MicroRNA Buffering of Gene Duplications and Aneuploidy

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2021

Mark D Reardon

School of Biological Sciences

Contents

Figures 5

Tables 7

Abstract 9

Declaration 10

Copyright statement 10

Acknowledgements 11

Journal format 12

Contributions 12

1 Introduction 13

1.1 MicroRNAs 13

1.1.1 The discovery of miRNAs 13

1.1.2 Genomics and conservation 14

1.1.3 Biogenesis and biological function 16

1.1.4 Target prediction, verification and interactions 19

1.1.5 Expression and selective avoidance 22

1.1.6 Loss of function 24

1.1.7 Regulatory motifs 24

1.2 Genetic variation 29

1.2.1 Evolutionary variation 29

1.2.2 Germline and somatic variation 31

1.3 Dosage compensation 33

1.3.1 Gene dosage balance hypothesis 33

1.3.2 Dosage compensation 33

1.3.3 Canalisation and morphological complexity 35

1.4 Disease 37

1.4.1 Dominant and recessive disease 37

1.4.2 Haploinsufficiency 38

1.5 Summary and aims of project 41

2 High-resolution CNVs reveal potential novel cancer driver genes 43

2.1 Abstract 43

2.2 Introduction 44

2.3 Methods 47

2.3.1 Construction of a high-resolution NCI-60 CNV map 47

2.3.2 Characterisation of CNVs 48

2.3.3 Well-known cancer-related genes ... 48

2.3.4 Calculation of metrics with a range of 'sliding window' sizes ... 49

2.3.5 Determining empirical significance of copy numbers by permutation testing ... 49

2.3.6 Enrichment analyses of cancer-related genes significantly affected by CNVs ... 49

2.3.7 Statistical analyses ... 49

2.4 Results ... 50

2.4.1 A high-resolution NCI-60 CNV map reveals widespread heterogeneity ... 50

2.4.2 Whole genome sequencing read depth analysis finds finer-detailed CNVs than aCGH ... 55

2.4.3 Cell lines are affected more by losses than by gains ... 57

2.4.4 Losses are longer than gains and more gene-dense ... 61

2.4.5 CNV hotspots preferentially occur in regions of higher gene density ... 63

2.4.6 NCI-60 cell lines experience tumour suppressor loss and oncogene gain ... 66

2.4.7 The majority of gene copy numbers are not significant by permutation testing ... 69

2.4.8 Genes with significant copy numbers in CNV hotspots are possible driver genes ... 71

2.5 Discussion ... 74

3 Transient *TP53* repression could lead to *Oncomir-1* activation ... 79

3.1 Abstract ... 79

3.2 Introduction ... 80

3.3 Methods ... 82

3.3.1 Seed/locus miRNA families ... 82

3.3.2 GO and pathway enrichment ... 82

3.3.3 Cancer-related miRNAs and transcription factors ... 83

3.3.4 Statistical analyses ... 83

3.4 Results ... 84

3.4.1 Individual miRNA CNV profiles are confounded by genomic proximity effects ... 84

3.4.2 Seed/locus families enable CNV analyses without genomic proximity bias ... 89

3.4.3 Multi-precursor and non-redundant seed/loci are lost more than expected ... 93

3.4.4 Seed/locus clusters are enriched for cancer-related processes ... 95

3.4.5 Partial loss of miRNA biogenesis pathway genes suggests global miRNA depletion ... 98

3.4.6 Consistent-gain miRNAs are enriched for cell cycle control and signalling ... 101

3.4.7 Target-increasing miRNA/target interactions are enriched for signalling ... 103

3.4.8 C-MYC repression of TP53 could be required for sustained Oncomir-1 activation ... 107

3.5 Discussion ... 114

4 Haploinsufficiency explains the heritable dominant disease burden ... 119

Statement on previous work ... 119

4.1 Abstract ... 120

4.2 Introduction ... 121

4.3 Methods 125

4.3.1 Building human-centric paralog trees from cross-species gene trees 125

4.3.2 Allocating duplication ages to genes in paralog families 126

4.3.3 Disease association, haploinsufficiency, verified miRNA targets and duplications 128

4.4 Results 129

4.4.1 FFS gene ages are based on sequence divergence as well as duplications 129

4.4.2 Genes originating from whole genome events are clearly visible with FFS 132

4.4.3 The peak of miRNA/target interactions are between genes from Vertebrata 134

4.4.4 FFS clearly shows cellular processes in their evolutionary context 135

4.4.5 Less diverged paralogs are older and more haploinsufficient 138

4.4.6 WGD is correlated with a sudden decrease in haploinsufficiency of new genes 141

4.5 Discussion 144

5 Discussion 149

5.1 CNVs affect a wide range of cancer-related processes 150

5.2 Novel candidate cancer driver genes 151

5.3 Haploinsufficiency leads to dominant disease gene retention 152

5.4 MicroRNA derepression occurs widely in cancer 154

5.5 *Oncomir-1* plays a significant role in sustaining the cancer phenotype 155

5.6 Conclusions 158

6 Supplementary Information 159

References 171

Word count: 65,795

Last updated: 20/08/21

Figures

Figure 1.1 - MicroRNA biogenesis 16
Figure 1.2 - Conserved miRNA target sites 21
Figure 1.3 - MicroRNA regulatory motifs 27
Figure 1.4 - Stoichiometric imbalance in dosage compensation 34
Figure 1.5 - Non-linear relationships between genotype and phenotype 38
Figure 1.6 - Stoichiometric imbalance in haploinsufficiency 39
Figure 2.1 - readDepth models, bin sizes and GC content correction 51
Figure 2.2 - High resolution NCI-60 panel-wide CNV map 53
Figure 2.3 - Chromosome 17 CNVs 54
Figure 2.4 - Comparison of whole genome sequencing-based CNVs to aCGH-based CNVs ... 56
Figure 2.5 - CNV frequencies and the proportion of the mappable genome affected 58
Figure 2.6 - Sex chromosomes are affected differently by CNVs to autosomes 59
Figure 2.7 - Aneuploidy frequencies 60
Figure 2.8 - CNV lengths and gene densities 62
Figure 2.9 - Sliding window analyses reveal possible CNV hotspots 64
Figure 2.10 - Gene-dense regions are gained more and lost more than low density regions 65
Figure 2.11 - Tumour suppressors are lost more often than oncogenes 67
Figure 2.12 - Consistently gained genes are more likely to be oncogenes 68
Figure 2.13 - Empirically significantly gained genes on MCF7 chromosome 17 70
Figure 2.14 - CNV hotspots with significant cancer gene copy number changes 72
Figure 3.1 - MicroRNA precursor and seed families cluster by CNV profile 85
Figure 3.2 - Multi-chromosome genomic proximity explains some CNV clusters 87
Figure 3.3 - Seed copy number ratios confirm bias against sex chromosome gains 88
Figure 3.4 - Seed/locus miRNA family construction avoids genomic proximity bias 90
Figure 3.5 - Seed/locus families exhibit CNV clustering across multiple loci 92
Figure 3.6 - Multi-precursor seed/loci are lost more than expected with fewer gains 93
Figure 3.7 - Seed/locus families without redundant seeds are lost more than expected 95
Figure 3.8 - MicroRNA biogenesis gene CNVs indicate global miRNA depletion 99
Figure 3.9 - MicroRNAs with consistent gains are enriched for cancer-related processes .. 102
Figure 3.10 - The majority of miRNAs and their targets have the same CNV type 104
Figure 3.11 - Oncomir and tumour suppressor CNV profile clusters include Oncomir-1 108
Figure 3.12 - Oncomir-1 CNVs imply selection pressure to retain Oncomir-1 function 109

Figure 3.13 - Oncomir-1 seed families form feedback loops with transcription factors 110
Figure 3.14 - CNVs lead to an increase in Oncomir-1 seed dosage in nearly all cell lines 112
Figure 4.1 - Building human-centric paralog trees from cross-species gene tree events 125
Figure 4.2 - Paralog dating methods assign a wide range of ages to genes 127
Figure 4.3 - The phylogenetic tree of the Ensembl Compara cross-species gene tree taxa . 130
Figure 4.4 - The choice of paralog dating algorithm greatly influences the gene ages 131
Figure 4.5 - The FFS age allocation algorithm applied to the Tetraspanin paralog family ... 131
Figure 4.6 - FFS identifies the ohnologs and separates the miRNA/protein-coding ages 133
Figure 4.7 - MicroRNAs mostly target protein-coding genes originating in Vertebrata 134
Figure 4.8 - FFS gene ages differentiate the evolutionary histories of cellular processes ... 137
Figure 4.9 - Less diverged paralogs are older and more haploinsufficient.......................... 139
Figure 4.10 - Dominant disease genes are the most haploinsufficient 140
Figure 4.11 - The overall burden of dominant disease plateaus after WGD 142
Figure 6.1 - The distributions of cell line gains and losses for 1 Mb sliding windows 160
Figure 6.2 - Chromosome arm-level aneuploidies... 163
Figure 6.3 - A cluster of seed/locus families affected by different CNVs 166
Figure 6.4 - Gene/GO mappings as annotated in Ensembl and after ‘cascading’ 168
Figure 6.5 - Seed dosage of Oncomir-1 paralogs in various CNV conditions....................... 169
Figure 6.6 - Examples of cell lines with gene ages which vary with CNV type 170

Tables

Table 2.1 - Oncogene and tumour suppressor clusters.. 68

Table 2.2 - Some significant cancer-related gene CNVs are putative drivers of oncogenesis 69

Table 2.3 - Gained oncogenes and lost tumour suppressors 71

Table 3.1 - Seed/locus clusters are enriched for cancer-related processes 97

Table 3.2 - RBP11-a/b/c subunits are mainly in unmappable regions of the genome 100

Table 3.3 - Enriched terms and pathways for mostly gained miRNAs................................. 102

Table 3.4 - Defining MTIs which are expected to have a net effect on protein levels 106

Table 3.5 - Enriched GO terms and Reactome pathways for net protein change MTIs 107

Table 6.1 - NCI-60 cell lines 159

Table 6.2 - Enriched biological process GO terms for gain hotspots 161

Table 6.3 - Enriched Reactome pathways for gain hotspots 161

Table 6.4 - Enriched biological process GO terms for loss hotspots................................... 162

Table 6.5 - Enriched Reactome pathways for loss hotspots... 163

Table 6.6 - miRNA seed clusters on different chromosomes 164

Table 6.7 - High-spread miRNA seed clusters with precursors in different loci 165

Table 6.8 - Seed/locus families with the same CNV profile despite differing CNVs 167

Table 6.9 - GO and Reactome functional keywords 169

Table 6.10 - Unique Ensembl Compara taxon names and approximate ages 170

Intentionally blank page.

Abstract

The study of genomic variation is vital for our understanding of the gene dosage changes which occur widely in cancer. These dosage changes are a key alteration to the cancer cell whereby tumours bypass protective cellular mechanisms in order to grow and proliferate.

In order to accurately determine the gene dosage changes in a representative range of cancers we used sequencing data for the NCI-60 cell lines to create a high-resolution map of copy number variants (CNVs). Our analysis of CNV hotspots in the cancer genome suggests novel candidate cancer driver genes. These driver genes are enriched for roles in proliferation, angiogenesis and apoptosis, consistent with the hypothesis that tumour cells must escape various protective mechanisms before they gain the hallmarks of cancer.

The presence of paralogs associated with heritable dominant diseases in the human genome is a paradox, since purifying selection would be expected to remove them, and yet they are found throughout the metazoa. To study the role of whole genome duplications (WGDs) in the persistence of these disease-associated genes we implemented a new gene dating method, which provides a more detailed perspective on gene duplications than has been previously possible. We propose that the WGDs in the vertebrate ancestor led to a switch from recessive to dominant disease specifically because of the haploinsufficiency of the retained ohnologs, rather than due to more general dosage sensitivity.

Tumour cells survive gene dosage alterations which are lethal to normal cells, so buffering mechanisms such as miRNAs must be key to these processes. A handful of miRNAs have been annotated with oncogenic and tumour suppressor roles. The most important of these is the *mir-17~92* cluster, also called *Oncomir-1*. We find widespread derepression of cancer-related processes and pathways caused by the frequent loss of tumour suppressor miRNAs, as well as by global miRNA depletion resulting from the disruption of miRNA biogenesis. We propose a novel mechanism whereby transient *C-MYC* elevation leads to *TP53* repression via *mir-663a/1228*, which then allows *Oncomir-1* to repress *TP53* and *PTEN* in a sustained manner. This bistable switch could potentially be reversed with *Oncomir-1* antagonists.

The work presented in this thesis advances our understanding of the role of miRNAs in buffering gene dosages changes in cancer and points the way to possible new interventions for *Oncomir-1*-dependent tumours.

Declaration

With the exception of Chapter 4, no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Chapter 4 is a substantial reworking of and extension to prototype algorithms and hypotheses first advanced in MSc dissertation “Visualisation of the Evolution of Molecular Interactions in Time” (Reardon 2016). Further information is provided at the start of Chapter 4.

Copyright statement

Acknowledgements

I’d like to thank my supervisors Professor Sam Griffiths-Jones, Dr Matthew Ronshaugen and Professor David Robertson for their guidance, patience and enthusiasm. Without their support, knowledge and wisdom this thesis would have been even worse.

I’d also like to thank the previous and current members of the Griffiths-Jones and Robertson labs for their encouragement, interesting discussions, friendship and for propping up various bars with me over the years. In particular I’d like to thank Dr Alex Martin-Geary and Dr David Newman on all those counts.

My family and friends have provided me with continuous encouragement and support (and quite often free beer) throughout my PhD and for this I will always be grateful. My parents in particular have always encouraged me to pursue my ambitions with confidence and I can honestly say I wouldn’t be here without them.

My gratitude also to the Biotechnology and Biological Sciences Research Council for funding my PhD (grant number BB/M011208/1).

Last, but very much not least, I would like to thank my amazing wife for her love, kindness, friendship and support (and for putting up with me while I wrote this thesis). Antonia, I really couldn’t have done this without you, and I love you dearly.

Journal format

This thesis is presented in journal format and consists therefore of a general introduction (Chapter 1), three manuscripts (Chapters 2-4) and a general discussion of the findings and significance of the work (Chapter 5).

Chapter 1 introduces miRNAs and covers the discovery, biogenesis, evolution, function and effects of miRNAs. We then discuss gene variation at both evolutionary and somatic levels, followed by gene dosage considerations and how dosage changes result in disease.

Chapter 2 is a manuscript in preparation, focussing on the creation and analysis of a high-resolution map of copy number variations (CNVs) in cancer-derived cell lines. We characterise the CNVs and resulting gene dosage changes and show that cancer genomes are dominated by partial losses, mainly affecting tumour suppressors, with less frequent gains affecting oncogenes. We develop a statistical framework for assessing the likelihood of CNVs affecting cancer genes and derive a list of novel candidate driver genes.

Chapter 3 is a manuscript in preparation, where we use our detailed map of cancer CNVs to investigate how miRNA buffering of gene dosage changes is disrupted in cancer. We find widespread derepression of cancer-related processes, caused both by specific mutations as well as by a more general depletion of miRNAs caused by disruption to the miRNA biogenesis machinery. We propose a new mechanism whereby transient oncogene-mediated repression of *TP53* can enable activation of well-known miRNA cluster *Oncomir-1*, leading to consistent and sustained repression of *TP53* and *PTEN* by *Oncomir-1*.

Chapter 4 is a manuscript in preparation, building as acknowledged above on prototype work carried out during my MSc (Reardon 2016), where we investigate the origins of heritable diseases over evolutionary timescales. We develop a novel gene paralog dating method and use it to distinguish ancient from more recent evolutionary events, leading to a more parsimonious explanation for the presence of dominant disease-associated genes than previously advanced.

Chapter 5 interprets the conclusions of the earlier chapters in the context of our current understanding of gene dosage buffering, both across evolutionary timescales and as mediated by miRNAs in cancer.

Contributions

All computational work was designed and implemented by Mark Reardon under the supervision of Professor Sam Griffiths-Jones, Dr Matthew Ronshaugen and Professor David Robertson. This thesis was written by Mark Reardon with feedback and advice from Professor Sam Griffiths-Jones and Dr Matthew Ronshaugen, with additional feedback from examiners Dr Jamie Ellingford and Professor Eduardo Eyras. All mistakes, errors and misunderstandings are due to Mark Reardon alone.

1 Introduction

1.1 MicroRNAs

1.1.1 The discovery of miRNAs

It was reported in 1991 that the downregulation of the levels of protein *lin-14*, which is implicated in the timing of cell fate decisions during larval transitions in *Caenorhabditis elegans*, was dependent on negative regulatory elements in the 3'UTR region of the *lin-14* gene, with deletion of several regions of the 3'UTR resulting in accumulation of the *lin-14* protein in cells in later larval stages than in wild type *C. elegans* (Wightman et al. 1991). The same study also found that the levels of *lin-14* mRNA were not altered in cells with inappropriate levels of *lin-14* protein and the authors proposed that this implied an unknown post-transcriptional process of protein downregulation (Wightman et al. 1991).

Two years later it was found that *lin-4*, another gene involved in the timing of cell fate decisions in *C. elegans*, encoded a pair of small RNAs instead of a protein, one of which had a region that was anti-sense complementary to the regions of the *lin-14* 3'UTR previously shown to be implicated in negative post-transcriptional gene regulation (Lee et al. 1993; Wightman et al. 1993). The sites in the 3'UTR region of the *lin-14* mRNA with this complementarity to the smaller *lin-4* RNA were shown to be necessary for the post-transcriptional downregulation of *lin-14* protein levels (Wightman et al. 1993) and were also shown to be conserved in closely related species *C. briggsae*, *C. remanei* and *C. vulgaris* (Lee et al. 1993; Wightman et al. 1993). The two RNAs produced from the *lin-4* gene were found to be 22 and 61 nucleotides long, with the longer RNA predicted to form the stem loop precursor of the shorter RNA (Lee et al. 1993).

For seven years the *lin-4* RNA was thought to be specific to nematodes as well as being the only gene of its type but then, seven years after the discovery of *lin-4*, another gene in *C. elegans* called *let-7* was also found to encode a similar RNA with complementarity to a region in the 3'UTR of *lin-41* that regulated the transition from late-larval to adult cell types (Reinhart et al. 2000; Slack et al. 2000). In addition to this discovery, *lin-4* was found to also target *lin-28* (Moss et al. 1997) and homologs of *let-7* were discovered in humans and *Drosophila melanogaster* among other bilaterian animals (Pasquinelli et al. 2000).

The 22 nt *lin-4* RNA is now known as the first member to be discovered of a class of small RNAs called microRNAs or miRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001), though they were originally referred to as short temporal RNAs or stRNAs due to their involvement in cell fate timing decisions (Lagos-Quintana et al. 2001; Lau et al. 2001). The change in nomenclature from stRNAs to miRNAs was prompted by the discovery that whereas *lin-4* and *let-7* were related to developmental timing decisions, the other short RNAs with similar features but unknown functions were expressed in specific cell types rather than at different developmental stages (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). A registry of the known miRNAs and their sequences called miRBase, that also controls the naming of newly discovered miRNA genes before publication, was set up to catalogue miRNAs (Griffiths-Jones 2004) and in March 2021 miRBase listed 38,589 known miRNAs in 271 species (Kozomara et al. 2019).

1.1.2 Genomics and conservation

While the majority of miRNA genes are found in intergenic regions of the genome a large minority are found in the introns of protein-coding genes, generally with the same orientation as the host gene, implying the coordinated expression of proteins and miRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Aravin et al. 2003; Lagos-Quintana et al. 2003; Lai et al. 2003; Lim et al. 2003b; Baskerville and Bartel 2005). The genomic relationship between miRNA and host gene in these cases is often deeply conserved, such as the miRNA *mir-7* which is found in the intron of *hnRNP K* in mammals and insects, implying that the association between miRNA and host gene is evolutionarily ancient (Aravin et al. 2003).

In contrast to the miRNA genes of *D. melanogaster*, over half of which are clustered together in the genome (Aravin et al. 2003), only about 1/3 of miRNAs in humans and nematodes are expressed from clusters (Lim et al. 2003a; Lim et al. 2003b). Exceptions to these include the orthologs of the *C. elegans lin-4* and *let-7* genes in humans which are clustered and indeed expressed from the same transcript, suggesting that the separation of these genes is nematode-specific (Aravin et al. 2003). MicroRNA genes in a cluster are in general however not necessarily related to each other in this way with little correlation between miRNA family and genomic location (Lagos-Quintana et al. 2001; Lau et al. 2001).

Among the clusters of human miRNAs are *mir-23~27* and *mir-17~92* (Lagos-Quintana et al. 2001).

MicroRNAs have since been found in plants (Reinhart et al. 2002), viruses (Cullen 2006) and green algae (Molnar et al. 2007; Zhao et al. 2007) as well as in a wide range of animals (Grimson et al. 2008; Shabalina and Koonin 2008). Genomic methods for predicting new miRNA genes include simple homology searches based on multiple sequence alignments (Pasquinelli et al. 2000; Lagos-Quintana et al. 2001; Lau et al. 2001), searching for potential stem loops in the vicinity of known miRNAs (Lau et al. 2001; Aravin et al. 2003) and the identification of conserved regions outside the known protein-coding genes that could also form stem loops when expressed as RNA (Lai et al. 2003; Lim et al. 2003a; Lim et al. 2003b).

The majority of miRNA orthologs are conserved between species to a degree proportional to their evolutionary distance so that, for example, most human and mouse orthologs and most *C. elegans* and *C. briggsae* orthologs are conserved (Lagos-Quintana et al. 2003; Lim et al. 2003a; Lim et al. 2003b). The *let-7* miRNA family (where each member has the same 'seed' sequence between 5' nucleotides 2 and 8) on the other hand has fifteen members in humans, four in *C. elegans* and just one in *D. melanogaster* (Pasquinelli et al. 2000; Aravin et al. 2003; Lai et al. 2003). Only one miRNA, *mir-100*, is known to be conserved in both bilaterians and cnidarians (Grimson et al. 2008), whereas across *Bilateria* there are an estimated 34 conserved miRNA families (Christodoulou et al. 2010). Despite this relative lack of deep conservation for miRNAs each significant eumetazoan clade acquired its own characteristic novel miRNAs, with large numbers of novel miRNA families at the split of the protostomes and deuterostomes and at the bases of the vertebrates and the primates (Sempere et al. 2006; Heimberg et al. 2008; Peterson et al. 2009; Berezikov 2011).

More conserved miRNAs tend to evolve at a slower rate than lineage-specific miRNA families (Meunier et al. 2013; Lyu et al. 2014; Ninova et al. 2014), with novel miRNAs coming into existence frequently but rapidly degenerating unless they gain a beneficial function (Nozawa et al. 2010; Lyu et al. 2014). Novel miRNAs are usually expressed in a tissue-specific pattern at low levels (Meunier et al. 2013) and only once a miRNA is fixed in a population and has a beneficial function does it tend to be expressed more widely and at higher levels (Chen and Rajewsky 2007; Meunier et al. 2013), presumably once target genes for which increased repression by the novel miRNA would be deleterious have been modified by purifying selection.

1.1.3 Biogenesis and biological function

The transcription of the majority of miRNAs that reside in the introns of protein-coding genes is regulated by the promotors and enhancers that regulate the expression of the host gene, but around a third of intronic miRNAs have independent promotors, as do all the intergenic miRNAs (Ozsolak et al. 2008).

RNA genes are variously transcribed by three RNA polymerases with, for example, ribosomal RNA genes transcribed by pol-I, short nucleolar RNAs by pol-II and transfer RNAs by pol-III (Ohler et al. 2004). MicroRNA genes have been shown to be expressed *in vitro* by both pol-II and pol-III (Zeng et al. 2002), but as the miRNA primary transcripts are often longer than those known to be processed by pol-III (Lee et al. 2002), it is likely that the majority of miRNAs are transcribed *in vivo* by pol-II (Lee et al. 2004) (Figure 1.1).

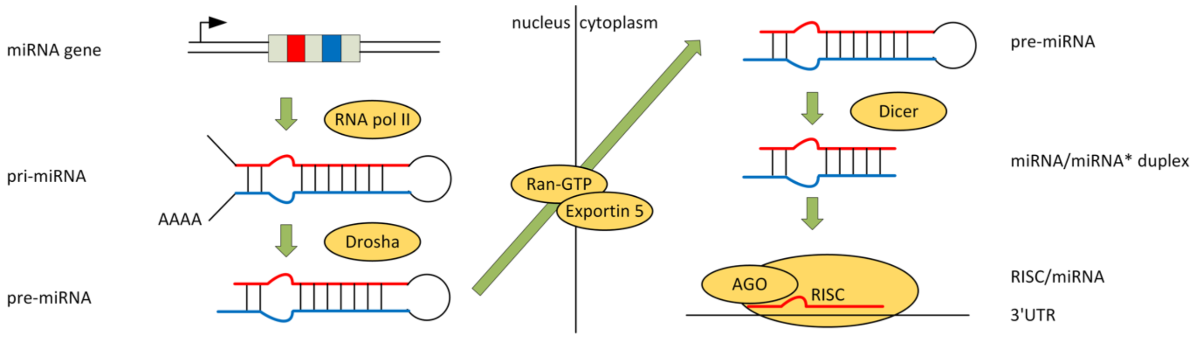


Figure 1.1 - MicroRNA biogenesis

MicroRNA genes are transcribed by RNA pol II into miRNA primary transcripts (pri-miRNAs) and then cleaved by *Drosha* into miRNA precursor hairpins (pre-miRNAs). The precursor is exported from the nucleus to the cytoplasm by *Exportin 5/Ran-GTP* where it is cleaved again by *Dicer* before the mature miRNA strand is incorporated into the RISC/*Argonaut* complex and bound to its target site in the 3'UTR of the mRNA. Mature miRNA strands are shown in red, miRNA* in blue and proteins in yellow.

Unlike small interfering RNAs (Reinhart and Bartel 2002; Ambros et al. 2003) and Piwi-interacting RNAs (Aravin et al. 2007), both of which are also small RNAs that provide target specificity to the RNA silencing machinery, miRNAs are processed from hairpin-like RNA stem loops called miRNA precursors (pre-miRNAs) (Grishok et al. 2001; Lee et al. 2002; Lee et al. 2003) (Figure 1.1). These stem loop precursors of miRNAs are processed from much longer primary transcripts called pri-miRNAs that can contain several miRNA encoding stem loops (Lee et al. 2002; Zeng et al. 2003), as first suggested by the co-expression of clustered miRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001) and the genomic overlap of large regions of expressed sequence tags and miRNA clusters (Lagos-Quintana et al. 2002). Further evidence for the existence of pri-miRNA as a primary transcript prior to processing into pre-miRNA stem loops came from an *in vitro* processing system developed in (Lee et al. 2002) which was used to show, both for clusters of miRNA genes and for single miRNA genes, that PCR products containing the pri-miRNAs are processed into ~65 and ~23 nucleotide RNA fragments, that the ~65 nt fragments are the precursors of the ~23 nt fragments and finally that the ~23 nt fragments are indeed the mature miRNAs (Lee et al. 2002).

The pre-miRNAs are processed from the pri-miRNAs in the nucleus by RNase III endonuclease *Drosha* in complex with dsRNA-binding *DGCR8* (Figure 1.1), as shown both by the *in vitro* cleavage of pri-miRNA into pre-miRNA by *Drosha* and by the increase of pri-miRNA and decrease of pre-miRNA *in vivo* when *Drosha* is subject to RNA interference (Lee et al. 2003; Gregory et al. 2004; Han et al. 2004; Han et al. 2006). In addition to recognising the secondary hairpin structure of the pre-miRNA, in bilaterians *Drosha* also recognises primary sequence elements upstream and downstream of the hairpin in order to distinguish the pre-miRNAs from the many other hairpin-like secondary structures (Auyeung et al. 2013). *Drosha* cleavage of the pri-miRNA leaves the 2 nt 3' overhang that is characteristic of RNase III endonucleases and which forms the base of the pre-miRNA stem loop (Lee et al. 2003; Han et al. 2004). Some intronic miRNAs called mirtrons bypass *Drosha* cleavage and are instead processed by the intronic splicing machinery into pre-miRNAs (Okamura et al. 2007). The pre-miRNAs are then exported from the nucleus to the cytoplasm by the enzymes *Exportin-5* and *Ran-GTP* (Yi et al. 2003) (Figure 1.1).

The PAZ domain of the cytoplasmic enzyme *Dicer* recognises the *Drosha*-cleaved 2 nt overhang of the pre-miRNA (Macrae et al. 2006), after which *Dicer* cleaves both strands of the pre-miRNA stem loop at a distance of approximately two helical turns from the base, to leave an RNA duplex about 22 nt long containing the mature miRNA and the corresponding fragment from the other arm of the pre-miRNA, known as the miRNA* (Hutvagner et al. 2001; Lee et al. 2003) (Figure 1.1).

The mature miRNA strand of the duplex is usually the strand with the 5′ end which can be most easily unwound from the duplex due to its relative internal instability within the RNA duplex (Khvorova et al. 2003; Kawamata et al. 2009), which is then incorporated into a protein complex called the RNA-induced silencing complex or RISC (Hammond et al. 2000; Kim et al. 2009; Kawamata et al. 2009) (Figure 1.1). The RISC contains a protein from the *Argonaute* family (Hammond et al. 2001) which, once bound to the mature miRNA, recognises the complementary sequence in the mRNA (Figure 1.1) and either cleaves and degrades the mRNA if the match between the entire miRNA and mRNA is nearly exact (Hutvagner and Zamore 2002) or inhibits ribosomal translation by remaining bound to the mRNA and interfering with ribosomal activity if the match is only exact in the miRNA seed region between nucleotides 2 and 8 at the miRNA 5′ end, sometimes with additional adjacent complementary and compensatory sites (Zeng et al. 2002; Zeng et al. 2003; Lim et al. 2005). These two mechanisms are known as mRNA cleavage and translational repression respectively.

Even in the case of translational repression however, there is still widespread degradation of mRNAs which are targeted by miRNAs (Baek et al. 2008; Selbach et al. 2008). The *GW182* protein bound to the RISC inhibits translation during the initiation phase by binding to the *PABPC* complex bound in turn to the mRNA poly(A) tail and thus preventing ribosome assembly (Ding and Grosshans 2009; Zdanowicz et al. 2009). *GW182* then directs the mRNA towards deadenylation by *CAF1/CCR4/NOT* (Behm-Ansmant et al. 2006). The deadenylated mRNA is then de-capped by enzyme *DCP2* (Rehwinkel et al. 2005) and finally degraded by the 5′-to-3′ exonuclease *XRN1* (Behm-Ansmant et al. 2006).

A possible explanation for the importance of seed region pairing is that the *Argonaute* protein presents these nucleotides pre-formed into an alpha helix to increase efficiency of binding to the cognate mRNA site (Bartel 2004; Mallory et al. 2004; Elkayam et al. 2012; Schirle and MacRae 2012). Seven nucleotide seed regions are the optimal length in this model as any longer and the RNA would present more than one complete alpha helix for binding to the mRNA with the attendant topological difficulties and any shorter seed would lead to reduced target site specificity (Bartel 2009). This mRNA-binding structure is also the basis for the enrichment of uracil and adenosine nucleotides at the start of the 5′ end of the mature miRNA in bilaterians (Hu et al. 2009).

MicroRNA targets in plants tend to have nearly perfect full-length complementarity (Rhoades et al. 2002) and have been shown to result therefore in mRNA cleavage (Llave et al. 2002; Rhoades et al. 2002). The targets found in (Rhoades et al. 2002) were mainly transcription factors that are associated with cell differentiation, implying a general mechanism in plant cell differentiation whereby miRNAs cleave the mRNA in order to sharpen the transition to a differentiated cell fate by stopping the production of the related protein more quickly (Rhoades et al. 2002).

The RISC-bound mature miRNA continues to function after mRNA cleavage has occurred and so can cleave additional mRNA molecules (Hutvagner and Zamore 2002). The presence of multiple mRNA target sites has little effect in the case of mRNA cleavage since, once cleaved at any site, the mRNA is rapidly degraded and so additional cleavages would have little effect on mRNA levels and hence gene expression (Doench et al. 2003). In contrast, multiple mRNA target sites in the cases of translational repression or transcript destabilisation result in the cooperative action of more than one RISC per mRNA, with the effects proportional to the number of sites, thus allowing the fine tuning of gene expression levels (Doench et al. 2003) and explaining the prevalence of protein-coding genes with multiple 3’UTR miRNA target sites.

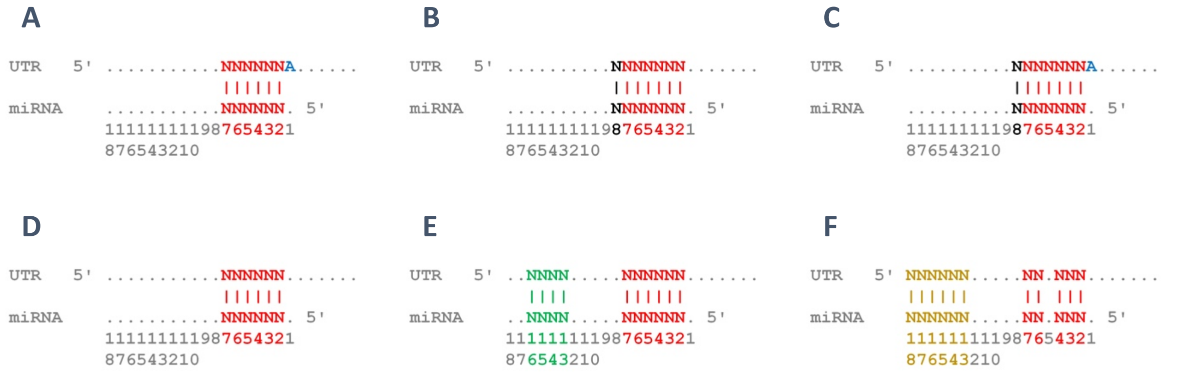
1.1.4 Target prediction, verification and interactions

In contrast to the behaviour of miRNAs in plants, while a few animal miRNAs initiate mRNA cleavage, such as the *mir-196* cleavage of *HOXB8*, *HOXC8* and *HOXD8* (Yekta et al. 2004) or the *mir-127/136* cleavage of *Rtl1/Peg11* (Davis et al. 2005), animal miRNAs tend not to have such complete complementarity to their mRNA target sites and so miRNA target prediction methods additionally take into account inter-species conservation of the seed region (Enright et al. 2003; Lewis et al. 2003; Stark et al. 2003). Despite the greater noise from these methods many of the predictions have been experimentally validated (Lewis et al. 2003; Stark et al. 2003) but, unlike in plants, are found to be less likely to target transcription factors and more likely to target a wider range of biological processes and so have further-reaching effects than simply increasing the speed of the cessation of protein production (Stark et al. 2003).

The first miRNA to be found, *lin-4* in *C. elegans*, was found to be complementary at its 5’ end to its target site in the *lin-14* 3’UTR (Wightman et al. 1993). It has since been shown that not only is perfect complementarity to nucleotides 2 to 8 from the 5’ end of the miRNA predictive of translational repression (Lai 2002) but also that these seed regions are highly conserved in the ortholog mRNAs of other metazoan species (Lewis et al. 2003; Lim et al. 2003b; Stark et al. 2003) and target prediction using these regions is more effective than using any other region of the same length (Lewis et al. 2003). While the majority of functional miRNA target sites are in mRNA 3’UTRs, target sites also exist in 5’UTRs and open reading frames, with the latter more common (Kloosterman et al. 2004; Lytle et al. 2007).

The earliest methods of miRNA target site prediction searched for mRNA sites with near-perfect complementarity to the entire miRNA (Rhoades et al. 2002), an approach that largely succeeded in plant genomes, but which rarely works in animal genomes. Requiring complementarity to just six or seven contiguous nucleotides of the seed region between nucleotides 2 and 8 at the 5’ end of the miRNA (Figure 1.2A-D), restricting mRNA candidate sites to those with high interspecies conservation and ranking candidate sites by the folding free energy of the miRNA/target site duplex resulted in greatly improved detection of miRNA target sites, albeit with wide disagreement between the results from each algorithm due to the allowance of RNA ‘wobble-pairing’ (Enright et al. 2003; Lewis et al. 2003; Stark et al. 2003; John et al. 2004; Kiriakidou et al. 2004).

A subsequent refinement to the TargetScan algorithm of (Lewis et al. 2003) restricted seed matching to strict Watson-Crick pairing and replaced the use of folding free energy to rank target candidates with the consideration of conserved nucleotides adjacent to the seed region, specifically an ‘anchoring’ adenosine residue at the 3’ end of the seed’s cognate region in the mRNA 3’UTR (Figure 1.2A/C) (Lewis et al. 2005). The adoption of similar refinements to those in TargetScan (Lewis et al. 2005) by other algorithms such as PicTar (Lall et al. 2006) and EMBL (Stark et al. 2005) have led to much less disagreement between the results from each. Experimental confirmation of the importance of strict Watson-Crick base-pairing of the miRNA seed region was provided by an analysis using quantitative mass spectrometry of mRNA and protein levels after miRNA transfections and knockdowns (Baek et al. 2008) and also by a microarray-based analysis of mRNA levels after similar manipulation of miRNA levels (Selbach et al. 2008).



A
UTR 5' ....NNNNNNA......
||||||
miRNA ....NNNNNN. 5'
11111111119876543 21
876543210
B
UTR 5' ....NNNNNNN.......
|||||||
miRNA ....NNNNNNN. 5'
11111111119876543 21
876543210
C
UTR 5' ....NNNNNNNA......
|||||||
miRNA ....NNNNNNN. 5'
11111111119876543 21
876543210
D
UTR 5' ....NNNNNN.......
||||||
miRNA ....NNNNNN. 5'
11111111119876543 21
876543210
E
UTR 5' ..NNNN.....NNNNNN.......
|||| ||||||
miRNA ..NNNN.....NNNNNN. 5'
11111111119876543 21
876543210
F
UTR 5' NNNNNN.....NN.NNN.......
|||||| || |||
miRNA NNNNNN.....NN.NNN. 5'
11111111119876543 21
876543210

Figure 1.2 - Conserved miRNA target sites

Types of conserved miRNA target sites, with matched seed nucleotides shown in red, additional matching nucleotides in black, anchoring adenosine residues in blue, matched 3'-supplementary nucleotides in green and matched 3'-compensatory nucleotides in yellow. (**A**) 7 nt site with six seed nucleotide matches and an anchoring adenosine residue. (**B**) 7 nt site with six seed nucleotide matches and an additional matching nucleotide at 5' position 8. (**C**) 8 nt site with six seed nucleotides matching as well as both an anchoring adenosine and an additional position 8 match. (**D**) Marginal six nucleotide target site. (**E**) Marginal six nucleotide target site with 3'-supplementary matches. (**F**) Mismatched seed region with 3'-compensatory matches.

Crucially, restricting the mRNA sites to those that are conserved across multiple species (Lewis et al. 2003) increased the chances that these sites are biologically functional and so facilitated the development of algorithms that did not depend on a training set of known targets (Lewis et al. 2003; Lewis et al. 2005). This also allowed the common features of miRNA target recognition to be defined, primarily the conserved Watson-Crick pairing between nucleotides 2-8 at the 5' end of the miRNA, consistent with the observations that the 5' end of the miRNA is the most highly conserved (Lim et al. 2003b) and that mutations in these regions were the most likely to disrupt miRNA regulation (Doench and Sharp 2004; Brennecke et al. 2005; Lai et al. 2005a).

The conserved strict Watson-Crick pairing of the seed alone raises the specificity signal above false positive noise (Brennecke et al. 2005; Lewis et al. 2005) though false positives still occur and, while the specificity can be further improved by requiring eight nucleotide seed matches (Figure 1.2C), this results in greatly decreased sensitivity to valid target sites (Lewis et al. 2005). Conversely, the target prediction sensitivity can be increased by reducing the length of the conserved seed match to just six nucleotides (Figure 1.2D/E) but only at the cost of reduced specificity caused by an increase in false positive matches (Lewis et al. 2005).

A further refinement to the TargetScan algorithm (Friedman et al. 2009) that incorporated more genomes and had a more sophisticated model of conservation found that when six nucleotide targets were included in addition to seven and eight nucleotide targets, more than 60% of human protein-coding genes have been under negative selection pressure to maintain targeting by miRNAs with an average of more than 400 conserved targets per miRNA family (Friedman et al. 2009). Confirmation that this level of miRNA targeting of protein-coding genes is observed *in vivo* is provided by the finding that when miRNAs are transfected into HeLa cells, the expression profile of hundreds of genes is shifted towards that observed in cells of the tissues in which the miRNAs are preferentially expressed (Lim et al. 2005). Similar results were seen in a study that used chemically modified oligonucleotides complementary to miRNAs to silence the regulatory effect of miRNAs in mice (Krutzfeldt et al. 2005).

In addition to the predominant miRNA seed-based mRNA target sites there are conserved 3'-supplementary sites (Figure 1.2E) where Watson-Crick pairing of 5' miRNA nucleotides 13-16 increases efficacy of miRNA/mRNA binding (Grimson et al. 2007). Further sites known as 3'-compensatory sites (Figure 1.2F) are longer regions of miRNA nucleotide pairing at 5' miRNA nucleotides 13-18, at least six nucleotides long and in some cases long enough to induce mRNA cleavage (Yekta et al. 2004), that compensate for single nucleotide mismatches in the seed region (Grimson et al. 2007). These 3'-compensatory sites are thought to be conserved despite their rarity because they allow a 3'UTR site with slightly mismatched seed pairing to different members of a miRNA family to be targeted by different family members that are expressed during different stages of development, facilitating greater temporal refinement of miRNA regulation (Brennecke et al. 2005; Lewis et al. 2005).

The relatively high false-positive rates of *in silico* miRNA target prediction methods have motivated the recent development of *in vivo* high-throughput methods for the experimental discovery of target sites, using the cross-linking immunoprecipitation (CLIP) (Darnell 2010; Hafner et al. 2010; Grosswendt et al. 2014) and cross-linking, ligation and sequencing of hybrids (CLASH) (Kudla et al. 2011; Helwak et al. 2013) methods to discover the actual binding sites of miRNA-incorporating RISCs.

1.1.5 Expression and selective avoidance

The first miRNAs to be discovered, *lin-4* and *let-7*, along with their homologs in other species, are expressed at different stages of development (Lee et al. 1993; Moss et al. 1997;

Pasquinelli et al. 2000; Reinhart et al. 2000; Slack et al. 2000). Other miRNAs are tissue-specific, such as *mir-122* which is expressed in liver cells (Lagos-Quintana et al. 2002) and *mir-1* which is expressed in the heart (Lee and Ambros 2001; Lagos-Quintana et al. 2002). Different regions and developmental stages of organs can have distinct miRNA expression patterns, such as in the mammalian brain (Krichevsky et al. 2003). The numbers of miRNA molecules expressed in a given cell can vary widely, from less than 1000 *miR-124* per adult nematode cell to more than 50,000 *miR-2*, *miR-52* and *miR-58* (Lim et al. 2003b).

The necessity of the use of inter-species conservation as well as seed region complementarity in all successful target prediction methods (Lewis et al. 2003; Stark et al. 2003; Lall et al. 2006; Friedman et al. 2009), which is not a mechanism available to cells, implies that there must be other determining features of miRNA targeting *in vivo*, such as relying on co-expression to limit the possible targets as with transcription factors (Rhoades et al. 2002; Farh et al. 2005). The effects of the combinations of different miRNA target sites on each mRNA (Lewis et al. 2003) and the effects of additional features adjacent to the seed-based target site (Grimson et al. 2007) act to increase the specificity of miRNA translational repression.

Even non-conserved target sites in mRNAs are functional when the cognate miRNA is present simultaneously in the cell (Farh et al. 2005; Krutzfeldt et al. 2005; Baek et al. 2008; Selbach et al. 2008) and are ten times more prevalent in 3'UTRs than conserved sites (Farh et al. 2005), yet the expression profiles of non-conserved mRNA targets and their cognate miRNAs are strongly anti-correlated *in vivo* (Farh et al. 2005). Over sufficient evolutionary time 7 nt regions of 3'UTRs will accumulate mutations that make them a match for miRNAs expressed in the same cell, causing immediate down-regulation by the co-expressed targeting miRNA, with the result that the mutation often fails to become fixed in the population due to selective disadvantage, in a process known as selective avoidance (Farh et al. 2005; Stark et al. 2005).

Protein-coding genes with 3'UTRs under evolutionary pressure to selectively avoid a co-expressed miRNA are known as 'anti-targets' of that miRNA (Bartel and Chen 2004) and have significantly fewer sites complementary to that miRNA than would be expected by chance (Farh et al. 2005). The numbers of conserved targets and anti-targets are of similar magnitude and so miRNAs have probably had an impact on most mammalian genes via one or other of these mechanisms (Farh et al. 2005).

1.1.6 Loss of function

The widespread targeting of mRNA 3'UTRs by multiple miRNAs (Friedman et al. 2009), whether from the same seed-based family or not, is one possible explanation for the observation that knockout of any given miRNA rarely results in an obvious phenotypical difference (Miska et al. 2007). This targeting redundancy, as exhibited for example by the co-targeting in *C. elegans* of *hbl-1* by three members of the *let-7* family (*mir-48*, *mir-84* and *mir-241*), where it was shown that at least two of the three miRNAs had to be deleted before a phenotypical difference to the wildtype was observed (Abbott et al. 2005), means that it might be necessary to silence not just an entire miRNA family but also all members of the co-targeting miRNAs' families in order for the targeted mRNA to experience significantly reduced miRNA-induced repression and hence exhibit a phenotype.

An additional explanation for the frequent lack of miRNA loss-of-function phenotype is that the regulation of expression levels by a given miRNA might only be essential when there are environmental stresses that the miRNA regulation would otherwise counteract. An example of this occurs in the development of the *Drosophila* eye where *mir-7*, in a negative feedback loop with transcription factor *Yan*, controls the change from progenitor cell to photoreceptor cell (Li and Carthew 2005). It was shown that in normal developmental conditions there is no phenotypical consequence of *mir-7* knockout but, when the temperature was fluctuated during the larval stage, *mir-7* knockout *Drosophila* had sensory organ defects (Li et al. 2009).

1.1.7 Regulatory motifs

Genes expressed in specific tissues have longer 3'UTRs with more miRNA target sites than genes associated with core cellular processes (Stark et al. 2005) and miRNA expression becomes more varied both during the development of embryos (Ji et al. 2009; Thomson et al. 2006) and in more complex organisms (Heimberg et al. 2008; Lee et al. 2007). Together with the observation that the expression profiles of miRNAs and their targets are anti-correlated across neighbouring tissues (Farh et al. 2005), this suggests that miRNAs increase the robustness of developmental transitions and aid in the maintenance of cell fate decisions via the mechanism of selective avoidance (Farh et al. 2005; Stark et al. 2005).

A regulatory network that can reinforce such anti-correlated tissue-specific expression is the coherent feed-forward loop in conjunction with a feedback loop (Mangan et al. 2003). In the coherent feed-forward loop an intermittently expressed transcription factor inhibits the transcription of a target gene and activates a miRNA that also inhibits the translation of the target gene, thus reinforcing the decision against transient fluctuations in the levels of the transcription factor due to the relatively long-lasting nature of the miRNA (Figure 1.3A) (Mangan et al. 2003). A tissue in which the transcription factor is expressed will therefore have reliably little target gene expression. The addition of a feedback loop where the target gene inhibits the miRNA reverses the effect such that a tissue in which the target gene is already expressed will not be able to express the miRNA and so the intermittent expression of the transcription factor will not be able to robustly inhibit the target gene (Figure 1.3A) (Mangan et al. 2003). An example of this type of composite network occurs in granulopoiesis in bone marrow, where *C/EBPα* inhibits the transcription of *E2F1* and activates *mir-223* which inhibits the translation of *E2F1*, with the feedback loop component coming from the inhibition of *mir-223* by *E2F1* (Figure 1.3A) (Pulikkan et al. 2010). Another variant of coherent feed-forward loop occurs when a transcription factor activates the transcription of the target gene and inhibits a miRNA which is an inhibitor of the target gene, thus reinforcing the activation of the target gene (Mangan and Alon 2003).

Robustness to extrinsic noise such as transcription factor levels can be provided by incoherent feed-forward loops in which a transcription factor activates the transcription of a target gene at the same time as activating a miRNA that inhibits the translation of the target gene (Figure 1.3B), with the effect that transient fluctuations in the transcription factor levels are counteracted by the corresponding increase or decrease in the levels of the miRNA, thus decoupling expression levels of the target gene from the levels of the transcription factor (Mangan and Alon 2003). An example of this occurs when *C-MYC* is expressed which activates expression of *miR-17-5p* and *miR-20a*, both of which inhibit translation of *E2F1* which is also activated by *C-MYC* (Figure 1.3B) (O'Donnell et al. 2005). Incoherent feed-forward loops have also been characterised as ‘sign-sensitive accelerators’ that increase the speed of transition of expression levels in just one direction, from off to on for example, in contrast to coherent feed-forward loops which act as ‘sign-sensitive delays’ (Mangan and Alon 2003).

A simpler mechanism for increasing robustness to transcriptional noise known as ‘weak buffering’ is the interaction between miRNA and mRNA copy numbers in the cell (Mukherji et al. 2011). When there are sufficiently more miRNA molecules than mRNA transcripts the protein output is negligible since the mRNA are all silenced, but when the level of miRNA decreases or the mRNA level increases past a threshold the protein output not only increases from zero but does so in a manner dependent on the rate of transcription (Paulsson 2004), thus avoiding oversensitivity to transcriptional noise at low levels of transcription when the noise in relative terms would be expected to be highest (Mukherji et al. 2011).

Another type of network which reinforces cell fate decisions in conjunction with the feed-forward/feedback network is the mutual negative feedback loop, where two elements inhibit each other (Figure 1.3C). In the example of granulopoiesis discussed above there is an additional mutual negative feedback loop involving *mir-223* and another transcription factor *NFI-A* which competes with *C/EBPα* to bind to the *mir-223* promoter and, prior to granulopoiesis, inhibits expression of *mir-223* (Figure 1.3C) (Fazi et al. 2005). After retinoic acid causes the cell to transition into a granulocyte, *C/EBPα* out-competes *NFI-A* to bind to the *mir-223* promoter and activates *mir-223* expression, which in turn represses transcription factor *NFI-A* expression and reinforces the cellular decision (Fazi et al. 2005).

In addition to the robustness conferred by miRNAs to developmental transitions and cell fate decisions, miRNAs also buffer some of the intrinsic variation in gene expression that arises from the stochastic nature of both transcription and translation events (Ozbudak et al. 2002; Raser and O'Shea 2005; Blake et al. 2006). Elevated transcription rates reduce the noise in transcription over time with translation only linearly amplifying any noise and so an increase in transcription events with a simultaneous decrease in translation events smooths expressed protein levels over time (Paulsson 2004). The mutual negative feedback loop has this buffering effect against the stochastic nature of transcription where increased transcription of a transcription factor leads to increased transcription of the transcription factor’s inhibitory miRNA, such as in the homeostatic regulation of *MeCP2* levels in neurons where *MeCP2* activates *BDNF* which activates *mir-132*, an inhibitor of *MeCP2* (Klein et al. 2007).

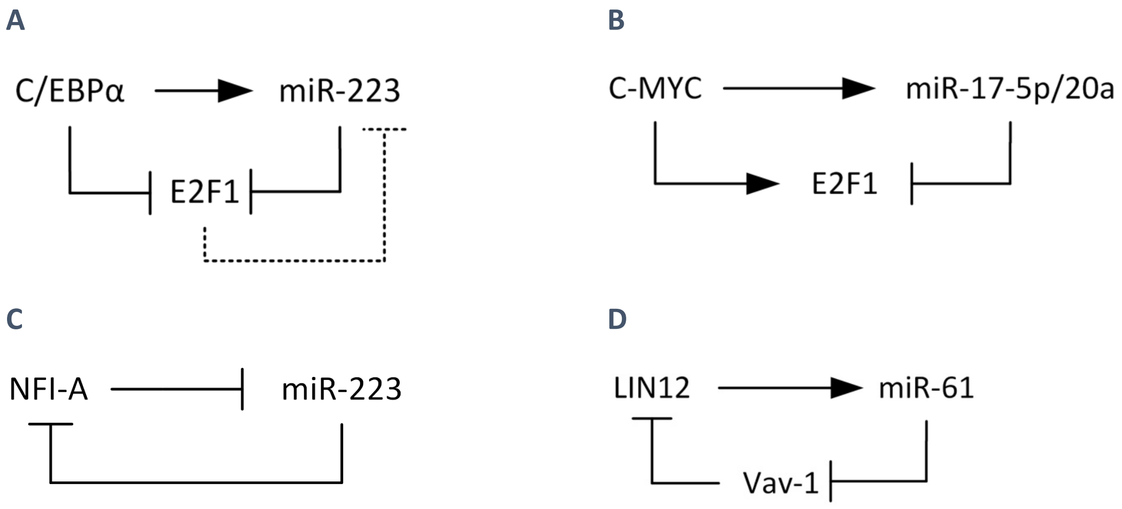


Figure 1.3 - MicroRNA regulatory motifs

Transcription factor, miRNA and target gene regulatory motifs with activation shown as an arrowhead and repression shown as a bar. (**A**) Coherent feed-forward loop with additional negative feedback loop (dotted line). (**B**) Incoherent feed-forward loop. (**C**) Mutual negative feedback loop. (**D**) Indirect positive feedback loop.

Similar in structure to the mutual negative feedback loop is the indirect positive feedback loop in which a transcription factor activates a miRNA that inhibits the translation of a second transcription factor which is an inhibitor of the first transcription factor (Figure 1.3D), a regulatory motif in which a miRNA can indirectly increase rather than decrease expression levels. This occurs during the differentiation of *C. elegan*s vulval cells when *LIN12* activates *mir-61* which inhibits *Vav-1* which in turn inhibits *LIN12*, with the consequence that *LIN12* activation is self-reinforcing (Figure 1.3D) (Yoo and Greenwald 2005).

While these bi-stable feedback loops are beneficial during development they can also be detrimental when the miRNA expression is lost, as can occur in the negative feedback loop between transcription factor *ZEB1* and the *mir-200* family of miRNAs that reinforces the transitions between epithelial and mesenchymal cell types during development (Bracken et al. 2008; Burk et al. 2008; Paterson et al. 2008). When *mir-200* expression is lost, as occurs in some carcinomas, the cell can switch back to a mesenchymal state resulting in an increased propensity to metastasize (Paterson et al. 2008; Gibbons et al. 2009).

Diploid organisms are able to tolerate many single copy recessive loss-of-function mutations and yet, despite the similarly small change in expression level of a miRNA's target of less than 2-fold (Baek et al. 2008), miRNAs and their cognate mRNA target sites are clearly under selective pressure to maintain or avoid interactions. This apparent contradiction can be explained by the observations that some miRNAs have more than one target site in a 3'UTR, such as *let-7* and its target *HMGA2* (Mayr et al. 2007), and also that multiple co-expressed miRNAs often have the same target and so increase the repressive effect (Friedman et al. 2009), additively if with sufficiently widely spaced target sites in the 3'UTR but multiplicatively when the target sites are in close proximity (Bartel 2009; Mukherji et al. 2011). Additional mechanisms for multiplying the repressive effect of a miRNA are positive feedback loops (Yoo and Greenwald 2005) and the targeting of multiple components of a complex or pathway (Linsley et al. 2007).

1.2 Genetic variation

1.2.1 Evolutionary variation

The largest-scale mutation known to occur is the addition of an entire set of chromosomes, known as polyploidization, with the resulting species or individual organisms being referred to as triploid (three sets of chromosomes), tetraploid (four sets of chromosomes) and so on (Otto and Whitton 2000). Most eukaryotic organisms are diploid although their gametes are generally haploid with just one set of chromosomes (Otto and Whitton 2000). Polyploidy is much more common in plants than in animals, occurring in more than 30% of plant species (Masterson 1994) with only occasional examples in animals (Lewis 1979). Polyploidy in animals occurs more frequently in invertebrates than in vertebrates (Otto and Whitton 2000). When polyploidy arises as the result of the combination of chromosomes from more than one species it is known as allopolyploidy and polyploids that are formed from multiple sets of chromosomes from the same species are known as autopolyploids (Otto and Whitton 2000).

Autopolyploidy occurs relatively frequently in mammals with 5% of spontaneously aborted human foetuses showing triploidy or tetraploidy, which rarely leads to viable foetuses (Creasy et al. 1976). In contrast, the rate of single gene duplications has been estimated at around 10^{-8} per gene per generation (Lynch 2007). Aneuploidy, the variation in the number of a single chromosome as opposed to the entire set of chromosomes, is four times more common than polyploidy in spontaneously aborted human foetuses (Creasy et al. 1976).

Two rounds of whole genome duplications (WGDs), where an error during meiosis leads to twice the usual number of chromosomes, are thought to have occurred in early vertebrate history (Ohno 1970). The retained gene copies resulting from WGD events, now known after their definition by Sasumu Ohno as ohnologs (Wolfe 2000), have been shown to be more essential, in that their deletion causes embryonic sterility or lethality, than genes arising from small-scale duplications (SSDs) (Makino et al. 2009). Genes that are not known to have been duplicated by either mechanism are known as singletons and have been shown to be as essential as ohnologs (Makino et al. 2009).

In addition to being more essential than SSDs, ohnologs have been shown to have more interaction partners and to occupy a more central position in the genome's interaction network than SSDs (Huminiecki and Heldin 2010), as well as being more conserved (Lynch and Hagner 2015), with SSDs tending to be on the periphery of the interaction network at first and only gradually acquiring interactions and increasing in pleiotropy and essentiality (Zhang et al. 2015). A gene's position within the interaction network is also predictive of disease association, with peripheral, co-expressed and interacting SSDs tending to be associated with similar diseases (Goh et al. 2007) and the more central ohnologs tending to be associated with somatic cancers (Garcia-Alonso et al. 2014).

Retained genes that are duplicates of other genes in the same species, known as paralogs in contrast to the orthologs that are duplicated by speciation events (Fitch 1970), tend to experience asymmetrical pressure from purifying selection after duplication. A duplicate can continue to support the ancestral function with the other copy acquiring a completely new function over time in a process called neofunctionalisation (Rastogi and Liberles 2005). Alternatively, one copy splits aspects of the function with the other copy in a process called subfunctionalisation (Braun and Liberles 2003) or the duplicate simply becomes non-functionalised as mutations cause pseudogenisation (Lynch and Conery 2000).

In order to analyse the shared evolutionary history of genes that have undergone duplication and speciation using inter-species comparative genomics, phylogenetic trees are created from multiple sequence alignments by a variety of methods such as the maximum likelihood approach (Felsenstein 1981) and the neighbour-joining method (Saitou and Nei 1987). In conjunction with the fossil record and theories of molecular substitution rates during genetic drift (Benton and Ayala 2003; Benton and Donoghue 2007; Archibald 2003), the approximate divergence ages of taxa in the phylogenetic trees can be derived, forming the basis for the assignment of approximate dates to gene duplication events despite these events having occurred in the distant past and in species that are mostly extinct.

A variety of methods have been used to deal with the ambiguity inherent in assigning dates to genes that have undergone repeated duplication and speciation events. One such method assigns to each gene the age of its last common ancestor (LCA) with the consequence that the ages of all genes in a paralog family are weighted to the most evolutionarily ancient date in the tree (Domazet-Loso and Tautz 2008). A similar approach, the duplicate common ancestor (DCA) method, assigns the date of the oldest duplication in a gene's history (Dickerson and Robertson 2012). A method that is instead weighted to evolutionarily recent dates assigns the date of the most recent duplication (MRD) in each gene's history (Dickerson and Robertson 2012). A drawback of the DCA and MRD methods is that they rely on the presence of duplications in the genes' histories and so must fall back to the LCA method or not assign a date for singleton genes.

1.2.2 Germline and somatic variation

In addition to the genetic variation that humans have inherited from ancestral species there are frequent and often heritable copy number variations (CNVs) from the usual two copies of each autosomal gene, with an average of 11 CNVs in each healthy human individual (Sebat et al. 2004). These CNVs are caused by deletions, duplications and movements of DNA that encompass entire genes and range in size from a few thousand bases to millions of bases long (Sebat et al. 2004). In March 2021 the Database of Genomic Variation catalogued nearly 10 million different human CNVs, covering the majority of every chromosome (MacDonald et al. 2014).

CNVs are caused by errors during meiosis, mitosis and DNA repair with a general mechanism for the creation of non-repeating CNVs called microhomology-mediated break-induced replication (MMBIR) (Slack et al. 2006). In MMBIR a stall occurs during DNA replication causing disengagement of the DNA replication fork from the template DNA strand, followed by annealing to another nearby fork in the genome that has homology at the 3' end (Slack et al. 2006). This will result in a deletion if the new fork is upstream of the stall or a duplication if downstream, with the orientation of the newly incorporated DNA of a duplication determined by whether the new fork incorporates the lagging strand, leading to the copied DNA having its original orientation, or the leading strand, in which case the duplicated DNA is reversed (Slack et al. 2006).

The existence of proximate duplicated regions of DNA with very high sequence identity, such as those arising from a CNV duplication, can catalyse the creation by the DNA replication machinery of other repeated CNVs between the two duplicated regions in another mechanism called nonallelic homologous recombination (NAHR) (Stankiewicz and Lupski 2002). If the two duplicated regions have the same orientation, then duplications or deletions of the intervening region can occur, whereas regions with opposite orientations will cause the inversion of the region between the two duplications (Stankiewicz and Lupski 2002).

The smallest scale variation that can occur is the single nucleotide polymorphism (SNP), thought to be the most common genetic variation in humans, with at least 11 million SNPs occurring in the human population with a minor allele frequency (MAF) of >1%, 7 million of which have a MAF of >5% (Kruglyak and Nickerson 2001).

Alleles of SNPs that are close to each other in the genome are often correlated, due to their relatively high frequency of occurrence when compared to recombination crossover points during DNA replication, in a process known as linkage disequilibrium (LD) (Slatkin 2008). The International HapMap Project found in 2004 that most of the SNPs with a MAF of >5% were correlated in this way and could be grouped into LD ‘bins’, with about half a million such LD bins observed in individuals of European or Asian descent and about a million in individuals with African ancestry (International HapMap et al. 2007).

1.3 Dosage compensation

1.3.1 Gene dosage balance hypothesis

Early studies investigating the difference between the consequences of the addition of one chromosome versus the addition of an entire set of chromosomes found that creating an aneuploid genotype by adding just one chromosome was generally deleterious and yet adding an entire set of chromosomes to create a polyploid genotype had little phenotypic effect (Blakeslee et al. 1920). These aneuploid phenotypes were later shown to be due to the disturbance of the stoichiometric relationships of macromolecular complexes, signalling pathways and transcription factors (Birchler and Newton 1981), in a phenomenon now known as the gene dosage balance hypothesis (Papp et al. 2003).

Most of the expression levels of the genes affected by aneuploidy were varied by an amount directly or inversely proportional to the change in copy number although some varied considerably more (Birchler and Newton 1981), consistent with the concept of a critical concentration and hence sigmoidal relationship between genotype and phenotype discussed below in relation to haploinsufficiency (Birchler and Newton 1981). Also supporting the gene dosage balance hypothesis is the finding that the likelihood of the maintenance of a copy number variation in the human population is negatively correlated with the number of protein domains in the affected genes that interact with other proteins, suggesting again the deleterious effects of interfering with the stoichiometric relationships between interacting proteins (Schuster-Bockler et al. 2010).

1.3.2 Dosage compensation

Dosage compensation occurs in macromolecular complexes, regardless of absolute expression level, if a copy number variation with an inverse effect on a gene's expression level also affects the gene's interaction partners, so that the effects of decreasing (or increasing) the levels of an inverse regulator at the same time as decreasing (or increasing) the levels of the interactors cancels out (Birchler and Newton 1981).

Similarly, an α-β-γ trimer can experience reduced yield due to an increase in β since unfinished α-β or β-γ dimers will sequester the available α and γ monomers (Veitia 2002). Therefore, if the normally expressed levels of β are indeed limiting the levels of α-β-γ by titration (Figure 1.4A), then a loss-of function in a β allele will be compensated for by an increase in α-β-γ trimers as the limiting factor is reduced (Figure 1.4B), a phenomenon known as the inverse dosage effect, which has been observed in monosomic aneuploids with doubled expression levels and trisomic aneuploids with two-thirds expression levels (Guo and Birchler 1994).

Another mechanism for avoiding excessive and potentially titrating levels of a monomer is if the aggregation into oligomers or macromolecular complexes masks a degradation signal on the monomer and so the bound monomer avoids degradation by the proteasomes and the excess free monomer is degraded (Asher et al. 2006). This is consistent with the observation that the protein abundances in trisomic and tetrasomic cells were similar to diploid cells even though the mRNA levels varied proportionally with copy number (Stingele et al. 2012).

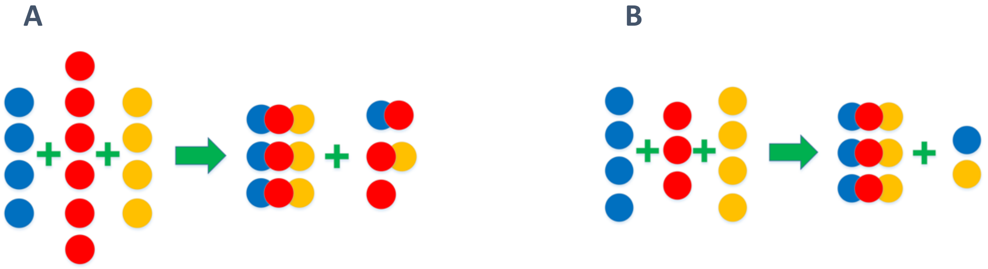


Figure 1.4 - Stoichiometric imbalance in dosage compensation

Stoichiometric imbalance scenarios for protein complexes, with leading α monomers shown in blue, bridging β monomers in red and trailing γ monomers in yellow. (**A**) The normally expressed levels of β limit by titration the yield of α-β-γ trimers. (**B**) A heterozygous null mutation in β produces the same yield of α-β-γ by reducing the titrating α-β or β-γ dimers.

The inverse dosage effect also occurs in transcription but only when the copy numbers of the negatively regulating gene and its target are co-varied, either if the α-β-γ trimer described above is a transcription factor and so the increase in β leads to a decrease in α-β-γ and hence in transcription (Veitia 2002) or if a transcriptional repressor such as a miRNA and its target experience the same copy number variation. Additionally, if the level of a transcription factor is limiting on the expression of the activated gene, then a heterozygous deletion of the target gene will lead to increased transcription of the remaining allele due to the reallocation of the available transcription factors (Veitia et al. 2013).

1.3.3 Canalisation and morphological complexity

As reviewed above, miRNA-based coherent feedforward loops act as a failsafe in cell fate decisions (Stark et al. 2005) and help to enforce different expression patterns in neighbouring cells of different tissue types (Farh et al. 2005) by the mechanism of selective avoidance (Bartel and Chen 2004). Robustness to extrinsic noise and the fine tuning of expression levels are provided by miRNA-based incoherent feedforward loops (Mangan et al. 2003).

The overall effect of these various miRNA-mediated forms of robustness to noise, together with the buffering effects of protein-folding chaperones such as *Heat Shock Protein 90* (Queitsch et al. 2002), is one of canalisation where phenotypical outcomes become more consistent despite fluctuations in both intrinsic and extrinsic signals (Waddington 1959). More formally, canalisation is defined as a reduction in the variance of phenotypic traits and hence stabilisation of the overall phenotype, caused by buffering at a genetic level that has evolved under the pressure of natural selection (Gibson and Wagner 2000).

One consequence of canalisation is that so-called cryptic genetic variations can accumulate without affecting phenotype until the canalization-conferring network is disrupted, thus contributing to evolutionary innovation by acting as a store of variability until mutation, the weakening of protein-folding chaperone activity or environmental changes disrupt the canalisation (Gibson and Dworkin 2004). An example of this disruption of canalisation occurs in *D. melanogaster* when the *mir-9a/senseless* network is disrupted, either by mutation-induced *mir-9a* homozygosity or by mutation of the *miR-9a* binding sites in the *senseless* 3' UTR, with the reduction in copy number of *miR-9a* being sufficient to lead to a two- or three-fold decrease in the heritability, and so an increase in the variance, of the number of bristles on the scutellum (Cassidy et al. 2013).

Over evolutionary time this canalisation, and therefore increased heritability of phenotype, has been proposed as an explanation for the sudden appearance and stability of the metazoan body plans that appear in the fossil record around 550 million years ago in the Cambrian explosion (Peterson et al. 2009). A phylogenetic analysis of 24 invertebrate and vertebrate taxa with a common ancestor 800 million years ago found that miRNA families were acquired at every node in the tree and that each lineage had at least one novel miRNA family (Peterson et al. 2009). Together with the discovery that morphological variation decreases within taxa over geological time (Webster 2007), it was argued that this meant that the miRNA-induced canalisation of phenotypical traits allows increases in morphological complexity by increasing the heritability of phenotypical traits via the increased robustness of gene expression levels to noise (Peterson et al. 2009).

This model of miRNA acquisition acting to increase the heritability of phenotypical traits and therefore allowing for evolutionary innovation is distinct from the scenario of cryptic genetic variations being released and subjected to selection when canalisation is disrupted and suggests an explanation for the seeming contradiction between this model and miRNA loss-of-function (LOF) and gain-of-function (GOF) studies. GOF studies often show that acquiring a novel miRNA or expressing it in different tissues or at different developmental stages is detrimental due to the alteration of the expression levels of many genes (Farh et al. 2005) and that losing miRNAs or entire families often has no phenotypic effect due to targeting redundancy (Miska et al. 2007). However, it is important to note that these are the effects on individuals; when the effect of acquiring miRNA families at a population level and over evolutionary time is understood to be one of increased precision of heritable traits, then there is no reason that fixing of a novel miRNA family in a population should be deleterious, so long as the novel miRNA is initially expressed at a low level in a tissue-specific manner (Meunier et al. 2013), until it gains a beneficial function and can then start to increase in pleiotropy (Chen and Rajewsky 2007; Meunier et al. 2013). Similarly, the loss of a miRNA would only be visible at the population level as an increase in the variance of the phenotypical traits which had been under the influence of the lost miRNA, even if it became fixed in the population (Peterson et al. 2009).

1.4 Disease

1.4.1 Dominant and recessive disease

The classical theory of dominant and recessive traits was proposed in the 19th century to explain how the 'factors' influencing a trait, the colour of flowers for example, were expressed in a dominant or recessive manner, with the recessive factor's influence hidden by that of the dominant factor (Mendel 1866). Diseases associated with mutations to a single gene are now known as Mendelian diseases. Genes where a mutation to a single allele causes the disease are known as dominant disease genes and genes where both alleles have to be mutated before a phenotype occurs are referred to as recessive (Furney et al. 2006). For example, diseases where the mutation is to a transcription factor gene are mainly dominant and so a mutation to a single allele is sufficient to cause the phenotype, whereas diseases caused by mutations to enzyme genes are generally recessive and so both alleles need to be mutated before the disease occurs (Jimenez-Sanchez et al. 2001).

Unsurprisingly given that the theory was proposed before modern genetic theory, the classical theory of dominance is a qualitative simplification of the actual quantitative situation, where co-dominant alleles are the result of a linear relationship between genotype and phenotype and dominance and recessiveness occur as the result of a non-linear sigmoidal relationship (Figure 1.5A) (Veitia 2002). The levels of phenotypic penetrance and expressivity for a heterozygous mutation, respectively the percentage of individuals that exhibit the phenotype and the severity of the expressed phenotype, depend on how close the genotype is to the inflexion point of the sigmoidal curve and on the influence of the environment or other genes in the affected gene's network (Figure 1.5B) (Veitia 2002).

Apparently healthy humans can have thousands of SNPs affecting protein-coding genes and dozens of heterozygous mutations (1000 Genomes Project et al. 2010) including loss-of-function mutations or 'dominant-negative' mutations where, for example, a mutant protein deleteriously affects a protein complex (Herskowitz 1987). The dominant-negative effect offers an explanation for the preferential retention of multi-domain protein paralogs after whole genome duplication (Gibson and Spring 1998), since a protein A and its paralog B that form dimers would form 100% active dimers (16 out of 16 combinations of A and B) if all four paralogs were functional, 56% active dimers (9 out of 16) if only three paralogs were functional and just 25% active dimers (4 out of 16) if only two paralogs were functional, with the increasingly deleterious effects being actively selected against (Perez-Perez et al. 2009).

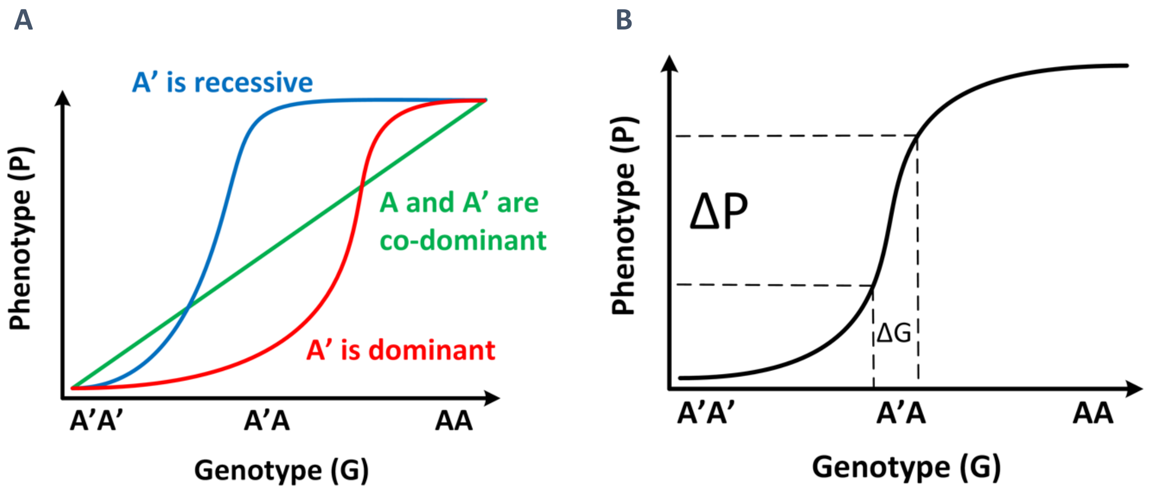


Figure 1.5 - Non-linear relationships between genotype and phenotype

(**A**) The relationships between genotype dosage and phenotypic trait when alleles A and A’ are co-dominant and have equal effect on the phenotype (straight green line), when allele A’ is recessive and so the heterozygote has a similar phenotype to AA (sigmoidal blue line) and when allele A’ is dominant and so the heterozygote has a similar phenotype to A’A’ (sigmoidal red line). (**B**) A small change in genotype dosage, ΔG, resulting from environmental influence or other interacting genes, can result in a large change in phenotype, ΔP, when the genotype is near the sigmoidal inflexion point.

Mutations at different loci can also combine their effects such that while each mutation is recessive their combined phenotype exhibits a dominant trait, such combined effects are known as digenic inheritance for mutations in two loci, trigenic for three loci and, generally, oligogenic (Stearns and Botstein 1988).

1.4.2 Haploinsufficiency

In contrast to the recessive disease-causing mutations in enzyme-encoding genes, heterozygous loss-of-function in loci that encode monomers of macromolecular complexes or that encode transcription factors can result in dominant phenotypes in a phenomenon known as haploinsufficiency (Seidman and Seidman 2002).

The assembly of a macromolecular complex from a monomer, whether it occurs by building successively larger oligomers or by nucleation, implies a critical concentration of the monomer for the reactions to occur (Oosawa and Kasai 1962) and hence a sigmoidal relationship between genotype and phenotype, leading to dominant haploinsufficiency when the expressed level of monomer from just one allele is below the critical concentration (Veitia 2002). Haploinsufficiency can also occur in complexes if one of the molecules acts as a bridge between multiple instances of another, such as in the trimer α-β-α (Figure 1.6A) where a heterozygous loss-of-function mutation in β leads to a proportional decrease in the trimer (Figure 1.6B) but a similar mutation in α leads to much lower trimer expression because incomplete α-β and β-α dimers sequester the available α molecules (Figure 1.6C) (Veitia 2003a).

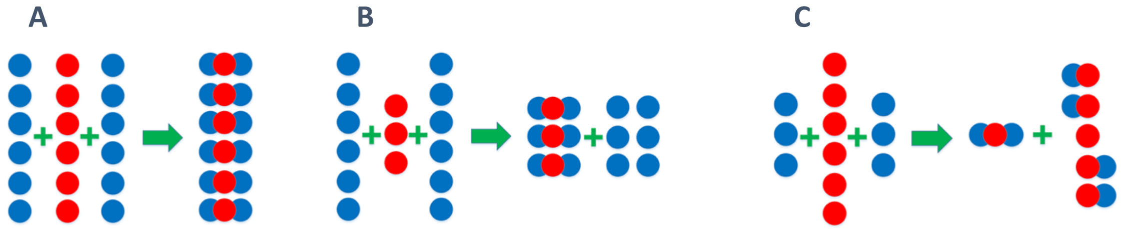


Figure 1.6 - Stoichiometric imbalance in haploinsufficiency

The effects of heterozygous null mutations in monomers of α-β-α trimers. (**A**) Balanced production of the monomers. (**B**) A null mutation in bridge monomer β. (**C**) A null mutation in flanking monomer α. Flanking α monomers are shown in blue and bridging β monomers are shown in red.

A similar sigmoidal relationship exists between transcription levels of a gene and the concentration of its activating transcription factors when there are multiple transcription factor binding sites (Veitia 2003b). This sigmoidal relationship results not only in cooperativity between the transcription factors as the bound ones attract the others, but also results in transcriptional synergy as the transcription factors cooperate to attract the transcription machinery, resulting however in haploinsufficiency if the loss-of-function mutation causes the concentration of the transcription factor to fall below the sigmoidal inflexion point (Veitia 2003b). An increase in the number of binding sites moves the threshold to a lower transcription factor concentration and so, if a heterozygous mutation causes loss of function in one allele of the transcription factor, only the target genes with a threshold above the reduced concentration will exhibit haploinsufficiency (Veitia 2003b).

Haploinsufficiency can also occur in tissues and at developmental stages where one of the two alleles at a locus is randomly silenced in a process called monoallelic gene expression, leading to tissues with a mosaic of cells expressing each allele and with the consequence that a heterozygous loss-of-function mutation would either have no effect in a cell that is not expressing that allele or lead to total loss of gene product in a cell that has not silenced that allele (Nutt and Busslinger 1999). Genomic imprinting is a similar process where either the maternal or paternal allele is silenced, with the same susceptibility to haploinsufficiency as in monoallelic gene expression, and is thought to provide a selective advantage in a population in a rapidly changing environment (Beaudet and Jiang 2002).

1.5 Summary and aims of project

Cancer cells frequently have widespread gene dosage changes caused by copy number variations and these dosage changes disrupt a wide range of cellular processes, leading to the cancer phenotype. Accurate determination of the endpoints of CNVs and of the resulting dosage changes is essential for understanding the causes of cancer. Previous studies of the CNVs occurring in the cell lines of the well-characterised NCI-60 panel have been based on relatively low-resolution aCGH assays or whole exome sequencing analyses and so we will use read depth analysis of whole genome sequencing data to produce a higher-resolution map of the CNVs in the NCI-60 cell lines than has been previously calculated from aCGH assays.

We will use this more accurate map of cancer CNVs to comprehensively characterise the CNV landscape in the NCI-60 cancer cell lines and to investigate the effects of gene dosage changes on cancer-related processes. We will develop statistical methods for determining the CNVs which are causal in oncogenesis and compare the genes found by these analyses to those already known to be cancer driver genes in the hope of discovering novel cancer drivers.

Cancer cells survive widespread dosage changes which are deleterious to normal cells and so we will investigate the roles of miRNAs in buffering the effects of CNVs in cancer-derived cell lines. We will use our high-resolution NCI-60 CNV map to look for global patterns of miRNA gain and loss in cancer genomes and we will analyse the effects of these CNVs on the regulation by miRNAs of cancer processes and pathways.

The gene dosage buffering mechanisms available to cancer cells have evolved over long timescales since the advent of multicellularity and so we will also investigate the evolutionary histories of gene dosage compensation mechanisms. We will analyse the surprisingly common retention of multiple copies of disease-associated genes in metazoan genomes in an effort to understand gene dosage compensation at an evolutionary level.

This project aims to increase our understanding of the gene dosages changes which occur widely in cancer, both at a cellular level and over evolutionary timescales, in order to advance our knowledge of the roles which miRNAs play in gene dosage compensation.

Intentionally blank page.

2 High-resolution CNVs reveal potential novel cancer driver genes

2.1 Abstract

Tumour cells and the cell lines derived from them are characterised by widespread gene dosage changes caused by point mutations, copy number variations (CNVs) and aneuploidies. The accurate determination of CNV endpoints and absolute copy numbers is essential for understanding the causes of cancer and for its treatment. Efforts to do so in the well-characterised NCI-60 cell line panel have so far been limited to relatively low-resolution array-based comparative genomic hybridisation or whole exome sequencing methods. We have used more accurate whole genome sequencing alignment read depths to produce a higher-resolution map of CNVs in the NCI-60 cell lines.

We find the CNV landscape in cancer-derived cell lines to be dominated by partial losses, primarily affecting tumour suppressors, alongside less frequent gains affecting oncogenes. The high resolution of our data allows us to accurately determine the locations of CNV 'hotspots': regions that are gained or lost in many cell lines, and which could therefore possibly be under the influence of selective pressure in cancer. We have also developed a statistical measure of how likely each gene's copy number is in each cell line. We determined the gained oncogenes and lost tumour suppressors occurring in CNV hotspots more than expected by chance and, by comparison to genes which are known to be drivers of oncogenesis, derived a list of candidate novel driver genes.

These candidate novel driver genes are enriched for involvement in cancer-related processes such as increased cell growth and proliferation, angiogenesis, upregulated gene transcription, control of the cell cycle and apoptosis. Our results are consistent with the hypothesis that tumour cells must escape the constraints of various cellular mechanisms that ordinarily protect against cancer before they can gain the various hallmarks of cancer. The more accurate map of CNVs in cancer cell lines that we have created will form the basis of ongoing work to understand gene dosage changes in cancer.

2.2 Introduction

Somatic mutations that affect genes involved in cell growth and survival are frequently implicated in the development of cancer (Hanahan and Weinberg 2000, 2011; Li et al. 2011; Budczies et al. 2016). At the smallest scales, these mutations include single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) (International HapMap 2003). Larger mutations affecting entire genes include amplifications and deletions known as copy number variations (CNVs), inversions and gene fusion events (Yang et al. 2016). Often observed at the chromosome scale are the gains and losses of chromosome arms or entire chromosomes called aneuploidies (Torres et al. 2008) and wide-scale rearrangements collectively known as chromothripsis, consisting of multiple inversions, amplifications and deletions caused by a single disruptive event (Cortes-Ciriano et al. 2020).

At the time of the completion of the first drafts of the human reference genome (Lander et al. 2001; Venter et al. 2001) it was thought that SNPs were the main causes of cancer development (Sachidanandam et al. 2001; International HapMap 2003), but it was soon recognised that CNVs and aneuploidies, leading to altered dosage of entire genes and the resulting stoichiometric disruption of gene interaction networks, are key to understanding the genesis and progression of cancer (Iafrate et al. 2004; Henrichsen et al. 2009). Gene dosage changes disrupt and vary protein interactions and so affect biological processes involved in cancer such as cell cycle progression, DNA damage detection and repair, apoptosis and cell proliferation (Conrad et al. 2006; Henrichsen et al. 2009; Kalluri and Weinberg 2009; Chaffer and Weinberg 2011). Common mechanisms of this disruption include the loss of tumour suppressor genes and the gain of genes which, when upregulated, lead to cancer and are thus known as oncogenes (Santarius et al. 2010; Budczies et al. 2016; Zhao and Zhao 2016).

However, not all dosage changes lead to cancer development and so it is important to distinguish between ‘driver’ CNVs that are causal in oncogenesis and ‘passenger’ CNVs that become fixed in a tumour clonal population due to the coincidental presence of a driver mutation (Greenman et al. 2007). A common method for differentiating drivers from passengers is to analyse mutations in a wide range of tumours or tumour-derived cell lines from different tissues under the hypothesis that genomic regions containing drivers will experience a higher frequency of CNVs than regions predominantly containing passenger

CNVs (Beroukhim et al. 2010; Bignell et al. 2010; Pleasance et al. 2010; Martincorena et al. 2017; Rheinbay et al. 2020).

The extensive heterogeneity of mutations that occur in tumours derived from different tissues means that moving beyond a tumour- or tissue-specific characterisation of cancer processes to a more general understanding requires a representative range of cancer genomes. The NCI-60 panel of cell lines (Shoemaker 2006), derived from human tumour samples from nine tissue types, has been extensively analysed and so is a resource which is well-suited to integrative analysis due to its detailed characterisation by many previous studies (Lorenzi et al. 2009; Li et al. 2011; Varma et al. 2014).

Previous panel-wide NCI-60 CNV studies have used array-based comparative genomic hybridisation (aCGH) to determine CNVs by the comparison of the relative intensities of fluorophores bound to specific regions of DNA in matched tumour/normal samples (Lorenzi et al. 2009; Beroukhim et al. 2010; Bignell et al. 2010; Varma et al. 2014). aCGH-based methods require a matched normal sample and can only achieve a CNV endpoint resolution of about 10^5 nucleotides (Lai et al. 2005b; Yoon et al. 2009). The endpoints and copy numbers of CNVs can be determined at much higher resolution from the relative read depths of short reads from whole genome sequencing of a single tumour sample aligned to a reference genome (Chiang et al. 2009; Duan et al. 2013; Bishara et al. 2015; Cortes-Ciriano et al. 2020). Previous sequencing-based analyses of NCI-60 CNVs have used only whole *exome* sequencing reads (Reinhold et al. 2014) and so are unable to determine the copy number of regions outside of the exons of protein-coding genes.

We reanalysed previously published NCI-60 whole genome sequencing datasets (Turner et al. 2017) and used the varying read depths of the sequencing reads when aligned to the reference genome (Miller et al. 2011) to create a CNV map of the NCI-60 cell line genomes with both higher endpoint resolution and more accurate copy numbers than previously calculated from aCGH assays (Varma et al. 2014). We show that this method finds a variety of additional and smaller CNVs than can be detected with aCGH. Even with relatively low-coverage sequencing of between 0.4x and 3.2x, we achieve an order of magnitude better resolution than aCGH, leading to a higher-confidence map of CNVs in the NCI-60 cell lines. We also find that most NCI-60 cell lines are affected by aneuploidy, consistent with earlier studies (Torres et al. 2008).

We show that CNVs preferentially occur in regions of high gene density with frequent losses of tumour suppressor genes and gains of oncogenes. We identify putative novel oncogenesis driver genes by focussing on cancer-related genes that are located in CNV 'hotspots' and which are affected by CNVs more than expected by chance, and we show that these driver genes are enriched for cancer-related processes and pathways associated with the 'hallmarks' of cancer (Hanahan and Weinberg 2000, 2011).

2.3 Methods

2.3.1 Construction of a high-resolution NCI-60 CNV map

2.3.1.1 Cell line read mapping

Previously published Illumina sequencing data derived from the NCI-60 cell lines, deposited under project accession PRJNA338012 (Turner et al. 2017), were obtained from the Sequence Read Archive (SRA) (Sayers et al. 2020), using SRA Toolkit v2.8.2 (Sayers et al. 2020). Five of these cell lines, MDA-MB-435, MDA-N, MCI-ADR-RES, SNB19 and U251, have been reported as being misidentified (Garraway et al. 2005; Lorenzi et al. 2009), and so were excluded from this study. Each dataset was aligned to the reference human genome (GRCh38.p12) with BWA v0.7.15 (Li and Durbin 2010) and the alignments were checked with FastQC v0.11.7 (Andrews 2018), with all sequence quality checks passed.

2.3.1.2 Mapability and GC content annotation

All possible reads of the same length as the NCI-60 reads (50 bp for MCF7 and 100 bp for the rest) were generated from the reference genome and aligned back to the reference genome with BWA. The mapability annotation was calculated by counting the proportion of positions in each 100 bp window where a uniquely mapped read starts. The GC content annotation was also calculated in 100 bp windows across the genome by counting the number of uniquely mapped guanine or cytosine bases in each 100 bp window.

2.3.1.3 CNV detection with readDepth

CNVs were calculated for each cell line with R package readDepth v0.9.8.5 with default parameters (Miller et al. 2011). The first step of the readDepth algorithm calculated raw read depths which were then scaled by the mean mapability score for each bin to obtain the mapability-corrected read depths, discarding bins with a mean mapability score below 75%. The readDepth algorithm then scaled the mapability-corrected read depths by the difference between the mean LOESS-fitted number of reads for bins at each level of GC content and an adjusted genome-wide median, iteratively calculated such that the adjustments have no effect on the actual genome-wide median (see section 2.4.1 for an illustration of the method).

The mapability- and GC-corrected read depths for contiguous bins were then coalesced by readDepth into segments of similar read depth using the circular binary segmentation (CBS) method (Olshen et al. 2004). The segments' read depths were converted into absolute copy numbers relative to the genome-wide median read depth to obtain a list of putative CNVs for the cell line.

We observed occasional false positive low mapability CNVs spanning centromeres and extending from telomeres and so we developed a custom post-readDepth step (discussed further in section 2.4.1) which removed CNVs with a mean mapability score below 75% and split the remaining CNVs at unmappable bins to obtain a conservative CNV list. Finally, the loss and gain thresholds, also converted into absolute copy numbers relative to the genome-wide median read depth, were applied in a custom final step to classify the conservative CNVs as losses, unaffected regions or gains.

2.3.2 Characterisation of CNVs

Human protein-coding gene locations were downloaded from Ensembl on 29/9/20 (release 100) (Yates et al. 2020). The overlaps between CNV locations and gene locations were calculated using R package GenomicRanges v1.40 (Lawrence et al. 2013). Aneuploidies were defined as the chromosome arms with more than 75% of their mappable genome affected by CNVs in the same direction (i.e., at least 75% lost or 75% gained). CNV gene densities were calculated as the number of protein-coding genes in each CNV divided by the CNV length. Centromere locations were taken from the R package GWASTools v1.36 (Gogarten et al. 2012), telomeres from the UCSC genome annotation downloads on 13/6/20 (Haeussler et al. 2019) and fragile sites from the HumCFS database on 24/6/20 (Kumar et al. 2019).

2.3.3 Well-known cancer-related genes

A dataset of protein-coding genes that are known to be involved in cancer-related processes as tumour suppressors and/or oncogenes was downloaded from the COSMIC Cancer Gene Census (CGC) at the Wellcome Sanger Institute on 7/5/20 (release 91) (Sondka et al. 2018). Analyses were performed using 'tier 1' CGC genes only (genes with strong evidence of functional involvement in cancer-related processes and with concordant mutations in cancer samples). Heatmaps of the copy numbers of CGC genes across the NCI-60 cell line panel were generated using the heatmap.2 function of R package gplots v3.0.3 (Warnes et al. 2020).

2.3.4 Calculation of metrics with a range of 'sliding window' sizes

Various metrics were calculated for adjacent regions of the genome for window sizes 10 kb, 100 kb, 1 Mb, 2 Mb, 5 Mb, 10 Mb, 20 Mb and 50 Mb. CNV frequencies were calculated as the number of cell lines in which a gain or loss started in each contiguous window. Gene densities were calculated as the number of genes starting in each contiguous window divided by the window length. Median gene lengths were calculated as the median length of the genes which started in each contiguous window.

2.3.5 Determining empirical significance of copy numbers by permutation testing

The significance of genes' CNVs across the NCI-60 cell line panel were determined by permutation testing. The CNVs in each cell line's chromosomes were randomly shuffled 1000 times and the resulting copy numbers of the GCG genes were recorded, generating a 1000-value distribution of copy numbers for each gene in each cell line. An empirical two-sided p-value was then derived for the actual copy number of each gene in each cell line from the location of the actual copy number in its distribution of permutated copy numbers. The false discovery rate was controlled with the Benjamini & Hochberg method (Benjamini and Hochberg 1995).

2.3.6 Enrichment analyses of cancer-related genes significantly affected by CNVs

Gene Ontology (GO) biological process term and Reactome pathway enrichment analyses were performed using the PANTHER tool at http://geneontology.org with default parameters (Fishers' exact test and calculated false discovery rate) on 30/12/20 (Mi et al. 2020).

2.3.7 Statistical analyses

All statistical analyses were performed in R. Correlations were calculated using the Pearson product-moment correlation test. Distributions were compared with the Kolmogorov-Smirnov test. Contingency table tests and goodness of fit tests were performed with the X^2 test.

2.4 Results

2.4.1 A high-resolution NCI-60 CNV map reveals widespread heterogeneity

We downloaded previously published (Turner et al. 2017) whole genome sequencing reads for 55 of the cell lines (Supplementary Table 6.1) in the NCI-60 panel (Shoemaker 2006) from NCBI's Sequence Read Archive (SRA) (Kodama et al. 2012) and aligned them to the GRCh38 human reference genome. These alignments were processed with the readDepth R package (Miller et al. 2011) to determine regions with statistically significant CNVs at a much higher resolution than has been previously calculated from aCGH assays (Varma et al. 2014).

The readDepth algorithm uses a binning procedure on the aligned reads to determine the relative read depth across the genome because the coverage of the whole genome sequencing data used is too low for base pair-level resolution. A model distribution of read depth frequencies containing haploid, diploid and triploid peaks is generated from the actual distribution of read depths and readDepth then iteratively determines the smallest bin size that leads to optimal separation of the model peaks while controlling the false discovery rate, specified as FDR = 0.01 in this project (Figure 2.1A). Within the 55 datasets, the chosen bin size decreases with increasing sequencing coverage, approaching 10.3 Kb for cell lines with a coverage of 3x (median bin size 20.6 Kb, maximum 41.2 Kb, Figure 2.1B).

The read depth thresholds for loss and gain relative to the median depth across the genome are then calculated by readDepth as the read depths that are midway between the haploid/diploid and diploid/triploid peaks respectively (Figure 2.1A). Once the optimal bin size and the loss and gain thresholds have been determined the raw read depths are calculated by readDepth by counting the uniquely mapped reads starting in each bin.

The readDepth algorithm then scales these raw read depths by the mean mapability score for each bin to obtain the mapability-corrected read depths, discarding those for bins with a mean mapability score below 75% to avoid distortions caused by repetitive regions to which short sequencing reads cannot be uniquely aligned (Chiang et al. 2009). The bias caused by GC content in the number of reads generated by the sequencing platform (Bentley et al. 2008) is estimated in readDepth by LOESS local regression and the read depths are then scaled by the difference between the mean LOESS-fitted number of reads for bins at each level of GC content and an adjusted genome-wide median, iteratively calculated such that the adjustments have no effect on the actual genome-wide median (Figure 2.1C/D).

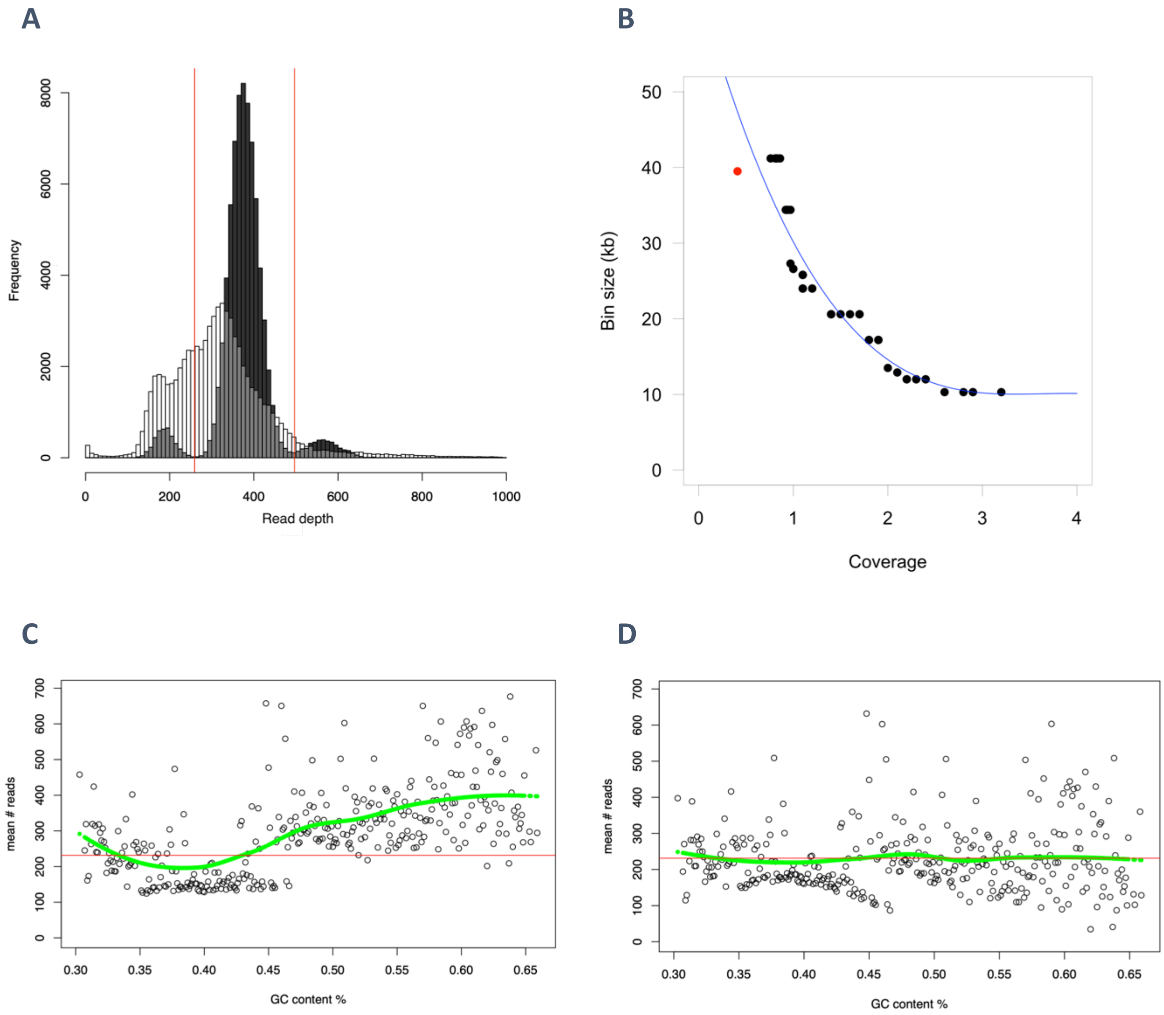


Figure 2.1 - readDepth models, bin sizes and GC content correction

(**A**) The distribution of actual read depth frequencies (outlined white bars) and model read depth frequencies (black bars) for the MCF7 cell line with the optimal bin size of 39.5 Kb. The haploid, diploid and triploid peaks in the model distribution are negative binomial distributions based on the mean, variance and number of reads in the actual distribution. The diploid peak in the model distribution is at a greater read depth (378) than the median actual read depth (311) because the model distribution is calculated on the expectation that the cell line would be entirely diploid in the absence of CNVs, as it is female-derived, but MCF7 has more losses than gains. The red vertical lines are the loss and gain thresholds between the model peaks. (**B**) The bin sizes (y axis) for the NCI-60 cell lines of various levels of sequencing coverage between 0.4x and 3.2x (x axis) as calculated by the readDepth package. The fitted curve is a third order polynomial that asymptotically approaches a bin size of approximately 10 Kb as coverage increases past 3x. MCF7 is the cell line indicated by a red dot. (**C**) The mean mapability-corrected read depths in MCF7 for bins with GC content in increments of 0.1%, with a LOESS regression line in green indicating the GC content-derived bias in the number of reads generated by the sequencing platform. (**D**) The LOESS-corrected mean read depths in MCF7 with the green regression line now showing that most of the bias from GC content has been removed. The red horizontal lines represent an adjusted genome-wide median read depth iteratively calculated such that the LOESS correction has no effect on the actual genome-wide median read depth.

The mapability- and GC-corrected read depths for contiguous bins are then coalesced by readDepth into segments of similar read depth using the circular binary segmentation (CBS) method (Olshen et al. 2004). The segments' read depths are converted by readDepth into absolute copy numbers relative to the genome-wide median read depth to obtain a list of putative CNVs for the cell line.

Close inspection of the putative CNVs in each cell line revealed that the CBS method used by readDepth introduced occasional false positive CNVs, generally spanning centromeres or extending from telomeres, caused by segment boundaries adjacent to regions of low mapability. We developed a custom post-readDepth processing step that removed CNVs with a mean mapability score below the aforementioned 75% mapability threshold, so removing all the false positive CNVs, and split the remaining CNVs at unmappable bins to obtain a conservative CNV list.

Finally, the loss and gain thresholds, also converted into absolute copy numbers relative to the genome-wide median read depth, were applied in a custom final step to classify the conservative CNVs as losses, unaffected regions or gains. CNVs with an absolute copy number of less than half the loss threshold for diploid chromosomes or less than the loss threshold for haploid chromosomes were designated as complete losses. CNVs in diploid chromosomes with an absolute copy number between the loss threshold and half of the loss threshold are partial losses, CNVs with an absolute copy number between the loss and gain thresholds are unaffected regions and CNVs with an absolute copy number above the gain threshold are gains.

When the CNVs for the entire NCI-60 panel are viewed together it is immediately apparent that there is widespread heterogeneity, with gains and losses varying from single bins (the smallest distinguishable region in each cell line) to whole chromosome arm aneuploidies (Figure 2.2). There are distinct differences between the individual cell lines with, for example, colon cancer-derived cell line HCT15 the least affected and a breast cancer cell line, T47D, the most affected by CNVs (Figure 2.2). In addition, the chromosomes have widely varying numbers of CNVs with, for example, chromosome 4 being mostly affected by losses and chromosome 7 mostly by gains (Figure 2.2). The sex chromosomes are affected differently than autosomes, with female X chromosomes experiencing frequent partial loss of the majority of the chromosome, unlike the male X chromosomes which have far fewer CNVs (Figure 2.2).

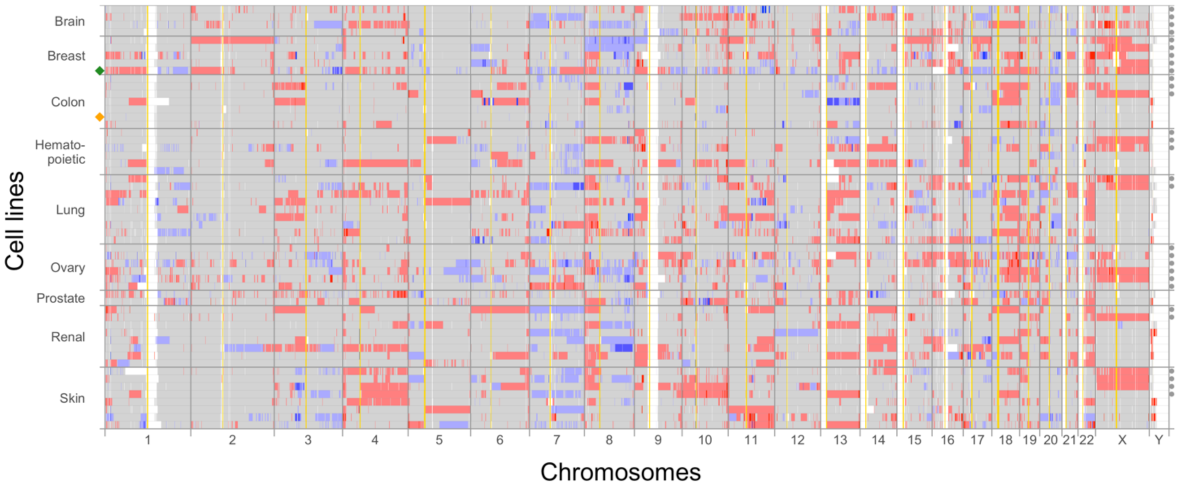


Figure 2.2 - High resolution NCI-60 panel-wide CNV map

NCI-60 panel-wide CNV map with rows that represent individual cell lines and columns for each chromosome. The cell line rows are grouped by tissue. Dark and light red areas represent complete and partial losses of the genome respectively, blue areas are gains with darker blues representing greater fold changes, grey areas are unchanged and white areas are where the cell lines' genomes are unmappable due to repetitive sequences or where regions with a low mean mapability have been removed. The vertical yellow lines represent the centromeres. The green marker on the left edge indicates cell line T47D and the orange marker indicates cell line HCT15. The cell lines derived from female patient tumour samples are indicated by a grey marker on the right edge.

Zooming in to show chromosome 17 at a higher resolution (Figure 2.3A) illustrates how much more detail is available when determining CNVs from whole genome sequencing data rather than from array-based comparative genomic hybridization (aCGH) as in previous studies such as (Varma et al. 2014). At this level of detail, the CNV heterogeneity is even more apparent with the short arm of chromosome 17 frequently experiencing partial loss (light red regions in the left quarter of Figure 2.3A) next to the unmappable centromeric region (white). There are also regions of complete loss detected (dark red) as well as widely varying levels of gains (shades of blue). Intriguing regions of possible CNV 'hotspots' are also visible, such as the small region just to the left of the centromere that is gained in most of the cell lines.

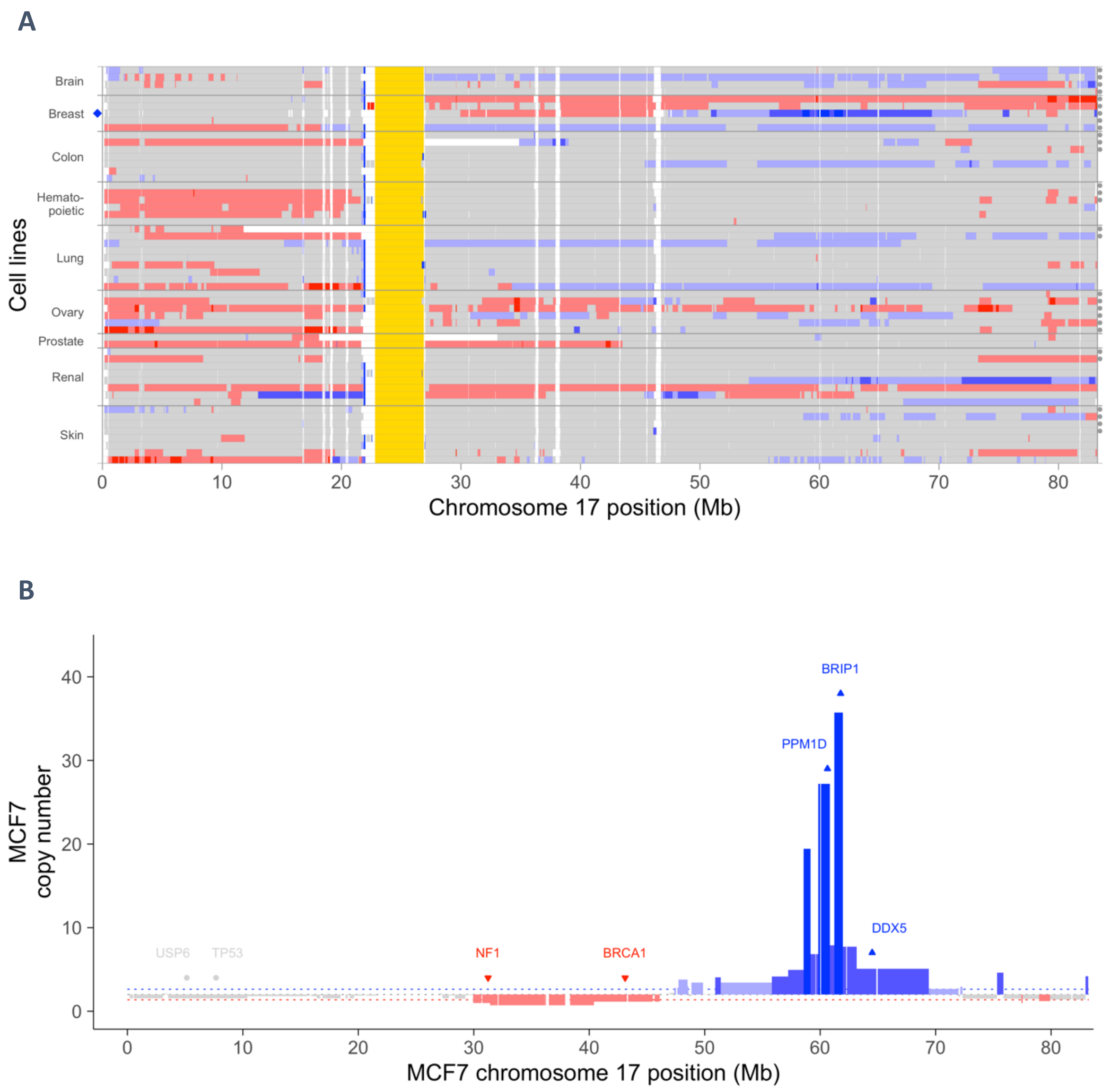


Figure 2.3 - Chromosome 17 CNVs

(**A**) Same as Figure 2.2 but now zoomed to show just chromosome 17 to illustrate the level of detail available. The blue marker on the left edge indicates cell line MCF7. (**B**) CNVs for chromosome 17 of the breast cancer-derived cell line MCF7. The x axis is the genomic position within chromosome 17 and the y axis is the absolute copy number. CNVs are drawn from the expected diploid read depth (grey dashed line at y = 2) to the measured read depth. Grey regions are those between the gain and loss thresholds and so statistically indistinguishable from diploid, red regions are losses and blue regions are gains (with >4 and >8-fold gains as progressively darker blues). The loss threshold is shown as a red dashed line (at approximately y = 1.3) and the gain threshold as a blue dashed line (at approximately y = 2.6). Selected unaltered genes are labelled grey circles, lost tumour suppressors are labelled red downwards-pointing triangles and gained oncogenes are labelled blue upwards-pointing triangles.

When looking at the detailed CNVs for a single chromosome in a cell line, such as chromosome 17 in breast cancer-derived cell line MCF7 (Figure 2.3B), the scale of copy number variation in cell line genomes is evident. While the short arm of this chromosome (between 0 and approximately 22 Mb) is unaffected by CNVs the longer arm has a wide range of variation, with a region of partial loss between 30 and 47 Mb followed by a region of gains that vary from a single extra copy at 70 Mb to a peak of 33 extra copies at 62 Mb.

A variety of cancer-related genes are affected by CNVs in MCF7 chromosome 17, including tumour suppressors *NF1* and *BRCA1* which are partially lost and oncogenes *PPM1D* and *DDX5* which have 25 and three extra copies respectively (Figure 2.3B). Oncogene *USP6* and tumour suppressor *TP53* are unaltered in this cell line while the gene with the most gains (33 extra copies) on chromosome 17 of MCF7 is tumour suppressor *BRIP1*, an interaction partner of *BRCA1* (Figure 2.3B). *BRCA1* is lost in three of the breast cancer-derived cell lines and two of the ovarian cancer cell lines (Figure 2.3A), consistent with its known role in susceptibility to these cancers (Futreal et al. 1994).

2.4.2 Whole genome sequencing read depth analysis finds finer-detailed CNVs than aCGH

Previous studies of CNVs in the NCI-60 cell lines used aCGH microarrays and we wanted to compare our results to these. We extracted two measures of genomic instability, the proportion of the genome gained and lost and the number of gains and losses, from Table 1 of (Varma et al. 2014) and compared these metrics with our results (Figure 2.4).

When comparing the total proportion of the mappable genome affected by CNVs from our results to the total proportion of the genome gained and lost as calculated by (Varma et al. 2014) we find that the results are well-correlated (Pearson product-moment correlation, r (53) = 0.792, p = 6.2 x 10^{-13}, Figure 2.4C). We find however that we detect somewhat less of the genomes as being gained (r (53) = 0.778, p = 2.6 x 10^{-12}, Figure 2.4A) but more as being lost (r (53) = 0.559, p = 9.24 x 10^{-6}, Figure 2.4B).

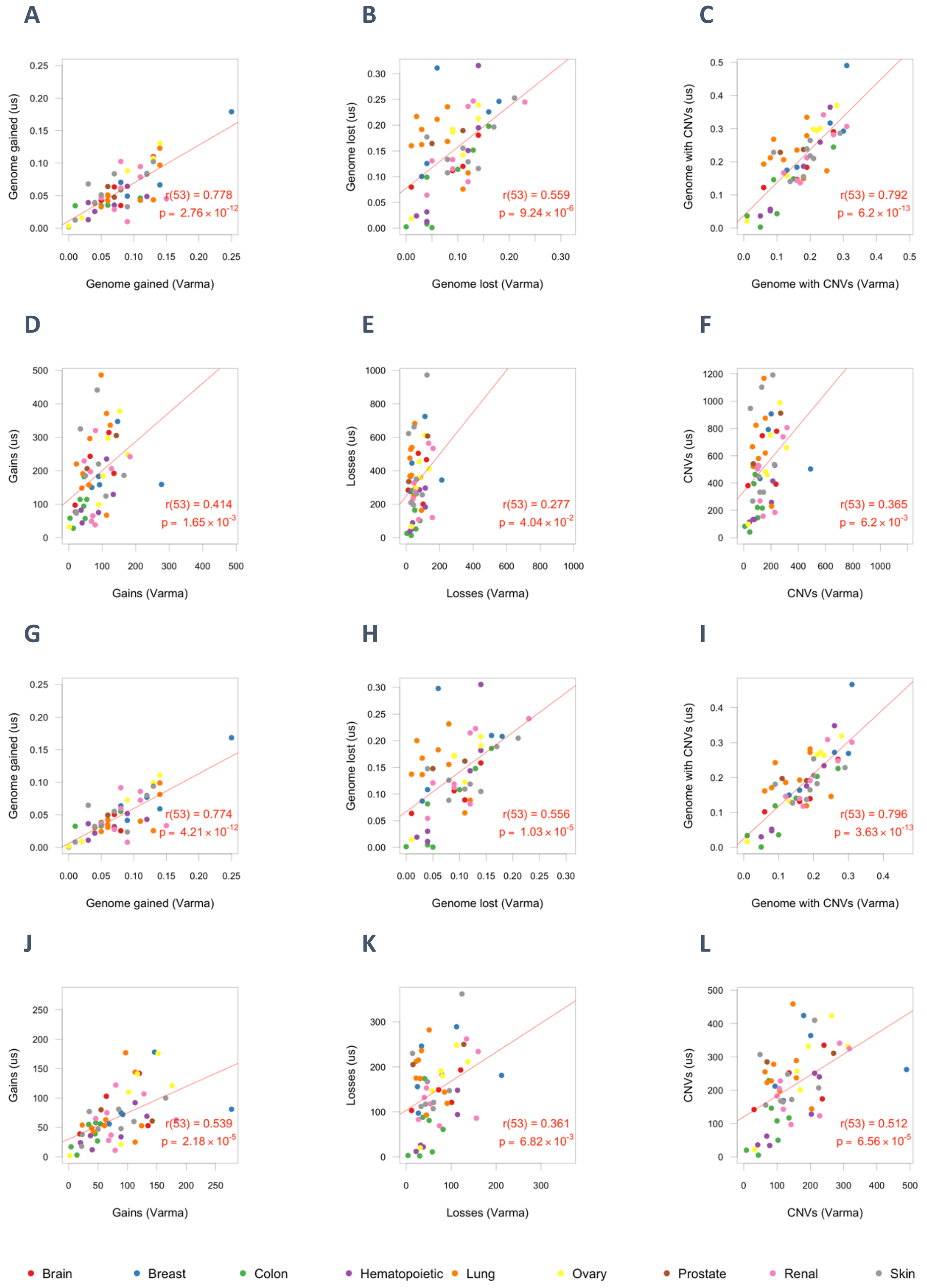


Figure 2.4 - Comparison of whole genome sequencing-based CNVs to aCGH-based CNVs

Comparisons between the results of our method and the results in (Varma et al. 2014). In each graph the x axis is the value from (Varma et al. 2014), the y axis is the result from our method and each point is a cell line, coloured by tissue of origin (see legend). The red line is the slope as calculated by Pearson's product moment correlation, the value and p-value of which are embedded in each plot. (**A-C**) The proportions of the genome gained, lost and in total. (**D-F**) The numbers of gains, losses and all CNVs. (**G-I**) The proportions of the genome gained, lost and in total for the largest 50% of CNVs detected by our method. (**J-L**) The numbers of gains, losses and all CNVs for the largest 50% of CNVs detected by our method.

Looking at the total number of gains and losses instead however (Figure 2.4F), we detect both more gains than the previous study (r (53) = 0.414, p = 1.65 x 10^{-3}, Figure 2.4D) and more losses (r (53) = 0.277, p = 4.04 x 10^{-2}, Figure 2.4E). We hypothesised that with our high-resolution approach we might be detecting smaller CNVs than is possible with aCGH and so we repeated the comparison after removing progressively more of the smaller CNVs from our results up to the smallest 50% of CNVs.

Removing the smallest half of our CNVs resulted in improved correlation with the data from (Varma et al. 2014) (Figure 2.4G-L) as did removing the smallest 10%, 20%, 30% and 40% smallest CNVs, albeit with intermediate effects (data not shown), thus supporting the hypothesis that our method can detect smaller CNVs than aCGH-based methods.

2.4.3 Cell lines are affected more by losses than by gains

To quantify the degree to which cell lines are affected by CNVs we plotted the proportions of the mappable genome affected by CNVs and the numbers of CNVs for the cell line genomes grouped by tissue of origin (Figure 2.5). While there is considerable variation of losses and gains within and between the tissue groups, both for the proportion of the mappable genome affected (Figure 2.5A/B) and the number of CNVs (Figure 2.5C/D), it is clear that on both measures the cell lines are more affected by losses (Figure 2.5B/D) than by gains (Figure 2.5A/C).

To further investigate the differences seen between the autosomes and sex chromosomes in the NCI-60 panel-wide CNV map (Figure 2.2) we separated the CNVs by chromosome type and compared for each the proportions of the mappable genome affected by gains and losses (Figure 2.6A). The overall pattern of more losses than gains is seen for each type of chromosome except for male X chromosomes, which are much less affected by either gains or losses (Figure 2.6A). While the maximum proportion of the mappable genome on autosomes affected by losses is 0.33 (median 0.14), the female X chromosomes have much higher proportions of the mappable genome lost, up to partial loss of the entire chromosome in some cases (Figure 2.6A). Male X chromosomes on the other hand are almost entirely unaffected by gain or loss, though the male Y chromosomes experience a similar range of losses (maximum 0.85) to the female X chromosomes, albeit with a much lower median proportion of 0.04 (Figure 2.6A).

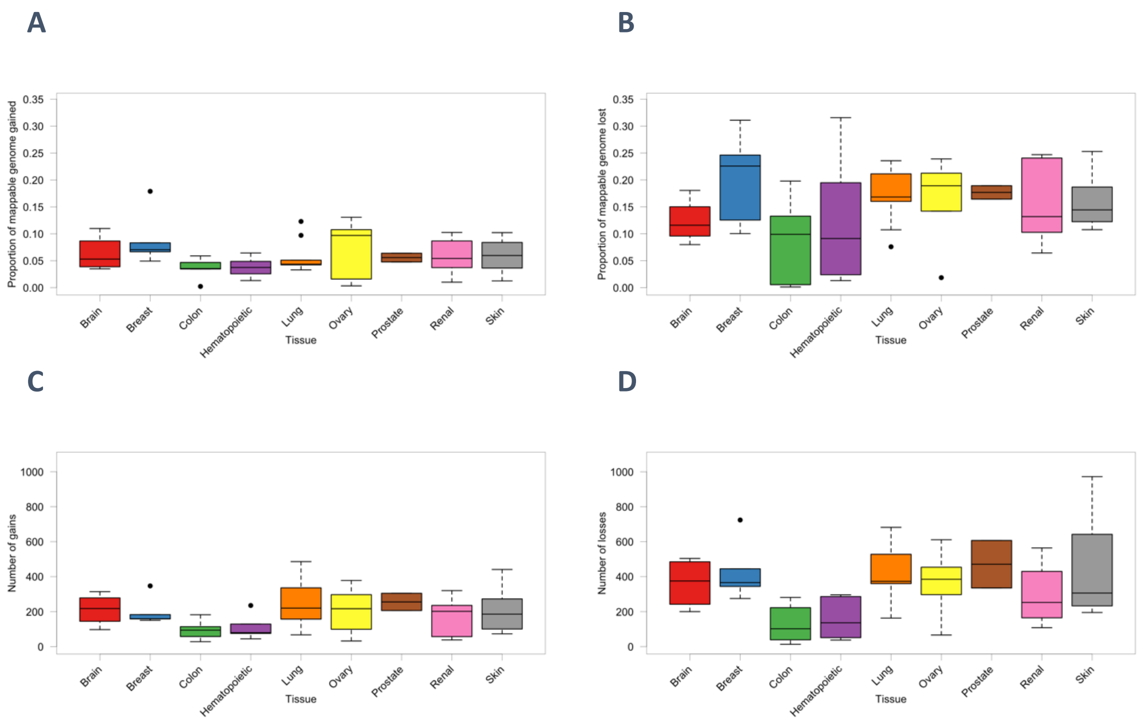


Figure 2.5 - CNV frequencies and the proportion of the mappable genome affected

The proportion of mappable genome (**A**) gained and (**B**) lost and the number of (**C**) gains and (**D**) losses for cell lines grouped by tissue.

When the proportions of the different types of chromosomes that are affected by CNVs are separated by tissue of origin (Figure 2.6B-E) we see a broadly similar pattern to the panel-wide results, with the caveat that not all the cell lines have both male and female-derived variants. On the autosomes, for each tissue of origin, the median proportion of the mappable genome lost is higher than that gained, and for each tissue apart from brain, ovary, prostate and skin the losses have much higher variance than the gains (Figure 2.6B). A similar pattern is exhibited by the female X chromosomes though the gains are practically non-existent with a maximum gain proportion of 0.017 for ovarian gains (Figure 2.6C).

Male X chromosomes on the other hand have much lower proportions of their mappable genomes affected by CNVs with a maximum gain proportion of just 0.08 for renal-derived cell lines being very much an outlier (Figure 2.6D). The patterns of more losses than gains still hold for male X chromosomes however, except for lung cancer and renal-derived cell lines where the pattern is reversed (Figure 2.6D). Interestingly, male Y chromosomes experience a similar range of CNVs to the female X chromosomes with losses again affecting a much higher proportion of the mappable genome than gains (Figure 2.6E).

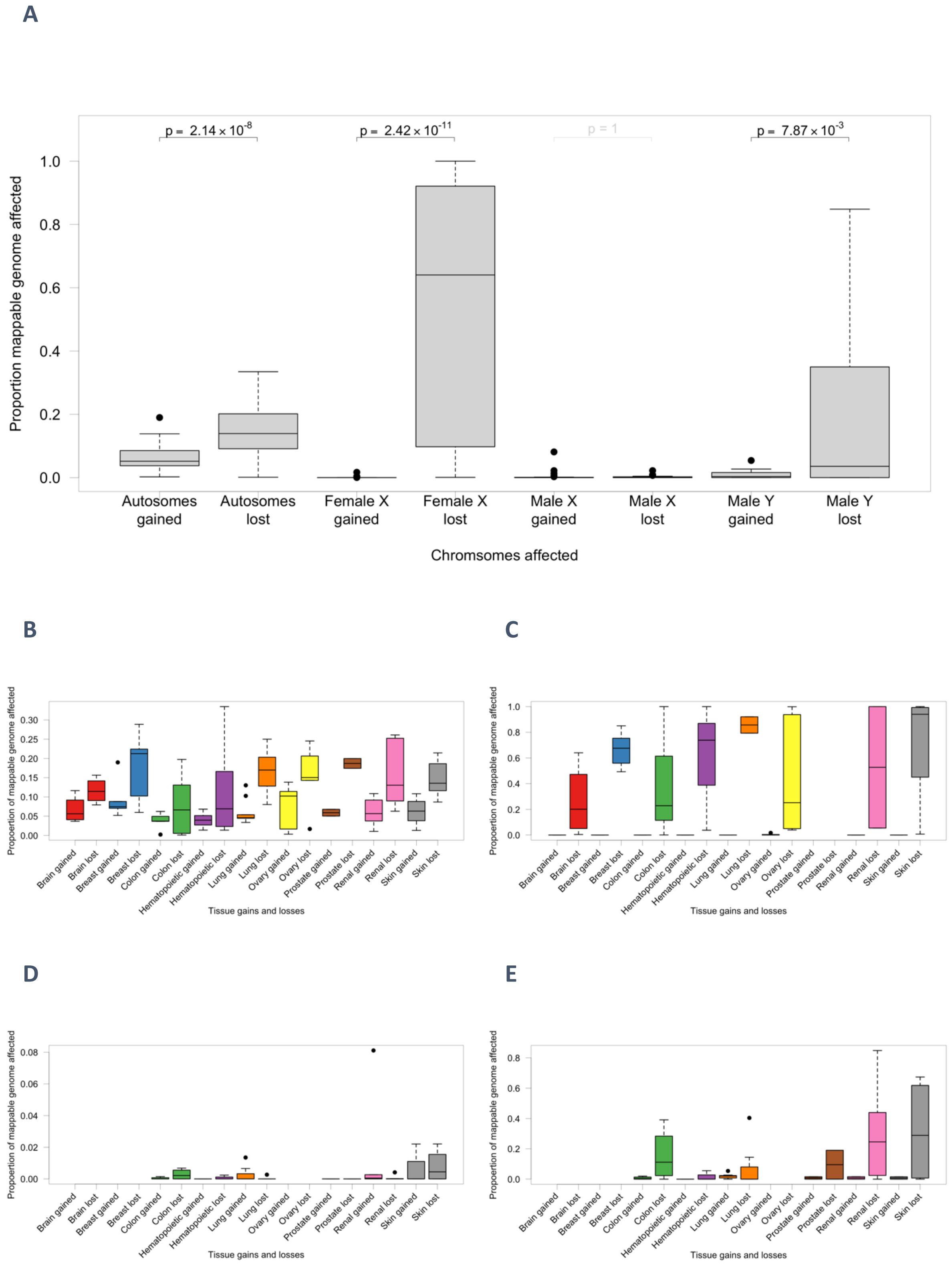


Figure 2.6 - Sex chromosomes are affected differently by CNVs to autosomes

(**A**) The proportions of the mappable genomes gained and lost for, from left to right, autosomes, female X chromosomes, male X chromosomes and male Y chromosomes. The p-values are from two-sided Kolmogorov-Smirnov tests and are black when significant at $\alpha = 0.05$ and light grey otherwise. When grouped by tissue of origin, the proportions of the mappable genomes gained and lost for (**B**) autosomes, (**C**) female X chromosomes, (**D**) male X chromosomes and (**E**) male Y chromosomes.

Cancer cells are known to be affected by widespread aneuploidies that are lethal to normal cells (Torres et al. 2008; Sheltzer and Amon 2011) and so we wanted to see how prevalent aneuploidy is in the NCI-60 panel. We defined aneuploid as chromosome arms with more than 75% of their mappable genome consistently affected by CNVs (i.e., at least 75% lost or 75% gained). We find that across different chromosomes there is considerable variation in the number of cell lines with aneuploidies and that the majority of aneuploidies are losses rather than gains (Figure 2.7). The short arm of chromosome 21 is the most affected with gains in 26 out of 55 cell lines and it is also the only acrocentric autosome with short arm aneuploidies (Figure 2.7). In 9 out of 13 submetacentric autosomes the short arm has more aneuploidies than the long arm (Figure 2.7). All but three of the cell lines (HCT15, HCC2998 and SR) had at least one aneuploidy (Supplementary Figure 6.2).

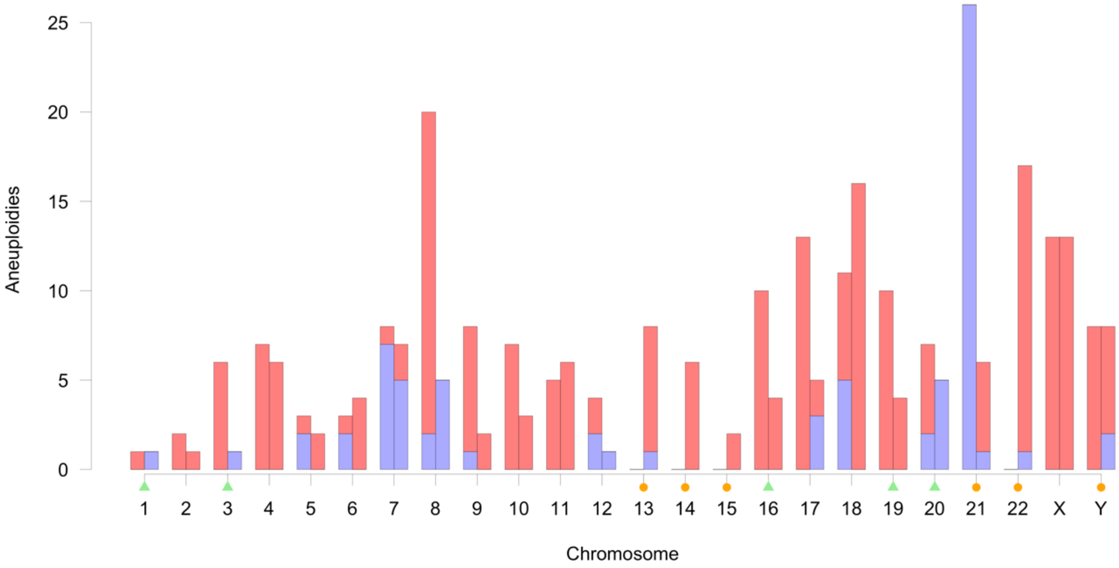


Figure 2.7 - Aneuploidy frequencies

The number of aneuploidies affecting each chromosome arm across the NCI-60 cell line panel. For each chromosome the left bar is the short arm (lowest genomic coordinates) and the right bar is the long arm (highest genomic coordinates). Gains are shown in blue and losses in red. Metacentric chromosomes are marked with green triangles and acrocentric chromosomes are marked with orange circles.

2.4.4 Losses are longer than gains and more gene-dense

The distributions of the lengths of CNVs across the entire NCI-60 panel are skewed with outlier lengths of up to 73 Mb but median lengths of just 0.35 Mb for gains and 0.49 Mb for losses. This pattern of losses being longer than gains holds when the CNV lengths are grouped by tissue of origin (Figure 2.8A) except for CNVs in brain, breast and ovarian cancer-derived cell lines, though the distributions of gains and losses for each tissue except brain are significantly different overall by the two-sided Kolmogorov-Smirnov test (Figure 2.8A).

The gene density distributions of CNVs, defined as the number of protein-coding genes that start in each CNV divided by the CNV length, are also skewed across the entire NCI-60 panel. Gains have a median gene density of 6.07 genes/Mb with outliers up to 291 genes/Mb and losses have a similar distribution with a median gene density of 6.94 genes/Mb and outliers up to 378 genes/Mb. CNVs in general however, whether gains or losses, are in regions of higher gene density than the rest of the genome, which has a background median gene density of 5.75 genes/Mb. The gene density distributions of CNVs when grouped by tissue of origin follow a similar pattern of higher gene density in CNVs than in the rest of the genome (Figure 2.8B) though not all tissues have similar gene densities for gains and losses. In cell lines derived from breast and ovarian cancer the losses have higher gene density than the gains but in lung cancer the opposite is the case with higher gene density gains than losses (Figure 2.8B).

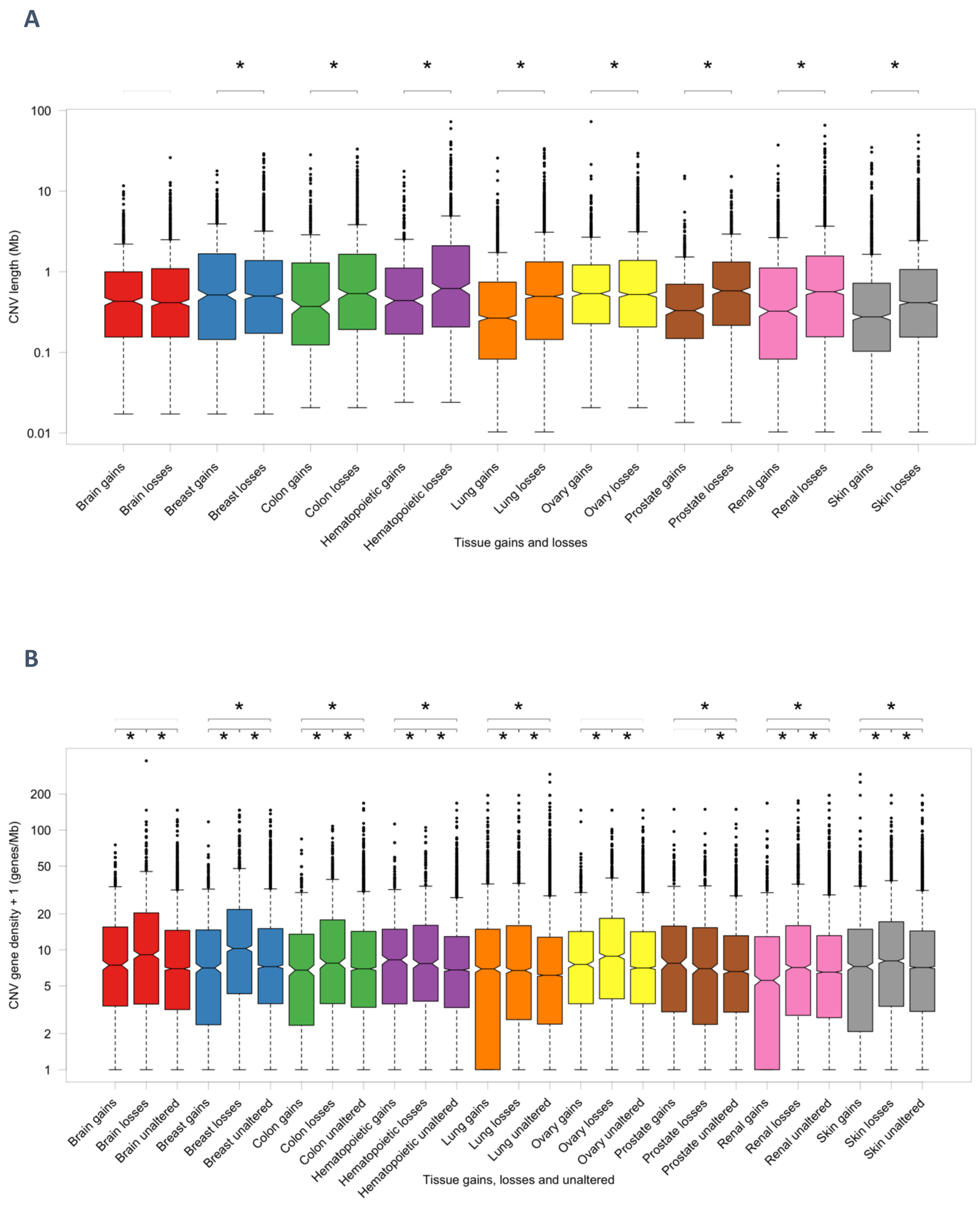


Figure 2.8 - CNV lengths and gene densities

(**A**) The length distributions of CNVs affecting each tissue of origin and, within each tissue, gains and losses from left to right. (**B**) The gene density distributions of CNVs affecting each tissue of origin (with a 'plus one prior' in order to use a log scale with zero densities) and, within each tissue, gains, losses and unaffected from left to right. The p-values from two-sided Kolmogorov-Smirnov tests between the pairs of distributions are shown as black stars and bars when significant at α = 0.05 and light grey bars otherwise.

2.4.5 CNV hotspots preferentially occur in regions of higher gene density

Having observed the frequently gained region of the short arm of chromosome 17 next to the centromere (Figure 2.3A), we hypothesised that there might be CNV 'hotspots' throughout the genome where multiple cell lines are similarly affected by gains or losses. We further hypothesised that these hotspots might preferentially occur in regions of high gene density because such CNVs would tend to lead to wider perturbations in gene interaction networks and so to a higher possibility of disrupting tumour-suppressive mechanisms.

We performed a sliding window analysis with various window sizes of the number of cell lines affected by CNVs and the gene density in each window. We calculated the number of cell lines affected by CNVs, the protein-coding gene density and the median protein-coding gene length across the genome for adjacent windows sized between 0.01 Mb and 50 Mb, comprising window resolutions chosen to range from slightly smaller than the smallest detectable CNVs (0.013 Mb in renal cell line 786-0) to slightly larger than the smallest chromosome (chromosome 21 with a length of 46.7 Mb).

When we plot these measures for chromosome 1 at 1 Mb resolution as an example, we can see that most of the regions that are gained or lost in many cell lines occur in fragile sites, 23 of which from the HumCFS dataset cover 140 Mb of chromosome 1 alone (Figure 2.9A), and coincide with regions of high gene density (Figure 2.9B) but low median gene length (Figure 2.9C), especially in the peri-centromeric and peri-telomeric regions. Similar patterns are observed for the other chromosomes and, apart from the first 1Mb window on chromosome 4 which is gained in ten cell lines and lost in 29, none of the gain or loss hotspots occurred in the same location (data not shown).

We defined as CNV hotspots the top 100 most frequently gained and top 100 most frequently lost regions at each resolution, and we observed that most of these hotspots occurred in the same place irrespective of window size (data not shown). The pattern of gain and loss hotspots across the chromosomes is similar to the pattern of gains and losses observed when viewing the entire CNV map (Figure 2.2) with, for example, losses predominant on chromosomes 4, 13, 18 and X.

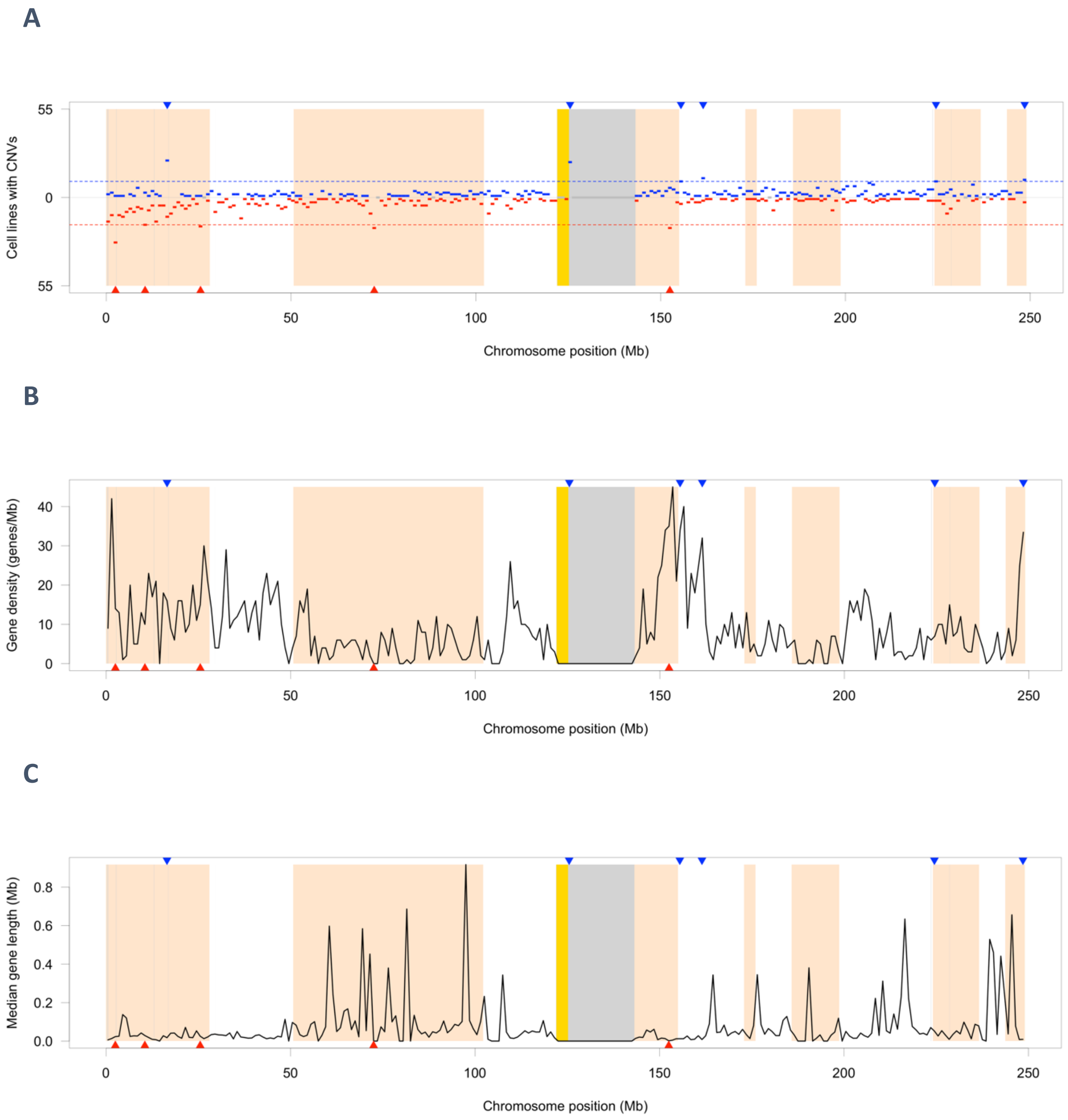


Figure 2.9 - Sliding window analyses reveal possible CNV hotspots

1 Mb sliding window analyses of chromosome 1 with a light orange background for fragile sites, yellow background for centromeres and grey background for unmappable poly N regions (hard masked as N in the reference genome). Windows in the top 100 most gained and lost are marked with blue and red triangles respectively. (**A**) The number of cell lines with CNVs in each 1 Mb window, with gains in blue above the zero line and losses in red below the zero line. The lowest number of gains in a CNV hotspot is shown as a blue dashed line at y = 10 and the lowest number of losses in a CNV hotspot is shown as a red dashed line at y = -17. (**B**) The gene density in each 1 Mb window, calculated as the number of protein-coding genes starting in each window. (**C**) The median protein-coding gene length in each 1 Mb window.

Regions at the highest resolution of 0.01 Mb can only contain a handful of genes at most and the lowest resolution of 50 Mb means that such regions are likely to have large numbers of cell lines with both gains and losses. Accordingly, we chose to focus on the intermediate 1 Mb sliding window resolution as the smallest window size and so highest resolution which clearly showed the prevailing pattern of CNV hotspots. At this scale there are 3,102 windows across the genome (including 24 partial windows, one at the end of each chromosome) and defining CNV hotspots as the top 100 most frequently gained or lost of these windows resulted in very conservative thresholds of 10 cell lines gained and 17 cell lines lost, allowing us to focus on the regions most affected by CNVs (Figure 2.9A, Supplementary Figure 6.1).

To further investigate our hypothesis that CNVs preferentially occur in regions of higher gene density we compared the distributions of the number of cell lines with gains or losses in regions of low, medium and high protein-coding gene density. We chose the lower (9 genes/Mb) and upper (26 genes/Mb) interquartile values of the distribution of gene densities at the 1 Mb resolution as the cut-offs between low and medium gene density regions and medium and high gene density regions respectively. The comparisons of the distributions of the number of cell lines affected by gains and losses in regions of low, medium and high gene density show clearly that high gene density regions have more cell lines with CNVs, both gains and losses, than low gene density regions (Figure 2.10).

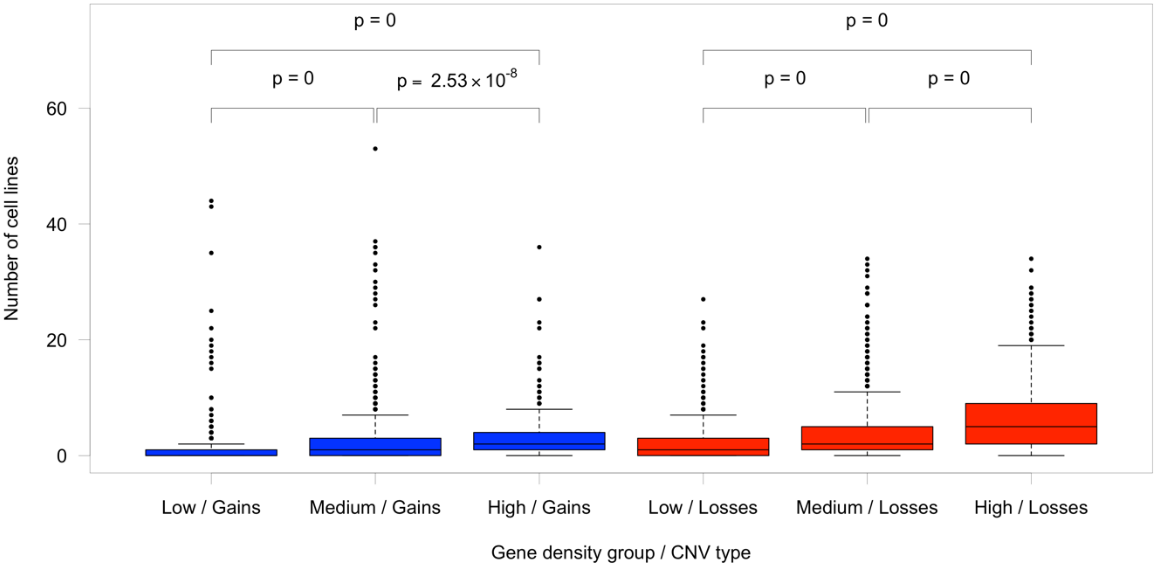


Figure 2.10 - Gene-dense regions are gained more and lost more than low density regions

The distributions of the number of cell lines affected by gains (blue) and losses (red) for regions of low, medium and high protein-coding gene density when calculated with a 1 Mb window size. The distributions are, from left to right, the number of cell lines with gains in low, medium and high gene density regions and the number of cell lines with losses in low, medium and high gene density regions. The p-values are from two-sided Kolmogorov-Smirnov tests and are black when significant at $\alpha = 0.05$ and light grey otherwise.

These distributions are significantly different as calculated by the Kolmogorov-Smirnov test for both gains and losses at sliding window resolutions between 0.1 Mb and 10 Mb (results shown only for 1 Mb, Figure 2.10). For example, at the 1 Mb window size, the median number of cell lines affected by gains are 0, 1 and 2 for low, medium and high gene densities respectively whereas the median number of cell lines affected by losses are 1, 2 and 5 for low, medium and high gene densities respectively, suggesting that losses affect regions of high protein-coding gene density more than gains do (Figure 2.10).

2.4.6 NCI-60 cell lines experience tumour suppressor loss and oncogene gain

The NCI-60 cell lines are all derived from cancers and so we hypothesised that the evolutionary driver for the prevalence of CNVs in gene-dense regions could be the presence of cancer-related genes in these regions. We downloaded expert-curated lists of known protein-coding oncogenes and tumour suppressor genes from the COSMIC Cancer Gene Census (CGC) at the Wellcome Sanger Institute (Sondka et al. 2018) and determined the copy numbers of these genes across the NCI-60 panel.

The distributions of the copy numbers of genes that are solely oncogenes (n = 209) and solely tumour suppressors (n = 197) across the NCI-60 panel are similarly skewed with a median oncogene copy number of 1.83 (maximum 32.8) and a median tumour suppressor copy number of 1.78 (maximum 35.7) (Figure 2.11A). The panel-wide oncogene and tumour suppressor distributions are however significantly different (two-sided Kolmogorov-Smirnov test, D = 0.06, $p < 2.2 \times 10^{-16}$) with the tumour suppressors' lower median copy number indicating that, on a panel-wide basis, they are lost more than oncogenes (Figure 2.11A). When the copy number distributions are examined per-tissue, tumour suppressors are lost significantly more than oncogenes for all tissues of origin except colon and ovary (Figure 2.11B).

We plotted the NCI-60 panel-wide oncogene and tumour suppressor CNVs as a clustered heatmap and calculated the largest column clusters of mainly losses and gains (Figure 2.12). There is a statistically significant relationship (X^2 (1, N = 143) = 14.7, $p = 1.24 \times 10^{-4}$) between the numbers of genes that are oncogenes or tumour suppressors (Table 2.1) in the two column clusters underlined in Figure 2.12, with tumour suppressors more likely to be in the column cluster dominated by losses (underlined in red in Figure 2.12) and oncogenes more likely to be in the column cluster dominated by gains (underlined in blue in Figure 2.12).

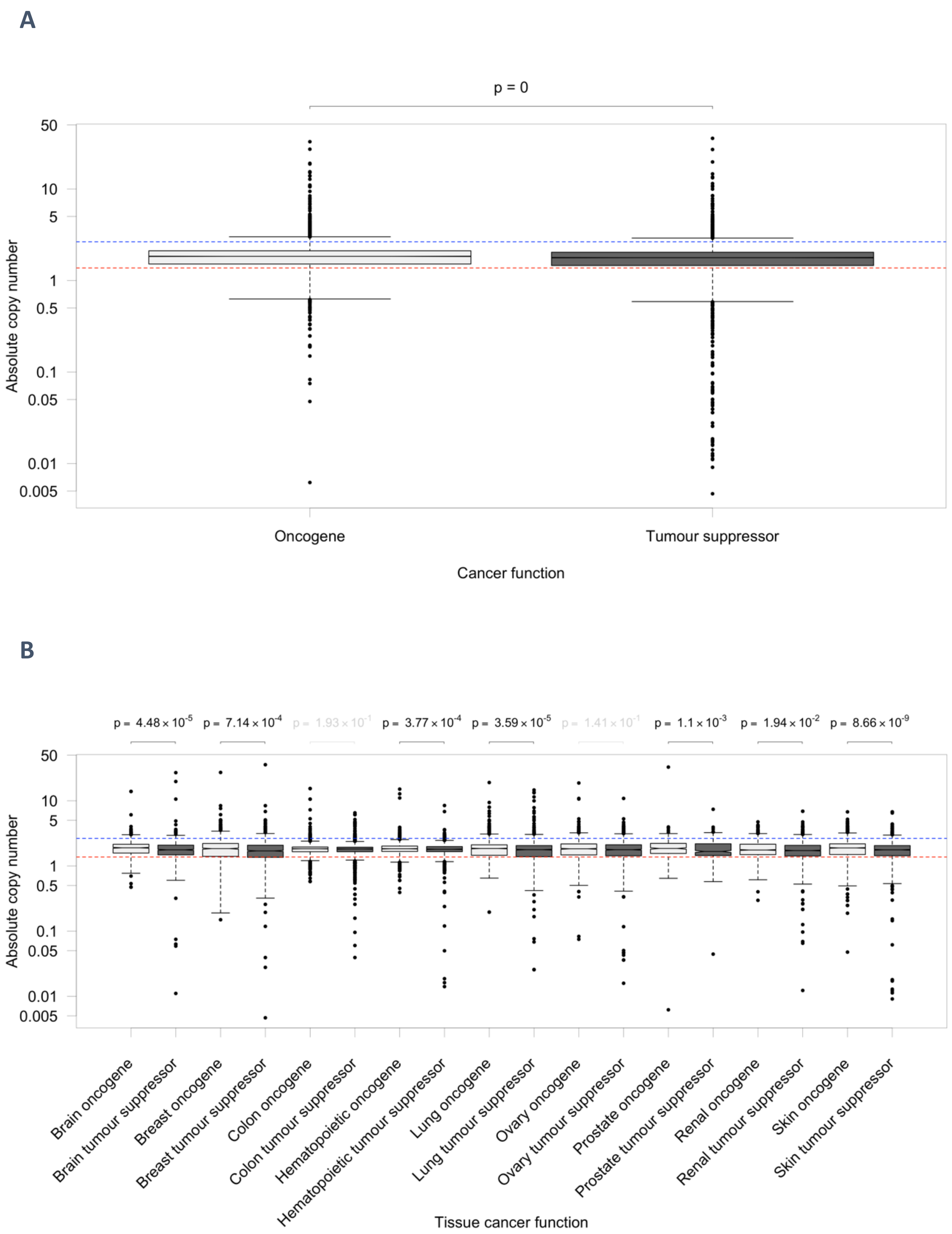


Figure 2.11 - Tumour suppressors are lost more often than oncogenes

(**A**) The panel-wide distributions of the copy numbers of cancer-related genes that are solely oncogenes (light grey) and solely tumour suppressors (dark grey). (**B**) The per-tissue distributions of the copy numbers of cancer-related genes that are solely oncogenes (light grey) and solely tumour suppressors (dark grey). The p-values are from two-sided Kolmogorov-Smirnov tests and are black when significant at $\alpha = 0.05$ and light grey otherwise. The red and blue dashed horizontal lines are the mean loss and gain thresholds respectively.

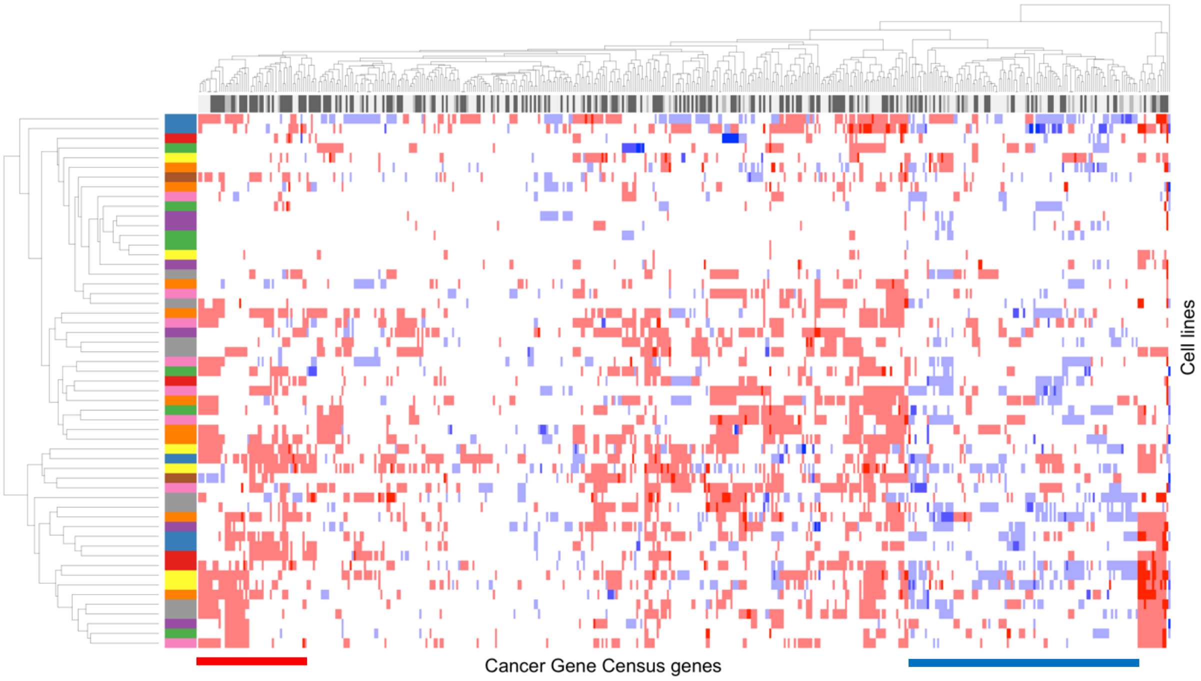


Figure 2.12 - Consistently gained genes are more likely to be oncogenes

A clustered heatmap of the copy numbers affecting 475 cancer-related genes (columns) from the Cancer Gene Census across 55 cell lines (rows) from the NCI-60 cell line panel. The row colours in the y axis dendrogram correspond to the tissue of origin and the column colours in the x axis dendrogram correspond to the genes' known cancer functions, with light grey for oncogenes, dark grey for tumour suppressors and intermediate grey for genes with both functions. The heatmap colours are red for losses, blue for gains and white for unaltered. The largest rooted column cluster of mostly losses is underlined in red and the largest rooted column cluster of mostly gains is underlined in blue.

	Column cluster	
	Losses	**Gains**
Oncogene	16	61
Tumour suppressor	35	31

Table 2.1 - Oncogene and tumour suppressor clusters

The contingency table of the numbers of genes that are solely oncogenes or tumour suppressors in each of the underlined column clusters in Figure 2.12.

2.4.7 The majority of gene copy numbers are not significant by permutation testing

With significant numbers of oncogenes commonly affected by gains as well as significant numbers of tumour suppressors commonly lost, we hypothesised that these cancer-related genes with CNVs in many cell lines could indicate common mechanisms of oncogenesis or cell line immortalisation. However, given the large number of genes and cell lines under consideration, there is a possibility of genes being gained or lost in many cell lines entirely by chance and so we wanted to know which genes in which cell lines had copy numbers which were significantly different to expectation.

To determine the significance of each copy number we performed a permutation test by randomly shuffling the CNVs on each cell line's chromosomes 1,000 times and recording the resulting copy numbers of the CGC genes. This resulted in a distribution of 1,000 copy numbers for each gene in each cell line and we derived a two-sided p-value from each actual copy number's location in its distribution of permutated copy numbers. We used the Benjamini & Hochberg method to control the false discovery rate of these p-values.

The majority (81%) of FDR-corrected copy numbers of cancer-related genes across the NCI-60 panel are not significantly different from what would be expected at $\alpha = 0.05$ by this test (Table 2.2). For example, of the 30 cancer-related genes on chromosome 17, 22 are affected by CNVs in the MCF7 cell line (12 gained and 10 lost) but only two of these, *BRIP1* and *PPM1D*, have FDR-corrected copy numbers that are significant at $\alpha = 0.05$ on the permutation test (Figure 2.13). Interestingly, while both of these genes are gained in MCF7, with 35 and 27 copies respectively, *BRIP1* (Figure 2.13A) is a tumour suppressor whereas *PPM1D* (Figure 2.13B) is an oncogene. Only the latter gene's increased copy number in MCF7 therefore makes sense as a putative driver of oncogenesis or cell line immortalisation.

	Oncogene	Both	Tumour suppressor
Complete loss	37 (53%)	25 (64%)	90 (68%)
Partial loss	200 (12%)	99 (15%)	273 (15%)
Unaltered	1,559 (18%)	582 (21%)	1,578 (19%)
Gain	237 (28%)	77 (30%)	173 (30%)

Table 2.2 - Some significant cancer-related gene CNVs are putative drivers of oncogenesis

The numbers of cancer-related gene copy numbers that are significant on the permutation test at $\alpha = 0.05$ (with the percentage of the total number of gene copy numbers in each category in brackets). In total, 4,930 gene copy numbers out of 26,125 (19%) are significant on the permutation test.

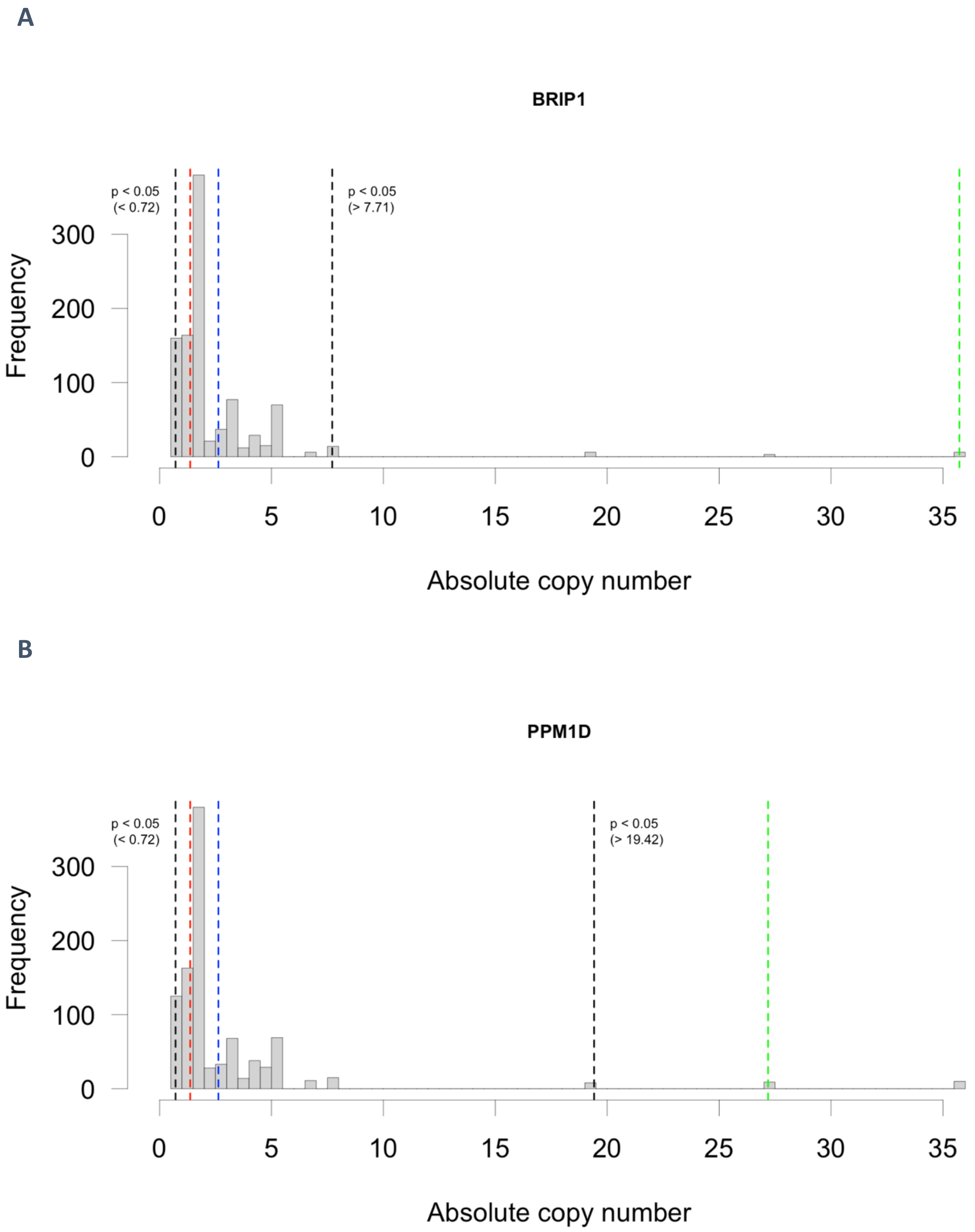


Figure 2.13 - Empirically significantly gained genes on MCF7 chromosome 17

The distributions of copy numbers that arise from 1,000 permutations of CNVs for (**A**) tumour suppressor *BRIP1* and (**B**) oncogene *PPM1D* on chromosome 17 of cell line MCF7 are shown as a frequency histogram, with black vertical dashed lines at the significance thresholds of p = 0.025 and p = 0.975 (two-sided significance at α = 0.05) and red and blue vertical dashed lines at the loss and gain thresholds respectively. The actual copy numbers of the genes in MCF7 are shown as green vertical dashed lines.

2.4.8 Genes with significant copy numbers in CNV hotspots are possible driver genes

Having established that there are hotspots of gain and loss in the NCI-60 cell line genomes and also that there are cancer-related genes with copy number changes unlikely to have occurred by chance, we tested if the intersection of these hotspots and significantly affected cancer genes could indicate putative drivers of oncogenesis or cell line immortalisation. We calculated the intersection of the top 100 most gained and lost 1 Mb regions of the cell line genomes (Figure 2.14A) with the locations of the cancer-related genes with FDR-corrected empirically significant copy number changes (Table 2.2). We find that 16 of the gain hotspots and 21 of the loss hotspots intersected with cancer-related genes with significant copy numbers in at least one cell line, with the gains concentrated on chromosomes 6, 7, 8 and 17 and the losses mainly on chromosomes 16, 22 and X (Figure 2.14B).

Out of the 17 cancer-related genes with significant copy number changes that are gained and 27 that are lost in these hotspots, there are ten gained oncogenes and 17 lost tumour suppressors. These genes are potential drivers of cancer (Table 2.3). Of these, five out of ten of the gained oncogenes and 11 out of 17 of the lost tumour suppressors are not already known to be driver genes (Martincorena et al. 2017), and we therefore suggest these could be novel cancer driver genes.

Gained oncogenes	Lost tumour suppressors
CDK6	ATP2B3
DDX5	AXIN1
FCGR2B	BMPR1A
H3C12	CDKN2A
IRF4	FOXO4
MET	GATA1
MYC	LZTR1
PPM1D	MED12
RAC1	NCOR1
RAD21	NCOR2
	PML
	PPARG
	RPL10
	SMAD2
	TNFRSF14
	TRAF7
	TSC2

Table 2.3 - Gained oncogenes and lost tumour suppressors

The oncogenes that are gained and the tumour suppressors that are in lost in hotspots. Genes in red are those which are not known driver genes according to (Martincorena et al. 2017) and so are potentially novel cancer driver genes.

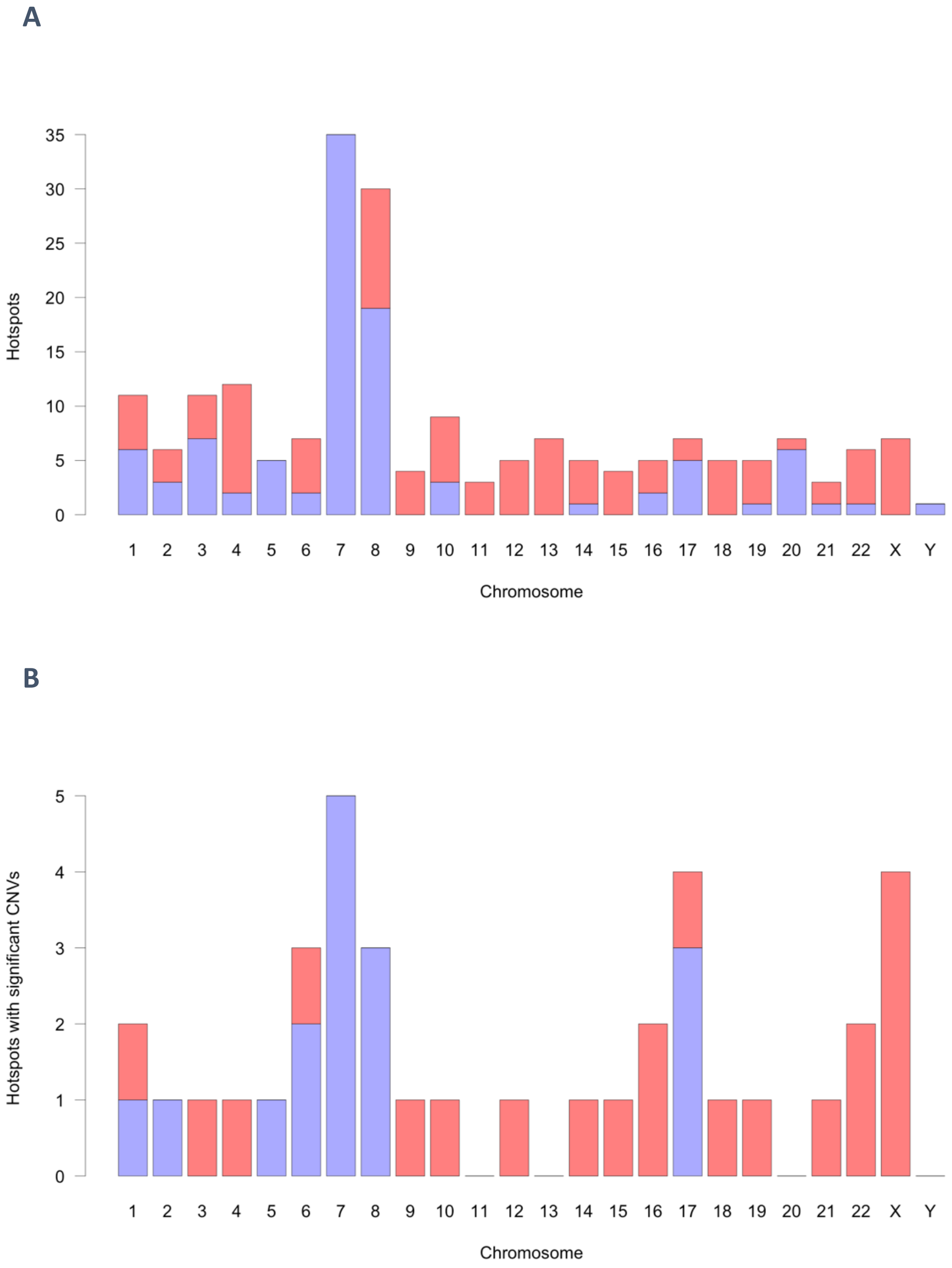


Figure 2.14 - CNV hotspots with significant cancer gene copy number changes

(**A**) The frequencies of gain and loss hotspots on each chromosome and (**B**) the frequencies of gain and loss hotspots on each chromosome which contain cancer-related genes which are significantly affected by CNVs. The frequency of gain hotspots is shown in blue, the frequency of loss hotspots is in red.

Notable among the gained oncogenes are the tyrosine kinase receptor *MET* and GTPase *RAC1* (Table 2.3), both associated with increased cell motility, epithelial-mesenchymal transition (EMT) and metastasis (Stallings-Mann et al. 2012). Also gained in a CNV hotspot on chromosome 8 is the transcription factor *MYC* (Table 2.3), the over-expression of which leads to increased expression of genes involved in cell proliferation, cell growth, apoptosis and DNA replication (Dominguez-Sola et al. 2007; Mannava et al. 2008).

The lost tumour suppressors include the bone morphogenetic protein receptor *BMPR1A*, cyclin-dependent kinase inhibitor *CDKN2A* and signal transducer *SMAD2* (Table 2.3), members of the TGF-β signalling pathway that in healthy cells controls the cell cycle to promote apoptosis or prevent proliferation, but the loss of which in cancer causes increased cell proliferation and angiogenesis (Blobe et al. 2000).

In order to refine our understanding of the cancer-related processes affected by CNVs we performed enrichment analyses of the Gene Ontology (GO) terms and Reactome pathways associated with our putative drivers of cancer (Table 2.3). Ten biological process GO terms are enriched with an FDR-corrected P value of less than 0.05 for the ten gained oncogenes under consideration, including terms associated with regulation of cellular response to stress, chromosome organisation, response to external stimulus and gene transcription and expression (Supplementary Table 6.2). When analysed at the pathway level, these ten gained oncogenes are enriched for 11 Reactome pathways, including pathways involved in cell attachment and motility, MAPK family signalling cascades, Wnt signalling and the cell cycle (Supplementary Table 6.3).

The 17 lost tumour suppressors have 42 enriched biological process GO terms, including multiple terms associated with apoptosis, gene silencing by miRNAs, BMP signalling, cellular senescence, cell cycle arrest, cell growth, cell proliferation and response to stress (Supplementary Table 6.4). The 18 enriched Reactome pathways for the 17 lost tumour suppressors include pathways related to TGF-β signalling, the cell cycle and gene transcription and expression (Supplementary Table 6.5).

2.5 Discussion

Cancer cells experience widespread gene dosage changes caused by point mutations, copy number variations and aneuploidies. Clarifying the causes and effects of these dosage changes is key to understanding and treating cancer (International HapMap 2003; Torres et al. 2008; Yang et al. 2016). The high variability of cancer genomes means that a representative range of cancers in different tissues must be analysed in order to move closer to elucidating general mechanisms of oncogenesis (Iafrate et al. 2004; Henrichsen et al. 2009). Whereas previous studies used array-based comparative genomic hybridisation (Lorenzi et al. 2009; Beroukhim et al. 2010; Bignell et al. 2010; Varma et al. 2014) or whole exome sequencing analyses (Reinhold et al. 2014) to characterise copy number variations in cancer, we have used whole *genome* sequencing datasets to build a more detailed picture of CNVs in the NCI-60 panel (Figure 2.2, Figure 2.3) than has been previously presented. While the analyses of extrachromosomal DNA oncogene copy numbers presented in (Turner et al. 2017) were based on CNVs calculated from the same Illumina data as our analyses and with the same readDepth software, their raw NCI-60 CNVs have not been published and, moreover, were calculated with a less stringent false discovery cut-off than our NCI-60 CNVs (FDR = 0.05 instead of our FDR = 0.01) and with a model overdispersion set to a value less well-suited to Illumina data (overdispersion = 1 instead of our overdispersion = 3). Our analyses of the 'hotspots' in this higher-resolution CNV map allow us to identify possible novel oncogenesis driver genes (Table 2.3).

Consistent with previous studies (Beroukhim et al. 2010; Varma et al. 2014), we found wide variation in levels of copy number variations across tissues and chromosomes, ranging from small 10 Kb gains and losses to aneuploidies of entire chromosome arms (Figure 2.7, Supplementary Figure 6.2). While complete loss of regions of the genome is rare in our data, partial losses dominate the CNV map, along with peaks of gains with tens of extra copies of genes (Figure 2.6, Figure 2.8A). Many of these CNVs coincide with known cancer-related genes. We find that partial losses of tumour suppressor genes and gains of oncogenes dominate (Figure 2.11), as one would intuitively expect in cancer and as previously found by others (Lorenzi et al. 2009; Beroukhim et al. 2010; Bignell et al. 2010; Varma et al. 2014). Unlike earlier aCGH-based results, our sequencing-based approach means that we can detect smaller CNV hotspots, such as the common short arm aneuploidy on chromosome 21 (Figure 2.7) that was undetected by (Varma et al. 2014). By focussing on regions that are similarly affected in many cell lines (Figure 2.9) and on cancer-related genes that are affected more than expected by chance (Figure 2.13), we determined a list of candidate oncogenesis driver genes, which includes known driver genes (Martincorena et al. 2017), but also suggests potentially novel cancer drivers (Table 2.3).

The functions of the gained candidate driver genes are enriched for multiple cancer-related processes such as increased cell growth, motility and proliferation leading to metastasis as well as cellular responses to stress and external stimuli (Supplementary Table 6.3). The lost tumour suppressors on the other hand are enriched for processes associated with control of the cell cycle such as TGF-β and Wnt signalling cascades, apoptosis and cellular senescence, in addition to processes associated with angiogenesis and miRNA-mediated gene silencing (Supplementary Table 6.5). Taken together, these results support the hypothesis that cells must bypass multiple protective mechanisms in order to gain the various 'hallmarks' of cancer and so develop into full-blown metastasising tumours (Hanahan and Weinberg 2000, 2011).

In addition to looking at significantly affected cancer-related genes with concordant copy number changes (gained oncogenes and lost tumour suppressors) in CNV hotspots, we intend also to investigate the CNV hotspots which do not contain possible driver genes as defined with our current criteria. For example, hotspots with cancer-related genes that are not significantly affected and also hotspots with genes which are not currently known to be cancer-related are worthy of consideration since these regions are consistently gained and lost across the NCI-60 cell lines, presumably under the influence of somatic selection (Bignell et al. 2010).

Our definition of CNV hotspots as the top 100 most gained or lost 1 Mb regions of the genome is probably both overly conservative and prone to false positives. An analysis based on susceptibility to nonallelic homologous recombination (NAHR) events would allow us to predict regions prone to genomic instability (Stankiewicz and Lupski 2002), since these are known to be a prominent source of CNVs (Mills et al. 2011). The presence of low copy repeats (LCRs) can catalyse the formation of CNVs (Liu et al. 2012), especially if the LCRs have a high density of *PMDR9*-binding motifs and are thus likely to cause crossover events (Myers et al. 2008). Future work will redefine CNV hotspots as regions flanked by paralogous LCRs containing frequent *PMDR9*-binding motifs and repeat the intersection with the cancer-related genes affected more than expected by chance (Figure 2.13) to investigate how this affects the selection of candidate driver genes.

The low coverage, 0.4x to 3.2x, of the sequencing data used in this study means that the resolution of the detected CNV breakpoints varies from just 41.2 kb to at best 10.3 kb respectively when analysed with readDepth (Miller et al. 2011). This could be greatly improved by using deeper sequencing, such as the 32x coverage used in a recent study which discovered a range of much smaller novel CNVs in nearly 15 thousand individual genomes (Collins et al. 2020).

Another limitation of our sequencing data is that it is all derived from the Illumina short read sequencing platform and so we are unable to detect CNVs accurately in repetitive regions of the genome (Chiang et al. 2009) and cannot detect inversions or translocations at all. Recently developed single-molecule platforms generate much longer contiguous reads up to hundreds of kilobases long, enabling the direct detection of CNVs in repetitive regions, nucleotide-level resolution of CNV breakpoints and the detection of inversions and translocations (Goodwin et al. 2016). These technologies include two competing long-read sequencing platforms developed by Pacific Biosciences (Chaisson et al. 2015) and Oxford Nanopore Technologies (Cretu Stancu et al. 2017), as well as an 'optical mapping' system developed by Bionano (Lam et al. 2012), which uses restriction enzymes to cleave at known sites an immobilised single DNA molecule which can then be imaged and compared to a reference restriction map. The investigation of the NCI-60 cell lines with these technologies would lead to a much more accurate picture of the perturbations to gene interaction networks in these cancer-derived cell lines.

We chose the readDepth algorithm primarily because it was shown to result in more accurate breakpoint and copy number estimations when compared to other available short-read algorithms (Duan et al. 2013), a choice subsequently validated by a study which found readDepth to have high precision for duplications (>89%) and deletions (>95%) across a range of CNV sizes (Kosugi et al. 2019). However, we have not yet formally quantified the precision (true positives / calls) or recall (true positives / actual positives) for either duplications or deletions in our data. Such sensitivity analyses could be implemented either by reference to a simulated genome, where the locations and copy numbers of introduced

CNVs are known, or by comparison to previously well-characterised CNV datasets such as those for NA12878, a well-studied patient genome with matching parental genomes, together with derived lymphoblastoma genome GM12878 (1000 Genomes Project et al. 2010). We will also compare our CNV calls to those for the NCI-60 cell lines which are now available in the Database of Genomic Variants (MacDonald et al. 2014). All CNV detection algorithms have characteristic biases and consequently none detect all CNVs of all sizes accurately (Pabinger et al. 2014), and so future work will merge the CNVs from our readDepth-based short-read analyses with CNVs called using other algorithms, leading in principle to increased precision, albeit probably at the cost of decreased recall (Kosugi et al. 2019).

There is little available data on the stage of tumour from which the NCI-60 cell lines were taken. The widespread aneuploidy that is seen in our data could be a late-stage consequence of, for example, *TP53* inactivation, rather than causal in oncogenesis (Torres et al. 2008). If so, it would be interesting to subtract the aneuploidies from the CNV map to see what difference that would make to the candidate driver genes and related analyses. In addition, normal tissue samples are not available for the NCI-60 cell lines to the best of our knowledge and consequently we are unable to determine germline CNVs from somatic CNVs. Reanalysis either with matched tumour/normal samples or by constructing a model of statistically significant germline CNVs that could be subtracted from the signal in our data (Bignell et al. 2010) could be used to resolve this question.

We have shown that 55 datasets across nine tissues allows us new insight into the role of CNVs and cancer genes. The application of our processing pipeline to work on much larger sets of cell line reads, such as the Cancer Cell Line Encyclopedia with its 1,457 cell lines (as of 3/1/21) derived from tumours in a much wider range of tissues (Ghandi et al. 2019), would be very exciting.

Intentionally blank page.

3 Transient *TP53* repression could lead to *Oncomir-1* activation

3.1 Abstract

Despite the wide variation in gene content in cancer cells caused by copy number variations and aneuploidies, tumour cells survive dosage changes that cause normal cells to undergo apoptosis, often exhibiting enhanced fitness, suggesting that buffering mechanisms must be key to cancer cells' tolerance of these dosage changes. One such dosage compensation mechanism occurs through the repressive activities of miRNAs, which negatively regulate as much as 60% of human protein-coding genes via post-transcriptional repression of mRNAs.

We have used a detailed map of CNVs affecting the NCI-60 cell lines which we generated in chapter 2 to investigate CNVs containing miRNAs in cancer-derived cell lines. Our findings include the widespread derepression of cancer-related processes and pathways caused by the frequent loss of pleiotropic miRNA clusters as well as by more global miRNA depletion resulting from disruption of miRNA biogenesis. We also observe surprisingly frequent loss of non-redundant miRNAs which ordinarily regulate cancer-associated processes, which further suggests that cancer cells benefit from escaping miRNA-mediated regulation.

Our investigation of the CNVs affecting the oncomir cluster *mir-17~92* on chromosome 13, also known as *Oncomir-1*, together with its paralogs and the related transcription factors, leads us to propose a new mechanism by which *TP53* and *PTEN* repression could be sustained in a range of cancers. The transient *C-MYC*-induced repression of *TP53* via *mir-663a* and *mir-1228* implied by the copy number changes in NCI-60 cell lines could lead to sufficient temporary derepression of *Oncomir-1* that *C-MYC* activation of *Oncomir-1* can become dominant, leading to stable repression of *TP53* and *PTEN* and the avoidance of *C-MYC*-induced apoptosis by elevated *Oncomir-1*.

3.2 Introduction

Tumour cells and cell lines derived from them are characterised by widespread genomic instability with frequent copy number variants (CNVs) and aneuploidies causing large changes in gene dosage (Iafrate et al. 2004; Torres et al. 2008; Henrichsen et al. 2009; Yang et al. 2016). However, despite these changes being ordinarily lethal to normal cells (Torres et al. 2008; Sheltzer and Amon 2011), cancerous cells not only survive but often do so with enhanced fitness (Sheltzer and Amon 2011). This suggests there must be mechanisms buffering the effects of these dosage changes.

MicroRNAs (or miRNAs) are short non-coding RNAs that function as post-transcriptional regulators, normally repressors, of protein-coding genes (Bartel 2004). MicroRNAs are transcribed in the nucleus as primary miRNA transcripts containing short hairpin loops, both from the introns of protein-coding genes and from intergenic loci (Lee et al. 2002). The hairpin loops are excised from the primary transcripts by the enzyme *Drosha* (Lee et al. 2002; Lee et al. 2003; Zeng et al. 2003) and exported as precursor miRNAs from the nucleus to the cytoplasm by *RanGTP/Exportin-5* (Yi et al. 2003), where they are further cleaved by the *Dicer* enzyme into a short RNA duplex (Lee et al. 2003). One arm of this duplex is then selected as the mature miRNA (Hammond et al. 2000) to guide the RNA-induced silencing complex (RISC) to bind by sequence complementarity to protein-coding mRNAs (Lewis et al. 2003), primarily in the 3'UTR region, leading then to either inhibition of mRNA translation into protein (Olsen and Ambros 1999; Ding and Grosshans 2009; Zdanowicz et al. 2009) or to the degradation of the mRNA by the 5'-to-3' mRNA decay pathway (Rehwinkel et al. 2005; Behm-Ansmant et al. 2006; Baek et al. 2008).

The primary target recognition site (or seed) of the mature miRNA is only six to eight nucleotides long and so an individual miRNA can have target sites in many mRNAs and thus function pleiotropically (Miska et al. 2007). While the effects of post-transcriptional repression by individual miRNAs are relatively small, often with only around a two-fold decrease in expression (Baek et al. 2008), the co-operative effects of multiple co-expressed miRNAs mean that miRNAs influence most developmental processes (Bartel 2004). The results of miRNA repression of protein-coding gene expression include tuning of gene expression levels to reduce the influence of transcriptional noise (Miska et al. 2007), cell fate decisions and tissue differentiation (Stark et al. 2005), stabilizing or increasing the precision of phenotype inheritance through a process known as canalisation (Hornstein and Shomron 2006; Peterson et al. 2009) and, more generally, dosage compensation (Sheltzer and Amon 2011).

Using the high-resolution CNV map generated in chapter 2 we have analysed the CNVs containing miRNAs across the NCI-60 cell line panel. MicroRNAs are often expressed from polycistronic loci (Lee et al. 2002), and these miRNAs are likely to be affected by the same CNVs purely because of their genomic proximity to each other. We have therefore developed a novel method of avoiding the multiple counting of miRNA CNVs that groups the miRNAs by their seed sequence as well as by their genomic locations into seed/locus families.

We show that miRNAs in multi-precursor families are lost more frequently and gained less often than those in single-precursor families, consistent with the greater disruption to the cell that gains of multiple pleiotropically acting miRNAs would cause. Counterintuitively, we find however that miRNA families with members expressed from genes in multiple loci in the genome (duplicated miRNAs) and which are therefore functionally redundant are lost less than non-redundant families.

We infer global depletion of miRNAs from our observations of partial losses of components of the miRNA biogenesis pathway in many cell lines, consistent with earlier studies and with the genomic instability characteristic of cancer (Lin and Gregory 2015). We find that miRNAs which are consistently gained across the NCI-60 panel are enriched for cancer-related processes and pathways.

The CNVs of well-known oncomir cluster *mir-17~92*, also known as *Oncomir-1*, together with the CNVs of the related oncogenic transcription factors, are indicative of tumorigenesis in the vast majority of cell lines. We propose that *Oncomir-1* and its paralogs integrate the CNV signals from their transcription factors and, acting via *mir-663a/1228*, ensure that *TP53* and *PTEN* expression remains low once *Oncomir-1* expression has been elevated.

3.3 Methods

3.3.1 Seed/locus miRNA families

Human protein-coding gene locations were downloaded from Ensembl on 29/9/20 (release 100) (Yates et al. 2020) and human miRNA locations and sequences were downloaded from miRBase on 12/8/20 (release 22.1) (Kozomara et al. 2019). The overlaps between CNV locations and gene locations were calculated using R package GenomicRanges v1.40 (Lawrence et al. 2013). In analyses where a gene overlapped more than one CNV and a single effective copy number was required then the lowest copy number was used, since loss of part of a gene means it is likely to be non-functional and gain of part of a gene means that if it is transcribed/translated at all then there will just be extra fragments of the protein.

Heatmaps were generated using the heatmap.2 function of R package gplots v3.0.3 (Warnes et al. 2020). Heatmap clusters were extracted by parsing the heatmap layout returned by the heatmap.2 function and combining adjacent and identical columns and rows into clusters.

MicroRNA seeds were calculated from 5' nucleotides two to eight inclusive. MicroRNA seed CNVs were calculated by comparing the actual number of mature miRNAs with the seed to the expected number for each cell line, taking the patient sex into account for miRNAs on chromosomes X and Y.

To avoid multiple counting precursor miRNAs which are affected by the same CNVs purely because of their genomic proximity, adjacent miRNA precursors were combined into loci if the precursors were less than a 'maximum gap size' apart (see section 3.4.2 for an illustration of the construction of loci). A range of maximum gap sizes were investigated between one nucleotide and 10^8 nucleotides; the median CNV length (6.2×10^5 nucleotides) was chosen as the maximum gap size to focus on so as to relate the loci to the CNVs that affect them. The distinct miRNA seeds in each resulting locus were then split into miRNA 'seed/locus' families containing the precursor miRNAs with those seeds in each locus.

3.3.2 GO and pathway enrichment

The mappings of Gene Ontology terms (Ashburner et al. 2000; Consortium 2021) to genes for cellular compartments, molecular functions and biological processes were downloaded from Ensembl on 6/2/21 (release 103). The mappings of Reactome pathways (Jassal et al. 2020) to genes were downloaded from Ensembl on 6/2/21 (release 103). Verified miRNA/target interactions were downloaded from miRTarBase (Chou et al. 2018) on 6/2/21.

A pipeline was constructed for miRNA functional enrichment analyses (discussed further in section 3.4.4):

a) Genes in each ontology category were additionally mapped to all higher categories to reduce the influence of genes of uncertain function.
b) Genes mapped to each category were replaced with the miRNAs experimentally confirmed in miRTarBase to target the category's genes.
c) Fisher's Exact test was performed with the miRNAs of interest and each category's targeting miRNAs to find over-represented categories and p-values were corrected for multiple testing with the Benjamini & Hochberg method (Benjamini and Hochberg 1995).
d) The enriched categories were summarised by counting the matches to lists of keywords associated with metastasis, cell cycle, expression, signalling and development (Supplementary Table 6.9).

3.3.3 Cancer-related miRNAs and transcription factors

Lists of miRNAs which act as oncomirs and/or tumour suppressors were manually curated by extensive literature search (Motofeanu 2019).

Transcription factors confirmed to bind to miRNAs were downloaded from the supplementary information of an experiment in the ENCODE project on 24/2/21 (Gerstein et al. 2012).

3.3.4 Statistical analyses

All statistical analyses were performed in R. Correlations were calculated using the Pearson product-moment correlation test. Distributions were compared with the Kolmogorov-Smirnov test. Contingency table tests and goodness of fit tests were performed with the X^2 test.

3.4 Results

3.4.1 Individual miRNA CNV profiles are confounded by genomic proximity effects

We downloaded locations and sequences for precursor and mature miRNAs from miRBase (Kozomara et al. 2019) and calculated the overlaps of these with the CNVs of the NCI-60 cell line panel which we generated in chapter 2. The CNVs of individual miRNAs are calculated by comparing the actual copy number determined by overlap with CNVs to that expected for the chromosome in each cell line (2 for autosomal and female X chromosomes, 1 for male X and Y chromosomes). When these individual miRNA CNVs are viewed as a hierarchically clustered heatmap (Figure 3.1A) it is clear that there is no clustering by tissue of origin, as shown by the mixing of row colours, but there are obvious column clusters of miRNAs with identical CNV profiles across the NCI-60 panel (Figure 3.1A, clusters 'a' to 'd' among others). The only noticeable row-based pattern is the horizontal white band of mainly unaltered miRNAs; this consists however of the cell lines which generally have only very sparse CNVs anyway and so it is not surprising that the miRNAs in these cell lines are also mainly unaltered.

There are 1,216 distinct loss and gain patterns of individual miRNA precursor CNVs, 305 of which form column clusters which contain more than one miRNA. The largest of these clusters (Figure 3.1A, cluster 'd') consists of 73 precursor miRNAs which are in regions of the genome which are unmappable and so are indistinguishable from the expected copy numbers. The next three largest miRNA clusters (Figure 3.1A, clusters 'a', 'b' and 'c'), consisting of 22, 42 and 37 miRNAs respectively with the same CNVs across all the cell lines and are all in close genomic proximity to each other. MicroRNA precursors that are in close genomic proximity will tend to be affected by the same CNVs and hence no biological or functional implications can be inferred just from their identical CNV profiles.

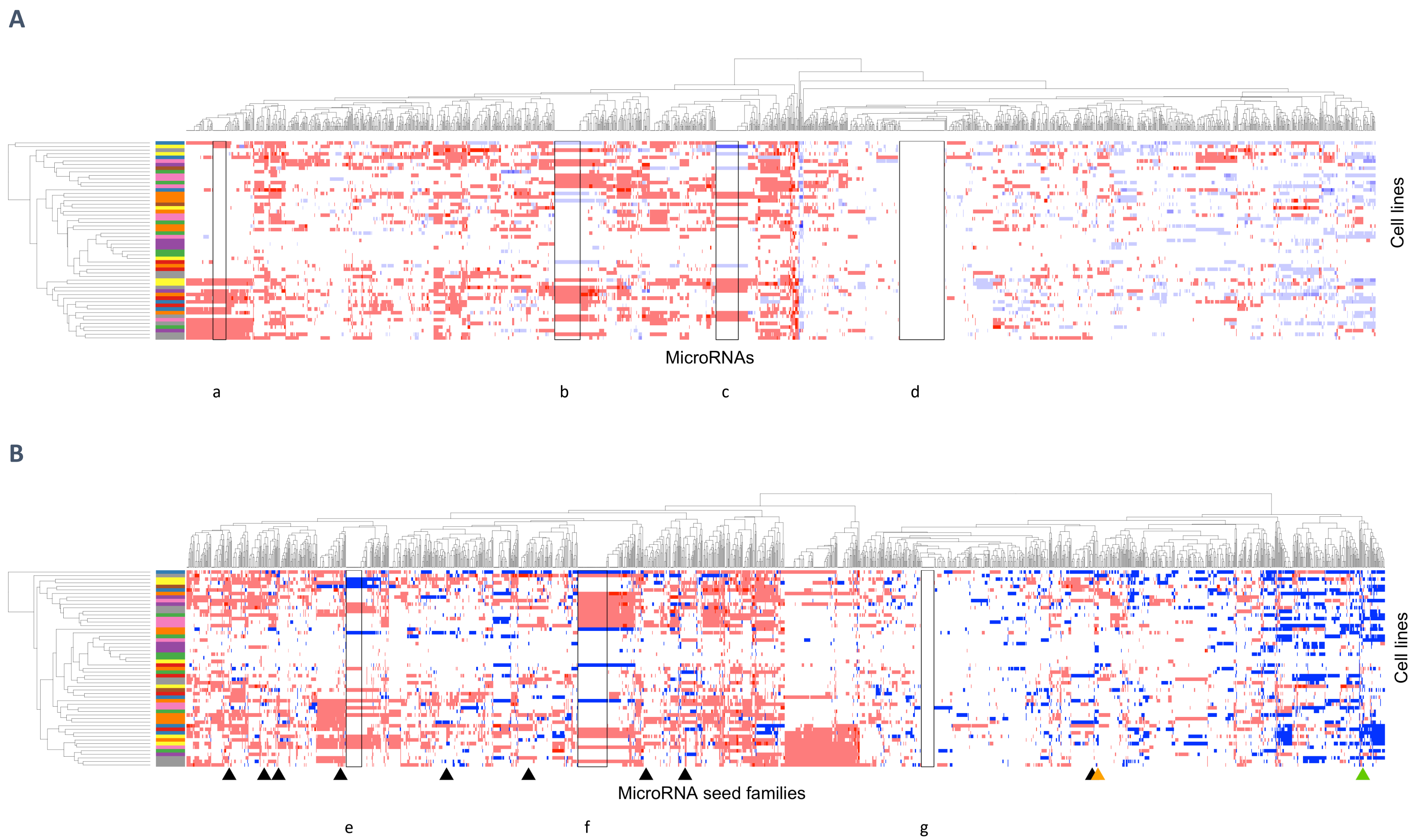


Figure 3.1 - MicroRNA precursor and seed families cluster by CNV profile

Hierarchically clustered heatmaps indicating (**A**) individual miRNA CNVs and (**B**) miRNA seed family CNVs, clustered by CNV profile across the NCI-60 panel (columns) and by cell line (rows). The row colours indicate the tissue of origin. Losses are red, unaltered or unmappable miRNAs are white and gains are blue. Larger clusters discussed in the text are outlined in black and smaller notable clusters are marked with triangles. The orange triangle indicates the *mir-103* cluster, the green triangle indicates the *mir-30* cluster.

The CNVs of individual miRNAs are thus confounded by these genomic proximity effects and so we looked next at the effective dosage of each miRNA seed, as this relates more closely to the cumulative repressive effects of the miRNAs. We calculated the seed CNVs as the differences between the actual and expected total counts of mature miRNAs with each seed in each cell line, based on the underlying miRNA precursor CNVs. As with the individual miRNAs, seed CNVs do not cluster by tissue of origin but there are again clusters of seeds with identical CNV profiles across the cell lines (n = 399) (Figure 3.1B). The unmappable miRNAs' seeds group together again, this time as the third largest group (Figure 3.1B, cluster 'g'), and the two largest clusters (Figure 3.1B, clusters 'e' and 'f'), consisting of 28 seeds on chromosome 19 and 52 seeds on chromosome 14 respectively, are again in close genomic proximity within each cluster.

There are however 11 groups of two or three seeds which have identical seed CNV profiles, but which are spread across more than one chromosome and so cannot be affected directly by the same CNVs (Figure 3.1B, clusters marked with triangles; Supplementary Table 6.6). Taking the *mir-30* paralog seeds as an example, while the two seeds in the cluster occur on three different chromosomes, they do so in the same three adjacent pairs of loci across the chromosomes (Figure 3.1B, green triangle; Figure 3.2A; Supplementary Table 6.6). The six *mir-30* paralogs all have the same 5' seed of GUAAACA and one paralog in each pair of loci also has the 3' seed of UUUCAGU (Figure 3.2A). Importantly, neither seed occurs anywhere else in the genome and so, since the instances of each seed are relatively close together on each chromosome (Figure 3.2A), they experience the same set of CNVs across the three chromosomes and hence have the same seed dosage profile across the NCI-60 panel.

The *mir-103* paralog seeds on the other hand each occur in different loci but, crucially, these loci are not just in close genomic proximity again within each chromosome but actually overlap on different DNA strands and so they also have the same CNV and seed dosage profiles (Figure 3.2B; Figure 3.1B, orange triangle; Supplementary Table 6.6).

We quantified this idea of genomic proximity as the maximum 'per-chromosome inter-locus spread', defined as the largest region spanned by any two adjacent miRNA precursors on each chromosome, ranging from just 78 nucleotides for the *mir-103a/b* paralogs in the same location but on opposite strands (Figure 3.2B) to 6.48 x 10^5 nucleotides for the *mir-33/6777/6889* group (Supplementary Table 6.6). In order to relate this metric to the CNVs that affect the cell lines we chose the median CNV length of 6.2 x 10^5 nucleotides as a conservative threshold below which we would consider two loci to be in close genomic proximity.

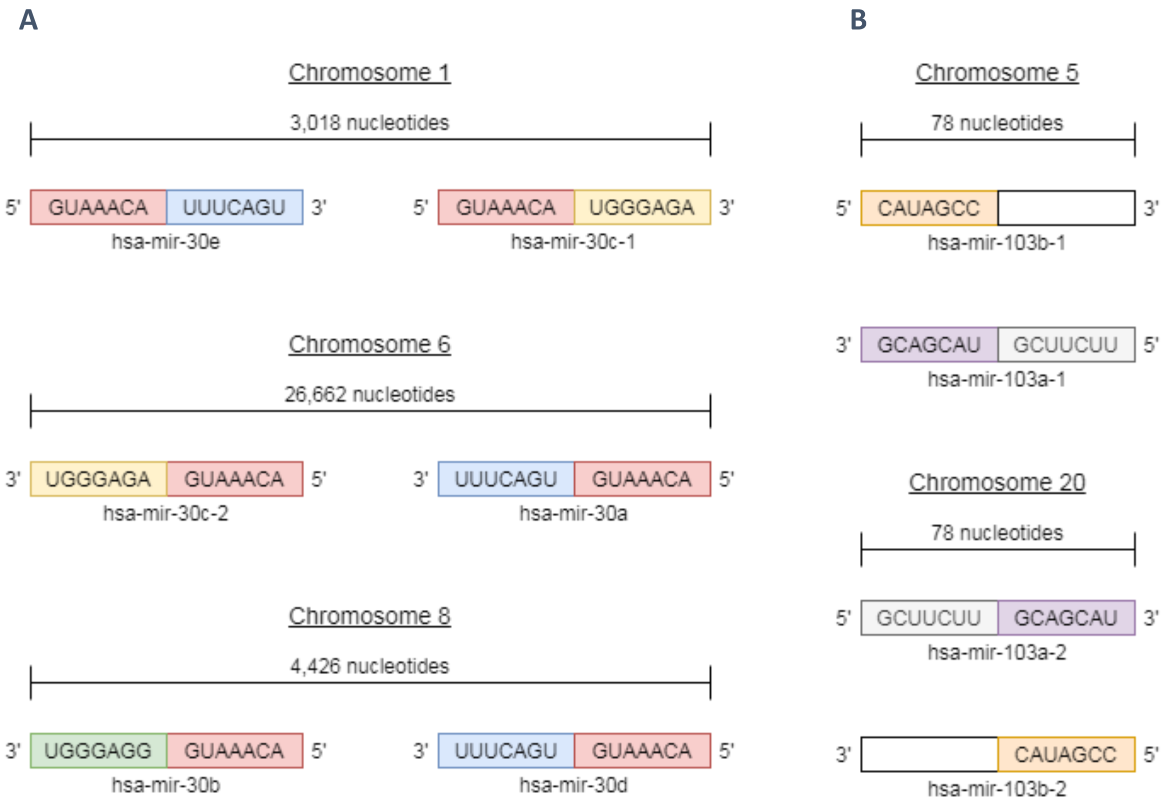


Figure 3.2 - Multi-chromosome genomic proximity explains some CNV clusters

The genomic locations, orientations and seed sequences of (**A**) the *mir-30* paralogs and (**B**) the *mir-103* paralogs. The 5' and 3' mature miRNA seeds are coloured according to their seed sequences. *hsa-mir-103b-1* and *hsa-mir-103b-2* do not have 3' mature miRNAs annotated in miRBase.

To narrow our focus to miRNAs with seeds which cluster together without confounding genomic proximity effects we removed clusters where all the seeds are either unmappable, in the same locus (n = 1,099) or have a maximum per-chromosome inter-locus spread below the median CNV length of 6.2 x 10^5 nucleotides (n = 1,200), resulting in 36 seed CNV clusters with identical CNV profiles which cannot be explained by the variants of genomic proximity effect discussed above (Supplementary Table 6.7, analysed further in section 3.4.4).

The repressive effects of miRNAs on their targets are determined in part by the effective concentration of the miRNAs and, since the miRNAs' targets are determined by their seed sequence, we wanted to see how CNVs altered the effective dosage of each miRNA seed.

We calculated the copy number ratio for each miRNA seed as the ratio of actual to expected seeds in each cell line. The maximum copy number ratio generally decreases with increasing number of expected copies (Figure 3.3A), which is unsurprising as a randomly gained miRNA represents a larger relative gain for smaller seed families.

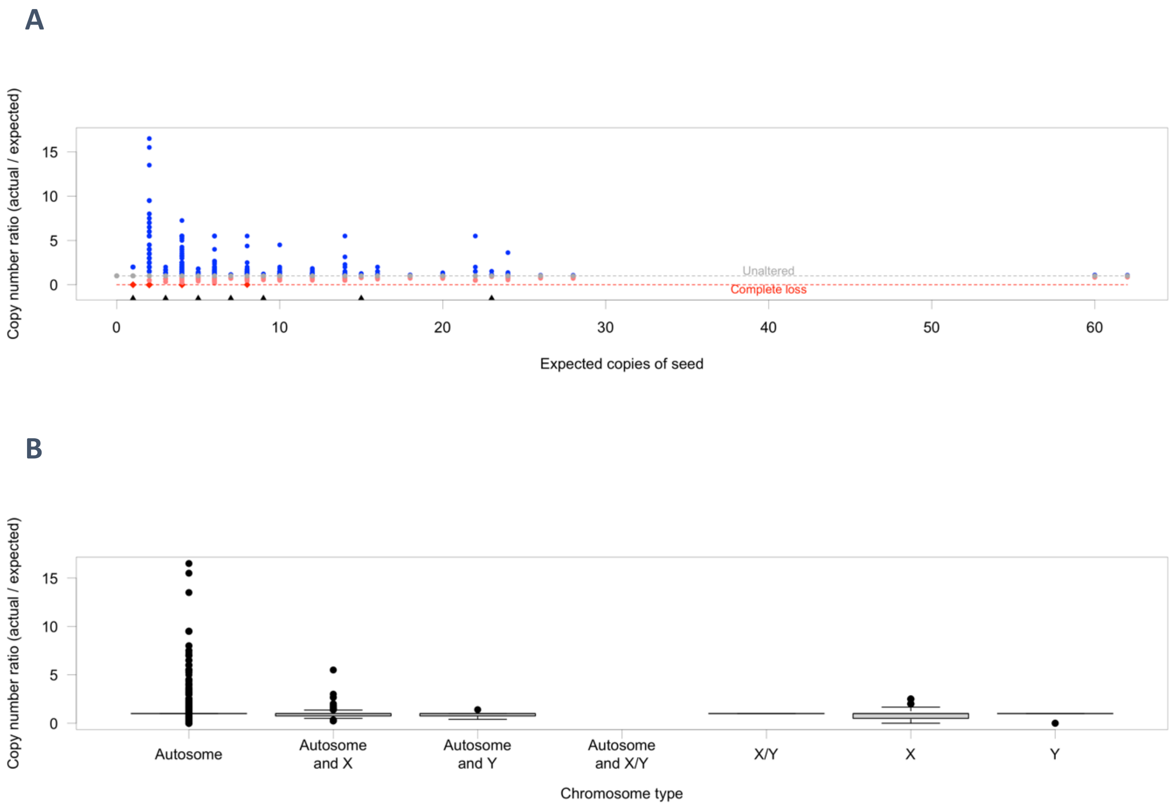


Figure 3.3 - Seed copy number ratios confirm bias against sex chromosome gains

(**A**) The copy number ratios of miRNA seed families (y axis) plotted against the expected number of copies (x axis). Gains with a copy number ratio above one are shown in blue, unaltered in grey, partial losses in light red and complete losses as dark red diamonds. The grey dashed line indicates a copy number ratio of 1 (unaltered) and the red dashed line indicates complete loss. The black triangles mark the seed families with an odd number of expected copies. (**B**) The distributions of miRNA seed family copy number ratios when grouped by the chromosome types on which they occur.

Seeds families with more than eight expected copies are never completely lost, consistent with the lower chances of larger families losing all their copies (Figure 3.3A). Strikingly, seeds with an odd number of expected copies are gained much less than those with an even number of expected copies and, apart from the seeds with only a single expected copy (on chromosomes X or Y in male cell lines), are never completely lost (Figure 3.3A).

Since the only way that a seed can have an odd number of expected copies is to have at least one copy on a sex chromosome in a male-derived cell line, we grouped the copy number ratios of miRNA seed families by the types of chromosomes on which they occur (Figure 3.3B). While seed families that have copies only on autosomes have copy number ratios of up to 16.5, the seeds that occur only on sex chromosomes never have a copy number ratio of more than 2.5 (Figure 3.3B). As we saw in section 2.4.3, the sex chromosomes are affected much less by gains than the autosomes (Figure 2.6A), thus explaining the lower gains of miRNA seed families with an odd number of expected copies (Figure 3.3A).

3.4.2 Seed/locus families enable CNV analyses without genomic proximity bias

We wanted to be able to investigate if miRNAs with redundant seeds are affected by CNVs differently than singleton miRNAs and to understand the dosage changes that occur in cancer-derived cell lines without double or multiple counting miRNAs with the same seed that are in close genomic proximity. We also wanted to be able to analyse genomic proximity and seed dosage independently. We therefore combined miRNA precursor loci that are adjacent by comparing the gaps between them to a range of maximum gap sizes and then considered these combined loci with the distinct mature miRNA seeds within them to be ‘seed/locus’ miRNA families (Figure 3.4). This allows us to differentiate between copies of a seed which are in the same effective location (seed ‘B’ in locus 1 in Figure 3.4 for example), which we therefore expect to be affected by the same CNVs, and between copies of a seed which are in different effective locations (such as seed ‘A’ in loci 1, 2 and 4 in Figure 3.4) but still have the same CNV profile, which could be evidence of selection for gain and/or loss of the genes containing these seeds. As well as taking genomic location into account and so avoiding multiple counting of seeds in genomic proximity, this method also allows us to determine accurately the effective dosage of each seed within the cancer genome.

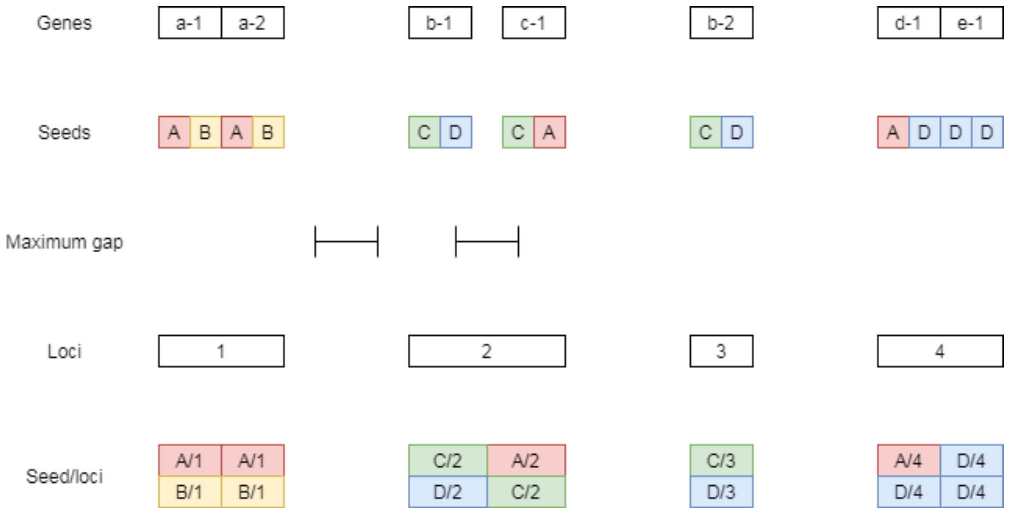


Figure 3.4 - Seed/locus miRNA family construction avoids genomic proximity bias

The mock genes a-1 to e-1 contain the seeds A to D in various combinations and their loci are combined when the gap between them is less that the maximum gap, such that the loci of genes b-1 and c-1 are combined in the example above. The seeds are then allocated to the combined loci in which they occur, resulting in seed/locus families with different seeds such as A/1 and B/1, both in locus 1, or families with the same seed, like A/1, A/2 and A/4, which are spread across three loci. Seed/locus B/1 is an example of a non-redundant seed/locus as its seed, B, does not occur in any other combined loci. The other seed/loci are all redundant as there are copies elsewhere in the genome of their seeds.

The proportion of miRNA precursors that are allocated to multi-precursor loci by this method varies considerably with the chosen gap size (Figure 3.5A). Even with the smallest possible gap size of one nucleotide there are 58 loci (6%) which contain more than one precursor because the precursors overlap on different DNA strands. As the gap size increases the proportion of loci which contain more than one precursor also increases until, with a gap size of 10^8 nucleotides, all the precursors are grouped together into one locus per chromosome (Figure 3.5A). There is a small excess above the sigmoidal random expectation background of multi-precursor loci for gap sizes between about 1 kb and 10 kb (Figure 3.5A, red dots), reflecting presumably the precursors which are transcribed from the same primary transcripts or host genes.

The total number of loci for the various gap sizes ranges from 1,859 loci for a gap size of one nucleotide to 24 loci (one per chromosome) for a gap size of 10^8 nucleotides (Figure 3.5B). When these loci are combined with the 2,090 distinct miRNA seeds to form seed/locus miRNA families the number of families which result for each gap size is dominated by the number of seeds and so the resulting seed/loci vary much less in number than the loci, from 2,843 seed/loci for a gap size of one nucleotide down to 2,582 seed/loci for a gap size of 10^8 nucleotides (Figure 3.5B). We wanted to avoid multiple counting miRNAs with the same seed that are affected by the same CNVs and so we chose the median CNV length of 6.2 x

10^5 nucleotides as the gap size to focus on, resulting in 923 loci and 2,677 distinct seed/locus families (Figure 3.5B).

We calculated the CNVs affecting each seed/locus family in a similar way as for the CNVs of the seed-based families, by comparing the actual and expected numbers of mature miRNAs with each seed in each cell line. We used the overlap between the combined loci and the CNVs to determine the actual copy numbers for each seed/locus and so a given seed can now have multiple CNV profiles if it occurs in different loci (Figure 3.5C). Once again, we see little to no clustering by tissue of origin but there are clear clusters of miRNA seed/loci with identical CNV profiles (Figure 3.5C). The five largest such clusters include a cluster of the seed/loci containing the miRNAs in unmappable regions of the genome again (Figure 3.5C, cluster 'e'), as well as a cluster dominated by partial and complete losses (Figure 3.5C, cluster 'd'), which was not apparent in the individual miRNA precursor or seed-based heatmaps discussed earlier. As with the two largest clusters (Figure 3.5C, clusters 'b' and 'c'), the seed/locus families in this cluster of complete and partial losses occur in a single locus and so it is not surprising that the miRNAs have the same CNV profile.

The fifth largest cluster however (Figure 3.5C, cluster 'a') consists of 26 miRNA seed/locus families spread across four non-adjacent loci and, importantly, the miRNAs in this cluster are not always affected by the same CNV in each cell line and so, despite the miRNAs occurring across 2.2 x10^7 nucleotides of chromosome X, the cluster cannot be explained simply by the length of the CNVs (Supplementary Figure 6.3). After removing the cluster of unmappable miRNAs, 799 single-locus clusters and 808 clusters where the miRNAs are affected by the same CNVs in each cell line, there are 36 clusters of seed/locus families which have identical CNVs across the NCI-60 panel (Supplementary Table 6.8) and which might therefore be evidence of selection for gain or loss of particular miRNAs in cancer. Interestingly, there is considerable though incomplete overlap between the groups of miRNAs in these clusters and the miRNAs in the 36 seed based CNV profile clusters discussed in the previous section (overlaps highlighted in Supplementary Table 6.8, analysed further in section 3.4.4).

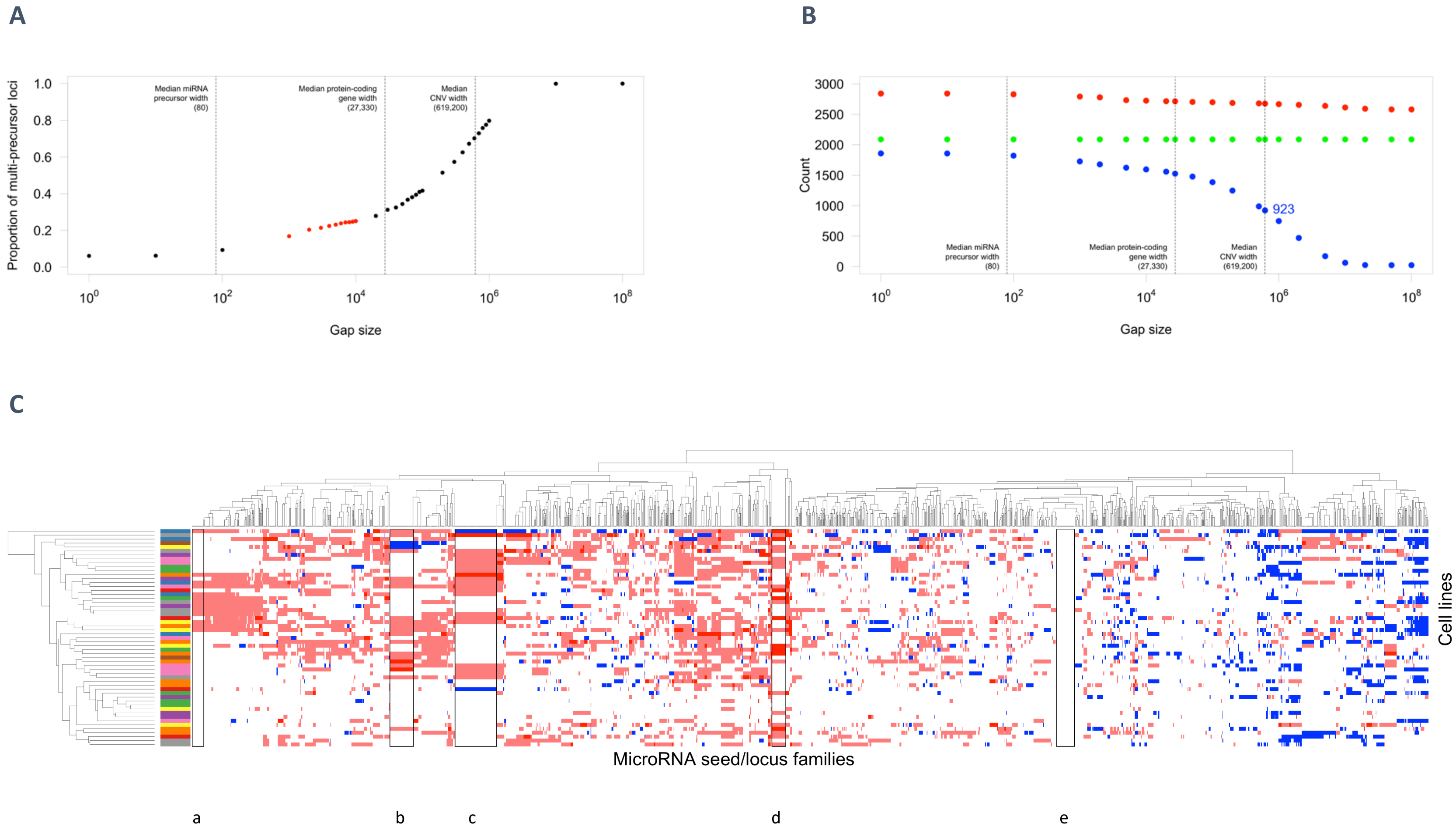


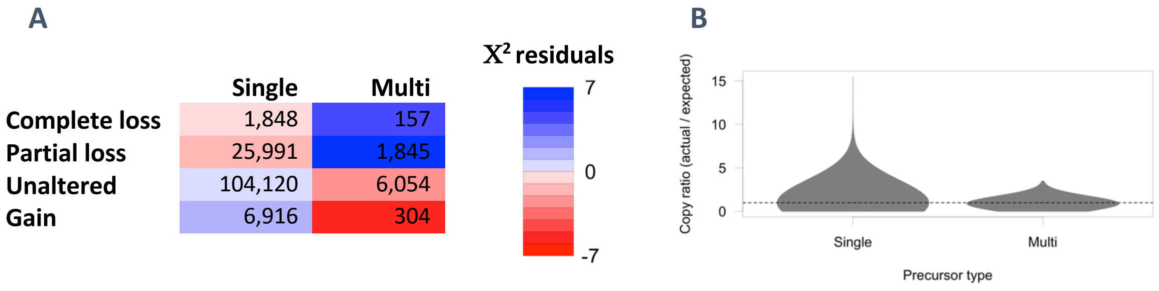
Figure 3.5 - Seed/locus families exhibit CNV clustering across multiple loci

(**A**) The proportion of precursor miRNAs which are in multi-precursor loci for a range of gap sizes. Gap sizes between 10^3 and 10^4 are shown in red. (**B**) The number of seed/locus families (red points) which result from combining the 2,090 individual seeds (green points) and loci generated for each gap size (blue points). The number of loci (923) for a gap size of the median CNV length (6.2 x 10^5 nucleotides) is shown in blue. (**C**) Heatmap indicating miRNA seed/locus family CNVs, clustered by CNV profile across the NCI-60 panel (columns) and by cell line (rows). The row colours indicate the tissue of origin. Losses are red, unaltered or unmappable miRNAs are white and gains are blue. Clusters discussed in the text are outlined in black.

3.4.3 Multi-precursor and non-redundant seed/loci are lost more than expected

We hypothesised that seed/locus families which contain multiple precursors would be more likely to contain miRNAs with different seeds and therefore repress a greater range of targets than single precursor seed/locus families and we wanted to see if this would affect their copy numbers. We divided the seed/locus families' CNVs in the NCI-60 cell lines into those for families with just one precursor (n = 138,875) and those with more than one precursor (n = 8,360).

There is a significant relationship between CNV type and whether a seed/locus family has multiple precursors (X^2 (3, N = 147,235) = 100, $p < 2.2 \times 10^{-16}$), with multi-precursor seed/locus families lost more and gained less than expected (Figure 3.6A). The single-precursor seed/locus families on the other hand not only exhibit the inverse relationship with CNV type of being lost less and gained more often than expected (Figure 3.6A), but also have a much wider range of gains than do the multi-precursor seed/locus families (two-sided Kolmogorov-Smirnov test, D = 0.04, $p = 9.26 \times 10^{-11}$, Figure 3.6B).



	Single	Multi
Complete loss	1,848	157
Partial loss	25,991	1,845
Unaltered	104,120	6,054
Gain	6,916	304

Figure 3.6 - Multi-precursor seed/loci are lost more than expected with fewer gains

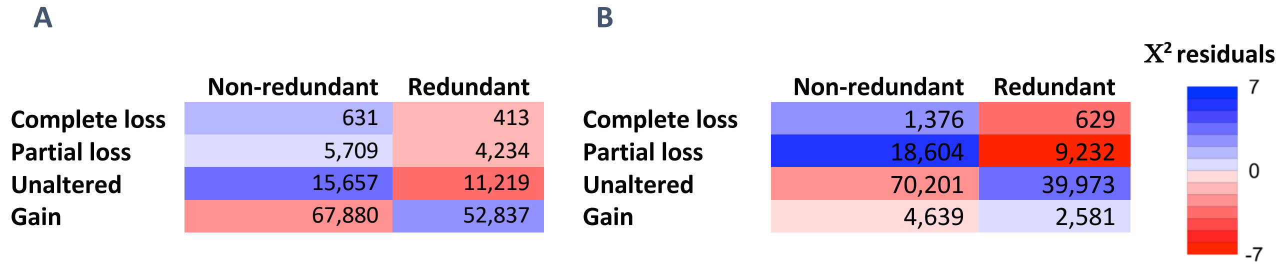
(**A**) The numbers of single and multi-precursor seed/locus family CNVs. The cells are shaded from blue for over-represented to red for under-represented, based on the X^2 residuals of the contingency table, which are calculated as $\frac{observed-expected}{\sqrt{expected}}$. (**B**) Copy number ratios (actual / expected seed copies) for single and multi-precursor seed/locus families. The dashed line at y = 1 indicates that actual copies = expected copies.

Both results are consistent with selection pressure on cancer cells not only to avoid the increased disruption that gains of multiple miRNAs would cause but also to benefit from the widespread derepression that results from the loss of multi-precursor seed/locus families, similar to the known advantages to cancer cells of the global depletion of miRNAs caused by partial *Dicer* knockdown (Lin and Gregory 2015).

Having investigated how the broadness of miRNA targeting affects CNVs by dividing miRNA seed/locus families into single and multi-precursor families, we wanted to then address the related hypothesis that cells should be able to tolerate loss of miRNAs with redundant seeds more readily than loss of miRNAs without a copy of their seed elsewhere in the genome, since this would result in a reduction of miRNA-mediated repression rather than a total loss. In addition to this we wanted to quantify the effects of using our novel miRNA seed/locus family concept to avoid biases arising from the multiple counting of miRNAs in close genomic proximity. We realised that we could compare the effects of seed redundancy on CNVs calculated for miRNA precursors irrespective of their genomic proximity to the effects of seed redundancy on CNVs for miRNA seed/locus families where, by definition, the seed/loci are unlikely to be affected by the same CNVs merely due to their location in the genome.

Accordingly, we divided individual precursor miRNA CNVs into those for miRNA precursors without a copy of their seeds elsewhere in the genome (n = 89,877) and those with a redundant seed (n = 68,703). We similarly partitioned miRNA seed/locus family CNVs into those for families without a redundant seed in another locus (n = 94,820) and those with a redundant seed (n = 52,415).

Surprisingly and counterintuitively, we found both for miRNA precursors (X^2 (3, N = 158,580) = 45.3, p = 7.8 x 10^{-10}, Figure 3.7A) and for miRNA seed/locus families (X^2 (3, N = 147,235) = 110, p < 2.2 x 10^{-16}, Figure 3.7B) that while seed redundancy does indeed have a significant effect on CNV type, those without a redundant seed elsewhere in the genome are lost more and gained less than expected. However, by investigating the NCI-60 cell lines we have selected for cells which have already become cancerous, and so perhaps the inversion of our initial hypothesis is to be expected since these cells are already radically altered from the norm, making miRNAs with non-redundant seeds which are nevertheless lost worthy of further study.



A

	Non-redundant	Redundant
Complete loss	631	413
Partial loss	5,709	4,234
Unaltered	15,657	11,219
Gain	67,880	52,837

B

	Non-redundant	Redundant
Complete loss	1,376	629
Partial loss	18,604	9,232
Unaltered	70,201	39,973
Gain	4,639	2,581

X^2 residuals
7
0
-7

Figure 3.7 - Seed/locus families without redundant seeds are lost more than expected

(**A**) The numbers of CNVs for miRNA precursors with and without copies of the miRNA seeds elsewhere in the genome. (**B**) The numbers of CNVs for miRNA seed/locus families with and without copies of the seed/loci seeds in other loci which are not in close genomic proximity. The cells are shaded from blue for over-represented to red for under-represented, based on the X^2 residuals of the contingency tables, which are calculated as $\frac{observed-expected}{\sqrt{expected}}$.

The effects of grouping miRNA precursors into miRNA seed/locus families can also be clearly seen, with a much sharper separation of over-represented losses from under-represented unaltered/gained for non-redundant miRNA seed/loci (Figure 3.7B) when compared to miRNA precursors (Figure 3.7A). This suggests that the seed/locus method designates some CNVs that would have been unaltered as partially lost instead, because in combined CNVs losses dominate unaltered and gains (see method in section 3.3.1).

3.4.4 Seed/locus clusters are enriched for cancer-related processes

We have identified several sets of miRNAs with potentially significant CNVs (36 seed clusters, 36 seed/locus clusters and non-redundant seed/loci which are lost more than expected). We wanted to see if they are functionally related in order to test the hypothesis that selection pressure is driving cancerous cells to gain or lose miRNAs which regulate specific relevant processes or pathways.

A 'standard' gene ontology or pathway over-representation analysis of the targets of a list of miRNAs, where the union of the miRNAs' targets are compared to the genes in Gene Ontology (GO) categories or to genes known to participate in pathways, has been shown to produce false positives (Bleazard et al. 2015), with many cancer-related categories consistently shown as over-represented even with randomly chosen miRNAs. A subsequent study showed that this 'miRNA targeting bias', whether due to an underlying biological reason or due to sampling bias in the pathways studied, can be avoided by 'inverting' the methodology and converting the GO categories' or pathways' genes into lists of miRNAs that target these genes, so that the hypergeometric over-representation test of the miRNAs of interest can then be performed directly against targeting miRNAs (Godard and van Eyll 2015).

An over-representation analysis cannot be performed simply using the annotation of genes to GO categories as downloaded from Ensembl as genes are mapped only to their most specific *known* GO category. Genes whose biological processes are completely unknown are mapped simply to the root term of the ontology and genes with uncertain function are mapped to the top layers of the ontologies, leading to over-representation of these categories in enrichment results, which we will call 'ontological uncertainty bias'. This could at least partially explain why multiple online miRNA over-representation tools, such as DIANA-miRPath (Vlachos et al. 2015), often show high-level terms such as the root term 'biological process' itself to be enriched irrespective of the query gene list.

We downloaded the gene/GO annotations from Ensembl and, to reduce this ontological uncertainty bias, we traversed the ontologies' directed acyclic graphs (DAGs) depth-first and assigned the genes annotated to each category recursively to their ancestral categories, resulting in a gene/GO mapping which is 'cascaded' up to the root node according to the structure of each ontology's DAG (Supplementary Figure 6.4). This has the effect of reducing the influence of the genes of more uncertain function at the top of the ontologies because these high-level terms now also contain the genes from lower, more specific terms.

We additionally downloaded the annotation of genes to Reactome pathways (Jassal et al. 2020) from Ensembl as another source of functional information on which to perform enrichment analyses. While the pathways in Reactome are organised hierarchically in a similar manner to the GO ontologies, the genes annotated to any particular pathway by Ensembl are also annotated to the parent pathways and so these Reactome annotations do not suffer from the equivalent of GO ontological uncertainty bias.

We created a miRNA functional enrichment pipeline that avoids the miRNA targeting and ontological uncertainty biases described above. The pipeline first converts the lists of genes assigned to each cascaded GO category or Reactome pathway into lists of miRNAs that are experimentally confirmed to target the genes, using the results of experiments curated by the online resource miRTarBase (Chou et al. 2018). The pipeline then performs the hypergeometric over-representation test on the miRNAs of interest directly against the experimentally confirmed targeting miRNAs for each category or pathway, followed by correction for multiple testing.

We processed the miRNA sets identified at the start of this section through the pipeline and summarised the resulting enriched GO terms and Reactome pathways by counting the number of times various cancer-related keywords (Supplementary Table 6.9) occurred in the enriched terms and pathways (Table 3.1).

Description	**Number of enriched terms (median depth)**				**Number of enriched terms containing functional keywords**				
	CC	**MF**	**BP**	**Reactome**	**Metastasis**	**Cell cycle**	**Expression**	**Signalling**	**Development**
Seed/locus cluster miRNAs									
X (118 - 140)	27 (4)	20 (7)	1,017 (5)	192	17	54	39	75	28
X (146 - 147)	7 (5)	1 (4)	43 (4)	43	-	-	-	2	-
1 (220 - 226)	1 (5)	-	-	-	-	-	-	-	-
X (151 - 152)	-	-	3 (5)	-	-	-	-	-	-
13 (50 - 52)	-	-	173 (6)	-	2	-	1	2	3
3 (159 - 160)	-	49 (5)	-	1	-	-	1	-	-
Multi-precursor seed/locus miRNAs									
Complete loss	187 (4)	281 (6)	1,645 (5)	209	21	68	35	93	38
Partial loss	527 (4)	886 (5)	3,847 (6)	665	65	140	95	269	56
Unaltered	461 (4)	770 (5)	3,649 (6)	635	63	136	92	256	54
Gain	276 (4)	494 (5)	3,482 (6)	522	54	131	87	229	59
Redundant seed/locus miRNAs									
Complete loss	302 (4)	304 (5)	-	-	-	-	-	-	-
Partial loss	648 (4)	870 (5)	2,660 (5)	343	38	67	48	146	81
Unaltered	623 (4)	829 (5)	2,622 (5)	367	33	88	52	148	82
Gain	430 (4)	491 (4)	3,221 (6)	498	46	110	74	222	45

Table 3.1 - Seed/locus clusters are enriched for cancer-related processes

The enriched GO terms and Reactome pathways for various sets of miRNAs. The seed/locus cluster descriptions show the genomic location as chromosome name followed by the start and end locations in millions of nucleotides in brackets. Seed/locus clusters discussed further in the main text are highlighted with a grey background. The number of enriched GO terms are shown in three columns: CC = cellular compartment, MF = molecular function and BP = biological process, with the number of enriched terms followed by the median term depth in brackets. The Reactome column lists the number of enriched pathways. The functional enrichment columns (metastasis, cell cycle, expression, signalling and development) show the number of enriched GO terms at the median term depth and the Reactome pathways containing at least one keyword for each cancer-related functional category (Supplementary Table 6.9). There were no seed clusters, single-precursor seed/loci or non-redundant seed/loci with any enriched terms.

The largest seed/locus cluster with enriched terms extends over approximately 22 Mb on chromosome X, is partially lost in 13 cell lines and is enriched for the most biological process terms and pathways in this particular analysis, with an emphasis on terms and pathways related to signalling and the cell cycle (Table 3.1). Notable miRNAs in this region on chromosome X include the *mir-106a~363* oncomir cluster, a paralog of the well-known *Oncomir-1* on chromosome 13.

The other seed/locus cluster with multiple enriched cancer-related terms, on chromosome 13, is characterised by partial losses in 9 cell lines and contains *mir-15a* and *mir-16-1*, which in normal cells are apoptosis-related tumour suppressors which target *BCL-2* (Cimmino et al. 2005).

There were no enriched terms or pathways at all for any of the 36 seed clusters, any of the single-precursor seed/loci or for any of the non-redundant seed/loci (Table 3.1). However, in the single/multi-precursor and non/redundant seed/locus analyses we have only divided the seed/locus CNVs into 8 different groups and we note that the stratifications with no enriched terms at all (single precursor and non-redundant seed/loci) comprise the majorities of the CNVs in those analyses (Figure 3.6A and Figure 3.7B). This suggests that the sensitivity of our enrichment method is greatly reduced for larger groups of miRNAs, because of the inclusion of the majority of terms and pathways due to pleiotropic miRNA targeting of protein-coding genes, and hence the relative enrichment of none.

3.4.5 Partial loss of miRNA biogenesis pathway genes suggests global miRNA depletion

We determined that multi-precursor seed/loci are lost more often than expected (Figure 3.6A), suggesting that cancer cells are selecting for the wider miRNA derepression that the loss of multiple miRNAs implies (Kumar et al. 2007; Lin and Gregory 2015), and so we wanted to see if we could identify more general global miRNA derepression in the copy number variations in cancer-derived cell lines.

We calculated the CNVs for the main components of the miRNA biogenesis pathway - RNA polymerase II, *Drosha*, *Exportin-5*, *Dicer* and *Argonaute* - along with the CNVs for these complexes' known interaction partners (Figure 3.8).

RNA polymerase II subunits (*RBP1-12*) experience widespread partial loss and occasional complete loss, especially affecting *RBP1*, *RBP5*, *RBP6* and *RBP10* (Figure 3.8). The *RBP11-a/b/c* subunits are expressed from a 2 x 10^5 nucleotide region of chromosome 7 and so the fact that the CNV patterns of *RBP11-b/c* are identical is not surprising, though it is interesting that *RBP11-b/c* are the only RNA polymerase II subunits that are never affected by CNVs (Figure 3.8) unlike the adjacent *RBP11-a*. Further investigation reveals however that the three *RBP11* genes are mainly in regions of chromosome 7 that are unmappable (Table 3.2), with *RBP11-b* and *RBP11-c* unmappable in all cell lines and *RBP11-a* unmappable in 12 out of 55 cell lines because of unmappable regions extending further in those cell lines due to differing readDepth bin sizes and removal of unmappable CNVs (see CNV detection method in section 2.3.1.3).

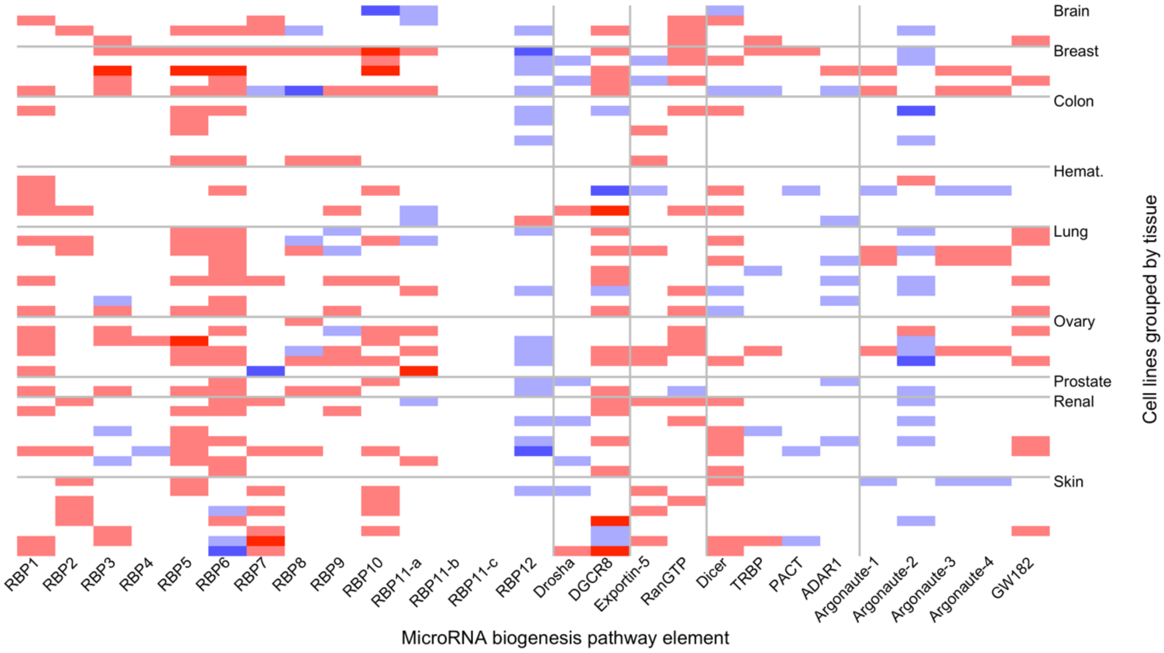


Figure 3.8 - MicroRNA biogenesis gene CNVs indicate global miRNA depletion

The CNVs of miRNA biogenesis genes on one row per cell line, grouped into rows by tissue of origin ('Hemat.' = Hematopoietic) and by columns into, from left to right, genes associated with RNA polymerase II, *Drosha*, *Exportin*, *Dicer* and *Argonaute*. Complete losses are dark red, partial losses are light red, unaltered (or unmappable) are white and gains are shades of blue from light blue for doubled, medium blue for four-fold gain and dark blue for eight-fold gain or more.

The two components of the nuclear Microprocessor complex, *Drosha* and *DGCR8*, are also affected by CNVs across the NCI-60 panel, with partial *Drosha* loss in two cell lines and gains in six cell lines (Figure 3.8). Interestingly, both under and over-expression of *Drosha* occur in cancer with increased expression known to promote cell migration (Sugito et al. 2006; Muralidhar et al. 2011) and decreased expression correlated with globally decreased miRNA expression (Kumar et al. 2007). *DGCR8* is also at least partially lost in the majority and completely lost in three of the NCI-60 cell lines (Figure 3.8), consistent with studies showing that increased tumour growth is correlated with *DGCR8* knockdown (Kumar et al. 2007).

Both *Exportin-5* and the associated protein *RanGTP*, together responsible for the export of precursor miRNAs from the nucleus to the cytoplasm, are also partially lost in the majority of cell lines (Figure 3.8). Down-regulation of these proteins reduces export of miRNA precursors and leads to a build-up of miRNA precursors in the nucleus which interferes with miRNA biogenesis (Melo et al. 2010).

	RBP11-a	*RBP11-b*	*RBP11-c*
Unmappable	12	55	55
Complete loss	1	0	0
Partial loss	6	0	0
Unaltered	30	0	0
Gain	6	0	0

Table 3.2 - **RBP11-a/b/c** ***subunits are mainly in unmappable regions of the genome***

The numbers of cell lines in which the three variants of *RBP11* are either in unmappable regions of the genome or are affected by detectable CNV types.

The next step in miRNA biogenesis, the processing of miRNA precursors into RNA duplexes by *Dicer* in association with *TRBP*, *PACT* and *ADAR1*, is also widely affected by CNVs with *Dicer* partially lost in 16 and gained in four cell lines and *TRBP*, *PACT* and *ADAR1* only sparsely affected by CNVs, mainly by gains (Figure 3.8). These losses of *Dicer* in particular are known to increase the growth rate of cells and are implicated more generally in tumorigenesis (Kumar et al. 2007).

In addition to its role in processing miRNA precursors into RNA duplexes, *Dicer* aids in the formation of the RNA-induced silencing complex (RISC) along with the *Argonaute* and *GW182* proteins (Gregory et al. 2005). In contrast to the majority of the primary elements of the miRNA biogenesis pathway that we've examined here so far, *Argonaute-2* is predominately gained rather than lost, though its cofactor *GW182* is partially lost in several cell lines (Figure 3.8). It is possible that these gains of *Argonaute-2* go some way to compensating for the otherwise fairly consistent loss of the other miRNA biogenesis pathway components, at least in the 40 out of 55 cell lines in which the other miRNA biogenesis-specific components (*Drosha*, *Exportin-5*, *Dicer* and associated cofactors) are affected by losses (Figure 3.8). The three other *Argonaute* proteins in the human genome, *Argonaute-1*, *Argonaute-3* and *Argonaute-4* are adjacent in a 2.6 x10^5 nucleotide region of chromosome 1, which explains their identical CNV profile (Figure 3.8).

Taken together, these various CNVs affecting the miRNA biogenesis pathway are indicative of global miRNA depletion in cancer-derived cell lines, consistent with earlier studies (Kumar et al. 2007; Lin and Gregory 2015), and implying a general relaxation of miRNA constraints, leading to loss of cell differentiation, increased proliferation and tumour growth.

3.4.6 Consistent-gain miRNAs are enriched for cell cycle control and signalling

In the previous chapter we observed that there are distinct hotspots of gains and losses across the cell lines which preferentially occur in gene-dense regions of the genome (Section 2.4.5) and we hypothesised that the regions which are mostly gained or mostly lost in the majority of cell lines might be under selection pressure because of the miRNAs in these regions.

Accordingly, we defined miRNAs which are mostly gained as those whose seed/locus CNVs across the cell lines have a ratio of losses to gains of 0.05 or less and similarly defined mostly lost miRNAs as those with a gain to loss ratio of less than 0.05 (Figure 3.9). So that we could distinguish between miRNAs meeting these criteria which are affected in few cell lines from those which are affected in many cell lines we further divided the mostly gained and mostly lost miRNAs into four groups each along the ranges of gains (0 – 22) and losses (0 – 44) respectively (Figure 3.9).

To see if the mostly gained and mostly lost miRNAs are functionally related, we used our miRNA functional enrichment method developed above (Section 3.4.4) to analyse each of the mostly gained and mostly lost groups of miRNAs, both as separate groups and as combined groups with miRNAs with fewer gains or losses successively removed.

None of the mostly lost miRNAs were functionally enriched for any GO terms or Reactome pathways, either as separate groups or when combined into all, top 75%, top 50% or top 25% most lost (Table 3.3). The groups of mostly gained miRNAs with between six and ten gains and between six and 21 gains were both enriched for multiple cancer-related GO terms and Reactome pathways, though none of the other combinations of mostly gained miRNA groups were (Table 3.3). Neither group of mostly gained miRNAs were enriched for any cellular compartment or molecular function GO terms (Table 3.3).

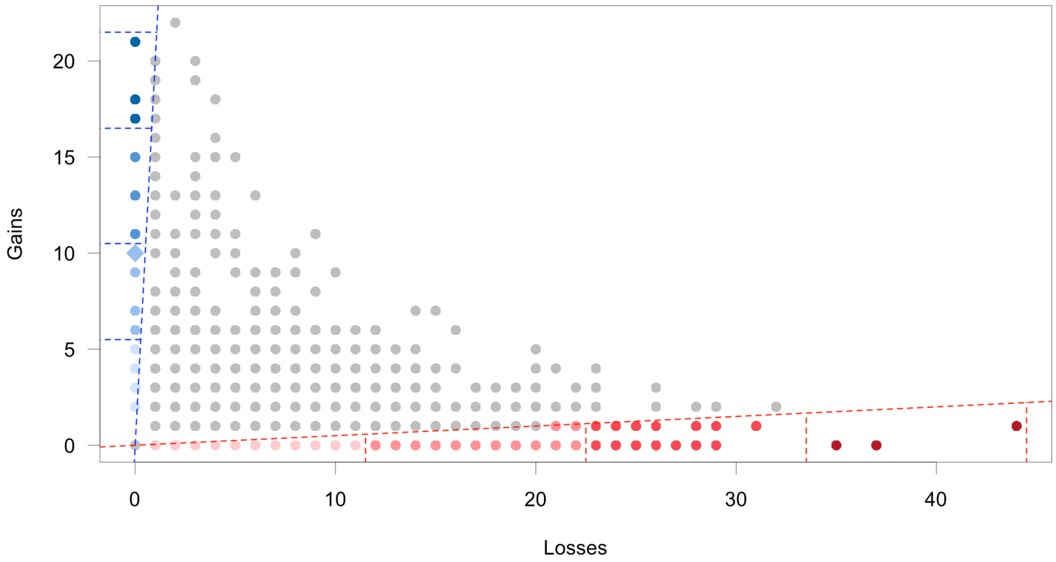


Figure 3.9 - MicroRNAs with consistent gains are enriched for cancer-related processes

The numbers of cell lines in which miRNA seed/locus families are lost (x axis) and gained (y axis) with miRNAs which are mostly gained (losses / gains < 0.05) highlighted in blue and miRNAs which are mostly lost (gains / losses < 0.05) highlighted in red. The dashed blue line at y = 20x is the cut-off for mostly gained and the red dashed line at y = x / 20 is the cut-off for mostly lost. The ranges of mostly gained and mostly lost are further divided into quarters with dashed lines and the colouring of the miRNAs in each quarter are increasingly saturated to indicate increasing number of gains and losses. The blue diamond at 10 gained and 0 lost indicates *mir-15b/mir-16-2*.

Description	Number of enriched terms (median depth)				Number of enriched terms containing functional keywords				
	CC	MF	BP	Reactome	Metastasis	Cell cycle	Expression	Signalling	Development
Mostly gained miRNAs									
6 – 10 gains	-	-	1,737 (6)	379	25	106	47	121	23
6 – 21 gains	-	-	297 (6)	150	5	63	27	39	2
Mostly lost miRNAs									
-	-	-	-	-	-	-	-	-	-

Table 3.3 - Enriched terms and pathways for mostly gained miRNAs

The enriched GO terms and Reactome pathways for mostly gained miRNAs. The number of enriched GO terms are shown in three columns: CC = cellular compartment, MF = molecular function and BP = biological process, with the number of enriched terms followed by the median term depth in brackets. The Reactome column lists the number of enriched pathways. The functional enrichment columns (metastasis, cell cycle, expression, signalling and development) show the number of enriched GO terms at the median term depth and Reactome pathways containing at least one keyword for each cancer-related functional category (Supplementary Table 6.9). There were no other groups of mostly gained miRNAs and no groups of mostly lost miRNAs with any enriched GO terms or Reactome pathways.

Notable among the miRNAs which are mostly gained and enriched for cancer-related processes are *mir-15b* and *mir-16-2* on chromosome 3, which are gained in ten cell lines but lost in none (Figure 3.9). We saw the paralogs of these miRNAs, *mir-15a* and *mir-16-1*, in the seed/locus CNV profile cluster on chromosome 13 (Table 3.1) but, interestingly, the chromosome 13 paralogs are partially lost in nine cell lines and gained in only three. The losses of *mir-15a* and *mir-16-1* on chromosome 13 are potentially compensated for by gains of *mir-15b* and *mir-16-2* on chromosome 3 in three of the cell lines - RPMI-8226, OVCAR3 and UACC62. Both *mir-15/16* clusters have several upstream *TP53* binding sites and have also been confirmed to target the *TP53* 3' UTR (Fabbri et al. 2011), forming a feedback regulatory loop, in addition to targeting *BCL2* and hence acting as tumour suppressors in normal cells (Cimmino et al. 2005).

3.4.7 Target-increasing miRNA/target interactions are enriched for signalling

In order to investigate the net effects of CNVs on motifs involving miRNAs, such as the mir-*15/16* and *TP53* feedback loop identified in the previous section, it is necessary to understand the interplay between miRNA and target CNVs. We used the dataset of strongly functional miRNA/target interactions (MTIs) curated by miRTarBase (Chou et al. 2018) to match miRNAs to their targets and we added the CNVs for each MTI's miRNA and target in each cell line so that we could investigate the correlation between miRNA and target CNVs on an NCI-60 panel-wide basis.

The MTIs' target CNVs extend over a greater range of gains than the miRNA CNVs, up to a maximum of 390 copies for DNA mismatch repair protein *MSH3*, unlike the miRNA CNVs which have a maximum copy number of just 19 copies for *mir-301a* (Figure 3.10A). The bulk of MTIs have both miRNA and target CNVs which vary between partial loss and five copies (Figure 3.10A), with outliers extending to complete loss as well as the aforementioned maximum gains.

Focusing on the bulk of the MTIs, between partial loss and five copies, there is a striking pattern of MTIs with the same CNV type for both miRNA and target between the mean loss and gain thresholds, primarily the MTIs where both miRNA and target are unaltered (Figure 3.10B).

Separating out the MTIs with the same CNV type for both miRNA and target, or consistent-CNV MTIs, clearly shows that there are additional groups of consistent-CNV MTIs where both miRNA and target are gained (above the mean gain threshold in both dimensions) and where both are lost (below the mean loss threshold in both dimensions) (Figure 3.10C).

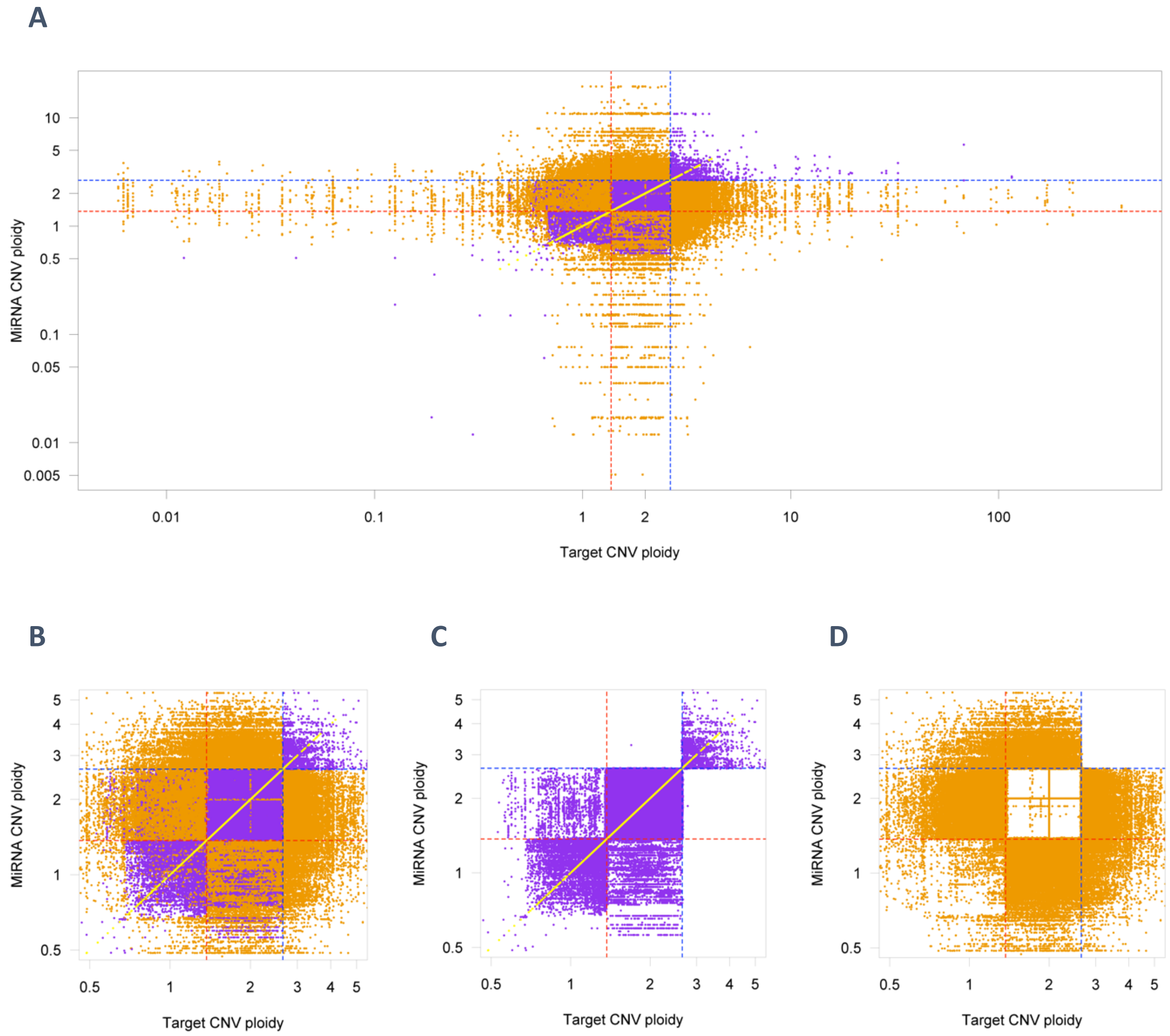


Figure 3.10 - The majority of miRNAs and their targets have the same CNV type

(**A**) The copy numbers of miRNAs (y axis) and their targets (x axis) for all strongly functional verified miRNA/target interactions (MTIs) in all cell lines. Both axes are log scale. The mean gain threshold for all cell lines is shown as blue dashed lines and the mean loss threshold is shown as red dashed lines. MTIs where the miRNA CNV type is different to the target CNV type are shown as orange dots, MTIs where the miRNA CNV type is the same as the target CNV type are purple and MTIs where the miRNA copy number is identical to the target copy number are yellow. (**B**) Zoomed plot to show the main bulk of MTIs. (**C**) Zoomed plot to show just the MTIs with the same CNV type as well as MTIs with identical miRNA and target copy numbers. (**D**) Zoomed plot to show just the MTIs with different CNV types.

There are also sparser regions of consistent-CNV MTIs where the miRNA copy number is less than the mean loss threshold and the target copy number is between the mean loss and gain thresholds, which nevertheless have the same CNV type: these are the haploid miRNAs (on chromosome X and Y in male-derived cell lines) which target diploid protein-coding genes, and so both miRNA and target can have a CNV type of ‘unaltered’ despite having different miRNA and target copy numbers (Figure 3.10C). There is a similar group of consistent-CNV MTIs where a diploid miRNA with a CNV type of unaltered targets a haploid protein-coding gene which is also unaltered, these lie between the mean loss and gain thresholds in the miRNA CNV dimension but below the mean loss threshold in the target CNV dimension (Figure 3.10C).

Additionally, there is a straight line of MTIs with identical copy numbers for both miRNA and target, caused by MTIs where the miRNA and protein-coding target are in the same genomic region and so are affected by the same CNVs (Figure 3.10C). There are no consistent-CNV MTIs which are above the mean gain threshold in one dimension but below the mean loss threshold in the other dimension (Figure 3.10C).

When we consider only the MTIs with different CNV types for miRNA and target, the inconsistent-CNV MTIs, the regions of miRNA/target CNV space which had lots of consistent-CNV MTIs are relatively sparse (Figure 3.10D), though the consistent-CNV and inconsistent-CNV MTIs overlap to an extent. In the central region between the mean loss and gain thresholds in both dimensions there are two crossing straight lines of inconsistent-CNV MTIs at a copy number of exactly two, these are the inconsistent-CNV MTIs where one or both of the miRNA or target are in an unmappable diploid region of the genome in that particular cell line and so are assigned the expected diploid copy number as a default since we have no evidence that they are altered from the norm (Figure 3.10D). There are no inconsistent-CNV MTIs which are above the mean gain threshold in both dimensions (Figure 3.10D).

In total there are 435,001 MTI CNVs across the cell lines, 266,829 (61.3%) have the same CNV type for both miRNA and target CNVs and 4,438 (1.02%) have identical copy numbers for both miRNA and target CNVs. Despite the apparent correlation implied by the structure of the data as visualised (Figure 3.10A), there is no statistically significant correlation between miRNA and target copy numbers (Pearson's product-moment correlation, r(434,999) = 2 x 10^{-3}, p = 0.18). There are however only 5,689 MTI CNVs with either a miRNA or target copy number not between 0.5 and five (the range of the zoomed plots in Figure 3.10B/C/D); without these outliers there is a weak but statistically significant correlation between miRNA and target copy numbers (Pearson's product-moment correlation, r(429,310) = 7.2 x 10^{-3}, p = 2.6 x 10^{-6}).

Not all of these MTI CNVs would be expected to produce a net change in protein level, for example when a gain of both miRNA and target cancels out, and we hypothesised that it is the MTIs which do lead to a net increase or decrease in protein level which would be under selection pressure in cancer-derived cell lines. We observed that the vast majority of consistent-CNV MTIs have both miRNA and target copy numbers between 0.5 and five (Figure 3.10A/B/C) and that only inconsistent-CNV MTIs extend beyond this range in either the miRNA or target CNV dimension (Figure 3.10A). These observations led us to allocate the MTIs into two groups that would be expected to lead to an increase in protein levels and two groups that would be expected to lead to a decrease, corresponding to the four 'arms' of MTIs with excess copy numbers of just one of either miRNA or target (Figure 3.10A, Table 3.4).

'Arm' of Figure 3.10A	Target copy number	miRNA copy number	Expected change in protein
Right	> 5	< mean gain threshold	Increase
Lower	> mean loss threshold	< 0.5	Increase
Left	< 0.5	> mean loss threshold	Decrease
Upper	< mean gain threshold	> 5	Decrease

Table 3.4 - Defining MTIs which are expected to have a net effect on protein levels

The parameters for defining the ranges of target and miRNA copy numbers to divide the MTIs into subsets which would be expected to have a net effect on protein levels.

We extracted the unique mature miRNAs for the MTIs expected to lead solely to an increase in protein levels (n = 136) and those expected to lead solely to a decrease in protein levels (n = 119) from the MTI groups expected to have a net effect on protein levels and, to see if they are functionally related, we analysed these miRNAs with our functional enrichment pipeline (Section 3.4.4).

The miRNAs for MTIs expected to lead to an increase in protein levels are enriched for biological process GO terms and Reactome pathways related to cancerous processes, as are the miRNAs for MTIs expected to lead to a decrease in protein levels (Table 3.5). Both sets of miRNAs are enriched for more signalling terms than the other categories but, while the miRNAs from MTIs expected to lead to an increase in protein levels have more enriched terms for each functional category than the miRNAs expected to decrease protein levels, the higher number of terms is not statistically significant (X^2 (4, N = 605) = 4.70, p = 0.32), suggesting that both groups of miRNAs are equally relevant to cancerous processes (Table 3.5).

Expected protein change	Number of enriched terms (median depth)				Number of enriched terms containing functional keywords				
	CC	MF	BP	Reactome	Metastasis	Cell cycle	Expression	Signalling	Development
Increase	-	-	2,276 (5)	398	31	80	48	159	78
Decrease	-	-	1,654 (5)	161	20	56	24	75	34

Table 3.5 - Enriched GO terms and Reactome pathways for net protein change MTIs

The enriched GO terms and Reactome pathways for miRNAs involved in MTIs expected to have a net effect on protein levels. The number of enriched GO terms are shown in three columns: CC = cellular compartment, MF = molecular function and BP = biological process, with the number of enriched terms followed by the median term depth in brackets. The Reactome column lists the number of enriched pathways. The functional enrichment columns (metastasis, cell cycle, expression, signalling and development) show the number of enriched GO terms at the median term depth and Reactome pathways containing at least one keyword for each cancer-related functional category (Supplementary Table 6.9).

3.4.8 C-MYC *repression of* TP53 *could be required for sustained* Oncomir-1 *activation*

We have previously identified groups of miRNAs which are enriched for cancer-related processes, many of which are oncomirs and tumour suppressors, frequently including *Oncomir-1* and its two paralogs (Table 3.1). Furthermore, there are groups of paralogs, *mir-15/16* for example, where one group is mostly gained (Figure 3.9) and the other mostly lost (Table 3.1) despite having almost identical seeds. We decided therefore to narrow our focus to oncomirs and tumour suppressor miRNAs and curated a list of these miRNAs by literature search. We then calculated the CNVs of these oncomirs and tumour suppressor miRNAs across the NCI-60 cell lines (Figure 3.11).

There are no obvious tissue-based row patterns in the heatmap, conforming to the overall CNV patterns we've seen so far, but there are several column clusters of between three and six miRNAs with identical CNV profiles across the cell lines (Figure 3.11, clusters 'a' to 'f'). The largest two clusters are *Oncomir-1* and its main paralog on chromosomes 13 and X (Figure 3.11, clusters 'd' and 'a' respectively). *Oncomir-1*'s other partial paralog on chromosome 7 is also one of the oncomir clusters (Figure 3.11, cluster 'f'). The three *Oncomir-1* paralog clusters, together with the other three highlighted clusters (Figure 3.11, clusters 'b', 'c' and 'e'), all oncomirs or capable of behaving as either oncomirs or tumour suppressors depending on the context, are all polycistronic and so it is unsurprising that they have identical CNV profiles. The other 14 clusters of two miRNAs each are also either polycistronic or from adjacent loci, again explaining their clustering (Figure 3.11).

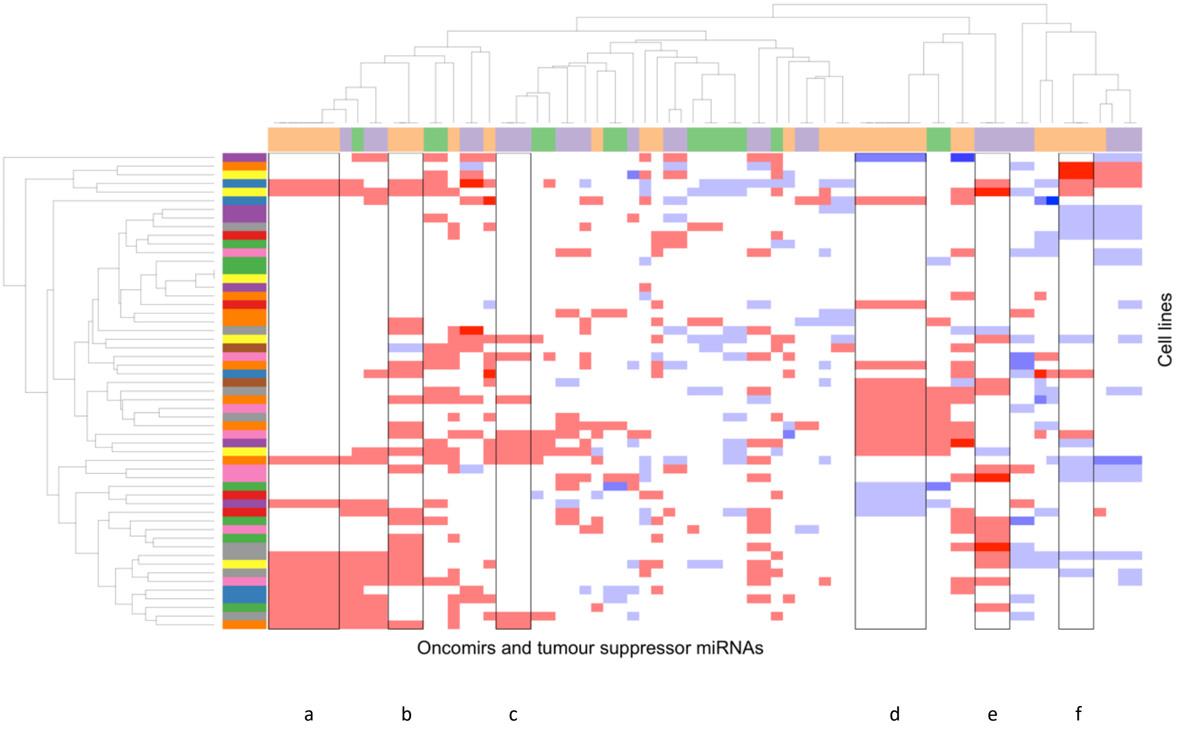


Figure 3.11 - Oncomir and tumour suppressor CNV profile clusters include **Oncomir-1**

Heatmap indicating oncomir and tumour suppressor miRNA CNVs, clustered by CNV profile across the NCI-60 panel (columns) and by cell line (rows). The row colours indicate the tissue of origin and the column colours indicate whether the miRNA is an oncomir (orange), a tumour suppressor (green) or both (purple). Losses are red, unaltered or unmappable miRNAs are white and gains are blue. Clusters discussed in the text are outlined in black.

Focusing now specifically on the CNVs of *Oncomir-1* on chromosome 13 and its paralogs on chromosomes X and 7, we can see that while the two main paralogs of six oncomirs each are partially lost in a total of 25 cell lines, they are never lost at the same time (Figure 3.12). The chromosome X *Oncomir-1* paralogs are partially lost only in female-derived cell lines along with, interestingly, the partial loss of *Xist*, the primary non-coding RNA involved in X-inactivation in placental mammals, in every case (Figure 3.12). This occurs despite *Xist* at approximate location on chromosome X of 74 Mb being affected by different CNVs to the *Oncomir-1* paralog at 134 Mb (Figure 3.12) and implies that the partial loss of *Xist* might lead to derepression of X transcription and may lead to compensation for the partial loss of the *Oncomir-1* paralog to some extent. Taken together, these results are suggestive of selection pressure to retain at least one of the two main *Oncomir-1* paralogs in cell lines.

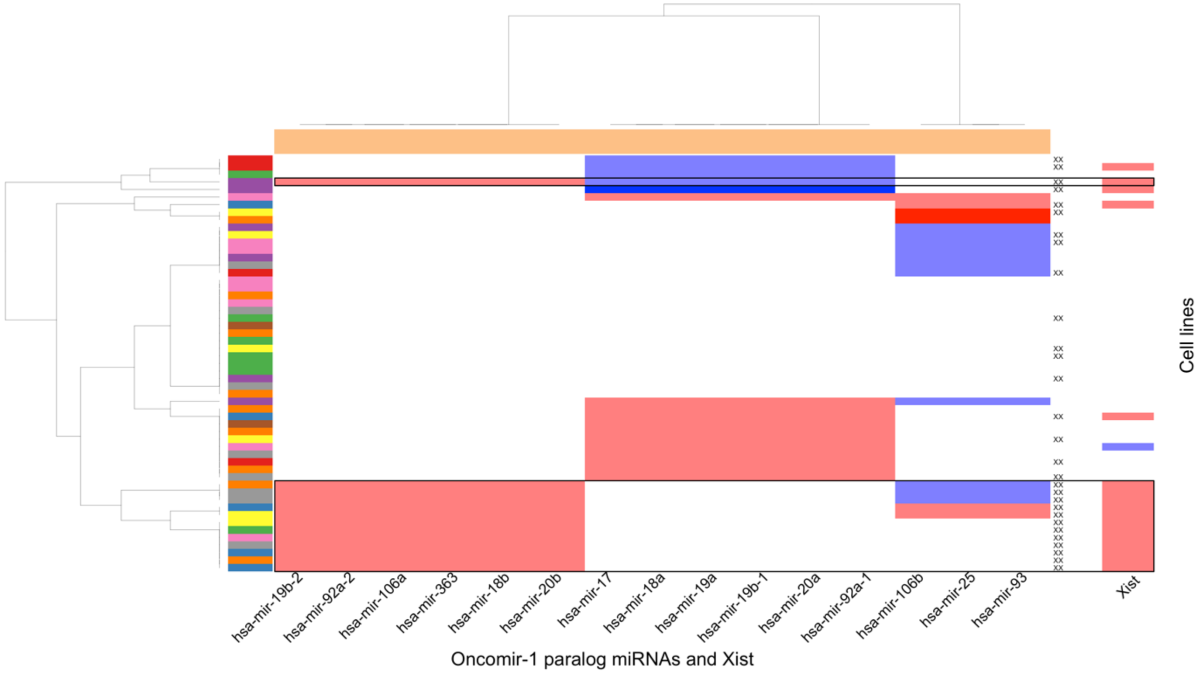


Figure 3.12 -* Oncomir-1 *CNVs imply selection pressure to retain* Oncomir-1 *function

Heatmap indicating *Oncomir-1* paralog and *Xist* CNVs, clustered by *Oncomir-1* CNV profile across the NCI-60 panel (columns) and by cell line (rows). The row colours indicate the tissue of origin and the column colours indicate whether the gene is an oncomir (orange) or non-coding RNA (white). Losses are red, unaltered or unmappable genes are white and gains are blue. The cell lines derived from cancers in female patients are marked with ‘XX’ to the right of the oncomir-1 paralogs. The three *Oncomir-1* column clusters are, from left to right, *Oncomir-1* paralog on chromosome X, *Oncomir-1* on chromosome 13 and *Oncomir-1* partial paralog on chromosome 7. The *Xist* CNVs on the right are not influencing the heatmap’s column dendrogram. The cell lines outlined in black are those with partial loss in both the *Oncomir-1* paralog on chromosome X and with partial loss of *Xist* (also on chromosome X).

The *Oncomir-1* paralogs form complex feedback loops of repression and activation with various oncogenic transcription factors (Mogilyansky and Rigoutsos 2013) and so understanding why the two main *Oncomir-1* paralogs are never lost at the same time requires knowledge of the network of *Oncomir-1*’s interaction partners and how these are affected in cancer-derived cell lines. We derived an interaction network for *Oncomir-1* and its activating and inactivating transcription factors based on data from an experiment in the ENCODE project that mapped the transcription factors that bind to miRNAs (Gerstein et al. 2012) and integrated this with verified miRNA-to-transcription factor interactions from miRTarBase (Figure 3.13A). The specificity of the interactions of the oncomir-1 paralogs’ miRNAs is determined by their seed sequences, which vary considerably in the three *Oncomir-1* paralogs, and which can be grouped into the *miR-17*, *miR-18*, *miR-19* and *miR-92* seed families (Figure 3.13A). The two main *Oncomir-1* paralogs on chromosomes 13 and X each contain members of all four seed families, albeit at varying dosages, but the chromosome 7 partial *Oncomir-1* paralog only contains members of the *miR-17* and *miR-92* families and so interacts with fewer of the transcription factors (Figure 3.13A).

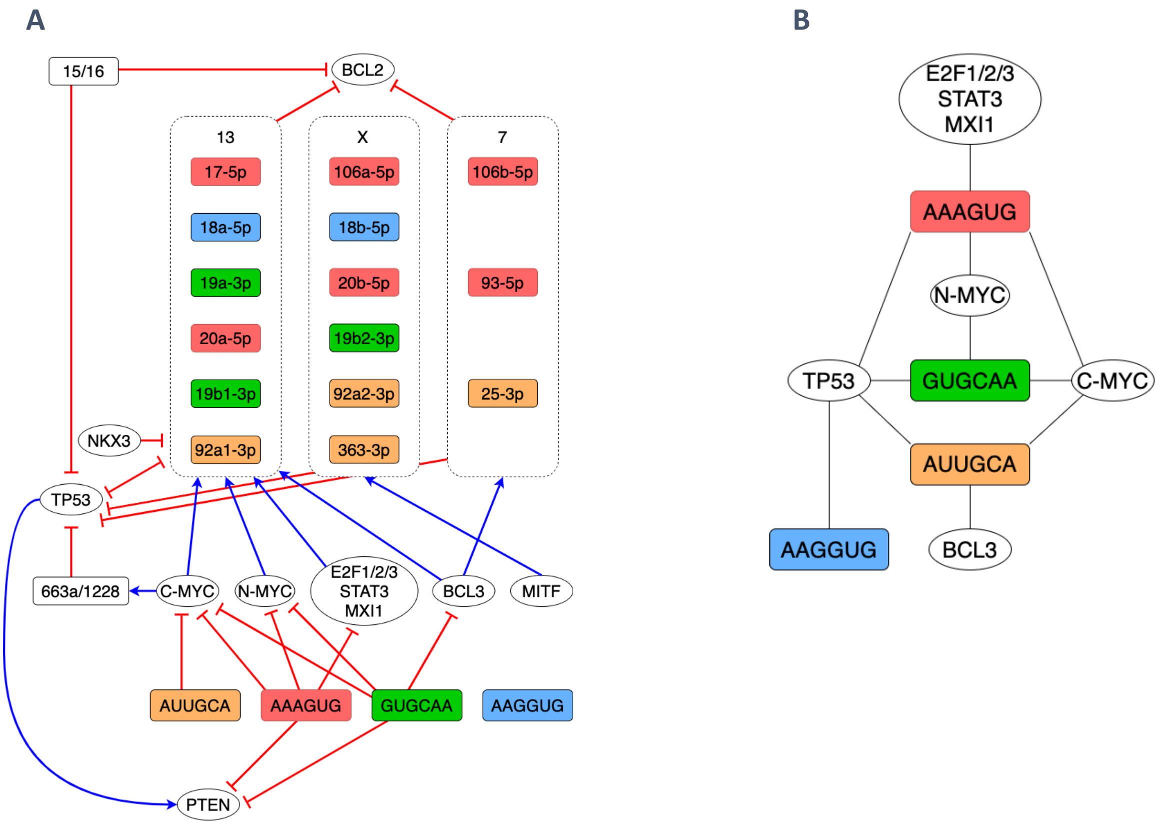


Figure 3.13** - Oncomir-1 **seed families form feedback loops with transcription factors

(**A**) A map of selected interactions between members of *Oncomir-1* paralogs, other oncomirs and related transcription factors. Transcription factors are shown as ellipses and miRNAs as rectangles. The mature miRNAs are grouped by dotted outlines into the three oncomir-1 paralog polycistronic genes, labelled at the top with the chromosome name, and the colours of the mature miRNAs match to the seed sequence shown in the bottom line of miRNA seeds. Activation is shown as blue arrows based on data from ENCODE and repression is shown as red crossbars based on data from ENCODE and miRTarBase. If a red repression line extends from a seed to a transcription factor, then all miRNAs with that specific seed target the transcription factor. If a red repression line extended from a polycistronic gene to a transcription factor, then all miRNAs in that gene target the transcription factor. (**B**) Auto-regulatory loops where a miRNA represses a transcription factor which also activates the miRNA. The miRNA seeds are shown as coloured rectangles and transcription factors are shown as ellipses. A line between a miRNA seed and a transcription factor shows the existence of an auto-regulatory loop between miRNAs with that seed and the transcription factor.

The transcription factor-to-miRNA interactions from the ENCODE data show not only that multiple oncogenic transcription factors and cofactors such as *C-MYC*, *BCL3* and the *E2Fn* family bind upstream of the *Oncomir-1* clusters and activate transcription but also show that tumour suppressors such as *NKX3* and *TP53* can inactivate *Oncomir-1* (Figure 3.13A). These interactions between transcription factors and miRNAs form feedback loops where a miRNA represses a transcription factor which in turn activates the miRNA, which we will call ‘auto-regulatory loops' (Figure 3.13B). The four different seed families present in the *Oncomir-1* paralogs form auto-regulatory loops with the related transcription factors to differing degrees, ranging from a single feedback loop between *TP53* and members of the *miR-18* seed family to the more central *miR-17* seed family which has direct auto-regulatory feedback loops with all the transcription factors except the coactivator *BCL3* (Figure 3.13B).

To understand therefore how these interaction networks are perturbed in cancer cell lines it is necessary to integrate detailed seed dosage patterns for the *Oncomir-1* paralogs together with the CNVs of the related transcription factors. Based on the CNVs of the *Oncomir-1* paralogs we calculated the actual dosage of each of the four seed families in each cell line and derived an effective 'seed dosage CNV' from the comparison of the actual dosage to the expected dosage (Figure 3.14). We can see from the single cell line where the chromosome X paralog is lost and the chromosome 13 paralog is gained that the two main *Oncomir-1* paralogs partially compensate for each other, leading to no change in the dosage of the *miR-17* and *miR-18* seed families (Figure 3.14).

Gains of the chromosome 7 paralog can also partially compensate for seed dosage loss of either main *Oncomir-1* paralog by compensating for losses of members of the *miR-17* and *miR-92* seed families in the other paralogs (Figure 3.14). Even partial loss of both chromosome 13 and chromosome X paralogs together would lead to a halving of the *miR-18* and *miR-19* seed families' dosage (Figure 3.13A, Supplementary Figure 6.5), which would lead to considerable derepression of *TP53* (Figure 3.13B), which is possibly why this combination of losses never occurs in the actual cell line CNVs (Figure 3.12 and Figure 3.14). The absence of cell lines where both main *Oncomir-1* paralogs lose a copy implies that the *Oncomir-1* paralogs together are effectively haploinsufficient.

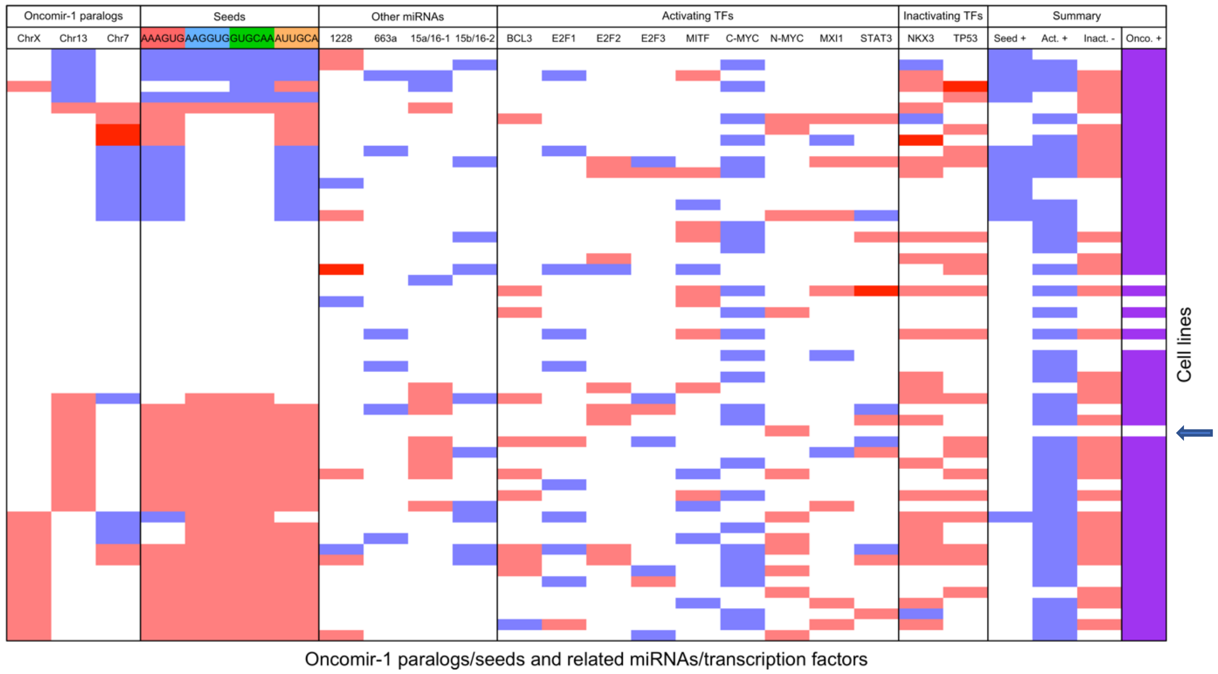


Figure 3.14 - CNVs lead to an increase in* Oncomir-1 *seed dosage in nearly all cell lines

CNVs of *Oncomir-1* paralogs, other oncomirs and related transcription factors with partial losses in light red, complete losses in dark red, unaltered in white and gains in blue. The rows are the CNVs in each cell line, shown in the same order as Figure 3.12 for comparison, and the columns are grouped from left to right into CNVs for the *Oncomir-1* paralogs, the seeds contained in the paralogs (with the same seed colours as Figure 3.13), other oncomirs, the activating and inactivating transcription factors and summary columns. The seed CNVs are calculated by comparing the exact actual seed dosage, based on the *Oncomir-1* paralogs' CNVs, to the expected seed dosage in each cell line and are blue if increased and red if decreased. The activating and inactivating transcription factors (TFs) are based on data from ENCODE and are blue if gained and red if lost. The 'Seed +' summary column is blue if any *Oncomir-1* seed is gained in that cell line and white otherwise. The 'Act. +' summary column is blue if any activating transcription factor is gained in that cell line and the 'Inact. -' summary column is red if any inactivating transcription factor is lost in that cell line. The final 'Onco. +' summary column integrates the preceding summary columns and is purple if any of the preceding summary columns is red or blue and indicates whether the cell line's CNVs would be expected to lead to an increase in expression of any *Oncomir-1* paralog seed families. The arrow on the right indicates prostate cancer cell line DU145.

The most striking pattern in the activating transcription factor CNVs is that *C-MYC* is gained in 25 of the cell lines but never lost (Figure 3.14). *E2F1* is gained in nine cell lines and lost only once, *N-MYC* on the other hand is lost in 12 cell lines but never gained and the other activating transcription factors have sparser patterns of both gain and loss (Figure 3.14). The frequent gains of *C-MYC* are interesting because this transcription factor, as well as directly activating the chromosome 13 *Oncomir-1* paralog, also activates *mir-663a* and *mir-1228*, which then repress *TP53* and so reduce *TP53*'s repression of the chromosome 13 *Oncomir-1* paralog (Figure 3.13A). It is possible that gain of *C-MYC* is therefore required to relax *TP53* repression of *Oncomir-1* via *mir-663a* or *mir-1228* before activation of *Oncomir-1* can be initiated or sustained. There are no cell lines where both *mir-663a* and *mir-1228* are completely lost (Figure 3.14), meaning that this potential mechanism is available in all the cell lines, and additionally *mir-663a* and *mir-1228* are directly gained in six and three cell lines respectively (Figure 3.14), directly repressing *TP53* further. The *mir-15a/16-1* and *mir-15b/16-2* paralogs also repress *TP53* and there are gains of one or both of these oncomir paralogs in 13 out of 55 cell lines (24%) (Figure 3.14).

In addition to the indirect repression of *TP53* via *mir-663a/1228* caused by *C-MYC* gains and the direct repression by the *mir15/16* paralogs, *TP53* is also directly affected by partial losses in 18 cell lines and completely lost in another (Figure 3.14). The other transcription factor which in normal cells represses the chromosome 13 *Oncomir-1* paralog, *NKX3*, is also lost in 21 cell lines and completely lost in another (Figure 3.14). These CNVs together lead to the derepression of *Oncomir-1* in a total of 31 out of 55 cell lines (Figure 3.14), in addition to the other direct oncogenic effects of *TP53* loss, such as a reduction in the ability to initiate apoptosis in cancerous cells (Bernstein et al. 2002).

In summary, none of the four miRNA seed families are completely lost and where any are partially lost there is almost always a compensatory gain of an activating transcription factor or loss of an inactivating transcription factor, the one exception being prostate cancer cell line DU145 (Figure 3.14, arrowed). There is a net increase in dosage of at least one seed family by direct CNV gain in 13 out of 55 cell lines (24%), indirectly by gain of an activating transcription factor in 41 out of 55 cell lines (75%) and indirectly by loss of an inactivating transcription factor in 31 out of 55 cell lines (56%) (Figure 3.14). In total, 50 out of 55 cell lines (91%) have CNVs that would be expected to lead to an increase in the expression of at least one of the four miRNA seed families found in the *Oncomir-1* paralogs and so lead to a decrease in *TP53* (Figure 3.14).

3.5 Discussion

Tumour cells and cell lines experience widespread changes to gene dosage caused by copy number variations (Iafrate et al. 2004; Torres et al. 2008; Henrichsen et al. 2009; Yang et al. 2016) and yet, unlike normal cells, not only survive but even thrive (Sheltzer and Amon 2011), which implies the existence of mechanisms which buffer the ordinarily deleterious effects of these dosage changes. In this study we've examined miRNA-mediated post-transcriptional buffering of the protein-coding gene dosage changes caused by CNVs in the NCI-60 cancer-derived cell line panel. We used a new method to avoid double-counting miRNAs with CNVs which are identical across the cell lines merely because of their genomic proximity and were able therefore to determine accurately which cancer-related processes involving miRNAs are consistently affected in cell lines, leading us to propose a novel *TP53* repression mechanism mediated by *mir-17~92*, better known as *Oncomir-1*.

The large clusters of miRNA precursors with identical CNVs across the NCI-60 panel (Figure 3.1A) are caused primarily by the genomic proximity of the precursors in each cluster, which means that they experience the same CNVs. The dosage changes of each distinct miRNA seed are also confounded in a similar manner (Figure 3.1B, Figure 3.2) and so it is clearly necessary to remove or at least reduce this bias before it is possible to see if any miRNAs are consistently perturbed in cell lines. Our method of combining miRNA precursors into loci which are expected to be affected by similar CNVs due to their proximity and then splitting these loci into distinct 'seed/locus families' (Figure 3.4) allows us to detect groups of miRNAs which have identical CNVs for reasons other than genomic proximity (Figure 3.5C), and which might therefore be evidence of consistent somatic selection pressure on miRNAs in cancer. In total we found 36 such groups of miRNAs which are in distinct loci, affected by different CNVs and which contain miRNAs known to be associated with cancer (Supplementary Table 6.8).

We also found that seed/locus families containing multiple miRNA precursors, such as the *Oncomir-1* paralog on chromosome X mentioned above, are lost significantly more often but have a lower range of gains than seed/locus families with just one precursor (Figure 3.6), consistent with selection pressure to deregulate a wide range of processes (Kumar et al. 2007). Similarly, we observed that the miRNA biogenesis pathway is disrupted in the majority of cell lines (Figure 3.8), which would imply widespread derepression of protein-coding genes caused by global miRNA depletion, changes known to lead to a loss of cell differentiation, the promotion of cell migration and increased tumour growth (Sugito et al. 2006; Kumar et al. 2007; Muralidhar et al. 2011; Lin and Gregory 2015). Surprisingly however, we found that seed/locus families without functional redundancy are lost more than expected (Figure 3.7B), further suggesting that the loss of miRNA regulation of processes is advantageous for tumour cells.

The largest of the clusters of seed/locus family CNVs with enriched GO terms, spread across 22 Mb of chromosome X and with partial losses in 13 cell lines (Table 3.1), contains *mir-106a~363* which is a paralog of the well-known chromosome 13 oncomir cluster *mir-17~92*. The miRNAs in this cluster are enriched for many cancer-related processes, especially those related to signalling and control of the cell cycle (Table 3.1), as is another cluster of frequent losses on chromosome 13 containing *mir-15a* and *mir-16-1* (Table 3.1), which are apoptosis-related tumour suppressors which target *BCL2* (Cimmino et al. 2005). Interestingly, the *mir-15/16* paralogs, *mir-15b* and *mir-16-2* on chromosome 3, are similarly enriched for processes associated with signalling and control of the cell cycle (Table 3.3) but are differently affected by CNVs, being gained in ten cell lines but never lost (Figure 3.9). Also enriched for cancer-related processes are the miRNAs with CNVs which, together with the miRNAs' targets' CNVs, would be expected to lead to large net changes in expression (Figure 3.10, Table 3.5).

A limitation of this study is that we have only considered copy number variations and have not yet incorporated expression data in the NCI-60 cell lines, so we are assuming that gene expression is linearly proportional to gene copy number. While mRNA abundance has been shown to be broadly correlated with CNVs in aneuploid cell lines (Stingele et al. 2012; Dephoure et al. 2014; Zhao and Zhao 2016; Shao et al. 2019), as has protein abundance for the majority of genes (Stingele et al. 2012; Dephoure et al. 2014), approximately 20% of protein-coding genes affected by copy number gains are detected at near disomic levels (Stingele et al. 2012; Dephoure et al. 2014), mainly protein kinases and subunits of macromolecular complexes. This protein buffering is mediated primarily by protein degradation (Stingele et al. 2012; Dephoure et al. 2014), with p62-dependent autophagy a dominant factor (Stingele et al. 2012). Other post-translational buffering mechanisms include the inverse dosage effect (Guo and Birchler 1994), where the formation of complexes is limited by monomer titration caused by the sequestration of excess monomers by incomplete complexes, and by the masking of degradation signals on monomers only when incorporated into the mature complex, leaving the unincorporated monomers vulnerable to protein degradation (Asher et al. 2006).

At a post-transcriptional level, miRNA-mediated tuning of protein expression levels (Sheltzer and Amon 2011) is another frequent cause of protein levels which are decoupled from gene copy numbers, a mechanism which is partially accounted for in our analyses. We also need to more thoroughly investigate the ability of miRNA paralogs to compensate for each other's CNVs, since we've seen that *mir-15/16* paralogs for example have very different patterns of gain and loss despite having broadly the same targets.

Our analyses could be improved therefore by the incorporation of NCI-60 mRNA and miRNA expression levels derived from RNA-seq experiments in addition to protein levels derived from mass spectrometry. Rather than assuming a linear relationship between gene and expressed copy numbers, we would be able to refine our models by using *actual* copy numbers at the transcriptomic, regulatory and proteomic levels. Such data is now readily available, such as those provided by a recent study which integrated RNA-seq data for the NCI-60 cell lines into online resource CellMiner (Reinhold et al. 2019), but there is an important caveat to add regarding the integration of cell line data from different studies. Cancer-derived cell lines are genomically unstable, varying dramatically across laboratories and even between cell line passages within the same laboratory (Kleensang et al. 2016; Liu et al. 2019), and so it would be preferable, though much more expensive, to perform the analyses at each 'omics' layer on the same cell line cultures, to ensure that the genomic, transcriptomic and proteomic data are from the same biological entities and so directly relatable.

In order to understand the conflicting CNV patterns affecting oncomirs and tumour suppressor miRNAs we narrowed our focus to the CNVs of just these miRNAs and found that *Oncomir-1* and its paralog on chromosome X are never lost at the same time (Figure 3.12), suggesting that the expression of the four distinct miRNA seeds in these oncomir clusters is essential for cell lines. In every female-derived cell line in which the chromosome X *Oncomir-1* paralog is lost there are different CNVs also causing partial loss of *Xist* (Figure 3.12), the main long non-coding RNA involved in X-inactivation in placental mammals, leading to possible compensation for the loss of genes on one copy of chromosome X by the reduction of X-inactivation of the other X chromosome. While this implies selection pressure in cancer cells to maintain dosage of X chromosome genes it does not however explain why the two main *Oncomir-1* paralogs are never lost simultaneously. In future studies we intend to investigate whether *Dicer* CNVs coincide with *Xist* CNVs since, in mice at least, *Dicer* deletion also prevents the accumulation of *Xist* (Ogawa et al. 2008).

To understand why the two *Oncomir-1* paralogs are never even partially lost at the same time we constructed a detailed network of the *Oncomir-1* paralogs together with the transcription factors known to activate and inactivate the *Oncomir-1* paralogs (Figure 3.13A). We grouped the distinct *Oncomir-1* miRNA seeds into four families – *miR-17*, *miR-18*, *miR-19* and *miR-92* – and found that they form 'auto-regulatory' feedback loops with their transcription factors (Figure 3.13B). Of these feedback loops, the most central to the network are the negative feedback loops involving *TP53* (Figure 3.13B), which is also repressed by the *mir-15/16* paralogs discussed earlier. As the partial *Oncomir-1* paralog on chromosome 7 can be lost at the same time as either of the main paralogs on chromosomes 13 and X (Figure 3.12) this suggests that rather than the overall *TP53*-repressing seed dosage from all the *Oncomir-1* paralogs, it is the seeds which do not occur on the chromosome 7 paralog which are under the most selection pressure to be maintained in cancer cells, specifically the *miR-18* and *miR-19* seed families, the latter being the second most central miRNA seed in the auto-regulatory loop network (Figure 3.13B). Interestingly, the *miR-19* seed family members are known to specifically target the tumour suppressor *PTEN* (Mu et al. 2009; Olive et al. 2009), which negatively regulates the Akt signalling pathway, indicating that this is one of the many tumorigenic consequences of increased *Oncomir-1* expression.

Further analysis of oncomirs and tumour suppressor miRNAs activated by the same transcription factors as those that activate the *Oncomir-1* paralogs revealed that *C-MYC* activates two more miRNAs which also repress *TP53*: *mir-663a* and *mir-1228* (Figure 3.13A). *C-MYC* is gained in 25 out of 55 cell lines but is never lost (Figure 3.14) and so it is possible that the gain of *C-MYC* is required in those cell lines in order to transiently repress *TP53* via *mir-663a* and/or *mir-1228* before *Oncomir-1* activation can be initiated or sustained, leading in turn to sustained *TP53* repression. It is also possible that *C-MYC*-induced *TP53* repression is limited to specific cancer types since our findings here conflict with recent studies which found that *mir-663a* acts as a tumour suppressor in hepatocellular carcinomas (Zhang et al. 2018) and colon cancer (Kuroda et al. 2017). As well as being activated by *C-MYC*, *Oncomir-1* also represses *C-MYC* in a negative feedback loop (Figure 3.13A) which, since *C-MYC* over-expression can trigger apoptosis (Hoffman and Liebermann 2008), could mean that *Oncomir-1* protects the cancer cell against *C-MYC*-induced apoptosis while preserving the oncogenic trigger from transient *C-MYC* activation.

We intend in future work to model the *Oncomir-1* paralogs and transcription factors (Figure 3.13A) as a system of ordinary differential equations in order to model the system's dynamic response to various perturbations such as CNVs and miRNA-mediated repression, and to investigate whether transient *C-MYC* repression of *TP53* via *mir-663a/1228* could indeed lead to a switch from steady-state *TP53* repression of *Oncomir-1* to the reverse, as our results so far suggest. As an initial experimental test of this hypothesis we will transfect NCI-60 cell line cultures with miRNA 'sponges' (Ebert et al. 2007) to sequester *Oncomir-1*-derived miRNAs. If these cell lines are indeed *Oncomir-1*-dependent, then we would expect to observe widespread apoptosis caused by the restoration of the *TP53* and *PTEN* pathways.

The centrality of *Oncomir-1* and its paralogs to *TP53* repression is further indicated by the balancing of even partial loss of any *Oncomir-1* miRNA seed by the gain of an activating transcription factor or loss of an inactivating transcription factor in all but one cell line (prostate cancer-derived DU145, Figure 3.14). Furthermore, in the vast majority of cell lines (50 out of 55) there are CNVs affecting either *Oncomir-1* or its interacting transcription factors (Figure 3.14) which would be expected to lead to increased *Oncomir-1* expression and consequently to increased *TP53* and *PTEN* repression.

4 Haploinsufficiency explains the heritable dominant disease burden

Statement on previous work

The paralog dating method used in this chapter, Furthest from Singleton (FFS), is based on an original algorithm developed for my MSc dissertation (Reardon 2016) and extensively modified here as described below.

Work carried out during MSc

- Prototype tree building and age allocation algorithms developed.
- Comparison of gene ages resulting from FFS to gene ages resulting from last common ancestor (LCA) and most recent duplication (MRD) algorithms.
- Comparison of gene ages by duplication type (ohnolog, small-scale duplication (SSD) and singleton).
- Comparison of miRNA/target ages based on predicted miRNA/target interactions.

Work carried out during PhD

- Refined FFS tree building algorithm to remove speciation, dubious duplication and non-bifurcating duplication events before building trees, resulting in trees topologically similar to Ensembl's gene trees.
- Presented FFS method as a poster at SMBE 2019.
- Rebuilt ages with latest Ensembl data.
- Repeated comparison of gene ages resulting from FFS, LCA and MRD methods.
- Repeated comparison of gene ages by duplication type.
- Repeated comparison of miRNA/target ages but this time with verified miRNA/target interactions from miRTarBase.
- New analyses of gene ages and haploinsufficiency for terminal paralog pairs and by disease association.
- New analyses of gene ages for sets of genes identified earlier in this thesis.
- New analyses of haploinsufficiency and disease over evolutionary time.

Repeats of results from MSc dissertation with new data

- Figure 4.4 - The choice of paralog dating algorithm greatly influences the gene ages
- Figure 4.5 - The FFS age allocation algorithm applied to the Tetraspanin paralog family
- Figure 4.6 - FFS identifies the ohnologs and separates the miRNA/protein-coding ages
- Figure 4.7 - MicroRNAs mostly target protein-coding genes originating in Vertebrata
- Supplementary Table 6.10 - Unique Ensembl Compara taxon names and approximate ages

4.1 Abstract

The presence in the human genome of multiple copies of ancient non-essential genes associated with heritable genetic diseases is a paradox, since purifying selection would be expected to remove these genes, and yet they are found throughout the metazoa. Understanding the origins and persistence of genes associated with heritable disease is clearly an important aspect of disease aetiology and has been extensively studied. Several related theories based on dosage sensitivity or compensating paralogs have been previously advanced to explain the persistence of non-essential disease genes in the context of whole genome duplications, two rounds of which occurred in our early vertebrate ancestors.

The study of disease over evolutionary timescales requires the accurate dating of the origins of the genes implicated in disease but the methods of determining the approximate age of genes developed so far have suffered from characteristic biases towards ancient or recent taxa, depending on the approach used. We have developed a novel method of dating genes which takes into account likely similarity to ancestral function as well as the topology of evolutionary events inferred from the cross-species gene tree to avoid these biases, leading to a more nuanced perspective on the evolution of inheritable diseases.

The increased temporal resolution afforded by our new method allows us to clearly distinguish between evolutionary events such as the inheritance of ancient core transcriptional machinery and the more recent miRNA-specific processes necessary for metazoan tissue differentiation and stable body plans. We show that more newly created genes associated with recessive disease than dominant disease are retained until the time of the whole genome duplications, after which there is an excess of dominant disease genes caused by biased retention of haploinsufficient ohnologs and a subsequent paucity of new haploinsufficient genes. Together with our observation that far more of the ohnologs are subsequently duplicated than previous shown, this leads us to propose that the haploinsufficiency of the retained ohnologs alone is a more parsimonious explanation for the retention of the dominant disease-associated genes than general dosage sensitivity or the masking of deleterious mutations by compensating paralogs.

4.2 Introduction

Following on from our investigations in the preceding chapters into copy number variation and the roles of miRNAs in buffering the somatic dosage changes that occur in cancer-derived cell lines, we wanted to also investigate the dosage compensation that occurs on evolutionary timescales and relate this to inheritable diseases.

Recent large-scale whole exome and whole genome sequencing efforts have made it clear that the modern human population harbours many structural variants (SVs) greater than 50 nucleotides long, including gene dosage-affecting copy number gains and losses (Karczewski et al. 2020). Surprisingly high variation was discovered by the Genome Aggregation Database (gnomAD) project (Karczewski et al. 2020) - the median human genome was found to have more than seven thousand SVs, mostly small and rare (allele frequency less than 1%), with deletions, duplications and insertions comprising the majority of variants and leading to the alteration by SV of a median 180 genes per genome (Collins et al. 2020). Overall, 37% of autosomal genes are affected by at least one loss-of-function variation and 24% by at least one copy number gain (Collins et al. 2020).

This population-level copy number variation forms a large part of the raw material that natural selection can operate on, leading to fixation or loss of paralogs in a population caused by the advantageous or deleterious effects respectively of gene dosage changes. Early theories of gene dosage compensation over evolutionary time proposed that duplicate genes confer redundancy and thus can compensate for loss-of-function mutations in paralogous genes (Gu et al. 2003; Hsiao and Vitkup 2008; Plata and Vitkup 2014; Su et al. 2014). However, duplications of genes such as oncogenes in which gain-of-function is deleterious are also seen in the human genome (McLysaght et al. 2014), and so these simple compensation models are clearly incomplete. Indeed, multiple studies have shown that there are many dominant disease-associated genes which are ancient and have multiple paralogs in the human genome (Domazet-Loso and Tautz 2008; Cai et al. 2009; Dickerson and Robertson 2012), despite the purifying selection which would be expected to remove these genes from the genome (Furney et al. 2006; Cai et al. 2009). One possible explanation for this could be co-dependence between interacting paralogs that together support the ancestral function, increasing fragility rather than robustness since mutation of one paralog would also disrupt the other, and so leading to selection against the loss of either paralog (Diss et al. 2017).

Later work explored the differences between genes duplicated by small-scale duplications (SSDs) and genes duplicated by whole genome duplication (WGD), two rounds of which occurred in our early vertebrate ancestors (Ohno 1970; Makino and McLysaght 2010; Singh et al. 2012). Genes created by WGD, known as the ohnologs, are by definition initially dosage-balanced since the genes are duplicated along with all their interacting genes, thus preserving interaction stoichiometry. It has also been hypothesised that WGD could confer an immediate fitness increase by reducing stochastic noise in expression rates, because the presence of duplicates would reduce the net effects of transient alterations to expression of one copy of a gene (Pires and Conant 2016).

Whole genome duplication therefore offers possible explanations for the presence of multiple copies of genes of ancient origin which have the potential for deleterious gain-of-function mutations. One such explanation is that dosage-balanced genes were safely duplicated by WGD along with their interaction partners and it was the subsequent loss of non-dosage-balanced genes which led to an excess of dosage-balanced dominant disease-associated genes in the human genome which cannot safely be gained or lost (Makino and McLysaght 2010). Another related theory is that because the WGD event was effectively a speciation event, caused by the inability of the tetraploid offspring to breed with diploids in the surrounding population, this led to a population bottleneck followed by retention of ohnologs prone to deleterious mutations because the continued presence of a functional copy of the gene masked the deleterious mutation (Singh et al. 2012).

A clear understanding of the evolution of genes over evolutionary timescales requires an accurate assignment of ages to genes. One such method of gene dating involves the use of gene conservation to find the last common ancestor (LCA) of gene orthologs across extant species followed by the assignment of the approximate age of the taxon in which the common ancestor occurred (Domazet-Loso and Tautz 2008). Another method assigns to each gene the age of the gene’s most recent duplication (MRD) in the cross-species tree (Dickerson and Robertson 2012). However, these methods weight the assigned ages to ancient and recent taxa respectively and do not take into account the entire duplication history of a family of paralogs nor the sequence divergence and hence likely similarity of genes to the inferred ancestral gene.

Consequently, we have developed a new method of dating gene paralogs based for the first time on all the relevant information in the cross-species gene tree, which we call Furthest from Singleton (FFS). The first stage of our method extracts human-centric duplication trees from the cross-species gene tree in Ensembl Compara (Yates et al. 2020), preserving normalised branch lengths so that similarity to the inferred ancestral sequence can be ascertained in addition to the duplication topology. The second stage allocates the ages of the taxa of the common ancestor and the subsequent duplications to the extant paralogs, starting with the gene which is most diverged from the common ancestor (which would have been a singleton had it not diverged, hence the algorithm's name).

FFS assigns more genes to the taxa around the time of the two rounds of whole genome duplication than the other paralog dating methods and allows us to distinguish genes which are effectively ancestral in function from those which have diverged. We find that most miRNA/target interactions are also between genes created at the time of WGD in the taxon of *Vertebrata*, consistent with the avoidance of the widespread disruption which would otherwise be caused by widespread miRNA duplications without concurrent duplication of their targets.

The greater temporal resolution that comes from using sequence divergence together with duplication topology means the FFS method assigns ages to genes on the autosomes and allosomes which are broadly consistent with the evolution of therian sex determination at the split of the eutherian mammals and the monotremes (Veyrunes et al. 2008). In addition, FFS assigns ages to genes specific to mRNA silencing in general and miRNA post-transcriptional silencing in particular which are more recent than the ages assigned to the core cellular processes of transcription and nuclear RNA export, consistent with the evolution of metazoan tissue differentiation and stable body plans occurring later in evolutionary time than the core transcriptional processes necessary for the single-celled eukaryotic ancestors of the metazoa.

We show that the ohnologs are more likely to be haploinsufficient when mutated than SSDs or singleton genes, consistent with haploinsufficient genes leading to dominant phenotypes in the context of heterozygous loss-of-function and with the ohnologs' association with dominant disease (Makino and McLysaght 2010). Unlike previous studies however, we show that the majority of ohnologs have in fact been subsequently duplicated (Makino and McLysaght 2010). We also find that genes associated with recessive diseases are slightly older than those associated with dominant diseases and we have used the increased accuracy of FFS gene dating to show that there is a shift from recessive to dominant disease association at *Vertebrata* as well as showing that the haploinsufficiency of newly created genes decreases sharply after *Euteleostomi*.

These results lead us to propose that it is specifically the haploinsufficiency enrichment of the retained ohnologs that explains the excess of dominant disease-associated genes originating in whole genome duplication, rather than post-WGD retention of dosage-balanced genes more generally (Makino and McLysaght 2010) or the retention of ohnologs prone to deleterious mutations by paralogous compensation (Singh et al. 2012).

4.3 Methods

4.3.1 Building human-centric paralog trees from cross-species gene trees

The Furthest from Singleton (FFS) paralog dating method is a two-phase algorithm, consisting of the building of human-centric paralog trees for each gene family followed by the allocation of duplication ages to the genes.

The duplication and speciation event histories in the Ensembl Compara cross-species gene tree (release 103) were downloaded on 7/3/21 for all human protein-coding and miRNA genes using the Ensembl Perl API (Yates et al. 2020). The event history for each gene was downloaded by starting with the gene's node in the cross-species gene tree and traversing the events in the tree, parent node by parent node, up to the root of the gene tree and recording each event's taxon, approximate age as calculated by Ensembl and phylogenetic branch length (Figure 4.1A).

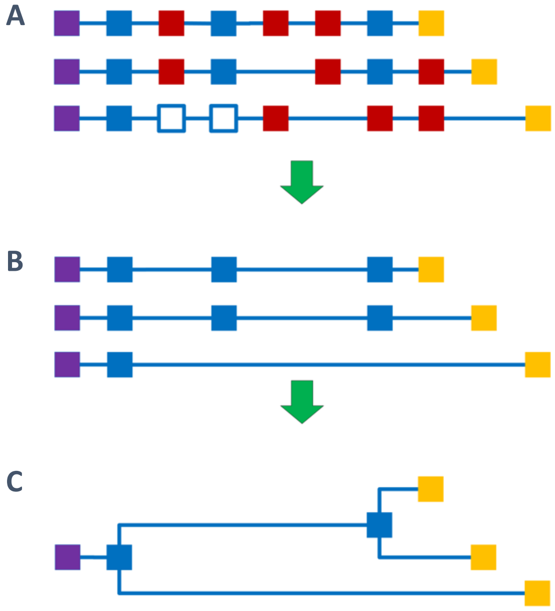


Figure 4.1 - Building human-centric paralog trees from cross-species gene tree events

The stages of building a human-centric paralog tree, illustrated by the merging of three model gene's event histories. Human genes are shown as yellow nodes, speciation events are red, duplication events are blue, dubious duplication events are outlined and root events, which can be either speciations or duplications, are purple. The distances between the nodes represent phylogenetic branch length or sequence divergence. (**A**) The event histories for each gene in a paralog family are downloaded from the gene up to the root. (**B**) The event histories are stripped of all but duplication events, retaining the branch lengths of the removed events, apart from the root event which is retained whether it is a speciation or duplication. (**C**) The duplication histories are combined into a single paralog tree by merging events where identical and branching where different.

Each gene's event history was then stripped of speciation and 'dubious' duplication events (the latter being events where Ensembl are not sure if the event was a speciation or duplication event), preserving the overall branch lengths between retained events by adding the branch lengths of the removed events to the next youngest event (Figure 4.1B). The root event is never removed and can be a speciation instead of a duplication (Figure 4.1B).

Each gene's paralogs, if any, were downloaded from Ensembl on 7/3/21 and grouped into families. The duplication histories for the genes in each paralog family were then combined into a single human-centric paralog tree started by merging the oldest event in every event history, the root event, and then merging events from each event history where both the event type and the branch length to the next youngest event are identical in every gene history and branching where different (Figure 4.1C). Non-bifurcating duplication nodes in the resulting tree were then removed, preserving the removed nodes' branch lengths as before (Figure 4.1C). Each pair of genes with a duplication as the shared parent node were saved as a list of terminal paralog pairs.

The genes' event histories were also processed to calculate the genes' ages by the last common ancestor (LCA) and the most recent duplication (MRD) methods, by taking for each gene the age of the root event and the age of the youngest duplication event in each event history respectively. The distinct taxon names and ages from the genes' events were extracted as the list of unique taxa in Ensembl (Supplementary Table 6.10).

4.3.2 Allocating duplication ages to genes in paralog families

The human-centric paralog tree for each gene family was then traversed in order to assign the ages from the root node and internal duplication nodes to the leaf gene nodes, taking into account not only the topological structure of the tree but also the relative branch lengths as a proxy for sequence divergence.

The assignment of duplication ages to the genes starts with the undated gene with the longest total branch length between it and the root node, which is the gene most diverged from the inferred ancestral sequence (Algorithm 4.1). The tree is then traversed towards the root considering each duplication in turn and, if the duplication has any undated genes apart from the current gene, then the current gene is assigned the duplication's age (Algorithm 4.1). This is repeated for each undated gene in order of decreasing total branch length until they have all been given an age, with the least diverged gene taking the age of the root event (Algorithm 4.1).

```
while there are genes in the paralog tree without ages...
  set the subject to be the most diverged undated gene
  for each duplication from the subject’s ancestor back up to the root...
    if the duplication has other undated genes...
      the subject takes the duplication's age
  if the subject is still undated...
    the subject takes the root event’s age
```

Algorithm 4.1 - Allocation of duplication ages to genes

The FFS age allocation algorithm allocates the ages of the root and subsequent duplication events in each paralog tree to the genes, starting with the gene which most diverged from the inferred ancestral sequence.

Whereas the LCA method assigns the oldest event’s age to every gene and so weights the overall age distribution to more ancient taxa (Figure 4.2A), the MRD method weights paralog ages to more recent taxa (Figure 4.2B), and neither method uses all the duplication nodes in the tree nor takes sequence divergence into account.

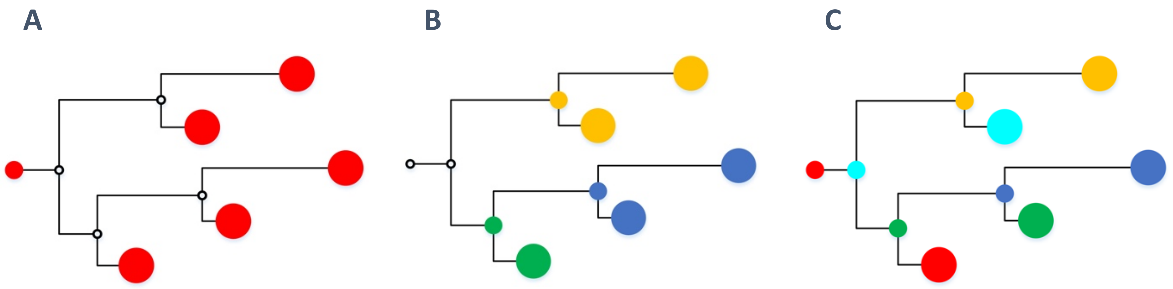


Figure 4.2 - Paralog dating methods assign a wide range of ages to genes

A model phylogenetic tree for a family of five paralogs illustrating the (**A**) Last Common Ancestor, (**B**) Most Recent Duplication and (**C**) Furthest from Singleton gene dating methods. The small circular nodes are the root event or subsequent duplication events, and the large circular nodes are genes, with matching colours showing which event’s ages are assigned to which genes (the small empty circles are events which are unused in the method).

The FFS method on the other hand allocates a wider range of ages to genes (Figure 4.2C), based on their inferred similarity to the ancestral form as well as their duplication topology, allowing paralogs that are effectively ancestral in function to be distinguished from those which have diverged. Both paralogs created by a duplication have the same origin and duplication date but, because of subsequent divergence, one is less related to the ancestor and so the FFS method gives the more diverged paralog a new origin date at the time of duplication (Figure 4.2C). Conversely, the less diverged paralog of a pair which result from a duplication will be assigned an older age *if one is available in the topology* (Figure 4.2C).

4.3.3 Disease association, haploinsufficiency, verified miRNA targets and duplications

Gene disease association status was downloaded on 21/3/21 from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al. 2002; McKusick-Nathans Institute of Genetic Medicine 2021). Each gene was annotated as dominant or recessive by searching for the keywords 'dominant' and 'recessive' respectively in the OMIM dataset's phenotype field. Haploinsufficiency scores were downloaded on 8/3/21 from Decipher (Firth et al. 2009; Huang et al. 2010). Verified miRNA/target interactions were downloaded from miRTarBase (Chou et al. 2018) on 6/2/21 and filtered to those flagged as strongly functional.

The gene paralogs which are likely to have arisen in the two rounds of whole genome duplication (WGD) thought to have occurred in ancestral vertebrate species were downloaded on 7/3/21 from the online Ohnologs database (Singh and Isambert 2020). The genes in the 'strict' ohnolog category were annotated as ohnologs, ohnologs which have a subsequent duplication later than *Vertebrata* (approximately 615 million years ago according to current Ensembl data) were annotated as 'ohnolog/SSD', genes with no duplications were annotated as singletons and the rest were annotated as small-scale duplications (SSDs).

4.4 Results

4.4.1 FFS gene ages are based on sequence divergence as well as duplications

The somatic evolution that occurs in cancer is mediated to a large degree by copy number variations and aneuploidies. To complement our analyses in earlier chapters of these somatic mutations we wanted to investigate the germline evolutionary histories of the genes involved in cancer as well as in other inheritable diseases. The methods used to assign approximate ages to genes in previous studies vary widely in approach and result in gene age distributions with characteristic weightings to either ancient or recent taxa for the last common ancestor (LCA) or most recent duplication (MRD) methods respectively.

Our new method of paralog dating is called Furthest from Singleton (FFS) because its age allocation phase starts with the gene which is most diverged from the inferred ancestral sequence of the gene which would have been a singleton had it not been duplicated (see methods). By considering the paralogs from most diverged to least, in the context of the remaining unallocated duplication ages in the paralog tree as each gene is considered, we guarantee as far as possible that within the confines of the topology the paralogs' ages are inversely correlated with their divergence from the ancestral form. The insight behind this algorithm came from the consideration of the scale-invariant nature of bifurcating paralog trees.

Briefly, the algorithm parses human gene event histories from the Ensembl Compara cross-species tree and then processes and combines these histories for each paralog group to reconstruct human-centric paralog duplication trees. The approximate ages of the taxa for the root event and subsequent duplication events are then assigned to the genes based on their divergence from the inferred ancestor and on the duplication topology, allowing us to distinguish between genes likely to be supporting the ancestral function and those which have diverged to support other functions.

We calculated the ages of the protein-coding and miRNA genes using the LCA, MRD and FFS methods, which resulted in ages from each method for 19,806 protein-coding genes and for 1,002 miRNA genes, comprising the genes with gene trees in Ensembl Compara. The genes' ages vary from zero for genes which are novel in *Homo sapiens* to 1,105 million years for genes originating in the *Opisthokonta* taxon at the split of animals and fungi (Figure 4.3).

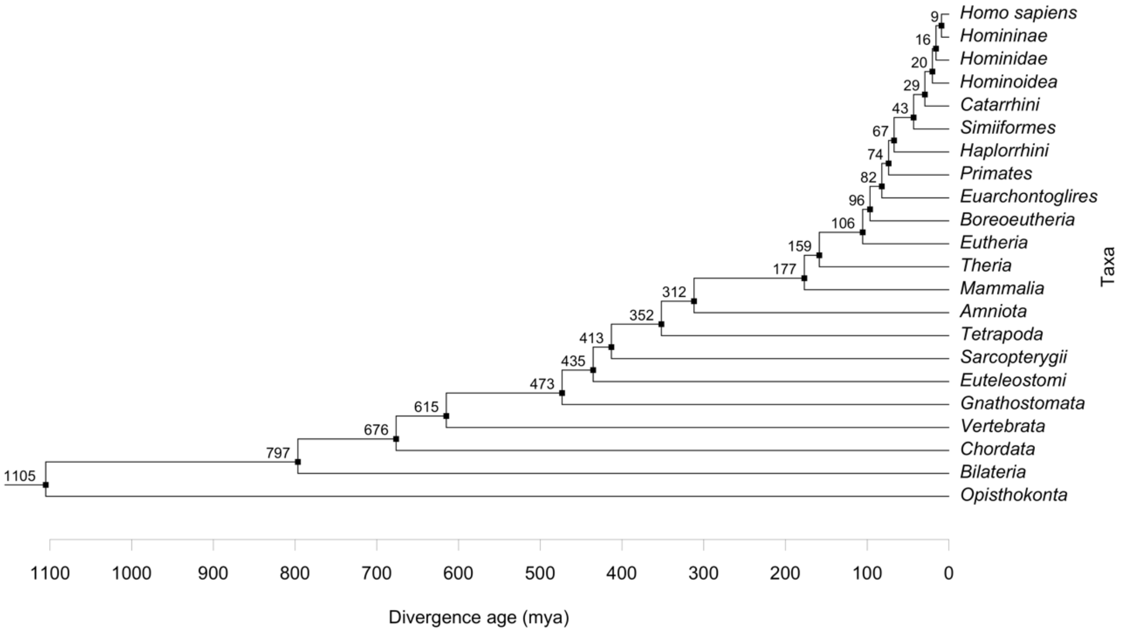


Figure 4.3 - The phylogenetic tree of the Ensembl Compara cross-species gene tree taxa

The phylogenetic tree of the unique duplication events ancestral to humans in the Ensembl Compara cross-species gene tree, generated from the FFS gene histories. The names of the taxa are on the right. The approximate age in millions of years (x axis) of each divergence is shown next to the divergence node (see also Supplementary Table 6.10).

The different dating algorithms result in very different distributions of gene ages. As expected, the LCA method dates many genes to the most ancient taxa in the Ensembl Compara tree of *Opisthokonta* (6,120 genes, 1,105 mya) and *Bilateria* (8,189 genes, 797 mya) (Figure 4.4). The MRD method on the other hand results in very few genes at *Opisthokonta* or *Bilateria* but instead allocates most genes to taxa around *Gnathostomata* (2,055 genes, 473 mya) and *Eutheria* (2,119 genes, 106 mya) (Figure 4.4). The MRD method does not assign any age to genes which have not been created by duplication (6,448 genes). FFS is intermediate between LCA and MRD at *Opisthokonta* (1,803 genes) and *Bilateria* (3,332 genes) and has the most genes at the next three ancient taxa of *Chordata* (1,796 genes, 676 mya), *Vertebrata* (3,974 genes, 615 mya) and *Gnathostomata* (3,222 genes) (Figure 4.4). All three algorithms have smaller peaks in the resulting age distributions at *Eutheria* and both MRD and FFS also have peaks around *Amniota* (312 mya), *Simiiformes/Catarrhini* (43/29 mya), *Homininae* (9 mya) and *Homo sapiens* (Figure 4.4).

Taking the *Tetraspanin* paralog family, one of the larger families, as an illustrative example, we can see clearly the effects of using the FFS algorithm to assign gene ages (Figure 4.5).

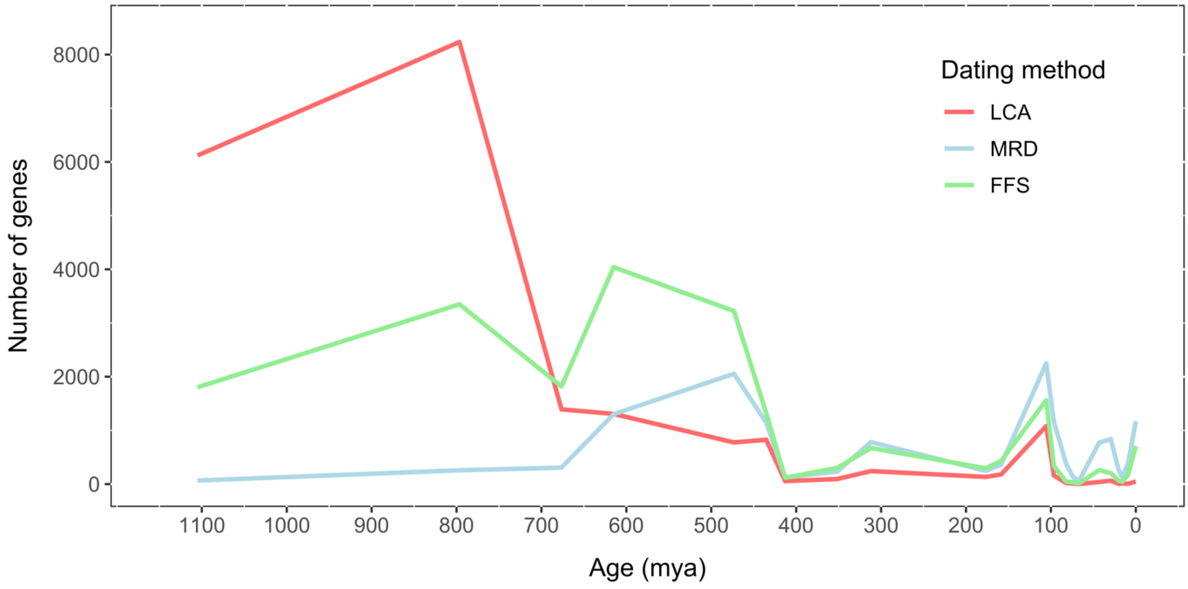


Figure 4.4 - The choice of paralog dating algorithm greatly influences the gene ages

The numbers of genes (y axis) dated to each taxon (x axis in millions of years) for the Last Common Ancestor (red), Most Recent Duplication (blue) and Furthest from Singleton (green) dating methods.

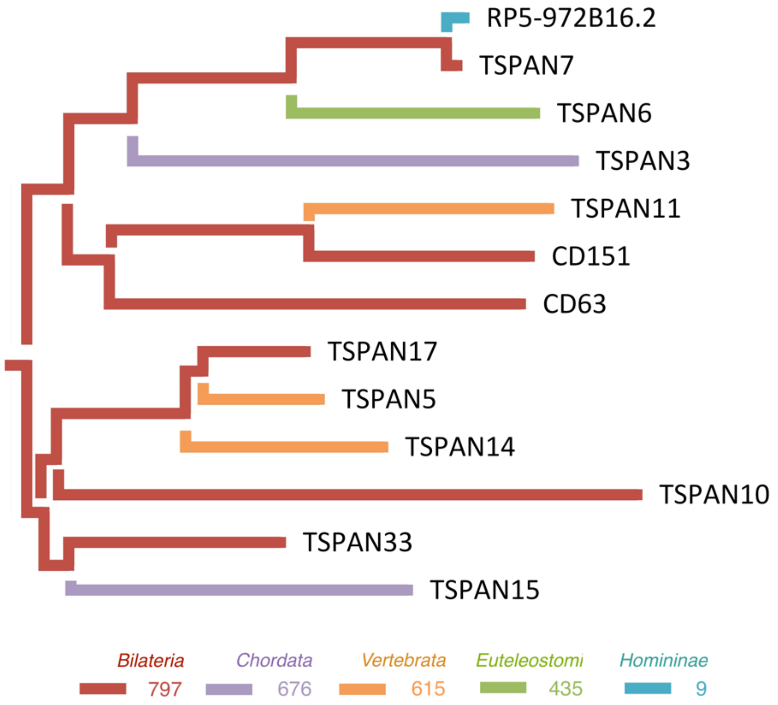


Figure 4.5 - The FFS age allocation algorithm applied to the* Tetraspanin *paralog family

The phylogenetic tree of the human *Tetraspanin* paralog family, annotated to show how the ages of the duplication nodes are assigned by the FFS algorithm to the genes. The colour of each gene's history line represents the taxon and approximate age (in millions of years, value from Ensembl shown in legend) assigned by Ensembl Compara to the duplication event (at the left-hand end of each line) that sets the age of the gene (at the right-hand end of each line). The branch lengths are proportional to the sequence divergence.

Pairs of genes which both originated from the same duplication event, such as *TSPAN5* and *TSPAN17*, are assigned ages such that the less diverged gene is assigned an older age if one is available in the topology, consistent with its assumed greater similarity to the ancestral form of the gene (Figure 4.5). *TSPAN10* on the other hand has only *Bilateria* duplication events in its history and so must be at least this age, despite being the most diverged paralog in the tree (Figure 4.5).

4.4.2 Genes originating from whole genome events are clearly visible with FFS

Genes are created by a variety of processes, such as the ohnologs (n = 4,918) created by whole genome duplication (WGD) of which a majority in our data experience later duplication (n = 3,303), small-scale duplications (SSDs) (n = 10,587) or the singletons created by *de novo* mutations, inversions and gene fusions (n = 5,410). The LCA method counterintuitively dates most of the ohnologs to between *Opisthokonta* and *Bilateria* despite the WGD events from which they originated occurring later at approximately *Vertebrata* (Singh and Isambert 2020) and simply ignores duplication events entirely (Figure 4.6A). The MRD method dates the ohnologs that have since duplicated (ohnolog/SSDs) to much more recently than the WGD events, along with the SSDs, and cannot assign an age to singleton genes at all (Figure 4.6B).

FFS clearly dates the retained ohnologs to a median age of *Vertebrata* whether they have subsequently been duplicated or not (Figure 4.6C), consistent with the origins of these genes in the two whole genome duplication events which occurred in an early ancestor of the vertebrates (Singh and Isambert 2020). The median FFS-derived age of the genes which originated from small-scale duplications is also at *Vertebrata* (Figure 4.6C), making this the taxon with the most genes assigned to it by this dating method (Figure 4.4). The singleton genes have a median age of *Gnathostomata* and an interquartile range of *Chordata* to *Mammalia* (Figure 4.6C).

The LCA and MRD dating methods give very different results for the ages of protein-coding genes and miRNAs, with the MRD method dating both sets of genes mainly between *Eutheria* and *Euteleostomi* (Figure 4.6E), whereas LCA assigns most miRNA to *Eutheria* and the protein-coding genes to between *Chordata* and *Opisthokonta* (Figure 4.6D). FFS dates the miRNAs mainly to *Eutheria*, similar to the LCA miRNA ages, but dates the bulk of the protein-coding genes to between *Bilateria* and *Euteleostomi* with a median age of

Vertebrata (Figure 4.6F). Interestingly, none of the dating methods assign any miRNAs to taxa between *Homininae* and *Boreoeutheria*, which is possibly related to the fact that 46% of miRNAs (837 out of 1,839 miRNAs with entries in Ensembl as well as miRBase) do not have a gene tree in Ensembl 103, unlike protein-coding genes where only 0.4% (86 out of 19,806) do not have a gene tree.

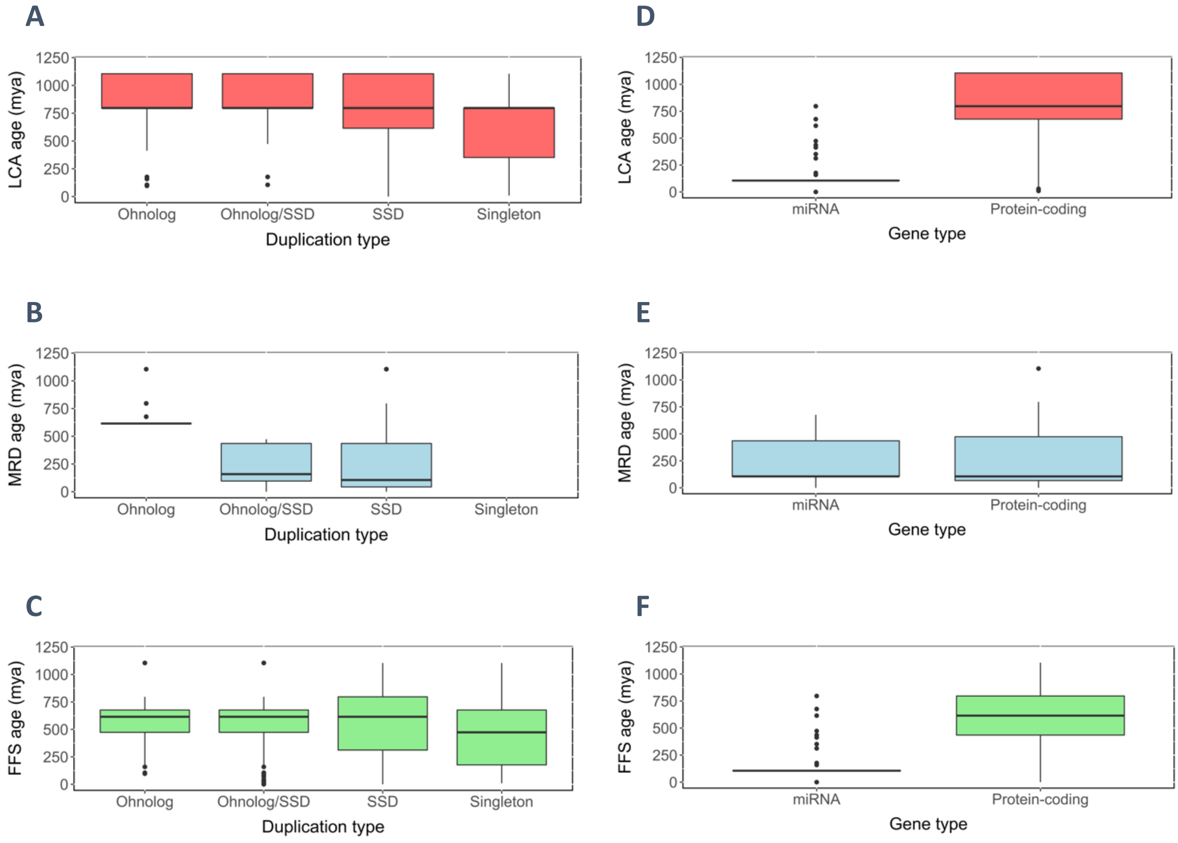


Figure 4.6 - FFS identifies the ohnologs and separates the miRNA/protein-coding ages

The distributions of ages for each duplication type (ohnolog, ohnolog/SSD, SSD and singleton) as calculated by the (**A**) Last Common Ancestor (LCA), (**B**) Most Recent Duplication (MRD) and (**C**) Furthest from Singleton paralog dating methods. Also shown are the distributions of ages for miRNAs and protein-coding genes as calculated by the (**D**) LCA, (**E**) MRD and (**F**) FFS paralog dating methods.

Our novel FFS paralog dating method assigns ages to the ohnologs which are consistent with their origins, unlike the other methods, and is able to clearly distinguish miRNAs from protein-coding genes. As this new method takes into account the gene trees' topologies as well as the likely similarity to the ancestral form it is more able to distinguish genes which are effectively ancestral in function from those which have diverged; consequently, all further analyses in this chapter will be based on gene ages derived from the Furthest from Singleton method.

4.4.3 *The peak of miRNA/target interactions are between genes from* Vertebrata

We have observed the very different distributions of ages for miRNAs and protein-coding genes, with median ages at *Eutheria* and *Vertebrata* respectively (Figure 4.6F), and we wanted to see whether miRNAs and their protein-coding targets tend to be created at the same time. Alternative scenarios to contemporaneous miRNA/target creation include new miRNAs repressing pre-existing targets and new targets being repressed by pre-existing miRNAs. We took the confirmed miRNA/target interactions from miRTarBase and plotted the number of interactions at each miRNA and target age as a 3D surface (Figure 4.7).

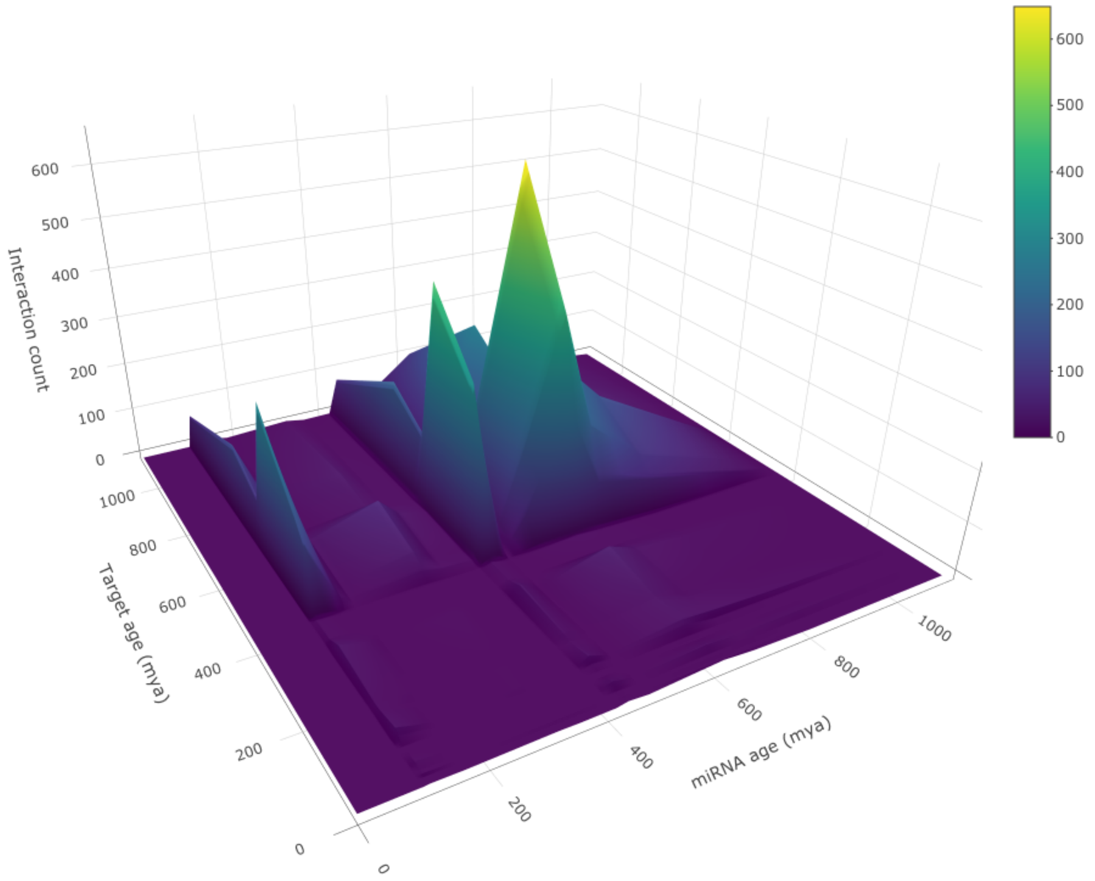


Figure 4.7 - MicroRNAs mostly target protein-coding genes originating in **Vertebrata**

A 3D surface of the number of miRNA/target interactions (vertical axis) at each miRNA age and target age (horizontal axes).

Despite there being four peaks in the creation rate of miRNAs with a main peak at *Eutheria* (747 miRNAs) and secondary peaks at *Homo sapiens* (61 miRNAs), *Euteleostomi* (51 miRNAs) and *Vertebrata* (67 miRNAs), only three of these peaks are visible in the miRNA/target interaction surface, with just one confirmed interaction for miRNAs which originated in humans meaning that there is no visible interaction peak at *Homo sapiens* (Figure 4.7).

The largest peak of miRNA/target interactions occurs for both miRNAs and targets from *Vertebrata* (615 mya) with secondary peaks of miRNA from *Euteleostomi* (435 mya) and *Eutheria* (106 mya) interacting with targets from *Vertebrata* (Figure 4.7). The similarity of shape of these peaks at each miRNA age is caused by the overall age distribution of protein-coding genes, with the majority originating between *Bilateria* and *Euteleostomi* with a median age of *Vertebrata* (Figure 4.6F). The creation of large numbers of miRNAs is likely to cause widespread disruption unless the stoichiometric balance of the interactions is maintained, and so it is likely that the peak of interactions occurring between both miRNAs and targets originating at *Vertebrata* is a consequence of the two rounds of whole-genome duplication which occurred in that era, which by definition create stoichiometrically balanced interaction networks.

4.4.4 FFS clearly shows cellular processes in their evolutionary context

The LCA and MRD paralog dating methods can't be used for fine-grained temporal analyses because of their weighting of gene ages to ancient and recent taxa respectively and so are only useful for relatively limited genome-wide analyses. FFS on the other hand uses sequence divergence as well as duplication topology in order to assign ages to the paralogs from internal nodes in the paralog tree and so we hypothesised that we might be able to use this increased temporal resolution to see process-specific evolutionary histories.

We preface this narrowing of focus however with a genome-wide analysis of miRNA and protein-coding gene ages based on the type of chromosome on which they occur. We saw in section 2.4.3 that the sex chromosomes rarely experience gains in cancer, consistent with their effectively haploid dosage, but can often have aneuploid X chromosome loss in female-derived cell lines because of the redundant copy of chromosome X and can also experience loss of chromosome Y in male-derived cell lines since there are relatively few genes on chromosome Y (Figure 2.6). We see an equally striking difference in the FFS-derived ages of the genes on the autosomes and sex chromosomes with the autosomal genes having an older median age of *Vertebrata* compared to genes on chromosome X (median age *Gnathostomata*, inter-quartile range *Eutheria* to *Chordata*) and to genes on chromosome Y (median age *Homininae*) (Figure 4.8A). The ages of the sex chromosome genes are broadly consistent with the origins of the X/Y system of therian sex determination in the divergence of a pair of autosomes at the split of eutherian mammals and the monotremes (Veyrunes et al. 2008), followed by subsequent degeneration of the Y chromosome (Graves 2006).

In another genome-wide analysis we investigated the ages of genes which experience CNVs across the NCI-60 panel and found that the age distributions of genes which are lost, unaltered and gained are practically identical, with a median age of *Vertebrata* and an inter-quartile range of *Euteleostomi* to *Opisthokonta* for all types of CNV, though the means vary slightly by around 18 million years (Figure 4.8B). This analysis however simply exposes a limitation of dating genes to only a limited quantised range of 22 distinct ages from Ensembl Compara: the large number of gene CNV values across the NCI-60 panel (1.3 x 10^6 gene CNVs) means that on a panel-wide basis the age distributions of large numbers of entities converge in the limit to the overall gene age distribution (Figure 4.4). There is some variation in the ages of genes affected by different CNV types on a per-cell line basis (Supplementary Figure 6.6) but not enough to influence the genome and panel-wide result (Figure 4.8B).

Narrowing our focus now to genes with oncogenic or tumour suppressive effects we find that while the protein-coding oncogenes and tumour suppressors have the same median age of *Vertebrata* the ranges vary a little with tumour suppressors having a greater inter-quartile range of *Gnathostomata* to *Bilateria* compared to the oncogenes which have only outliers older than *Vertebrata* (Figure 4.8C). The miRNAs associated with cancer on the other hand exhibit the opposite pattern with both oncomirs and tumour suppressor miRNAs again having a median age of *Vertebrata* but the tumour suppressors having an inter-quartile range which extends to the younger taxon of *Euteleostomi* (Figure 4.8D). The vast majority (96%) of miRNAs are not associated with cancer and these miRNAs have a median age of *Eutheria* (Figure 4.8D), which is the same age distribution as all miRNAs in general (Figure 4.6F).

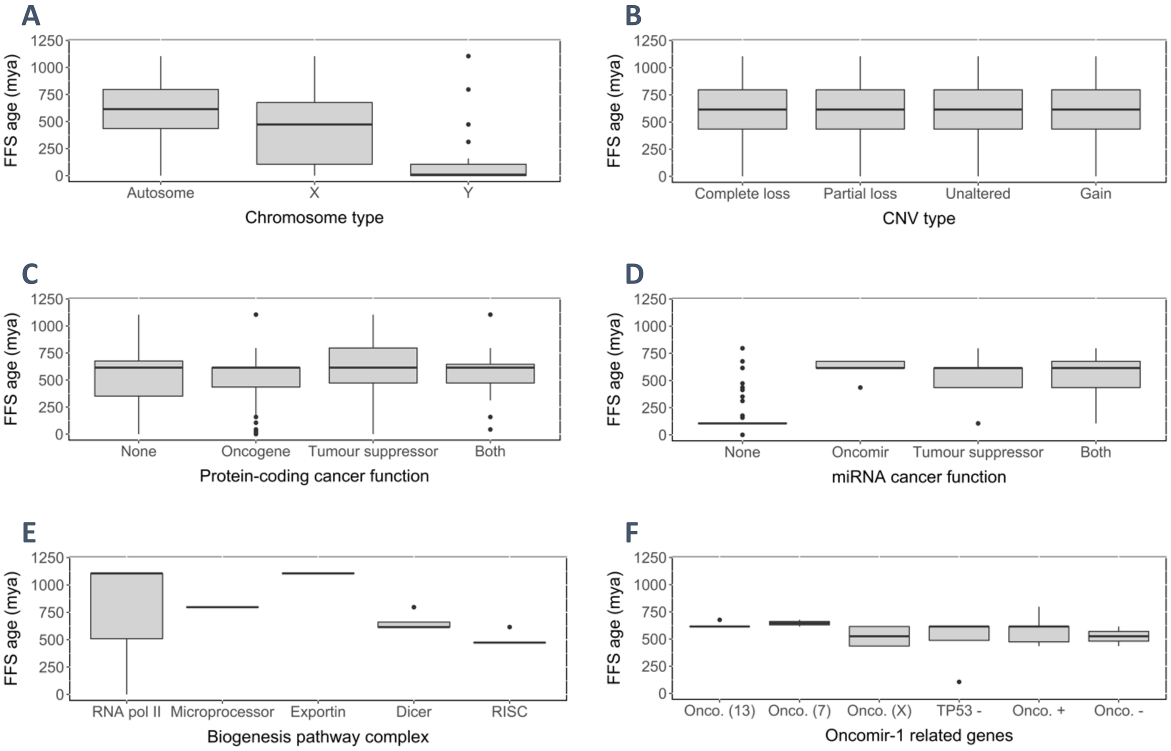


Figure 4.8 - FFS gene ages differentiate the evolutionary histories of cellular processes

The age distributions assigned by the Furthest from Singleton paralog dating method to various sets of genes. (**A**) The ages of genes which are located on autosomes and sex chromosomes. (**B**) The ages of genes which are affected by various types of CNV. (**C**). The age of protein-coding genes which are associated with cancer. (**D**) The ages of miRNAs which are associated with cancer. (**E**) The ages of genes which form components of the miRNA biogenesis pathway. (**F**). The ages of genes associated with *Oncomir-1* and its paralogs (shown as 'Onco.' followed by the name of the chromosome), miRNAs which repress *TP53* ('TP53 –'), transcription factors which activate *Oncomir-1* and/or its paralogs ('Onco. +') and transcription factors which inactivate *Oncomir-1* and/or its paralogs ('Onco. –').

While the miRNAs associated with cancer have a median age of *Vertebrata* (Figure 4.8D), the same as the protein-coding genes associated with cancer (Figure 4.8C), the genes which form the various components of the miRNA biogenesis pathway are generally much older (Figure 4.8E). The genes for the 12 subunits of RNA polymerase II and the genes necessary for export of RNA hairpin loops from the nucleus to the cytoplasm (*Exportin-5* and *RanGTP*) are almost all dated to *Opisthokonta* (Figure 4.8E). Genes associated with the nucleus-located *Drosha* and the cytoplasm-located *Dicer* and RISC complexes on the other hand are younger, with median ages of *Bilateria*, *Vertebrata* and *Gnathostomata* respectively (Figure 4.8E). The ages of the genes associated with the miRNA biogenesis pathway indicate that the mRNA silencing mechanism in general and miRNA-mediated mRNA repression in particular have been under selection pressure later than the more core cellular processes of transcription and export of RNA from the nucleus to the cytoplasm, consistent with the latter's necessity in the single-celled ancestor of *Opisthokonta* and the former's necessity for tissue differentiation and stable body plans in metazoa.

In contrast to the wide range of gene ages that we see in the miRNA biogenesis pathway (Figure 4.8E) the ages of the genes in *Oncomir-1*, its paralogs and its interaction partners are more tightly clustered around *Vertebrata* (Figure 4.8F). Of the six precursor miRNAs in *Oncomir-1*, *mir-92a-1* is dated to *Chordata* and the rest to *Vertebrata* (Figure 4.8F), consistent with the more ancient origins of *mir-92a-1* which is conserved in flies unlike the other *Oncomir-1* members which are conserved only in vertebrates (Wang et al. 2016). The dating to *Chordata* of *mir-92a-1*, rather than to *Bilateria* as implied by its conservation in flies, is caused by the Ensembl gene tree for this miRNA not currently (in Ensembl version 103) having any events older than *Chordata*. The *Oncomir-1* paralog on chromosome X has a wider range of ages with two precursors (*mir-19b-2* and *mir-92a-2*) assigned an age of *Vertebrata*, two precursors (*mir-20b* and *mir-363*) assigned an age of *Euteleostomi* and two undatable (*mir-106a* and *mir-18b*) because they lack gene trees in the current version of Ensembl Compara (Figure 4.8F). These slightly more recent ages for the members of the chromosome 13 *Oncomir-1* paralog are consistent with the presumed creation of the chromosome X and chromosome 7 paralogs during the two rounds of whole genome duplication in vertebrates (Tanzer and Stadler 2004).

4.4.5 Less diverged paralogs are older and more haploinsufficient

We have observed so far that the genes associated with cancer are dated to around *Vertebrata* (Figure 4.8C/D), which is when two rounds of whole genome duplication are thought to have occurred (Singh and Isambert 2020), and we wanted to see what else we could determine about disease over evolutionary time using our new paralog dating method. Using the pairs of paralogs (n = 3,384) which are leaf nodes of a duplication in the FFS duplication trees, which we call terminal paralog pairs, we calculated the relative branch length of each member of a terminal paralog pair by comparison to the other member's branch length as either short, long or the same (a difference of < 10^{-6} normalised to the paralogs' tree's unitless maximum branch length).

Paralogs which are less diverged than their partner have the shorter branch length of the pair and so are much more likely to be assigned an older age by the FFS age allocation algorithm if an older age is available in the duplication topology. Unsurprisingly we therefore find that, at least in part as a consequence of the FFS age allocation algorithm, paralogs with the shorter branch length of a terminal paralog pair are indeed older than their sibling paralogs, with a median age of *Vertebrata* for short branch paralogs compared to *Gnathostomata* for long branch paralogs, which also have an inter-quartile range which extends to more recent taxa (Figure 4.9A). Terminal paralog pairs with the same branch length have much more recent FFS-assigned ages (median *Homo sapiens*), consistent with being created by duplication relatively recently and so having not had the evolutionary time necessary to have diverged (Figure 4.9A).

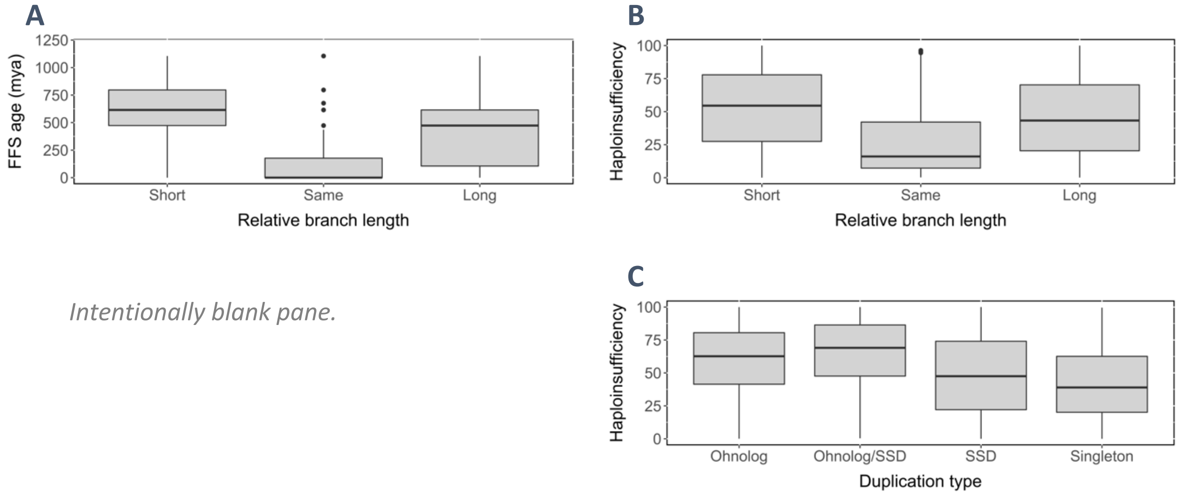


Figure 4.9 - Less diverged paralogs are older and more haploinsufficient

(**A**) The gene age distributions of members of terminal paralogs pairs which have shorter, longer or the same relative branch length. The haploinsufficiency prediction scores, where zero is completely haplosufficient and 100 is completely haploinsufficient, of (**B**) members of terminal paralogs pairs which have shorter, longer or the same relative branch length and of (**C**) genes which are ohnologs, ohnolog/SSDs, SSDs or singletons.

The age distribution of the short branch paralogs with a median age of *Vertebrata* (Figure 4.9A) is strikingly similar to that of the ohnologs, whether or not they have subsequently been duplicated (Figure 4.6C). The ohnologs have been shown to be enriched for dosage-balanced genes (Makino and McLysaght 2010), and so we hypothesised that as the short branch paralogs have a similar age distribution to the ohnologs, the short branch paralogs would be more haploinsufficient than the longer branch paralogs. We used the haploinsufficiency prediction scores from Decipher (Firth et al. 2009; Huang et al. 2010) to calculate how likely members of terminal paralog pairs are to be haploinsufficient and found that the paralogs with the shorter relative branch length are indeed more likely to be haploinsufficient than paralogs with the longer relative branch length (Figure 4.9B).

Paralogs with the same branch length are much less likely to be predicted to be haploinsufficient (Figure 4.9B), consistent with having the same interaction partners as each other because of their sequence similarity. We then directly compared the haploinsufficiency prediction scores for genes designated as ohnologs to those which are SSDs or singletons and confirmed that the ohnologs, whether subsequently duplicated or not, are predicted to be more likely to be haploinsufficient than SSDs or singletons (Figure 4.9C), consistent with earlier studies (Makino and McLysaght 2010).

Haploinsufficient genes lead to a dominant phenotype in the context of heterozygous loss-of-function (Kondrashov and Koonin 2004) and so we hypothesised that, as the more conserved short branch paralogs are older (Figure 4.9A) and more likely to be haploinsufficient (Figure 4.9B), genes associated with dominant disease should be similarly older and more haploinsufficient. We investigated the FFS-derived ages for genes annotated as dominant or recessive in the phenotypes downloaded from OMIM (McKusick-Nathans Institute of Genetic Medicine 2021) and found that while the ages of genes associated with dominant or recessive disease are fairly similar it is in fact the recessive disease-associated genes which are slightly older, with a median age of *Chordata* versus *Vertebrata* for the dominant disease-associated genes (Figure 4.10A).

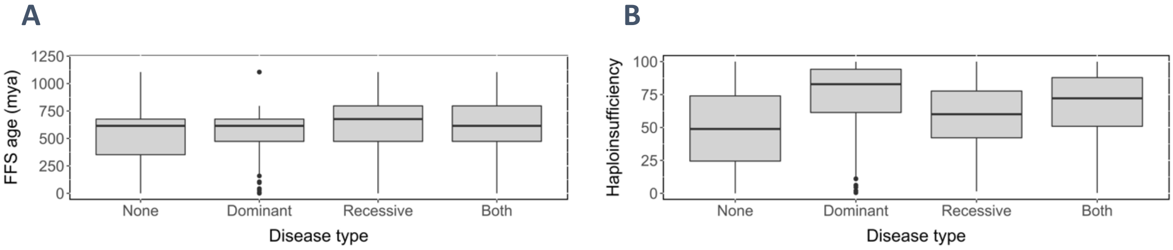


Figure 4.10 - Dominant disease genes are the most haploinsufficient

(**A**) The FFS-assigned age distributions for genes associated with dominant and recessive diseases. (**B**) The predicted haploinsufficiency scores for genes associated with dominant and recessive diseases.

Genes associated with disease in general are predicted to be more likely to be haploinsufficient than genes not associated with disease (Figure 4.10B) and, as we hypothesised above, genes associated with dominant disease are the most likely to be haploinsufficient (Figure 4.10B), consistent with earlier work (Kondrashov and Koonin 2004; Makino and McLysaght 2010). Interestingly, genes associated with recessive disease are also predicted to be more haploinsufficient than genes not associated with disease, albeit less so than genes associated with dominant disease (Figure 4.10B).

4.4.6 WGD is correlated with a sudden decrease in haploinsufficiency of new genes

Our results so far are consistent with the contribution of the ohnologs to the bulk of inheritable dominant disease, as a consequence of the biased retention of haploinsufficient genes, and we've also shown that genes associated with recessive disease are slightly older than the ohnologs. Consequently, we wanted to see if the increased temporal resolution of the FFS gene dating method could determine whether there is an increase in the rate of retention of genes associated with dominant diseases relative to recessive diseases at the time of the ohnologs.

The median haploinsufficiency prediction scores of the ohnologs at each distinguishable taxon are consistently high over evolutionary time from *Opisthokonta* to approximately *Euteleostomi*, at which time they start to decline in predicted haploinsufficiency (Figure 4.11A). SSDs and singleton genes on the other hand start with a similarly high haploinsufficiency score at *Opisthokonta* but decline fairly steadily over time to the present, with the exception of the dramatic spike in the singleton median haploinsufficiency score at *Haplorrhini* (Figure 4.11A).

Interestingly, it is at *Gnathostomata*/*Euteleostomi*, the taxa after the two rounds of whole genome duplication at *Vertebrata*, that the rate of gene retention decreases sharply irrespective of the dating method used (Figure 4.4) and we see a corresponding decrease in the retention of disease-associated genes at *Euteleostomi* (Figure 4.11B). For both dominant and recessive disease-associated genes there is a fairly constant rate of increase in the cumulative frequency of genes retained until *Euteleostomi* when the rate plateaus (Figure 4.11B). There are more genes associated with recessive disease than with dominant at all taxa, primarily as a consequence of the excesses in recessive disease-associated gene retention over dominant disease-associated genes at *Opisthokonta*, *Bilateria* and *Chordata*, after which the retention rates of dominant and recessive disease-associated genes track each other more closely (Figure 4.11B).

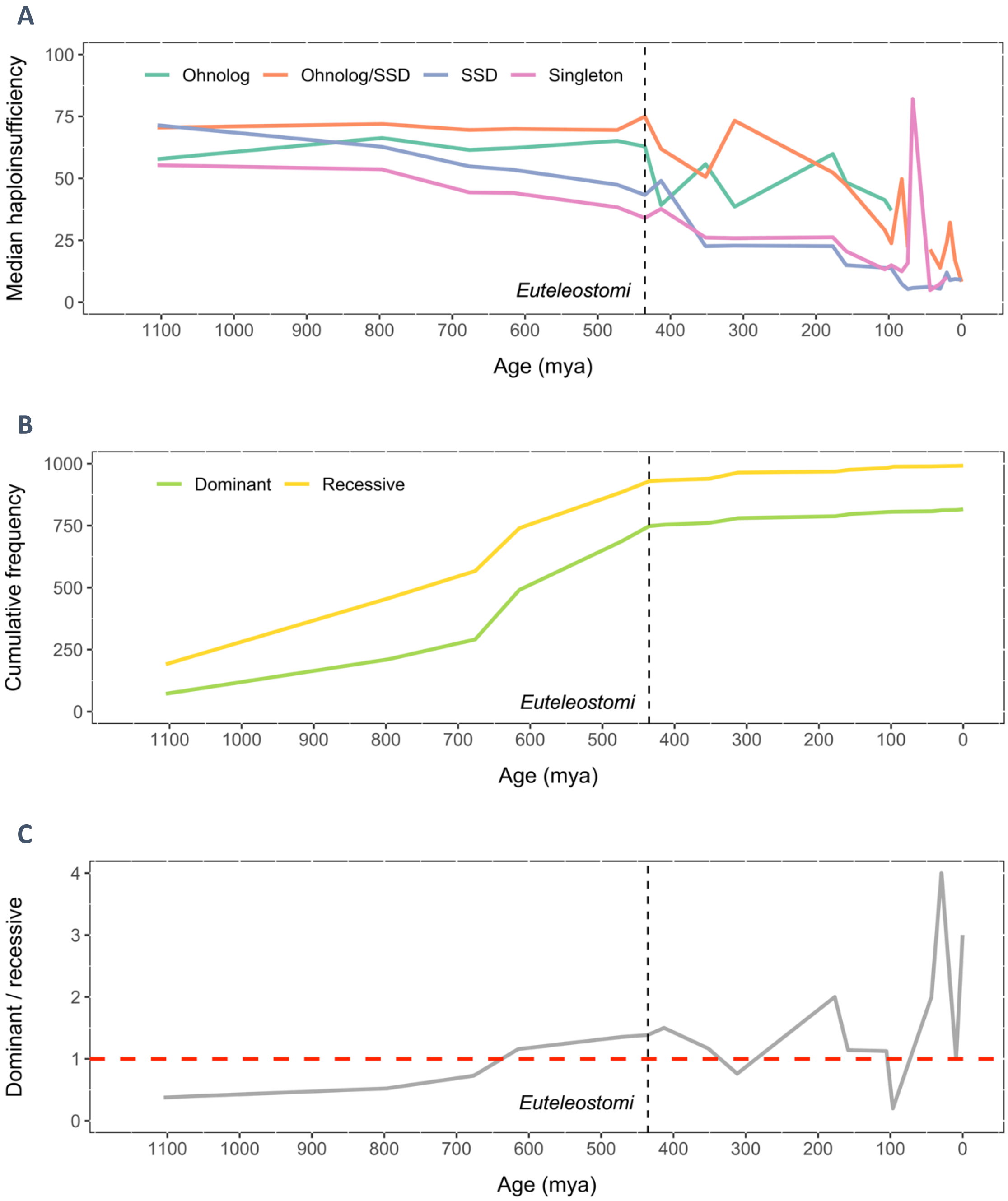


Figure 4.11 - The overall burden of dominant disease plateaus after WGD

(**A**) The median haploinsufficiency scores (y axis) over evolutionary time (x axis) for ohnologs (green line), ohnolog/SSDs (orange), SSDs (blue) and singletons (magenta). (**B**) The cumulative frequency of genes (y axis) associated with dominant disease (green) and recessive disease (yellow) over evolutionary time (x axis). (**C**) The ratio of gene creation (y axis) between dominant and recessive disease-associated genes over evolutionary time (x axis), with the red horizontal dashed line at y = 1 indicating balanced creation of dominant and recessive disease-associated genes. The black vertical dashed line in each pane indicates the *Euteleostomi* taxon at approximately 435 million years ago.

We can see the change from an excess of recessive disease-associated genes to a deficit between *Chordata* and *Vertebrata* more clearly in the ratio of dominant to recessive disease-associated genes retained from each taxon (Figure 4.11C). The rate of disease gene retention after *Euteleostomi* decreases by an order of magnitude to single digits per taxon for both dominant and recessive disease genes, and as a consequence the sharper oscillations in the ratio of dominant to recessive gene retention after this era are likely dominated by stochastic fluctuations (Figure 4.11C).

In summary, because our novel FFS algorithm uses duplication topology in conjunction with subsequent paralog divergence, we are able to elucidate the consequences of whole genome duplication with more temporal accuracy than in previous studies. Genes created after whole genome duplication decrease sharply in predicted haploinsufficiency (Figure 4.11A) and it is striking that cancer-associated genes in general (Figure 4.8D) and more specifically genes related to *Oncomir-1* in particular (Figure 4.8F) date to around *Vertebrata*, adding further evidence to the hypothesis that some cancers are associated with dominant disease mutations of the haploinsufficient ohnologs, which are retained because they are dosage-balanced.

4.5 Discussion

Various systems of gene dosage compensation have been proposed to evolve over time, such as redundant paralogs compensating for loss-of-function mutations (Gu et al. 2003; Hsiao and Vitkup 2008; Plata and Vitkup 2014; Wang et al. 2016) and stoichiometrically balanced duplication of dosage-sensitive genes in the context of whole genome duplications (WGD) (Makino and McLysaght 2010; Singh et al. 2012). Many of these ohnologs originating from the two rounds of WGD in early vertebrates are associated with dominant diseases (Domazet-Loso and Tautz 2008; Cai et al. 2009; Dickerson and Robertson 2012), despite the purifying selection that would be expected to purge these genes from the genome over time (Furney et al. 2006; Cai et al. 2009).

Accurate dating of genes allows the evolution of dosage compensation to be investigated and multiple dating methods have been proposed, such as dating the genes to the taxon of the last common ancestor (LCA) of the genes' orthologs in other species or to the taxon of each gene's most recent duplication (MRD) (Domazet-Loso and Tautz 2008; Dickerson and Robertson 2012). These methods distort the evolutionary picture however by weighting gene ages towards ancient or recent taxa respectively, leaving a dearth of genes around the period of interest near the time of the two rounds of WGD. Our new Furthest from Singleton (FFS) gene dating method takes account of the entire paralog duplication topology as well as the relative divergence of each gene from the inferred ancestor, resulting in a more nuanced distribution of gene ages with increased temporal resolution, based on all the relevant evidence in the cross-species gene tree. Our analyses of this new perspective on gene evolution leads us to propose a subtly different explanation to previous studies for the association of the retained ohnologs with dominant diseases, namely that ohnolog haploinsufficiency is a more parsimonious hypothesis than those previously advanced.

FFS dates more genes to *Chordata*, *Vertebrata* and *Gnathostomata*, the taxa from just before to just after the two WGD events in vertebrate evolutionary history (Singh and Isambert 2020), than do the previously described LCA and MRD methods (Figure 4.4). FFS dates the majority of the ohnologs, whether subsequently duplicated or not, to the taxon *Vertebrata* (Figure 4.6C), consistent with the synteny-based work which defined the ohnologs (Singh and Isambert 2020). As FFS assigns ages based on sequence divergence within the duplication topology, the dating of the ohnologs to *Vertebrata* is evidence that

FFS is assigning ages to genes based on their ‘functional age’, rather than assuming in effect that the genes’ functions must be ancestral (LCA) or novel (MRD). This does not explain however why there are some outliers in the FFS-derived pure ohnolog ages (i.e., ohnologs which have not subsequently duplicated) which appear younger than *Vertebrata* (Figure 4.6C); none of these ‘young ohnologs’ have event trees which go back as far as *Vertebrata* in Ensembl 103 whereas they presumably did in Ensembl 84, the version with which the ohnolog designation was calculated (Singh and Isambert 2020). This will be investigated further in future work.

We find that miRNAs with confirmed target protein-coding gene interactions originate mainly at *Vertebrata*, as do these miRNAs’ targets, with smaller groups of miRNAs originating at *Euteleostomi* and *Eutheria* and also predominately targeting the protein-coding genes dated to *Vertebrata* (Figure 4.7). Interestingly, this is a similar but temporally shifted view of bursts of miRNA creation to that shown in a study examining the sudden creation of the majority of metazoan body plans during the so-called Cambrian explosion (Peterson et al. 2009). Unlike our focus on individual miRNAs, this previous work dated the peak bursts of miRNA *family* creation to around *Bilateria*/*Chordata* (at the base of the protostomes and deuterostomes), *Vertebrata* and *Primates*, using an LCA-like algorithm (Peterson et al. 2009). It is possible that the FFS method provides a finer-detailed view than was available with the data or LCA-like approach used in the previous study and so we’re now seeing evidence for the divergence of the miRNAs since the origins of their families. Another explanation for the discrepancy is our use of a human-centric ‘slice’ of the Ensembl Compara cross-species gene tree rather than all known clades. More significantly, our data lacks any of the primate-specific miRNAs apart from those dated to *Homo sapiens* and so we cannot yet tell for certain how the distribution of miRNA/target interactions which we observe relates to this earlier work (Peterson et al. 2009). Obtaining the primate-specific miRNA gene trees from Ensembl Compara will clearly be a priority in our future work.

The ‘functional ages’ assigned by the Furthest from Singleton method mean that we can shed some light on process-specific evolutionary histories. We observe that while the genes on the autosomes have a median age of *Vertebrata*, the X chromosome genes originate mainly in *Gnathostomata* and the Y chromosome genes are much more recent, dating to *Homininae* at the common ancestor of humans and chimpanzees (Figure 4.8A), consistent with the origins of therian X/Y sex determination (Veyrunes et al. 2008) and Y chromosome degeneration respectively (Graves 2006). Both miRNAs (Figure 4.8D) and protein-coding genes (Figure 4.8C) associated with cancer date to around *Vertebrata* in our analyses, as do the genes associated with *Oncomir-1* and its related transcription factors and paralogs (Figure 4.8F), similar to the results from the earlier analyses of Cambrian-era miRNAs *mir-15/16* and *mir-17~92* (Peterson et al. 2009). We also see an interesting distinction between ancient core cellular processes such as transcription and nuclear RNA export and more recent miRNA silencing-specific complexes (Figure 4.8E), indicating that the miRNA-specific machinery has been under selection pressure much more recently than the basal transcription processes, also at around the time of the Cambrian explosion in the taxon of *Vertebrata*, when it would be necessary for tissue differentiation and the canalisation leading to stable body plans (Peterson et al. 2009).

The consistent dating of genes associated with cancer, a group of diseases often associated with dominant mutations, to around the time of the two rounds of whole genome duplication in the taxon *Vertebrata* led us to investigate the distribution of dominant disease over time. It has been previously suggested that the enrichment of the ohnologs for dominant disease associations is because of the higher post-WGD retention of dosage-sensitive genes (Makino and McLysaght 2010; Singh et al. 2012) and so we wanted to see how haploinsufficiency varies at around the time of the two rounds of WGD in our early vertebrate ancestors. We observe that as well as being older (partially at least as a consequence of the FFS age allocation algorithm) (Figure 4.9A), the less diverged paralogs are predicted to be more likely to be haploinsufficient when mutated (Figure 4.9B), as are the ohnologs (Figure 4.9C), consistent with earlier studies (Makino et al. 2009; Makino and McLysaght 2010) and with our observations of the ohnologs' dominant disease enrichment.

The initially identical paralogs which resulted from whole genome duplications, in conjunction with the increased precision of heritability that the evolution of the miRNAs enabled, allowed natural selection the freedom to rapidly explore morphological space during the Cambrian explosion (Peterson et al. 2009). This would appear however to have been at the cost of an increased dominant disease burden, in large part due to the biased retention of haploinsufficient genes as previously discussed. We unsurprisingly find that genes associated with dominant diseases are the most highly predicted to be haploinsufficient (Figure 4.10B) but, while the less diverged paralogs with higher predictions of haploinsufficiency are older (Figure 4.9A), we find the recessive disease-associated genes in general are slightly older than the dominant disease-associated genes, dating to *Chordata* as opposed to a median age of *Vertebrata* (Figure 4.10A).

In this study we have defined the disease association status of genes by reference to the genes' OMIM phenotypes (McKusick-Nathans Institute of Genetic Medicine 2021) and have only classified the disease-associated genes as dominant or recessive. Our results would be more nuanced if the dominant disease-associated genes were stratified further into those that are known to be haploinsufficient, dominant-negative or prone to gain-of-function mutations. A related improvement would be to take the type of gene product into account, since mutations to enzymes are generally recessive and mutations to transcription factors and monomers of multimeric complexes are usually dominant (Jimenez-Sanchez et al. 2001).

We have used the Decipher (Huang et al. 2010) predictions of haploinsufficiency in our analyses, which are derived from a linear discriminant analysis of several genomic and evolutionary properties of known haploinsufficient and presumed haplosufficient genes (genes with heterozygous loss but no phenotypical change in multiple control individuals in genome-wide association studies). Of the properties considered by this model, proximity in the network of gene interactions to known haploinsufficient genes was the most predictive of haploinsufficiency, followed by promotor conservation, embryonic expression and human/macaque dN/dS ratios (Huang et al. 2010). We will repeat our haploinsufficiency-based analyses with another continuous predictive measure of haploinsufficiency developed by the Genome Aggregation Database (gnomAD), the loss-of-function observed/expected upper bound fraction (LOEUF) metric (Karczewski et al. 2020). This metric, derived from the analysis of 125,748 whole exome sequences and 15,708 whole genome sequences, is applicable to all protein-coding genes, unlike the Decipher scores which are applicable only to the 12,443 genes with values for the four properties selected by Decipher (Huang et al. 2010). The LOEUF metric is calculated from the ratio of observed to expected loss-of-function variants and the large sample sizes of gnomAD mean that the LOEUF metric can be calculated for all the protein-coding genes (Karczewski et al. 2020).

Interestingly, the ohnologs are consistently highly predicted to be haploinsufficient from *Opisthokonta* past the likely time of the two rounds of WGD until *Euteleostomi*, at which point the predicted likelihood of haploinsufficiency decreases fairly steadily until the present (Figure 4.11A). The rate of gene retention also decreases sharply from *Euteleostomi* irrespective of the gene dating method used (Figure 4.4), leading to a plateauing of the cumulative frequencies of both dominant and recessive disease-associated genes at that time, with consistently more genes associated with recessive disease than with dominant disease at all taxa because of the early enrichment for recessive diseases in ancient taxa (Figure 4.11B).

The excess of recessive over dominant disease-associated gene retention switches to a deficit at *Vertebrata* (Figure 4.11C), consistent with the biased retention of dosage-sensitive genes post-WGD previously suggested (Makino and McLysaght 2010; Singh et al. 2012). However, unlike previous studies which found that ohnologs rarely experience SSDs post-WGD (Makino et al. 2009; Makino and McLysaght 2010), we find that when analysed in the context of the modern cross-species gene tree, which is based on the sequences of far more species than available to the previous studies, together with the more functionally based gene dating of the FFS algorithm, 67% of the ohnologs are in fact subsequently duplicated after *Vertebrata* (3,303 out of 4,918) and so are not as sensitive to dosage gains as previously thought. In conclusion therefore, while the study of extant human genes in the context of the cross-species gene tree cannot definitively explain the paradox of inheritable disease-associated paralogs, we propose that the haploinsufficiency of the ohnologs is a more parsimonious explanation than more general dosage sensitivity for the biased retention of the ohnologs associated with dominant disease.

5 Discussion

We have focussed in this thesis on understanding how gene dosage changes are buffered by miRNAs in cancer cells and how dominant diseases caused by dosage imbalances evolved across evolutionary timescales.

Efforts to map the gene dosage changes arising from copy number variations in the well-characterised NCI-60 cell line panel have to date been conducted using relatively low-resolution array-based comparative genomic hybridisation methods or whole exome sequencing analyses. We have used short read alignments from whole genome sequencing data to generate in Chapter 2 a higher-resolution map of CNVs in cancer than has been previously calculated from aCGH assays. Our analyses of these CNVs show *inter alia* that tumour suppressor genes are lost more than expected and we derive a list of candidate novel driver genes which are enriched for involvement in multiple cancer-related processes.

Tumour cells survive large variations in gene dosage which cause non-tumour cells to undergo apoptosis and so there must be mechanisms in cancer which buffer the consequences of the gene dosage alterations. We hypothesised that miRNA-mediated buffering might be one such mechanism. We investigated the CNVs of miRNAs in the cancer cell lines in Chapter 3 and found that widespread losses of miRNAs lead to the derepression of genes involved in multiple cancer-related processes and pathways. Our analyses of the *mir-17~92* cluster of miRNAs lead us to propose a novel mechanism, which potentially occurs generally in cancer cells. We suggest that transient *C-MYC*-induced repression of *TP53* can, acting via *mir-17~92* and other miRNAs, result in sustained *TP53* and *PTEN* repression while shielding the cancer cell from the apoptosis which would normally result from *C-MYC* overexpression.

In addition to investigating gene dosage changes in cancer cells, we wanted to see if the evolutionary history of disease genes could explain the surprising persistence of such genes in metazoan genomes. We developed in Chapter 4 a novel method for assigning ages to paralogs based on their duplication history and subsequent divergence. We then used this new perspective to investigate evolutionarily distinct processes such as the inheritance of basal cellular processes and the metazoan evolution of miRNA gene repression. We also see a change from recessive to dominant disease association around the time of the whole genome duplications that occurred in our vertebrate ancestors. We suggest a subtly different model of disease gene retention in the human genome to those previously advanced.

5.1 CNVs affect a wide range of cancer-related processes

Widespread gene dosage changes caused by copy number variations and aneuploidies are characteristic of the cancer cell and understanding the causes of these dosage changes is key to understanding and treating the cancer phenotype (International HapMap et al. 2007; Torres et al. 2008; Yang et al. 2016). It is necessary to study a representative range of cancers in order to elucidate general features of cancer cells and so we have analysed copy number variations in the often-studied NCI-60 cell line panel (Iafrate et al. 2004; Shoemaker 2006; Henrichsen et al. 2009). Unlike the previous such work to date, which used aCGH assays or whole exome sequencing to determine the end points and copy numbers of gained and lost regions of the genome in cancer (Lorenzi et al. 2009; Beroukhim et al. 2010; Varma et al. 2014), we have used the relative read depths of aligned short reads from whole genome sequencing data to create a more accurate map of CNVs in the NCI-60 cell lines than previously seen.

Unsurprisingly given that the cell lines under investigation were derived from tumour cells, and consistent with earlier studies (Beroukhim et al. 2010; Varma et al. 2014), we find that there are frequent dosage changes to genes associated with cancer processes. The CNV landscape is dominated by losses of tumour suppressors with less frequent gains of oncogenes and there are aneuploidies affecting the majority of chromosomes. There are few regions where the cell line genomes are completely lost however, indicating that the dosage changes are mainly perturbations to existing interaction stoichiometries rather than complete disruptions to interaction networks. Unlike previous studies (Lorenzi et al. 2009; Bignell et al. 2010; Varma et al. 2014), we are able to detect smaller CNVs with an endpoint resolution of just 10^4 nucleotides even with relatively low coverage sequencing and moreover we can calculate accurate absolute copy numbers instead of just the fold changes characteristic of aCGH-based studies.

5.2 Novel candidate cancer driver genes

Mutations occur in all cells undergoing mitosis and it is important to distinguish between the driver mutations which are causal in oncogenesis and positively selected for because of the growth advantages they provide to the cell and the passenger mutations which become fixed in a tumour clonal population simply because of driver gene proximity (Greenman et al. 2007). It was previously suggested that driver mutations could be identified by calculating CNV frequencies under the hypothesis that more CNVs will affect regions dominated by driver mutations than regions containing mainly passengers (Beroukhim et al. 2010; Bignell et al. 2010; Pleasance et al. 2010; Martincorena et al. 2017; Rheinbay et al. 2020). We therefore focused on the regions of the genome which were the most affected by CNVs across the NCI-60 panel, defining the 100 most often gained and 100 most often lost as CNV hotspots. We analysed the genes in the resulting gain and loss hotspots by performing a permutation test on the CNVs affecting the genes in each hotspot in order to focus on cancer-related genes which are affected by CNVs more than expected by chance. These candidate oncogenesis driver genes include previously known driver genes (Martincorena et al. 2017) but we also find potentially novel cancer drivers, which are enriched for cancer-related processes.

The candidate driver genes which are tumour suppressors and lost more than expected by chance are enriched for processes such as apoptosis, senescence and the Wnt and TGF-β cell cycle signalling cascades whereas the gained oncogene drivers are more associated with enhanced cell growth, proliferation and metastasis. For example, the tumour suppressor *FOXO4*, which is lost in 15 cell lines, regulates cell cycle progression and its expression levels are inversely correlated with severity of non-small cell lung cancer (Xu et al. 2014). Gains of cell division protein kinase *CDK6* on the other hand are associated with increased proliferation and angiogenesis (Diaz-Moralli et al. 2013) and are commonly upregulated in medulloblastomas (Silber et al. 2013), though in our data the *CDK6* gains are predominantly in cell lines derived from renal, hematopoietic and skin cancers. These results are consistent with the widely held hypothesis that for cancer cells to gain the various hallmarks of cancer they must disrupt a range of protective mechanisms in order to attain the cancer phenotype (Hanahan and Weinberg 2011).

Our analyses so far implicitly assume that the prominent aneuploidies which we observe across the cell lines are a cause of cancer but, since we do not know the tumour stages from which the NCI-60 cell lines were taken, we cannot be sure that these are not in fact a late-stage consequence of, for example, *TP53* inactivation (Torres et al. 2008). It would be interesting to reanalyse our data after removing the aneuploidies from the CNV map as this would remove the influence of the aneuploidies from the calculation of CNV hotspot locations, thereby changing the regions of the genome from which we draw our candidate novel driver genes. A subsequent comparison of the number of known driver genes found with and without the influence of the aneuploidies would indicate whether removing the aneuploidies could increase the sensitivity of our method in determining cancer driver genes. We have also only analysed the 55 NCI-60 cell lines which are derived from cancers in just nine tissues and so future work to apply our pipeline to the much larger set of cell lines in the Cancer Cell Line Encyclopedia (Ghandi et al. 2019) will be extremely interesting.

5.3 Haploinsufficiency leads to dominant disease gene retention

We have observed widespread gene dosage changes in cancer, changes which in normal cells trigger apoptosis. That cancer cells survive suggests the existence of mechanisms that buffer the ordinarily deleterious consequences of dosage imbalance (Torres et al. 2008; Sheltzer and Amon 2011). The buffering mechanisms available to cancer cells arose in large part over evolutionary timescales and so we have investigated the evolutionary histories of gene dosage compensation mechanisms which buffer dominant disease mutations such as those associated with cancer.

We wanted to understand the evolutionary histories of disease-associated genes without the weightings to ancient or recent taxa caused by previously proposed paralog dating methods and so we developed a novel method which takes into account paralog divergence as well as the topology of paralog duplication trees when calculating gene ages. Our new method, Furthest from Singleton, results in gene ages which correspond better to their known origins, such as dating the ohnologs to *Vertebrata*, consistent with their creation by whole genome duplication events (Singh and Isambert 2020).

MicroRNAs mediate post-transcriptional buffering of gene dosage changes by providing targeting specificity to the mRNA silencing machinery, but miRNA gains are often deleterious because of the widespread resulting disruption to expression levels, unless they are duplicated along with their targets and so overall interaction stoichiometry is preserved. We calculated the ages of the miRNAs and their verified protein-coding targets and found that the majority of the miRNAs with experimentally confirmed target interactions are dated by FFS to *Vertebrata*, as are the majority of their targets, consistent with these miRNAs and target genes being created by whole genome duplication events.

In further evidence that FFS assigns sensible ages to genes, the ages assigned to the miRNAs by our method result in a similar distribution of miRNA ages to that shown by a study which found that novel miRNA families were created at the divergences of the major metazoan lineages, especially during the Cambrian explosion (Peterson et al. 2009). This analysis is currently limited however by the lack of primate-specific miRNA gene trees in Ensembl Compara v103 and so we will pursue this in our future work.

In addition to analysing the evolutionary history of miRNA-mediated gene dosage buffering we were able to determine the ages of genes involved in specific processes and systems, such as the more ancient origins of the genes on the autosomes with a median taxon of *Vertebrata* compared to the X chromosome genes which have a younger median taxon of *Gnathostomata*, consistent with the origins of sex determination in therian mammals (Veyrunes et al. 2008). As expected, the genes associated with transcription and nuclear export during miRNA biogenesis are much older than those associated specifically with mRNA silencing and miRNA biogenesis, indicating that the miRNA-specific machinery has been shaped by more recent selection pressures than the basal transcription processes, because of its association with miRNAs during the Cambrian explosion and increased morphological complexity which arose at that time (Peterson et al. 2009).

The majority of genes associated with cancer, whether miRNA or protein-coding, also date to around the time of the vertebrate common ancestor. Since cancer is often triggered by dominant mutations, we analysed the ages of genes associated with dominant diseases. Previous studies suggested that the enrichment of the ohnologs for dominant diseases was caused by the post-WGD retention of dosage-sensitive genes while the non-dosage-balanced genes could be more safely lost (Makino and McLysaght 2010; Singh et al. 2012), and so we also incorporated analysis of how haploinsufficiency varies over evolutionary timescales. Our analysis shows that the older and more conserved paralogs are more likely to be haploinsufficient, as are the ohnologs. The ohnologs are consistently predicted to be likely to be haploinsufficient until two taxa after the whole genome duplications at approximately *Euteleostomi*, after which the predicted haploinsufficiency of newly created genes decreases sharply. The cumulative frequencies of retained genes associated with recessive and dominant diseases also plateau at this time because the overall rate of gene retention falls significantly, but the ratio of dominant to recessive disease-associated genes changes at *Vertebrata* from an excess of recessive diseases to a deficit, consistent with the post-WGD retention of dominant disease-associated ohnologs.

These changes in the relative burdens of recessive and dominant diseases are consistent with the previously advanced theories of post-WGD retention of dominant disease genes due to dosage sensitivity (Makino et al. 2009; Makino and McLysaght 2010). However, unlike these earlier studies which found that few ohnologs experienced subsequent duplications, we find instead that two thirds of the ohnologs have been duplicated in taxa more recent than *Vertebrata*. It will be interesting to investigate this further in an effort to understand whether this is because of the now much greater number of species in the Ensembl cross-species gene tree than available to the previous studies or whether this is because FFS assigns ages based on sequence divergence in addition to paralog duplication topology. However, since it seems that the ohnologs do experience frequent duplications after all, they are not therefore generally sensitive to dosage gains. We propose instead that the greater haploinsufficiency of the ohnologs is sufficient to explain their retention rather than a more general dosage sensitivity, a subtly different and more parsimonious explanation than previous hypothesised (Makino and McLysaght 2010).

5.4 MicroRNA derepression occurs widely in cancer

We have already seen that buffering of gene dosage changes must be key to the survival of cancer cells. Our finding that the bulk of experimentally confirmed miRNA/target interactions are between genes created in the whole genome duplications contemporaneously with the dominant disease-associated ohnologs led us to focus on miRNA-mediated buffering of gene dosage changes in the cancer cell.

We reasoned that miRNAs which are pivotal in oncogenesis would tend to be consistently affected by dosage changes such as CNVs, but we also realised that miRNAs expressed from adjacent loci would tend to experience the same CNVs simply as a function of their close genomic proximity. We grouped the miRNAs into distinct seed/locus families to avoid multiple counting miRNAs and were therefore able to accurately determine which miRNAs are consistently affected by CNVs in cell lines and so might be under selective pressure in cancer.

Seed/locus families containing multiple miRNA precursors are lost significantly more often than families with just one precursor but have a lower range of gains, consistent with selection pressure acting to deregulate a wide variety of processes in cancer cells (Kumar et al. 2007). We also observe miRNA biogenesis pathway disruption in the majority of cell lines which implies a widespread deregulation of processes caused by global miRNA depletion, previously shown to lead to increased cell migration and tumour growth (Kumar et al. 2007; Lin and Gregory 2015). Additionally, seed/locus families without redundant copies of their precursors elsewhere in the genome are lost more often than expected, also consistent with the hypothesis that tumour cells benefit from relaxing miRNA regulation.

We have so far only considered CNVs in cell lines and have not yet added transcriptomic data to our models, because we wanted to carefully consider the effects of CNVs alone. Adding expression data is a priority for future work. While we expect CNVs to broadly correlate with mRNA levels in cell lines (Stingele et al. 2012), the level of expressed protein is frequently decoupled from the number of copies of a gene, not least because of miRNA-mediated post-transcriptional tuning of protein expression (Sheltzer and Amon 2011).

5.5 *Oncomir-1* plays a significant role in sustaining the cancer phenotype

The largest group of miRNA seed/locus families with an identical pattern of CNVs across the cell lines that is not explained by genomic proximity contains *mir-106a~363* on chromosome X. This cluster is a paralog of the *mir-17~92* oncomir cluster on chromosome 13, which is better known as *Oncomir-1*. The miRNAs in these oncomir clusters are enriched for cancer-related processes such as control of the cell cycle, apoptosis and signalling. Another cluster on chromosome 13, which contains *mir-15a* and *mir-16-1*, tumour suppressor miRNAs which target *BCL2* and induce apoptosis (Cimmino et al. 2005), is similarly mainly affected by losses. Interestingly, while the paralogs of these miRNAs, *mir-15b* and *mir-16-2* on chromosome 3, are also enriched for signalling and cell cycle control, they are differently affected by CNVs with gains in ten cell lines but with no losses.

In order to understand these seemingly contradictory CNV patterns we narrowed our focus to the CNVs of just oncomirs and tumour suppressor miRNAs. The most striking result is that *Oncomir-1* on chromosome 13 and its paralog on chromosome X are never lost together in the same cell line, suggesting that at least some of the miRNAs expressed from these polycistronic miRNA clusters are essential to cancer cells. While we found potentially compensatory losses of *Xist* in every female-derived cell line in which the chromosome X *Oncomir-1* paralog was also partially lost, this only offers a possible explanation for selection pressure in cancer cells to maintain X chromosome gene dosage, and does not address the question of why the two *Oncomir-1* paralogs are never simultaneously lost even partially.

We constructed a detailed interaction network of the *Oncomir-1* paralogs, including an additional partial paralog on chromosome 7, together with the transcription factors known to interact with these miRNA clusters. Grouping the miRNAs into four seed-based families revealed that the transcription factors and seed families form auto-regulatory feedback loops, where a transcription factor activates transcription of a miRNA, but the miRNA represses the translation of the transcription factor. The most central to the *Oncomir-1* interaction network of these auto-regulatory loops are those involving *TP53*, which is also repressed by the *mir-15/16* paralogs discussed above.

Even partial loss of any *Oncomir-1* miRNA seed is balanced by the gain of an *Oncomir-1* activating transcription factor or the loss of an *Oncomir-1* inactivating transcription factor in all but one of the cell lines (prostate cancer-derived DU145), further indicating the centrality of *Oncomir-1* and its paralogs to *TP53* repression. Furthermore, there are CNVs affecting *Oncomir-1* either directly or via its transcription factors which are consistent with increased *Oncomir-1* expression and hence *TP53* repression in the vast majority of cell lines (91%).

The partial *Oncomir-1* paralog on chromosome 7 only contains precursors of miRNAs with the *miR-17* and *miR-92* seeds. The chromosome 7 paralog can be lost at the same time as the two main paralogs on chromosomes 13 and X, which suggests that it is the miRNAs with the seeds which are *not* expressed from the chromosome 7 paralog, *miR-18* and *miR-19*, which are under the most selection pressure to be retained in cancer cells. The *miR-19* family miRNAs contain the second-most central seed in the auto-regulatory network and miRNAs containing this seed, in addition to targeting *TP53*, are also known to target *PTEN*, a tumour suppressor which negatively regulates the Akt signalling pathway (Mu et al. 2009; Olive et al. 2009).

C-MYC is the most consistently affected transcription factor in *Oncomir-1*'s interaction network, gained in 25 out of 55 cell lines, and further analysis of *C-MYC*'s interactions revealed that in addition to directly activating transcription of *Oncomir-1*, *C-MYC* also activates *mir-663a* and *mir-1228* which, indirectly via *TP53* repression, act to derepress *Oncomir-1*. Since full *TP53* expression would prevent the activation of *Oncomir-1* it seems likely that transient *C-MYC* expression would, acting via *mir-663a/1228*, repress *TP53* while simultaneously activating *Oncomir-1*, thus converting a transient oncogenic signal from *C-MYC* overexpression into sustained *TP53* repression by *Oncomir-1*. Furthermore, while *C-MYC* overexpression can trigger apoptosis (Hoffman and Liebermann 2008), the resulting consistent expression of *Oncomir-1* would lead to down-tuning of *C-MYC* levels via the *miR-17*, *miR-19* and *miR-92* family miRNAs, protecting the cancer cell from the otherwise potentially apoptotic effects of *C-MYC* overexpression.

We therefore propose a novel mechanism whereby transient *C-MYC* elevation leads to *TP53* repression via *mir-663a/1228* for long enough that *Oncomir-1* can then repress *TP53* in a sustained manner, in addition to repressing *PTEN* via the *miR-19* family miRNAs and avoiding *C-MYC*-induced apoptosis. This network forms a bistable switch in the majority of cancers which once activated by *C-MYC* leads to sustained repression of *TP53* and *PTEN*, but which could be potentially reversed with *Oncomir-1* antagonists such as miRNA 'sponges' (Ebert et al. 2007), in order to sequester the miRNAs expressed from *Oncomir-1* and its paralogs.

Future work could investigate this mechanism on two fronts. Firstly, as a proof-of-concept, we will transfect tumour cell line cultures and matching normal tissue cell cultures with miRNA sponges which sequester various combinations of the miRNAs expressed from *Oncomir-1*, to confirm whether the mechanism functions as we expect *in vivo*. In parallel with these experiments, we will construct a system of ordinary differential equations to model the network's response to various perturbations such as CNVs and miRNA sponges. This systems biology approach will guide the experimental investigation, which can in turn be used to update the theoretical model, enabling us to refine our knowledge of this system of *Oncomir-1* interactions in a broad range of cancers.

5.6 Conclusions

Taken as a whole, the work presented in this thesis advances our understanding of the role of miRNAs in buffering the gene dosage changes which arise from copy number variations in cancer. We have shown that read depth analysis of whole genome sequencing data reveals detailed CNVs and hence gene dosage changes in the NCI-60 cell lines, and we have derived a list of candidate novel driver genes for further investigation. We have also been able to add to the understanding of the mechanisms behind the retention of ancient disease-associated paralogs in the human genome. Our investigation of the dosage changes affecting miRNAs and their targets has enabled us to propose a novel mechanism, potentially widespread across cancer types, whereby the interplay between *Oncomir-1* and its paralogs, transcription factors and other miRNAs comprise a bistable switch in oncogenesis. When this system is in the conformation favourable to cancer progression it sustains *TP53* and *PTEN* repression and enables the avoidance of apoptosis. We believe this switch could be potentially reversed with *Oncomir-1* antagonists, leading to regained control over the cell cycle and thereby inducing apoptosis in *Oncomir-1*-dependent cancer cells.

6 Supplementary Information

Accession	Name	Read length	Sex	Tissue	Disease	Age	Coverage
SRR4009203	UACC62	100	female	Skin	Melanoma		0.82
SRR4009225	H322M	100	male	Lung	Lung	52	2.2
SRR4009236	OVCAR4	100	female	Ovary	Ovary	42	1.6
SRR4009238	MDA-MB-231	100	female	Breast	Breast	51	1.4
SRR4009239	HOP62	100	female	Lung	Lung	60	1.8
SRR4009240	RPMI-8226	100	male	Hematopoietic	Hematopoietic	61	0.92
SRR4009241	RXF-393	100	male	Renal	Renal	54	1.5
SRR4009242	SF539	100	female	Brain	Glioblastoma	34	1.8
SRR4009243	BT549	100	female	Breast	Breast	72	1.9
SRR4009245	CAKI-1	100	male	Renal	Renal	49	2.2
SRR4009246	OVCAR5	100	female	Ovary	Ovary	67	0.97
SRR4009247	H23	100	male	Lung	Lung	51	1.4
SRR4009248	H460	100	male	Lung	Lung		2.1
SRR4009249	Hs578T	100	female	Breast	Breast	74	1.4
SRR4009250	SNB75	100	female	Brain	Glioblastoma	78	1.7
SRR4009251	TK10	100	male	Renal	Renal	43	1.1
SRR4009252	SF268	100	female	Brain	Glioblastoma	24	1.9
SRR4009253	M14	100	female	Skin	Melanoma		2.3
SRR4009254	SKMEL2	100	male	Skin	Melanoma	60	1.9
SRR4009256	786-0	100	male	Renal	Renal	58	2.8
SRR4009258	UACC257	100	female	Skin	Melanoma		2.9
SRR4009259	T47D	100	female	Breast	Breast	54	1.9
SRR4009260	SKMEL-28	100	male	Skin	Melanoma	51	1.4
SRR4009261	UO31	100	female	Renal	Renal		0.94
SRR4009262	HCC2998	100	female	Colon	Colon		1.1
SRR4009263	OVCAR3	100	female	Ovary	Ovary	60	0.76
SRR4009265	H522	100	male	Lung	Lung	60	2.6
SRR4009267	KM12	100	female	Colon	Colon		1.2
SRR4009270	DU145	100	male	Prostate	Prostate	69	0.81
SRR4009272	SKMEL5	100	female	Skin	Melanoma	24	1.1
SRR4009273	MCF7	50	female	Breast	Breast	69	0.41
SRR4009274	HOP92	100	female	Lung	Lung	62	2.4
SRR4009275	OVCAR8	100	female	Ovary	Ovary	64	1.7
SRR4009277	PC3	100	male	Prostate	Prostate	62	2
SRR4009278	SKOV3	100	female	Ovary	Ovary	64	0.94
SRR4009279	H226	100	male	Lung	Lung		2.9
SRR4009280	Sw620	100	male	Colon	Colon	51	1.7
SRR4009281	IGROV1	100	female	Ovary	Ovary	47	1.1
SRR4009282	A549	100	male	Lung	Lung	58	0.97
SRR4009283	MOLT-4	100	male	Hematopoietic	Hematopoietic	19	1.1
SRR4009284	HT29	100	female	Colon	Colon	44	0.83
SRR4009286	HCT15	100	male	Colon	Colon		1.1
SRR4009287	HCT116	100	male	Colon	Colon		1
SRR4009289	SN12C	100	male	Renal	Renal	43	2.2
SRR4009290	HL-60	100	female	Hematopoietic	Hematopoietic	36	1.1
SRR4009291	CCRF-CEM	100	female	Hematopoietic	Hematopoietic	4	1.2
SRR4009292	A498	100	female	Renal	Renal	52	0.86
SRR4009293	COLO205	100	male	Colon	Colon	70	1
SRR4009294	LOXIMVI	100	male	Skin	Melanoma	58	1.1
SRR4009295	ACHN	100	male	Renal	Renal	22	2.4
SRR4009296	SR	100	male	Hematopoietic	Hematopoietic	11	1.1
SRR4009297	K562	100	female	Hematopoietic	Hematopoietic	53	1.1
SRR4009310	EKVX	100	male	Lung	Lung		3.2
SRR4009321	MALME3M	100	male	Skin	Melanoma	43	2.3
SRR4009329	SF295	100	female	Brain	Glioblastoma	67	1.9

Table 6.1 - NCI-60 cell lines

The 55 NCI-60 cell lines analysed in this thesis. The accession numbers refer to sequencing runs deposited in the Sequence Read Archive (Sayers et al. 2020) under project accession PRJNA338012 (Turner et al. 2017). The name, read length, sex, tissue of origin, disease type, age and sequencing coverage fields were curated by manual inspection of the SRA information pages for each sequencing run.

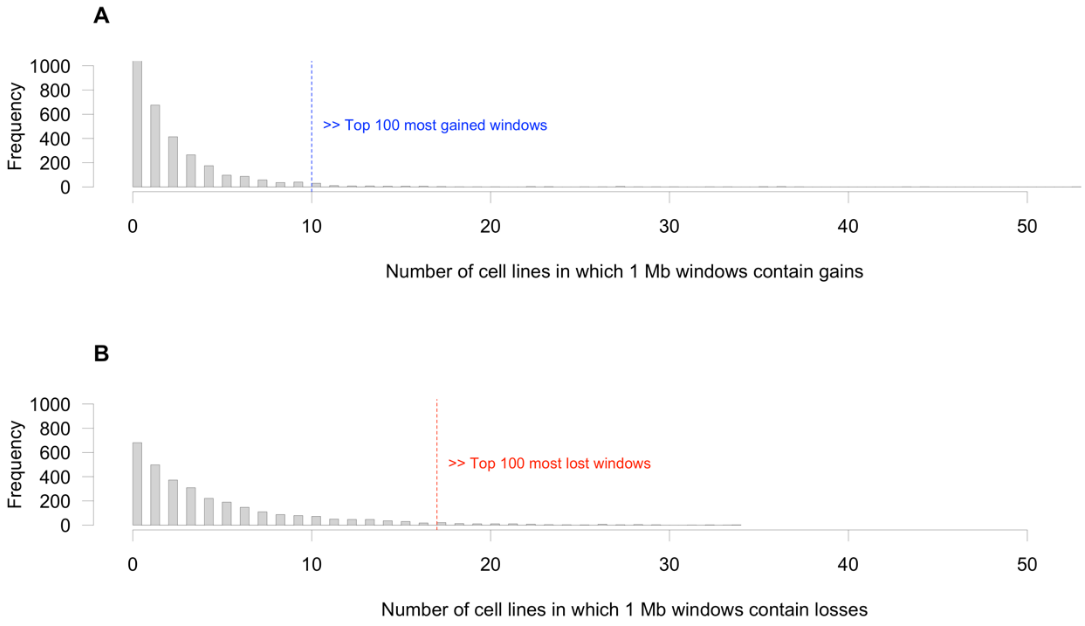


Figure 6.1 - The distributions of cell line gains and losses for 1 Mb sliding windows

The distributions of cell line gains (**A**) and losses (**B**) for 1 Mb sliding windows across the entire genome. The vertical dashed lines are the thresholds at or above which the 1 Mb windows contain enough cell lines with gains (**A**) or losses (**B**) to be included in the top 100 gain or loss CNV hotspots.

GO term	Enrichment
negative regulation of monocyte differentiation (GO:0045656)	> 100
hepatocyte growth factor receptor signaling pathway (GO:0048012)	> 100
positive regulation of microtubule polymerization (GO:0031116)	> 100
positive regulation of binding (GO:0051099)	34.91
regulation of cellular response to stress (GO:0080135)	14.92
chromosome organization (GO:0051276)	9.71
negative regulation of gene expression (GO:0010629)	7.14
interspecies interaction between organisms (GO:0044419)	6.78
response to external stimulus (GO:0009605)	5.69
regulation of transcription by RNA polymerase II (GO:0006357)	5.51

Table 6.2 - Enriched biological process GO terms for gain hotspots

The biological process GO terms that are enriched for gain hotspots in the NCI-60 cell lines. All terms are significant at FDR P < 0.05.

Reactome pathway	Enrichment
Sema4D mediated inhibition of cell attachment and migration (R-HSA-416550)	> 100
MET activates RAP1 and RAC1 (R-HSA-8875555)	> 100
Transcriptional regulation of granulopoiesis (R-HSA-9616222)	71.02
Estrogen-dependent gene expression (R-HSA-9018519)	69.81
Formation of the beta-catenin:TCF transactivating complex (R-HSA-201722)	68.65
Constitutive Signaling by Aberrant PI3K in Cancer (R-HSA-2219530)	66.44
RHO GTPases activate PKNs (R-HSA-5625740)	65.38
MAPK family signaling cascades (R-HSA-5683057)	22.39
Cell Cycle, Mitotic (R-HSA-69278)	16.68
Cytokine Signaling in Immune system (R-HSA-1280215)	10.01
Generic Transcription Pathway (R-HSA-212436)	8.61

Table 6.3 - Enriched Reactome pathways for gain hotspots

The Reactome pathways that are enriched for gain hotspots in the NCI-60 cell lines. All terms are significant at FDR P < 0.05.

GO term	Enrichment
common-partner SMAD protein phosphorylation (GO:0007182)	> 100
dorsal/ventral axis specification (GO:0009950)	> 100
paraxial mesoderm morphogenesis (GO:0048340)	> 100
negative regulation of telomerase activity (GO:0051974)	> 100
embryonic brain development (GO:1990403)	> 100
positive regulation of smooth muscle cell apoptotic process (GO:0034393)	> 100
negative regulation of production of miRNAs involved in gene silencing by miRNA (GO:1903799)	> 100
negative regulation of androgen receptor signaling pathway (GO:0060766)	> 100
retinoic acid receptor signaling pathway (GO:0048384)	> 100
granulocyte differentiation (GO:0030851)	86.53
response to immobilization stress (GO:0035902)	83.55
ribosomal large subunit assembly (GO:0000027)	80.76
cellular senescence (GO:0090398)	75.72
endoderm development (GO:0007492)	48.46
positive regulation of cell cycle arrest (GO:0071158)	43.27
activation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0006919)	40.38
transforming growth factor beta receptor signaling pathway (GO:0007179)	36.71
negative regulation of angiogenesis (GO:0016525)	34.61
regulation of BMP signaling pathway (GO:0030510)	34.61
transcription initiation from RNA polymerase II promoter (GO:0006367)	26.63
stem cell population maintenance (GO:0019827)	25.78
negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0090101)	25.59
cell cycle arrest (GO:0007050)	25.42
regulation of ubiquitin-dependent protein catabolic process (GO:2000058)	22.16
negative regulation of cell growth (GO:0030308)	19.54
cell fate commitment (GO:0045165)	19.46
rhythmic process (GO:0048511)	17.75
negative regulation of catabolic process (GO:0009895)	15.48
negative regulation of cell population proliferation (GO:0008285)	13.89
gland development (GO:0048732)	11.65
heart development (GO:0007507)	11.56
chordate embryonic development (GO:0043009)	9.87
tube morphogenesis (GO:0035239)	9.12
regulation of cellular response to stress (GO:0080135)	8.78
negative regulation of transcription by RNA polymerase II (GO:0000122)	7.87
regulation of protein phosphorylation (GO:0001932)	7.45
protein modification by small protein conjugation or removal (GO:0070647)	7.08
positive regulation of transcription by RNA polymerase II (GO:0045944)	6.99
positive regulation of protein modification process (GO:0031401)	6.82
epithelium development (GO:0060429)	6.48
positive regulation of developmental process (GO:0051094)	5.69
regulation of intracellular signal transduction (GO:1902531)	4.77

Table 6.4 - Enriched biological process GO terms for loss hotspots

The biological process GO terms that are enriched for loss hotspots in the NCI-60 cell lines. All terms are significant at FDR P < 0.05.

Reactome pathway	Enrichment
Loss of MECP2 binding ability to the NCoR/SMRT complex (R-HSA-9022537)	> 100
NR1H2 & NR1H3 regulate gene expression to control bile acid homeostasis (R-HSA-9623433)	> 100
Downregulation of SMAD2/3:SMAD4 transcriptional activity (R-HSA-2173795)	> 100
FOXO-mediated transcription of cell cycle genes (R-HSA-9617828)	> 100
Constitutive Signaling by AKT1 E17K in Cancer (R-HSA-5674400)	93.19
Nuclear Receptor transcription pathway (R-HSA-383280)	88.64
Notch-HLH transcription pathway (R-HSA-350054)	86.53
Regulation of MECP2 expression and activity (R-HSA-9022692)	80.76
FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes (R-HSA-9615017)	80.76
NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux (R-HSA-9029569)	67.30
Transcriptional regulation of white adipocyte differentiation (R-HSA-381340)	58.38
NOTCH1 Intracellular Domain Regulates Transcription (R-HSA-2122947)	53.84
PPARA activates gene expression (R-HSA-1989781)	42.51
HCMV Early Events (R-HSA-9609690)	36.71
SUMO E3 ligases SUMOylate target proteins (R-HSA-3108232)	23.15
PIP3 activates AKT signaling (R-HSA-1257604)	19.54
Ub-specific processing proteases (R-HSA-5689880)	17.82
Transcriptional regulation by RUNX1 (R-HSA-8878171)	17.82

Table 6.5 - Enriched Reactome pathways for loss hotspots

The Reactome pathways that are enriched for loss hotspots in the NCI-60 cell lines. All terms are significant at FDR P < 0.05.

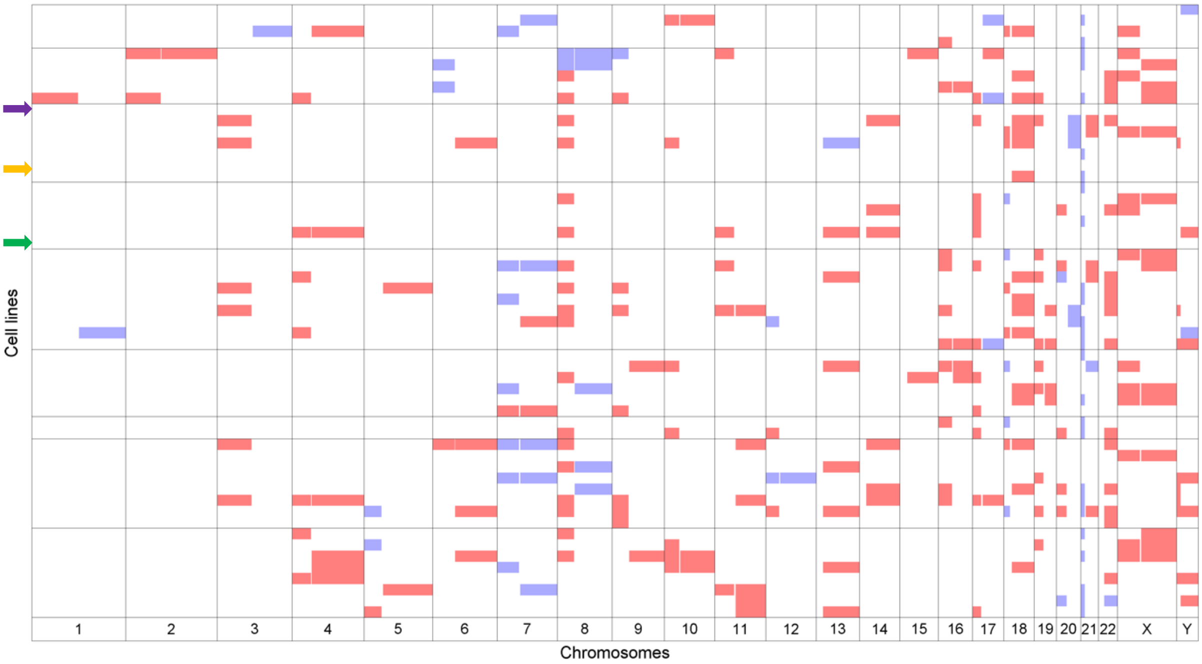


Figure 6.2 - Chromosome arm-level aneuploidies

The arm-level aneuploidies affecting each cell line of the NCI-60 panel with gains shown in blue and losses in red. The three cell lines with no aneuploidies, HCC2998, HCT15 and SR, are marked with purple, orange and green arrows respectively.

miRNAs	Chromosomes	Seeds	# Precursors	Max Spread
miR-23/24	9, 19	GCCUACU, GGCUCAG, GGGUUCC	4	881
miR-193/365	16, 17	AAUGCCC, ACUGGCC, GGGACUU	4	15,526
miR-4253/4684/6862	1, 16	GGGCAUG, GUUGCAA, UCUCUAC	4	333,340
miR-33/6777/6889	17, 22	CGGGGAG, UGCAUUG	4	648,021
miR-744/10396	17, 21	GCGGGGC, UGUUGCC	2	98
miR-9	1, 5, 15	CUUUGGU, UAAAGCU	3	90
miR-192/194/215	1, 11	GUAACAG, UGACCUA	4	389
miR-19	13, X	GUGCAAA, GUUUUGC	3	388
miR-6784/6862	16, 17	CCGGGGC, CUCACCC	3	333,340
miR-103	5, 20	CAUAGCC, GCUUCUU	4	78
miR-30	1, 6, 8	GUAAACA, UUUCAGU	6	26,662

Table 6.6 - miRNA seed clusters on different chromosomes

Groups of mature miRNA seeds with identical CNV profiles across the NCI-60 panel with precursor miRNAs that occur on different chromosomes and so cannot be directly affected by the same CNVs. The field ‘Max Spread’ is the largest region spanned by any two adjacent miRNA precursors on each chromosome.

miRNAs	Chromosomes	# Seeds	# Precursors	Max Spread
miR-548/4779	2	2	2	50,721,704
miR-561/4785/6888	2	5	3	14,555,377
miR-3126/3682/5000	2	4	3	10,620,896
miR-3936/6830/12130	5	4	3	8,054,928
miR-4477	9	2	4	7,528,633
miR-1303/3141/12125	5	3	3	6,784,959
miR-146/5003/10523	5	4	3	6,214,904
miR-448/504/764/1264/1298/1911/1912	X	9	7	4,004,726
miR-4431/4433/5192	2	6	4	3,879,442
miR-579/580	5	3	2	3,753,611
miR-548/583/2277	5	4	3	3,597,935
miR-215/664/6741	1	4	3	2,902,145
miR-1225/6511	16	3	7	2,708,981
miR-3691/5683/5689	6	4	3	2,645,781
miR-558/4263/4765	2	3	3	2,319,482
miR-218/12115	4	2	2	1,987,111
miR-18/20/92/106/424/450/503/505/542/934/1277	X	18	13	1,794,820
miR-3655/5197/6831	5	4	3	1,558,150
miR-449/582/5687	5	5	4	1,511,019
miR-339/4648	7	3	2	1,504,213
miR-4636/10397	5	3	2	1,348,616
miR-4259/5187	1	3	2	1,327,283
miR-101/4665	9	2	2	1,157,608
miR-4772/5696	2	3	2	1,122,917
miR-15/3919	3	2	2	1,122,040
miR-3925/5690	6	3	2	957,796
miR-4451/4452/5705	4	3	3	789,058
let-7, miR-548/6125/10527	12	4	4	787,416
miR-3679/5590	2	2	2	730,749
miR-6782/6783	17	3	2	726,911
miR-630/12135	15	2	2	712,763
miR-5584/6079	1	3	2	706,931
miR-198/5682	3	2	2	654,048
miR-33/6777/6889	17, 22	2	4	648,021
miR-3121/12116	1	3	2	634,467
miR-567/9900	3	2	2	624,381

Table 6.7 - High-spread miRNA seed clusters with precursors in different loci

Groups of mature miRNA seeds with identical CNV profiles across the NCI-60 panel with precursor miRNAs that occur in different loci and are far enough apart that they are unlikely to be directly affected by the same CNVs. The field ‘Max Spread’ is the largest region spanned by any two adjacent miRNA precursors on each chromosome.

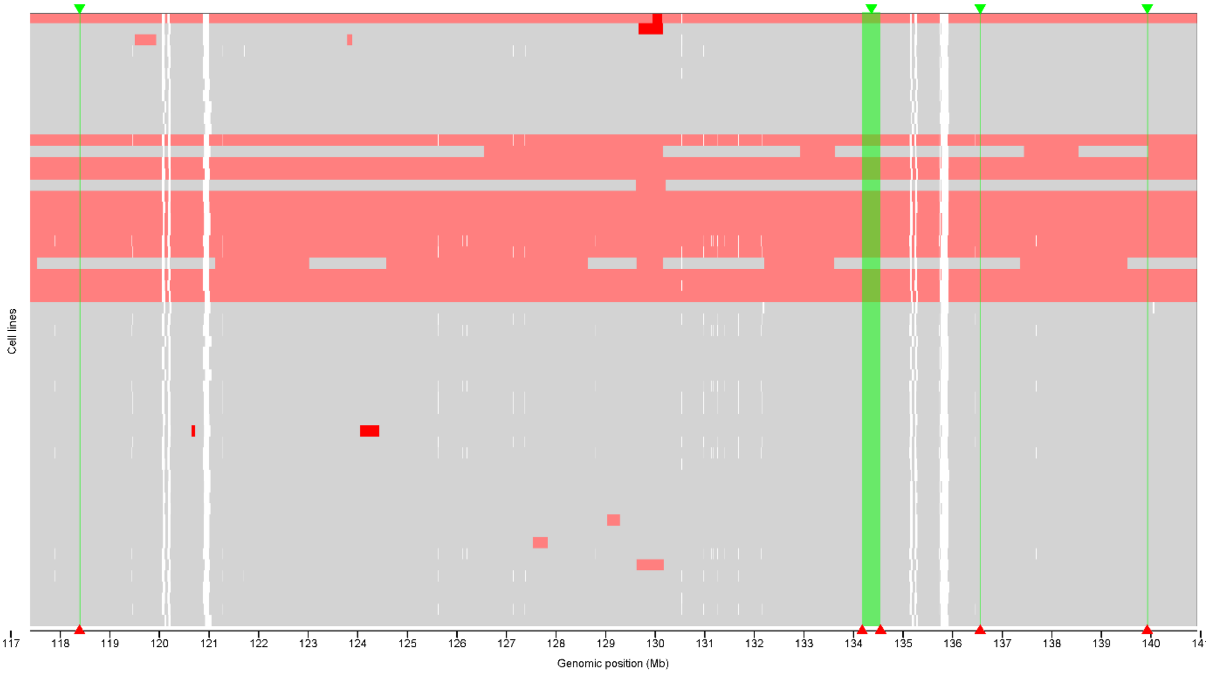


Figure 6.3 - A cluster of seed/locus families affected by different CNVs

The CNVs that affect the fifth largest cluster (in Figure 3.5C) of 26 miRNA seed/locus families with identical CNV profiles on chromosome X despite being in four non-adjacent loci spread across 2.2 x 10^7 nucleotides and despite being affected by different CNVs in some of the cell lines. The rows are individual cell lines (in the same order as in Figure 3.5C) with regions that are completely lost shown in dark red, partially lost in light red, unaffected in grey and unmappable in white (there are no gains in this area of chromosome X). The four seed/locus family loci are shown as green highlights topped with green indicator triangles at the midpoint and the actual miRNA precursors that form the boundaries of the loci are indicated with red triangles below the loci. The x axis is the genomic position within chromosome X.

miRNAs	Chromosome	# Seed/loci	# Loci	Spread
miR-18/19/20/92/106/363/424/450/503/505/542/934/1277	X	26	4	21,537,838
miR-506/507/508/509/510/513/514/888/890/891/892	X	26	2	1,291,979
miR-448/504/764/1264/1298/1911/1912/3672	X	12	3	24,028,359
miR-194/215/320/664/4742/6741	1	10	3	5,804,290
miR-325/374/384/421/545	X	10	2	2,787,125
miR-449/548/582/5687	5	9	3	4,533,172
miR-105/224/452/767/4330	X	9	2	1,226,271
miR-1285/4431/4433/4434/5192	2	8	4	17,550,484
miR-15/16/3613/4703/5693	13	8	2	1,556,253
miR-4315/6782/6783/6784	17	7	2	1,267,665
miR-548/4430/4435/4771/4780	2	6	3	54,664,084
miR-3126/3682/5000	2	6	3	21,241,793
miR-568/4446/4796/8076	3	6	2	1,311,408
miR-561/4785/6888	2	5	3	29,110,754
miR-1289/3936/6830/12130	5	5	3	17,171,964
miR-3691/5683/5689/7853	6	5	3	5,291,562
miR-3655/5197/6831	5	5	2	3,116,301
miR-548/559/4264/8080	2	4	3	32,488,909
miR-874/3661/5692	5	4	3	3,421,893
miR-4462/5690/7111	6	4	2	2,084,915
miR-361/548/1321	X	4	2	1,677,956
miR-101/4665	9	4	2	1,157,608
miR-640/3188/3189	19	4	2	1,153,082
miR-15/16/3919	3	4	2	1,122,180
miR-554/4257/6878	1	4	2	1,053,547
miR-198/5682/6529	3	4	2	965,152
miR-4779/6071/8485	2	3	2	35,496,937
miR-558/4263/4765	2	3	2	4,638,965
miR-3124/3916	1	3	2	1,624,477
miR-4259/5187	1	3	2	1,327,283
miR-4772/5696	2	3	2	1,122,917
miR-5586/9718	14	3	2	1,001,750
miR-613/614/1244	12	3	2	803,967
miR-320/548	X	2	2	34,286,418
miR-630/12135	15	2	2	712,763
miR-567/9900	3	2	2	624,381

Table 6.8 - Seed/locus families with the same CNV profile despite differing CNVs

Groups of miRNA seed/locus families with identical CNV profiles across the NCI-60 panel but with precursor miRNAs that occur in different loci and are far enough apart that they are unlikely to be directly affected by the same CNVs and which are also affected by different CNVs in each cell line. The miRNAs highlighted in red also occur in the seed-based miRNA clusters detailed in Supplementary Table 6.7.

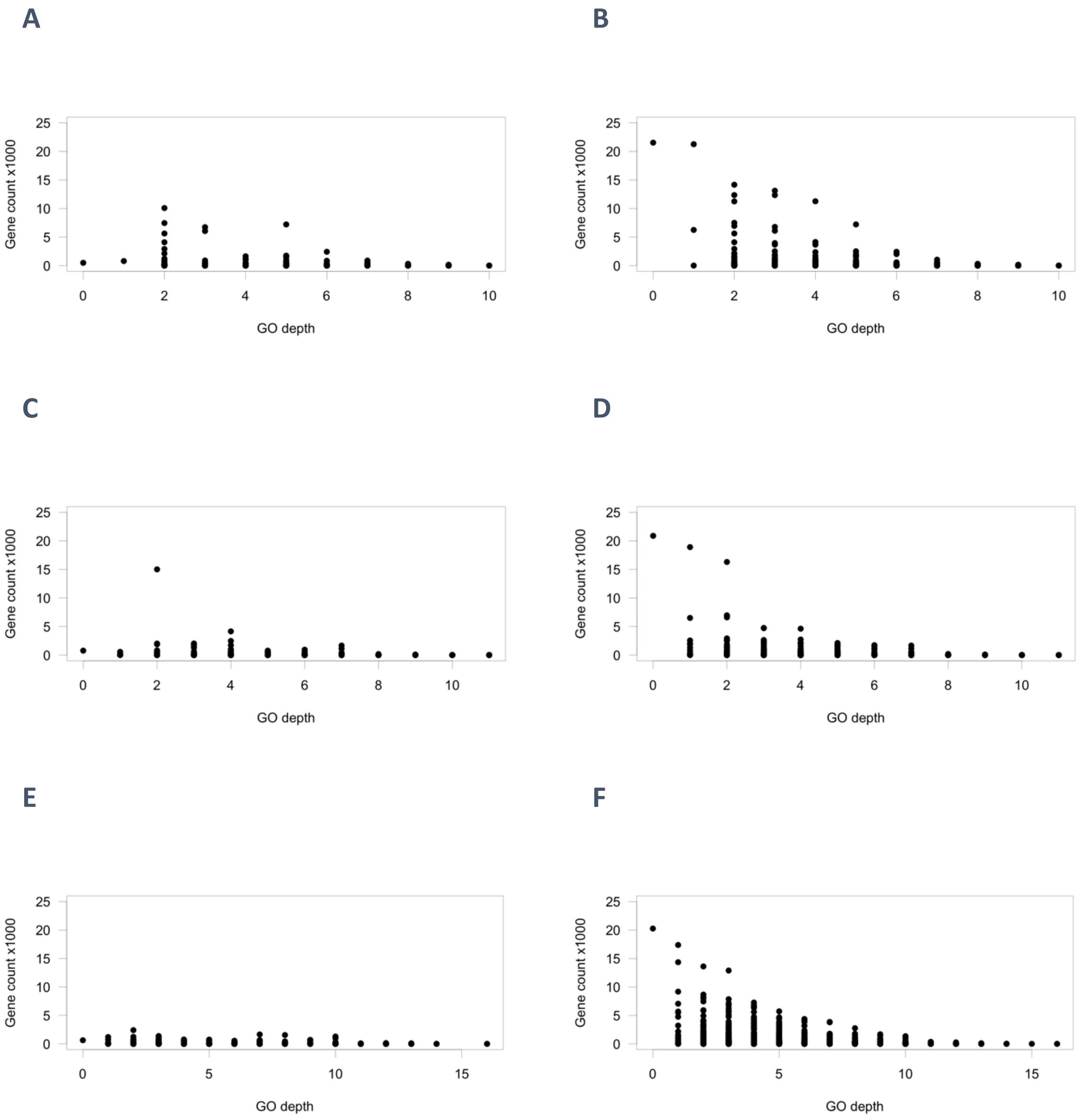


Figure 6.4 - Gene/GO mappings as annotated in Ensembl and after 'cascading'

The numbers of genes allocated to categories at each depth of the cellular compartment gene ontology, (**A**) as annotated by Ensembl and (**B**) after traversing the ontology depth-first and assigning genes to their ancestral categories as well, resulting in a 'cascaded' ontology. (**C**) Annotated and (**D**) cascaded gene counts at each depth of the molecular function gene ontology. (**E**) Annotated and (**F**) cascaded gene counts at each depth of the biological process gene ontology.

Category	**Keywords**
Metastasis	metastasis, migration, proliferation, epithelial to mesenchymal transition, motility, cell adhesion
Cell cycle	cell cycle, apoptosis, apoptotic, cell death, mitotic, mitosis, checkpoint, transforming growth factor, cell growth, meiosis, meiotic, DNA repair, p53, DNA replication, senescence, PTEN, AKT, checkpoint
Expression	expression, transcription, translation
Signalling	signaling, BMP, Wnt, SMAD, MAPK, TGF, transforming growth factor
Development	development, angiogenesis, axonogenesis, morphogenesis, vasculogenesis

Table 6.9 - GO and Reactome functional keywords

Selected keywords occurring in GO term and Reactome pathway descriptions related to five broad cancer-related categories. Keyword spellings are American to match the descriptions.

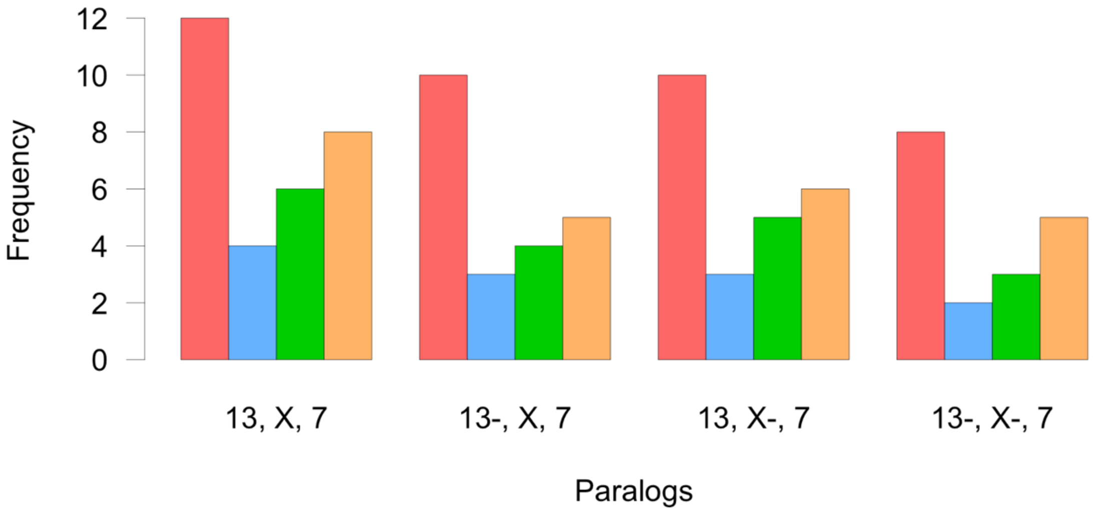


Figure 6.5 - Seed dosage of* Oncomir-1 *paralogs in various CNV conditions

The numbers of copies of each distinct seed in the *Oncomir-1* paralogs under different CNV conditions. Each group is labelled with the names of the chromosomes on which the paralogs occur, followed with a minus sign to indicate which paralog has lost a single copy. The red bars indicate *miR-17* seed copies, blue are *miR-18* seed copies, green are *miR-19* seed copies and orange are *miR-92* seed copies.

Name	Age (years x 10^6)
Opisthokonta	1,105.1
Bilateria	796.6
Chordata	676.4
Vertebrata	615
Gnathostomata	473.3
Euteleostomi	435.3
Sarcopterygii	413
Tetrapoda	351.8
Amniota	311.9
Mammalia	176.9
Theria	158.6
Eutheria	105.5
Boreoeutheria	96.5
Euarchontoglires	82.1
Primates	73.8
Haplorrhini	67.1
Simiiformes	43.2
Catarrhini	29.4
Hominoidea	20.2
Hominidae	15.8
Homininae	9.1
Homo sapiens	0

Table 6.10 - Unique Ensembl Compara taxon names and approximate ages

The unique taxon names and approximate duplication ages (as determined by Ensembl) extracted from the gene event histories downloaded from Ensembl Compara as part of the FFS tree building algorithm.

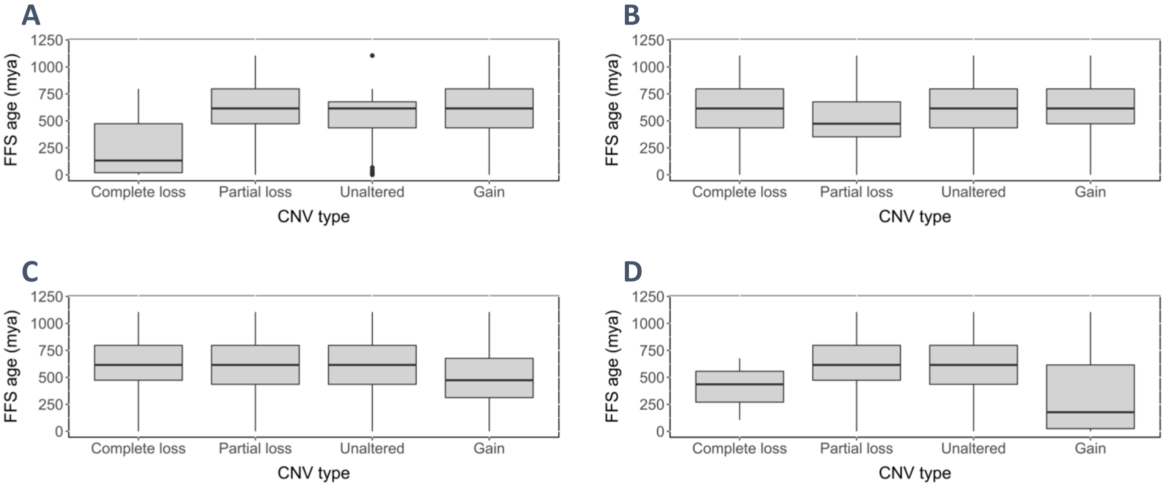


Figure 6.6 - Examples of cell lines with gene ages which vary with CNV type

(**A**) Genes which are completely lost in renal cancer cell line CAKI-1 are younger than the rest. (**B**) Genes which are partially lost in colon cancer cell line KM12 are younger than the rest. (**C**) Genes which are gained in prostate cancer cell line PC3 are younger than the rest. (**D**) Genes which are completely lost or gained in ovarian cancer cell line IGROV1 are younger than the rest.

References

1000 Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. 'A map of human genome variation from population-scale sequencing', *Nature*, 467: 1061-73.

Abbott, A. L., E. Alvarez-Saavedra, E. A. Miska, N. C. Lau, D. P. Bartel, H. R. Horvitz, and V. Ambros. 2005. 'The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans', *Dev Cell*, 9: 403-14.

Ambros, V., R. C. Lee, A. Lavanway, P. T. Williams, and D. Jewell. 2003. 'MicroRNAs and other tiny endogenous RNAs in C. elegans', *Curr Biol*, 13: 807-18.

Andrews, Simon. 2018. 'FastQC v0.11.7', Accessed 6/4/18. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aravin, A. A., G. J. Hannon, and J. Brennecke. 2007. 'The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race', *Science*, 318: 761-4.

Aravin, A. A., M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl. 2003. 'The small RNA profile during Drosophila melanogaster development', *Dev Cell*, 5: 337-50.

Archibald, D. J. 2003. 'Timing and biogeography of the eutherian radiation: fossils and molecules compared', *Mol Phylogenet Evol*, 28: 350-9.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nat Genet*, 25: 25-9.

Asher, G., N. Reuven, and Y. Shaul. 2006. '20S proteasomes and protein degradation "by default"', *Bioessays*, 28: 844-9.

Auyeung, V. C., I. Ulitsky, S. E. McGeary, and D. P. Bartel. 2013. 'Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing', *Cell*, 152: 844-58.

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: 64-71.

Bartel, D. P. 2004. 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell*, 116: 281-97.

———. 2009. 'MicroRNAs: target recognition and regulatory functions', *Cell*, 136: 215-33.

Bartel, D. P., and C. Z. Chen. 2004. 'Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs', *Nat Rev Genet*, 5: 396-400.

Baskerville, S., and D. P. Bartel. 2005. 'Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes', *RNA*, 11: 241-7.

Beaudet, A. L., and Y. H. Jiang. 2002. 'A rheostat model for a rapid and reversible form of imprinting-dependent evolution', *Am J Hum Genet*, 70: 1389-97.

Behm-Ansmant, I., J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde. 2006. 'mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes', *Genes Dev*, 20: 1885-98.

Benjamini, Yoav, and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*, 57: 289-300.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. Catenazzi M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. 2008. 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456: 53-9.

Benton, M. J., and F. J. Ayala. 2003. 'Dating the tree of life', *Science*, 300: 1698-700.

Benton, M. J., and P. C. Donoghue. 2007. 'Paleontological evidence to date the tree of life', *Mol Biol Evol*, 24: 26-53.

Berezikov, E. 2011. 'Evolution of microRNA diversity and regulation in animals', *Nat Rev Genet*, 12: 846-60.

Bernstein, C., H. Bernstein, C. M. Payne, and H. Garewal. 2002. 'DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis', *Mutat Res*, 511: 145-78.

Beroukhim, R., C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y. J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M. S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson. 2010. 'The landscape of somatic copy-number alteration across human cancers', *Nature*, 463: 899-905.

Bignell, G. R., C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, and M. R. Stratton. 2010. 'Signatures of mutation and selection in the cancer genome', *Nature*, 463: 893-8.

Birchler, J. A., and K. J. Newton. 1981. 'Modulation of protein levels in chromosomal dosage series of maize: the biochemical basis of aneuploid syndromes', *Genetics*, 99: 247-66.

Bishara, A., Y. Liu, Z. Weng, D. Kashef-Haghighi, D. E. Newburger, R. West, A. Sidow, and S. Batzoglou. 2015. 'Read clouds uncover variation in complex regions of the human genome', *Genome Res*, 25: 1570-80.

Blake, W. J., G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins. 2006. 'Phenotypic consequences of promoter-mediated transcriptional noise', *Mol Cell*, 24: 853-65.

Blakeslee, A. F., J. Belling, and M. E. Farnham. 1920. 'Chromosomal Duplication and Mendelian Phenomena in Datura Mutants', *Science*, 52: 388-90.

Bleazard, T., J. A. Lamb, and S. Griffiths-Jones. 2015. 'Bias in microRNA functional enrichment analysis', *Bioinformatics*, 31: 1592-8.

Blobe, G. C., W. P. Schiemann, and H. F. Lodish. 2000. 'Role of transforming growth factor beta in human disease', *N Engl J Med*, 342: 1350-8.

Bracken, C. P., P. A. Gregory, N. Kolesnikoff, A. G. Bert, J. Wang, M. F. Shannon, and G. J. Goodall. 2008. 'A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition', *Cancer Res*, 68: 7846-54.

Braun, F. N., and D. A. Liberles. 2003. 'Retention of enzyme gene duplicates by subfunctionalization', *Int J Biol Macromol*, 33: 19-22.

Brennecke, J., A. Stark, R. B. Russell, and S. M. Cohen. 2005. 'Principles of microRNA-target recognition', *PLoS Biol*, 3: e85.

Budczies, J., M. Bockmayr, C. Denkert, F. Klauschen, S. Groschel, S. Darb-Esfahani, N. Pfarr, J. Leichsenring, M. L. Onozato, J. K. Lennerz, M. Dietel, S. Frohling, P. Schirmacher, A. J. Iafrate, W. Weichert, and A. Stenzinger. 2016. 'Pan-cancer analysis of copy number changes in programmed death-ligand 1 (PD-L1, CD274) - associations with gene expression, mutational load, and survival', *Genes Chromosomes Cancer*, 55: 626-39.

Burk, U., J. Schubert, U. Wellner, O. Schmalhofer, E. Vincan, S. Spaderna, and T. Brabletz. 2008. 'A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells', *EMBO Rep*, 9: 582-9.

Cai, J. J., E. Borenstein, R. Chen, and D. A. Petrov. 2009. 'Similarly strong purifying selection acts on human disease genes of all evolutionary ages', *Genome Biol Evol*, 1: 131-44.

Cassidy, J. J., A. R. Jha, D. M. Posadas, R. Giri, K. J. Venken, J. Ji, H. Jiang, H. J. Bellen, K. P. White, and R. W. Carthew. 2013. 'miR-9a minimizes the phenotypic impact of genomic diversity by buffering a transcription factor', *Cell*, 155: 1556-67.

Chaffer, C. L., and R. A. Weinberg. 2011. 'A perspective on cancer cell metastasis', *Science*, 331: 1559-64.

Chaisson, M. J., J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, and E. E. Eichler. 2015. 'Resolving the complexity of the human genome using single-molecule sequencing', *Nature*, 517: 608-11.

Chen, K., and N. Rajewsky. 2007. 'The evolution of gene regulation by transcription factors and microRNAs', *Nat Rev Genet*, 8: 93-103.

Chiang, D. Y., G. Getz, D. B. Jaffe, M. J. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. 2009. 'High-resolution mapping of copy-number alterations with massively parallel sequencing', *Nat Methods*, 6: 99-103.

Chou, C. H., S. Shrestha, C. D. Yang, N. W. Chang, Y. L. Lin, K. W. Liao, W. C. Huang, T. H. Sun, S. J. Tu, W. H. Lee, M. Y. Chiew, C. S. Tai, T. Y. Wei, T. R. Tsai, H. T. Huang, C. Y. Wang, H. Y. Wu, S. Y. Ho, P. R. Chen, C. H. Chuang, P. J. Hsieh, Y. S. Wu, W. L. Chen, M. J. Li, Y. C. Wu, X. Y. Huang, F. L. Ng, W. Buddhakosai, P. C. Huang, K. C. Lan, C. Y. Huang, S. L. Weng, Y. N. Cheng, C. Liang, W. L. Hsu, and H. D. Huang. 2018. 'miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions', *Nucleic Acids Res*, 46: D296-D302.

Christodoulou, F., F. Raible, R. Tomer, O. Simakov, K. Trachana, S. Klaus, H. Snyman, G. J. Hannon, P. Bork, and D. Arendt. 2010. 'Ancient animal microRNAs and the evolution of tissue identity', *Nature*, 463: 1084-8.

Cimmino, A., G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini, and C. M. Croce. 2005. 'miR-15 and miR-16 induce apoptosis by targeting BCL2', *Proc Natl Acad Sci U S A*, 102: 13944-9.

Collins, R. L., H. Brand, K. J. Karczewski, X. Zhao, J. Alfoldi, L. C. Francioli, A. V. Khera, C. Lowther, L. D. Gauthier, H. Wang, N. A. Watts, M. Solomonson, A. O'Donnell-Luria, A. Baumann, R. Munshi, M. Walker, C. W. Whelan, Y. Huang, T. Brookings, T. Sharpe, M. R. Stone, E. Valkanas, J. Fu, G. Tiao, K. M. Laricchia, V. Ruano-Rubio, C. Stevens, N.

Gupta, C. Cusick, L. Margolin, Team Genome Aggregation Database Production, Consortium Genome Aggregation Database, K. D. Taylor, H. J. Lin, S. S. Rich, W. S. Post, Y. I. Chen, J. I. Rotter, C. Nusbaum, A. Philippakis, E. Lander, S. Gabriel, B. M. Neale, S. Kathiresan, M. J. Daly, E. Banks, D. G. MacArthur, and M. E. Talkowski. 2020. 'A structural variation reference for medical and population genetics', *Nature*, 581: 444-51.

Conrad, D. F., T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard. 2006. 'A high-resolution survey of deletion polymorphism in the human genome', *Nat Genet*, 38: 75-81.

Consortium, Gene Ontology. 2021. 'The Gene Ontology resource: enriching a GOld mine', *Nucleic Acids Res*, 49: D325-D34.

Cortes-Ciriano, I., J. J. Lee, R. Xi, D. Jain, Y. L. Jung, L. Yang, D. Gordenin, L. J. Klimczak, C. Z. Zhang, D. S. Pellman, Pcawg Structural Variation Working Group, P. J. Park, and Pcawg Consortium. 2020. 'Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing', *Nat Genet*, 52: 331-41.

Creasy, M. R., J. A. Crolla, and E. D. Alberman. 1976. 'A cytogenetic study of human spontaneous abortions using banding techniques', *Hum Genet*, 31: 177-96.

Cretu Stancu, M., M. J. van Roosmalen, I. Renkens, M. M. Nieboer, S. Middelkamp, J. de Ligt, G. Pregno, D. Giachino, G. Mandrile, J. Espejo Valle-Inclan, J. Korzelius, E. de Bruijn, E. Cuppen, M. E. Talkowski, T. Marschall, J. de Ridder, and W. P. Kloosterman. 2017. 'Mapping and phasing of structural variation in patient genomes using nanopore sequencing', *Nat Commun*, 8: 1326.

Cullen, B. R. 2006. 'Viruses and microRNAs', *Nat Genet*, 38 Suppl: S25-30.

Darnell, R. B. 2010. 'HITS-CLIP: panoramic views of protein-RNA regulation in living cells', *Wiley Interdiscip Rev RNA*, 1: 266-86.

Davis, E., F. Caiment, X. Tordoir, J. Cavaille, A. Ferguson-Smith, N. Cockett, M. Georges, and C. Charlier. 2005. 'RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus', *Curr Biol*, 15: 743-9.

Dephoure, N., S. Hwang, C. O'Sullivan, S. E. Dodgson, S. P. Gygi, A. Amon, and E. M. Torres. 2014. 'Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast', *Elife*, 3: e03023.

Diaz-Moralli, S., M. Tarrado-Castellarnau, A. Miranda, and M. Cascante. 2013. 'Targeting cell cycle regulation in cancer therapy', *Pharmacol Ther*, 138: 255-71.

Dickerson, J. E., and D. L. Robertson. 2012. 'On the origins of Mendelian disease genes in man: the impact of gene duplication', *Mol Biol Evol*, 29: 61-9.

Ding, X. C., and H. Grosshans. 2009. 'Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins', *EMBO J*, 28: 213-22.

Diss, G., I. Gagnon-Arsenault, A. M. Dion-Cote, H. Vignaud, D. I. Ascencio, C. M. Berger, and C. R. Landry. 2017. 'Gene duplication can impart fragility, not robustness, in the yeast protein interaction network', *Science*, 355: 630-34.

Doench, J. G., C. P. Petersen, and P. A. Sharp. 2003. 'siRNAs can function as miRNAs', *Genes Dev*, 17: 438-42.

Doench, J. G., and P. A. Sharp. 2004. 'Specificity of microRNA target selection in translational repression', *Genes Dev*, 18: 504-11.

Domazet-Loso, T., and D. Tautz. 2008. 'An ancient evolutionary origin of genes associated with human genetic diseases', *Mol Biol Evol*, 25: 2699-707.

Dominguez-Sola, D., C. Y. Ying, C. Grandori, L. Ruggiero, B. Chen, M. Li, D. A. Galloway, W. Gu, J. Gautier, and R. Dalla-Favera. 2007. 'Non-transcriptional control of DNA replication by c-Myc', *Nature*, 448: 445-51.

Duan, J., J. G. Zhang, H. W. Deng, and Y. P. Wang. 2013. 'Comparative studies of copy number variation detection methods for next-generation sequencing technologies', *PLoS One*, 8: e59128.

Ebert, M. S., J. R. Neilson, and P. A. Sharp. 2007. 'MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells', *Nat Methods*, 4: 721-6.

Elkayam, E., C. D. Kuhn, A. Tocilj, A. D. Haase, E. M. Greene, G. J. Hannon, and L. Joshua-Tor. 2012. 'The structure of human argonaute-2 in complex with miR-20a', *Cell*, 150: 100-10.

Enright, A. J., B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. 2003. 'MicroRNA targets in Drosophila', *Genome Biol*, 5: R1.

Fabbri, M., A. Bottoni, M. Shimizu, R. Spizzo, M. S. Nicoloso, S. Rossi, E. Barbarotto, A. Cimmino, B. Adair, S. E. Wojcik, N. Valeri, F. Calore, D. Sampath, F. Fanini, I. Vannini, G. Musuraca, M. Dell'Aquila, H. Alder, R. V. Davuluri, L. Z. Rassenti, M. Negrini, T. Nakamura, D. Amadori, N. E. Kay, K. R. Rai, M. J. Keating, T. J. Kipps, G. A. Calin, and C. M. Croce. 2011. 'Association of a microRNA/TP53 feedback circuitry with pathogenesis and outcome of B-cell chronic lymphocytic leukemia', *JAMA*, 305: 59-67.

Farh, K. K., A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. 2005. 'The widespread impact of mammalian MicroRNAs on mRNA repression and evolution', *Science*, 310: 1817-21.

Fazi, F., A. Rosa, A. Fatica, V. Gelmetti, M. L. De Marchis, C. Nervi, and I. Bozzoni. 2005. 'A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis', *Cell*, 123: 819-31.

Felsenstein, J. 1981. 'Evolutionary trees from DNA sequences: a maximum likelihood approach', *J Mol Evol*, 17: 368-76.

Firth, H. V., S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter. 2009. 'DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources', *Am J Hum Genet*, 84: 524-33.

Fitch, W. M. 1970. 'Distinguishing homologous from analogous proteins', *Syst Zool*, 19: 99-113.

Friedman, R. C., K. K. Farh, C. B. Burge, and D. P. Bartel. 2009. 'Most mammalian mRNAs are conserved targets of microRNAs', *Genome Res*, 19: 92-105.

Furney, S. J., M. M. Alba, and N. Lopez-Bigas. 2006. 'Differences in the evolutionary history of disease genes affected by dominant or recessive mutations', *BMC Genomics*, 7: 165.

Futreal, P. A., Q. Liu, D. Shattuck-Eidens, C. Cochran, K. Harshman, S. Tavtigian, L. M. Bennett, A. Haugen-Strano, J. Swensen, Y. Miki, and et al. 1994. 'BRCA1 mutations in primary breast and ovarian carcinomas', *Science*, 266: 120-2.

Garcia-Alonso, L., J. Jimenez-Almazan, J. Carbonell-Caballero, A. Vela-Boza, J. Santoyo-Lopez, G. Antinolo, and J. Dopazo. 2014. 'The role of the interactome in the maintenance of deleterious variability in human populations', *Mol Syst Biol*, 10: 752.

Garraway, L. A., H. R. Widlund, M. A. Rubin, G. Getz, A. J. Berger, S. Ramaswamy, R. Beroukhim, D. A. Milner, S. R. Granter, J. Du, C. Lee, S. N. Wagner, C. Li, T. R. Golub, D. L. Rimm, M. L. Meyerson, D. E. Fisher, and W. R. Sellers. 2005. 'Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma', *Nature*, 436: 117-22.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder. 2012. 'Architecture of the human regulatory network derived from ENCODE data', *Nature*, 489: 91-100.

Ghandi, M., F. W. Huang, J. Jane-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, 3rd, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, and W. R. Sellers. 2019. 'Next-generation characterization of the Cancer Cell Line Encyclopedia', *Nature*, 569: 503-08.

Gibbons, D. L., W. Lin, C. J. Creighton, Z. H. Rizvi, P. A. Gregory, G. J. Goodall, N. Thilaganathan, L. Du, Y. Zhang, A. Pertsemlidis, and J. M. Kurie. 2009. 'Contextual extracellular cues promote tumor cell EMT and metastasis by regulating miR-200 family expression', *Genes Dev*, 23: 2140-51.

Gibson, G., and I. Dworkin. 2004. 'Uncovering cryptic genetic variation', *Nat Rev Genet*, 5: 681-90.

Gibson, G., and G. Wagner. 2000. 'Canalization in evolutionary genetics: a stabilizing theory?', *Bioessays*, 22: 372-80.

Gibson, T. J., and J. Spring. 1998. 'Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins', *Trends Genet*, 14: 46-9; discussion 49-50.

Godard, P., and J. van Eyll. 2015. 'Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy', *Nucleic Acids Res*, 43: 3490-7.

Gogarten, S. M., T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, D. Levine, T. Lumley, S. C. Nelson, K. Rice, J. Shen, R. Swarnkar, B. S. Weir, and C. C. Laurie. 2012. 'GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies', *Bioinformatics*, 28: 3329-31.

Goh, K. I., M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. 2007. 'The human disease network', *Proc Natl Acad Sci U S A*, 104: 8685-90.

Goodwin, S., J. D. McPherson, and W. R. McCombie. 2016. 'Coming of age: ten years of next-generation sequencing technologies', *Nat Rev Genet*, 17: 333-51.

Graves, J. A. 2006. 'Sex chromosome specialization and degeneration in mammals', *Cell*, 124: 901-14.

Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. 2007. 'Patterns of somatic mutation in human cancer genomes', *Nature*, 446: 153-8.

Gregory, R. I., T. P. Chendrimada, N. Cooch, and R. Shiekhattar. 2005. 'Human RISC couples microRNA biogenesis and posttranscriptional gene silencing', *Cell*, 123: 631-40.

Gregory, R. I., K. P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. 2004. 'The Microprocessor complex mediates the genesis of microRNAs', *Nature*, 432: 235-40.

Griffiths-Jones, S. 2004. 'The microRNA Registry', *Nucleic Acids Res*, 32: D109-11.

Grimson, A., K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. 2007. 'MicroRNA targeting specificity in mammals: determinants beyond seed pairing', *Mol Cell*, 27: 91-105.

Grimson, A., M. Srivastava, B. Fahey, B. J. Woodcroft, H. R. Chiang, N. King, B. M. Degnan, D. S. Rokhsar, and D. P. Bartel. 2008. 'Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals', *Nature*, 455: 1193-7.

Grishok, A., A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello. 2001. 'Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing', *Cell*, 106: 23-34.

Grosswendt, S., A. Filipchyk, M. Manzano, F. Klironomos, M. Schilling, M. Herzog, E. Gottwein, and N. Rajewsky. 2014. 'Unambiguous identification of miRNA:target site interactions by different types of ligation reactions', *Mol Cell*, 54: 1042-54.

Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. H. Li. 2003. 'Role of duplicate genes in genetic robustness against null mutations', *Nature*, 421: 63-6.

Guo, M., and J. A. Birchler. 1994. 'Trans-acting dosage effects on the expression of model gene systems in maize aneuploids', *Science*, 266: 1999-2002.

Haeussler, M., A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn, and W. J. Kent. 2019. 'The UCSC Genome Browser database: 2019 update', *Nucleic Acids Res*, 47: D853-D58.

Hafner, M., M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. 2010. 'PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins', *J Vis Exp*.

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: 293-6.

Hammond, S. M., S. Boettcher, A. A. Caudy, R. Kobayashi, and G. J. Hannon. 2001. 'Argonaute2, a link between genetic and biochemical analyses of RNAi', *Science*, 293: 1146-50.

Hamosh, A., A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. 2002. 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res*, 30: 52-5.

Han, J., Y. Lee, K. H. Yeom, Y. K. Kim, H. Jin, and V. N. Kim. 2004. 'The Drosha-DGCR8 complex in primary microRNA processing', *Genes Dev*, 18: 3016-27.

Han, J., Y. Lee, K. H. Yeom, J. W. Nam, I. Heo, J. K. Rhee, S. Y. Sohn, Y. Cho, B. T. Zhang, and V. N. Kim. 2006. 'Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex', *Cell*, 125: 887-901.

Hanahan, D., and R. A. Weinberg. 2000. 'The hallmarks of cancer', *Cell*, 100: 57-70.

———. 2011. 'Hallmarks of cancer: the next generation', *Cell*, 144: 646-74.

Heimberg, A. M., L. F. Sempere, V. N. Moy, P. C. Donoghue, and K. J. Peterson. 2008. 'MicroRNAs and the advent of vertebrate morphological complexity', *Proc Natl Acad Sci U S A*, 105: 2946-50.

Helwak, A., G. Kudla, T. Dudnakova, and D. Tollervey. 2013. 'Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding', *Cell*, 153: 654-65.

Henrichsen, C. N., E. Chaignat, and A. Reymond. 2009. 'Copy number variants, diseases and gene expression', *Hum Mol Genet*, 18: R1-8.

Herskowitz, I. 1987. 'Functional inactivation of genes by dominant negative mutations', *Nature*, 329: 219-22.

Hoffman, B., and D. A. Liebermann. 2008. 'Apoptotic signaling by c-MYC', *Oncogene*, 27: 6462-72.

Hornstein, E., and N. Shomron. 2006. 'Canalization of development by microRNAs', *Nat Genet*, 38 Suppl: S20-4.

Hsiao, T. L., and D. Vitkup. 2008. 'Role of duplicate genes in robustness against deleterious human mutations', *PLoS Genet*, 4: e1000014.

Hu, H. Y., Z. Yan, Y. Xu, H. Hu, C. Menzel, Y. H. Zhou, W. Chen, and P. Khaitovich. 2009. 'Sequence features associated with microRNA strand selection in humans and flies', *BMC Genomics*, 10: 413.

Huang, N., I. Lee, E. M. Marcotte, and M. E. Hurles. 2010. 'Characterising and predicting haploinsufficiency in the human genome', *PLoS Genet*, 6: e1001154.

Huminiecki, L., and C. H. Heldin. 2010. '2R and remodeling of vertebrate signal transduction engine', *BMC Biol*, 8: 146.

Hutvagner, G., J. McLachlan, A. E. Pasquinelli, E. Balint, T. Tuschl, and P. D. Zamore. 2001. 'A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA', *Science*, 293: 834-8.

Hutvagner, G., and P. D. Zamore. 2002. 'A microRNA in a multiple-turnover RNAi enzyme complex', *Science*, 297: 2056-60.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. 2004. 'Detection of large-scale variation in the human genome', *Nat Genet*, 36: 949-51.

International HapMap, Consortium. 2003. 'The International HapMap Project', *Nature*, 426: 789-96.

International HapMap, Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan,

A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. 2007. 'A second generation human haplotype map of over 3.1 million SNPs', *Nature*, 449: 851-61.

Jassal, B., L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. 2020. 'The reactome pathway knowledgebase', *Nucleic Acids Res*, 48: D498-D503.

Ji, Z., J. Y. Lee, Z. Pan, B. Jiang, and B. Tian. 2009. 'Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development', *Proc Natl Acad Sci U S A*, 106: 7028-33.

Jimenez-Sanchez, G., B. Childs, and D. Valle. 2001. 'Human disease genes', *Nature*, 409: 853-5.

John, B., A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. 2004. 'Human MicroRNA targets', *PLoS Biol*, 2: e363.

Kalluri, R., and R. A. Weinberg. 2009. 'The basics of epithelial-mesenchymal transition', *J Clin Invest*, 119: 1420-8.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Consortium Genome Aggregation Database, B. M. Neale, M. J. Daly, and D. G. MacArthur. 2020. 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581: 434-43.

Kawamata, T., H. Seitz, and Y. Tomari. 2009. 'Structural determinants of miRNAs for RISC loading and slicer-independent unwinding', *Nat Struct Mol Biol*, 16: 953-60.

Khvorova, A., A. Reynolds, and S. D. Jayasena. 2003. 'Functional siRNAs and miRNAs exhibit strand bias', *Cell*, 115: 209-16.

Kim, V. N., J. Han, and M. C. Siomi. 2009. 'Biogenesis of small RNAs in animals', *Nat Rev Mol Cell Biol*, 10: 126-39.

Kiriakidou, M., P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. 2004. 'A combined computational-experimental approach predicts human microRNA targets', *Genes Dev*, 18: 1165-78.

Kleensang, A., M. M. Vantangoli, S. Odwin-DaCosta, M. E. Andersen, K. Boekelheide, M. Bouhifd, A. J. Fornace, Jr., H. H. Li, C. B. Livi, S. Madnick, A. Maertens, M. Rosenberg, J. D. Yager, L. Zhao, and T. Hartung. 2016. 'Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function', *Sci Rep*, 6: 28994.

Klein, M. E., D. T. Lioy, L. Ma, S. Impey, G. Mandel, and R. H. Goodman. 2007. 'Homeostatic regulation of MeCP2 expression by a CREB-induced microRNA', *Nat Neurosci*, 10: 1513-4.

Kloosterman, W. P., E. Wienholds, R. F. Ketting, and R. H. Plasterk. 2004. 'Substrate requirements for let-7 function in the developing zebrafish embryo', *Nucleic Acids Res*, 32: 6284-91.

Kodama, Y., M. Shumway, R. Leinonen, and Collaboration International Nucleotide Sequence Database. 2012. 'The Sequence Read Archive: explosive growth of sequencing data', *Nucleic Acids Res*, 40: D54-6.

Kondrashov, F. A., and E. V. Koonin. 2004. 'A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications', *Trends Genet*, 20: 287-90.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. 'Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing', *Genome Biol*, 20: 117.

Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones. 2019. 'miRBase: from microRNA sequences to function', *Nucleic Acids Res*, 47: D155-D62.

Krichevsky, A. M., K. S. King, C. P. Donahue, K. Khrapko, and K. S. Kosik. 2003. 'A microRNA array reveals extensive regulation of microRNAs during brain development', *RNA*, 9: 1274-81.

Kruglyak, L., and D. A. Nickerson. 2001. 'Variation is the spice of life', *Nat Genet*, 27: 234-6.

Krutzfeldt, J., N. Rajewsky, R. Braich, K. G. Rajeev, T. Tuschl, M. Manoharan, and M. Stoffel. 2005. 'Silencing of microRNAs in vivo with 'antagomirs'', *Nature*, 438: 685-9.

Kudla, G., S. Granneman, D. Hahn, J. D. Beggs, and D. Tollervey. 2011. 'Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast', *Proc Natl Acad Sci U S A*, 108: 10010-5.

Kumar, M. S., J. Lu, K. L. Mercer, T. R. Golub, and T. Jacks. 2007. 'Impaired microRNA processing enhances cellular transformation and tumorigenesis', *Nat Genet*, 39: 673-7.

Kumar, R., G. Nagpal, V. Kumar, S. S. Usmani, P. Agrawal, and G. P. S. Raghava. 2019. 'HumCFS: a database of fragile sites in human chromosomes', *BMC Genomics*, 19: 985.

Kuroda, K., T. Fukuda, M. Krstic-Demonacos, C. Demonacos, K. Okumura, H. Isogai, M. Hayashi, K. Saito, and E. Isogai. 2017. 'miR-663a regulates growth of colon cancer cells, after administration of antimicrobial peptides, by targeting CXCR4-p21 pathway', *BMC Cancer*, 17: 33.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. 'Identification of novel genes coding for small expressed RNAs', *Science*, 294: 853-8.

Lagos-Quintana, M., R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl. 2003. 'New microRNAs from mouse and human', *RNA*, 9: 175-9.

Lagos-Quintana, M., R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. 2002. 'Identification of tissue-specific microRNAs from mouse', *Curr Biol*, 12: 735-9.

Lai, E. C. 2002. 'Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation', *Nat Genet*, 30: 363-4.

Lai, E. C., B. Tam, and G. M. Rubin. 2005a. 'Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs', *Genes Dev*, 19: 1067-80.

Lai, E. C., P. Tomancak, R. W. Williams, and G. M. Rubin. 2003. 'Computational identification of Drosophila microRNA genes', *Genome Biol*, 4: R42.

Lai, W. R., M. D. Johnson, R. Kucherlapati, and P. J. Park. 2005b. 'Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data', *Bioinformatics*, 21: 3763-70.

Lall, S., D. Grun, A. Krek, K. Chen, Y. L. Wang, C. N. Dewey, P. Sood, T. Colombo, N. Bray, P. Macmenamin, H. L. Kao, K. C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky. 2006. 'A genome-wide map of conserved microRNA targets in C. elegans', *Curr Biol*, 16: 460-71.

Lam, E. T., A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P. Y. Kwok. 2012. 'Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly', *Nat Biotechnol*, 30: 771-6.

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, J. Szustakowki, and Consortium International Human Genome Sequencing. 2001. 'Initial sequencing and analysis of the human genome', *Nature*, 409: 860-921.

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: 858-62.

Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. 2013. 'Software for computing and annotating genomic ranges', *PLoS Comput Biol*, 9: e1003118.

Lee, C. T., T. Risom, and W. M. Strauss. 2007. 'Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny', *DNA Cell Biol*, 26: 209-18.

Lee, R. C., and V. Ambros. 2001. 'An extensive class of small RNAs in Caenorhabditis elegans', *Science*, 294: 862-4.

Lee, R. C., R. L. Feinbaum, and V. Ambros. 1993. 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14', *Cell*, 75: 843-54.

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: 415-9.

Lee, Y., K. Jeon, J. T. Lee, S. Kim, and V. N. Kim. 2002. 'MicroRNA maturation: stepwise processing and subcellular localization', *EMBO J*, 21: 4663-70.

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: 4051-60.

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: 15-20.

Lewis, B. P., I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. 2003. 'Prediction of mammalian microRNA targets', *Cell*, 115: 787-98.

Lewis, W. H. 1979. 'Polyploidy: biological relevance', *Basic Life Sci*, 13: 1-544.

Li, H., and R. Durbin. 2010. 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics*, 26: 589-95.

Li, S. D., T. Tagami, Y. F. Ho, and C. H. Yeang. 2011. 'Deciphering causal and statistical relations of molecular aberrations and gene expressions in NCI-60 cell lines', *BMC Syst Biol*, 5: 186.

Li, X., and R. W. Carthew. 2005. 'A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye', *Cell*, 123: 1267-77.

Li, X., J. J. Cassidy, C. A. Reinke, S. Fischboeck, and R. W. Carthew. 2009. 'A microRNA imparts robustness against environmental fluctuation during development', *Cell*, 137: 273-82.

Lim, L. P., M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel. 2003a. 'Vertebrate microRNA genes', *Science*, 299: 1540.

Lim, L. P., N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. 2005. 'Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs', *Nature*, 433: 769-73.

Lim, L. P., N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. 2003b. 'The microRNAs of Caenorhabditis elegans', *Genes Dev*, 17: 991-1008.

Lin, S., and R. I. Gregory. 2015. 'MicroRNA biogenesis pathways in cancer', *Nat Rev Cancer*, 15: 321-33.

Linsley, P. S., J. Schelter, J. Burchard, M. Kibukawa, M. M. Martin, S. R. Bartz, J. M. Johnson, J. M. Cummins, C. K. Raymond, H. Dai, N. Chau, M. Cleary, A. L. Jackson, M. Carleton, and L. Lim. 2007. 'Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression', *Mol Cell Biol*, 27: 2240-52.

Liu, P., C. M. Carvalho, P. J. Hastings, and J. R. Lupski. 2012. 'Mechanisms for recurrent and complex human genomic rearrangements', *Curr Opin Genet Dev*, 22: 211-20.

Liu, Y., Y. Mi, T. Mueller, S. Kreibich, E. G. Williams, A. Van Drogen, C. Borel, M. Frank, P. L. Germain, I. Bludau, M. Mehnert, M. Seifert, M. Emmenlauer, I. Sorg, F. Bezrukov, F. S. Bena, H. Zhou, C. Dehio, G. Testa, J. Saez-Rodriguez, S. E. Antonarakis, W. D. Hardt, and R. Aebersold. 2019. 'Multi-omic measurements of heterogeneity in HeLa cells across laboratories', *Nat Biotechnol*, 37: 314-22.

Llave, C., Z. Xie, K. D. Kasschau, and J. C. Carrington. 2002. 'Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA', *Science*, 297: 2053-6.

Lorenzi, P. L., W. C. Reinhold, S. Varma, A. A. Hutchinson, Y. Pommier, S. J. Chanock, and J. N. Weinstein. 2009. 'DNA fingerprinting of the NCI-60 cell line panel', *Mol Cancer Ther*, 8: 713-24.

Lynch, M. 2007. 'The frailty of adaptive hypotheses for the origins of organismal complexity', *Proc Natl Acad Sci U S A*, 104 Suppl 1: 8597-604.

Lynch, M., and J. S. Conery. 2000. 'The evolutionary fate and consequences of duplicate genes', *Science*, 290: 1151-5.

Lynch, M., and K. Hagner. 2015. 'Evolutionary meandering of intermolecular interactions along the drift barrier', *Proc Natl Acad Sci U S A*, 112: E30-8.

Lytle, J. R., T. A. Yario, and J. A. Steitz. 2007. 'Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR', *Proc Natl Acad Sci U S A*, 104: 9667-72.

Lyu, Y., Y. Shen, H. Li, Y. Chen, L. Guo, Y. Zhao, E. Hungate, S. Shi, C. I. Wu, and T. Tang. 2014. 'New microRNAs in Drosophila--birth, death and cycles of adaptive evolution', *PLoS Genet*, 10: e1004096.

MacDonald, J. R., R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer. 2014. 'The Database of Genomic Variants: a curated collection of structural variation in the human genome', *Nucleic Acids Res*, 42: D986-92.

Macrae, I. J., K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna. 2006. 'Structural basis for double-stranded RNA processing by Dicer', *Science*, 311: 195-8.

Makino, T., K. Hokamp, and A. McLysaght. 2009. 'The complex relationship of gene duplication and essentiality', *Trends Genet*, 25: 152-5.

Makino, T., and A. McLysaght. 2010. 'Ohnologs in the human genome are dosage balanced and frequently associated with disease', *Proc Natl Acad Sci U S A*, 107: 9270-4.

Mallory, A. C., B. J. Reinhart, M. W. Jones-Rhoades, G. Tang, P. D. Zamore, M. K. Barton, and D. P. Bartel. 2004. 'MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region', *EMBO J*, 23: 3356-64.

Mangan, S., and U. Alon. 2003. 'Structure and function of the feed-forward loop network motif', *Proc Natl Acad Sci U S A*, 100: 11980-5.

Mangan, S., A. Zaslaver, and U. Alon. 2003. 'The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks', *J Mol Biol*, 334: 197-204.

Mannava, S., V. Grachtchouk, L. J. Wheeler, M. Im, D. Zhuang, E. G. Slavina, C. K. Mathews, D. S. Shewach, and M. A. Nikiforov. 2008. 'Direct role of nucleotide metabolism in C-MYC-dependent proliferation of melanoma cells', *Cell Cycle*, 7: 2392-400.

Martincorena, I., K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. 2017. 'Universal Patterns of Selection in Cancer and Somatic Tissues', *Cell*, 171: 1029-41 e21.

Masterson, J. 1994. 'Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms', *Science*, 264: 421-4.

Mayr, C., M. T. Hemann, and D. P. Bartel. 2007. 'Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation', *Science*, 315: 1576-9.

McKusick-Nathans Institute of Genetic Medicine. 2021. "Online Mendelian Inheritance in Man (OMIM)." In.: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

McLysaght, A., T. Makino, H. M. Grayton, M. Tropeano, K. J. Mitchell, E. Vassos, and D. A. Collier. 2014. 'Ohnologs are overrepresented in pathogenic copy number mutations', *Proc Natl Acad Sci U S A*, 111: 361-6.

Melo, S. A., C. Moutinho, S. Ropero, G. A. Calin, S. Rossi, R. Spizzo, A. F. Fernandez, V. Davalos, A. Villanueva, G. Montoya, H. Yamamoto, S. Schwartz, Jr., and M. Esteller. 2010. 'A genetic defect in exportin-5 traps precursor microRNAs in the nucleus of cancer cells', *Cancer Cell*, 18: 303-15.

Mendel, G. 1866. 'Versuche über Pflanzenhybriden', *Verhandlungen des naturforschenden Vereines in Brünn*.

Meunier, J., F. Lemoine, M. Soumillon, A. Liechti, M. Weier, K. Guschanski, H. Hu, P. Khaitovich, and H. Kaessmann. 2013. 'Birth and expression evolution of mammalian microRNA genes', *Genome Res*, 23: 34-45.

Mi, H., D. Ebert, A. Muruganujan, C. Mills, L. P. Albou, T. Mushayamaha, and P. D. Thomas. 2020. 'PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API', *Nucleic Acids Res*.

Miller, C. A., O. Hampton, C. Coarfa, and A. Milosavljevic. 2011. 'ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads', *PLoS One*, 6: e16327.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and Project Genomes. 2011. 'Mapping copy number variation by population-scale genome sequencing', *Nature*, 470: 59-65.

Miska, E. A., E. Alvarez-Saavedra, A. L. Abbott, N. C. Lau, A. B. Hellman, S. M. McGonagle, D. P. Bartel, V. R. Ambros, and H. R. Horvitz. 2007. 'Most Caenorhabditis elegans microRNAs are individually not essential for development or viability', *PLoS Genet*, 3: e215.

Mogilyansky, E., and I. Rigoutsos. 2013. 'The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease', *Cell Death Differ*, 20: 1603-14.

Molnar, A., F. Schwach, D. J. Studholme, E. C. Thuenemann, and D. C. Baulcombe. 2007. 'miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii', *Nature*, 447: 1126-9.

Moss, E. G., R. C. Lee, and V. Ambros. 1997. 'The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA', *Cell*, 88: 637-46.

Motofeanu, Corina 2019. "Oncomirs and tumour suppressors from literature review." In.

Mu, P., Y. C. Han, D. Betel, E. Yao, M. Squatrito, P. Ogrodowski, E. de Stanchina, A. D'Andrea, C. Sander, and A. Ventura. 2009. 'Genetic dissection of the miR-17~92 cluster of microRNAs in Myc-induced B-cell lymphomas', *Genes Dev*, 23: 2806-11.

Mukherji, S., M. S. Ebert, G. X. Zheng, J. S. Tsang, P. A. Sharp, and A. van Oudenaarden. 2011. 'MicroRNAs can generate thresholds in target gene expression', *Nat Genet*, 43: 854-9.

Muralidhar, B., D. Winder, M. Murray, R. Palmer, N. Barbosa-Morais, H. Saini, I. Roberts, M. Pett, and N. Coleman. 2011. 'Functional evidence that Drosha overexpression in cervical squamous cell carcinoma affects cell phenotype and microRNA profiles', *J Pathol*, 224: 496-507.

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. 'A common sequence motif associated with recombination hot spots and genome instability in humans', *Nat Genet*, 40: 1124-9.

Ninova, M., M. Ronshaugen, and S. Griffiths-Jones. 2014. 'Fast-evolving microRNAs are highly expressed in the early embryo of Drosophila virilis', *RNA*, 20: 360-72.

Nozawa, M., S. Miura, and M. Nei. 2010. 'Origins and evolution of microRNA genes in Drosophila species', *Genome Biol Evol*, 2: 180-9.

Nutt, S. L., and M. Busslinger. 1999. 'Monoallelic expression of Pax5: a paradigm for the haploinsufficiency of mammalian Pax genes?', *Biol Chem*, 380: 601-11.

O'Donnell, K. A., E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell. 2005. 'c-Myc-regulated microRNAs modulate E2F1 expression', *Nature*, 435: 839-43.

Ogawa, Y., B. K. Sun, and J. T. Lee. 2008. 'Intersection of the RNA interference and X-inactivation pathways', *Science*, 320: 1336-41.

Ohler, U., S. Yekta, L. P. Lim, D. P. Bartel, and C. B. Burge. 2004. 'Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification', *RNA*, 10: 1309-22.

Ohno, Susumu. 1970. *Evolution by Gene Duplication* (Springer-Verlag Berlin Heidelberg).

Okamura, K., J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai. 2007. 'The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila', *Cell*, 130: 89-100.

Olive, V., M. J. Bennett, J. C. Walker, C. Ma, I. Jiang, C. Cordon-Cardo, Q. J. Li, S. W. Lowe, G. J. Hannon, and L. He. 2009. 'miR-19 is a key oncogenic component of mir-17-92', *Genes Dev*, 23: 2839-49.

Olsen, P. H., and V. Ambros. 1999. 'The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation', *Dev Biol*, 216: 671-80.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. 'Circular binary segmentation for the analysis of array-based DNA copy number data', *Biostatistics*, 5: 557-72.

Oosawa, F., and M. Kasai. 1962. 'A theory of linear and helical aggregations of macromolecules', *J Mol Biol*, 4: 10-21.

Otto, S. P., and J. Whitton. 2000. 'Polyploid incidence and evolution', *Annu Rev Genet*, 34: 401-37.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. 2002. 'Regulation of noise in the expression of a single gene', *Nat Genet*, 31: 69-73.

Ozsolak, F., L. L. Poling, Z. Wang, H. Liu, X. S. Liu, R. G. Roeder, X. Zhang, J. S. Song, and D. E. Fisher. 2008. 'Chromatin structure analyses identify miRNA promoters', *Genes Dev*, 22: 3172-83.

Pabinger, S., A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski. 2014. 'A survey of tools for variant analysis of next-generation genome sequencing data', *Brief Bioinform*, 15: 256-78.

Papp, B., C. Pal, and L. D. Hurst. 2003. 'Dosage sensitivity and the evolution of gene families in yeast', *Nature*, 424: 194-7.

Pasquinelli, A. E., B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Muller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. 2000. 'Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA', *Nature*, 408: 86-9.

Paterson, E. L., N. Kolesnikoff, P. A. Gregory, A. G. Bert, Y. Khew-Goodall, and G. J. Goodall. 2008. 'The microRNA-200 family regulates epithelial to mesenchymal transition', *ScientificWorldJournal*, 8: 901-4.

Paulsson, J. 2004. 'Summing up the noise in gene networks', *Nature*, 427: 415-8.

Perez-Perez, J. M., H. Candela, and J. L. Micol. 2009. 'Understanding synergy in genetic interactions', *Trends Genet*, 25: 368-76.

Peterson, K. J., M. R. Dietrich, and M. A. McPeek. 2009. 'MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion', *Bioessays*, 31: 736-47.

Pires, J. C., and G. C. Conant. 2016. 'Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage in the Evolution of Genomes', *Annu Rev Genet*, 50: 113-31.

Plata, G., and D. Vitkup. 2014. 'Genetic robustness and functional evolution of gene duplicates', *Nucleic Acids Res*, 42: 2405-14.

Pleasance, E. D., R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. Ordonez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-

Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. 2010. 'A comprehensive catalogue of somatic mutations from a human cancer genome', *Nature*, 463: 191-6.

Pulikkan, J. A., V. Dengler, P. S. Peramangalam, A. A. Peer Zada, C. Muller-Tidow, S. K. Bohlander, D. G. Tenen, and G. Behre. 2010. 'Cell-cycle regulator E2F1 and microRNA-223 comprise an autoregulatory negative feedback loop in acute myeloid leukemia', *Blood*, 115: 1768-78.

Queitsch, C., T. A. Sangster, and S. Lindquist. 2002. 'Hsp90 as a capacitor of phenotypic variation', *Nature*, 417: 618-24.

Raser, J. M., and E. K. O'Shea. 2005. 'Noise in gene expression: origins, consequences, and control', *Science*, 309: 2010-3.

Rastogi, S., and D. A. Liberles. 2005. 'Subfunctionalization of duplicated genes as a transition state to neofunctionalization', *BMC Evol Biol*, 5: 28.

Reardon, Mark. 2016. 'Visualisation of the Evolution of Molecular Interactions in Time', University of Manchester.

Rehwinkel, J., I. Behm-Ansmant, D. Gatfield, and E. Izaurralde. 2005. 'A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing', *RNA*, 11: 1640-7.

Reinhart, B. J., and D. P. Bartel. 2002. 'Small RNAs correspond to centromere heterochromatic repeats', *Science*, 297: 1831.

Reinhart, B. J., F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. 2000. 'The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans', *Nature*, 403: 901-6.

Reinhart, B. J., E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel. 2002. 'MicroRNAs in plants', *Genes Dev*, 16: 1616-26.

Reinhold, W. C., S. Varma, F. Sousa, M. Sunshine, O. D. Abaan, S. R. Davis, S. W. Reinhold, K. W. Kohn, J. Morris, P. S. Meltzer, J. H. Doroshow, and Y. Pommier. 2014. 'NCI-60 whole exome sequencing and pharmacological CellMiner analyses', *PLoS One*, 9: e101670.

Reinhold, W. C., S. Varma, M. Sunshine, F. Elloumi, K. Ofori-Atta, S. Lee, J. B. Trepel, P. S. Meltzer, J. H. Doroshow, and Y. Pommier. 2019. 'RNA Sequencing of the NCI-60: Integration into CellMiner and CellMiner CDB', *Cancer Res*, 79: 3514-24.

Rheinbay, E., M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshoj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzos, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar,

D. Lee, K. V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uuskula-Reimand, C. Wadelius, L. Wadi, X. Yao, C. Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael, M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Campbell, N. Lopez-Bigas, Pcawg Drivers, Group Functional Interpretation Working, Pcawg Structural Variation Working Group, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, G. Getz, and Pcawg Consortium. 2020. 'Analyses of non-coding somatic drivers in 2,658 cancer whole genomes', *Nature*, 578: 102-11.

Rhoades, M. W., B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel. 2002. 'Prediction of plant microRNA targets', *Cell*, 110: 513-20.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, and S. N. P. Map Working Group International. 2001. 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature*, 409: 928-33.

Saitou, N., and M. Nei. 1987. 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Mol Biol Evol*, 4: 406-25.

Santarius, T., J. Shipley, D. Brewer, M. R. Stratton, and C. S. Cooper. 2010. 'A census of amplified and overexpressed human cancer genes', *Nat Rev Cancer*, 10: 59-64.

Sayers, E. W., J. Beck, J. R. Brister, E. E. Bolton, K. Canese, D. C. Comeau, K. Funk, A. Ketter, S. Kim, A. Kimchi, P. A. Kitts, A. Kuznetsov, S. Lathrop, Z. Lu, K. McGarvey, T. L. Madden, T. D. Murphy, N. O'Leary, L. Phan, V. A. Schneider, F. Thibaud-Nissen, B. W. Trawick, K. D. Pruitt, and J. Ostell. 2020. 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res*, 48: D9-D16.

Schirle, N. T., and I. J. MacRae. 2012. 'The crystal structure of human Argonaute2', *Science*, 336: 1037-40.

Schuster-Bockler, B., D. Conrad, and A. Bateman. 2010. 'Dosage sensitivity shapes the evolution of copy-number varied regions', *PLoS One*, 5: e9474.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. 2004. 'Large-scale copy number polymorphism in the human genome', *Science*, 305: 525-8.

Seidman, J. G., and C. Seidman. 2002. 'Transcription factor haploinsufficiency: when half a loaf is not enough', *J Clin Invest*, 109: 451-5.

Selbach, M., B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. 2008. 'Widespread changes in protein synthesis induced by microRNAs', *Nature*, 455: 58-63.

Sempere, L. F., C. N. Cole, M. A. McPeek, and K. J. Peterson. 2006. 'The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint', *J Exp Zool B Mol Dev Evol*, 306: 575-88.

Shabalina, S. A., and E. V. Koonin. 2008. 'Origins and evolution of eukaryotic RNA interference', *Trends Ecol Evol*, 23: 578-87.

Shao, Xin, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. 2019. 'Copy number variation is highly correlated with differential gene expression: a pan-cancer study', *BMC Medical Genetics*, 20: 175.

Sheltzer, J. M., and A. Amon. 2011. 'The aneuploidy paradox: costs and benefits of an incorrect karyotype', *Trends Genet*, 27: 446-53.

Shoemaker, R. H. 2006. 'The NCI60 human tumour cell line anticancer drug screen', *Nat Rev Cancer*, 6: 813-23.

Silber, J., R. Hashizume, T. Felix, S. Hariono, M. Yu, M. S. Berger, J. T. Huse, S. R. VandenBerg, C. D. James, J. G. Hodgson, and N. Gupta. 2013. 'Expression of miR-124 inhibits growth of medulloblastoma cells', *Neuro Oncol*, 15: 83-90.

Singh, P. P., S. Affeldt, I. Cascone, R. Selimoglu, J. Camonis, and H. Isambert. 2012. 'On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates', *Cell Rep*, 2: 1387-98.

Singh, P. P., and H. Isambert. 2020. 'OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates', *Nucleic Acids Res*, 48: D724-D30.

Slack, A., P. C. Thornton, D. B. Magner, S. M. Rosenberg, and P. J. Hastings. 2006. 'On the mechanism of gene amplification induced under stress in Escherichia coli', *PLoS Genet*, 2: e48.

Slack, F. J., M. Basson, Z. Liu, V. Ambros, H. R. Horvitz, and G. Ruvkun. 2000. 'The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor', *Mol Cell*, 5: 659-69.

Slatkin, M. 2008. 'Linkage disequilibrium--understanding the evolutionary past and mapping the medical future', *Nat Rev Genet*, 9: 477-85.

Sondka, Z., S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes. 2018. 'The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers', *Nat Rev Cancer*, 18: 696-705.

Stallings-Mann, M. L., J. Waldmann, Y. Zhang, E. Miller, M. L. Gauthier, D. W. Visscher, G. P. Downey, E. S. Radisky, A. P. Fields, and D. C. Radisky. 2012. 'Matrix metalloproteinase induction of Rac1b, a key effector of lung cancer progression', *Sci Transl Med*, 4: 142ra95.

Stankiewicz, P., and J. R. Lupski. 2002. 'Genome architecture, rearrangements and genomic disorders', *Trends Genet*, 18: 74-82.

Stark, A., J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. 2005. 'Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution', *Cell*, 123: 1133-46.

Stark, A., J. Brennecke, R. B. Russell, and S. M. Cohen. 2003. 'Identification of Drosophila MicroRNA targets', *PLoS Biol*, 1: E60.

Stearns, T., and D. Botstein. 1988. 'Unlinked noncomplementation: isolation of new conditional-lethal mutations in each of the tubulin genes of Saccharomyces cerevisiae', *Genetics*, 119: 249-60.

Stingele, S., G. Stoehr, K. Peplowska, J. Cox, M. Mann, and Z. Storchova. 2012. 'Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells', *Mol Syst Biol*, 8: 608.

Su, Z., J. Wang, and X. Gu. 2014. 'Effect of duplicate genes on mouse genetic robustness: an update', *Biomed Res Int*, 2014: 758672.

Sugito, N., H. Ishiguro, Y. Kuwabara, M. Kimura, A. Mitsui, H. Kurehara, T. Ando, R. Mori, N. Takashima, R. Ogawa, and Y. Fujii. 2006. 'RNASEN regulates cell proliferation and affects survival in esophageal cancer patients', *Clin Cancer Res*, 12: 7322-8.

Tanzer, A., and P. F. Stadler. 2004. 'Molecular evolution of a microRNA cluster', *J Mol Biol*, 339: 327-35.

Thomson, J. M., M. Newman, J. S. Parker, E. M. Morin-Kensicki, T. Wright, and S. M. Hammond. 2006. 'Extensive post-transcriptional regulation of microRNAs and its implications for cancer', *Genes Dev*, 20: 2202-7.

Torres, E. M., B. R. Williams, and A. Amon. 2008. 'Aneuploidy: cells losing their balance', *Genetics*, 179: 737-46.

Turner, K. M., V. Deshpande, D. Beyter, T. Koga, J. Rusert, C. Lee, B. Li, K. Arden, B. Ren, D. A. Nathanson, H. I. Kornblum, M. D. Taylor, S. Kaushal, W. K. Cavenee, R. Wechsler-Reya, F. B. Furnari, S. R. Vandenberg, P. N. Rao, G. M. Wahl, V. Bafna, and P. S. Mischel. 2017. 'Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity', *Nature*, 543: 122-25.

Varma, S., Y. Pommier, M. Sunshine, J. N. Weinstein, and W. C. Reinhold. 2014. 'High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner', *PLoS One*, 9: e92047.

Veitia, R. A. 2002. 'Exploring the etiology of haploinsufficiency', *Bioessays*, 24: 175-84.

———. 2003a. 'Nonlinear effects in macromolecular assembly and dosage sensitivity', *J Theor Biol*, 220: 19-25.

———. 2003b. 'A sigmoidal transcriptional response: cooperativity, synergy and dosage effects', *Biol Rev Camb Philos Soc*, 78: 149-70.

Veitia, R. A., S. Bottani, and J. A. Birchler. 2013. 'Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation', *Trends Genet*, 29: 385-93.

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. Di 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. Wang, A. Wang, X. Wang, J. Wang, M. 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. 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. Guigo, 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. 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: 1304-51.

Veyrunes, F., P. D. Waters, P. Miethke, W. Rens, D. McMillan, A. E. Alsop, F. Grutzner, J. E. Deakin, C. M. Whittington, K. Schatzkamer, C. L. Kremitzki, T. Graves, M. A. Ferguson-Smith, W. Warren, and J. A. Marshall Graves. 2008. 'Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes', *Genome Res*, 18: 965-73.

Vlachos, I. S., K. Zagganas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, and A. G. Hatzigeorgiou. 2015. 'DIANA-miRPath v3.0: deciphering microRNA function with experimental support', *Nucleic Acids Res*, 43: W460-6.

Waddington, C. H. 1959. 'Canalization of development and genetic assimilation of acquired characters', *Nature*, 183: 1654-5.

Wang, Y., J. Luo, H. Zhang, and J. Lu. 2016. 'microRNAs in the Same Clusters Evolve to Coordinately Regulate Functionally Related Genes', *Mol Biol Evol*, 33: 2232-47.

Warnes, Gregory R., Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. 2020. "gplots: Various R Programming Tools for Plotting Data." In.

Webster, M. 2007. 'A Cambrian peak in morphological variation within trilobite species', *Science*, 317: 499-502.

Wightman, B., T. R. Burglin, J. Gatto, P. Arasu, and G. Ruvkun. 1991. 'Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development', *Genes Dev*, 5: 1813-24.

Wightman, B., I. Ha, and G. Ruvkun. 1993. 'Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans', *Cell*, 75: 855-62.

Wolfe, K. 2000. 'Robustness--it's not where you think it is', *Nat Genet*, 25: 3-4.

Xu, M. M., G. X. Mao, J. Liu, J. C. Li, H. Huang, Y. F. Liu, and J. H. Liu. 2014. 'Low expression of the FoxO4 gene may contribute to the phenomenon of EMT in non-small cell lung cancer', *Asian Pac J Cancer Prev*, 15: 4013-8.

Yang, L., M. S. Lee, H. Lu, D. Y. Oh, Y. J. Kim, D. Park, G. Park, X. Ren, C. A. Bristow, P. S. Haseley, S. Lee, A. Pantazi, R. Kucherlapati, W. Y. Park, K. L. Scott, Y. L. Choi, and P. J. Park. 2016. 'Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing', *Am J Hum Genet*, 98: 843-56.

Yates, A. D., P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, J. C. Marugan, C. Cummins, C. Davidson, K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore, M. Nuhn, D. N. Oheh, A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, I. Isley G, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion, F. Cunningham, K. L. Howe, D. R. Zerbino, and P. Flicek. 2020. 'Ensembl 2020', *Nucleic Acids Res*, 48: D682-D88.

Yekta, S., I. H. Shih, and D. P. Bartel. 2004. 'MicroRNA-directed cleavage of HOXB8 mRNA', *Science*, 304: 594-6.

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: 3011-6.

Yoo, A. S., and I. Greenwald. 2005. 'LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans', *Science*, 310: 1330-3.

Yoon, S., Z. Xuan, V. Makarov, K. Ye, and J. Sebat. 2009. 'Sensitive and accurate detection of copy number variants using read depth of coverage', *Genome Res*, 19: 1586-92.

Zdanowicz, A., R. Thermann, J. Kowalska, J. Jemielity, K. Duncan, T. Preiss, E. Darzynkiewicz, and M. W. Hentze. 2009. 'Drosophila miR2 primarily targets the m7GpppN cap structure for translational repression', *Mol Cell*, 35: 881-8.

Zeng, Y., E. J. Wagner, and B. R. Cullen. 2002. 'Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells', *Mol Cell*, 9: 1327-33.

Zeng, Y., R. Yi, and B. R. Cullen. 2003. 'MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms', *Proc Natl Acad Sci U S A*, 100: 9779-84.

Zhang, C., B. Chen, A. Jiao, F. Li, N. Sun, G. Zhang, and J. Zhang. 2018. 'miR-663a inhibits tumor growth and invasion by regulating TGF-beta1 in hepatocellular carcinoma', *BMC Cancer*, 18: 1179.

Zhang, W., P. Landback, A. R. Gschwend, B. Shen, and M. Long. 2015. 'New genes drive the evolution of gene interaction networks in the human and mouse genomes', *Genome Biol*, 16: 202.

Zhao, M., and Z. Zhao. 2016. 'Concordance of copy number loss and down-regulation of tumor suppressor genes: a pan-cancer study', *BMC Genomics*, 17 Suppl 7: 532.

Zhao, T., G. Li, S. Mi, S. Li, G. J. Hannon, X. J. Wang, and Y. Qi. 2007. 'A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii', *Genes Dev*, 21: 1190-203.