferation rate cannot explain a smaller colony size. These genes and maybe part of the overlap with the $3T3^{supp+}$ may be false positives in one or both screens or they may have different effects in each of the screens.

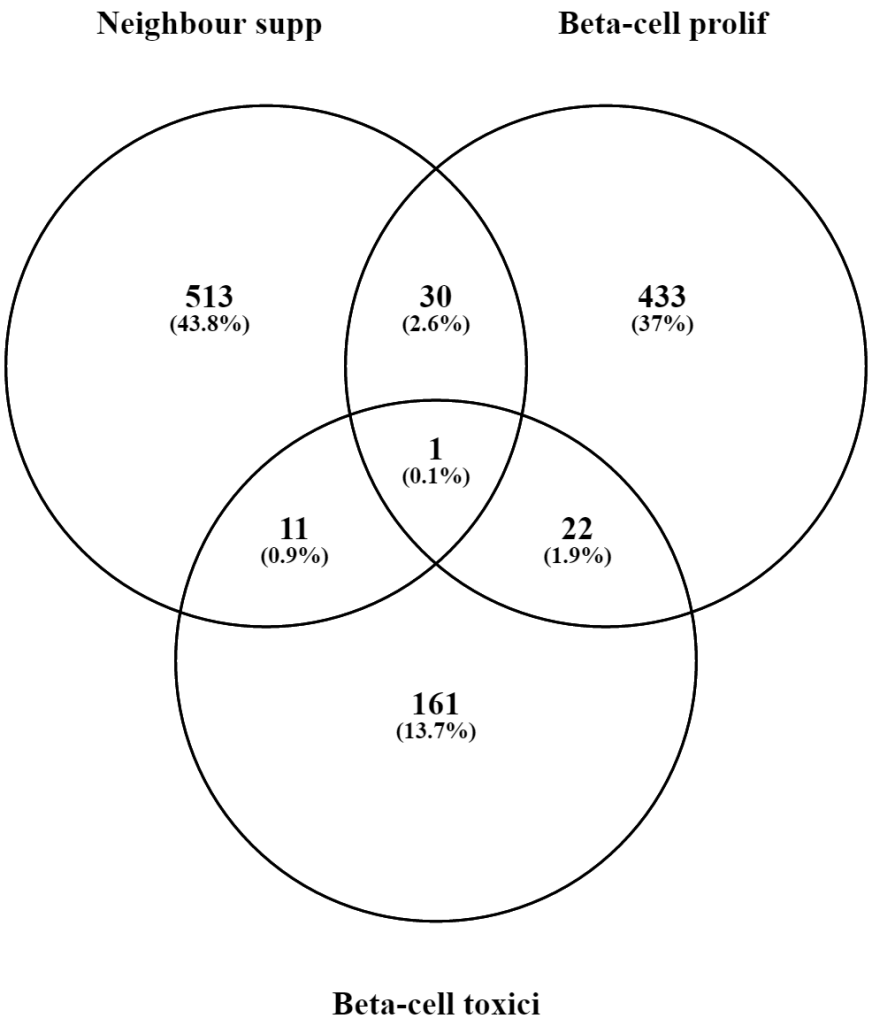



*Figure 5.1. Venn diagram depicting the overlap in the GOI lists from the three screens presented in this thesis.*
*"Neighbour supp": GOI list from both co-cultures. "Beta-cell prolif": β-cell proliferation GOI list. "Beta-cell toxici": β-cell resistance to lipotoxicity GOI list.*

| GOI | Neighbour suppression GOI list |
|---|---|
| *Ankhd1* | $3T3^{supp+}$ |
| *Lpar1* | $3T3^{supp+}$ |
| *Best1* | $3T3^{supp+}$ |
| *Tp53* | $3T3^{supp+}$ |
| *Osbpl8* | $3T3^{supp+}$ |
| *F11r* | $3T3^{supp+}$ |
| *Ppp2cb* | $3T3^{supp+}$ |
| *Wnt4* | $3T3^{supp+}$ |
| *Tuba4a* | $3T3^{supp+}$ |
| *Fbxo30* | $3T3^{supp+}$ |
| *Zfp606* | $3T3^{supp+}$ |
| *Ogdhl* | $3T3^{supp+}$ |
| *Flna* | $3T3^{supp+}$ |
| *Ube4a* | $3T3^{supp-}$ |
| *Capza2* | $3T3^{supp-}$ |
| *Rnaseh2b* | $3T3^{supp-}$ |
| *Slc25a23* | $3T3^{supp-}$ |
| *Sema6d* | $3T3^{supp-}$ |
| *Emd* | $3T3^{supp-}$ |
| *Ttc7a* | $3T3^{supp-}$ |
| *Cxadr* | $3T3^{supp-}$ |
| *Arhgap44* | $3T3^{supp-}$ |
| *Zbtb20* | $3T3^{supp-}$ |
| *Oaz2* | $3T3^{supp-}$ |
| *Pabpc1l* | $3T3^{supp-}$ |
| *Atp8a2* | $3T3^{supp-}$ |
| *Ppia* | $3T3^{supp-}$ |
| *H3f3c* | $3T3^{supp-}$ |
| *Clic5* | $3T3^{supp-}$ |
| *Sipa11* | $3T3^{supp-}$ |

*Table 5.2. GOI detected in the neighbour suppression screen and the β-cell proliferation screen.*

### 5.2.2 Overlap in the functional and gene set enrichment analysis

To investigate whether the same functional terms and gene sets may have been detected in the neighbour suppression and β-cell screens I examined the overlap in functional enrichment as determined by FunRich and gene set enrichment analysed with GSEA. I did not consider the functional enrichment analysis performed with the other approaches as they were done with different backgrounds (mouse genome for the neighbour suppression screen and human genome for the β-cell screens).

The overlap for the functional enrichment analysis performed using FunRich is depicted in Figure 5.2. The combined β-proliferation list results were used for the β-cell proliferation categories. All categories with an uncorrected p-value under 0.05 were used to calculate the overlap.

No functional categories were found enriched in all the GOI lists. Six were enriched in both the $3T3^{supp-}$ and the β-cell proliferation GOI lists, all of them transcription factors: FOS, FOSB, JUN, JUNB, JUND and NKX2-1. “Signal transduction” is enriched in both neighbour suppression GOI lists and the β-cell proliferation list. The $3T3^{supp+}$ and β-cell proliferation GOI lists share five enrichment functional categories: “Nuclear body”, “Receptor signalling complex scaffold activity”, “Cell communication” and two transcription factors: POU6F1 and IRF1. One category was found to be enriched in both β-cell screens and the $3T3^{supp-}$ GOI lists: “Centrosome”. Four categories are enriched in both β-cell screens and the $3T3^{supp+}$ screens: “Cytoplasm”, “Nucleus” and transcription factors ELF3 and HOXB4. The resistance to lipotoxicity and $3T3^{supp+}$ GOI lists share two enriched functional categories: “Extracellular matrix structural constituent” and transcription factor ARID3A. Finally the overlap between the resistance to lipotoxicity and $3T3^{supp-}$ is also composed of two functional categories: transcription factor FOXA1 and “Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism”.

Most of the shared functional terms are broad and refer to either cellular components such as “Nucleus” and “Cytoplasm” or to transcription factors. However there are other categories such as “Cell communication” which are logical for both GOI lists: cell-cell contact is known to be required for neighbour suppression (Mehta, Bertram et al. 1986, Alexander, Ichikawa et al. 2004) and

therefore its enrichment in 3T3$^{supp+}$ was expected, and β-cells function is improved *in vitro* when cells are in close contact as 3D structures (Beattie, Rubin et al. 1996). The “Extracellular matrix structural constituent” is also logical for the 3T3$^{supp+}$ GOI list as components of the extracellular matrix are differentially expressed in 3T3$^{supp+}$ and 3T3$^{supp-}$ cells (Allard et al, unpublished data). Decreasing the expression of genes in this category may perhaps contribute to the resistance to lipotoxicity by altering the uptake of palmitate.

For the gene set enrichment overlap I used the combined β-cell proliferation and the “Genes screened, ranked by (3T3supp+-3T3supp-) z-scores, duplicates removed using (3T3supp+-3T3supp-) z-scores” criteria for ordering the neighbour suppression GOIs (case 7 in section 3.3.11.5).

The overlap for enriched gene sets is represented in Figure 5.3. No gene sets were found to be enriched in all the screens, while seven are common to the resistance to lipotoxicity and the neighbour suppression and eight gene sets are enriched in both the β-cell proliferation and the neighbour suppression screen.

The seven gene sets shared between the resistance to lipotoxicity in β-cells screen and the neighbour suppression screen are
“HAMAI_APOPTOSIS_VIA_TRAIL_UP”,
“FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_OK_VS_DONOR_UP”,
“MORI_LARGE_PRE_BII_LYMPHOCYTE_UP”,
“SHEN_SMARCA2_TARGETS_UP”,
“KIM_ALL_DISORDERS_CALB1_CORR_UP”,
“BROWNE_HCMV_INFECTION_10HR_UP” and
“NEMETH_INFLAMMATORY_RESPONSE_LPS_DN”.

Enrichment in the “apoptosis via TRAIL” gene set in both screens may indicate a role for apoptosis in neighbour suppression and an involvement of the TRAIL (TNF-related apoptosis-inducing ligand) pathway or part of it in the apoptotic response of β-cells to palmitate.

The gene sets enriched in both the β-cell proliferation and the neighbour suppression screens are
“KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION”, “MUELLER_PLURINET”, “ST_P38_MAPK_PATHWAY”,
“ZAMORA_NOS2_TARGETS_UP”,

“KEGG_VIBRIO_CHOLERAE_INFECTION”, “BIOCARTA_WNT_PATHWAY”, “WHITEFORD_PEDIATRIC_CANCER_MARKERS” and “LASTOWSKA_NEUROBLASTOMA_COPY_NUMBER_DN”.

The “Biocarta Wnt Pathway” is perhaps the most interesting set in this case as the Wnt pathway is known to be involved in β-cell proliferation and aberrantly regulated in cancer. Signalling through p38 has also been suggested to be involved in β-cell proliferation in MIN6 cells (Burns, Squires et al. 2000) and in many aspects of tumourigenesis including signal transduction in tumour cells (Koul, Pal et al. 2013). Therefore the overlap in both screens is not unexpected and may not be a result of off-target effects of the targeting sequences. In the case of the β-cell proliferation it may indicate that signalling through p38 is also involved in proliferation of R7T1 cells after decreasing the expression of the oncogene. This suggests a possible role of p38 signalling in β-cell proliferation independent of oncogene expression. Signal transduction through p38 in tumour cells may also be involved in the response of tumour cells to signals from normal cells.

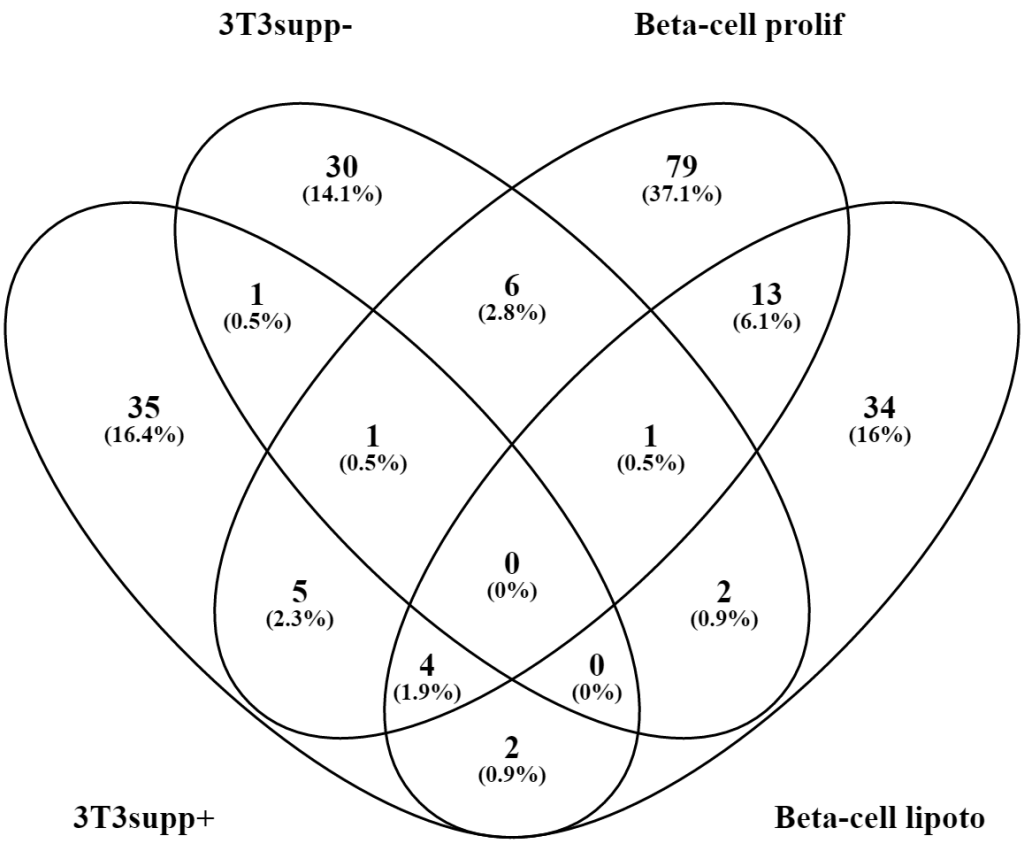



*Figure 5.2 Venn diagram representing the overlap in functional enrichment between the neighbour suppression screen GOI lists (3T3supp+ and 3T3supp-) and the β-cell proliferation (Beta-cell prolif) and resistance to lipotoxicity (Beta-cell lipoto).*

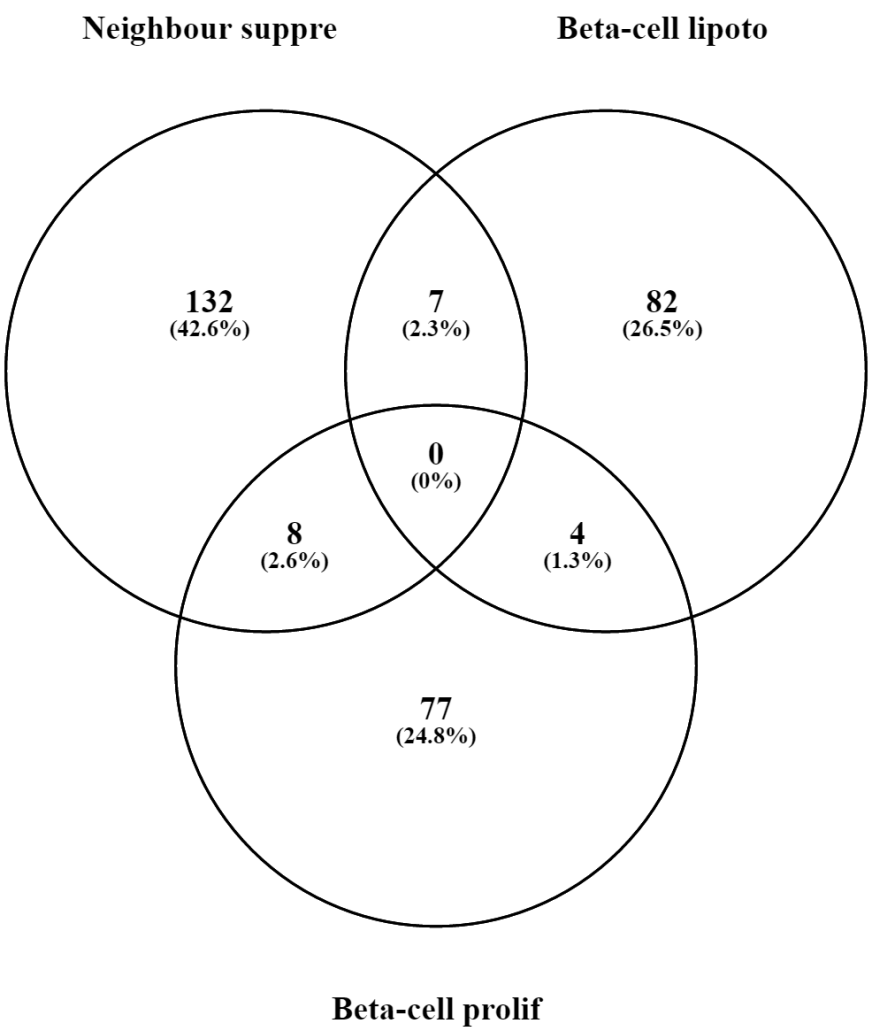



*Figure 5.3. Venn diagram depicting the overlap in enriched gene sets as determined by GSEA in the neighbour suppression (Neighbour suppre), and β-cell resistance to lipotoxicity (Beta-cell lipoto) and proliferation (Beta-cell prolif) screens.*

## 5.3 Limitations of the screens

In the discussion sections of Chapter 3 and Chapter 4 and above in this chapter I have analysed some of the limitations encountered by the screening process. Although extensive optimisation was implemented in both the functional assays and the analysis process, several limitations remain.

### 5.2.1. *In vitro* model used

The use of mouse cell lines has been extensively discussed through the results and discussion chapters. I have already discussed why the use of primary human cells was not a feasible option for the screening processes. I acknowledge that using a mouse model may be a challenge for translation of results into human biology. However, there were two main reason for this decision.

Firstly, previous work in the biology of neighbour suppression has been done using the cell lines in Chapter 3 (Stoker, Shearer et al. 1966, Allard, Stoker et al. 2003, Flaberg, Guven et al. 2012, Nicholls 2015) and by using the same model the results of this screen can be integrated with the data available from these previous studies. Furthermore, as discussed in Chapter 3, the availability of a non-suppressive cell line ($3T3^{supp-}$) is a strong advantage which allows the study of neighbour suppression by detecting genes necessary for the response to suppression but also those which can elicit a similar response in absence of suppression.

Secondly, the screen in Chapter 4 required the use of a reversibly immortalised β-cell line. A reversibly immortalised human β-cell line was developed in 2014 (Scharfmann, Pechberty et al. 2014), when the screening process had already been started.

### 5.2.2. Limitations derived from the use of RNAi and the NKI library

RNAi is now a well-established method for the study of functional genomics. However, there are aspects of the technology which need to be considered when designing experiments and drawing conclusions from the results produced.

Principally, RNAi targeting sequences in libraries are designed using predictive computer algorithms and their performance is unknown, which may lead to the knockdown of unwanted genes (Boutros, Bras et al. 2006) and to variable levels of knockdown (Sims, Mendes-Pereira et al. 2011).

The shRNA targeting sequences in the NKI library have not been validated. Although the theoretical composition of the library is approximately 28,000 constructs targeting 14,000 genes, some of those constructs may not provide sufficient reduction of gene expression or may have off-target effects, meaning that the real coverage of the library is unknown. Furthermore, the sequences in the library are not equally represented (Nicholls 2015) and so some of the sequences may not have been transfected or transduced, further reducing the coverage of the library.

RNAi libraries are now available with a high level of redundancy (Kampmann, Horlbeck et al. 2015) which may alleviate the impact of off-target effects on coverage. The NKI library was designed with a two-fold redundancy which is not optimal for coverage in case of off-target effects but as discussed earlier it allows a genome-wide scale screening process easier than high-redundancy libraries.

### 5.2.3. Limitations of the sequencing and analysis

The number of sequencing reads per targeting sequence (sequencing depth) achieved in the screens is lower than the recommended levels (Hu, Kim et al. 2009, Sims, Mendes-Pereira et al. 2011). Many sequencing reads in Chapter 3 and the pilot project in Chapter 4 did not contain an exact match to the primer used for amplification and so they could not be assigned to a sample, reducing the sequencing depth obtained.

In the process of data analysis functional enrichment was used to produce a list of potential mechanisms or pathways involved in both phenotypes studied. As I discussed in Chapter 3 and Chapter 4, these functional enrichments are relative to the whole genome. The correct background for these analyses would be the list of genes for which the library contains targeting sequences producing a significant level of knockdown. This list is not available and it is unlikely to be available not only because it would require extensive validation of each

sequence but also because the “significant level of knockdown” may be different for each phenotype assayed.

### 5.2.4. Validation

As previously discussed, RNAi screens require validation of the genes identified. For the screen in Chapter 3 I performed some validation experiments using other targeting sequences for the genes of interest (GOI). While some of the GOI showed the same effect as detected in the screen, the experiments were very variable in terms of effect size.

In Chapter 4 I did not perform any validation experiments, which limits the conclusions that can be extracted from the results. Validation work is ongoing at the moment.

## 5.4 Future work

There are two main lines of future work for the screens in this thesis: *in silico* and *in vitro.*

The analysis pipeline can be further refined by implementing some changes and additions to the methods:

- Allow mismatches in the identification of primers to sort sequencing reads into samples. This could increase the sequencing depth in Chapter 3.
- Allow mismatches between targeting sequences detected in the sequencing data and the targeting sequences in the library. This is already implemented in packages such as edgeR (Dai, Sheridan et al. 2014) and could increase the sequencing depth in both screens.
- Predict efficiency of the targeting sequences would allow a better understanding and selection of GOI. Algorithms such as sherWOOD (Knott, Maceli et al. 2014) could be used to inform on the potential effects of the targeting sequences.

After refining the analysis the main objective on the i*n vitro* work should be validation experiments to produce an informative list of genes involved in the phenotypes studied in both screens.

For the genes in the neighbour suppression screen for which a second targeting sequence has produced a difference in tumour cell colony size, the first step would be to confirm that the targeted gene is expressed in the SVE cell line and that the targeting sequence does reduce its expression. This could be done using qRTPCR. The protein levels before and after introduction of the targeting sequences also need to be examined, for example using western blots. If the gene is expressed and the targeting sequence reduces expression and protein levels, the pathway through which the gene is altering colony size should be examined. This could be done by investigation of genes known to operate upstream or downstream the candidate gene.

As the *in vitro* model of neighbour suppression that I have used for the screen is composed of two co-culture conditions, the candidate genes could be

overexpressed in the tumour cells to observe the effect of overexpression on the other co-culture.

Then, the same process should be performed in another *in vitro* co-culture model of neighbour suppression to investigate whether the effect is translatable to other cell types and not specific to $3T3^{supp+}$, $3T3^{supp-}$ and SVE6 cells. It should also be investigated if the effect is independent of the oncogene used to transform the cells. For this, tumorigenic cell lines transformed with other oncogenes could be used.

Further validation experiments to translate the uncovered mechanisms to human cancer biology should be performed by repeating the validation experiments in human cell lines or primary cells.

*In vivo* validation should also be performed to investigate whether the same effect is observed. For this, mice may be injected with tumorigenic cells carrying the targeting sequence for the validated genes or with a control sequence. Tumour growth in the control and experimental conditions should then be compared.

The β-cell screens require more work as no validation has yet been performed. The first step would be to select a number of GOI from both screens, design new targeting sequences and evaluate if any changes are present in proliferation rate or resistance to death by palmitate after their delivery to the cells. This target validation is currently ongoing using R7T1 cells. The same process planned for the neighbour suppression GOI validation above should then be performed: assessment of expression and protein levels before and after the introduction of the targeting sequences, investigation of the pathway involved in the effect and validation experiments in other cell lines. Those cell lines should be reversibly immortalized, or ideally primary cells. A reversibly immortalized human β-cell line is available (Scharfmann, Pechberty et al. 2014) which could be used to validate the results in a human *in vitro* model of β-cells.

*In vivo* experiments could be performed by producing conditional knockdown mice (i.e. the targeting sequence is expressed under the control of a β-cell specific promoter). The mice should then be exposed to diabetogenic conditions

and their glycaemia levels could be measured along with number, proliferation and apoptosis rate in their β-cells post-mortem. Changes in these measurements with respect to control mice (i.e. without expression of the targeting sequence) would indicate whether the effect of the GOI is validated *in vivo*.

## 5.5 Conclusions

RNAi has been extensively used in the recent years to study the molecular effectors underlying a wide array of phenotypes. Designing, optimising and implementing an RNAi genome-wide screen is a complex process which requires attention to detail. Genome-wide screens have significantly evolved in every aspect from experimental aspects to analysis, and efforts should be made to refine even further these approaches.

In Chapter 3 I presented a genome-wide screen for the detection of molecular effectors involved in the response of tumorigenic cells to suppressive cues from their neighbouring non-tumorigenic cells. The results from that screen provide a basic insight into the possible mechanisms responsible for the response of tumorigenic cells to their surrounding normal tissue in the early stages of tumorigenesis. From a methodological point of view, it introduces a pooled screen in a co-culture *in vitro* model and an analysis pipeline which accounts for one control condition and two experimental conditions which are also controls for each other.

In Chapter 4 I described the (to my knowledge) first genome-wide RNAi screen in a reversibly immortalised model of β-cells. Upon validation of the results obtained, this screen could inform on the mechanisms of proliferation in β-cells and how to potentially manipulate their growth, and also on the mechanism of resistance to lipotoxicity.

***Appendices***

## Appendix 1. Reannotation of the shRNA library using BLAST

- Bulding the RefSeq mRNA reference to align the library to:

I downloaded a fasta file with the refseq mrna database for Mus musculus from NCBI nucleotide (http://www.ncbi.nlm.nih.gov/nuccore, search "Mus musculus"[porgn:__txid10090] , check "mRNA(78059)" and "RefSeq(78059)", "Send to: File, Fasta"

I made the database from the fasta in Standalone Blast from Window's command line:

C:\Program Files\NCBI\blast-2.2.30+\db>makeblastdb -in mouserefseqmrna.fasta -dbtype "nucl" -out mouserefseqmrna

Output on command line:

"Building a new DB, current time: 03/17/2015 14:27:39

New DB name: mouserefseqmrna

New DB title: mouserefseqmrna.fasta

Sequence type: Nucleotide

Keep Linkouts: T

Keep MBits: T

Maximum file size: 1000000000B

Adding sequences from FASTA; added 78059 sequences in 5.26591 seconds."

- Blasting the library against the reference database:

blastn -query libquery.fasta -out libblastmousemrnadb.xml -outfmt "6 qseqid nident evalue stitle" -db "mouserefseqmrna" -max_target_seqs 1 -num_threads 4 -task blastn-short -evalue 1000 -word_size 7

## Appendix 2. Grep commands for the neighbour suppression screen (Chapter 3)

Format of command:

grep “primer” “Original file to grep from” > “File to move the grepped lines to”

Examples:

grep TCGAGGTCGACGGTA
/mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/A_TCGAAG_L006_R1_001.fastq
> /mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/MreadsAinc.txt

grep ATCTCTCGAGGTCGACGGTA
/mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/A_TCGAAG_L006_R1_001.fastq
> /mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/EreadsAinc.txt

grep GAGGTCGACGGTA
/mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/A_TCGAAG_L006_R1_001.fastq
> /mnt/Data10/lb428/Fibroblast_sequences/Project_2053/Sample_A/raw_illumina_reads/PreadsAinc.txt

## Appendix 3. Python script to extract targeting sequences

```
#Arguments: input file, output file. Prints a sequence of 19 bp before the beginning of the loop sequence - the shRNA 19mer in NKI library - and counts how many times that sequence is in the file

#1) check arguments, outputs usage error message if incorrect
import sys
import re
if len(sys.argv)!=3:
    sys.stderr.write("USAGE: %s <InFile> <OutFile>" % sys.argv[0])
    sys.exit(1)

#2) Open input and output file
inputf= open(sys.argv[1])
outf=open (sys.argv[2], 'w+')

#3) Re-direct output to file
sys.stdout=outf
from collections import Counter

#4) Initiate variables
fseq=""
seq=""
seq_list=[]
loopseqs=[]
setlist=[]
```

```
#5) Loop through lines in input
for line in inputf:
    #6) Search for exact matches of hairpin
    seq=re.search("(.19TCTCTTGAA)", line)
    #7) If hairpin is found, select the previous 19bp as a sequence and append to the list of sequences found
    if seq is not None:
        fseq=seq.group()[:19]
        seq_list.append(fseq)
    seq=""
#8) Count occurrence of sequences
seq_dict=Counter(seq_list)
for sequence in seq_list:
    loopseqs.append(str('%s %d' %(sequence, seq_dict[sequence])))

seqset=set(loopseqs)
setlist=list(seqset)

#9) Print to file sequences and occurrence
print ("\n".join(setlist))

#10) Close files
inputf.close()
outf.close()
```

## Appendix 4. Stata do-file for transformation and calculation of z-scores

```
replace Preads=. if Preads<51

levelsof pool, local(levels)

gen Prpmpool=.

gen Mrpmpool=.

gen Erpmpool=.

foreach l of local levels {
        egen totalpoolP=sum(Preads) if pool==`l'
        replace Prpmpool=(Preads/totalpoolP)*1000000 if pool==`l'
        drop totalpoolP
        egen totalpoolM=sum(Mreads) if pool==`l'
        replace Mrpmpool=(Mreads/totalpoolM)*1000000 if pool==`l'
        drop totalpoolM
        egen totalpoolE=sum(Ereads) if pool==`l'
        replace Erpmpool=(Ereads/totalpoolE)*1000000 if pool==`l'
        drop totalpoolE
    }
gen EtoPfoldrpm=Erpmpool/Prpmpool

gen MtoPfoldrpm=Mrpmpool/Prpmpool

gen lnMtoPfoldrpm=log(MtoPfoldrpm)

gen lnEtoPfoldrpm=log(EtoPfoldrpm)


levelsof pool, local(levels)

gen lnEtoPfoldmed =.

gen lnMtoPfoldmed=.
```

```
foreach l of local levels {
        qui sum lnEtoPfoldrpm if pool==`l', detail
        qui replace lnEtoPfoldmed = r(p50) if pool ==  `l'
        qui sum lnMtoPfoldrpm if pool==`l', detail
        qui replace lnMtoPfoldmed = r(p50) if pool ==  `l'

    }
bysort pool: egen lnEtoPfoldmad=mad(lnEtoPfoldrpm)
bysort pool: egen lnMtoPfoldmad=mad(lnMtoPfoldrpm)
gen zscorelnEtoP=(lnEtoPfoldrpm-lnEtoPfoldmed)/lnEtoPfoldmad
gen zscorelnMtoP=(lnMtoPfoldrpm-lnMtoPfoldmed)/lnMtoPfoldmad
```

## Appendix 5. Random sampling approach for recursive partitioning results

```
#Program to get a distribution of the number of genes  from a number of random groups that are significant in our recursive partitioning datasets.

#Written for my fibroblast screen data but can be used for any other type of input as long as we know sample size and have an input file organized in lines

#outputs a matrix (repetitions, number of datasetsx2)

import sys

import random

#1) Set the sample size and number of repetitions; and a header for the file (to discern the datasets)


size=312

repetitions=500


header="taylorprim_highisgood taylorprim_lowisgood taylorall_highisgood taylorall_lowisgood shed_highisgood shed_lowisgood loitamox_highisgood loitamox_lowisgood loinot_highisgood loinot_lowisgood des3_highisgood des3_lowisgood des1_highisgood des1_lowisgood desall_highisgood desall_lowisgood"


def check(n, variable, value):
    if lineS[n]==value:
        variable+=1
        return variable
    else:
        return variable
```

#2) Open the output file and set the standard output to the file so data is printed to it directly

```
with open (r"\\isad.isadroot.ex.ac.uk\UOE\User\Fibroblast screen\Recursive partitioning\smallMallscreenedpoolsrandom.txt", "w+") as outfile:
    sys.stdout=outfile
    print (header)
    n=0
    while n<repetitions:
        with open (r"\\isad.isadroot.ex.ac.uk\UOE\User\Fibroblast screen\Recursive partitioning\allscreenedpoolsresults.txt") as infile:
            sample=[]
            genenames=[]
            tayphigh=0
            tayplow=0
            tayahigh=0
            tayalow=0
            shedhigh=0
            shedlow=0
            loithigh=0
            loitlow=0
            loinhigh=0
            loinlow=0
            des3high=0
```

```
des3low=0
des1high=0
des1low=0
desahigh=0
desalow=0
sample=random.sample(infile.readlines(), size)
for line in sample:
    line = line.strip()
    lineS=line.split("\t")
    if len(lineS)>1:
        if lineS[0] not in genenames:
            genenames.append(lineS[0])
            tayphigh=check(1, tayphigh, "TRUE")
            tayplow=check(1, tayplow, "FALSE")
            tayahigh=check(2, tayahigh, "TRUE")
            tayalow=check(2, tayalow, "FALSE")
            shedhigh=check(3, shedhigh, "TRUE")
            shedlow=check(3, shedlow, "FALSE")
            loithigh=check(4, loithigh, "TRUE")
            loitlow=check(4, loitlow, "FALSE")
            loinhigh=check(5, loinhigh, "TRUE")
            loinlow=check(5, loinlow, "FALSE")
            des3high=check(6, des3high, "TRUE")
            des3low=check(6, des3low, "FALSE")
            des1high=check(7, des1high, "TRUE")
            des1low=check(7, des1low, "FALSE")
```

```
                desahigh=check(8, desahigh, "TRUE")
                desalow=check(8, desalow, "FALSE")

        print (tayphigh, tayplow, tayahigh, tayalow, shedhigh, shedlow, loithigh,
loitlow, loinhigh, loinlow, des3high, des3low, des1high, des1low, desahigh,
desalow)
        n+=1
```

## Appendix 6. Random sampling approach for detection of “hub” genes in network analysis

```
#Takes a list of nodes and outputs the number of interactors in a number of random samples
import sys
import random
from collections import Counter

#1) Input a list of nodes with gene name and ensembl ID separated by tab
samplesize= 592 #integer of size=hit list size-1
repetitions= 1000 #number of random samples to draw
#open output file
with open (r"PATH_TO_FILE/FILENAME"w+") as outfile:
    sys.stdout=outfile
    with open (r "PATH_TO_FILE/LIBRARY FILENAME”) as librarygenes:
        libraryg=[]
        for gene in librarygenes.readlines():
            gene=gene.strip()
            libraryg.append(gene)
    with open (r "PATH_TO_FILE/SAMPLEINTERACTIONSFILENAME”) as nodelist:
        for line in nodelist:
            line=line.strip()
            if len(line.split("\t"))>2:
                node=line.split("\t")[2]
                nodesymbol=line.split("\t")[0]
```

```
            geneinteractions=int(line.split("\t")[1])



        # Find interactors in file and build list of interactors
            interactors=[]
            with open (r" PATH_TO_FILE/INTERACTION DATABASE
FILENAME ") as interactions:
                for interaction in interactions:
                    interaction=interaction.strip()

                    if interaction.split("\t")[0] == node:
                        if interaction.split("\t")[1] in libraryg:
                            interactors.append(interaction.split("\t")[1])

                    if interaction.split("\t")[1] == node:
                        if interaction.split("\t")[0] in libraryg:

                            interactors.append(interaction.split("\t")[0])
                possibleinteractions=len(interactors)



                    # For 1 to n:
                repeat=0
```

```
            highconnected=0
            while repeat<repetitions:
            #open the file with the gene names and ensemblIDs of the
screened genes
                with open (r" PATH_TO_FILE/SCREENED GENES FILENAME ")
as library:
        #draw random sample from library ensembl ids
                sample=random.sample(library.readlines(), samplesize)
                interactorsinsample=[]
                numberinteractors=0


        #count interactors in random sample
                for line in sample:
                   line=line.strip()
                   if len(line.split("\t"))>1:
                      if line.split("\t")[1] in interactors:
                         if line.split("\t")[1] not in interactorsinsample:
                            interactorsinsample.append(line.split("\t")[1])
                            numberinteractors+=1
        #output node name and number of interactors in random sample
                if numberinteractors>geneinteractions:
                   highconnected+=1




                repeat+=1
```

```
                if repeat==repetitions-1:
                    print (nodesymbol, possibleinteractions, geneinteractions,
highconnected)
```

## Appendix 7. Stata do-file for transformation and calculation of z-scores in the β-cell screen.

```
levelsof pool, local(levels)

gen T0rpmpool=.

gen T7rpmpool=.

gen T14rpmpool=.

gen Controlrpmpool=.

gen Palmitaterpmpool=.

foreach l of local levels {

        egen totalpoolT0=sum(T0reads) if pool==`l'

        replace T0rpmpool=(T0reads/totalpoolT0)*1000000 if pool==`l'

        drop totalpoolT0

        egen totalpoolT7=sum(T7reads) if pool==`l'

        replace T7rpmpool=(T7reads/totalpoolT7)*1000000 if pool==`l'

        drop totalpoolT7

        egen totalpoolT14=sum(T14reads) if pool==`l'

        replace T14rpmpool=(T14reads/totalpoolT14)*1000000 if pool==`l'

        drop totalpoolT14

        egen totalpoolPalm=sum(Palmitatereads) if pool==`l'

        replace Palmitaterpmpool=(Palmitatereads/totalpoolPalm)*1000000 if
pool==`l'

        drop totalpoolPalm

        egen totalpoolControl=sum(Controlreads) if pool==`l'

        replace Controlrpmpool=(Controlreads/totalpoolControl)*1000000 if
pool==`l'

        drop totalpoolControl
```

```
    }
gen T7toT0foldrpm=T7rpmpool/T0rpmpool
gen T14toT0foldrpm=T14rpmpool/T0rpmpool
gen PalmtoContfoldrpm=Palmitaterpmpool/Controlrpmpool
gen lnT7toT0foldrpm=log(T7toT0foldrpm)
gen lnT14toT0foldrpm=log(T14toT0foldrpm)
gen lnPalmtoContfoldrpm=log(PalmtoContfoldrpm)

levelsof pool, local(levels)
gen lnT7toT0foldmed =.
gen lnT14toT0foldmed=.
gen lnPalmtoContfoldmed=.



foreach l of local levels {
        qui sum lnT7toT0foldrpm if pool==`l', detail
        qui replace lnT7toT0foldmed = r(p50) if pool ==  `l'
        qui sum lnT14toT0foldrpm if pool==`l', detail
        qui replace lnT14toT0foldmed = r(p50) if pool ==  `l'
        qui sum lnPalmtoContfoldrpm if pool==`l', detail
        qui replace lnPalmtoContfoldmed = r(p50) if pool ==  `l'

    }
bysort pool: egen lnT7toT0foldmad=mad(lnT7toT0foldrpm)
bysort pool: egen lnT14toT0foldmad=mad(lnT14toT0foldrpm)
bysort pool: egen lnPalmtoContfoldmad=mad(lnPalmtoContfoldrpm)
```

```
gen zscorelnT7toT0=(lnT7toT0foldrpm-lnT7toT0foldmed)/lnT7toT0foldmad

gen zscorelnT14toT0=(lnT14toT0foldrpm-lnT14toT0foldmed)/lnT14toT0foldmad

gen zscorelnPalmtoCont=(lnPalmtoContfoldrpm-
lnPalmtoContfoldmed)/lnPalmtoContfoldmad


gen PalmtoT0foldrpm=Palmitaterpmpool/T0rpmpool

gen lnPalmtoT0foldrpm=log(PalmtoT0foldrpm)

levelsof pool, local(levels)

gen lnPalmtoT0foldmed=.

foreach l of local levels {
        qui sum lnPalmtoT0foldrpm if pool==`l', detail
        qui replace lnPalmtoT0foldmed = r(p50) if pool ==  `l'
        }

bysort pool: egen lnPalmtoT0foldmad=mad(lnPalmtoT0foldrpm)

gen zscorelnPalmtoT0=(lnPalmtoT0foldrpm-
lnPalmtoT0foldmed)/lnPalmtoT0foldmad
```

## Appendix 8. Grep commands for the β-cell screen (Chapter 4)

Format of command:

grep “shRNA loop” “Original file to grep from” > “File to move the grepped lines to”

Examples:

grep TCTCTTGAA /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta2/raw_illumina_reads/Beta2_AAGAGGCA-CTCTCTAT_L002_R1_001.fastq > /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta2/raw_illumina_reads/T7loopFW.txt

grep TCTCTTGAA /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta3/raw_illumina_reads/Beta3_AAGAGGCA-TATCCTCT_L002_R1_001.fastq > /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta3/raw_illumina_reads/T14loopFW.txt

grep TCTCTTGAA /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta4/raw_illumina_reads/Beta4_AAGAGGCA-AGAGTAGA_L002_R1_001.fastq > /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta4/raw_illumina_reads/PalmloopFW.txt

grep TCTCTTGAA /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta5/raw_illumina_reads/Beta5_AAGAGGCA-GTAAGGAG_L002_R1_001.fastq > /mnt/Data10/lb428/Fibroblast_sequences/project2235/Sample_Beta5/raw_illumina_reads/ContloopFW.txt

## Appendix 9. List of functional terms found to be enriched in the 3T3$^{supp+}$, 3T3$^{supp-}$ and combined GOI lists using DAVID

| |
|---|
| phosphoprotein |
| acetylation |
| GO:0005829~cytosol |
| nucleus |
| GO:0043085~positive regulation of catalytic activity |
| GO:0044093~positive regulation of molecular function |
| cytoplasm |
| GO:0045944~positive regulation of transcription from RNA polymerase II promoter |
| metal-binding |
| nucleotide-binding |

## Appendix 10. List of functional terms found to be enriched in the 3T3$^{supp+}$ and combined GOI lists using DAVID

| |
|---|
| GO:0006793~phosphorus metabolic process |
| GO:0006796~phosphate metabolic process |
| isopeptide bond |
| GO:0030029~actin filament-based process |
| ubl conjugation |
| Transcription |
| GO:0016310~phosphorylation |
| GO:0048538~thymus development |
| GO:0010648~negative regulation of cell communication |
| GO:0030036~actin cytoskeleton organization |
| GO:0000159~protein phosphatase type 2A complex |
| mmu05410:Hypertrophic cardiomyopathy (HCM) |
| GO:0032553~ribonucleotide binding |
| GO:0032555~purine ribonucleotide binding |
| GO:0001822~kidney development |
| transcription regulation |
| GO:0003700~transcription factor activity |
| GO:0009968~negative regulation of signal transduction |
| GO:0017076~purine nucleotide binding |
| GO:0001558~regulation of cell growth |
| GO:0000166~nucleotide binding |
| kinase |
| GO:0001882~nucleoside binding |
| GO:0030016~myofibril |
| GO:0006350~transcription |

| |
|---|
| GO:0045449~regulation of transcription |
| GO:0032559~adenyl ribonucleotide binding |
| GO:0008287~protein serine/threonine phosphatase complex |
| GO:0006357~regulation of transcription from RNA polymerase II promoter |
| GO:0003677~DNA binding |
| GO:0001883~purine nucleoside binding |
| GO:0005604~basement membrane |
| GO:0043292~contractile fiber |
| GO:0030554~adenyl nucleotide binding |
| cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO) |
| GO:0030054~cell junction |
| GO:0043009~chordate embryonic development |
| GO:0009792~embryonic development ending in birth or egg hatching |
| GO:0006355~regulation of transcription, DNA-dependent |
| GO:0044459~plasma membrane part |
| GO:0043408~regulation of MAPKKK cascade |
| serine/threonine-protein kinase |
| domain:CTCK |
| centromere |
| actin-binding |
| IPR003349:Transcription factor jumonji, JmjN |
| GO:0051252~regulation of RNA metabolic process |
| mmu05414:Dilated cardiomyopathy |
| GO:0007015~actin filament organization |
| GO:0030528~transcription regulator activity |
| domain:JmjN |
| GO:0005626~insoluble fraction |

| |
|---|
| cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin) |
| IPR006207:Cystine knot, C-terminal |
| GO:0000267~cell fraction |
| GO:0001756~somitogenesis |
| GO:0016323~basolateral plasma membrane |
| SM00545:JmjN |
| GO:0044420~extracellular matrix part |
| GO:0007548~sex differentiation |
| GO:0006979~response to oxidative stress |

## Appendix 11. List of functional terms found to be enriched in the 3T3$^{supp-}$ and combined GOI lists using DAVID

| |
|---|
| dna-binding |
| GO:0010557~positive regulation of macromolecule biosynthetic process |
| GO:0009891~positive regulation of biosynthetic process |
| GO:0042995~cell projection |
| GO:0045941~positive regulation of transcription |
| GO:0051173~positive regulation of nitrogen compound metabolic process |
| GO:0050768~negative regulation of neurogenesis |
| DNA binding |
| GO:0033674~positive regulation of kinase activity |
| GO:0016477~cell migration |
| GO:0010721~negative regulation of cell development |
| GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process |
| GO:0010628~positive regulation of gene expression |
| GO:0051347~positive regulation of transferase activity |
| GO:0031328~positive regulation of cellular biosynthetic process |
| GO:0010604~positive regulation of macromolecule metabolic process |
| GO:0048870~cell motility |
| GO:0051674~localization of cell |
| GO:0006970~response to osmotic stress |
| GO:0043549~regulation of kinase activity |
| GO:0006928~cell motion |
| GO:0051338~regulation of transferase activity |
| zinc finger region:C2H2-type 3 |
| IPR013087:Zinc finger, C2H2-type/integrase, DNA-binding |

| |
|---|
| GO:0045859~regulation of protein kinase activity |
| GO:0030182~neuron differentiation |
| GO:0048732~gland development |
| GO:0051495~positive regulation of cytoskeleton organization |

***Bibliography***

"Cancer incidence statistics." Retrieved 06/06/2016, 2016, from http://www.cancerresearchuk.org/health-professional/cancer-statistics.

Alemán, L. M., J. Doench and P. A. Sharp (2007). "Comparison of siRNA-induced off-target RNA and protein effects." RNA **13**(3): 385-395.

Alexander, D. B., H. Ichikawa, J. F. Bechberger, V. Valiunas, M. Ohki, C. C. Naus, T. Kunimoto, H. Tsuda, W. T. Miller and G. S. Goldberg (2004). "Normal cells control the growth of neighboring transformed cells independent of gap junctional communication and SRC activity." Cancer Res **64**(4): 1347-1358.

Ali, S. H. and J. A. DeCaprio (2001). "Cellular transformation by SV40 large T antigen: interaction with host proteins." Seminars in Cancer Biology **11**(1): 15-22.

Aly, H., N. Rohatgi, C. A. Marshall, T. C. Grossenheider, H. Miyoshi, T. S. Stappenbeck, S. J. Matkovich and M. L. McDaniel (2013). "A Novel Strategy to Increase the Proliferative Potential of Adult Human β-Cells While Maintaining Their Differentiated Phenotype." PLoS ONE **8**(6): e66131.

Allard, D., M. Stoker and E. Gherardi (2003). "A G2/M cell cycle block in transformed cells by contact with normal neighbors." Cell Cycle **2**(5): 484-487.

Allison, J., H. Thomas, D. Beck, J. L. Brady, A. M. Lew, A. Elefanty, H. Kosaka, T. W. Kay, D. C. S. Huang and A. Strasser (2000). "Transgenic overexpression of human Bcl-2 in islet β cells inhibits apoptosis but does not prevent autoimmune destruction." International Immunology **12**(1): 9-17.

Angus, S. P. and J. R. Nevins (2012). "A role for Mediator complex subunit MED13L in Rb/E2F-induced growth arrest." Oncogene **31**(44): 4709-4717.

Ansieau, S. and A. Leutz (2002). "The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif." J Biol Chem **277**(7): 4906-4910.

Ansieau, S. and A. Sergeant (2003). "[BS69 and RACK7, a potential novel class of tumor suppressor genes]." Pathol Biol (Paris) **51**(7): 397-399.

Araten, D. J., D. W. Golde, R. H. Zhang, H. T. Thaler, L. Gargiulo, R. Notaro and L. Luzzatto (2005). "A quantitative measurement of the human somatic mutation rate." Cancer Res **65**(18): 8111-8117.

Arif, A., J. Jia, R. Mukhopadhyay, B. Willard, M. Kinter and P. L. Fox (2009). "Two-Site Phosphorylation of EPRS Coordinates Multimodal Regulation of Noncanonical Translational Control Activity." Molecular Cell **35**(2): 164-180.
Arnoldo, A., S. Kittanakom, L. E. Heisler, A. B. Mak, A. I. Shukalyuk, D. Torti, J. Moffat, G. Giaever and C. Nislow (2014). "A genome scale overexpression screen to reveal drug activity in human cells." Genome Medicine **6**(4): 1-16.
Arnould, S., P. Chames, C. Perez, E. Lacroix, A. Duclert, J.-C. Epinat, F. Stricher, A.-S. Petit, A. Patin, S. Guillier, S. Rolland, J. Prieto, F. J. Blanco, J. Bravo, G. Montoya, L. Serrano, P. Duchateau and F. Pâques (2006). "Engineering of Large Numbers of Highly Specific Homing Endonucleases that Induce Recombination on Novel DNA Targets." Journal of Molecular Biology **355**(3): 443-458.
Ashworth, J., J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, B. L. Stoddard and D. Baker (2006). "Computational redesign of endonuclease DNA binding and cleavage specificity." Nature **441**(7093): 656-659.
Atkinson, M. A., J. A. Bluestone, G. S. Eisenbarth, M. Hebrok, K. C. Herold, D. Accili, M. Pietropaolo, P. R. Arvan, M. Von Herrath, D. S. Markel and C. J. Rhodes (2011). "How Does Type 1 Diabetes Develop?: The Notion of Homicide or β-Cell Suicide Revisited." Diabetes **60**(5): 1370-1379.
Baek, D., J. Villen, C. Shin, F. D. Camargo, S. P. Gygi and D. P. Bartel (2008). "The impact of microRNAs on protein output." Nature **455**(7209): 64-71.
Bargmann, C. I., M. C. Hung and R. A. Weinberg (1986). "Multiple independent activations of the neu oncogene by a point mutation altering the transmembrane domain of p185." Cell **45**(5): 649-657.
Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-1712.
Barton, G. M. and R. Medzhitov (2002). "Retroviral delivery of small interfering RNA into primary cells." Proc Natl Acad Sci U S A **99**(23): 14943-14945.
Bassett, A. R., L. Kong and J.-L. Liu (2015). "A Genome-Wide CRISPR Library for High-Throughput Genetic Screening in Drosophila Cells." Journal of Genetics and Genomics **42**(6): 301-309.
Bassik, M. C., R. J. Lebbink, L. S. Churchman, N. T. Ingolia, W. Patena, E. M. LeProust, M. Schuldiner, J. S. Weissman and M. T. McManus (2009). "Rapid

creation and quantitative monitoring of high coverage shRNA libraries." Nat Methods **6**(6): 443-445.
Bauer, D. E., S. C. Kamran, S. Lessard, J. Xu, Y. Fujiwara, C. Lin, Z. Shao, M. C. Canver, E. C. Smith, L. Pinello, P. J. Sabo, J. Vierstra, R. A. Voit, G.-C. Yuan, M. H. Porteus, J. A. Stamatoyannopoulos, G. Lettre and S. H. Orkin (2013). "An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level." Science **342**(6155): 253-257.
Beattie, G. M., J. S. Rubin, M. I. Mally, T. Otonkoski and A. Hayek (1996). "Regulation of proliferation and differentiation of human fetal pancreatic islet cells by extracellular matrix, hepatocyte growth factor, and cell-cell contact." Diabetes **45**(9): 1223-1228.
Been, R. A., M. A. Linden, C. J. Hager, K. J. DeCoursin, J. E. Abrahante, S. R. Landman, M. Steinbach, A. L. Sarver, D. A. Largaespada and T. K. Starr (2014). "Genetic Signature of Histiocytic Sarcoma Revealed by a Sleeping Beauty Transposon Genetic Screen in Mice." PLoS ONE **9**(5): e97280.
Bender, A. M., L. S. Collier, F. J. Rodriguez, C. Tieu, J. D. Larson, C. Halder, E. Mahlum, T. M. Kollmeyer, K. Akagi, G. Sarkar, D. A. Largaespada and R. B. Jenkins (2010). "Sleeping Beauty-mediated somatic mutagenesis implicates CSF1 in the formation of high grade astrocytomas." Cancer research **70**(9): 3557-3565.
Bender, W. and N. Davidson (1976). "Mapping of poly(A) sequences in the electron microscope reveals unusual structure of type C oncornavirus RNA molecules." Cell **7**(4): 595-607.
Berenblum, I. (1941). "The cocarcinogenic action of croton resin." Cancer research **1**: 44-50.
Berezikov, E., V. Guryev, J. van de Belt, E. Wienholds, R. H. Plasterk and E. Cuppen (2005). "Phylogenetic shadowing and computational identification of human microRNA genes." Cell **120**(1): 21-24.
Bernal-Mizrachi, E., S. Fatrai, J. D. Johnson, M. Ohsugi, K. Otani, Z. Han, K. S. Polonsky and M. A. Permutt (2004). "Defective insulin secretion and increased susceptibility to experimental diabetes are induced by reduced Akt activity in pancreatic islet beta cells." J Clin Invest **114**(7): 928-936.
Berns, K., E. M. Hijmans, J. Mullenders, T. R. Brummelkamp, A. Velds, M. Heimerikx, R. M. Kerkhoven, M. Madiredjo, W. Nijkamp, B. Weigelt, R. Agami,

W. Ge, G. Cavet, P. S. Linsley, R. L. Beijersbergen and R. Bernards (2004). "A large-scale RNAi screen in human cells identifies new components of the p53 pathway." Nature **428**(6981): 431-437.

Berns, K., H. M. Horlings, B. T. Hennessy, M. Madiredjo, E. M. Hijmans, K. Beelen, S. C. Linn, A. M. Gonzalez-Angulo, K. Stemke-Hale, M. Hauptmann, R. L. Beijersbergen, G. B. Mills, M. J. van de Vijver and R. Bernards (2007). "A Functional Genetic Approach Identifies the PI3K Pathway as a Major Determinant of Trastuzumab Resistance in Breast Cancer." Cancer Cell **12**(4): 395-402.

Bernstein, E., A. A. Caudy, S. M. Hammond and G. J. Hannon (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-366.

Bibikova, M., D. Carroll, D. J. Segal, J. K. Trautman, J. Smith, Y. G. Kim and S. Chandrasegaran (2001). "Stimulation of homologous recombination through targeted cleavage by chimeric nucleases." Mol Cell Biol **21**(1): 289-297.

Bibikova, M., M. Golic, K. G. Golic and D. Carroll (2002). "Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases." Genetics **161**(3): 1169-1175.

Bignami, M., S. Rosa, S. A. La Rocca, G. Falcone and F. Tato (1988). "Differential influence of adjacent normal cells on the proliferation of mammalian cells transformed by the viral oncogenes myc, ras and src." Oncogene **2**(5): 509-514.

Birmingham, A., E. M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville, E. Maksimova, K. Robinson, J. Karpilow, W. S. Marshall and A. Khvorova (2006). "3[prime] UTR seed matches, but not overall identity, are associated with RNAi off-targets." Nat Meth **3**(3): 199-204.

Birmingham, A., L. M. Selfors, T. Forster, D. Wrobel, C. J. Kennedy, E. Shanks, J. Santoyo-Lopez, D. J. Dunican, A. Long, D. Kelleher, Q. Smith, R. L. Beijersbergen, P. Ghazal and C. E. Shamu (2009). "Statistical methods for analysis of high-throughput RNA interference screens." Nat Methods **6**(8): 569-575.

Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt and U. Bonas (2009). "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors." Science **326**(5959): 1509-1512.

Boden, G. and G. I. Shulman (2002). "Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and beta-cell dysfunction." Eur J Clin Invest **32: 14-23**.

Bolotin, A., B. Quinquis, A. Sorokin and S. D. Ehrlich (2005). "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin." Microbiology **151**(Pt 8): 2551-2561.

Bosco, E. E., Y. Nakai, R. F. Hennigan, N. Ratner and Y. Zheng (2010). "NF2-deficient cells depend on the Rac1-canonical Wnt signaling pathway to promote the loss of contact inhibition of proliferation." Oncogene **29**(17): 2540-2549.

Boutros, M., L. P. Bras and W. Huber (2006). "Analysis of cell-based RNAi screens." Genome Biol **7**(7): R66.

Boyce, M., K. F. Bryant, C. Jousse, K. Long, H. P. Harding, D. Scheuner, R. J. Kaufman, D. Ma, D. M. Coen, D. Ron and J. Yuan (2005). "A selective inhibitor of eIF2alpha dephosphorylation protects cells from ER stress." Science **307**(5711): 935-939.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet and E. Coissac (2016). "obitools: a unix-inspired software package for DNA metabarcoding." Mol Ecol Resour **16**(1): 176-182.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). Classification and Regression Trees, Chapman and Hall/CRC.

Breusegem, S. Y. and M. N. Seaman (2014). "Genome-wide RNAi screen reveals a role for multipass membrane proteins in endosome-to-golgi retrieval." Cell Rep **9**(5): 1931-1945.

Broderick, C. L., G. S. Brooke, R. D. DiMarchi and G. Gold (1991). "Human and rat amylin have no effects on insulin secretion in isolated rat pancreatic islets." Biochemical and Biophysical Research Communications **177**(3): 932-938.

Brouns, S. J., M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin and J. van der Oost (2008). "Small CRISPR RNAs guide antiviral defense in prokaryotes." Science **321**(5891): 960-964.

Brown, M. L. and A. L. Schneyer (2010). "Emerging roles for the TGFbeta family in pancreatic beta-cell homeostasis." Trends Endocrinol Metab **21**(7): 441-448.

Brummelkamp, T. R., R. Bernards and R. Agami (2002). "Stable suppression of tumorigenicity by virus-mediated RNA interference." Cancer Cell **2**(3): 243-247.

Brummelkamp, T. R., R. Bernards and R. Agami (2002). "A system for stable expression of short interfering RNAs in mammalian cells." Science **296 (5567): 550-553**.

Brummelkamp, T. R., A. W. M. Fabius, J. Mullenders, M. Madiredjo, A. Velds, R. M. Kerkhoven, R. Bernards and R. L. Beijersbergen (2006). "An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors." Nat Chem Biol **2**(4): 202-206.

Burgess, D. J., J. Doles, L. Zender, W. Xue, B. Ma, W. R. McCombie, G. J. Hannon, S. W. Lowe and M. T. Hemann (2008). "Topoisomerase levels determine chemotherapy response in vitro and in vivo." Proc Natl Acad Sci U S A **105**(26): 9053-9058.

Burleigh, A., S. McKinney, J. Brimhall, D. Yap, P. Eirew, S. Poon, V. Ng, A. Wan, L. Prentice, L. Annab, J. C. Barrett, C. Caldas, C. Eaves and S. Aparicio (2015). "A co-culture genome-wide RNAi screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation." Breast Cancer Res **17**: 4.

Burns, C. J., P. E. Squires and S. J. Persaud (2000). "Signaling through the p38 and p42/44 mitogen-activated families of protein kinases in pancreatic beta-cell proliferation." Biochem Biophys Res Commun **268**(2): 541-546.

Buschman, M. D., P. A. Bromann, P. Cejudo-Martin, F. Wen, I. Pass and S. A. Courtneidge (2009). "The novel adaptor protein Tks4 (SH3PXD2B) is required for functional podosome formation." Mol Biol Cell **20**(5): 1302-1311.

Buteau, J., S. Foisy, C. J. Rhodes, L. Carpenter, T. J. Biden and M. Prentki (2001). "Protein Kinase Cζ Activation Mediates Glucagon-Like Peptide-1–Induced Pancreatic β-Cell Proliferation." Diabetes **50**(10): 2237-2243.

Butler, A. E., J. Janson, S. Bonner-Weir, R. Ritzel, R. A. Rizza and P. C. Butler (2003). "Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes." Diabetes **52**(1): 102-110.

Butler, A. E., J. Janson, W. C. Soeller and P. C. Butler (2003). "Increased β-Cell Apoptosis Prevents Adaptive Increase in β-Cell Mass in Mouse Model of Type 2 Diabetes." Evidence for Role of Islet Amyloid Formation Rather Than Direct Action of Amyloid **52**(9): 2304-2314.

Büttner, D. and S. Y. He (2009). "Type III Protein Secretion in Plant Pathogenic Bacteria." Plant Physiology **150**(4): 1656-1664.

Cairns, J. (1975). "Mutation selection and the natural history of cancer." Nature **255**(5505): 197-200.
Clark, A., C. A. Wells, I. D. Buley, J. K. Cruickshank, R. I. Vanhegan, D. R. Matthews, G. J. Cooper, R. R. Holman and R. C. Turner (1988). "Islet amyloid, increased A-cells, reduced B-cells and exocrine fibrosis: quantitative changes in the pancreas in type 2 diabetes." Diabetes Res **9**(4): 151-159.
Cnop, M., J. C. Hannaert, A. Hoorens, D. L. Eizirik and D. G. Pipeleers (2001). "Inverse relationship between cytotoxicity of free fatty acids in pancreatic islet cells and cellular triglyceride accumulation." Diabetes **50**(8): 1771-1777.
Coghill, E. L., A. Hugill, N. Parkinson, C. Davison, P. Glenister, S. Clements, J. Hunter, R. D. Cox and S. D. M. Brown (2002). "A gene-driven approach to the identification of ENU mutants in the mouse." Nat Genet **30**(3): 255-256.
Colombi, M., K. D. Molle, D. Benjamin, K. Rattenbacher-Kiser, C. Schaefer, C. Betz, A. Thiemeyer, U. Regenass, M. N. Hall and C. Moroni (2011). "Genome-wide shRNA screen reveals increased mitochondrial dependence upon mTORC2 addiction." Oncogene **30**(13): 1551-1565.
Collier, L. S., C. M. Carlson, S. Ravimohan, A. J. Dupuy and D. A. Largaespada (2005). "Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse." Nature **436**(7048): 272-276.
Collinet, C., M. Stoter, C. R. Bradshaw, N. Samusik, J. C. Rink, D. Kenski, B. Habermann, F. Buchholz, R. Henschel, M. S. Mueller, W. E. Nagel, E. Fava, Y. Kalaidzidis and M. Zerial (2010). "Systems survey of endocytosis by multiparametric image analysis." Nature **464**(7286): 243-249.
Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang (2013). "Multiplex genome engineering using CRISPR/Cas systems." Science **339**(6121): 819-823.
Conner, S. D. and S. L. Schmid (2002). "Identification of an adaptor-associated kinase, AAK1, as a regulator of clathrin-mediated endocytosis." The Journal of Cell Biology **156**(5): 921-929.
Cooper, G. J., B. Leighton, G. D. Dimitriadis, M. Parry-Billings, J. M. Kowalchuk, K. Howland, J. B. Rothbard, A. C. Willis and K. B. Reid (1988). "Amylin found in amyloid deposits in human type 2 diabetes mellitus may be a hormone that regulates glycogen metabolism in skeletal muscle." Proceedings of the National Academy of Sciences **85**(20): 7763-7766.

Cooper, S. and N. Brockdorff (2013). "Genome-wide shRNA screening to identify factors mediating Gata6 repression in mouse embryonic stem cells." Development **140**(19): 4110-4115.
Cradick, T. J., E. J. Fine, C. J. Antico and G. Bao (2013). "CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity." Nucleic Acids Res **41**(20): 9584-9592.
Croisé, P., L. Brunaud, P. Tóth, S. Gasman and S. Ory (2016). "Inhibition of Cdc42 and Rac1 activities in pheochromocytoma, the adrenal medulla tumor." Small GTPases: 1-6.
Cruz, C., A. Glavic, M. Casado and J. F. de Celis (2009). "A Gain-of-Function Screen Identifying Genes Required for Growth and Pattern Formation of the Drosophila melanogaster Wing." Genetics **183**(3): 1005-1026.
Cryns, K., A. Shamir, N. Van Acker, I. Levi, G. Daneels, I. Goris, J. A. Bouwknecht, L. Andries, S. Kass, G. Agam, H. Belmaker, Y. Bersudsky, T. Steckler and D. Moechars (2008). "IMPA1 is essential for embryonic development and lithium-like pilocarpine sensitivity." Neuropsychopharmacology **33**(3): 674-684.
Chang, L., Z. Zhang, J. Yang, S. H. McLaughlin and D. Barford (2014). "Molecular architecture and mechanism of the anaphase-promoting complex." Nature **513**(7518): 388-393.
Chao, J.-R., E. Parganas, K. Boyd, C. Y. Hong, J. T. Opferman and J. N. Ihle (2008). "Hax1-mediated processing of HtrA2 by Parl allows survival of lymphocytes and neurons." Nature **452**(7183): 98-102.
Charles, A. M., E. Eschwe`ge, N. Thibult, J.-R. Claude, J.-M. Warnet, E. G. Rosselin, J. Girard and B. Balkau (1997). "The role of non-esterified fatty acids in the deterioration of glucose tolerance in Caucasian subjects: results of the Paris Prospective Study." Diabetologia **40**(9): 1101-1106.
Chen, H., X. Gu, Y. Liu, J. Wang, S. E. Wirt, R. Bottino, H. Schorle, J. Sage and S. K. Kim (2011). "PDGF signalling controls age-dependent proliferation in pancreatic [bgr]-cells." Nature **478**(7369): 349-355.
Chen, L., L. Stuart, T. K. Ohsumi, S. Burgess, G. K. Varshney, A. Dastur, M. Borowsky, C. Benes, A. Lacy-Hulbert and E. V. Schmidt (2013). "Transposon activation mutagenesis as a screening tool for identifying resistance to cancer therapeutics." BMC Cancer **13**(1): 1-15.

Chen, T., M. Li, Y. Ding, L. S. Zhang, Y. Xi, W. J. Pan, D. L. Tao, J. Y. Wang and L. Li (2009). "Identification of zinc-finger BED domain-containing 3 (Zbed3) as a novel Axin-interacting protein that activates Wnt/beta-catenin signaling." J Biol Chem **284**(11): 6683-6689.

Chiu, Y.-L., A. Ali, C.-y. Chu, H. Cao and T. M. Rana (2004). "Visualizing a Correlation between siRNA Localization, Cellular Uptake, and RNAi in Living Cells." Chemistry & Biology **11**(8): 1165-1175.

Cho, H., J. Mu, J. K. Kim, J. L. Thorvaldsen, Q. Chu, E. B. Crenshaw, K. H. Kaestner, M. S. Bartolomei, G. I. Shulman and M. J. Birnbaum (2001). "Insulin Resistance and a Diabetes Mellitus-Like Syndrome in Mice Lacking the Protein Kinase Akt2 (PKBβ)." Science **292**(5522): 1728-1731.

Choudhury, A. I., H. Heffron, M. A. Smith, H. Al-Qassab, A. W. Xu, C. Selman, M. Simmgen, M. Clements, M. Claret, G. Maccoll, D. C. Bedford, K. Hisadome, I. Diakonov, V. Moosajee, J. D. Bell, J. R. Speakman, R. L. Batterham, G. S. Barsh, M. L. Ashford and D. J. Withers (2005). "The role of insulin receptor substrate 2 in hypothalamic and beta cell function." J Clin Invest **115**(4): 940-950.

Christian, M., T. Cermak, E. L. Doyle, C. Schmidt, F. Zhang, A. Hummel, A. J. Bogdanove and D. F. Voytas (2010). "Targeting DNA Double-Strand Breaks with TAL Effector Nucleases." Genetics **186**(2): 757-761.

Chuang, T. H., X. Xu, V. Kaartinen, N. Heisterkamp, J. Groffen and G. M. Bokoch (1995). "Abr and Bcr are multifunctional regulators of the Rho GTP-binding protein family." Proceedings of the National Academy of Sciences **92**(22): 10282-10286.

Dai, Z., J. Sheridan, L. Gearing, D. Moore, S. Su, R. Dickins, M. Blewitt and M. Ritchie (2014). shRNA-seq data analysis with edgeR [version 1; referees: 3 approved].

Damcott, C. M., T. I. Pollin, L. J. Reinhart, S. H. Ott, H. Shen, K. D. Silver, B. D. Mitchell and A. R. Shuldiner (2006). "Polymorphisms in the Transcription Factor 7-Like 2 (TCF7L2) Gene Are Associated With Type 2 Diabetes in the Amish." Replication and Evidence for a Role in Both Insulin Secretion and Insulin Resistance **55**(9): 2654-2659.

Das, S., S. Ghosal, J. Chakrabarti and K. Kozak (2013). "SeedSeq: Off-Target Transcriptome Database." BioMed Research International **2013**: 9.

Datler, C. and S. Grimm (2013). "Reconstitution of CKMT1 expression fails to rescue cells from mitochondrial membrane potential dissipation: Implications for controlling RNAi experiments." Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1833**(12): 2844-2855.
Dégano, P., R. A. Silvestre, M. Salas, E. Peiró and J. Marco (1993). "Amylin inhibits glucose-induced insulin secretion in a dose-dependent manner. Study in the perfused rat pancreas." Regulatory Peptides **43**(1–2): 91-96.
Del Guerra, S., R. Lupi, L. Marselli, M. Masini, M. Bugliani, S. Sbrana, S. Torri, M. Pollera, U. Boggi, F. Mosca, S. Del Prato and P. Marchetti (2005). "Functional and molecular defects of pancreatic islets in human type 2 diabetes." Diabetes **54**(3): 727-735.
Deleu, L., F. Fuks, D. Spitkovsky, R. Horlein, S. Faisst and J. Rommelaere (1998). "Opposite transcriptional effects of cyclic AMP-responsive elements in confluent or p27KIP-overexpressing cells versus serum-starved or growing cells." Mol Cell Biol **18**(1): 409-419.
Delic, S., N. Lottmann, K. Jetschke, G. Reifenberger and M. J. Riemenschneider (2012). "Identification and functional validation of CDH11, PCSK6 and SH3GL3 as novel glioma invasion-associated candidate genes." Neuropathol Appl Neurobiol **38**(2): 201-212.
Deltcheva, E., K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel and E. Charpentier (2011). "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III." Nature **471**(7340): 602-607.
Deltour, S., S. Pinte, C. Guerardel and D. Leprince (2001). "Characterization of HRG22, a human homologue of the putative tumor suppressor gene HIC1." Biochem Biophys Res Commun **287**(2): 427-434.
Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse and C. Sotiriou (2007). "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series." Clin Cancer Res **13**(11): 3207-3214.
Devroe, E. and P. A. Silver (2002). "Retrovirus-delivered siRNA." BMC Biotechnol **2**: 15.

Diaz, A. A., H. Qin, M. Ramalho-Santos and J. S. Song (2015). "HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis." Nucleic Acids Res **43**(3): e16.

Do, O. H., J. E. Gunton, H. Y. Gaisano and P. Thorn (2016). "Changes in beta cell function occur in prediabetes and early disease in the Lepr (db) mouse model of diabetes." Diabetologia **59**(6): 1222-1230.

Dolberg, D. S., R. Hollingsworth, M. Hertle and M. J. Bissell (1985). "Wounding and its role in RSV-mediated tumor formation." Science **230**(4726): 676-678.

Dopie, J., E. K. Rajakyla, M. S. Joensuu, G. Huet, E. Ferrantelli, T. Xie, H. Jaalinoja, E. Jokitalo and M. K. Vartiainen (2015). "Genome-wide RNAi screen for nuclear actin reveals a network of cofilin regulators." J Cell Sci **128**(13): 2388-2400.

Dor, Y., J. Brown, O. I. Martinez and D. A. Melton (2004). "Adult pancreatic [beta]-cells are formed by self-duplication rather than stem-cell differentiation." Nature **429**(6987): 41-46.

Dorr, C., C. Janik, M. Weg, R. A. Been, J. Bader, R. Kang, B. Ng, L. Foran, S. R. Landman, M. G. O'Sullivan, M. Steinbach, A. L. Sarver, K. A. Silverstein, D. A. Largaespada and T. K. Starr (2015). "Transposon Mutagenesis Screen Identifies Potential Lung Cancer Drivers and CUL3 as a Tumor Suppressor." Mol Cancer Res **13**(8): 1238-1247.

Dosil, M. and X. R. Bustelo (2004). "Functional characterization of Pwp2, a WD family protein essential for the assembly of the 90 S pre-ribosomal particle." J Biol Chem **279**(36): 37385-37397.

Doyon, Y., J. M. McCammon, J. C. Miller, F. Faraji, C. Ngo, G. E. Katibah, R. Amora, T. D. Hocking, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov and S. L. Amacher (2008). "Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases." Nat Biotech **26**(6): 702-708.

Du, C., B. Ge, Z. Liu, K. Fu, W. C. Chan and T. W. McKeithan (2006). "PCR-based generation of shRNA libraries from cDNAs." BMC Biotechnology **6**(1): 1-11.

Dupraz, P., C. Rinsch, W. F. Pralong, E. Rolland, R. Zufferey, D. Trono and B. Thorens (1999). "Lentivirus-mediated Bcl-2 expression in bTC-tet cells improves resistance to hypoxia and cytokine-induced apoptosis while preserving in vitro and in vivo control of insulin secretion." Gene Therapy **6**: 1160-1169.

Dupuy, A. J., K. Akagi, D. A. Largaespada, N. G. Copeland and N. A. Jenkins (2005). "Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system." Nature **436**(7048): 221-226.
Dupuy, A. J., L. M. Rogers, J. Kim, K. Nannapaneni, T. K. Starr, P. Liu, D. A. Largaespada, T. E. Scheetz, N. A. Jenkins and N. G. Copeland (2009). "A modified Sleeping Beauty transposon system that can be used to model a wide variety of human cancers in mice." Cancer research **69**(20): 8150-8156.
Dutta, B., A. Azhir, L. H. Merino, Y. Guo, S. Revanur, P. B. Madhamshettiwar, R. N. Germain, J. A. Smith, K. J. Simpson, S. E. Martin, E. Beuhler and I. D. Fraser (2016). "An interactive web-based application for Comprehensive Analysis of RNAi-screen Data." Nat Commun **7**: 10578.
Dykes, I. M., K. L. van Bueren, R. J. Ashmore, T. Floss, W. Wurst, D. Szumska, S. Bhattacharya and P. J. Scambler (2014). "HIC2 Is a Novel Dosage-Dependent Regulator of Cardiac Development Located Within the Distal 22q11 Deletion Syndrome Region." Circulation Research **115**(1): 23-31.
Eagle, H. and E. M. Levine (1967). "Growth regulatory effects of cellular interaction." Nature **213**(5081): 1102-1106.
Easton, D. F., D. T. Bishop, D. Ford and G. P. Crockford (1993). "Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium." American Journal of Human Genetics **52**(4): 678-701.
Echeverri, C. J., P. A. Beachy, B. Baum, M. Boutros, F. Buchholz, S. K. Chanda, J. Downward, J. Ellenberg, A. G. Fraser, N. Hacohen, W. C. Hahn, A. L. Jackson, A. Kiger, P. S. Linsley, L. Lum, Y. Ma, B. Mathey-Prevot, D. E. Root, D. M. Sabatini, J. Taipale, N. Perrimon and R. Bernards (2006). "Minimizing the risk of reporting false positives in large-scale RNAi screens." Nat Methods **3**(10): 777-779.
Eddy, B. E., G. S. Borman, G. E. Grubbs and R. D. Young (1962). "Identification of the oncogenic substance in rhesus monkey kidney cell culture as simian virus 40." Virology **17**: 65-75.
Efanova, I. B., S. V. Zaitsev, B. Zhivotovsky, M. Köhler, S. Efendić, S. Orrenius and P.-O. Berggren (1998). "Glucose and Tolbutamide Induce Apoptosis in Pancreatic β-Cells: A PROCESS DEPENDENT ON INTRACELLULAR

Ca2+CONCENTRATION." Journal of Biological Chemistry **273**(50): 33501-33507.
Ehrlich, J. S., M. D. Hansen and W. J. Nelson (2002). "Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion." Dev Cell **3**(2): 259-270.
Eiberg, H., J. Mohr, K. Schmiegelow, L. S. Nielsen and R. Williamson (1985). "Linkage relationships of paraoxonase (PON) with other markers: indication of PON-cystic fibrosis synteny." Clin Genet **28**(4): 265-271.
Eid, C., M. Hémadi, N.-T. Ha-Duong and J.-M. El Hage Chahine (2014). "Iron uptake and transfer from ceruloplasmin to transferrin." Biochimica et Biophysica Acta (BBA) - General Subjects **1840**(6): 1771-1781.
Elbashir, S. M., J. Harborth, W. Lendeckel, A. Yalcin, K. Weber and T. Tuschl (2001). "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells." Nature **411**(6836): 494-498.
Elbashir, S. M., J. Martinez, A. Patkaniowska, W. Lendeckel and T. Tuschl (2001). "Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate." EMBO J **20**(23): 6877-6888.
Espinha, G., J. H. Osaki, Y. T. Magalhaes and F. L. Forti (2015). "Rac1 GTPase-deficient HeLa cells present reduced DNA repair, proliferation, and survival under UV or gamma irradiation." Molecular and Cellular Biochemistry **404**(1): 281-297.
Essafi, A., A. Webb, Rachel L. Berry, J. Slight, Sally F. Burn, L. Spraggon, V. Velecela, Ofelia M. Martinez-Estrada, John H. Wiltshire, Stefan G. E. Roberts, D. Brownstein, Jamie A. Davies, Nicholas D. Hastie and P. Hohenstein (2011). "A Wt1-Controlled Chromatin Switching Mechanism Underpins Tissue-Specific Wnt4 Activation and Repression." Developmental Cell **21**(3): 559-574.
Evellin, S., J. Nolte, K. Tysack, F. vom Dorp, M. Thiel, P. A. O. Weernink, K. H. Jakobs, E. J. Webb, J. W. Lomasney and M. Schmidt (2002). "Stimulation of Phospholipase C-ε by the M3Muscarinic Acetylcholine Receptor Mediated by Cyclic AMP and the GTPase Rap2B." Journal of Biological Chemistry **277**(19): 16805-16813.
Fan, C., G. Jiang, X. Zhang, Y. Miao, X. Lin, L. Luan, Z. Xu, Y. Zhang, H. Zhao, D. Liu and E. Wang (2015). "Zbed3 contributes to malignant phenotype of lung

cancer via regulating beta-catenin and P120-catenin 1." Mol Carcinog **54 Suppl 1**: E138-147.

Fatrai, S., L. Elghazi, N. Balcazar, C. Cras-Méneur, I. Krits, H. Kiyokawa and E. Bernal-Mizrachi (2006). "Akt Induces β-Cell Proliferation by Regulating Cyclin D1, Cyclin D2, and p21 Levels and Cyclin-Dependent Kinase-4 Activity." Diabetes **55**(2): 318-325.

Federici, M., M. Hribal, L. Perego, M. Ranalli, Z. Caradonna, C. Perego, L. Usellini, R. Nano, P. Bonini, F. Bertuzzi, L. N. Marlier, A. M. Davalli, O. Carandente, A. E. Pontiroli, G. Melino, P. Marchetti, R. Lauro, G. Sesti and F. Folli (2001). "High glucose causes apoptosis in cultured human pancreatic islets of Langerhans: a potential role for regulation of specific Bcl family genes toward an apoptotic cell death program." Diabetes **50**(6): 1290-1301.

Ferlay, S. I., D. Ervik M, M. C. Eser S, R. M., P. DM., F. D. and F. Bray. (2012). "Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11." from http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver and C. C. Mello (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-811.

Flaberg, E., H. Guven, A. Savchenko, T. Pavlova, V. Kashuba, L. Szekely and G. Klein (2012). "The architecture of fibroblast monolayers of different origin differentially influences tumor cell growth." Int J Cancer **131**(10): 2274-2283.

Flaberg, E., L. Markasz, G. Petranyi, G. Stuber, F. Dicso, N. Alchihabi, E. Olah, I. Csizy, T. Jozsa, O. Andren, J. E. Johansson, S. O. Andersson, G. Klein and L. Szekely (2011). "High-throughput live-cell imaging reveals differential inhibition of tumor cell proliferation by human fibroblasts." Int J Cancer **128**(12): 2793-2802.

Florez , J. C., K. A. Jablonski , N. Bayley , T. I. Pollin , P. I. W. de Bakker , A. R. Shuldiner , W. C. Knowler , D. M. Nathan  and D. Altshuler (2006). "TCF7L2 Polymorphisms and Progression to Diabetes in the Diabetes Prevention Program." New England Journal of Medicine **355**(3): 241-250.

Foulquier, F., M. Amyere, J. Jaeken, R. Zeevaert, E. Schollen, V. Race, R. Bammens, W. Morelle, C. Rosnoblet, D. Legrand, D. Demaegd, N. Buist, D. Cheillan, N. Guffon, P. Morsomme, W. Annaert, H. Freeze H , E.

Van Schaftingen, M. Vikkula and G. Matthijs (2012). "TMEM165 Deficiency Causes a Congenital Disorder of Glycosylation." Am J Hum Genet **91**(1): 15-26.
Friedrichsen, B. N., H. E. Richter, J. A. Hansen, C. J. Rhodes, J. H. Nielsen, N. Billestrup and A. Møldrup (2003). "Signal Transducer and Activator of Transcription 5 Activation Is Sufficient to Drive Transcriptional Induction of Cyclin D2 Gene and Proliferation of Rat Pancreatic β-Cells." Molecular Endocrinology **17**(5): 945-958.
Fritz, G., C. Brachetti, F. Bahlmann, M. Schmidt and B. Kaina (2002). "Rho GTPases in human breast tumours: expression and mutation analyses and correlation with clinical parameters." Br J Cancer **87**(6): 635-644.
Fu, A., A. C.-H. Ng, C. Depatie, N. Wijesekara, Y. He, G.-S. Wang, N. Bardeesy, F. W. Scott, R. M. Touyz, M. B. Wheeler and R. A. Screaton (2009). "Loss of Lkb1 in Adult β Cells Increases β Cell Mass and Enhances Glucose Tolerance in Mice." Cell Metabolism **10**(4): 285-295.
Fu, A., A. C. Ng, C. Depatie, N. Wijesekara, Y. He, G. S. Wang, N. Bardeesy, F. W. Scott, R. M. Touyz, M. B. Wheeler and R. A. Screaton (2009). "Loss of Lkb1 in adult beta cells increases beta cell mass and enhances glucose tolerance in mice." Cell Metab **10**(4): 285-295.
Fu, Y., J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung and J. D. Sander (2013). "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells." Nat Biotechnol **31**(9): 822-826.
Fuse, T., M. Tanikawa, M. Nakanishi, K. Ikeda, T. Tada, H. Inagaki, K. Asai, T. Kato and K. Yamada (2000). "p27Kip1 expression by contact inhibition as a prognostic index of human glioma." Journal of neurochemistry **74**(4): 1393-1399.
Galeev, R., A. Baudet, P. Kumar, A. Rundberg Nilsson, B. Nilsson, S. Soneji, T. Torngren, A. Borg, A. Kvist and J. Larsson (2016). "Genome-wide RNAi Screen Identifies Cohesin Genes as Modifiers of Renewal and Differentiation in Human HSCs." Cell Rep **14**(12): 2988-3000.
Gari, H. H., C. M. Gearheart, S. Fosmire, G. D. DeGala, Z. Fan, K. C. Torkko, S. M. Edgerton, M. S. Lucia, R. Ray, A. D. Thor, C. C. Porter and J. R. Lambert (2016). "Genome-wide functional genetic screen with the anticancer agent AMPI-109 identifies PRL-3 as an oncogenic driver in triple-negative breast cancers." Oncotarget **7**(13): 15757-15771.

Garneau, J. E., M. E. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadan and S. Moineau (2010). "The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA." Nature **468**(7320): 67-71.

Garofalo, R. S., S. J. Orena, K. Rafidi, A. J. Torchia, J. L. Stock, A. L. Hildebrandt, T. Coskran, S. C. Black, D. J. Brees, J. R. Wicks, J. D. McNeish and K. G. Coleman (2003). "Severe diabetes, age-dependent loss of adipose tissue, and mild growth deficiency in mice lacking Akt2/PKB beta." J Clin Invest **112**(2): 197-208.

Gazin, C., N. Wajapeyee, S. Gobeil, C.-M. Virbasius and M. R. Green (2007). "An elaborate pathway required for Ras-mediated epigenetic silencing." Nature **449**(7165): 1073-1077.

Genovesio, A., M. A. Giardini, Y. J. Kwon, F. de Macedo Dossin, S. Y. Choi, N. Y. Kim, H. C. Kim, S. Y. Jung, S. Schenkman, I. C. Almeida, N. Emans and L. H. Freitas-Junior (2011). "Visual genome-wide RNAi screening to identify human host factors required for Trypanosoma cruzi infection." PLoS One **6**(5): e19733.

Georgia, S. and A. Bhushan (2004). "Beta cell replication is the primary mechanism for maintaining postnatal beta cell mass." J Clin Invest **114**(7): 963-968.

Gerhard, D. S., L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brown-stein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei, S. Mathavan, P. H. Gunaratne, J. Wu, A. M. Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G.

Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak and J. Malek (2004). "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)." Genome Res **14**(10B): 2121-2127.

Geurts, A. M., G. J. Cost, Y. Freyvert, B. Zeitler, J. C. Miller, V. M. Choi, S. S. Jenkins, A. Wood, X. Cui, X. Meng, A. Vincent, S. Lam, M. Michalkiewicz, R. Schilling, J. Foeckler, S. Kalloway, H. Weiler, S. Ménoret, I. Anegon, G. D. Davis, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov, H. J. Jacob and R. Buelow (2009). "Knockout Rats via Embryo Microinjection of Zinc-Finger Nucleases." Science **325**(5939): 433-433.

Gilbert, D. F., G. Erdmann, X. Zhang, A. Fritzsche, K. Demir, A. Jaedicke, K. Muehlenberg, E. E. Wanker and M. Boutros (2011). "A Novel Multiplex Cell Viability Assay for High-Throughput RNAi Screening." PLoS ONE **6**(12): e28338.

Gilot, D., N. Le Meur, F. Giudicelli, M. Le Vée, D. Lagadic-Gossmann, N. Théret and O. Fardel (2011). "RNAi-Based Screening Identifies Kinases Interfering with Dioxin-Mediated Up-Regulation of CYP1A1 Activity." PLoS ONE **6**(3): e18261.

Goh, W. S., J. W. Seah, E. J. Harrison, C. Chen, C. M. Hammell and G. J. Hannon (2014). "A genome-wide RNAi screen identifies factors required for distinct stages of C. elegans piRNA biogenesis." Genes Dev **28**(7): 797-807.

Gonzalez, A., B. Merino, L. Marroqui, P. Neco, P. Alonso-Magdalena, E. Caballero-Garrido, E. Vieira, S. Soriano, R. Gomis, A. Nadal and I. Quesada (2013). "Insulin hypersecretion in islets from diet-induced hyperinsulinemic obese female mice is associated with several functional adaptations in individual beta-cells." Endocrinology **154**(10): 3515-3524.

Granot, Z., A. Swisa, J. Magenheim, M. Stolovich-Rain, W. Fujimoto, E. Manduchi, T. Miki, J. K. Lennerz, C. J. Stoeckert, Jr., O. Meyuhas, S. Seino, M. A. Permutt, H. Piwnica-Worms, N. Bardeesy and Y. Dor (2009). "LKB1 regulates pancreatic beta cell size, polarity, and function." Cell Metab **10**(4): 296-308.

Grant, S. F., G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K. P. Magnusson, G. B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A. Gylfason, J. Saemundsdottir, R. L. Wilensky, M. P. Reilly, D. J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J. R. Gulcher, A. Kong and K. Stefansson (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes." Nat Genet **38**(3): 320-323.

Gray, T. A., K. Azama, K. Whitmore, A. Min, S. Abe and R. D. Nicholls (2001). "Phylogenetic conservation of the makorin-2 gene, encoding a multiple zinc-finger protein, antisense to the RAF1 proto-oncogene." Genomics **77**(3): 119-126.

Green, I. C. and K. W. Taylor (1972). "Effects of pregnancy in the rat on the size and insulin secretory response of the islets of Langerhans." J Endocrinol **54**(2): 317-325.

Groves, C. J., E. Zeggini, J. Minton, T. M. Frayling, M. N. Weedon, N. W. Rayner, G. A. Hitman, M. Walker, S. Wiltshire, A. T. Hattersley and M. I. McCarthy (2006). "Association Analysis of 6,736 U.K. Subjects Provides Replication and Confirms TCF7L2 as a Type 2 Diabetes Susceptibility Gene With a Substantial Effect on Individual Risk." Diabetes **55**(9): 2640-2644.

Gu, S., L. Jin, F. Zhang, Y. Huang, D. Grimm, J. J. Rossi and M. A. Kay (2011). "Thermodynamic stability of small hairpin RNAs highly influences the loading process of different mammalian Argonautes." Proceedings of the National Academy of Sciences **108**(22): 9208-9213.

Gupta-Rossi, N., S. Ortica, V. Meas-Yedid, S. Heuss, J. Moretti, J.-C. Olivo-Marin and A. Israël (2011). "The Adaptor-associated Kinase 1, AAK1, Is a Positive Regulator of the Notch Pathway." Journal of Biological Chemistry **286**(21): 18720-18730.

Hakonen, E., J. Ustinov, I. Mathijs, J. Palgi, L. Bouwens, P. J. Miettinen and T. Otonkoski (2011). "Epidermal growth factor (EGF)-receptor signalling is needed for murine beta cell mass expansion in response to high-fat diet and pregnancy but not after pancreatic duct ligation." Diabetologia **54**(7): 1735-1743.

Halperin, F., X. Lopez, R. Manning, C. R. Kahn, R. N. Kulkarni and A. B. Goldfine (2012). "Insulin Augmentation of Glucose-Stimulated Insulin Secretion Is Impaired in Insulin-Resistant Humans." Diabetes **61**(2): 301-309.
Hall, A. (1998). "Rho GTPases and the Actin Cytoskeleton." Science **279**(5350): 509-514.
Hall, E., P. Volkov, T. Dayeh, K. Bacos, T. Rönn, M. D. Nitert and C. Ling (2014). "Effects of palmitate on genome-wide mRNA expression and DNA methylation patterns in human pancreatic islets." BMC Medicine **12**(1): 1-15.
Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey and M. C. King (1990). "Linkage of early-onset familial breast cancer to chromosome 17q21." Science **250**(4988): 1684-1689.
Hammond, S. M., E. Bernstein, D. Beach and G. J. Hannon (2000). "An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells." Nature **404**(6775): 293-296.
Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.
Harborth, J., S. M. Elbashir, K. Bechert, T. Tuschl and K. Weber (2001). "Identification of essential genes in cultured mammalian cells using small interfering RNAs." Journal of Cell Science **114**(24): 4557-4565.
Hashimoto, N., Y. Kido, T. Uchida, S. Asahara, Y. Shigeyama, T. Matsuda, A. Takeda, D. Tsuchihashi, A. Nishizawa, W. Ogawa, Y. Fujimoto, H. Okamura, K. C. Arden, P. L. Herrera, T. Noda and M. Kasuga (2006). "Ablation of PDK1 in pancreatic beta cells induces diabetes as a result of loss of beta cell mass." Nat Genet **38**(5): 589-593.
Hateboer, G., A. Gennissen, Y. F. Ramos, R. M. Kerkhoven, V. Sonntag-Buck, H. G. Stunnenberg and R. Bernards (1995). "BS69, a novel adenovirus E1A-associated protein that inhibits E1A transactivation." EMBO J **14**(13): 3159-3169.
Hayflick, L. (1965). "THE LIMITED IN VITRO LIFETIME OF HUMAN DIPLOID CELL STRAINS." Exp Cell Res **37**: 614-636.
Heit, J. J., A. A. Apelqvist, X. Gu, M. M. Winslow, J. R. Neilson, G. R. Crabtree and S. K. Kim (2006). "Calcineurin/NFAT signalling regulates pancreatic [beta]-cell growth and function." Nature **443**(7109): 345-349.

Heit, J. J., A. A. Apelqvist, X. Gu, M. M. Winslow, J. R. Neilson, G. R. Crabtree and S. K. Kim (2006). "Calcineurin/NFAT signalling regulates pancreatic beta-cell growth and function." Nature **443**(7109): 345-349.
Holmberg, C. G. and C. B. Laurell (1948). "Investigations in Serum Copper. II. Isolation of the Copper Containing Protein, and a Description of some of its Properties." Acta Chemica Scandinavica **2**.
Holst, C. R., G. J. Nuovo, M. Esteller, K. Chew, S. B. Baylin, J. G. Herman and T. D. Tlsty (2003). "Methylation of p16(INK4a) promoters occurs in vivo in histologically normal human mammary epithelia." Cancer Res **63**(7): 1596-1601.
Horiuchi, A., T. Imai, C. Wang, S. Ohira, Y. Feng, T. Nikaido and I. Konishi (2003). "Up-Regulation of Small GTPases, RhoA and RhoC, Is Associated with Tumor Progression in Ovarian Carcinoma." Lab Invest **83**(6): 861-870.
Hsu, P. D., D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao and F. Zhang (2013). "DNA targeting specificity of RNA-guided Cas9 nucleases." Nat Biotechnol **31**(9): 827-832.
Hsu, Y. C. and M. S. Perin (1995). "Human neuronal pentraxin II (NPTX2): conservation, genomic structure, and chromosomal localization." Genomics **28**(2): 220-227.
Hu, G., J. Kim, Q. Xu, Y. Leng, S. H. Orkin and S. J. Elledge (2009). "A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal." Genes & Development **23**(7): 837-848.
Hu, W. and H. Temin (1990). "Retroviral recombination and reverse transcription." Science **250**(4985): 1227-1233.
Huang, F., X. Tong, H. Deng, L. Fu and R. Zhang (2008). "Inhibition of the antigen-induced activation of RBL-2H3 cells by Gab2 siRNA." Cell Mol Immunol **5**(6): 433-438.
Huang, X., J. Y. Wang and X. Lu (2008). "Systems analysis of quantitative shRNA-library screens identifies regulators of cell adhesion." BMC Systems Biology **2**(1): 1-11.
Hull, R. L., K. Kodama, K. M. Utzschneider, D. B. Carr, R. L. Prigeon and S. E. Kahn (2005). "Dietary-fat-induced obesity in mice results in beta cell hyperplasia

but not increased insulin release: evidence for specificity of impaired beta cell adaptation." Diabetologia **48**(7): 1350-1358.
Hurlin, P. J., E. Steingrìmsson, N. G. Copeland, N. A. Jenkins and R. N. Eisenman (1999). "Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif." EMBO J **18**(24): 7019-7028.
Hwang, W. Y., Y. Fu, D. Reyon, M. L. Maeder, S. Q. Tsai, J. D. Sander, R. T. Peterson, J. R. J. Yeh and J. K. Joung (2013). "Efficient genome editing in zebrafish using a CRISPR-Cas system." Nat Biotech **31**(3): 227-229.
Illmensee, K. and B. Mintz (1976). "Totipotency and normal differentiation of single teratocarcinoma cells cloned by injection into blastocysts." Proc Natl Acad Sci U S A **73**(2): 549-553.
Inoki, K., Y. Li, T. Zhu, J. Wu and K.-L. Guan (2002). "TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling." Nat Cell Biol **4**(9): 648-657.
Iorns, E., T. M. Ward, S. Dean, A. Jegg, D. Thomas, N. Murugaesu, D. Sims, C. Mitsopoulos, K. Fenwick, I. Kozarewa, C. Naceur-Lombarelli, M. Zvelebil, C. M. Isacke, C. J. Lord, A. Ashworth, H. J. Hnatyszyn, M. Pegram and M. Lippman (2012). "Whole genome in vivo RNAi screening identifies the leukemia inhibitory factor receptor as a novel breast tumor suppressor." Breast Cancer Research and Treatment **135**(1): 79-91.
Irles, E., P. Neco, M. Lluesma, S. Villar-Pazos, J. C. Santos-Silva, J. F. Vettorazzi, P. Alonso-Magdalena, E. M. Carneiro, A. C. Boschero, A. Nadal and I. Quesada (2015). "Enhanced glucose-induced intracellular signaling promotes insulin hypersecretion: pancreatic beta-cell functional adaptations in a model of genetic obesity and prediabetes." Mol Cell Endocrinol **404**: 46-55.
Jackson, A. L., S. R. Bartz, J. Schelter, S. V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet and P. S. Linsley (2003). "Expression profiling reveals off-target gene regulation by RNAi." Nat Biotech **21**(6): 635-637.
Jackson, A. L., J. Burchard, J. Schelter, B. N. Chau, M. Cleary, L. Lim and P. S. Linsley (2006). "Widespread siRNA “off-target” transcript silencing mediated by seed region sequence complementarity." RNA **12**(7): 1179-1187.
Jarjour, J., H. West-Foyle, M. T. Certo, C. G. Hubert, L. Doyle, M. M. Getz, B. L. Stoddard and A. M. Scharenberg (2009). "High-resolution profiling of homing

endonuclease binding and catalytic specificity using yeast surface display." Nucleic Acids Res **37**(20): 6871-6880.
Javaherian, A., M. Vaccariello, N. E. Fusenig and J. A. Garlick (1998). "Normal keratinocytes suppress early stages of neoplastic progression in stratified epithelium." Cancer Res **58**(10): 2200-2208.
Jia, Y., L. Yuan, W. Hu, Y. Luo, L. Suo, M. Yang, S. Chen, Y. Wang, H. Liu, G. Yang and L. Li (2014). "Zinc-finger BED domain-containing 3 (Zbed3) is a novel secreted protein associated with insulin resistance in humans." J Intern Med **275**(5): 522-533.
Jiang, P., H. Wang, W. Li, C. Zang, B. Li, Y. J. Wong, C. Meyer, J. S. Liu, J. C. Aster and X. S. Liu (2015). "Network analysis of gene essentiality in functional genomics experiments." Genome Biology **16**(1): 1-10.
Jiang, W., D. Bikard, D. Cox, F. Zhang and L. A. Marraffini (2013). "RNA-guided editing of bacterial genomes using CRISPR-Cas systems." Nat Biotech **31**(3): 233-239.
Jiang, X., H. Pan, J. F. Nabhan, R. Krishnan, C. Koziol-White, R. A. Panettieri and Q. Lu (2012). "A novel EST-derived RNAi screen reveals a critical role for farnesyl diphosphate synthase in β2-adrenergic receptor internalization and down-regulation." The FASEB Journal **26**(5): 1995-2007.
Jiao, K., Y. Zhou and B. L. Hogan (2002). "Identification of mZnf8, a mouse Kruppel-like transcriptional repressor, as a novel nuclear interaction partner of Smad1." Mol Cell Biol **22**(21): 7633-7644.
Johnstone, K. A., E. Diakogiannaki, S. Dhayal, N. G. Morgan and L. W. Harries (2011). "Dysregulation of Hnf1b gene expression in cultured beta-cells in response to cytotoxic fatty acid." JOP **12**(1): 6-10.
Jolly, A. L., C. H. Luan, B. E. Dusel, S. F. Dunne, M. Winding, V. J. Dixit, C. Robins, J. L. Saluk, D. J. Logan, A. E. Carpenter, M. Sharma, D. Dean, A. R. Cohen and V. I. Gelfand (2016). "A Genome-wide RNAi Screen for Microtubule Bundle Formation and Lysosome Motility Regulation in Drosophila S2 Cells." Cell Rep **14**(3): 611-620.
Kahn, S. E., S. Zraika, K. M. Utzschneider and R. L. Hull (2009). "The beta cell lesion in type 2 diabetes: there has to be a primary functional abnormality." Diabetologia **52**(6): 1003-1012.

Kamai, T., T. Tsujii, K. Arai, K. Takagi, H. Asami, Y. Ito and H. Oshima (2003). "Significant association of Rho/ROCK pathway with invasion and metastasis of bladder cancer." Clinical Cancer Research **9**(7): 2632-2641.

Kampmann, M., M. A. Horlbeck, Y. Chen, J. C. Tsai, M. C. Bassik, L. A. Gilbert, J. E. Villalta, S. C. Kwon, H. Chang, V. N. Kim and J. S. Weissman (2015). "Next-generation libraries for robust RNA interference-based genome-wide screens." Proc Natl Acad Sci U S A **112**(26): E3384-3391.

Kaneko, K., K. Ueki, N. Takahashi, S. Hashimoto, M. Okamoto, M. Awazawa, Y. Okazaki, M. Ohsugi, K. Inabe, T. Umehara, M. Yoshida, M. Kakei, T. Kitamura, J. Luo, R. N. Kulkarni, C. R. Kahn, H. Kasai, L. C. Cantley and T. Kadowaki (2010). "Class IA phosphatidylinositol 3-kinase in pancreatic beta cells controls insulin secretion by multiple mechanisms." Cell Metab **12**(6): 619-632.

Karnik, S. K., H. Chen, G. W. McLean, J. J. Heit, X. Gu, A. Y. Zhang, M. Fontaine, M. H. Yen and S. K. Kim (2007). "Menin Controls Growth of Pancreatic ß-Cells in Pregnant Mice and Promotes Gestational Diabetes Mellitus." Science **318**(5851): 806-809.

Kerem, B., J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald and L. C. Tsui (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.

Khvorova, A., A. Reynolds and S. D. Jayasena (2003). "Functional siRNAs and miRNAs exhibit strand bias." Cell **115**(2): 209-216.

Kim, S. D., J. L. Liu, T. Roscioli, M. F. Buckley, G. Yagnik, S. A. Boyadjiev and J. Kim (2012). "Leucine-rich repeat, immunoglobulin-like and transmembrane domain 3 (LRIT3) is a modulator of FGFR1." FEBS Lett **586**(10): 1516-1521.

Kim, Y. G., J. Cha and S. Chandrasegaran (1996). "Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain." Proc Natl Acad Sci U S A **93**(3): 1156-1160.

Kloppel, G., M. Lohr, K. Habich, M. Oberholzer and P. U. Heitz (1985). "Islet pathology and the pathogenesis of type 1 and type 2 diabetes mellitus revisited." Surv Synth Pathol Res **4**(2): 110-125.

Knott, S. R. V., A. Maceli, N. Erard, K. Chang, K. Marran, X. Zhou, A. Gordon, O. E. Demerdash, E. Wagenblast, S. Kim, C. Fellmann and G. J. Hannon (2014). "A computational algorithm to predict shRNA potency." Mol Cell **56**(6): 796-807.

Knowlton, R. G., O. Cohen-Haguenauer, N. Van Cong, J. Frezal, V. A. Brown, D. Barker, J. C. Braman, J. W. Schumm, L.-C. Tsui, M. Buchwald and H. Donis-Keller (1985). "A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7." Nature **318**(6044): 380-382.

Koike-Yusa, H., Y. Li, E. P. Tan, M. D. C. Velasco-Herrera and K. Yusa (2014). "Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library." Nat Biotech **32**(3): 267-273.

Konarkowska, B., J. F. Aitken, J. Kistler, S. Zhang and G. J. S. Cooper (2006). "The aggregation potential of human amylin determines its cytotoxicity towards islet β-cells." FEBS Journal **273**(15): 3614-3624.

Kouhara, H., Y. R. Hadari, T. Spivak-Kroizman, J. Schilling, D. Bar-Sagi, I. Lax and J. Schlessinger (1997). "A lipid-anchored Grb2-binding protein that links FGF-receptor activation to the Ras/MAPK signaling pathway." Cell **89**(5): 693-702.

Koul, H. K., M. Pal and S. Koul (2013). "Role of p38 MAP Kinase Signal Transduction in Solid Tumors." Genes & Cancer **4**(9-10): 342-359.

Kretzschmar, M., J. Doody and J. Massagu (1997). "Opposing BMP and EGF signalling pathways converge on the TGF-[beta] family mediator Smad1." Nature **389**(6651): 618-622.

Kubota, N., K. Tobe, Y. Terauchi, K. Eto, T. Yamauchi, R. Suzuki, Y. Tsubamoto, K. Komeda, R. Nakano, H. Miki, S. Satoh, H. Sekihara, S. Sciacchitano, M. Lesniak, S. Aizawa, R. Nagai, S. Kimura, Y. Akanuma, S. I. Taylor and T. Kadowaki (2000). "Disruption of insulin receptor substrate 2 causes type 2 diabetes because of liver insulin resistance and lack of compensatory beta-cell hyperplasia." Diabetes **49**(11): 1880-1889.

Kulkarni, R. N., M. Holzenberger, D. Q. Shih, U. Ozcan, M. Stoffel, M. A. Magnuson and C. R. Kahn (2002). "[beta]-cell-specific deletion of the Igf1 receptor leads to hyperinsulinemia and glucose intolerance but does not alter [beta]-cell mass." Nat Genet **31**(1): 111-115.

Kumar, P., G. Goh, S. Wongphayak, D. Moreau and F. Bard (2013). "ScreenSifter: analysis and visualization of RNAi screening data." BMC Bioinformatics **14**(1): 1-12.

Kurahara, H., C. Bohl, S. Natsugoe, Y. Nishizono, S. Harihar, R. Sharma, T. Iwakuma and D. R. Welch (2016). "Suppression of pancreatic cancer growth

and metastasis by HMP19 identified through genome-wide shRNA screen." International Journal of Cancer **139**(3): 628-638.
Kushwaha, D., C. O'Leary, K. R. Cron, P. Deraska, K. Zhu, A. D. D'Andrea and D. Kozono (2015). "USP9X inhibition promotes radiation-induced apoptosis in non-small cell lung cancer cells expressing mid-to-high MCL1." Cancer Biol Ther **16**(3): 392-401.
Kwak, P. B. and Y. Tomari (2012). "The N domain of Argonaute drives duplex unwinding during RISC assembly." Nat Struct Mol Biol **19**(2): 145-151.
La Rocca, S. A., M. Grossi, G. Falcone, S. Alema and F. Tato (1989). "Interaction with normal cells suppresses the transformed phenotype of v-myc-transformed quail muscle cells." Cell **58**(1): 123-131.
Ladendorff, N. E., S. Wu and J. S. Lipsick (2001). "BS69, an adenovirus E1A-associated protein, inhibits the transcriptional activity of c-Myb." Oncogene **20**(1): 125-132.
Lalloyer, F., B. Vandewalle, F. Percevault, G. Torpier, J. Kerr-Conte, M. Oosterveer, R. Paumelle, J. C. Fruchart, F. Kuipers, F. Pattou, C. Fievet and B. Staels (2006). "Peroxisome proliferator-activated receptor alpha improves pancreatic adaptation to insulin resistance in obese mice and reduces lipotoxicity in human islets." Diabetes **55**(6): 1605-1613.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki,

P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen and J. Szustakowki (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Landi, A., J. Vermeire, V. Iannucci, H. Vanderstraeten, E. Naessens, M. Bentahir and B. Verhasselt (2014). "Genome-wide shRNA screening identifies host factors involved in early endocytic events for HIV-1-induced CD4 down-regulation." Retrovirology **11**: 118.

Lau, N. C., L. P. Lim, E. G. Weinstein and D. P. Bartel (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.
Le Novère, N., P.-J. Corringer and J.-P. Changeux (2002). "The diversity of subunit composition in nAChRs: Evolutionary origins, physiologic and pharmacologic consequences." Journal of Neurobiology **53**(4): 447-456.
Lee, K. Y., K. Y. Chan, K. S. Tsang, Y. C. Chen, H. F. Kung, P. C. Ng, C. K. Li, K. T. Leung and K. Li (2014). "Ubiquitous expression of MAKORIN-2 in normal and malignant hematopoietic cells and its growth promoting activity." PLoS One **9**(3): e92706.
Lee, N. K., J. P. Skinner, J. D. Zajac and H. E. MacLean (2011). "Ornithine decarboxylase is upregulated by the androgen receptor in skeletal muscle and regulates myoblast proliferation." Am J Physiol Endocrinol Metab **301**(1): E172-179.
Lee, S. B. and M. Esteban (1994). "The Interferon-induced Double-Stranded RNA-Activated Protein Kinase Induces Apoptosis." Virology **199**(2): 491-496.
Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim and V. N. Kim (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.
Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek and V. N. Kim (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-4060.
Lewis, B. P., C. B. Burge and D. P. Bartel (2005). "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets." Cell **120**(1): 15-20.
Li, D., Z. Qiu, Y. Shao, Y. Chen, Y. Guan, M. Liu, Y. Li, N. Gao, L. Wang, X. Lu, Y. Zhao and M. Liu (2013). "Heritable gene targeting in the mouse and rat using a CRISPR-Cas system." Nat Biotech **31**(8): 681-683.
Li, J., S. C. Ding, H. Cho, B. C. Chung, M. Gale, S. K. Chanda and M. S. Diamond (2013). "A Short Hairpin RNA Screen of Interferon-Stimulated Genes Identifies a Novel Negative Regulator of the Cellular Antiviral Response." mBio **4**(3).
Li, J., C. Yen, D. Liaw, K. Podsypanina, S. Bose, S. I. Wang, J. Puc, C. Miliaresis, L. Rodgers, R. McCombie, S. H. Bigner, B. C. Giovanella, M.

Ittmann, B. Tycko, H. Hibshoosh, M. H. Wigler and R. Parsons (1997). "PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer." Science **275**(5308): 1943-1947.

Lima, M. J., H. M. Docherty, Y. Chen, L. Vallier and K. Docherty (2012). "Pancreatic transcription factors containing protein transduction domains drive mouse embryonic stem cells towards endocrine pancreas." PLoS One **7**(5): e36481.

Lin, X., X. Ruan, M. G. Anderson, J. A. McDowell, P. E. Kroeger, S. W. Fesik and Y. Shen (2005). "siRNA-mediated off-target gene silencing triggered by a 7 nt complementation." Nucleic Acids Research **33**(14): 4527-4535.

Lindstrom, M. S. and M. Nister (2010). "Silencing of ribosomal protein S9 elicits a multitude of cellular responses inhibiting the growth of cancer cells subsequent to p53 activation." PLoS One **5**(3): e9578.

List, M., S. Schmidt, H. Christiansen, M. Rehmsmeier, Q. Tan, J. Mollenhauer and J. Baumbach (2016). "Comprehensive analysis of high-throughput screens with HiTSeekR." Nucleic Acids Res.

Liu, H., M. S. Remedi, K. L. Pappan, G. Kwon, N. Rohatgi, C. A. Marshall and M. L. McDaniel (2009). "Glycogen Synthase Kinase-3 and Mammalian Target of Rapamycin Pathways Contribute to DNA Synthesis, Cell Cycle Progression, and Proliferation in Human Islets." Diabetes **58**(3): 663-672.

Loi, S., B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. A. Foekens, J. G. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart and C. Sotiriou (2007). "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade." J Clin Oncol **25**(10): 1239-1246.

Lopez, X., A. Cypess, R. Manning, S. O'Shea, R. N. Kulkarni and A. B. Goldfine (2011). "Exogenous Insulin Enhances Glucose-Stimulated Insulin Response in Healthy Humans Independent of Changes in Free Fatty Acids." The Journal of Clinical Endocrinology & Metabolism **96**(12): 3811-3821.

Lorenzo, A., B. Razzaboni, G. C. Weir and B. A. Yankner (1994). "Pancreatic islet cell toxicity of amylin associated with type-2 diabetes mellitus." Nature **368**(6473): 756-760.

Lu, X. J., X. Qi, D. H. Zheng and L. J. Ji (2015). "Modeling cancer processes with CRISPR-Cas9." Trends Biotechnol **33**(6): 317-319.

Lu, Y., P. L. Herrera, Y. Guo, D. Sun, Z. Tang, D. LeRoith and J.-L. Liu (2004). "Pancreatic-Specific Inactivation of IGF-I Gene Causes Enlarged Pancreatic Islets and Significant Resistance to Diabetes." Diabetes **53**(12): 3131-3141.
Luciani, D. S., S. A. White, S. B. Widenmaier, V. V. Saran, F. Taghizadeh, X. Hu, M. F. Allard and J. D. Johnson (2013). "Bcl-2 and Bcl-xL suppress glucose signaling in pancreatic beta-cells." Diabetes **62**(1): 170-182.
Lund, H., R. F. Cowburn, E. Gustafsson, K. Stromberg, A. Svensson, L. Dahllund, D. Malinowsky and D. Sunnemark (2013). "Tau-tubulin kinase 1 expression, phosphorylation and co-localization with phospho-Ser422 tau in the Alzheimer's disease brain." Brain Pathol **23**(4): 378-389.
Luo, B., H. W. Cheung, A. Subramanian, T. Sharifnia, M. Okamoto, X. Yang, G. Hinkle, J. S. Boehm, R. Beroukhim, B. A. Weir, C. Mermel, D. A. Barbie, T. Awad, X. Zhou, T. Nguyen, B. Piqani, C. Li, T. R. Golub, M. Meyerson, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini and D. E. Root (2008). "Highly parallel identification of essential genes in cancer cells." Proceedings of the National Academy of Sciences **105**(51): 20380-20385.
Lupi, R., F. Dotta, L. Marselli, S. Del Guerra, M. Masini, C. Santangelo, G. Patane, U. Boggi, S. Piro, M. Anello, E. Bergamini, F. Mosca, U. Di Mario, S. Del Prato and P. Marchetti (2002). "Prolonged exposure to free fatty acids has cytostatic and pro-apoptotic effects on human pancreatic islets: evidence that beta-cell death is caspase mediated, partially dependent on ceramide pathway, and Bcl-2 regulated." Diabetes **51**(5): 1437-1442.
Maclean, N. and R. F. Ogilvie (1955). "Quantitative Estimation of the Pancreatic Islet Tissue in Diabetic Subjects." Diabetes **4**(5): 367-376.
Maedler, K., P. Sergeev, F. Ris, J. Oberholzer, H. I. Joller-Jemelka, G. A. Spinas, N. Kaiser, P. A. Halban and M. Y. Donath (2002). "Glucose-induced beta cell production of IL-1beta contributes to glucotoxicity in human pancreatic islets." J Clin Invest **110**(6): 851-860.
Maedler, K., G. A. Spinas, D. Dyntar, W. Moritz, N. Kaiser and M. Y. Donath (2001). "Distinct Effects of Saturated and Monounsaturated Fatty Acids on β-Cell Turnover and Function." Diabetes **50**(1): 69-76.
Maedler, K., G. A. Spinas, R. Lehmann, P. Sergeev, M. Weber, A. Fontana, N. Kaiser and M. Y. Donath (2001). "Glucose Induces β-Cell Apoptosis Via Upregulation of the Fas Receptor in Human Islets." Diabetes **50**(8): 1683-1690.

Maehama, T. and J. E. Dixon (1998). "The Tumor Suppressor, PTEN/MMAC1, Dephosphorylates the Lipid Second Messenger, Phosphatidylinositol 3,4,5-Trisphosphate." Journal of Biological Chemistry **273**(22): 13375-13378.

Maia, A. F., M. E. Tanenbaum, M. Galli, D. Lelieveld, D. A. Egan, R. Gassmann, C. E. Sunkel, S. van den Heuvel and R. H. Medema (2015). "Genome-wide RNAi screen for synthetic lethal interactions with the C. elegans kinesin-5 homolog BMK-1." Sci Data **2**: 150020.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville and G. M. Church (2013). "RNA-guided human genome engineering via Cas9." Science **339**(6121): 823-826.

Manning, B. D., A. R. Tee, M. N. Logsdon, J. Blenis and L. C. Cantley (2002). "Identification of the Tuberous Sclerosis Complex-2 Tumor Suppressor Gene Product Tuberin as a Target of the Phosphoinositide 3-Kinase/Akt Pathway." Molecular Cell **10**(1): 151-162.

Mao, X., K. D. Dillon, M. F. McEntee, A. M. Saxton and J. H. Kim (2014). "Islet Insulin Secretion, β-Cell Mass, and Energy Balance in a Polygenic Mouse Model of Type 2 Diabetes With Obesity." Journal of Inborn Errors of Metabolism & Screening **2**.

Marcotte, R., K. R. Brown, F. Suarez, A. Sayad, K. Karamboulas, P. M. Krzyzanowski, F. Sircoulomb, M. Medrano, Y. Fedyshyn, J. L. Y. Koh, D. van Dyk, B. Fedyshyn, M. Luhova, G. C. Brito, F. J. Vizeacoumar, F. S. Vizeacoumar, A. Datti, D. Kasimer, A. Buzina, P. Mero, C. Misquitta, J. Normand, M. Haider, T. Ketela, J. L. Wrana, R. Rottapel, B. G. Neel and J. Moffat (2012). "Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells." Cancer Discovery **2**(2): 172-189.

Mari, A., A. Tura, A. Natali, C. Anderwald, B. Balkau, N. Lalic, M. Walker and E. Ferrannini (2011). "Influence of Hyperinsulinemia and Insulin Resistance on In Vivo β-Cell Function." Their Role in Human β-Cell Dysfunction **60**(12): 3141-3147.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays." Genome Research **18**(9): 1509-1517.

Marr, R. A., H. Guan, E. Rockenstein, M. Kindy, F. H. Gage, I. Verma, E. Masliah and L. B. Hersh (2004). "Neprilysin regulates amyloid Beta peptide levels." J Mol Neurosci **22**(1-2): 5-11.

Marraffini, L. A. and E. J. Sontheimer (2008). "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA." Science **322**(5909): 1843-1845.

Martin, B. C., J. H. Warram, A. S. Krolewski, R. N. Bergman, J. S. Soeldner and C. R. Kahn (1992). "Role of glucose and insulin resistance in development of type 2 diabetes mellitus: results of a 25-year follow-up study." Lancet **340**(8825): 925-929.

Martinez, J., A. Patkaniowska, H. Urlaub, R. Luhrmann and T. Tuschl (2002). "Single-stranded antisense siRNAs guide target RNA cleavage in RNAi." Cell **110**(5): 563-574.

Martz, E. and M. S. Steinberg (1972). "The role of cell-cell contact in "contact" inhibition of cell division: a review and new evidence." J Cell Physiol **79**(2): 189-210.

Maserati, M., X. Dai, M. Walentuk and J. Mager (2014). "Identification of four genes required for mammalian blastocyst formation." Zygote **22**(3): 331-339.

Medina, D., F. Shepherd and T. Gropp (1978). "Enhancement of the tumorigenicity of preneoplastic mammary nodule lines by enzymatic dissociation." J Natl Cancer Inst **60**(5): 1121-1126.

Mehta, P. P., J. S. Bertram and W. R. Loewenstein (1986). "Growth inhibition of transformed cells correlates with their junctional communication with normal cells." Cell **44**(1): 187-196.

Meier, J. J., A. E. Butler, Y. Saisho, T. Monchamp, R. Galasso, A. Bhushan, R. A. Rizza and P. C. Butler (2008). "β-Cell Replication Is the Primary Mechanism Subserving the Postnatal Expansion of β-Cell Mass in Humans." Diabetes **57**(6): 1584-1594.

Mendes-Pereira, A. M., D. Sims, T. Dexter, K. Fenwick, I. Assiotis, I. Kozarewa, C. Mitsopoulos, J. Hakas, M. Zvelebil, C. J. Lord and A. Ashworth (2012). "Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen." Proceedings of the National Academy of Sciences **109**(8): 2730-2735.

Metukuri, M. R., P. Zhang, M. K. Basantani, C. Chin, R. E. Stamateris, L. C. Alonso, K. K. Takane, R. Gramignoli, S. C. Strom, R. M. O'Doherty, A. F. Stewart, R. C. Vasavada, A. Garcia-Ocaña and D. K. Scott (2012). "ChREBP Mediates Glucose-Stimulated Pancreatic β-Cell Proliferation." Diabetes **61**(8): 2004-2015.

Miele, C., G. A. Raciti, A. Cassese, C. Romano, F. Giacco, F. Oriente, F. Paturzo, F. Andreozzi, A. Zabatta, G. Troncone, F. Bosch, A. Pujol, H. Chneiweiss, P. Formisano and F. Beguinot (2007). "PED/PEA-15 Regulates Glucose-Induced Insulin Secretion by Restraining Potassium Channel Expression in Pancreatic β-Cells." Diabetes **56**(3): 622-633.

Miki, Y., J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding and et al. (1994). "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1." Science **266**(5182): 66-71.

Milo-Landesman, D., M. Surana, I. Berkovich, A. Compagni, G. Christofori, N. Fleischer and S. Efrat (2001). "Correction of Hyperglycemia in Diabetic Mice Transplanted With Reversibly Immortalized Pancreatic β Cells Controlled by the tet-on Regulatory System." Cell Transplantation **10**(7): 645-650.

Mockenhaupt, S., S. Grosse, D. Rupp, R. Bartenschlager and D. Grimm (2015). "Alleviation of off-target effects from vector-encoded shRNAs via codelivered RNA decoys." Proceedings of the National Academy of Sciences **112**(30): E4007-E4016.

Moffat, J., D. A. Grueneberg, X. Yang, S. Y. Kim, A. M. Kloepfer, G. Hinkle, B. Piqani, T. M. Eisenhaure, B. Luo and J. K. Grenier (2006). "A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen." Cell **124**.

Moffat, J., D. A. Grueneberg, X. Yang, S. Y. Kim, A. M. Kloepfer, G. Hinkle, B. Piqani, T. M. Eisenhaure, B. Luo, J. K. Grenier, A. E. Carpenter, S. Y. Foo, S. A. Stewart, B. R. Stockwell, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini and D. E. Root (2006). "A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen." Cell **124**(6): 1283-1298.

Mojica, F. J., C. Diez-Villasenor, J. Garcia-Martinez and E. Soria (2005). "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements." J Mol Evol **60**(2): 174-182.
Moore, J. K. and J. E. Haber (1996). "Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae." Mol Cell Biol **16**(5): 2164-2173.
Morioka, T., E. Asilmaz, J. Hu, J. F. Dishinger, A. J. Kurpad, C. F. Elias, H. Li, J. K. Elmquist, R. T. Kennedy and R. N. Kulkarni (2007). "Disruption of leptin receptor expression in the pancreas directly affects beta cell growth and function in mice." J Clin Invest **117**(10): 2860-2868.
Moscou, M. J. and A. J. Bogdanove (2009). "A Simple Cipher Governs DNA Recognition by TAL Effectors." Science **326**(5959): 1501-1501.
Nakamura, Y., R. Moriuchi, D. Nakayama, I. Yamashita, Y. Higashiyama, T. Yamamoto, Y. Kusano, S. Hino, T. Miyamoto and S. Katamine (1994). "Altered expression of a novel cellular gene as a consequence of integration of human T cell lymphotropic virus type 1." Journal of General Virology **75**(10): 2625-2633.
Nakayama, M., T. Ishibashi, H. O. Ishikawa, H. Sato, T. Usui, T. Okuda, H. Yashiro, H. Ishikawa, Y. Taikou, A. Minami, K. Kato, M. Taki, T. Aigaki, W. Gunji, M. Ohtsu, Y. Murakami, S.-i. Tanuma, A. Tsuboi, M. Adachi, J. Kuroda, T. Sasamura, T. Yamakawa and K. Matsuno (2014). "A gain-of-function screen to identify genes that reduce lifespan in the adult of Drosophila melanogaster." BMC Genetics **15**(1): 1-14.
Narod, S. A., J. Feunteun, H. T. Lynch, P. Watson, T. Conway, J. Lynch and G. M. Lenoir (1991). "Familial breast-ovarian cancer locus on chromosome 17q12-q23." Lancet **338**(8759): 82-83.
Neuhausen, S. L., J. Swensen, Y. Miki, Q. Llu, S. Tavtiglan, D. Shattuck-Eidens, A. Kamb, M. R. Hobbs, J. Gingrich, H. Shizuya, U.-J. Kim, C. Cochran, P. A. Futreal, R. W. Wiseman, H. T. Lynch, P. Tonin, S. Narod, L. Cannon-Albright, M. H. Skolnick and D. E. Goldgar (1994). "A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene BRCA1." Human Molecular Genetics **3**(11): 1919-1926.
Nica, A. C., H. Ongen, J.-C. Irminger, D. Bosco, T. Berney, S. E. Antonarakis, P. A. Halban and E. T. Dermitzakis (2013). "Cell-type, allelic, and genetic

signatures in the human pancreatic beta cell transcriptome." Genome Research **23**(9): 1554-1562.
Nicholls, A. J. (2015). Development of a shRNA screen to identify effectors of neighbour suppression of tumour cell growth. Doctor of Philosophy, Universities of Exeter and Plymouth.
Nykänen, A., B. Haley and P. D. Zamore (2001). "ATP Requirements and Small Interfering RNA Structure in the RNA Interference Pathway." Cell **107**(3): 309-321.
O'Brien, T. D., A. E. Butler, P. C. Roche, K. H. Johnson and P. C. Butler (1994). "Islet amyloid polypeptide in human insulinomas. Evidence for intracellular amyloidogenesis." Diabetes **43**(2): 329-336.
O'Brien, T. D., P. Westermark and K. H. Johnson (1990). "Islet amyloid polypeptide (IAPP) does not inhibit glucose-stimulated insulin secretion from isolated perfused rat pancreas." Biochemical and Biophysical Research Communications **170**(3): 1223-1228.
O'Connell, B. C., B. Adamson, J. R. Lydeard, M. E. Sowa, A. Ciccia, A. L. Bredemeyer, M. Schlabach, S. P. Gygi, S. J. Elledge and J. W. Harper (2010). "A Genome-wide Camptothecin Sensitivity Screen Identifies a Mammalian MMS22L-NFKBIL2 Complex Required for Genomic Stability." Molecular Cell **40**(4): 645-657.
O'Reilly, L. P., O. S. Long, M. C. Cobanoglu, J. A. Benson, C. J. Luke, M. T. Miedel, P. Hale, D. H. Perlmutter, I. Bahar, G. A. Silverman and S. C. Pak (2014). "A genome-wide RNAi screen identifies potential drug targets in a C. elegans model of alpha1-antitrypsin deficiency." Hum Mol Genet **23**(19): 5123-5132.
O'Reilly, P. G., S. Wagner, D. J. Franks, K. Cailliau, E. Browaeys, C. Dissous and L. A. Sabourin (2005). "The Ste20-like Kinase SLK Is Required for Cell Cycle Progression through G2." Journal of Biological Chemistry **280**(51): 42383-42390.
O'Donohue, M.-F., V. Choesmel, M. Faubladier, G. Fichant and P.-E. Gleizes (2010). "Functional dichotomy of ribosomal proteins during the synthesis of mammalian 40S ribosomal subunits." The Journal of Cell Biology **190**(5): 853-866.

Obchoei, S., S. M. Weakley, S. Wongkham, C. Wongkham, K. Sawanyawisuth, Q. Yao and C. Chen (2011). "Cyclophilin A enhances cell proliferation and tumor growth of liver fluke-associated cholangiocarcinoma." Mol Cancer **10**: 102.

Oehmig, A., A. Klotzbücher, M. Thomas, F. Weise, U. Hagner, R. Brundiers, D. Waldherr, A. Lingnau, A. Knappik, M. H. Kubbutat, T. O. Joos and H. Volkmer (2008). "A novel reverse transduction adenoviral array for the functional analysis of shRNA libraries." BMC Genomics **9**(1): 1-13.

Ohrt, T., D. Merkle, K. Birkenfeld, C. J. Echeverri and P. Schwille (2006). "In situ fluorescence analysis demonstrates active siRNA exclusion from the nucleus by Exportin 5." Nucleic Acids Res **34**(5): 1369-1380.

Ohshige, T., M. Iwata, S. Omori, Y. Tanaka, H. Hirose, K. Kaku, H. Maegawa, H. Watada, A. Kashiwagi, R. Kawamori, K. Tobe, T. Kadowaki, Y. Nakamura and S. Maeda (2011). "Association of New Loci Identified in European Genome-Wide Association Studies with Susceptibility to Type 2 Diabetes in the Japanese." PLoS ONE **6**(10): e26911.

Okada, A. and Y. Tomooka (2013). "A role of Sema6A expressed in oligodendrocyte precursor cells." Neuroscience Letters **539**: 48-53.

Okada, T., C. W. Liew, J. Hu, C. Hinault, M. D. Michael, J. Kŕtzfeldt, C. Yin, M. Holzenberger, M. Stoffel and R. N. Kulkarni (2007). "Insulin receptors in β-cells are critical for islet compensatory growth response to insulin resistance." Proceedings of the National Academy of Sciences **104**(21): 8977-8982.

Olsavsky, K. M., J. L. Page, M. C. Johnson, H. Zarbl, S. C. Strom and C. J. Omiecinski (2007). "Gene expression profiling and differentiation assessment in primary human hepatocyte cultures, established hepatoma cell lines, and human liver tissues." Toxicol Appl Pharmacol **222**(1): 42-56.

Ovcharenko, D., R. Jarvis, S. Hunicke-Smith, K. Kelnar and D. Brown (2005). "High-throughput RNAi screening in vitro: From cell lines to primary cells." RNA **11**(6): 985-993.

Paddison, P. J., J. M. Silva, D. S. Conklin, M. Schlabach, M. Li, S. Aruleba, V. Balija, A. O'Shaughnessy, L. Gnoj, K. Scobie, K. Chang, T. Westbrook, M. Cleary, R. Sachidanandam, W. R. McCombie, S. J. Elledge and G. J. Hannon (2004). "A resource for large-scale RNA-interference-based screens in mammals." Nature **428**.

Pan, C., C. Kumar, S. Bohl, U. Klingmueller and M. Mann (2009). "Comparative Proteomic Phenotyping of Cell Lines and Primary Cells to Assess Preservation of Cell Type-specific Functions." Mol Cell Proteomics **8**(3): 443-450.
Pan, F., L. Sun, D. B. Kardian, K. A. Whartenby, D. M. Pardoll and J. O. Liu (2007). "Feedback inhibition of calcineurin and Ras by a dual inhibitory protein Carabin." Nature **445**(7126): 433-436.
Pankow, J. S., B. B. Duncan, M. I. Schmidt, C. M. Ballantyne, D. J. Couper, R. C. Hoogeveen and S. H. Golden (2004). "Fasting Plasma Free Fatty Acids and Risk of Type 2 Diabetes." The Atherosclerosis Risk in Communities study **27**(1): 77-82.
Paolisso, G., P. A. Tataranni, J. E. Foley, C. Bogardus, B. V. Howard and E. Ravussin (1995). "A high concentration of fasting plasma non-esterified fatty acids is a risk factor for the development of NIDDM." Diabetologia **38**(10): 1213-1217.
Pardo, F. N., J. Altirriba, M. Pradas-Juni, A. Garcia, U. Ahlgren, A. Barbera, J. C. Slebe, A. J. Yanez, R. Gomis and R. Gasa (2012). "The role of Raf-1 kinase inhibitor protein in the regulation of pancreatic beta cell proliferation in mice." Diabetologia **55**(12): 3331-3340.
Park, S., X. Dong, T. L. Fisher, S. Dunn, A. K. Omer, G. Weir and M. F. White (2006). "Exendin-4 Uses Irs2 Signaling to Mediate Pancreatic β Cell Growth and Function." Journal of Biological Chemistry **281**(2): 1159-1168.
Pathan, M., S. Keerthikumar, C.-S. Ang, L. Gangoda, C. Y. J. Quek, N. A. Williamson, D. Mouradov, O. M. Sieber, R. J. Simpson, A. Salim, A. Bacic, A. F. Hill, D. A. Stroud, M. T. Ryan, J. I. Agbinya, J. M. Mariadason, A. W. Burgess and S. Mathivanan (2015). "FunRich: An open access standalone functional enrichment and interaction network analysis tool." PROTEOMICS **15**(15): 2597-2601.
Pattanayak, V., S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna and D. R. Liu (2013). "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity." Nat Biotechnol **31**(9): 839-843.
Paulsen, R. D., D. V. Soni, R. Wollman, A. T. Hahn, M. C. Yee, A. Guan, J. A. Hesley, S. C. Miller, E. F. Cromwell, D. E. Solow-Cordero, T. Meyer and K. A. Cimprich (2009). "A genome-wide siRNA screen reveals diverse cellular

processes and pathways that mediate genome stability." Mol Cell **35**(2): 228-239.
Peto, R., F. J. Roe, P. N. Lee, L. Levy and J. Clack (1975). "Cancer and ageing in mice and men." Br J Cancer **32**(4): 411-426.
Pettersson, M. and B. Ahren (1990). "Failure of islet amyloid polypeptide to inhibit basal and glucose-stimulated insulin secretion in model experiments in mice and rats." Acta Physiol Scand **138**(3): 389-394.
Pevsner, J. (2009). Bioinformatics and functional genomics. Hoboken, N.J., Wiley-Blackwell,**:** xxviii, 951 p., 958 p. of plates.
Pick, A., J. Clark, C. Kubstrup, M. Levisetti, W. Pugh, S. Bonner-Weir and K. S. Polonsky (1998). "Role of apoptosis in failure of beta-cell mass compensation for insulin resistance and beta-cell defects in the male Zucker diabetic fatty rat." Diabetes **47**(3): 358-364.
Pienta, K. J., N. McGregor, R. Axelrod and D. E. Axelrod (2008). "Ecological therapy for cancer: defining tumors using an ecosystem paradigm suggests new opportunities for novel cancer treatments." Transl Oncol **1**(4): 158-164.
Piro, S., M. Anello, C. Di Pietro, M. N. Lizzio, G. Patane, A. M. Rabuazzo, R. Vigneri, M. Purrello and F. Purrello (2002). "Chronic exposure to free fatty acids or high glucose induces apoptosis in rat pancreatic islets: possible role of oxidative stress." Metabolism **51**(10): 1340-1347.
Poungvarin, N., J. K. Lee, V. K. Yechoor, M. V. Li, T. Assavapokee, P. Suksaranjit, J. J. Thepsongwajja, P. K. Saha, K. Oka and L. Chan (2012). "Carbohydrate response element-binding protein (ChREBP) plays a pivotal role in beta cell glucotoxicity." Diabetologia **55**(6): 1783-1796.
Pourcel, C., G. Salvignol and G. Vergnaud (2005). "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies." Microbiology **151**(Pt 3): 653-663.
Prentki, M., E. Joly, W. El-Assaad and R. Roduit (2002). "Malonyl-CoA signaling, lipid partitioning, and glucolipotoxicity: role in beta-cell adaptation and failure in the etiology of diabetes." Diabetes **51 Suppl 3**: S405-413.
Pritsker, M., N. R. Ford, H. T. Jenq and I. R. Lemischka (2006). "Genomewide gain-of-function genetic screen identifies functionally active genes in mouse

embryonic stem cells." Proceedings of the National Academy of Sciences **103**(18): 6946-6951.
Pulvino, M., Y. Liang, D. Oleksyn, M. DeRan, E. Van Pelt, J. Shapiro, I. Sanz, L. Chen and J. Zhao (2012). "Inhibition of proliferation and survival of diffuse large B-cell lymphoma cells by a small-molecule inhibitor of the ubiquitin-conjugating enzyme Ubc13-Uev1A." Blood **120**(8): 1668-1677.
Rachdi, L., N. Balcazar, F. Osorio-Duque, L. Elghazi, A. Weiss, A. Gould, K. J. Chang-Chen, M. J. Gambello and E. Bernal-Mizrachi (2008). "Disruption of Tsc2 in pancreatic β cells induces β cell mass expansion and improved glucose tolerance in a TORC1-dependent manner." Proceedings of the National Academy of Sciences **105**(27): 9250-9255.
Rahier, J., Y. Guiot, R. M. Goebbels, C. Sempoux and J. C. Henquin (2008). "Pancreatic beta-cell mass in European subjects with type 2 diabetes." Diabetes Obes Metab **10 Suppl 4**: 32-42.
Reaven, G. M., C. Hollenbeck, C. Y. Jeng, M. S. Wu and Y. D. Chen (1988). "Measurement of plasma glucose, free fatty acid, lactate, and insulin for 24 h in patients with NIDDM." Diabetes **37**(8): 1020-1024.
Reers, C., S. Erbel, I. Esposito, B. Schmied, M. W. Buchler, P. P. Nawroth and R. A. Ritzel (2009). "Impaired islet turnover in human donor pancreata with aging." Eur J Endocrinol **160**(2): 185-191.
Reidelberger, R. D., U. Arnelo, L. Granqvist and J. Permert (2001). "Comparative effects of amylin and cholecystokinin on food intake and gastric emptying in rats." American Journal of Physiology - Regulatory, Integrative and Comparative Physiology **280**(3): R605-R611.
Ress, A. and K. Moelling (2005). "Bcr is a negative regulator of the Wnt signalling pathway." EMBO reports **6**(11): 1095-1100.
Riordan, J., J. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J. Chou and e. al. (1989). "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA." Science **245**(4922): 1066-1073.
Roberts, R. B., L. Min, M. K. Washington, S. J. Olsen, S. H. Settle, R. J. Coffey and D. W. Threadgill (2002). "Importance of epidermal growth factor receptor signaling in establishment of adenomas and maintenance of carcinomas during intestinal tumorigenesis." Proc Natl Acad Sci U S A **99**(3): 1521-1526.

Roeser, H. P., G. R. Lee, S. Nacht and G. E. Cartwright (1970). "The role of ceruloplasmin in iron metabolism." J Clin Invest **49**(12): 2408-2417.
Rommens, J., M. Iannuzzi, B. Kerem, M. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. Cole, D. Kennedy, N. Hidaka and e. al. (1989). "Identification of the cystic fibrosis gene: chromosome walking and jumping." Science **245**(4922): 1059-1065.
Root, D. E., N. Hacohen, W. C. Hahn, E. S. Lander and D. M. Sabatini (2006). "Genome-scale loss-of-function screening with a lentiviral RNAi library." Nat Meth **3**(9): 715-719.
Rose, R. B., K. E. Pullen, J. H. Bayle, G. R. Crabtree and T. Alber (2004). "Biochemical and structural basis for partially redundant enzymatic and transcriptional functions of DCoH and DCoH2." Biochemistry **43**(23): 7345-7355.
Rosen, L. E., H. A. Morrison, S. Masri, M. J. Brown, B. Springstubb, D. Sussman, B. L. Stoddard and L. M. Seligman (2006). "Homing endonuclease I-CreI derivatives with novel DNA target specificities." Nucleic Acids Res **34**(17): 4791-4800.
Rouet, P., F. Smih and M. Jasin (1994). "Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells." Proceedings of the National Academy of Sciences **91**(13): 6064-6068.
Rubin, H. (1960). "An analysis of the assay of Rous sarcoma cells in vitro by the infective center technique." Virology **10**: 29-49.
Rubin, H. (1971). "pH and population density in the regulation of animal cell multiplication." J Cell Biol **51**(3): 686-702.
Rulifson, I. C., S. K. Karnik, P. W. Heiser, D. ten Berge, H. Chen, X. Gu, M. M. Taketo, R. Nusse, M. Hebrok and S. K. Kim (2007). "Wnt signaling regulates pancreatic β cell proliferation." Proceedings of the National Academy of Sciences **104**(15): 6247-6252.
Russell, L. B., C. S. Montgomery and G. D. Raymer (1982). "Analysis of the Albino-Locus Region of the Mouse: IV. Characterization of 34 Deficiencies." Genetics **100**(3): 427-453.
Saci, A., L. C. Cantley and C. L. Carpenter (2011). "Rac1 Regulates the Activity of mTORC1 and mTORC2 and Controls Cellular Size." Mol Cell **42**(1): 50-61.

Sack, L. M., T. Davoli, Q. Xu, M. Li and S. J. Elledge (2016). "Sources of Error in Mammalian Genetic Screens." G3 (Bethesda).
Sambrook, J. and D. W. Russell (2006). "Purification of Nucleic Acids by Extraction with Phenol:Chloroform." Cold Spring Harbor Protocols **2006**(1): pdb.prot4455.
Sampath, P., B. Mazumder, V. Seshadri and P. L. Fox (2003). "Transcript-Selective Translational Silencing by Gamma Interferon Is Directed by a Novel Structural Element in the Ceruloplasmin mRNA 3′ Untranslated Region." Mol Cell Biol **23**(5): 1509-1519.
Sandberg, R. and I. Ernberg (2005). "The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles." Genome Biol **6**(8): R65.
Sandler, S. and M. Stridsberg (1994). "Chronic exposure of cultured rat pancreatic islets to elevated concentrations of islet amyloid polypeptide (IAPP) causes a decrease in islet DNA content and medium insulin accumulation." Regul Pept **53**(2): 103-109.
Santiago, Y., E. Chan, P.-Q. Liu, S. Orlando, L. Zhang, F. D. Urnov, M. C. Holmes, D. Guschin, A. Waite, J. C. Miller, E. J. Rebar, P. D. Gregory, A. Klug and T. N. Collingwood (2008). "Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases." Proceedings of the National Academy of Sciences **105**(15): 5809-5814.
Saxena, R., L. Gianniny, N. P. Burtt, V. Lyssenko, C. Giuducci, M. Sjögren, J. C. Florez, P. Almgren, B. Isomaa, M. Orho-Melander, U. Lindblad, M. J. Daly, T. Tuomi, J. N. Hirschhorn, K. G. Ardlie, L. C. Groop and D. Altshuler (2006). "Common Single Nucleotide Polymorphisms in TCF7L2 Are Reproducibly Associated With Type 2 Diabetes and Reduce the Insulin Response to Glucose in Nondiabetic Individuals." Diabetes **55**(10): 2890-2895.
Scacheri, P. C., O. Rozenblatt-Rosen, N. J. Caplen, T. G. Wolfsberg, L. Umayam, J. C. Lee, C. M. Hughes, K. S. Shanmugam, A. Bhattacharjee, M. Meyerson and F. S. Collins (2004). "Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells." Proc Natl Acad Sci U S A **101**(7): 1892-1897.
Scott, L. J., L. L. Bonnycastle, C. J. Willer, A. G. Sprau, A. U. Jackson, N. Narisu, W. L. Duren, P. S. Chines, H. M. Stringham, M. R. Erdos, T. T. Valle, J. Tuomilehto, R. N. Bergman, K. L. Mohlke, F. S. Collins and M. Boehnke (2006).

"Association of Transcription Factor 7-Like 2 (TCF7L2) Variants With Type 2 Diabetes in a Finnish Sample." Diabetes **55**(9): 2649-2653.
Scharfmann, R., S. Pechberty, Y. Hazhouz, M. von Bulow, E. Bricout-Neveu, M. Grenier-Godard, F. Guez, L. Rachdi, M. Lohmann, P. Czernichow and P. Ravassard (2014). "Development of a conditionally immortalized human pancreatic β cell line." The Journal of Clinical Investigation **124**(5): 2087-2098.
Schisler, J. C., P. T. Fueger, D. A. Babu, H. E. Hohmeier, J. S. Tessem, D. Lu, T. C. Becker, B. Naziruddin, M. Levy, R. G. Mirmira and C. B. Newgard (2008). "Stimulation of human and rat islet beta-cell proliferation with retention of function by the homeodomain transcription factor Nkx6.1." Mol Cell Biol **28**(10): 3465-3476.
Schmid-Burgk, J. L., D. Chauhan, T. Schmidt, T. S. Ebert, J. Reinhardt, E. Endl and V. Hornung (2016). "A Genome-wide CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) Screen Identifies NEK7 as an Essential Component of NLRP3 Inflammasome Activation." J Biol Chem **291**(1): 103-109.
Schmitt, T., C. Ogris and E. L. Sonnhammer (2014). "FunCoup 3.0: database of genome-wide functional coupling networks." Nucleic Acids Res **42**(Database issue): D380-388.
Schürmann, N., L. G. Trabuco, C. Bender, R. B. Russell and D. Grimm (2013). "Molecular dissection of human Argonaute proteins by DNA shuffling." Nat Struct Mol Biol **20**(7): 818-826.
Schwarz, D. S., G. Hutvagner, T. Du, Z. Xu, N. Aronin and P. D. Zamore (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell **115**(2): 199-208.
Selbach, M., B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin and N. Rajewsky (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.
Seshadri, R., R. J. Kutlaca, K. Trainor, C. Matthews and A. A. Morley (1987). "Mutation rate of normal and malignant human lymphocytes." Cancer Res **47**(2): 407-409.
Shalem, O., N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench and F. Zhang (2014). "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells." Science **343**(6166): 84-87.

Shedden, K., J. M. Taylor, S. A. Enkemann, M. S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, A. C. Chang, C. Q. Zhu, D. Strumpf, S. Hanash, F. A. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. E. Seshan, M. Meyerson, R. Kuick, K. K. Dobbin, T. Lively, J. W. Jacobson and D. G. Beer (2008). "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study." Nat Med **14**(8): 822-827.

Shen, B., J. Zhang, H. Wu, J. Wang, K. Ma, Z. Li, X. Zhang, P. Zhang and X. Huang (2013). "Generation of gene-modified mice via Cas9/RNA-mediated gene targeting." Cell Res **23**(5): 720-723.

Shen, C., A. K. Buck, X. Liu, M. Winkler and S. N. Reske (2003). "Gene silencing by adenovirus-delivered siRNA." FEBS Letters **539**(1-3): 111-114.

Shen, M., Y. Cai, Y. Yang, X. Yan, X. Liu and T. Zhou (2013). "Centrosomal protein FOR20 is essential for S-phase progression by recruiting Plk1 to centrosomes." Cell Res **23**(11): 1284-1295.

Shepherd, P. R., B. T. Navé and K. Siddle (1995). "Insulin stimulation of glycogen synthesis and glycogen synthase activity is blocked by wortmannin and rapamycin in 3T3-L1 adipocytes: evidence for the involvement of phosphoinositide 3-kinase and p70 ribosomal protein-S6 kinase." Biochem J **305**(Pt 1): 25-28.

Shimabukuro, M., Y. T. Zhou, M. Levi and R. H. Unger (1998). "Fatty acid-induced beta cell apoptosis: a link between obesity and diabetes." Proc Natl Acad Sci U S A **95**(5): 2498-2502.

Shukla, N., R. Somwar, R. S. Smith, S. Ambati, S. Munoz, M. Merchant, P. D'Arcy, X. Wang, R. Kobos, C. Antczak, B. Bhinder, D. Shum, C. Radu, G. Yang, B. S. Taylor, C. K. Ng, B. Weigelt, I. Khodos, E. de Stanchina, J. S. Reis-Filho, O. Ouerfelli, S. Linder, H. Djaballah and M. Ladanyi (2016). "Proteasome Addiction Defined in Ewing Sarcoma Is Effectively Targeted by a Novel Class of 19S Proteasome Inhibitors." Cancer Res.

Silva, J. M., M. Z. Li, K. Chang, W. Ge, M. C. Golding, R. J. Rickles, D. Siolas, G. Hu, P. J. Paddison, M. R. Schlabach, N. Sheth, J. Bradshaw, J. Burchard, A. Kulkarni, G. Cavet, R. Sachidanandam, W. R. McCombie, M. A. Cleary, S. J.

Elledge and G. J. Hannon (2005). "Second-generation shRNA libraries covering the mouse and human genomes." Nat Genet **37**(11): 1281-1288.
Silva, J. M., K. Marran, J. S. Parker, J. Silva, M. Golding, M. R. Schlabach, S. J. Elledge, G. J. Hannon and K. Chang (2008). "Profiling essential genes in human mammary cells by multiplex RNAi screening." Science **319**(5863): 617-620.
Simard, J., J. Feunteun, G. Lenoir, P. Tonin, T. Normand, V. Luu The, A. Vivier, D. Lasko, K. Morgan, G. A. Rouleau and et al. (1993). "Genetic mapping of the breast-ovarian cancer syndrome to a small interval on chromosome 17q12-21: exclusion of candidate genes EDH17B2 and RARA." Hum Mol Genet **2**(8): 1193-1199.
Simpson, C. D., R. Hurren, D. Kasimer, N. MacLean, Y. Eberhard, T. Ketela, J. Moffat and A. D. Schimmer (2012). "A genome wide shRNA screen identifies α/β hydrolase domain containing 4 (ABHD4) as a novel regulator of anoikis resistance." Apoptosis **17**(7): 666-678.
Sims, D., A. M. Mendes-Pereira and J. Frankum… (2011). "High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing." Genome ….
Smith, J. A., E. A. White, M. E. Sowa, M. L. Powell, M. Ottinger, J. W. Harper and P. M. Howley (2010). "Genome-wide siRNA screen identifies SMCX, EP400, and Brd4 as E2-dependent regulators of human papillomavirus oncogene expression." Proc Natl Acad Sci U S A **107**(8): 3752-3757.
Smithies, O., R. G. Gregg, S. S. Boggs, M. A. Koralewski and R. S. Kucherlapati (1985). "Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination." Nature **317**(6034): 230-234.
Song, W.-J., W. E. Schreiber, E. Zhong, F.-F. Liu, B. D. Kornfeld, F. E. Wondisford and M. A. Hussain (2008). "Exendin-4 Stimulation of Cyclin A2 in β-Cell Proliferation." Diabetes **57**(9): 2371-2381.
Sorenson, R. L. and T. C. Brelje (1997). "Adaptation of islets of Langerhans to pregnancy: beta-cell growth, enhanced insulin secretion and the role of lactogenic hormones." Horm Metab Res **29**(6): 301-307.
Stambolic, V., A. Suzuki, J. L. De la Pompa, G. M. Brothers, C. Mirtsos, T. Sasaki, J. Ruland, J. M. Penninger, D. P. Siderovski and T. W. Mak (1998).

"Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN." Cell **95**(1): 29-39.
Stern, P., S. Astrof, S. J. Erkeland, J. Schustak, P. A. Sharp and R. O. Hynes (2008). "A system for Cre-regulated RNA interference in vivo." Proceedings of the National Academy of Sciences **105**(37): 13895-13900.
Stewart, S. A., D. M. Dykxhoorn, D. Palliser, H. Mizuno, E. Y. Yu, D. S. An, D. M. Sabatini, I. S. Chen, W. C. Hahn, P. A. Sharp, R. A. Weinberg and C. D. Novina (2003). "Lentivirus-delivered stable gene silencing by RNAi in primary cells." RNA **9**(4): 493-501.
Stiles, B. L., C. Kuralwalla-Martinez, W. Guo, C. Gregorian, Y. Wang, J. Tian, M. A. Magnuson and H. Wu (2006). "Selective Deletion of Pten in Pancreatic β Cells Leads to Increased Islet Mass and Resistance to STZ-Induced Diabetes." Molecular and Cellular Biology **26**(7): 2772-2781.
Stoker, M. (1964). "REGULATION OF GROWTH AND ORIENTATION IN HAMSTER CELLS TRANSFORMED BY POLYOMA VIRUS." Virology **24**: 165-174.
Stoker, M. G. (1967). "Transfer of growth inhibition between normal and virus-transformed cells: autoradiographic studies using marked cells." J Cell Sci **2**(3): 293-304.
Stoker, M. G., M. Shearer and C. O'Neill (1966). "Growth inhibition of polyoma-transformed cells by contact with static normal fibroblasts." J Cell Sci **1**(3): 297-310.
Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proceedings of the National Academy of Sciences **100**(16): 9440-9445.
Strausberg, R. L., E. A. Feingold, L. H. Grouse, J. G. Derge, R. D. Klausner, F. S. Collins, L. Wagner, C. M. Shenmen, G. D. Schuler, S. F. Altschul, B. Zeeberg, K. H. Buetow, C. F. Schaefer, N. K. Bhat, R. F. Hopkins, H. Jordan, T. Moore, S. I. Max, J. Wang, F. Hsieh, L. Diatchenko, K. Marusina, A. A. Farmer, G. M. Rubin, L. Hong, M. Stapleton, M. B. Soares, M. F. Bonaldo, T. L. Casavant, T. E. Scheetz, M. J. Brownstein, T. B. Usdin, S. Toshiyuki, P. Carninci, C. Prange, S. S. Raha, N. A. Loquellano, G. J. Peters, R. D. Abramson, S. J. Mullahy, S. A. Bosak, P. J. McEwan, K. J. McKernan, J. A. Malek, P. H. Gunaratne, S. Richards, K. C. Worley, S. Hale, A. M. Garcia, L. J.

Gay, S. W. Hulyk, D. K. Villalon, D. M. Muzny, E. J. Sodergren, X. Lu, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madan, A. C. Young, Y. Shevchenko, G. G. Bouffard, R. W. Blakesley, J. W. Touchman, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. I. Krzywinski, U. Skalska, D. E. Smailus, A. Schnerch, J. E. Schein, S. J. Jones and M. A. Marra (2002). "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences." Proc Natl Acad Sci U S A **99**(26): 16899-16903.

Strezoska, Ž., A. Licon, J. Haimes, K. J. Spayd, K. M. Patel, K. Sullivan, K. Jastrzebski, K. J. Simpson, D. Leake, A. van Brabant Smith and A. Vermeulen (2012). "Optimized PCR Conditions and Increased shRNA Fold Representation Improve Reproducibility of Pooled shRNA Screens." PLoS ONE **7**(8): e42341.

Strickland, J. E., M. Ueda, H. Hennings and S. H. Yuspa (1992). "A model for initiated mouse skin: suppression of papilloma but not carcinoma formation by normal epidermal cells in grafts on athymic nude mice." Cancer Res **52**(6): 1439-1444.

Sun, X. J., P. Rothenberg, C. R. Kahn, J. M. Backer, E. Araki, P. A. Wilden, D. A. Cahill, B. J. Goldstein and M. F. White (1991). "Structure of the insulin receptor substrate IRS-1 defines a unique signal transduction protein." Nature **352**(6330): 73-77.

Sun, X. J., L. M. Wang, Y. Zhang, L. Yenush, M. G. Myers, Jr., E. Glasheen, W. S. Lane, J. H. Pierce and M. F. White (1995). "Role of IRS-2 in insulin and cytokine signalling." Nature **377**(6545): 173-177.

Swift, S., J. Lorens, P. Achacoso and G. P. Nolan (2001). "Rapid production of retroviruses for efficient gene delivery to mammalian cells using 293T cell-based systems." Curr Protoc Immunol **Chapter 10**: Unit 10.17C.

Tang, R. H., X. L. Zheng, T. E. Callis, W. E. Stansfield, J. He, A. S. Baldwin, D. Z. Wang and C. H. Selzman (2008). "Myocardin inhibits cellular proliferation by inhibiting NF-kappaB(p65)-dependent cell cycle progression." Proc Natl Acad Sci U S A **105**(9): 3362-3367.

Taylor, B. S., N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. A. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers

and W. L. Gerald (2010). "Integrative genomic profiling of human prostate cancer." Cancer Cell **18**(1): 11-22.

Temple, G., D. S. Gerhard, R. Rasooly, E. A. Feingold, P. J. Good, C. Robinson, A. Mandich, J. G. Derge, J. Lewis, D. Shoaf, F. S. Collins, W. Jang, L. Wagner, C. M. Shenmen, L. Misquitta, C. F. Schaefer, K. H. Buetow, T. I. Bonner, L. Yankie, M. Ward, L. Phan, A. Astashyn, G. Brown, C. Farrell, J. Hart, M. Landrum, B. L. Maidak, M. Murphy, T. Murphy, B. Rajput, L. Riddick, D. Webb, J. Weber, W. Wu, K. D. Pruitt, D. Maglott, A. Siepel, B. Brejova, M. Diekhans, R. Harte, R. Baertsch, J. Kent, D. Haussler, M. Brent, L. Langton, C. L. Comstock, M. Stevens, C. Wei, M. J. van Baren, K. Salehi-Ashtiani, R. R. Murray, L. Ghamsari, E. Mello, C. Lin, C. Pennacchio, K. Schreiber, N. Shapiro, A. Marsh, E. Pardes, T. Moore, A. Lebeau, M. Muratet, B. Simmons, D. Kloske, S. Sieja, J. Hudson, P. Sethupathy, M. Brownstein, N. Bhat, J. Lazar, H. Jacob, C. E. Gruber, M. R. Smith, J. McPherson, A. M. Garcia, P. H. Gunaratne, J. Wu, D. Muzny, R. A. Gibbs, A. C. Young, G. G. Bouffard, R. W. Blakesley, J. Mullikin, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, M. Hirst, T. Zeng, K. Tse, M. Moksa, M. Deng, K. Ma, D. Mah, J. Pang, G. Taylor, E. Chuah, A. Deng, K. Fichter, A. Go, S. Lee, J. Wang, M. Griffith, R. Morin, R. A. Moore, M. Mayo, S. Munro, S. Wagner, S. J. Jones, R. A. Holt, M. A. Marra, S. Lu, S. Yang, J. Hartigan, M. Graf, R. Wagner, S. Letovksy, J. C. Pulido, K. Robison, D. Esposito, J. Hartley, V. E. Wall, R. F. Hopkins, O. Ohara and S. Wiemann (2009). "The completion of the Mammalian Gene Collection (MGC)." Genome Res **19**(12): 2324-2333.

Terauchi, Y., I. Takamoto, N. Kubota, J. Matsui, R. Suzuki, K. Komeda, A. Hara, Y. Toyoda, I. Miwa, S. Aizawa, S. Tsutsumi, Y. Tsubamoto, S. Hashimoto, K. Eto, A. Nakamura, M. Noda, K. Tobe, H. Aburatani, R. Nagai and T. Kadowaki "Glucokinase and IRS-2 are required for compensatory β cell hyperplasia in response to high-fat diet–induced insulin resistance." The Journal of Clinical Investigation **117**(1): 246-257.

Terzaghi-Howe, M. (1987). "Inhibition of carcinogen-altered rat tracheal epithelial cell proliferation by normal epithelial cells in vivo." Carcinogenesis **8**(1): 145-150.

Thierry, A. and B. Dujon (1992). "Nested chromosomal fragmentation in yeast using the meganuclease I-Sce I: a new method for physical mapping of eukaryotic genomes." Nucleic Acids Research **20**(21): 5625-5631.
Todaro, G. J. and H. Green (1963). "Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines." J Cell Biol **17**: 299-313.
Todaro, G. J. and H. Green (1966). "Cell growth and the initiation of transformation by SV40." Proc Natl Acad Sci U S A **55**(2): 302-308.
Tsui, L. C., M. Buchwald, D. Barker, J. C. Braman, R. Knowlton, J. W. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic and et al. (1985). "Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker." Science **230**(4729): 1054-1057.
Van Aelst, L. and C. D'Souza-Schorey (1997). "Rho GTPases and signaling networks." Genes & Development **11**(18): 2295-2322.
Van Assche, F. A., L. Aerts and F. De Prins (1978). "A morphological study of the endocrine pancreas in human pregnancy." Br J Obstet Gynaecol **85**(11): 818-820.
van de Weijer, M. L., M. C. Bassik, R. D. Luteijn, C. M. Voorburg, M. A. M. Lohuis, E. Kremmer, R. C. Hoeben, E. M. LeProust, S. Chen, H. Hoelen, M. E. Ressing, W. Patena, J. S. Weissman, M. T. McManus, E. J. H. J. Wiertz and R. J. Lebbink (2014). "A high-coverage shRNA screen identifies TMEM129 as an E3 ligase involved in ER-associated protein degradation." Nature Communications **5**: 3832.
Varney, M. L. and R. K. Singh (2015). "VEGF-C-VEGFR3/Flt4 axis regulates mammary tumor growth and metastasis in an autocrine manner." Am J Cancer Res **5**(2): 616-628.
Vartiainen, M. K., E. M. Sarkkinen, T. Matilainen, M. Salminen and P. Lappalainen (2003). "Mammals have two twinfilin isoforms whose subcellular localizations and tissue distributions are differentially regulated." J Biol Chem **278**(36): 34347-34355.
Vasavada, R. C., L. Wang, Y. Fujinaka, K. K. Takane, T. C. Rosa, J. M. D. Mellado-Gil, P. A. Friedman and A. Garcia-Ocaña (2007). "Protein Kinase C-ζ Activation Markedly Enhances β-Cell Proliferation." An Essential Role in Growth Factor–Mediated β-Cell Mitogenesis **56**(11): 2732-2743.

Velazquez-Garcia, S., S. Valle, T. C. Rosa, K. K. Takane, C. Demirci, J. C. Alvarez-Perez, J. M. Mellado-Gil, S. Ernst, D. K. Scott, R. C. Vasavada, L. C. Alonso and A. Garcia-Ocaña (2011). "Activation of Protein Kinase C-ζ in Pancreatic β-Cells In Vivo Improves Glucose Tolerance and Induces β-Cell Expansion via mTOR Activation." Diabetes **60**(10): 2546-2559.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V.

Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The Sequence of the Human Genome." Science **291**(5507): 1304-1351.

Wang, C. H., P. T. Su, X. Y. Du, M. W. Kuo, C. Y. Lin, C. C. Yang, H. S. Chan, S. J. Chang, C. Kuo, K. Seo, L. L. Leung and Y. J. Chuang (2010). "Thrombospondin type I domain containing 7A (THSD7A) mediates endothelial cell migration and tube formation." J Cell Physiol **222**(3): 685-694.

Wang, T., J. J. Wei, D. M. Sabatini and E. S. Lander (2014). "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." Science **343**(6166): 80-84.

Wang, W.-H., L. L. Studach and O. M. Andrisani (2011). "Proteins ZNF198 and SUZ12 are down-regulated in hepatitis B virus (HBV) X protein-mediated hepatocyte transformation and in HBV replication." Hepatology **53**(4): 1137-1147.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D.

Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody and E. S. Lander (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

Wei, G., A. E. Schaffner, K. M. Baker, K. C. Mansky and M. C. Ostrowski (2003). "Ets-2 interacts with co-repressor BS69 to repress target gene expression." Anticancer Res **23**(3A): 2173-2178.

Weir, G. C. and S. Bonner-Weir (2004). "Five Stages of Evolving Beta-Cell Dysfunction During Progression to Diabetes." Diabetes **53**(suppl 3): S16-S21.

Weiss, W. A., S. S. Taylor and K. M. Shokat (2007). "Recognizing and exploiting differences between RNAi and small-molecule inhibitors." Nat Chem Biol **3**(12): 739-744.

Welters, H. J. and R. N. Kulkarni (2008). "Wnt signaling: relevance to β-cell biology and diabetes." Trends in Endocrinology & Metabolism **19**(10): 349-355.
Welters, H. J., A. Oknianska, K. S. Erdmann, G. U. Ryffel and N. G. Morgan (2008). "The protein tyrosine phosphatase-BL, modulates pancreatic β-cell proliferation by interaction with the Wnt signalling pathway." Journal of Endocrinology **197**(3): 543-552.
Wen, H., Y. Li, Y. Xi, S. Jiang, S. Stratton, D. Peng, K. Tanaka, Y. Ren, Z. Xia, J. Wu, B. Li, M. C. Barton, W. Li, H. Li and X. Shi (2014). "ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression." Nature **508**(7495): 263-268.
Wermke, M., A. Camgoz, M. Paszkowski-Rogacz, S. Thieme, M. von Bonin, A. Dahl, U. Platzbecker, M. Theis, G. Ehninger, S. Brenner, M. Bornhäuser and F. Buchholz (2015). "RNAi profiling of primary human AML cells identifies ROCK1 as a therapeutic target and nominates fasudil as an antileukemic drug." Blood **125**(24): 3760-3768.
Westbrook, T. F., E. S. Martin, M. R. Schlabach, Y. Leng, A. C. Liang, B. Feng, J. J. Zhao, T. M. Roberts, G. Mandel, G. J. Hannon, R. A. DePinho, L. Chin and S. J. Elledge (2005). "A Genetic Screen for Candidate Tumor Suppressors Identifies REST." Cell **121**(6): 837-848.
Westermark, P. and E. Wilander (1978). "The influence of amyloid deposits on the islet volume in maturity onset diabetes mellitus." Diabetologia **15**(5): 417-421.
Widschwendter, M., J. Berger, G. Daxenbichler, E. Muller-Holzner, A. Widschwendter, A. Mayr, C. Marth and A. G. Zeimet (1997). "Loss of retinoic acid receptor beta expression in breast cancer and morphologically normal adjacent tissue but not in the normal breast tissue distant from the cancer." Cancer Res **57**(19): 4158-4161.
Wilson, J. L., S. Dalin, S. Gosline, M. Hemann, E. Fraenkel and D. A. Lauffenburger (2016). "Pathway-based network modeling finds hidden genes in shRNA screen for regulators of acute lymphoblastic leukemia." Integrative Biology.
Withers, D. J., J. S. Gutierrez, H. Towery, D. J. Burks, J.-M. Ren, S. Previs, Y. Zhang, D. Bernal, S. Pons, G. I. Shulman, S. Bonner-Weir and M. F. White

(1998). "Disruption of IRS-2 causes type 2 diabetes in mice." Nature **391**(6670): 900-904.
Wong, C. C., I. Martincorena, A. G. Rust, M. Rashid, C. Alifrangis, L. B. Alexandrov, J. C. Tiffen, C. Kober, C. Chronic Myeloid Disorders Working Group of the International Cancer Genome, A. R. Green, C. E. Massie, J. Nangalia, S. Lempidaki, H. Döhner, K. Döhner, S. J. Bray, U. McDermott, E. Papaemmanuil, P. J. Campbell and D. J. Adams (2014). "Inactivating CUX1 mutations promote tumorigenesis." Nature genetics **46**(1): 10.1038/ng.2846.
Wu, Z., S. Shen, Z. Zhang, W. Zhang and W. Xiao (2014). "Ubiquitin-conjugating enzyme complex Uev1A-Ubc13 promotes breast cancer metastasis through nuclear factor-κB mediated matrix metalloproteinase-1 gene regulation." Breast Cancer Research **16**(4): 1-15.
Xiao, X., Z. Liu, R. Wang, J. Wang, S. Zhang, X. Cai, K. Wu, R. C. Bergan, L. Xu and D. Fan (2015). "Genistein suppresses FLT4 and inhibits human colorectal cancer metastasis." Oncotarget **6**(5): 3225-3239.
Xu, Q., G.-Y. Yang, N. Liu, P. Xu, Y.-L. Chen, Z. Zhou, Z.-G. Luo and X. Ding (2012). "P4-ATPase ATP8A2 acts in synergy with CDC50A to enhance neurite outgrowth." FEBS Letters **586**(13): 1803-1812.
Xu, S., S. M. Nam, J. H. Kim, R. Das, S. K. Choi, T. T. Nguyen, X. Quan, S. J. Choi, C. H. Chung, E. Y. Lee, I. K. Lee, A. Wiederkehr, C. B. Wollheim, S. K. Cha and K. S. Park (2015). "Palmitate induces ER calcium depletion and apoptosis in mouse podocytes subsequent to mitochondrial oxidative stress." Cell Death Dis **6**: e1976.
Yamagata, K., N. Oda, P. J. Kaisaki, S. Menzel, H. Furuta, M. Vaxillaire, L. Southam, R. D. Cox, G. M. Lathrop, V. V. Boriraj, X. Chen, N. J. Cox, Y. Oda, H. Yano, M. M. Le Beau, S. Yamada, H. Nishigori, J. Takeda, S. S. Fajans, A. T. Hattersley, N. Iwasaki, T. Hansen, O. Pedersen, K. S. Polonsky, R. C. Turner, G. Velho, J.-C. Chevre, P. Froguel and G. I. Bell (1996). "Mutations in the hepatocyte nuclear factor-1[alpha] gene in maturity-onset diabetes of the young (MODY3)." Nature **384**(6608): 455-458.
Yang, D., J. Xu, T. Zhu, J. Fan, L. Lai, J. Zhang and Y. E. Chen (2014). "Effective gene targeting in rabbits using RNA-guided Cas9 nucleases." Journal of Molecular Cell Biology **6**(1): 97-99.

Yang, J., J. Fan, Y. Li, F. Li, P. Chen, Y. Fan, X. Xia and S. T. Wong (2013). "Genome-Wide RNAi Screening Identifies Genes Inhibiting the Migration of Glioblastoma Cells." PLoS ONE **8**(4): e61915.

Yang, K. T., J. A. Bayan, N. Zeng, R. Aggarwal, L. He, Z. Peng, A. Kassa, M. Kim, Z. Luo, Z. Shi, V. Medina, K. Boddupally and B. L. Stiles (2014). "Adult-onset deletion of Pten increases islet mass and beta cell proliferation in mice." Diabetologia **57**(2): 352-361.

Yang, Q., K. Yamagata, K. Fukui, Y. Cao, T. Nammo, H. Iwahashi, H. Wang, I. Matsumura, T. Hanafusa, R. Bucala, C. B. Wollheim, J.-i. Miyagawa and Y. Matsuzawa (2002). "Hepatocyte Nuclear Factor-1α Modulates Pancreatic β-Cell Growth by Regulating the Expression of Insulin-Like Growth Factor-1 in INS-1 Cells." Diabetes **51**(6): 1785-1792.

Yang, S., S. Tutton, E. Pierce and K. Yoon (2001). "Specific Double-Stranded RNA Interference in Undifferentiated Mouse Embryonic Stem Cells." Mol Cell Biol **21**(22): 7807-7816.

Yang, X., J. S. Boehm, X. Yang, K. Salehi-Ashtiani, T. Hao, Y. Shen, R. Lubonja, S. R. Thomas, O. Alkan, T. Bhimdi, T. M. Green, C. M. Johannessen, S. Silver, C. Nguyen, R. R. Murray, H. Hieronymus, D. Balcha, C. Fan, C. Lin, L. Ghamsari, M. Vidal, W. C. Hahn, D. E. Hill and D. E. Root (2011). "A public genome-scale lentiviral expression library of human ORFs." Nat Methods **8**(8): 659-661.

Yatsu, A., N. Ohbayashi, K. Tamura and M. Fukuda (2013). "Syntaxin-3 Is Required for Melanosomal Localization of Tyrp1 in Melanocytes." Journal of Investigative Dermatology **133**(9): 2237-2246.

Yi, R., Y. Qin, I. G. Macara and B. R. Cullen (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." Genes Dev **17**(24): 3011-3016.

Yoshino, S., T. Hara, J. S. Weng, Y. Takahashi, M. Seiki and T. Sakamoto (2012). "Genetic Screening of New Genes Responsible for Cellular Adaptation to Hypoxia Using a Genome-Wide shRNA Library." PLoS ONE **7**(4): e35590.

Zaidi, M., H. K. Datta, P. J. Bevis, S. J. Wimalawansa and I. MacIntyre (1990). "Amylin-amide: a new bone-conserving peptide from the pancreas." Exp Physiol **75**(4): 529-536.

Zamore, P. D., T. Tuschl, P. A. Sharp and D. P. Bartel (2000). "RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals." Cell **101**(1): 25-33.

Zaslavskiy, M., C. Bertonati, P. Duchateau, A. Duclert and G. H. Silva (2014) "Efficient design of meganucleases using a machine learning approach." BMC bioinformatics **15**, 191 DOI: 10.1186/1471-2105-15-191.

Zechner, D., T. Muller, H. Wende, I. Walther, M. M. Taketo, E. B. Crenshaw, 3rd, M. Treier, W. Birchmeier and C. Birchmeier (2007). "Bmp and Wnt/beta-catenin signals control expression of the transcription factor Olig3 and the specification of spinal cord neurons." Dev Biol **303**(1): 181-190.

Zhang, C., L. Qi, D. J. Hunter, J. B. Meigs, J. E. Manson, R. M. van Dam and F. B. Hu (2006). "Variant of Transcription Factor 7-Like 2 (TCF7L2) Gene and the Risk of Type 2 Diabetes in Large Cohorts of U.S. Women and Men." Diabetes **55**(9): 2645-2648.

Zheng, M. Z., H. D. Qin, X. J. Yu, R. H. Zhang, L. Z. Chen, Q. S. Feng and Y. X. Zeng (2007). "Haplotype of gene Nedd4 binding protein 2 associated with sporadic nasopharyngeal carcinoma in the Southern Chinese population." J Transl Med **5**: 36.

Zhou, H., M. Xu, Q. Huang, A. T. Gates, X. D. Zhang, J. C. Castle, E. Stec, M. Ferrer, B. Strulovici, D. J. Hazuda and A. S. Espeseth (2008). "Genome-Scale RNAi Screen for Host Factors Required for HIV Replication." Cell Host & Microbe **4**(5): 495-504.

Zhou, N., B. Ding, M. Agler, M. Cockett and F. McPhee (2015). "Lethality of PAK3 and SGK2 shRNAs to Human Papillomavirus Positive Cervical Cancer Cells Is Independent of PAK3 and SGK2 Knockdown." PLoS ONE **10**(1): e0117357.

Zhou, Z., M. Patel, N. Ng, M. H. Hsieh, A. P. Orth, J. R. Walker, S. Batalov, J. L. Harris and J. Liu (2014). "Identification of synthetic lethality of PRKDC in MYC-dependent human cancers by pooled shRNA screening." BMC Cancer **14**: 944.

Zhu, L. H., A. Wang, P. Luo, X. Wang, D. S. Jiang, W. Deng, X. Zhang, T. Wang, Y. Liu, L. Gao, S. Zhang, X. Zhang, J. Zhang and H. Li (2014). "Mindin/Spondin 2 inhibits hepatic steatosis, insulin resistance, and obesity via interaction with peroxisome proliferator-activated receptor alpha in mice." J Hepatol **60**(5): 1046-1054.

Zierath, J. R., D. Galuska, Å. Engström, K. H. Johnson, C. Betsholtz, P. Westermark and H. Wallberg-Henriksson (1992). "Human islet amyloid polypeptide at pharmacological levels inhibits insulin and phorbol ester-stimulated glucose transport in in vitro incubated human muscle strips." Diabetologia **35**(1): 26-31.

Ziouzenkova, O., S. Perrey, L. Asatryan, J. Hwang, K. L. MacNaul, D. E. Moller, D. J. Rader, A. Sevanian, R. Zechner, G. Hoefler and J. Plutzky (2003). "Lipolysis of triglyceride-rich lipoproteins generates PPAR ligands: Evidence for an antiinflammatory role for lipoprotein lipase." Proceedings of the National Academy of Sciences **100**(5): 2730-2735.

Zuber, J., K. McJunkin, C. Fellmann, L. E. Dow, M. J. Taylor, G. J. Hannon and S. W. Lowe (2011). "Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi." Nat Biotechnol **29**(1): 79-83.

Zugasti, O., N. Thakur, J. Belougne, B. Squiban, C. L. Kurz, J. Soule, S. Omi, L. Tichit, N. Pujol and J. J. Ewbank (2016). "A quantitative genome-wide RNAi screen in C. elegans for antifungal innate immunity genes." BMC Biol **14**(1): 35.