



# THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

**Network-based visualisation and  
analysis of next-generation  
sequencing (NGS) data**



**Wan Fahmi Bin Wan Mohamad Nazarie**

**Thesis submitted for the degree of  
Doctor of Philosophy**

**2016**

## **Declaration of Originality**

I hereby declare that this thesis and the research work reported herein was composed and originated entirely by myself except where specifically indicated.

Wan Fahmi bin Wan Mohamad Nazarie

October 2016

## Abstract

Next-generation sequencing (NGS) technologies have revolutionised research into nature and diversity of genomes and transcriptomes. Since the initial description of these technology platforms over a decade ago, massively parallel RNA sequencing (RNA-seq) has driven many advances in the characterization and quantification of transcriptomes. RNA-seq is a powerful gene expression profiling technology enabling transcript discovery and provides a far more precise measure of the levels of transcripts and their isoforms than other methods e.g. microarray.

However, the analysis of RNA-seq data remains a significant challenge for many biologists. The data generated is large and the tools for its assembly, analysis and visualisation are still under development. Assemblies of reads can be inspected using tools such as the Integrative Genomics Viewer (IGV) where visualisation of results involves ‘stacking’ the reads onto a reference genome. Whilst sufficient for many needs, when the underlying variance of the genome or transcript assemblies is complex, this visualisation method can be limiting; errors in assembly can be difficult to spot and visualisation of splicing events may be challenging.

Data visualisation is increasingly recognised as an essential component of genomic and transcriptomic data analysis, enabling large and complex datasets to be better understood. An approach that has been gaining traction in biological research is based on the application of network visualisation and analysis methods. Networks consist of nodes connected by edges (lines), where nodes usually represent an entity and edge a relationship between them. These are now widely used for plotting experimentally or computationally derived relationships between genes and proteins.

The overall aim of this PhD project was to explore the use of network-based visualisation in the analysis and interpretation of RNA-seq data. In chapter 2, I describe the development of a data pipeline that has been designed to go from ‘raw’ RNA-seq data to a file format which supports data visualisation as a ‘DNA assembly graph’. In DNA assembly graphs, nodes represent sequence reads and edges denote a

homology between reads above a defined threshold. Following the mapping of reads to a reference sequence and defining which reads a map to a given loci, pairwise sequence alignments are performed between reads using MegaBLAST. This provides a weighted similarity score that is used to define edges between reads. Visualisation of the resulting networks is then carried out using BioLayout *Express*<sup>3D</sup> that can render large networks in 3-D, thereby allowing a better appreciation of the often-complex network structure. This pipeline has formed the basis for my subsequent work on the exploring and analysing alternative splicing in human RNA-seq data. In the second half of this chapter, I provide a series of tutorials aimed at different types of users allowing them to perform such analyses. The first tutorial is aimed at computational novices who might want to generate networks using a web-browser and pre-prepared data. Other tutorials are designed for use by more advanced users who can access the code for the pipeline through GitHub or via an Amazon Machine Image (AMI).

In chapter 3, the utility of network-based visualisations of RNA-seq data is explored using data processed through the pipeline described in Chapter 2. The aim of the work described in this chapter was to better understand the basic principles and challenges associated with network visualisation of RNA-seq data, in particular how it could be used to visualise transcript structure and splice-variation. These analyses were performed on data generated from four samples of human fibroblasts taken at different time points during their entry into cell division. One of the first challenges encountered was the fact that the existing network layout algorithm (Fruchterman-Reingold) implemented within BioLayout *Express*<sup>3D</sup> did not result in an optimal layout of the unusual graph structures produced by these analyses. Following the implementation of the more advanced layout algorithm FMMM within the tool, network structure could be far better appreciated. Using this layout method, the majority of genes sequenced to an adequate depth assemble into networks with a linear ‘corkscrew’ appearance and when representing single isoform transcripts add little to existing views of these data. However, in a small number of cases (~5%), the networks generated from transcripts expressed in human fibroblasts possess more complex structures, with ‘loops’, ‘knots’ and multiple ends being observed. In a

majority of cases examined, these loops were associated with alternative splicing events, a fact confirmed by RT-PCR analyses. Other DNA assembly networks representing the mRNAs for genes such as *MKI67* showed knot-like structures, which was found to be due to the presence of repetitive sequence within an exon of the gene. In another case, *CENPO* the unusual structure observed was due to reads derived from an overlapping gene of *ADCY3* gene present on the opposite strand with reads being wrongly mapped to *CENPO*. Finally, I explored the use of a network reduction strategy as an approach to visualising highly expressed genes such as *GAPDH* and *TUBA1C*. Having successfully demonstrated the utility of networks in analysing transcript isoforms in data derived from a single cell type I set out to explore its utility in analysing transcript variation in tissue data where multiple isoforms expressed by different cells within the tissue might be present in a given sample.

In chapter 4, I explore the analysis of transcript variation in an RNA-seq dataset derived from human tissue. The first half of this chapter describes the quality control of these data again using a network-based approach but this time based the correlation in expression between genes and samples. Of the 95 samples derived from 27 human tissues, 77 passed the quality control. A network was constructed using a correlation threshold of  $r \geq 0.9$ , which comprised 6,109 nodes (genes) and 1,091,477 edges (correlations) and clustered. Subsequently, the profile and gene content of each cluster was examined and enrichment of GO terms analysed. In the second half of this chapter, the aim was to detect and analyse alternative splicing events between different tissues using the rMATS tool. By using a false-discovery rate (FDR) cut-off of  $< 0.01$ , I found that in comparisons of brain vs. heart, brain vs. liver and heart vs. liver, the program reported 4,992, 4,804 and 3,990 splicing events, respectively. Of these events, only 78 splicing events (52 genes) with more than 50% of exon inclusion level and expression level more than FPKM 30. To further explore the sometimes-complex structure of transcripts diversity derived from tissue, RNA-seq assembly networks for *KLC1*, *SORBS2*, *GUK1*, and *TPM1* were explored. Each of these networks showed different types of alternative splicing events and it was sometimes difficult to determine the isoforms expressed between tissues using other

approaches. For instance, there is an issue in visualising the read assembly of long genes such as *KLC1* and *SORBS2*, using a Sashimi plots or even Vials, just because of the number of exons and the size of their genomic loci. In another case of *GUK1*, tissue-specific isoform expression was observed when a network of three tissues was combined. Arguably the most complex analysis is the network of *TPM1* where the unification step was employed for this highly expressed gene.

In chapter 5, I perform a usability testing for NGS Graph Generator web application and visualising RNA-seq assemblies as a network using BioLayout Express<sup>3D</sup>. This test was important to ensure that the application is well received and utilised by the user. Almost all participants of this usability test agree that this application would encourage biologists to visualise and understand the alternative splicing together with existing tools. The participants agreed that Sashimi plots rather difficult to view and visualise and perhaps would lose something interesting features. However, there were also reviews of this application that need improvements such as the capability to analyse big network in a short time, side-by-side analysis of network with Sashimi plot and Ensembl. Additional information of the network would be necessary to improve the understanding of the alternative splicing.

In conclusion, this work demonstrates the utility of network visualisation of RNA-seq data, where the unusual structure of these networks can be used to identify issues in assembly, repetitive sequences within transcripts and splice variation. As such, this approach has the potential to significantly improve our understanding of transcript complexity. Overall, this thesis demonstrates that network-based visualisation provides a new and complementary approach to characterise alternative splicing from RNA-seq data and has the potential to be useful for the analysis and interpretation of other kinds of sequencing data.

## Lay Summary

Alternative splicing is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene. In these recent years, sequencing of RNA (RNA-seq) has emerged as the favoured technology for the simultaneous measurement of transcript sequences and expression abundance. The data visualisation of RNA-seq data presents novel challenges and many methods have been developed for the purpose of building a network and visualising transcript variation. In the first part of my thesis, I developed a network-based pipeline for preparing for visualising RNA-seq datasets from a 'raw' data to a layout file which can be visualised using a network analysis tool, BioLayout *Express3D*. This pipeline formed the basis for my subsequent work on the exploration and analysis of alternative splicing in a single cell type and human tissue sample. I explore the optimal parameters for network analysis of these data and interpreting the resulting the DNA assembly graphs, in particular how this approach can be used to better define transcript variation and alternative splicing. Most of the network structure generated from human fibroblast data generate networks is in a linear structure but in a small number of cases with more complex structures, with 'loops', 'knots' and multiple ends being observed. Since alternative splicing in gene expression is thought to regulate many of the isoforms differences between tissues, visualising and analysing the splicing variant transcript of a gene responsible for these changes is an important goal of molecular biology. For this, I explore and analyse the alternative splicing divergence of human tissue gene and isoform expression using network-based approach and splice variant detector tool. I demonstrate that gene expression from the network analysis diverges extensively between tissues. In conclusion, the utility of this approach for RNA-seq data, including the unusual structure of these networks and how they can be used to identify issues in assembly, repetitive sequences within transcripts and splice variation. This approach has the potential to significantly improve our understanding of transcript complexity. In overall, the network-based visualisation can be an alternative way over the current existing visualisation platform to visualise and characterise alternative splicing of RNA-seq of such data.



# Acknowledgements

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

All praises due to Allah, the Most Gracious, the Most Merciful. Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. I am thankful to Allah, who supplied me with the courage, the guidance, and the love to complete this research.

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people. I would like to take this opportunity to express my sincere thanks to those who have involved and supported me towards the successful completion of my PhD study.

First and foremost, my deepest gratitude goes to my supervisor, Professor Tom Freeman, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. Thank you very much indeed, for your suggestions, criticisms and guidance. Your invaluable efforts, profound expertise, patience and kindness towards developing my professional knowledge, especially in the field of network analysis, biology and technical writing skills are highly appreciated.

This appreciation also goes to Tim Angus for his valuable help, input and kindness especially in the field of computer science. The most enjoyable part of this PhD has been meeting a wonderful group of fellow researchers especially to my collaborator at European Bioinformatics Institute (EBI), Hinxton, Cambridge; Dr Anton Enright, Dr Stijn van Dongen, Dr Harpreet Saini, Dr Matthew Davis and Dr Matloob Qureshi.

I would also like to thank my committee members, Professor Mick Watson (co-advisor), Professor John Wolliams (Chairman of PhD committee), Professor Kim Summers (Chairman of Thesis Writing Committee), and Dr. Tom Michael (former co-advisor) for serving as my committee members even at hardship. I also want to

thank you for letting my defence be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

To all my friends in the lab, Dr. Mark Barnett, Dr. David Sz-hau Chen, Dr. Bruno Giotti, Derek Wright, Tim Regan, Ajit Nirmal, Dr. Wu Zhaozong, Dr. Barbara Shih and Dr. Khoo Choon Kiat from The Roslin Institute, for your help and encouragement are greatly appreciated.

Special thanks also to my financial sponsor, Majlis Amanah Rakyat (MARA) for the scholarship and administrative support. Also thanks to the University of Edinburgh for the facilities and administrative support during my study. This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

Special appreciation goes to my beloved wife, Dr Nur Annies, for your love, encouragement, patience, faith in me and endless support throughout the years of effort. Who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries, for keeping me sane over the past few months. But most of all, thank you for being my best friend. I owe you everything.

Last, but not least, to my lovely family, especially to my dear parents Dato' Haji Wan Mohamad Nazarie bin Wan Mahmood and Datin Hajjah Norsina binti Panot, my family and my parent-in-law Tuan Haji Abd Hadi bin Ibrahim and Puan Hajjah Mazinah binti Sulaiman and family. Thank you for your love, support, and unwavering belief in me.

# Table of Contents

|   |          |
|---|----------|
| Declaration of Originality                                  | ii       |
| Abstract  | iii      |
| Lay Summary   | vii      |
| Acknowledgments   | viii     |
| Table of Contents   | x        |
| List of Figures   | xiv      |
| List of Tables  | xvii     |
| List of Acronyms and Abbreviations                          | xviii    |
| List of Publications  | xx       |
| <br>  |          |
| <b>Chapter 1 – Introduction</b>                             | <b>1</b> |
| 1.1 History of DNA sequencing                               | 1        |
| 1.1.1 The Sanger Method                                     | 1        |
| 1.1.2 Automated DNA sequencing                              | 3        |
| 1.2 Next-generation DNA sequencing                          | 4        |
| 1.3 Life cycle of an mRNA                                   | 5        |
| 1.3.1 Transcription   | 6        |
| 1.3.2 Transcription and mRNA processing                     | 7        |
| 1.3.3 Life of mRNA in cell                                  | 8        |
| 1.4 The splicing reaction                                   | 9        |
| 1.4.1 Major mechanisms in mRNA splicing                     | 10       |
| 1.4.2 The major spliceosome splicing                        | 11       |
| 1.4.3 Diversification of genes through alternative splicing | 14       |
| 1.4.4 Splicing regulation                                   | 17       |
| 1.5 Transcriptomic studies using RNA sequencing             | 20       |
| 1.5.1 RNA-seq vs microarrays                                | 21       |
| 1.5.2 RNA-seq workflow                                      | 22       |
| 1.5.3 RNA-seq using Illumina sequencing technology          | 24       |
| 1.5.4 Quality assessment of RNA-seq data                    | 27       |
| 1.5.5 Read mapping strategy                                 | 27       |
| 1.5.5.1 Alignment to the genome or transcriptome            | 29       |
| 1.5.5.2 <i>De novo</i> assembly                             | 31       |
| 1.5.6 The estimation of expression levels                   | 31       |
| 1.5.6.1 Gene expression levels                              | 33       |
| 1.5.6.2 Transcript expression levels                        | 34       |
| 1.5.7 Read count normalisation                              | 34       |
| 1.5.8 Differential alternative splicing                     | 35       |
| 1.5.9 Challenges in RNA-seq                                 | 35       |
| 1.6 Visualisation   | 36       |
| 1.6.1 Visualisation of co-expression gene network           | 37       |
| 1.6.1.1 Graph clustering                                    | 39       |
| 1.6.2 Network visualisation                                 | 40       |
| 1.6.3 Algorithms for sequence assemblies                    | 41       |

|   |  |            |
|---|--|------------|
| 1.6.3.1   | The Overlap-Layout-Consensus (OLC) algorithm                       | 42         |
| 1.6.3.2   | The de Bruijn Graph (DBG) algorithm                                | 42         |
| 1.6.4   | Visualisation of RNA-seq assemblies                                | 44         |
| 1.7   | Aims of the thesis   | 48         |
| <b>Chapter 2 - Development of an analysis pipeline for the network analysis of RNA-seq data</b> |  | <b>50</b>  |
| 2.1   | Introduction   | 50         |
| 2.2   | Correction of GraphNGS pipeline                                    | 52         |
| 2.3   | Network-based visualisation pipeline: an alternative solution      | 56         |
| 2.4   | Component of NGS Graph Generator                                   | 62         |
| 2.5   | Implementation network-based visualisation pipeline                | 64         |
| 2.5.1   | Web-based application  | 64         |
| 2.5.2   | Public repository  | 64         |
| 2.5.3   | Amazon machine image (AMI)   | 65         |
| 2.6   | Discussion   | 66         |
| <b>Chapter 3 - Network-based visualisation and analysis of RNA-seq data</b>                     |  | <b>70</b>  |
| 3.1   | Introduction   | 70         |
| 3.2   | Methods  | 71         |
| 3.2.1   | RNA-seq data used for these studies                                | 71         |
| 3.2.2   | RNA-seq data processing  | 71         |
| 3.2.3   | Network layout   | 72         |
| 3.2.4   | Optimisation of read comparison parameters                         | 73         |
| 3.2.5   | Collapsing of redundant reads                                      | 74         |
| 3.2.6   | Analysis of the network structure                                  | 74         |
| 3.2.7   | Validation of splice variant using RT-PCR                          | 74         |
| 3.3   | Results  | 75         |
| 3.3.1   | Optimisation of network visualisation ‘perfect’ overlap data       | 75         |
| 3.3.2   | Optimisation of read-to-read comparison similarity score Threshold | 79         |
| 3.3.3   | Network visualisation of transcripts                               | 82         |
| 3.3.3.1   | Network reduction  | 84         |
| 3.3.3.2   | Splice variant network structure                                   | 86         |
| 3.3.3.3   | Issues with assembly and internal repeats                          | 88         |
| 3.3.3.4   | Highly expressed gene network analysis                             | 91         |
| 3.3.3.5   | Internal splicing network structure                                | 95         |
| 3.3.3.6   | Three loops network structure                                      | 96         |
| 3.3.3.7   | Alternative splice network structure                               | 98         |
| 3.4   | Discussion   | 100        |
| <b>Chapter 4 – An analysis of transcript variation in human tissues</b>                         |  | <b>106</b> |
| 4.1   | Introduction   | 106        |
| 4.1.1   | Alternative splicing analysis                                      | 106        |

|  |  |            |
|--|--|------------|
| 4.1.2  | rMATS – one of the best tool to detect alternative splicing                | 108        |
| 4.1.3  | Visualising alternative splicing   | 109        |
| 4.2  | Methods  | 110        |
| 4.2.1  | Datasets   | 111        |
| 4.2.2  | Read alignment and quantification  | 111        |
| 4.2.3  | Quality control and data analysis  | 112        |
| 4.2.4  | Functional annotation  | 113        |
| 4.2.5  | Differential splicing  | 113        |
| 4.2.6  | Analysis of the network structure  | 114        |
| 4.3  | Results  | 115        |
| 4.3.1  | Quality control of human tissue atlas RNA-seq                              | 115        |
| 4.3.2  | Network construction and layout  | 115        |
| 4.3.3  | Analysis of alternative splicing between tissue using rMATS                | 126        |
| 4.3.4  | Network analysis of alternative splicing transcripts of human tissue atlas | 131        |
| 4.3.4.1  | Analysis of <i>KLC1</i>  | 131        |
| 4.3.4.2  | Analysis of <i>GUK1</i>  | 142        |
| 4.3.4.3  | Analysis of <i>SORBS2</i>  | 148        |
| 4.3.4.4  | Analysis of <i>TPM1</i>  | 156        |
| 4.3.5  | Comparing visualisation approaches   | 162        |
| 4.4  | Discussion   | 164        |
| <b>Chapter 5 – Evaluating the usability of network-based visualisation approach using NGS Graph Generator &amp; BioLayout Express<sup>3D</sup></b> |  | <b>169</b> |
| 5.1  | Introduction   | 169        |
| 5.2  | Method   | 170        |
| 5.2.1  | Test metrics   | 170        |
| 5.2.1.1  | Successful task completion   | 170        |
| 5.2.1.2  | Critical errors  | 170        |
| 5.2.1.3  | Non-critical errors  | 171        |
| 5.2.1.4  | Likes, dislikes, and future recommendations                                | 171        |
| 5.2.2  | Usability test   | 171        |
| 5.2.2.1  | Session introduction   | 171        |
| 5.2.2.2  | Pre-test briefing  | 171        |
| 5.2.2.3  | Tasks  | 172        |
| 5.2.2.4  | Questionnaire survey   | 173        |
| 5.2.2.5  | Participants   | 173        |
| 5.2.2.6  | Location and usability setup   | 174        |
| 5.3  | Results  | 174        |
| 5.3.1  | Successful task completion   | 174        |
| 5.3.2  | Finding a network figure of a gene   | 174        |
| 5.3.3  | Generating a layout file   | 176        |
| 5.3.4  | Opening BioLayout Express <sup>3D</sup>                                    | 178        |
| 5.3.5  | Visualising network using BioLayout Express <sup>3D</sup>                  | 178        |
| 5.3.6  | Determining alternative splicing in the network                            | 178        |
| 5.3.7  | Questionnaire survey   | 178        |

|   |   |            |
|---|---|------------|
| 5.3.7.1   | Question 1: What field are you working on?  | 178        |
| 5.3.7.2   | Question 2: What organism are you working on?   | 179        |
| 5.3.7.3   | Question 3: How do you analyse your RNA-seq data?   | 179        |
| 5.3.7.4   | Question 4: How do you visualise your RNA-seq data?   | 180        |
| 5.3.7.5   | Question 5: How do you find visualising your data using IGV/Sashimi plot?   | 181        |
| 5.3.7.6   | Question 6: Have you ever used BioLayout <i>Express</i> <sup>3D</sup> /Miru for visualisation?  | 181        |
| 5.3.7.7   | Question 7: Do you know that we can visualise RNA-seq assemblies of a gene as a network in BioLayout <i>Express</i> <sup>3D</sup> /Miru?  | 182        |
| 5.3.7.8   | Question 8: Here is an example of a RNA-seq data of <i>LRR1</i> gene using network-based visualisation. How do you find this network visualisation of <i>LRR1</i> gene compare to Sashimi plot in term of splice variant? | 183        |
| 5.3.7.9   | Question 9: Will you use NGS Graph Generator in the future?   | 184        |
| 5.3.7.10  | Question 10: What do you like about NGS Graph Generator? Explain  | 184        |
| 5.3.7.11  | Question 11: What do you dislike about NGS Graph  |            |
| 5.3.7.12  | Question 12: Can you give overall feedback, suggestions or recommendation for this application, NGS Graph Generator?  | 185        |
| 5.4   | Discussion and future recommendation works  | 187        |
| 5.5   | Conclusions   | 188        |
| <b>Chapter 6 – General discussion and conclusions</b> |   | <b>189</b> |
| <b>References</b>                                     |   | <b>200</b> |
| <b>Supplementary Materials</b>                        |   | <b>220</b> |
| <b>Supplementary Chapter 4</b>                        |   | <b>220</b> |
| <b>Supplementary Chapter 5</b>                        |   | <b>230</b> |

# List of Figures

## Chapter 1

|              |   |    |
|--------------|---|----|
| Figure 1.1:  | A schematic representation of Sanger sequencing                                 | 2  |
| Figure 1.2:  | Primary stages in the regulation of eukaryotic gene expression                  | 6  |
| Figure 1.3:  | Splicing by the major spliceosome   | 13 |
| Figure 1.4:  | Types of alternative splicing   | 16 |
| Figure 1.5:  | Regulation of splicing regulators   | 18 |
| Figure 1.6:  | An overview of library preparation and sequencing steps in an Illumina platform | 26 |
| Figure 1.7:  | An overview of the mapping algorithm implemented in TopHat                      | 30 |
| Figure 1.8:  | GenomicRanges overview  | 33 |
| Figure 1.9:  | Network visualisation and clustering of the pig transcriptome                   | 39 |
| Figure 1.10: | Differences between an OLC and a DBG for assembly                               | 44 |
| Figure 1.11: | Example Sashimi plot for an alternatively spliced exon                          | 47 |
| Figure 1.12: | Vials - Visualization of Alternative Splicing                                   | 48 |

## Chapter 2

|             |  |    |
|-------------|--|----|
| Figure 2.1: | The basic principle of GraphNGS pipeline                 | 53 |
| Figure 2.2: | Network visualisation of <i>COL5A1</i> transcript        | 53 |
| Figure 2.3: | Network transcript of <i>COL5A1</i>                      | 54 |
| Figure 2.4: | Correction of the GraphNGS pipeline                      | 55 |
| Figure 2.5: | Pipeline for network-based visualisation of RNA-seq data | 59 |
| Figure 2.6: | Framework of NGS Graph Generator                         | 62 |

## Chapter 3

|              |  |    |
|--------------|--|----|
| Figure 3.1:  | Optimisation of network layout   | 77 |
| Figure 3.2:  | Network layout - quality vs. speed   | 78 |
| Figure 3.3:  | Optimisation plot of ‘synthetic’ data of <i>COL5A1</i>   | 80 |
| Figure 3.4:  | Optimisation plot of ‘real’ data of four different complex genes                               | 81 |
| Figure 3.5:  | Typical networks of RNA-seq data derived from linear transcripts                               | 83 |
| Figure 3.6:  | Read unification of highly expressed genes <i>TUBA1C</i> and <i>GAPDH</i> in human fibroblasts | 85 |
| Figure 3.7:  | Splice variant visualisation and confirmation  | 87 |
| Figure 3.8:  | Complex gene network structure   | 90 |
| Figure 3.9:  | Network-based visualisation of <i>TPM1</i> of human fibroblast at 24 after serum refeeding     | 94 |
| Figure 3.10: | Network-based visualisation of <i>BUB3</i> of human fibroblast                                 | 96 |
| Figure 3.11: | Network-based visualisation of <i>FAM64A</i> of human fibroblast                               | 98 |

|                  |   |     |
|------------------|---|-----|
| Figure 3.12:     | Network-based visualisation of <i>NRM</i> of human fibroblast                                     | 100 |
| <b>Chapter 4</b> |   |     |
| Figure 4.1:      | Data analysis workflow  | 111 |
| Figure 4.2:      | Clustering of human tissue data sample  | 117 |
| Figure 4.3:      | Network visualisation and clustering of the human tissue atlas<br>RNA-seq data                    | 119 |
| Figure 4.4:      | Expression profile of top 20 clusters from human tissue atlas<br>RNA-seq data                     | 123 |
| Figure 4.5:      | Summary of different types of significant AS events   | 127 |
| Figure 4.6:      | Visualisation of <i>KLC1</i> transcript in the human brain  | 134 |
| Figure 4.7:      | Visualisation of AS gene of <i>KLC1</i> in the human heart  | 136 |
| Figure 4.8:      | Issue with DNA read-assembly  | 139 |
| Figure 4.9:      | Network-based visualisation of <i>APOPT1</i> transcript in the<br>brain and heart                 | 141 |
| Figure 4.10:     | Visualisation of <i>GUK1</i> transcript in human brain  | 145 |
| Figure 4.11:     | Visualisation of <i>GUK1</i> transcript in the heart  | 146 |
| Figure 4.12:     | Visualisation of <i>GUK1</i> transcript in liver  | 147 |
| Figure 4.13:     | Visualisation of the <i>SORBS2</i> transcript in the heart  | 152 |
| Figure 4.14:     | Network-based visualisation of the <i>SORBS2</i> transcript in<br>the liver                       | 153 |
| Figure 4.15:     | Network-based visualisation of the <i>SORBS2</i> transcript in<br>the brain                       | 153 |
| Figure 4.16:     | Vials – visualizing AS of genes   | 155 |
| Figure 4.17:     | Visualisation of RNA-seq data of <i>TPM1</i>  | 157 |
| Figure 4.18:     | Network-based visualisation of <i>TPM1</i> transcripts from<br>a different tissue                 | 160 |
| <b>Chapter 5</b> |   |     |
| Figure 5.1:      | Amount of people who completed each task  | 174 |
| Figure 5.2:      | Finding network assemblies of a gene  | 175 |
| Figure 5.3:      | <i>CERS5</i> gene information   | 176 |
| Figure 5.4:      | A network-based visualisation pipeline page   | 177 |
| Figure 5.5:      | Job sent confirmation page  | 178 |
| Figure 5.6:      | Result page   | 178 |
| Figure 5.7:      | Question 1: What field are you working on?  | 180 |
| Figure 5.8:      | Question 2 - What organism are you working on?  | 180 |
| Figure 5.9:      | Question 3: How do you analyse your RNA-seq data?   | 181 |
| Figure 5.10:     | Question 5: How do you visualise your RNA-seq data?   | 181 |
| Figure 5.11:     | Visualisation of RNA-seq data using IGV and Sashimi plots   | 182 |
| Figure 5.12:     | Question 5: How do you find visualising your data using<br>IGV/Sashimi plots?                     | 182 |
| Figure 5.13:     | Question 6: Have you ever used BioLayout<br><i>Express</i> <sup>3D</sup> /Miru for visualisation? | 183 |



|              |   |     |
|--------------|---|-----|
| Figure 5.14: | Question 7: Do you know that we can visualise RNA-seq data of a gene as a network in BioLayout <i>Express</i> <sup>3D</sup> /Mirus? | 184 |
| Figure 5.15: | Network visualisation of <i>LRR1</i>  | 185 |
| Figure 5.16: | Question 9: Will you use NGS Graph Generator in the future?   | 185 |

# List of Tables

## Chapter 4

|            |   |     |
|------------|---|-----|
| Table 4.1: | List of samples removed   | 118 |
| Table 4.2: | List of 20 largest gene clusters  | 124 |
| Table 4.3: | Differential splicing events in human tissue atlas ranked by FDR value                        | 128 |
| Table 4.4: | Summary of visualisation analysis of <i>KLC1</i> between the brain and heart tissue           | 138 |
| Table 4.5: | Summary of visualisation analysis of <i>GUK1</i> between the brain, heart, and liver tissue   | 148 |
| Table 4.6: | Summary of visualisation analysis of <i>SORBS2</i> between the heart, liver, and brain tissue | 156 |
| Table 4.7: | Summary of visualisation analysis of <i>TPM1</i> between the heart, liver, and brain tissue   | 162 |

## Chapter 5

|            |   |     |
|------------|---|-----|
| Table 5.1: | Usability test task   | 172 |
| Table 5.2: | Breakdown of participants who participated in the usability test and questionnaire survey with the familiarity of BioLayout <i>Express</i> <sup>3D</sup> /Miru software | 173 |

## Supplementary Chapter 4

|                          |  |     |
|--------------------------|--|-----|
| Supplementary Table 4.1: | Details list of 95 samples from human tissue atlas                                   | 222 |
| Supplementary Table 4.2: | Full list of differentially spliced events in human tissue atlas ranked by FDR value | 227 |

## List of Acronyms and Abbreviations

|             |   |
|-------------|---|
| 2-D         | 2-dimensional   |
| 3-D         | 3-dimensional   |
| A3SS        | alternative 3' splice sites   |
| A5SS        | alternative 5' splice sites   |
| AS          | alternative splicing  |
| APA         | alternative polyadenylation   |
| AMI         | Amazon Machine Image  |
| AWS         | Amazon Web Service  |
| BAM         | binary alignment/map  |
| BLAST       | Blast-Like Alignment Sequence Tool  |
| Bp          | basepairs   |
| cDNA        | complementary DNA   |
| CDSs        | coding sequences  |
| DAVID       | The Database for Functional Annotation,<br>Visualization and Integrated Discovery |
| DBG         | de Bruijn graph   |
| DNA         | deoxyribonucleic acid   |
| dNTPs       | <i>deoxynucleotides</i>   |
| ddNTPs      | <i>dideoxynucleotides</i>   |
| EBI         | European Bioinformatics Institute   |
| EC2         | Elastic Cloud Compute   |
| EJC         | exon-junction complex   |
| EM          | electron microscopy   |
| EMBL        | European Molecular Biology Laboratory   |
| ESE         | exonic splicing enhancers   |
| ESTs        | expressed sequence tags   |
| FDR         | false discovery rate  |
| FMMM        | Fast multipole multiple multilevel  |
| FPKM        | Fragment Per Kibobase of transcript per<br>Million mapped reads                   |
| F-R         | Fruchtermann-Reingold   |
| GENCODE     | Genome research of ENCyclopedia Of DNA<br>Elements                                |
| GFF         | general feature format  |
| GL          | graph layout  |
| GTF         | general transfer format   |
| GO          | gene ontology   |
| GUI         | graphical user interface  |
| <i>GUK1</i> | guanylate kinase 1  |

|         |  |
|---------|--|
| HTML    | hyper text markup language                               |
| IaaS    | Infrastructure as a Service                              |
| IGV     | Integrative Genomics Viewer                              |
| mRNA    | messenger ribonucleic acid                               |
| MXE     | mutually exclusive exon                                  |
| NHDF    | Normal human dermal fibroblast                           |
| NGS     | next-generation sequencing                               |
| NMD     | nonsense-mediated decay                                  |
| OGDF    | Open Graph Drawing Framework                             |
| OLC     | overlap-layout-consensus                                 |
| PCR     | Polymerase chain reaction                                |
| QC      | quality control  |
| RNA     | Ribonucleotides acid                                     |
| RNA-seq | RNA sequencing   |
| tRNAs   | transfer RNAs  |
| PDF     | portable document file                                   |
| PHP     | hypertext preprocessor                                   |
| PPT     | polypyrimidine tract                                     |
| Poly-A  | poly-adenylation   |
| PSI     | percentage spliced inclusion                             |
| RBP(s)  | RNA binding proteins                                     |
| rRNAs   | ribosomal RNAs   |
| RPKM    | Read Per Kibobase of transcript per Million mapped reads |
| RT-PCR  | reverse transcriptase polymerase chain reaction          |
| SaaS    | Software as a Service                                    |
| SAM     | sequence alignment/map                                   |
| SE      | skipped exon   |
| SFs     | splicing factors   |
| SMRT    | single molecule real time                                |
| snRNAs  | small nuclear RNA molecules                              |
| SR      | Serine-Rich  |
| TFs     | transcription factors                                    |
| TIGR    | The Institute for Genomic Research                       |
| MCL     | Markov clustering  |
| rMATS   | (replicate) Multivariate analysis of transcript splicing |
| TPM     | Transcript per million                                   |
| UCSC    | University of California Santa Carlo                     |
| Vials   | Visualizing alternative splicing                         |

# List of Publications

## In this study

1. **Fahmi W. Nazarie**, Tim Angus, Sz-Hau Chen, Mark Barnett, Karsten Klein, Harpreet Kaur, Mick Watson, Stijn van Dongen, Anton J. Enright, Tom C. Freeman. Network-based visualisation and analysis of RNA-seq data (2017). (Manuscript in preparation).

## Conferences

2. **Fahmi W. Nazarie**, Tim Angus, Sz-Hau Chen, Mark Barnett, Anton J. Enright, Tom C. Freeman. Network-based visualisation and analysis of RNA-seq data. *23<sup>rd</sup> Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and European Conference on Computational Biology (ECCB)*, Dublin, Ireland 10<sup>th</sup> – 14<sup>th</sup> July 2015.
3. **Fahmi W. Nazarie**, Tim Angus, Sz-Hau Chen, Mark Barnett, Anton J. Enright, Tom C. Freeman. Analysis of human fibroblast RNA-seq data using network-based visualisation approach. *13<sup>th</sup> European Conference on Computational Biology (ECCB)*, Strasbourg, France. 7<sup>th</sup>-10<sup>th</sup> September 2014.
4. **Fahmi W. Nazarie**, Tim Angus, Sz-Hau Chen, Mark Barnett, Anton J. Enright, Tom C. Freeman. Development a network-based visualisation pipeline of RNA-seq data. *Functional Genomics and Systems Biology*, Wellcome Trust Genome Campus, *Hinxton*, Cambridge, United Kingdom. 19<sup>th</sup>-28<sup>th</sup> June 2013.

## Previous study

5. Huan Yong Yap, Kamal Ghazali, **Wan Fahmi Wan Mohamad Nazarie**, Mohd Noor Mat Isa, Zunita Zakaria, Abdul Rahman Omar. Draft Genome Sequence of *Pasteurella multocida* subsp. *multocida* Strain PMTB, Isolated from a Buffalo. (2013). *Genome Announcements*. 5(1), e00872-13.
6. **Wan Fahmi Wan Mohamad Nazarie**, Mohd Noor Mat Isa, Zunita Zakaria, Abdul Rahman Omar. Genome Sequencing and Bioinformatic Analysis of *Pasteurella multocida* Serotype B: 2 Strain PMTB. *Proceedings of 21st Veterinary Association Malaysia (VAM) Congress 2009*. 7-9 August 2009. The Legend, Water Chalets, Port Dickson, Negeri Sembilan, Malaysia.

# Chapter 1 - Introduction

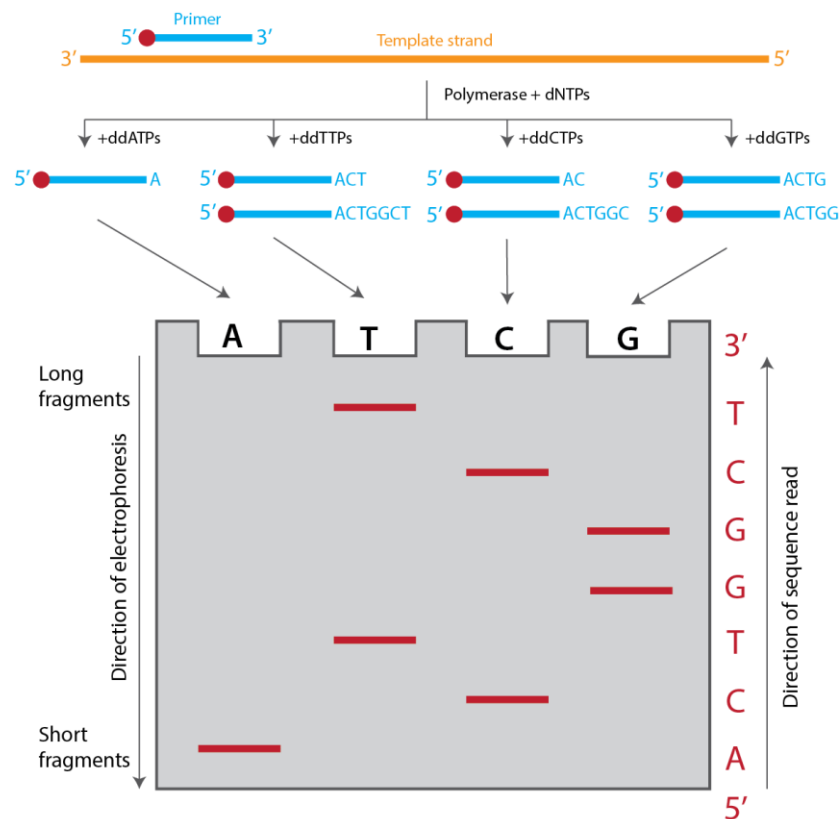
## 1.1 History of DNA sequencing

When the structure of DNA was discovered in 1953 by Watson and Crick (Watson and Crick, 1953), it was one of the most significant scientific discoveries of the 20th century. It was followed by the innovation of a new chemical degradation method for sequencing DNA (Maxam and Gilbert, 1977) and the enzymatic chain termination technique (Sanger et al., 1977a). These technologies would transform biology as a whole by providing a tool for complete sequencing genes and eventually entire genomes. The Sanger method was successfully applied to sequence 5,375 DNA nucleotides of bacteriophage  $\phi$ X174 (Sanger et al., 1977b) and later by a group from the Medical Research Council (MRC), Cambridge, decoding 172,282 base pairs of the Epstein–Barr virus (Baer et al., 1984).

### 1.1.1 The Sanger Method

The Sanger method (Sanger et al., 1977), also known as the dideoxynucleotide chain termination method. The method uses DNA polymerase to replicate DNA in the presence of a mix of the four deoxynucleotides (dNTPs), with a small amount of one of the four dideoxynucleotides (ddNTPs) for generating DNA fragments terminated at a specific nucleotide. It required single-stranded DNA molecules as a template, a DNA polymerase, DNA primer, normal dNTPs and dents that terminate DNA strand elongation. A primer was annealed to a specific region on the DNA template strand, which provided a starting point to synthesise a new DNA strand in the presence of DNA polymerase. The ddNTPs lack the 3'-hydroxyl group of dNTPs, which is required for phosphodiester bond formation between one nucleotide and the following nucleotide during DNA strand extension. The reaction was conducted in four separate tubes, each containing one of the four ddNTPs and the four normal dNTPs. All generated fragments had the same 5'-end, whereas the residue at the 3'-end was determined by the specific ddNTP used in the reaction. DNA fragments were labelled using a radioactive dNTP. The fragments resulting from these reactions were separated by size on thin denaturing slab polyacrylamide gels in four parallel

lanes. The bands on the autoradiogram were identified by virtue of  $^{35}\text{S}$ , which was the radioactive label on one dNTP. It showed the size of fragments terminated with a specific nucleotide and the sequence could be determined by assembling the fragments in order of size (**Figure 1.1**).



**Figure 1.1: A schematic representation of Sanger sequencing.** The Sanger method used dideoxynucleotides that terminate newly synthesised DNA fragments at specific bases either A, T, C or G. Then, the resulting fragments were resolved by electrophoresis on a denaturing polyacrylamide gel in four parallel lanes, and the DNA sequence was read. Figure redrawn from Rosenberg and Pascual (2014).

The initial Sanger sequencing method has been subjected to several significant improvements and developed remarkably over three decades. Cloning of DNA fragments into a plasmid vector was originally required in Sanger sequencing, but the polymerase chain reaction (PCR) (Saiki et al., 1988) for the amplification of specific DNA fragments *in vitro* has been extensively applied in the field of Sanger sequencing. The development of the technique for labelling of the chain terminator

ddNTPs with four different fluorescent dyes as an alternative to radioisotope labelling allowed sequencing in a single reaction tube and was also the foundation for the use of automated DNA sequencing instruments (Ansorge et al., 1987). The technology advancement of capillary electrophoresis together with highly sensitive detection of nucleotides and a high degree of parallelisation remarkably enhanced the throughput of Sanger sequencing data (Mardis, 2013). Automated DNA sequencing machines can sequence up to 384 fluorescently labelled samples in a single batch. The machines automatically carry out capillary electrophoresis for size separation, detection, and recording of dye fluorescence, and output data as fluorescent peak trace chromatograms. The read length of DNA fragments generated by Sanger sequencing is approximately 500 – 1000 base pairs (bp). The Human Genome Project (HGP) was only possible due to the innovative technological advancements as described above (International Human Genome Sequencing Consortium, 2004).

### 1.1.2 Automated DNA sequencing

Since the 1990s, DNA sequencing has almost always been carried out with semi-automated implementations of the Sanger technique (Shendure and Ji, 2008). In high-throughput production pipelines, the DNA to be sequenced is prepared either for shotgun *de novo* sequencing or a targeted sequencing approach. In the first approach, randomly fragmented DNA is cloned into a high copy number plasmid and bacterial artificial chromosomes (BACs) which are then transformed into *Escherichia coli*, while in the second approach, PCR amplification is carried out with primers that flank the target (Shendure and Ji, 2008).

The first genome of a free-living species ever sequenced was *Haemophilus influenzae*, which was published in 1995 (Fleischmann et al., 1995). The genome was sequenced at The Institute for Genomic Research (TIGR) using the whole-genome shotgun sequencing method. Data from this project, which included 1,830,137 bp of DNA and 1743 predicted genes, showed for the first time the full genetic complement of a bacterial organism. Numerous other bacteria were sequenced within five years of the publication, including *Mycobacterium tuberculosis* (Cole et al., 1998), one of the most important human bacterial



pathogens, *Escherichia coli* (Blattner et al., 1997) and the first archaeon, *Archaeoglobus fulgidus* (Klenk et al., 1997). In 2001, the first consensus sequence of the human genome was obtained by using Sanger sequencing (Lander et al., 2001; Venter et al., 2001) and the first individual human diploid sequence was published in 2007 (Levy et al., 2007). Other bacterial genome sequences, along with the large genomes of mammals such as human (Lander et al., 2001), mouse (Mouse Genome Sequencing Consortium, 2002) and chimpanzee (The Chimpanzee Sequencing and Analysis Consortium, 2005) also have been sequenced and characterised.

## 1.2 Next-generation DNA sequencing

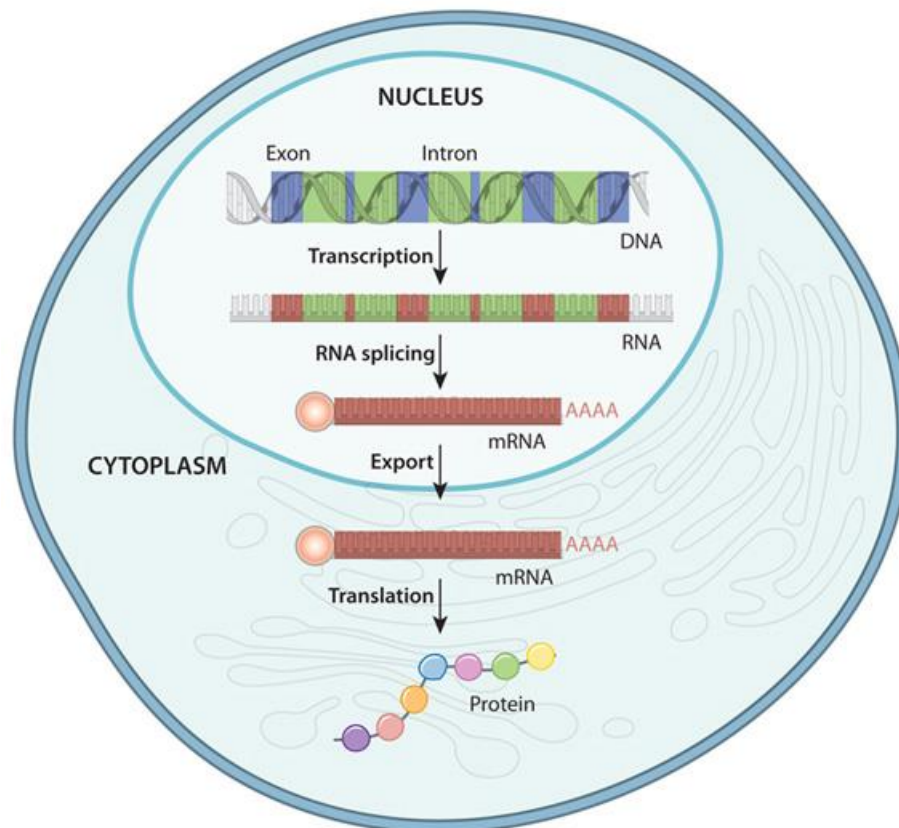
The second complete genome was sequenced using new next-generation sequencing (NGS) technology and marked the first human genome (James D. Watson) sequenced with this technology (Wheeler et al., 2008). NGS technologies have had an astounding impact on genomic research since first introduced to the market in 2005. The technologies have been used for general sequencing applications, such as genome sequencing and re-sequencing, and for novel applications previously unexplored by classic Sanger sequencing (van Dijk et al., 2014; Morozova and Marra, 2008). Using today's NGS sequencers, billions of DNA molecules can be sequenced, but the read lengths are relatively short, from 100 to 500 bp. The third-generation sequencer of the PacBio RS system, which was commercialised in 2010 by Pacific Bioscience, is the first single molecule real-time DNA sequencer (Eid et al., 2009). The read lengths can be very long, up to 20,000 bp, compared with the second-generation sequencers where read length depends on the size of the library fragments and the time of data generation. The error rate is higher (15%) compared to second-generation sequencers on a per read basis, but the accuracy achieved from multiple passes on a single molecule can exceed 99% (Rosenberg and Pascual, 2014).

These sequencing technologies include a number of methods that are commonly shared, such as template preparation, sequencing and imaging and data analysis. The exclusive combination of specific protocols distinguishes one technology from another and determines the type of data produced from different types of available

platforms (Metzker, 2010). These new technologies are rapidly expanding and have been modified to include sequencing and quantification of RNA as well as genomic sequencing. Details of the sequencing approach using the most common Illumina platform are given in Section 1.5. With the exponential increase in genomic and RNA sequencing data, imminent challenges include the development of vigorous protocols for generating sequencing libraries, building efficient new approaches for downstream analysis, handling of large data sets and often re-evaluation of experimental design (Buermans and den Dunnen, 2014; Shendure and Ji, 2008). The work described in this thesis examines the use of these sequencing technologies to understand the structure, processing, and expression of RNA molecules and how this has increased our knowledge of how information is transferred from the genomic DNA in the nucleus to the cell and its environment.

### **1.3 Life cycle of an mRNA**

Since this thesis focuses on next-generation sequencing approaches to understanding RNA, this section reviews current knowledge of the RNA molecule that conveys information from DNA to the ribosome for the synthesis of protein, the messenger RNA (mRNA). mRNAs carry the essential information for the synthesis of proteins. The half-lives of mRNAs are relatively short compared to other molecules in a cell, for instance, most mammalian mRNAs remain in the cell for roughly 9 hours while proteins can last for 46 hours (Schwanhäusser et al., 2011). These mRNA molecules are controlled by a complex regulatory system that determines which messages are finally expressed (**Figure 1.2**).



**Figure 1.2: Primary stages in the regulation of eukaryotic gene expression.** Gene expression commences with nuclear transcription of specific DNA loci such as genes which contain the information required for the synthesis of proteins required by the cell. After several processing stages, the transcription products are then converted into mature mRNAs which will be exported to the cytosol. At this stage, a stringent quality control mechanism is taking place where unprocessed RNAs and RNA fragments will be degraded. When the mRNA is in the cytosol, it will be recognised by ribosomes and translated into protein, and finally, this mRNA will be degraded (<http://www.nature.com/scitable/topicpage/gene-expression-14121669>).

### 1.3.1 Transcription

Transcription is the first step in determining the set of RNAs expressed in a cell and is a primary control point of gene expression (Mutalik et al., 2013; Porrua and Libri, 2013). During transcription, stretches of DNA (gene) known as the transcription unit are used as a template for the synthesis of RNA molecules (transcript). In eukaryote cells, depending on the type of gene being targeted, these reactions can be catalysed by three different enzymes, RNA polymerases I, II and III. Of the three RNA

polymerases, RNA polymerase II is responsible for the synthesis of RNAs from protein-coding genes. RNA polymerase I and III are involved in the transcription of other types of RNA: transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and various small RNAs (Paule and White, 2000).

Transcription by RNA polymerase II is a multi-step process that begins with the binding of several proteins to a promoter, which is a regulatory region located upstream of the gene (Fuda et al., 2009). These proteins facilitate the assembly of the polymerase and the formation of the transcription initiation complex. They are termed general transcription factors (TFs) due to their participation in recognition of most promoters. More specific TFs also exist, which are able to modulate the fate of transcription by binding to promoters as well as DNA regions that promote (enhancers) or inhibit (silencers) polymerase assembly, thus contributing to the regulation of gene expression levels (Vaquerizas et al., 2009). RNA polymerase II is then released from the large complex of proteins after conformational rearrangements, and moves along the DNA from the promoter region, with transcription entering the elongation phase (Kwak and Lis, 2013). The transition to this stage is not instantaneous, and in most cases, the polymerase stays at the promoter, producing short truncated transcripts which are known as abortive initiation. RNA is synthesised from the transcription start site (TSS) during elongation, and nucleotides complementary to the DNA strand are incorporated in the 5' to 3' direction. Finally, the polymerase transcribes through the cleavage and polyadenylation (poly-A) signals that indicate the end of the gene and is released from the DNA template (Kuehner et al., 2011).

### **1.3.2 Transcription and mRNA processing**

Before they are exported to the cytosol, mRNA molecules will undergo several modifications, which comprise 5' capping, the polyadenylation of the 3' end and the removal of non-coding intervening sequences (introns) through splicing (Darnell, 2013). A methylated guanine nucleotide ('cap') is added to the 5' end through an enzymatic reaction immediately after the RNA polymerase II has entered the elongation phase (Ramanathan et al., 2016). The cap not only protects the 5' end of

transcripts but also facilitates the differentiation of mRNAs from other RNA species such as the uncapped RNAs produced by RNA polymerase I and III (de Klerk and 't Hoen, 2015). Immediately after termination of transcription, the 3' end of the mRNAs is also modified through the addition of approximately 200-250 nucleotides (in mammalian cells), known as the poly-Adenosine (poly-A) tail (Jalkanen et al., 2014). RNA polyadenylation serves the purpose of extending the half-life of mRNA, and it has been exploited for research, where one of the most common RNA extraction protocols depends on the explicit selection of poly-A-tailed RNA species (Elkon et al., 2013). Splicing is a more complicated reaction where intronic regions of the pre-mRNA are removed, and the stretches of sequence that contain the essential information (exons) for protein synthesis are merged (Kornblihtt et al., 2013). The mechanism of splicing is discussed in more detail in section 1.4. Eventually, a mature mRNA product is produced.

### **1.3.3 Life of mRNA in cell**

When a mature mRNA molecule is processed in the nucleus, it undergoes selective export via the nuclear pore through multiple steps. mRNA export is subjected to strict quality control mechanisms or surveillance pathways that ensure fidelity and quality products where immature RNAs remain in the nucleus (Denti et al., 2013). These mechanisms depend on the identification of protein complexes that accompany the RNA molecules such as RNA binding proteins (RBPs), which act as markers for the completion status of the processing steps mentioned in section 1.3.2. For instance, the cap-binding and poly-A binding complexes are indicators of successful capping and polyadenylation reactions, respectively, while protein complexes (exon-junction complex - EJC) mark the completion of splicing similarly. The mRNA molecule is marked as immature because of the binding of RBPs involved in carrying out each of these steps.

Unprocessed mRNAs, together with the residues from the transcription and splicing reactions, will be degraded by the exosome, a multi-protein complex, which possesses ribonucleolytic activity (Pérez-Ortín et al., 2013). More quality control mechanisms exist to prevent translation where mRNAs are incorrectly exported or

when intact mRNAs are damaged in the cytosol. These mechanisms, for the most part, are inherent to the steps needed for the beginning of protein synthesis, for example, recognition of the 5' cap and poly-A tail by the translation initiation machinery. A separate system for surveillance known as nonsense-mediated decay (NMD) actively searches for abnormal mRNAs for degradation, in advance of translation. NMD targets mRNAs with premature stop codons, which could be a result of inaccuracies in the splicing reaction or genetic mutations (Pérez-Ortín et al., 2013). When NMD occurs, translation begins immediately after the 5' end of the mRNA emerges from the nuclear pore, when the EJC that encloses each splice-site would normally be removed from the mRNA. The mRNA stays bound to these complexes and is rapidly degraded (Alberts et al., 2002)

After the export process, mRNAs are localised to distinct regions within the cytosol based on the signals encoded in the 3' UTR regions, finally recognised by ribosomes and translated (Alberts et al., 2002). The binding process of ribosomes to the mRNAs competes with mRNA degradation, a process that starts immediately after transcripts are exported into the cytosol and involve the gradual shortening of the poly-A tail. mRNAs are eventually degraded when the poly-A tail reaches a critical length in the course of continued digestion from the 3' end or through the decapping process which is the removal of the 5' cap and subsequent 5' to 3' decay (Schoenberg and Maquat, 2012). Alternatively, a process of cytosolic polyadenylation can also happen, therefore having a positive impact on the mRNA half-life (Villalba et al., 2011). In general, these processes regulate mRNA stability and translation efficiency.

## 1.4 The splicing reaction

The splicing process was first discovered by Phillip Sharp and Richard J. Roberts in 1977 (Gelinis and Roberts, 1977), who detected, independently, that in the DNA, coding sequences were discontinuous and included intervening non-coding segments. In their studies, adenoviral mRNAs were hybridised with complementary single-stranded DNA fragments, and after careful observation using electron microscopy (EM), it was discovered that there were alternate double-stranded and single-stranded stretches in the hybrid that resulted. The conclusion from this was that, while the

initial RNA transcript is maturing, some regions of it are removed; this now brings separate parts of the mRNA together (Lodish et al., 2000). Splicing predominates in eukaryotes, but the splicing process has also been observed in all forms of life. Although splicing has been detected in prokaryotes, they do not have the major eukaryotic pathway to achieve this process (Alberts et al., 2002).

#### 1.4.1 Major mechanisms in mRNA splicing

Splicing is carried out in eukaryotes via the spliceosomal pathway, by which intron removal is orchestrated by a large complex of proteins and RNAs. This large complex is known as the spliceosome, and it has been defined as one of the most complicated types of machinery in the cell (Hoskins and Moore, 2012). Two types of spliceosome have been recognised, the major and minor spliceosome, which vary in their components and the properties of the introns they remove. The major spliceosome is involved in almost all the splicing events and is responsible for the removal of the introns that harbour specific signals indicating intron-exon boundaries (consensus splice site sequences) (**Figure 1.2a**) (Matera and Wang, 2014). On the other hand, a small set of introns which are different from the consensus introns is targeted by the minor spliceosome (non-canonical splicing) (Irimia and Roy, 2014).

Besides spliceosomal introns, another class of introns that undergo splicing in a protein-independent manner has also been discovered. These are self-splicing introns which are able to mediate the splicing reaction by acting as ribozymes through the RNA structure (Alberts et al., 2002). Even though these reactions are not usually seen in eukaryotes, they are involved in the splicing of certain rRNAs and organelle genes. Therefore showing that mRNAs are not the only RNA molecules that undergo splicing, and even non-coding RNAs like micro-RNAs, tRNAs and long non-coding RNAs can also undergo splicing (Cech and Steitz, 2014; Kelemen et al., 2013). Their presence is widely believed to support the RNA world hypothesis, which postulates that the initial building blocks of life were self-replicating RNAs (Robertson and Joyce, 2012).

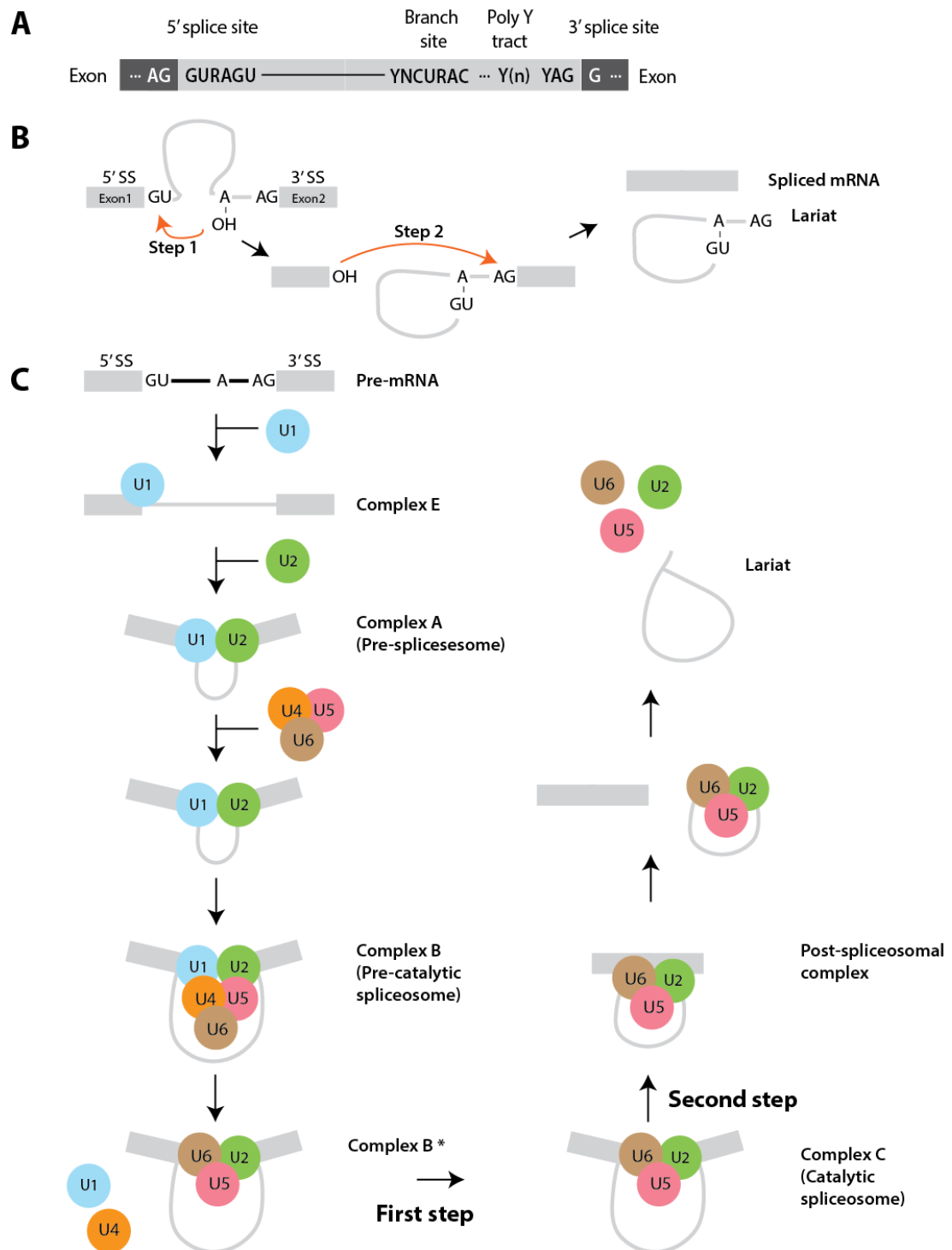
### 1.4.2 The major spliceosome splicing

The major spliceosome is composed of five different small nuclear RNA molecules (snRNAs) which are U1, U2, U4, U5 and U6, and several hundred proteins. Each of the snRNAs associated with a number of proteins to form complexes known as small nuclear ribonucleoproteins (snRNP) (Will and Lührmann, 2011). These snRNPs form the main components of the spliceosome and are involved in the detection of splice sites and branch-point sequences (**Figure 1.3**) within the introns, in addition to their role in catalysing the splicing reaction. One of the most prominent features of the spliceosome is the enzymatic reaction of intron removal from an RNA molecule (Kosmyna and Query, 2016). Splicing is the outcome of two transesterification reactions, (**Figure 1.3B**) (Will and Lührmann, 2011). In the first reaction, the 5' exon is cleaved from the intron through a nucleophilic attack of the 2'-OH group of the branch-site residue (located in the phosphate group of the GU dinucleotide at the 5' splice site), thus forming a branched intron known as a lariat. Meanwhile, in the second reaction, the two exons are joined together (ligated) therefore releasing the lariat.

The active catalytic site of the spliceosome needs to be formed before intron removal can take place. This requires numerous changes in its composition and conformation (**Figure 1.3C**) (Matera and Wang, 2014). Splice-site recognition is the first step of the splicing reaction. Base pairing between the U1 snRNP with the 5' splice site results in the formation of complex E (**Figure 1.3C**). The branch point and 3' splice site nucleotide sequences are recognised by the U2 snRNP, which then interacts with the U1 snRNP and forms complex A (pre-spliceosome), therefore bringing together both splice sites. U4, U5 and U6 snRNPs pre-assembled to form a complex called the U4/U6.U5 tri-snRNP which joins complex A to form complex B. Then complex B becomes activated due to a number of conformational changes and loss of U1 and U4 from the complex. The activated complex B initiates the first catalytic core step in the splicing reaction and directs the formation of complex C which includes the free 5' exon and lariat intermediate. Following further rearrangements, complex C executes the second catalytic step and creates a post-spliceosomal complex which consists of the spliced exons and the lariat. The lariat is finally released together with



the remaining snRNPs, which will be used for further cycles of splicing. The binding of the freshly created exon junction to a new complex of proteins (EJC) occurs after the previous steps; this now marks the effective conclusion of splicing at that particular location. This further aid in determining the fate of the mRNA molecule (Section 1.3.3).



**Figure 1.3: Splicing by the major spliceosome. (A) The major spliceosome recognises core splicing signals.** It recognises several core signals in the pre-mRNA transcript, which are the 5' splice sites and 3' splice sites, the branch point sequence (normally located between 15 and 50 nucleotides upstream of the 3' intron) and polypyrimidine tract. R represents a purine (A or G), Y represents a pyrimidine (U or C), and N represents to any nucleotide. Introns that harbour the consensus sequences

are referred to as U2-type introns because they are identified by the U2 snRNP. **(B) Steps in the splicing reaction process.** Splicing is the consequence of two transesterification reactions that engage the nucleotides from the pre-mRNA and snRNA molecules. **(C) Spliceosomal rearrangements during the splicing process.** As transcription progresses, some parts of the spliceosome are transferred from the polymerase tail to the nascent pre-mRNA, thus assisting the process of splice site recognition. Then, the spliceosome undergoes several compositional and conformational changes that direct the creation of the catalytic site, the cleavage of the intron and the final release of the splicing products. Figure redrawn from Will and Lührmann (2011).

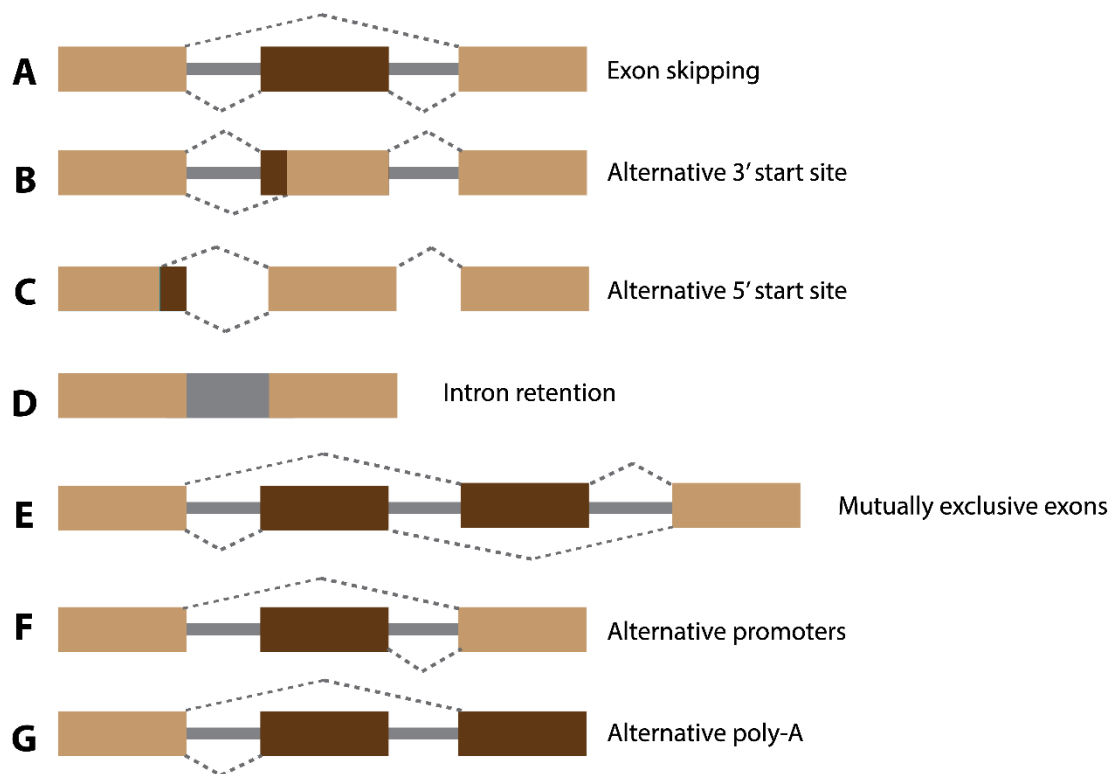
### 1.4.3 Diversification of genes through alternative splicing

Sometimes exons may be excluded from the final mRNA product during the splicing process, and likewise, some introns may fail to be removed. This results in the formation of mature alternative mRNA products from a particular genomic locus (**Figure 1.4**). This process is identified as alternative splicing (AS) event, and it is an important diversification mechanism of the message encoded in a gene (Kornblihtt et al., 2013). Approximately 85% of genes have been detected that produce more than one mature mRNA transcript in humans, and much scrutiny has been given to the relatively low number of genes in mammals. However, in comparison to lower eukaryotes such as *Caenorhabditis elegans*, mammals contain a considerably higher number of genes (Pan et al., 2008; Wang et al., 2008). Alternative splicing can result in the production of proteins with diverse biological function, structure, localisation and interaction capabilities. This is due to the possible differences in biological function between the distinct alternative transcripts (i.e. transmembrane domain, a calcium-binding domain, a protein binding domain, a targeting domain etc.) (Kalsotra and Cooper, 2011; Kelemen et al., 2013). Detection of alternative splicing products at the protein level validates the potential of the splicing process in increasing cellular protein diversity (Tress et al., 2008). It has also been proposed that a considerable amount of the detected alternative splicing products result merely from noise in the splicing process and has no function (Melamud and Moulton, 2009).

AS events can be categorised into four major subgroups. The first major type which accounts for nearly 40% of AS events in higher eukaryotes (Alekseyenko et al.,

2007; Keren et al., 2010) but is exceptionally rare in lower eukaryotes, is exon skipping or skipped exon (SE). It involves a type of exon, also known as a cassette exon, which is spliced out of the transcript together with its flanking introns (**Figure 1.4A**). The second and third types are alternative 3' splice site (3' SS) and 5' SS selection which account for 18.4% and 7.9% of all AS events in higher eukaryotes, respectively (**Figure 1.4B and C**). These types of AS events happen when more than two splice sites are detected at one end of an exon. The fourth type is the rarest AS event in both vertebrates and invertebrates, accounting for less than 5% of known events (Burgess, 2014; Kornblihtt et al., 2013), and is termed intron retention (**Figure 1.4D**). This type of AS event occurs when an intron remains in the mature mRNA transcript. In contrast to animals, intron retention is the most common AS events in plants, fungi, and protozoa (Syed et al., 2012; Xiong et al., 2012; Zhao et al., 2013).

Another three AS events that produce alternative transcript variants that occur infrequently include mutually exclusive exons (MXE) (**Figure 1.4E**) (Pohl et al., 2013), alternative promoter usage (**Figure 1.4F**) and alternative polyadenylation (APA) (**Figure 1.4G**) (Tian and Manley, 2013). Splicing of exons distinguishes MXE in an organised way where two or more splicing events are not independent. The name of “mutually exclusive” indicates that one out of two exons is retained, while the other one is spliced out (Pohl et al., 2013). The use of alternative promoters is a widespread phenomenon in humans and enables diversified transcriptional regulation of a single gene. This AS event serves as a molecular foundation for the complexity of systems in humans (Batut et al., 2013) where more than half of human genes are regulated by alternative promoters (Cooper et al., 2006; Kimura et al., 2006). APA is a well-known phenomenon to control gene expression, producing mRNAs with alternative 3' ends. This event occurs at the 3' end of most protein-coding genes and long non-coding RNAs. A number of studies have shown that a large proportion of these genes have more than one polyadenylation site (Elkon et al., 2013).



**Figure 1.4: Types of alternative splicing.** Variations in the splicing reaction can lead to message diversification (**A**) exon skipping, (**B**) alternative 3' splice sites, (**C**) alternative 5' splice sites, (**D**) intron retention and (**E**) mutually exclusive exons (MXE) (**F**) alternative promoters and (**G**) alternative polyadenylation. Constitutive exons are shown in light brown and alternatively spliced regions in dark brown. Introns are represented by solid grey lines and dashed lines indicate splicing options. Figure redrawn from Keren et al. (2010).

The biological importance of alternative splicing becomes clear with evidence of tissue-specific events (Chen et al., 2012; Naftelberg et al., 2015), and the role of AS in dynamic processes; proliferation (Chen et al., 2012), development (Kalsotra and Cooper, 2011), and differentiation (Pimentel et al., 2014). Therefore, the functional characterization and annotation of alternative transcript mRNA products are vital. The GENCODE project (Harrow et al., 2012) aims to annotate all evidence-based features in the human genome. One of the GENCODE gene sets, the so-called GENCODE Comprehensive, is rich in alternative splicing, novel CDSs, novel exons and high genomic coverage information (Frankish et al., 2015).

However, protein diversification mechanisms are not limited to alternative splicing. For instance, a process called trans-splicing allows exons from different genes to be merged during the splicing reaction (Kornblihtt et al., 2013). Similarly, RNA editing comprises a distinct mechanism where the information within an mRNA can be modified through the insertion, deletion or substitution of specific nucleotides (Peng et al., 2012). Lastly, post-translational modification processes also contribute to protein diversity. Inteins are segments of a protein that is able to excise themselves and join the remaining portions (the exteins) with a peptide bond in a process termed protein splicing (Bah and Forman-Kay, 2016; Shah and Muir, 2014).

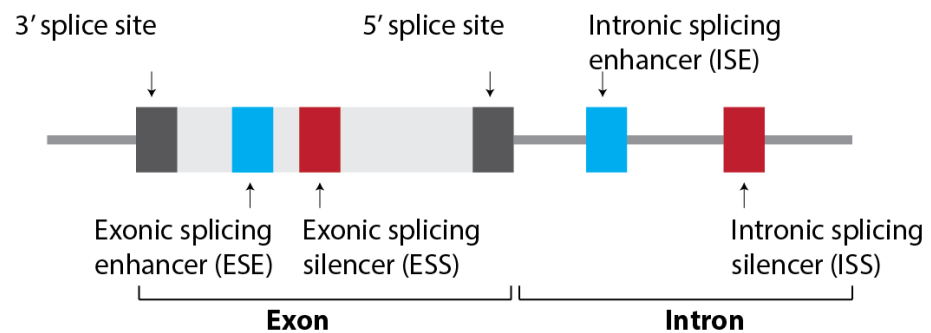
#### 1.4.4 Splicing regulation

Regulation of the alternative splicing process requires a core splicing mechanism that binds to splicing signals around the exon-intron junctions (splice sites). The core splicing signals comprise the canonical 5- and 3-splice site (located at the 5 and 3-ends of the intron, respectively), polypyrimidine tract (PPT) (located upstream of the 3-splice site), and the branch site (located upstream of the PPT) (Wang and Burge, 2008). Besides the core splicing signals, more elements are added to the regulation of splicing and defining exon-intron boundaries. Other cis-regulatory sequences, such as splicing regulatory elements (SREs), which can differ regarding location and effect, are normally present in the pre-mRNA (**Figure 1.5A**) (Matera and Wang, 2014). In most cases, SREs contribute to the recruitment of a set of proteins such as trans-acting splicing factors (SFs) that can function as repressors or activators of splicing, usually by manipulating spliceosome assembly.

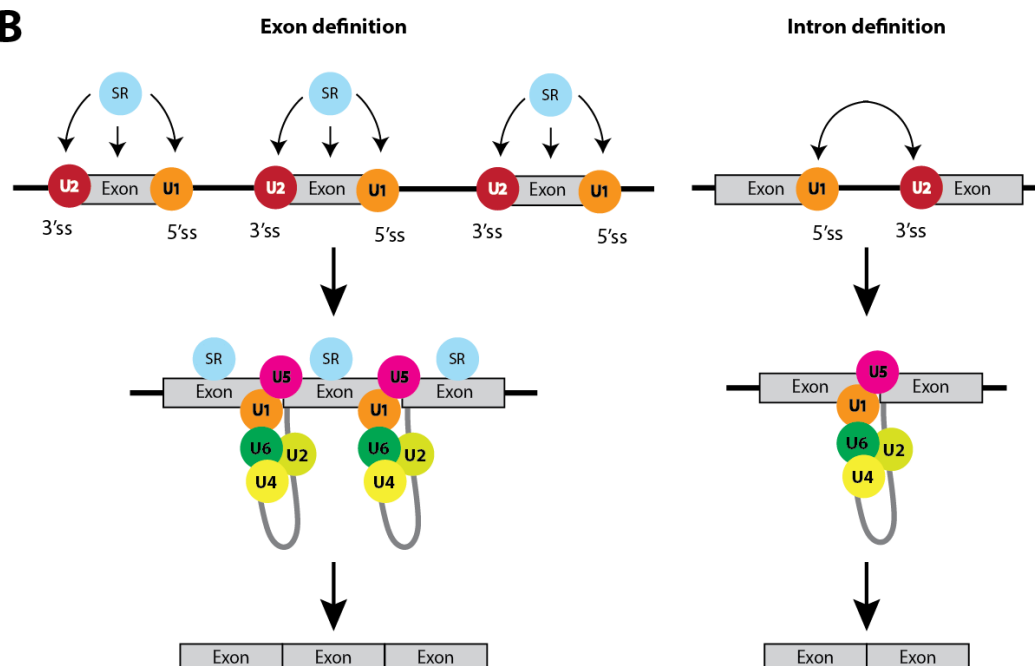
A process known as exon definition is one of the well-known roles of SREs because of their involvement in recognition of exon-intron junctions (**Figure 1.5B**) (De Conti et al., 2013). In higher eukaryotes, intron size commonly exceeds that of exons (Cooper, 2000). This could lead to splicing errors due to the existence of cryptic splice sites. Therefore, a class of SFs called Serine-Rich proteins (SR proteins) binds to exonic splicing enhancers (ESE) to promote the binding of snRNPs to the splice sites located at both ends of the same exon. Consequently, a ‘cross-exon’ recognition complex is formed which will ultimately lead to intron-spanning interactions via

spliceosomal rearrangements. On the other hand, for short introns and in lower eukaryotes, intron definition is the most common mechanism for splice site recognition (**Figure 1.5B**) (De Conti et al., 2013). In this situation, splice sites located on both ends of the same intron are directly identified without the help of specific SFs.

**A**



**B**



**Figure 1.5: Regulation of splicing regulators. (A) Regulation of intron removal process by cis-acting sequences.** Supplementary to the core splice signals such as 5' splice site, branch point, and 3' splice site, a number of regulatory sequences influence the splicing decision through trans-acting splicing factors (SFs). Typical SFs include SR proteins and hnRNPs, which promote and inhibit splicing, respectively. There is a high dependence on SRE activity: even if the same SF is

recruited, opposite functions can be produced from the same sequence, depending on the location, whether within exonic or intronic boundaries. Figure redrawn from Matera and Wang, (2014). **(B) Exon and intron definition.** In lower eukaryotes, intron definition is the common-mode of splice site recognition, with both 5' and 3' splice sites paired together at the ends of the intron. The longer intron size in higher eukaryotes could lead to the use of cryptic splice sites, so exon definition is used. In this instance, the identification of the splice sites that enclose such exons is mediated by SR proteins, to promote the formation of a 'cross-exon' recognition complex. Eventually, further spliceosomal rearrangements ensure intron-spanning interactions take place. Figure redrawn from Ast (2004).

A lot of these splicing events happen before transcription termination, in humans, resulting in a phenomenon which is called co-transcriptional splicing (Merkhofer et al., 2014). It has been suggested that the rate of transcription elongation can have an effect on the choice of splice site to be used; a fast elongation will afford an opportunity window for recognising strong splice sites. However, a slow elongation will afford an opportunity window for recognising splice sites that are weak (Kornblihtt et al., 2013; Naftelberg et al., 2015). Since instability in the concentration of core components of the spliceosome is known to influence the outcome of splicing, the regulation of splicing decisions will ultimately not be limited to the role played by the specific SFs (Karamysheva et al., 2015).

From all the processes mentioned above, it can be concluded that there is a likelihood that splicing would occur in a precise through flexible fashion (Djebali et al., 2012). Its precision is increased further by the numerous readjustments that are necessary before the reaction involving the removal of the actual intron can occur. The nonsense-mediated decay pathway helps to avoid errors in splicing. Conversely, the modification or accumulation of splice site mutations in the spliceosomal components can lead to serious phenotypic results and dysregulation of splicing has been linked to a number of diseases such as cancer (Ladomery, 2013; Padgett, 2012).



## 1.5 Transcriptomic studies using RNA sequencing

The work described in this section examines the use of the next generation sequencing technologies (described briefly in Section 1.2) to understand the structure, processing, and expression of RNA molecules and how this has increased our knowledge of how information is transferred from genomic DNA in the nucleus to the cell and its environment. NGS technologies were initially developed for genomic sequencing but have now been modified to sequence RNA molecules, providing information about both the sequence and the relative quantity of different RNAs.

RNA sequencing (RNA-seq) is the application of a variety of NGS methods which are also known as deep sequencing technologies due to the potential for high coverage of sequence to study RNA (Chu and Corey, 2012). It is also an approach to transcriptome profiling that uses next-generation sequencing technologies. Studies using this approach have already changed our view of the level and complexity of eukaryotic transcriptomes. RNA-seq also provides a far more accurate measurement of levels of transcripts and their isoforms than the other methods (Wang et al., 2009). It is also an approach to reveal the existence and quantity of RNA in a biological sample from different types and time points (Chu and Corey, 2012).

The popularity of RNA-seq for the study of transcriptomes has been increasing over the last decade (Mortazavi et al., 2008a; Wang et al., 2009). RNA-seq provides a much higher dynamic range than other approaches to studying gene expression patterns, and facilitates a much larger set of analyses, in contrast to microarrays, which have been the primary technology for the high-throughput comparison of transcriptome-wide expression levels across samples and conditions (Malone and Oliver, 2011). Other applications of RNA-seq, in addition to gene expression analysis, include the identification of novel transcribed regions (Wang et al., 2016), detection of fusion transcripts (Maher et al., 2009; Supper et al., 2013). Furthermore, detection of allele-specific expression (Berger et al., 2010), estimation of expression levels of different transcripts from the same gene (Trapnell et al., 2012) and study of differential splicing across conditions.

Besides that, RNA-seq is also useful in genome annotation and gene model building (Yandell and Ence, 2012), discovery of alternative promoters or polyadenylation (which are potentially involved in regulation), detection of enhancers through bidirectional transcription, finding novel transcribed elements such as micro RNAs and novel genes and looking at allelic imbalance in expression (Eswaran et al., 2013). Even though microarrays can be a cheaper platform, time- and cost-efficient to perform routine differential expression analysis at the gene level (Guo et al., 2013), the additional studies, the amount of data and the falling costs of sequencing explain the increasing popularity of RNA-seq. In this section, the typical steps required to sequence a transcriptome with an Illumina platform to generate RNA-seq data are outlined. Furthermore, a detailed description of the most commonly used methods to study the transcriptome composition of RNA-seq data will be provided.

### **1.5.1 RNA-seq vs microarrays**

RNA-seq methods involve the conversion of transcripts into complementary DNA (cDNA) which is sequenced directly in a massively parallel sequencing reaction (Mortazavi et al., 2008a). The expression levels of genes relative to another condition of interest or absolute levels can be quantified by counting the number of short sequencing reads mapping onto the reference genome (Marguerat and Bähler, 2010; Nagalakshmi et al., 2010). In contrast, the basic principle of microarrays is that labelled samples of transcribed RNAs are hybridised to immobilised complementary DNA probes representing target genes. The relative abundance of each transcript in samples can be assessed by measuring the signal intensity of the two distinct fluorescent dyes (two-colour arrays) or more commonly now by comparing the signal of one dye across arrays (Hegde et al., 2000; Pariset et al., 2009; Ramsay, 1998; Schena et al., 1995). Currently, there are two popular platforms in microarray technologies; Affymetrix and Illumina.

DNA microarrays have been employed as the main technology platform for transcriptome profiling studies since their development approximately 15 years ago. Microarrays are still widely used for whole transcriptome analysis, but currently, RNA-seq is rapidly becoming the favoured method of choice in certain

circumstances as it overcomes some of the inherent limitations of microarrays (Fu et al., 2009; Wang et al., 2009). Microarrays depend on the prerequisite knowledge of the reference transcriptome (to design probes), RNA-seq does not (Raz et al., 2011). There are several advantages of RNA-seq when compared to a microarray with regards to the data: it produces a very low background signal, a higher dynamic range of expression levels and more accurate quantification of transcript abundance (Wang et al., 2009). However, the efficiency of RNA-seq is also associated with various problems including the huge volume of data (big memory footprint), amount of rRNA in the sample, short reads, less base accuracy and variation of read density along the length of the transcript (Fu et al., 2009; Martin and Wang, 2011). Both RNA-seq and microarrays have their strengths and limitations but between them cover most needs for transcriptome research (van Vliet, 2010; Wang et al., 2009).

Estimating gene expression levels using RNA-seq data requires reads counts to be normalised in order to get meaningful expression estimates (Blencowe et al., 2009; Mortazavi et al., 2008a; Wang et al., 2008). There are two major reasons why RNA-seq data requires normalisation: Firstly, longer transcripts produce more reads compared to shorter transcripts during library construction, assuming the same abundance in the sample (Marioni et al., 2008). Secondly, different runs produce varying read depths thereby affecting the number of reads assigned to a given transcript (Marioni et al., 2008a; Mortazavi et al., 2008). To address these issues, normalisation is usually now performed to adjust transcript read counts by the length of the gene and the total number of mapped reads in the sample and expression levels are recorded as a number of reads per kilobase of transcript per million mapped reads (RPKM) metric. The fragment per kilobase of transcript per million mapped reads (FPKM) metric is used for both gene and isoform quantification of paired-end reads data (Trapnell et al., 2010). The read count normalisation is discussed in more detail in section 1.5.8.

### **1.5.2 RNA-seq workflow**

A typical RNA-seq workflow involves three main sections which are experimental biology, computational biology, and systems biology. The experimental section

includes the methods for RNA collection, first strand synthesis, and library construction, producing millions of short read sequences from the NGS sequencer. Various sequencing platforms have been implemented for RNA-seq studies which include Illumina Genome Analyzer GAIIx and HiSeq (Liu et al., 2012; Nagalakshmi et al., 2008), Roche 454 Life Science (Marioni et al., 2008), Applied BioSystems SOLiD (Eid et al., 2009), Ion Torrent Personal Genome Machine (PGM) (Rothberg et al., 2011), single molecule real-time (SMRT) machine PacBio RS System (Eid et al., 2009) and the nanopore technology-driven portable device MinION.

RNA preparation methods differ for different types of sequencing platforms, RNA sub-types, and sequencing purpose. Furthermore, sample quality is a major factor in obtaining good quality data and deriving biological insights from unbiased analyses. Selection of Poly-A mRNA with oligo-dT oligonucleotides has been used in a range of transcriptomic analyses including gene expression, variant detection and alternative splicing (Carrara et al., 2015; Tariq et al., 2011). For both random sequencing and single cell sequencing, the labelled molecules on Illumina sequencing platform can achieve a significant mRNA sequencing efficiency (Carrara et al., 2015).

In Illumina sequencing technology, typically the reaction is based on the use of modified versions of the four bases, which vary from the standard nucleotides because they incorporate a reversible terminator, in addition to a fluorescent dye. Therefore, throughout each sequencing cycle, and following the addition of the necessary reagents, elongation will be blocked after the successful incorporation of a single base and the identity of a nucleotide can be traced by measuring its fluorescent signal. Repetition of this process will lead to a set of images that are converted into a set of sequences or reads after interpretation using a base calling software (Das and Vikalo, 2013). The reads represent the set of molecules in the initial sample, and the length of the reads corresponds to the number of cycles performed throughout the sequencing reaction. Finally, the acquired sequence information is stored in a plain text file in a FASTQ format with the probability of a wrong base call at each position of the read given by metrics such as the Phred score (Cock et al., 2010).

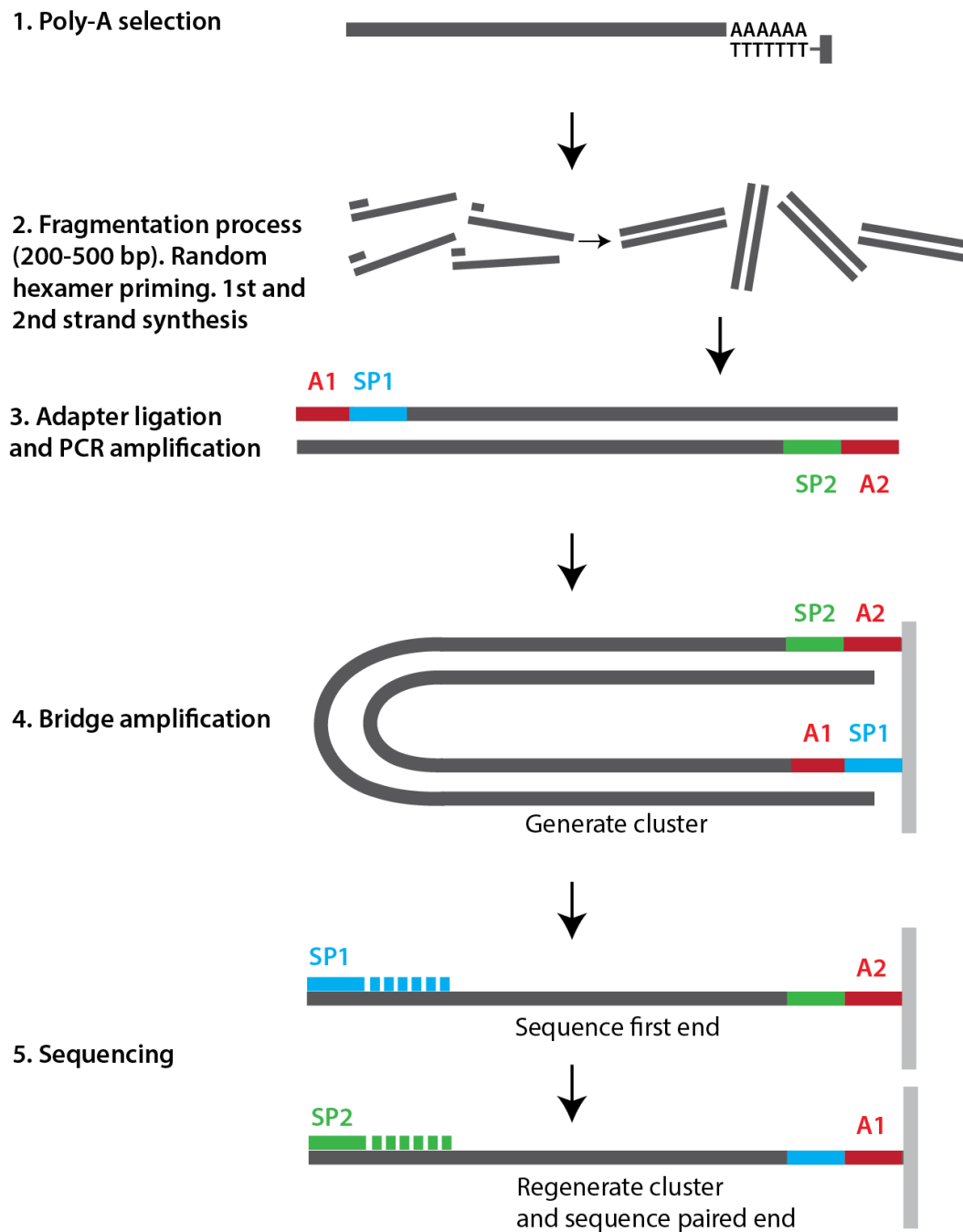
The raw reads are supplied as starting material for the computational biology analysis (Conesa et al., 2016). Initially, biological and technical contaminations are removed by preprocessing steps, followed by mapping qualified reads either to the genome or transcriptome. The mapped reads for each sample are subsequently indexed into three different levels: gene, exon, or transcript-level depending on the experimental purpose, to evaluate the abundance of reads (Martin and Wang, 2011). Then, the summarised data are estimated by statistical models to detect differentially expressed genes and alternative splicing events, or regulatory mechanisms are evaluated through integration analysis with the data set (Han et al., 2015). Lastly, network or pathway analyses are examined to obtain biological insight from the data (Freeman et al., 2007; Theocharidis et al., 2009).

### 1.5.3 RNA-seq using Illumina sequencing technology

The initial step in transcriptome sequencing is library preparation, which consists of extracting RNA from the starting material, converting it into a cDNA library and loading into the NGS sequencing machine (**Figure 1.6**) (van Dijk et al., 2014). Following the RNA extraction step, the RNA type of interest is usually enriched through either ribodepletion or polyadenylated RNA selection. In both cases, the aim is to reduce the concentration of ribosomal RNAs (rRNAs), the most abundant type of RNA in the cell. However, for polyadenylated RNA selection, the use of oligo-dT beads is needed, which facilitates the extraction of polyadenylated RNAs, therefore ensuring a high-quality representation of mRNAs (**Figure 1.6 - step 1**). In contrast, ribodepletion depends on the use of ribonucleases to exclusively digest rRNAs and has the advantage of not limiting the analyses to polyadenylated RNA. In fact, the term “total RNA” is usually used to refer to datasets produced with the rRNA depletion protocol, whereas those acquired with the former method are known as “poly-A-selected”. Poly-A selection has become the most popular selection amongst the presently available RNA-seq datasets because of the easier protocol and its competitive price, with the important exception of those studies aimed at characterising non-polyadenylated RNA types, such many non-coding RNA types which typically lack a poly-A tail. The extracted RNA is fragmented through hydrolysis with divalent cations and will be transcribed into first strand cDNA with

random hexamer primers. Then, there will then be a second strand synthesis (**Figure 1.6 - step 2**). This is followed by the ligation of adapter sequences at both ends of each cDNA fragment (**Figure 1.6 - step 3**). The adapters serve two different roles. Firstly, they facilitate the immobilisation of the cDNA fragments by hybridization to anchored complementary sequences in the flow cell where the sequencing will take place. Secondly, they serve as primers for the sequencing reaction. The resulting cDNA fragments are then size-selected (typically 300-500 bp) through gel electrophoresis to fit within the range required by the next-generation sequencing machine. cDNA fragments outside this range will be ignored; therefore alternative protocols for the study of small RNAs have been created (Zhuang et al., 2012), in which the cDNA library is amplified by the polymerase chain reaction (PCR).

Samples are loaded into a flow cell for transcriptome sequencing when the library preparation procedure has finished (Mardis, 2013). After this step, in order to increase the signal for the sequencing reaction, the starting material needs to be amplified once again through bridge amplification (**Figure 1.6 - step 4**). The process consists of the synthesis of fragments that are complementary to the hybridised cDNA molecules which bind and hybridise with neighbouring adapters (**Figure 1.6, step 4**), therefore facilitating subsequent rounds of synthesis. Consequently, many clusters with identical sequences will be formed, which is now ready for sequencing. Illumina platforms rely on sequencing by synthesis technology to read the base pair composition of each cDNA cluster (**Figure 1.6 - step 5**) (Bentley et al., 2008).



**Figure 1.6: An overview of library preparation and sequencing steps in an Illumina platform.** A representative paired-end workflow is illustrated here, which consists of ligation different adaptors at each end of the initial cDNA molecule. This enables sequencing each cDNA fragment from both ends, in two separate reactions, and has further advantages for the downstream bioinformatics analyses compared to single-end approaches. Figure redrawn from Mardis (2013).

### 1.5.4 Quality assessment of RNA-seq data

It is not a straightforward process to identify and quantify all RNA species from the reads sequenced since RNA-seq is a complex, several step process which involves sample library preparation, fragmentation, purification, amplification, and sequencing (Han et al., 2015). Hence, read quality assessment is the first step of the bioinformatics analysis pipeline of RNA-seq and is a crucial step before downstream analysis (Conesa et al., 2016). It is recommended and always necessary to filter data, removing low-quality sequences, any base contamination, or overrepresented sequences to ensure a coherent final data set (Wang, 2016). A variety of tools is currently available for this purpose. Read quality can be visualised graphically for example through FastQC (Andrews, 2010). Recently, Kraken, a flexible and efficient pre-processing tool was designed to streamline the analysis of next-generation sequencing data. It was developed for demultiplexing, trimming, removing redundancy and filtering short read sequencing data (Davis et al., 2013), while HTSeq was designed to deduce the base calling and evaluate base quality in every position as well as the overall read features (Anders et al., 2015).

### 1.5.5 Read mapping strategy

The next step in an RNA-seq analysis pipeline consists of allocation each sequencing read to a known gene or genomic sequence. The outcome is equivalent to discovering the loci that are expressed in a given sample (Conesa et al., 2016). In principle, two different strategies exist to perform this task where reads can be aligned to the reference genome or transcriptome, if available for the species of interest (Garber et al., 2011); otherwise, they can be directly assembled into contigs to produce contiguously expressed regions with the purpose of reconstructing the set of expressed transcripts (Flicek and Birney, 2009; Trapnell et al., 2012). SOAP (Eid et al., 2009), SOAP2 (Li et al., 2009), MAQ (Eid et al., 2009), Bowtie (Eid et al., 2009), BWA (Li and Durbin, 2009), STAR (Dobin et al., 2013) and Kallisto (Bray et al., 2016) are popular bioinformatics packages that can be implemented in the analyses for this purpose.



A different approach has also been taken, the reads that map to the intron-exon junctions help with the determination of alternative splicing variation models, and in the early days of RNA-seq promoted the development of a new generation of spliced alignment software, including BLAT (Fonseca et al., 2012; Kent, 2002), TopHat (Kim et al., 2013; Trapnell et al., 2009), and MapSplice (Wang et al., 2010).

Two evaluations of RNA-seq data analysis methods have been presented based on the performance of several spliced alignment programs. The evaluations focused on the quality of the alignments (Engström et al., 2013) and the computational methods for transcript reconstruction and quantification from human RNA-seq data (Steijger et al., 2013). The study compared 26 mapping protocols based on 11 programs and pipelines to measure the performance of current mapping software (Engström et al., 2013). The alignment evaluation showed that STAR aligner performed remarkably well based on various factors such as alignment yield, accuracy, nucleotide mismatches, exon junction discovery and suitability of alignments for transcript reconstruction. However, for transcript reconstruction evaluation, this study found that none of the protocols excelled at all metrics.

The strategy of mapping to a reference sequence is much easier than assembling contigs *de novo* and it is generally the method of choice when working with model organisms. Regardless of the strategy used, read mapping is usually the most time-consuming step of the analysis workflow, and to speed up this step, the existing tools use heuristic parameters such as the maximum number of allowed mismatches per read. The possibility of loss of information is high due to a decreased quality at the 3' end of the sequence read because of the difficulty in deducing the fluorescent signal as sequencing cycles build up (Minoche et al., 2011). For this reason, it is usually practical to perform a quality control and pre-filtering step to avoid such reads being discarded where read sequences can be shortened or trimmed based on the quality of the base calls (Trivedi et al., 2014). Similarly, in order to speed up the subsequent mapping process, reads with overall low quality can be removed (Dozmorov et al., 2015; Eid et al., 2009; Mortazavi et al., 2008). Genetic polymorphism is a great challenge for short reads that have a very high copy number and repetitive

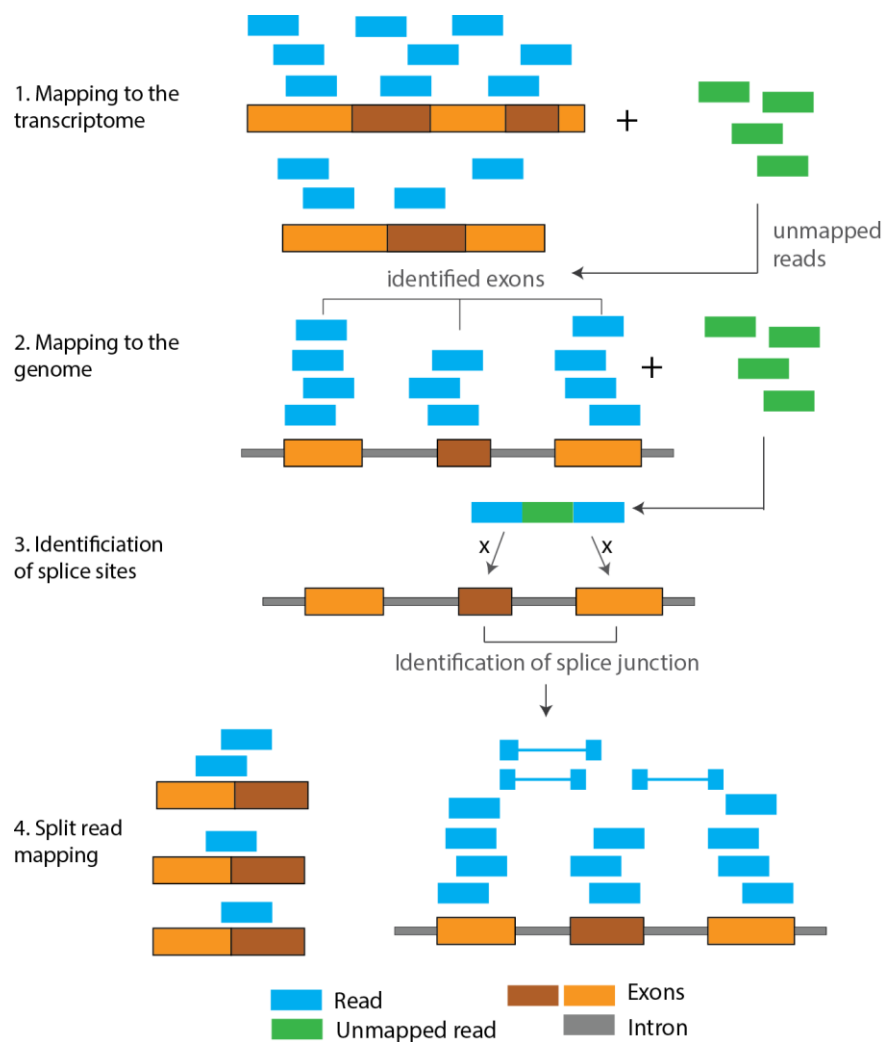
sequences. A longer read sequencer system such as the Roche 454 or PacBio sequence analyser may be required to overcome these challenges (Campbell et al., 2008; Hillier et al., 2008; Holt and Jones, 2008).

#### 1.5.5.1 Alignment to the genome or transcriptome

A general approach when a reference genome exists is to align or map the reads directly to that sequence (Conesa et al., 2016). Likewise, reads can be aligned to the transcriptome if a good annotation exists. Alignment to transcriptome has the advantage because alignment job is simplified due to lack of intronic sequences. This comes at the price of limiting the number of downstream analyses that can be performed. For example, alignment to the transcriptome is not compatible with the identification of novel expressed regions or the study of intronic expression levels. Therefore, a good compromise is the use of hybrid approaches, as implemented in TopHat and Tophat2 (Kim et al., 2013; Trapnell et al., 2009).

TopHat is a read mapping tool specifically designed for RNA-seq data since it facilitates alignment of the reads to the genome while considering the existence of splice junctions (**Figure 1.7**). The core of TopHat is based on Bowtie (Langmead et al., 2009), which is an independent algorithm for the alignment of short reads, and its main strength is the ability to detect exon-exon junctions without the need for any prior knowledge on the annotation. Nevertheless, TopHat will first attempt to map the reads to the derived transcriptome to simplify the search by providing such information. Reads that fail to align to the transcriptome will be then queried against the genome (**Figure 1.7 - step 1**). Reads can also be mapped to the genome directly (**Figure 1.7 - step 2**). Reads that fail to align in this initial stage, and those that map with low alignment scores, are subsequently used to build a database of potential splice junctions, by splitting them into smaller segments and realigning independently (**Figure 1.7 - step 3**). In this situation, every time a read appears to span several exons, a splice junction is reported, for example, when an internal fragment fails to align, or when two consecutive fragments from the same read do not align contiguously on a known genomic locus. Subsequently, the identified splice sites and their flanking sequences are concatenated into a novel transcriptome, which

is used to realign the set of unmapped reads (**Figure 1.7 - step 4**). With paired-end read data, each read is processed separately, and the alignments obtained are assessed in final phase by considering additional sources of information such as the orientation of the reads and fragment length. Eventually, all the information combined during the mapping process is reported in SAM/BAM format (Li et al., 2009).



**Figure 1.7: An overview of the mapping algorithm implemented in TopHat.** When an annotation file is provided, TopHat employs a hybrid approach to discover the genomic loci from which the detected reads could have originated. Otherwise, TopHat can directly align the reads to a reference genome. In both situations, the first step consists of discovering a set of expressed exons and is followed by the detection of splice junctions by using information from reads that span multiple exons. Figure redrawn from Kim et al. (2013).

### 1.5.5.2 *De novo* assembly

Where the species of interest lacks a reference genome it is necessary to perform *de novo* assembly. This approach can also be used where the genome composition of a sample is expected to vary greatly from the reference assembly, such as a cancer sample. The aim here is to assemble the reads into sets of expressed regions (contigs), by relying on their overlap. Nevertheless, the short-read length adds to the non-trivial problems. For example, lowly expressed regions are often difficult to solve even though the use of paired-end data can simplify the process. Trinity, developed by Garber et al. (2011), is one of the most popular software platforms to perform this task.

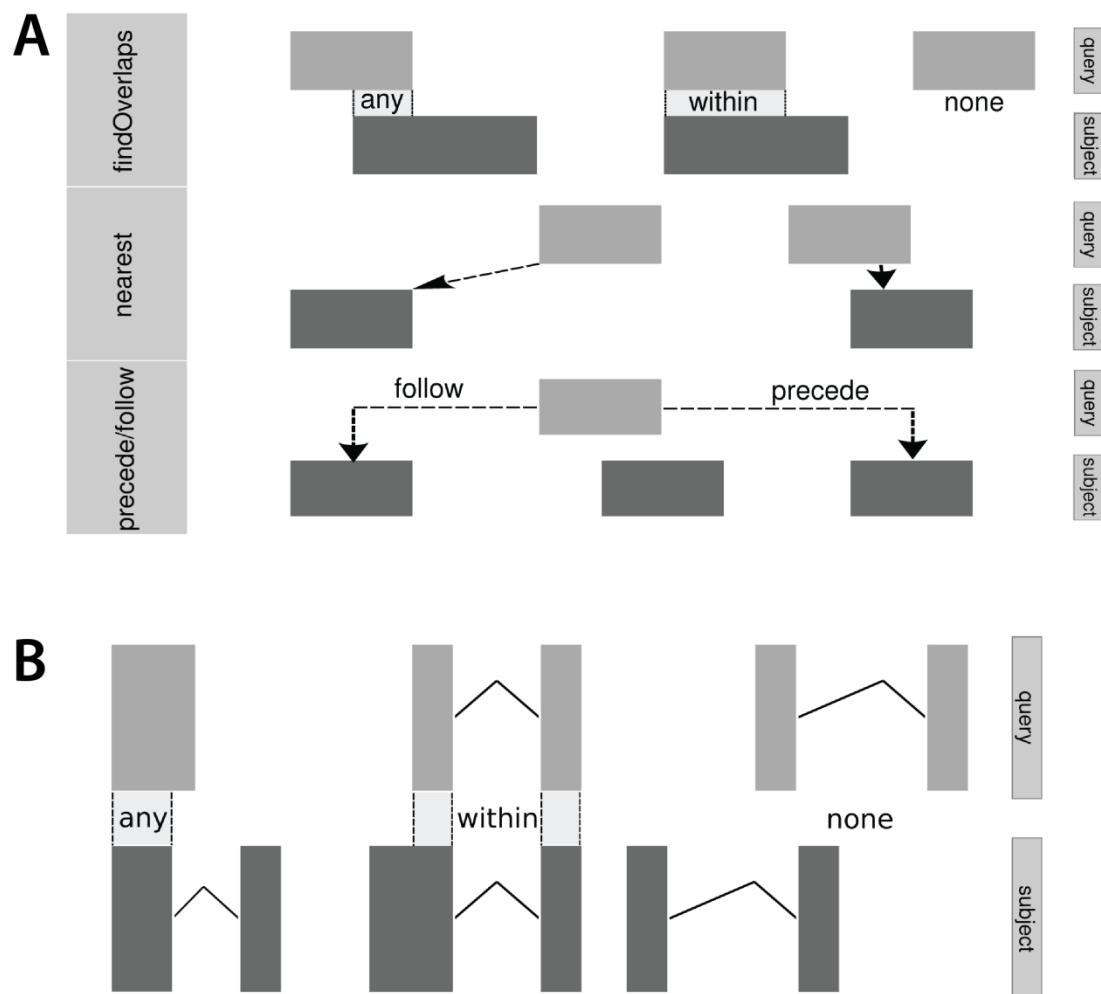
### 1.5.6 The estimation of expression levels

After the reads have been mapped to a specific location in the genome or transcriptome, the next step in the RNA-seq bioinformatics analysis pipeline consists of estimating expression levels for the features of interest, such as genes and transcripts. Like the mapping step, the quantification of expression levels can be accomplished by using existing annotation information, but it can also be performed independently, therefore enabling *de novo* identification of transcribed regions such as novel genes or unannotated transcripts within known gene loci.

Measurement of a gene's expression level means determining the number of RNA-seq reads that map to the gene (Mortazavi et al., 2008). Counting the reads that mapped to the gene based on the reference genome, will facilitate the subsequent steps. A library preparation method such as a strand-specific protocol is a factor for counting of reads. A bedtools (Quinlan and Hall, 2010) is one example of a tool for read counting from a bam file, it takes a feature file (GFF) and count reads in certain regions (e.g. all exons of a gene). It counts reads on both strands within specified regions by a default setting and it can also work in a strand-specific mode if required. HTseq is a specialised tool for counting reads although increasing its speed for read counting is necessary for the future (Anders et al., 2015). In addition, the gene model that hypothesises the structure of transcripts/isoforms expressed by a gene also affects the subsequent analysis. UCSC (Speir et al., 2016), Ensembl (Yates et al.,

2016) and RefSeq (Pruitt et al., 2014) are the most popular annotation databases for genome annotation. The selection of genome annotation databases directly affects gene expression estimation. One study (Zhao and Zhang, 2015) showed that the different definitions of gene models result in a discrepancy in gene quantification.

Furthermore, there is several R packages can be used to count reads. This includes a BioConductor package for the integrative statistical analysis of range-based genomic data (Lawrence et al., 2013). The main features include scalable data structures for annotated genomic ranges and genome-length vectors, and efficient algorithms for overlap detection, especially in RNA-seq data. The tools include the *IRanges*, *GenomicRanges*, and *GenomicFeatures* which are the core of this package. The *IRanges* class encodes only the start and end of ranges but not the chromosome, strand or other information that is critical in genomic applications. Whilst *GenomicRanges* class can support many of the same range operations as *IRanges* and concentrates them for genomic data. One of the important features in this *GenomicRanges* which provides special consideration to the chromosome as well as the strand is *findOverlaps* (**Figure 1.8**). The *findOverlaps* feature is specifically able to take advantage of the chromosome information when detecting overlaps (Lawrence et al., 2013). All the information retrieved from the genomic data is stored in a database. The database is called *TranscriptDb* class and the information stored includes the range of each exon, the coding range, the transcript ID, the gene ID, and metadata about the source of the transcript information. All this information is very important when creating a layout file for network visualising of a gene. The integration of this package will be described in Chapter 2.



**Figure 1.8: GenomicRanges overview. (A) Illustration of overlap (top) and adjacency (bottom) relationships.** The *any* mode detects hits with partial or complete overlap, while *within* requires that the query range represents a subregion of the subject range. **(B) Overlap computations between two *GRangesList* objects.** Each set of rectangles linked by solid lines represents a compound range, i.e., an element of the list. Ranges in the query (top) are being matched against ranges in the subject (bottom). The labels between them indicate the type of overlap (*any*, *within*, *none*). Figure redraw from Lawrence et al. (2013).

### 1.5.6.1 Gene expression levels

Estimation of read abundance can be determined by counting the total number of reads overlapping each locus at the gene level together with a complete annotation. Count-based tools such as DESeq2 (Love et al., 2014) and DEXSeq (Anders et al., 2012), comprise the starting point for most of the downstream analysis algorithms and can be performed with the tool htseq-count (Anders et al., 2015). There are some

challenges that need to be considered despite this obvious simplicity. The first challenge is the multi-mappers, and the second is repetitive or duplicated loci, both of which circumstances need to be handled with care to avoid over-estimated expression levels. Generally, approaches included uniformly distributing reads to all mapped positions (Trapnell et al., 2010), or probabilistically assigning them based on the coverage at each mapping locus (Trapnell et al. 2010; Turro et al. 2011), as first proposed by Mortazavi et al. (2008).

#### **1.5.6.2 Transcript expression levels**

Since many reads will overlap with exons that are shared across multiple isoforms of the same gene, the task of estimating expression levels becomes more complicated when dealing with individual transcripts. In general, the currently available algorithms depend on a variety of sources of information to statistically estimate transcript expression levels which are most relevant to reads mapping uniquely to one of the annotated transcripts within the chromosome region. Furthermore, reads that span two different exons become informative. For instance, splice junctions that involve skipped exons, tend to give unambiguous support for their skipping or inclusion. At this stage, the information from paired-end reads become most applicable where sequencing both ends of the cDNA fragment facilitates covering larger genomic regions, therefore increasing the chance that a given read pair is mapped across different exons i.e. spliced reads.

#### **1.5.7 Read count normalisation**

Normalisation is an important step in the analysis of RNA-seq data which has a strong impact on the detection of differentially expressed genes (Dillies et al., 2013). The most common measurement for expression level derived RNA-seq data is the Reads/Fragments per Kilobase per Million mapped reads (RPKMs or FPKMs, in the case of single-end or paired-end data, respectively) (Mortazavi et al., 2008). Data normalisation is one of the most important steps of data processing after getting the read counts. There are various aspects of the RNA-seq data that need to be considered including sequencing depth, transcript size, sequencing error rate, GC-content and insert size (Filloux et al., 2014; Li et al., 2014).

### 1.5.8 Differential alternative splicing

Several tools have been developed to identify and classify major alternative splicing events such as alternative 5' splice sites, mutually exclusive exons and skipped exons (see **Figure 1.4**). There are two methods for quantification of alternative splicing using RNA-seq data: count-based models and isoform resolution models, using rMATS (Shen et al., 2014) and Cuffdiff 2 (Trapnell et al., 2013) respectively. rMATS was developed to detect differential, alternative splicing events from datasets. Correction for multiple sample comparisons is vital because of the huge number of genes in an RNA-seq dataset. Therefore, the false discovery rate (FDR) (Benjamini and Hochberg, 1995) offers an attractive measure of control for multiple testing. It involves a statistical model that calculates the p-value and FDR for the differences in the isoform ratio of a gene between two conditions (Shen et al., 2012). rMATS uses read counts of RNA-seq data mapped to an exon junction and its two flanking exons, to measure the exon inclusion levels in two samples (percentage spliced inclusion [ $\Psi$ ]). Then, it compares values between samples to provide the probability of differential splicing expressed in  $\Psi$ . It classifies different types of alternative splicing (for example skipped exons and mutually exclusive exons; refer to **Figure 1.4**) and generates both p-value and magnitude ( $\Delta\Psi$ ) for each alternative spliced form in the result (Chen et al., 2013). Cuffdiff 2 measures expression using a negative binomial model for fragment counts at transcript level resolution, controlling for variability and read mapping ambiguity. It determines differentially expressed transcripts and genes and reports differential splicing and promoter changes (Trapnell et al., 2013).

### 1.5.9 Challenges in RNA-seq

Despite its many advantages, RNA-seq still provides challenges as it is not an established technology like expression microarrays. For instance, the PCR amplification step has been shown to lead to differential amplification of fragments with lower or higher GC content (Benjamini and Speed, 2012). Incorrect base calls can be made due to two situations: failure to block the elongation reaction and failure to remove the fluorescent dye during the sequencing step, which would apply to DNA sequencing as well (Metzker, 2010).



In most cases, to overcome these problems, alternative protocols or analyses have been introduced. For instance, several algorithms such as Cufflinks have been developed to correct for the potential biases from the random hexamer amplification step (Trapnell et al., 2010). Other library preparation methods have also been suggested to account for PCR bias, whereby random barcodes are used as molecular identifiers to quantify the absolute number of molecules (Shiroguchi et al., 2012). Some downstream analysis algorithms also include information on the probability of a wrong base call at specific positions of the read, as reported by the Phred score, for example as reported in Del Fabbro et al., (2013). A very widespread strategy to overcome limitations on the read length and try to span larger regions consists of sequencing each cDNA fragment from both ends (paired-end sequencing; Figure 1.6 and Section 1.5.3), as opposed to the single-end strategy, and can be accomplished through modified adapters (Mardis, 2013).

Alternative library preparation strategies can add more information to the experiment, especially for strand-specific protocols, which are able to provide strand information for each read (Levin et al., 2010). Likewise, multiplexing has become a broadly used approach to optimise the amount of data that can be acquired from each sequencing run, by allowing pooling of several different samples into a single lane of the flow cell using a range of sequence identifier bar codes (Wong et al., 2013).

## 1.6 Visualisation

This section reviews current knowledge of the visualisation platforms and their advantages and disadvantages. Data visualisation is one of the major challenges in the analysis of large biological datasets, particularly when dealing with large organised structures with diverse clusters (Rubel et al., 2010). This is vital when analysing 3-dimensional (3D) data sets. Typically, the first step in interpreting the data is to visualise a biological process or feature. Then, downstream process analysis can be continued when data is visualised and the quality of the data is assessed. Often the second step is to cluster the observations into different sub-clusters based on gene-gene interactions. The output from the clustering process requires visualisation. A number of 3D visualisation tools have been developed

where the software can be locally installed such as BioLayout *Express*<sup>3D</sup> (Theocharidis et al., 2009) or Cytoscape (Shannon et al., 2003). Furthermore, only machines with recent graphics card will be able to perform such analysis in a 3D environment.

### 1.6.1 Visualisation of co-expression gene network

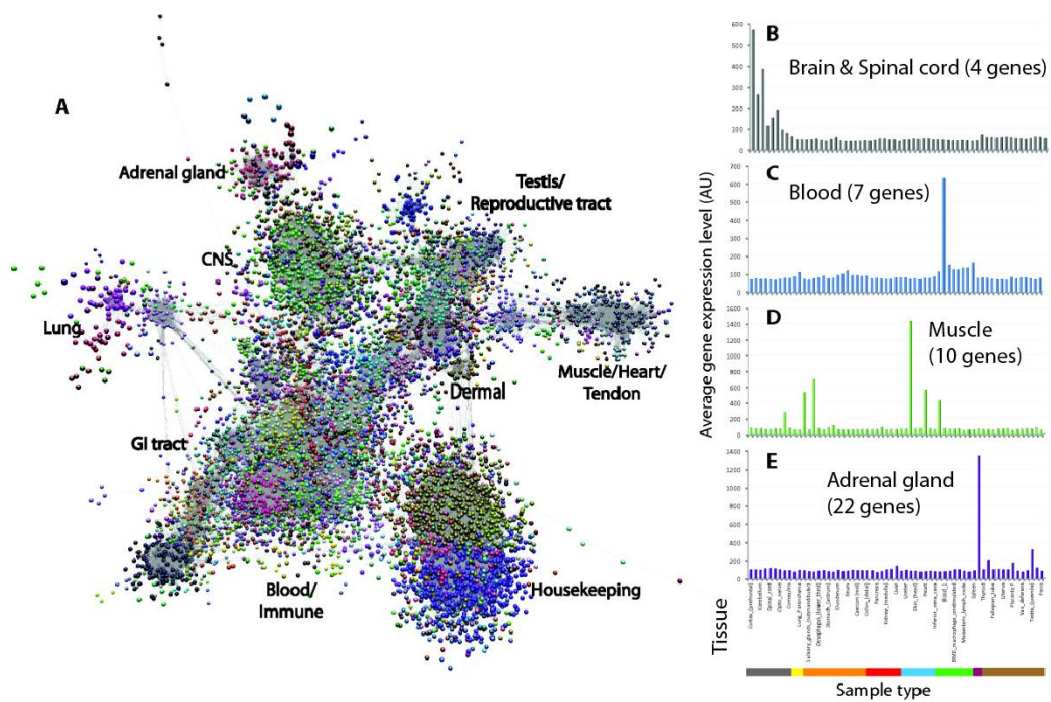
Visualisation and analysis of biological data as networks have been an increasingly important approach to explore and investigate a variety of biological relationships (Freeman et al., 2007). Network analysis has had a growing role in our efforts to comprehend the complexity of biological systems. NGS platforms have the ability to generate large datasets, and the relations or distance between biological components can be either measured experimentally or calculated (Theocharidis et al., 2009). Many studies have been successful using the approach in the study of sequence similarity (Enright et al., 2002), protein structure and protein interactions (Enright et al., 2003), and evolution (Li et al., 2003).

Incorporating biological data into a network model will enable one to exploit algorithms, techniques, ideas, and statistics previously developed in graph theory, engineering, computer science, and computational systems biology (Freeman et al., 2007). Networks are widely used in computer science. In network theory, a network usually consists of nodes connected by edges (lines). In biological networks, nodes usually represent an entity e.g. genes, transcripts, or proteins, while edges represent a relationship e.g. an experimentally determined similarity between entities (Bader and Enright, 2005; Miller et al., 2010).

Cytoscape (Shannon et al., 2003) is an open-source bioinformatics software platform for visualising molecular interaction networks together with gene expression profiles integration and other state data. It is one of the most popular available network visualisation tools which have been well supported in the community. The best features in Cytoscape is a plugin which is available for many analyses such as network and molecular profiling analyses, new layouts, additional file format support for expression and connection with databases and searching data in large networks.

However, the tool is unable to display and analyse very large networks and this tremendously limits its utility. When handling a large, highly structured network graph generated from expression or sequence data, Cytoscape strictly limits the ability to visualise and explore a network's topology. Nonetheless, the network analysis software BioLayout *Express*<sup>3D</sup> (Freeman et al., 2007; Theocharidis et al., 2009) was designed to analyse very large network graphs and provides a unique tool to analyse large complex expression datasets. It renders graphs in a 3D interactive OpenGL interface allowing a far better appreciation of complex graph structures which has proved to be extremely useful when dealing with large networks derived from microarray expression data. The statistical approach based on the transcript-to-transcript comparison of the microarray expression signal across different samples using a Pearson or Spearman correlation matrix is a basic principle of co-expression analysis (Freeman et al., 2007; Theocharidis et al., 2009).

Complex biological systems may be characterised and analysed as computable networks such as protein-protein interactions, genetic, biochemical, metabolic and cell signalling networks (Sevimoglu and Arga, 2014). Since BioLayout *Express*<sup>3D</sup> was released in 2007, it has been used to visualise gene expression using a network-based approach in many studies, such as a study of gene expression across a variety of tissues in the pig (a pig gene expression atlas) (Freeman et al., 2012), gene expression in human macrophages (Xue et al., 2014) and gene regulation in mammalian cells undergoing state changes (Arner et al., 2015). In the study of genes expressed in pig tissues (Freeman et al., 2012), an undirected network graph was built of a weighted pairwise transcript-to-transcript correlation matrix using a Pearson correlation threshold cut-off of  $r \geq 0.80$ . The resultant graph was large and highly structured (**Figure 1.9**). Network analysis of biological data has shown great promise, especially for microarray gene expression data, but no attention has been paid to RNA-seq data. These data are now abundant and of high quality, and consist of the type of high-dimensional data for which such approaches are well-suited (Freeman et al., 2007). Theoretically, the transformation of RNA-seq data into a network graph holds few challenges which depend on the read coverage, complexity of the graph and the splice variation of the selected gene.



**Figure 1.9: Network visualisation and clustering of the pig transcriptome.** (A) 3-D visualisation of a Pearson correlation graph of data derived from analysis of pig tissues and cells. Nodes represent individual probe sets on the array whilst the edges (lines) represent the correlations between individual measurements above the defined threshold. The network is comprised of 20,355 nodes (probe sets) and 1,251,575 edges (correlations  $\geq 0.8$ ). The complex topology of the network is a result of sets of coexpressed genes forming groups of high connectivity within the network. Clustering of the graph with the MCL algorithm (Section 1.6.1.1) was used to assign genes to groups based on coexpression. Areas of the network can be associated with genes expressed by specific tissue or cell populations. Plots of the average expression profile of genes shown on the right are (B) a profile of cluster 4 genes whose expression is restricted to brain and spinal cord (C) a profile of cluster 7 genes whose expression is highest in blood; (D) a profile of cluster 10 genes whose expression is restricted to skeletal muscle; (E) a profile of cluster 22 genes whose expression is highest in the adrenal gland (Freeman et al., 2012).

### 1.6.1.1 Graph clustering

High-throughput sequencing data need to be processed, analysed, and interpreted carefully. Clustering is one of the methods used to understand biological processes based on large data sets, particularly at the genomics level (Pirim et al., 2012). A cluster analysis step is commonly used for gene expression analysis. The process of identifying clusters of genes based on some aspect of biological similarity allows the data to be partitioned into smaller segments. Subsequently handling and analysis

becomes easier and more effective (Jiang et al., 2004). The MCL algorithm (Markov Cluster algorithm) was invented by van Dongen, (2000) and clustering of the nodes in BioLayout *Express*<sup>3D</sup> using the MCL algorithm was performed to assign genes to groups based on co-expression (Freeman et al., 2007). The MCL algorithm is one of the most effective graph-based clustering methods (Brohée and van Helden, 2006). This section explores the use of network-based visualisation of RNA-seq data to provide a complementary approach to understanding the accuracy of and differences in assembly algorithms, transcript structure, and splice variation.

### 1.6.2 Network visualisation

Generally, a graph  $G = (V, E)$  is used to draw information that can be represented as objects (the node set  $V$ ) and relations between those objects (the edge set of  $E$ ) (Hachul and Jünger, 2005). A major tool for analysing a graph is the automatic generation of layouts that visualise the graph and are easy to understand. A prominent type of algorithm to visualise graphs is the force-directed graph drawing method (Hachul and Jünger, 2007). It is based on assigning edges as springs and the nodes are electrically charged particles. The graph,  $G$ , is simulated as a physical system. The algorithm will try to place the nodes so that the total energy of the physical system is minimal (Harel and Koren, 2001; Herman et al., 2000). The graphs drawn using these methods are usually aesthetically pleasing and display symmetries, few edges crossings, uniformity of edge length and non-overlapping nodes and edges (Hachul and Jünger, 2007; Kobourov, 2012).

There are a few force-directed algorithms that are popular in graph layout as they are easy to implement and generally generate a ‘nice’ graph. These algorithms include versions published by Eades (Eades, 1984), Kamada-Kawai (Kamada and Kawai, 1989), Fruchterman-Reingold (Fruchterman and Reingold, 1991) and Davidson-Harel (Davidson and Harel, 1996). Both Eades and Fruchterman-Reingold algorithms are based on spring forces which are based on Hooke’s Law. There are repulsive forces between all nodes, but also attractive forces between nodes that are adjacent (Kobourov, 2012). Further improvements have been made to improve the speed and accuracy of these algorithms, including JIGGLE (Tunkelang, 1998),

FADE (Quigley and Eades, 2001), a hierarchical force-directed method for drawing large graphs (Gajer et al., 2001), a fast multi-scale method for drawing large graphs (Harel and Koren, 2001) and a multilevel algorithm for force-directed graph drawing (Walshaw, 2003). All these force-directed algorithms generate layouts of large graphs at reasonable times and the differences between them in terms of output are often small.

Initial development of BioLayout (Enright and Ouzounis, 2001) used a Fruchterman-Reingold algorithm (Fruchterman and Reingold, 1991) to construct a large network derived from biological data. This section will describe an alternative algorithm, the Fast Multipole Multilevel Method (FMMM) (Hachul and Jünger, 2005) for the purpose of visualising RNA-seq data. The FMMM algorithm is available through the Open Graph Drawing Framework (OGDF) (Chimani, 2007), an open source library that includes a variety of algorithms used in the drawing and analysis of graphs. The force-directed graph drawing algorithm developed by Hachul and Jünger (2005) and is a combination of an adequate multilevel technique to get the repulsive force between all pairs of nodes. The FMMM generates “pleasing” layouts and is comparatively fast (Godiyal et al., 2009). As part of the work described in Chapter 3, I have examined the use of two different algorithms on the layout of RNA-seq assembly graphs, Fruchterman-Reingold and FMMM algorithms.

### **1.6.3 Algorithms for sequence assemblies**

In RNA-seq assembly graphs, nodes represent sequence reads while edges denote a similarity overlap or homology between reads above a defined threshold. Overlaps must be pre-computed by a series of (computationally expensive) pairwise sequence alignments (Pop, 2009) as the first step in building a graph from such data (Miller et al., 2010). The evolution of assembly algorithms has accompanied the development of sequencing technologies. Currently, there are two widely used classes of algorithms: overlap–layout–consensus (OLC) and de-Bruijn-graph (DBG) (Zerbino and Birney, 2008).

### 1.6.3.1 The Overlap-Layout-Consensus (OLC) algorithm

OLC generally works in three steps: first overlaps (O) among all the reads are found, then it carries out a layout (L) of all the reads and overlaps information on a graph and finally, the consensus (C) sequence is inferred (Flicek and Birney, 2009). It is an assembly algorithm, initially developed by Staden (1980) and subsequently extended and elaborated upon by many scientists. OLC became successful with the wide application of Sanger sequencing technology (Section 1.1.1). Many widely used assembly programs adopted OLC, such as the Celera Assembler (Myers et al., 2000) and CAP3 (Huang and Madan, 1999). In the OLC algorithm, the identification of overlap between each pair of reads is explicit; typically by doing all-against-all pairwise read alignment (Flicek and Birney, 2009). As a result, the OLC algorithm constructs a reads graph, which places reads as nodes and assigns a link between two nodes when these two reads overlap by more than a cut-off length. The basic principle of the OLC method has been used in this thesis to construct a graph-based assembly of RNA-seq data.

### 1.6.3.2 The de Bruijn Graph (DBG) algorithm

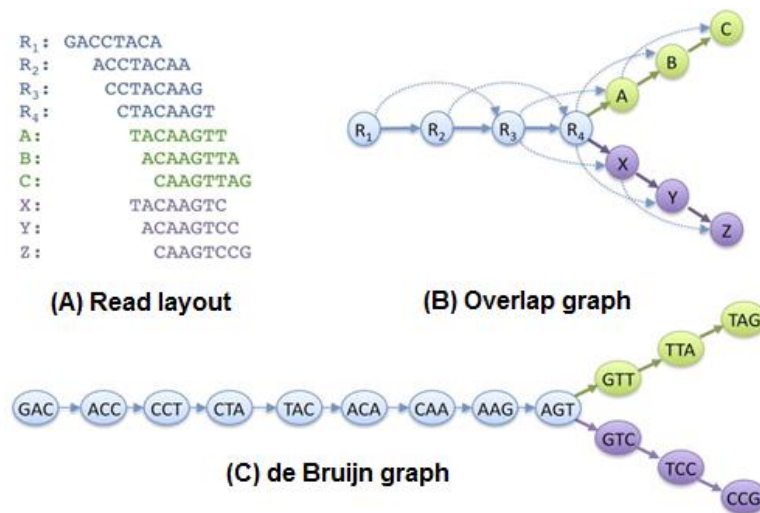
The method used to exploit the overlap information is different in the DBG algorithm than in the OLC algorithm. A DBG is a compact representation based on short words (k-mers) that is ideal for high coverage, very short read (25–50 bp) datasets where the k is 19 or higher (Zerbino and Birney, 2008). For every k-mer observed in the sequence set, a node is created, while edges are drawn between every pair of successive k-mers in a read if the k-mers overlap by  $k - 1$  bases. Some of the edges are therefore correlated with the single-step base difference of moving the fixed k-mer window along by one position.

The DBG formulation has properties that differ from OLC in important ways because of the use of k-mers to calculate the overlaps, even though it is seemingly like the read overlap graph used by traditional assembly programs which use OLB algorithms. There are three steps involved in DBG algorithms. The first step is that a read will be split across its component nodes. Secondly, a short sequence repeat will be a series of adjacent k-mers which many reads pass through. The graph will

diverge into the unique regions of the genome at the edges of the repeat. Finally, the graph can be constructed in an amount of computational time that scales linearly with the number of reads (Flicek and Birney, 2009). An overview of the DBG graph is shown in **Figure 1.10**.

Novák et al. (2010) first introduced the idea of using graph-based methods to visualise DNA assemblies. Their work was based on an all-to-all comparison of sequence reads to generate the similarities, which were used to build clusters of overlapping reads representing different repetitive elements of two plant genomes (pea and soybean) where read similarity exceeded a specified threshold. This method focused on the characterisation of the repetitive regions of plant genomes where it was argued the method could be used to better analyse the variability and evolutionary divergence of repeat families, as well as to discover and characterise novel elements. This all-to-all comparison can be performed using MegaBLAST (Morgulis et al., 2008) which is a part of the BLAST (Altschul et al., 1990) program in the NCBI toolkit. It is this basic approach that was developed in this thesis, and detailed explanation is given in Chapter 2.





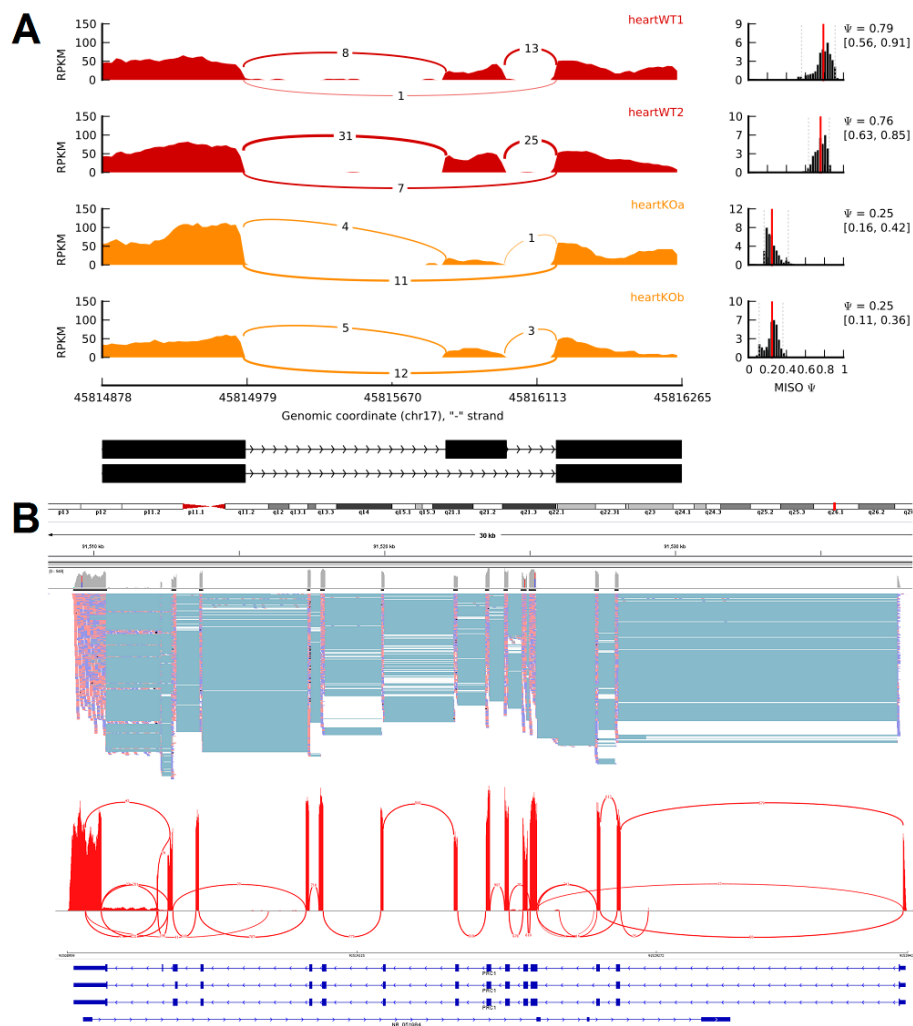
**Figure 1.10: Differences between an OLC and a DBG for assembly.** Based on the set of 10 8-bp reads (A), we can build an overlap graph (B) in which each read is a node, and overlaps  $> 5$  bp are indicated by directed edges. Transitive overlaps, which are implied by other longer overlaps, are shown as dotted edges. (C) In a DBG a node is created for every  $k$ -mer in all the reads; here the  $k$ -mer size is 3. Edges are drawn between every pair of successive  $k$ -mers in a read, where the  $k$ -mers overlap by  $k - 1$  bases. In both approaches, repeat sequences create a diversion in the graph. This example only considered the forward orientation of each sequence to simplify the figure. Figure redrawn from Schatz et al. (2010).

#### 1.6.4 Visualisation of RNA-seq assemblies

When next-generation sequencing came into the market in 2005, RNA-seq application became an option to generate enormous amounts of expression data. Various pipelines such as TopHat and Cufflinks are widely applied to analyse these datasets. The most important aspect in analysing RNA-seq data is the ability to visualise the complexity of AS. Previously, many tools have been developed to visualise alternative isoform from cDNAs and ESTs data (Bhasi et al., 2009). Visualisation for RNA-seq assemblies needs dedicated tools that efficiently process a large amount of data from multiple samples from different cells or tissues. Thus, several tools have been developed for the purpose to visualise alternative isoforms and events from RNA-seq data. However, accessing and handling the analytical output remain challenging for most researchers. To further analyse, visualise and interpret the RNA-seq data, the assemblies can be viewed using visualisation tools.

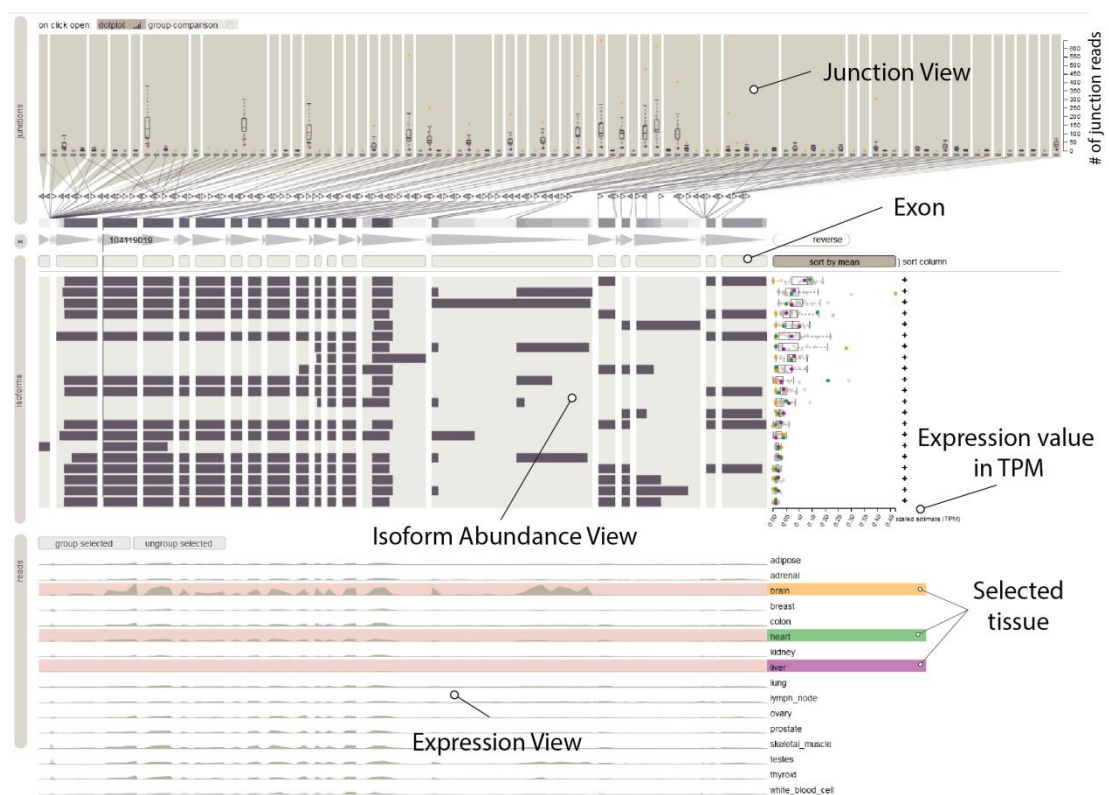
The simplest approach to visualise isoforms and events from such gene is to generate track files for a genome browser e.g. UCSC browser (Speir et al., 2016) or GBrowse (Donlin, 2002). A number of tools such as RSEM (Li and Dewey, 2011), SpliceGrapher (Rogers et al., 2012) and DiffSplice (Hu et al., 2013) can produce WIG files or GFF-like formats to be uploaded into these browsers. Today, perhaps the most popular tool to visualise RNA-seq data is Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). This software provides a versatile visualisation and exploration platform for DNA/RNA sequence data. It was developed to support a wide-range of data types, including NGS and array-based platforms. The distinguishing feature compared to other visualisation viewers such as Tablet (Milne et al., 2010), BamView (Carver et al., 2010) and Artemis (Carver et al., 2012), is that IGV can view data in multiple genomic regions simultaneously.

All these visualisation tools have limitations and the software is constantly under development. Most of the visualisation tools available provide assembly/alignment views as reads ‘stack’. Whilst this is sufficient for many needs when the underlying variances in the genome or transcript assemblies are complex, existing visualisation methods can be limiting. One approach that assists is a quantitative multi-sample visualisation of RNA-seq reads aligned to gene or locus annotations, called Sashimi plots, developed by Katz et al. (2013) (**Figure 1.11**). The Sashimi plots are made using alignments (stored in the SAM/BAM format) and gene model annotations (in GFF format), which can be visualised using the IGV browser, which enables swift and dynamic creation of the plots for any gene or locus suitable for exploratory analysis of alternatively spliced regions of the transcriptome. This tool produces publication-ready plots of isoform expression for RNA-seq analyses. The Sashimi plots provide a quantitative summary of genomic and splice junction reads together with the gene model annotations and read alignments. The quantitative and comparative visualisation of RNA-seq reads can be done across different samples to detect differentially expressed spliced exons and isoforms (Katz et al., 2010). However, none of these tools uses a network-based approach to visualise and identify alternatively spliced isoforms in RNA-seq data, which is described in this thesis.



**Figure 1.11: Example Sashimi plots for an alternatively spliced exon. (A)** Gene model annotation showing two transcripts where the middle exon is alternatively spliced. Sashimi plots for the black exons (shown in lower panel) is shown, where genomic reads are converted into read densities (measured in FPKM) and junction reads are plotted as arcs whose width is determined by the number of reads aligned to the junction spanning the exons connected by the arc (Katz et al., 2010). **(B)** Representative IGV view (upper panel) together with Sashimi coverage plot (lower panel) showing RNA-seq reads mapping to the *PRCI* locus from human fibroblasts at 24 h after serum refeeding (Freeman et al, unpublished data). The height of the bars represents overall read coverage. Splice junctions are displayed as loops. The number of reads observed for each junction is indicated within segments, and y-axis ranges for the number of reads per exon base are shown (read coverage, left). The plot suggests different isoforms expressed in the sample, as shown by the arcs connecting a pair of exons.

Recently, Strobel et al. (2016) developed a visual analysis tool called Vials (Visualizing Alternative Splicing) (**Figure 1.12**) to explore the publicly available datasets to determine the abundance of isoforms which are associated with coding regions of the gene and evidence for read junctions. The tool is scalable for the simultaneous analysis of numerous samples in multiple groups. Vials tool allows the researcher to identify patterns of isoform abundance in groups of samples for example tissues and the quality of the data can be determined. This tool is used to determine the isoforms and compare with the network analysis of a human tissue atlas described in Chapter 4.



**Figure 1.12: Vials - Visualisation of Alternative Splicing.** Vials showing isoforms for the Kinesin light chain 1 (*KLC1*) gene and data from the Illumina BodyMap 2.0 data. In this figure, there are three junctions; junction view, isoform abundance view and expression view. Three tissues of the brain, heart, and liver are selected from the expression view.

## 1.7 Aims of the thesis

Data visualisation is increasingly recognised as an essential component of genomic and transcriptomic data analysis, enabling large and complex datasets to be better understood. However, the analysis of RNA-seq data remains a significant challenge for many biologists. The data is large and the tools for its assembly, analysis, and visualisation are still under development. Currently, there are a number of software tools available for the visualisation of sequencing data, the most widely used of which is the Integrative Genomics Viewer (IGV).

However, visualisation of the data is still linear and basically involves read stacking onto the genome reference and visualisation of splicing events is difficult. The limitation of linear visualisation is that the data could not be interpreted at a glance e.g. to detect an alternative splicing. Difficulty to determine isoform expressed could potentially lead to data misinterpretation. Therefore, the work described in this thesis explores ways of visualising and examining RNA-seq data. Hence, the hypothesis is to provide an alternative visualisation of the sequencing data and therefore better analysis of exon structure and splice variants.

Ultimately, this new method of visualising RNA-seq data will become important in RNA-seq analysis to discover and explore the nature of the sequence. Therefore, the overall aim of this thesis was to explore the utility of network-based visualisation in the analysis and interpretation of RNA-seq data. In summary, the aims of this thesis were:

1. To develop a pipeline to go from ‘raw’ RNA-seq data to a layout file that can be visualised as an ‘RNA-seq assembly graph’ using the network analysis tool, BioLayout *Express*<sup>3D</sup>.
2. To better understand the basic principles and challenges associated with network visualisation of RNA-seq data, in particular how it could be used to visualise transcript structure and splice variation.

3. To explore the analysis of transcript variation in an RNA-seq dataset derived from human tissue including the quality control of the sample using a network-based approach, detect alternative splicing events and validate them using network-based visualisation.
  
4. To investigate the usability of network-based visualisation using NGS Graph Generator application thus enables review and feedback from users to improve the user experience for this application.

## Chapter 2 - Development of an analysis pipeline for the network analysis of RNA-seq data

### 2.1 Introduction

Data visualisation is a fundamental component of genomic and transcriptomic data analysis. However, diversity and size of the data sets produced by current sequencing and array-based transcriptome profiling methods present major challenges for analysis and visualisation. Several analytical and visualisation approaches have been used to analyse and display the relative abundance of a mixture of transcript isoforms in a sample as analysed using RNA-seq data (Refer Chapter 1 Section 1.5.3). To address the need for alternative methods to explore transcript isoform diversity in RNA-seq data, I developed a computational pipeline that uses genome-mapped short-read sequences to generate a network-based visualisation of these data. This web application, called “NGS Graph Generator”, can be accessed at <http://seq-graph.roslin.ed.ac.uk>, is written as a Bash script and maintained in a GitHub repository. The pipeline relies on several external libraries, many of which enable fast and efficient processing of short read data, namely: SAMtools, R programming language, its Python wrapper and *GenomicRanges* for read overlap. This pipeline has been incorporated into a web application which can be used to generate DNA assembly networks from RNA-seq data.

This chapter describes the NGS Graph Generator pipeline used to generate networks from RNA-seq data, the results of which will be described in subsequent chapters. The package provides a network layout file to enable visualisation of read assemblies the result for a given gene or transcript. This pipeline processes data from a sequence mapping file (i.e. BAM file) to a network layout file which can then be visualised using network visualisation software, BioLayout *Express*<sup>3D</sup> or Miru from Kajeka Ltd, Edinburgh, United Kingdom.

In this chapter, the use of network-based visualisations of sequencing data was further explored, applying the fundamental principles first described by Novak et al. (2010) to RNA-seq data. The aim of this chapter was to develop a complementary approach to understanding differences between RNA assembly algorithms as well as to better understand transcript structure and splice variation. In doing so, a platform that supports the improved interpretation of complex transcript isoforms is provided. This approach will be useful in the exploration and discovery of new insights from sequence data.

The objectives of this work were:

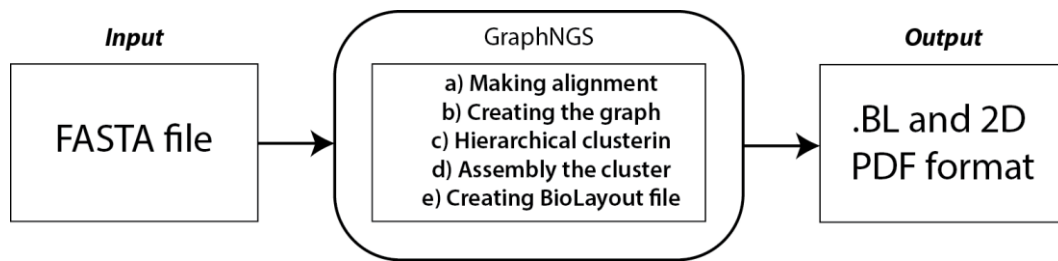
- a) To develop a seamless informatics pipeline from a sequence file of RNA-seq data to a network visualisation file.
- b) To implement the pipeline into a web-based application tool to allow a user to run the pipeline via a web interface.
- c) To create an Amazon Cloud Image (AMI) to allow the pipeline to be run in a cloud environment.



## 2.2 Correction of GraphNGS pipeline

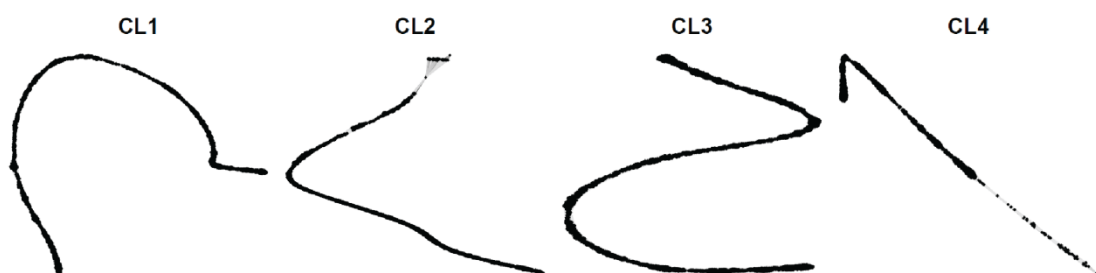
GraphNGS pipeline (Faulkner et al, unpublished, Roslin Institute) was initially developed to analyse and visualise NGS data using a network-based approach. The pipeline was based on the work of Novák et al. (2010), which creates networks from NGS data i.e. genomics or transcriptomics data. This GraphNGS pipeline contains several Python and R scripts to process the FASTA files to create a network input file that can be visualised on BioLayout *Express*<sup>3D</sup>. The pipeline written in Python produced two types of cluster outputs which are in a portable document file (PDF) format in 2D, as described by Novak et al. (2010), and .BL format which can be imported into BioLayout *Express*<sup>3D</sup> where the network can be visualised in 3D format. The output from GraphNGS can also be visualised on a SeqGraper visualisation tool (Novák et al., 2010).

Several steps are involved in this GraphNGS pipeline (**Figure 2.1**), which is a) making the alignment; b) creating the graph; c) hierarchical clustering; d) assembly of the cluster, and e) creating the BioLayout file. The first step of this pipeline is making the alignment, which uses the MegaBLAST to perform a read-to-read comparison. The MegaBLAST scripts create a table which reports the similarity metrics for each pair of sequences (*suffix\_megablast.txt*). This MegaBLAST output is used to create another table (*suffix\_pairwise.txt*) which has three columns (Node A, Node B, Weight) using the megablast2ncol scripts in the second step. The third step is using an R script named fgclust4.4.r to compute the weight of the edge between two reads. The fourth step creates contigs (cluster) and class sets for each cluster in BioLayout format. In the last step, two outputs are generated, which is in PDF (2D) and .BL format.



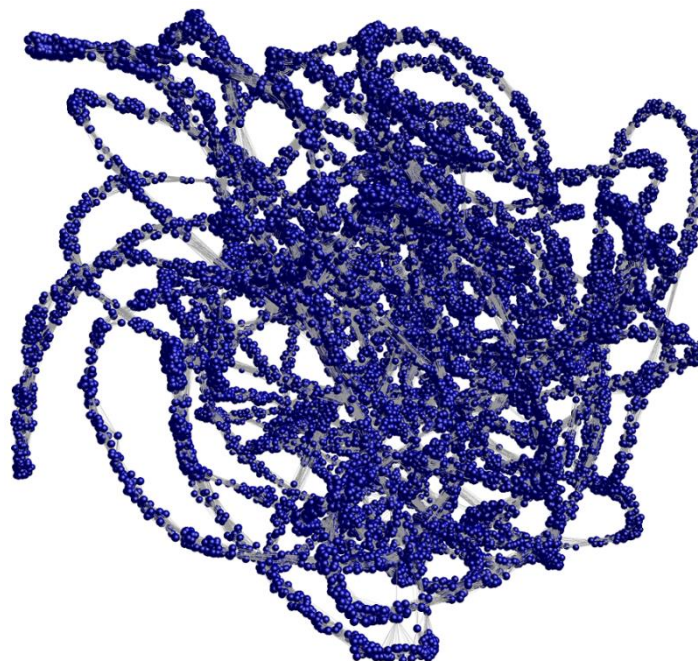
**Figure 2.1: The basic principle of GraphNGS pipeline.** The pipeline that uses FASTA read as an input to process the read, cluster and produce outputs in a .BL format that can be visualised using BioLayout *Express*<sup>3D</sup> as well as a PDF format.

In order to ensure the GraphNGS pipeline was working, an RNA-seq data of human fibroblasts were used. The details analysis of this sample will be described in Chapter 3. However, for this purpose, reads mapping to *COL5A1* was used. A quality control was performed and the data were aligned to the human reference genome (GRCh37.71) using TopHat (Trapnell et al., 2009). The *COL5A1* is large (8,468 bp), containing 66 exons, and is one of the most highly expressed genes in human fibroblasts. Around 40,170 reads were mapped to this gene locus (t24 h sample) and these were put through the GraphNGS pipeline. The output file from GraphNGS was laid out and produced four components using SeqGrapheR (**Figure 2.2**).

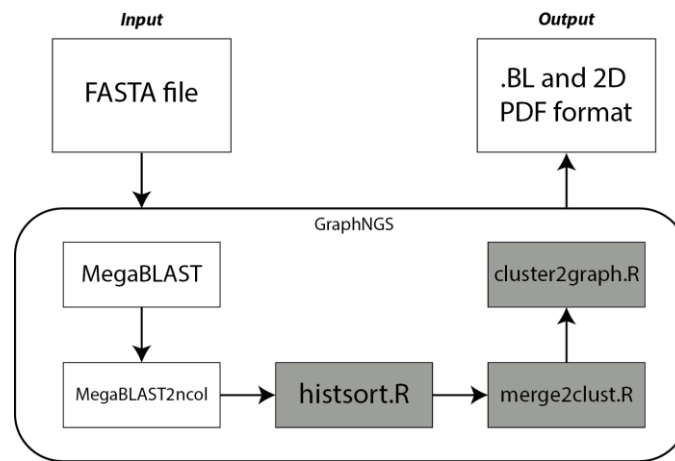


**Figure 2.2: Network visualisation of the *COL5A1* transcript.** In this case, *COL5A1* was used. The output generated using GraphNGS pipeline using SeqGrapheR. The network transcript of *COL5A1* separated into four components. The transcript network of *COL5A1* generated after the correcting the GraphNGS pipeline.

As *COL5A1* (**Figure 2.3**) has a high coverage of reads and when visualised in IGV there were no weak links or gaps in the sequence, the network layout should not, in theory, break up, thus this result was quite unexpected. However, it was believed that additional code could solve the error of producing multiple components to the *COL5A1* transcript network. This code was `hitsort2cluster.r`, `merge2cluster.r`, and `cluster2graph` where the four different components produced by the GraphNGS pipeline were merged (**Figure 2.4**). The `hitsort2cluster` takes the input file of paired reads and weights and creates a file which defines clusters. The two different outputs generated are “*INPUT.cls*” and “*INPUT.tcls*”. The `.cls` file is generated based on hierarchical agglomerative algorithms while `.tcls` file is generated based on connected components’ algorithms.



**Figure 2.3: Network transcript of *COL5A1*.** This network consists of 40,170 nodes and 802,182 edges mapped on this gene, and crucially they assemble as one network. This network is visualised using the Fruchterman-Reingold algorithm in *BioLayout Express<sup>3D</sup>*.



**Figure 2.4: Correction of the GraphNGS pipeline.** The pipeline was corrected by adding the code available at SeqGrapheR developed by Novák et al. (2010). Three R scripts (grey box) were added into the pipeline and produced the desired network transcript of NGS data.

In order to make it more reliable for the analysis of NGS data, especially for RNA-seq data, this pipeline has been corrected in several ways. This includes testing the pipeline and adding some scripts from the SeqGrapheR development site. The outputs from `histsort2clusters.R` were subjected to another step to merge the clusters. This step used the `merge2clusters.R` script to merge the cluster that generated from the previous step. This merging step generated a file that contains all connected cluster in a GL format. This GL format is a binary format which only could be opened using SeqGrapheR software developed by Novák et al. (2010). SeqGrapheR is a package that provides an interactive GUI for visualisation of DNA sequence clusters. The output GL format was imported into SeqGrapheR, with a network layout connected component being generated, suggesting that this GL format could be used as a layout file for BioLayout; however, this format it cannot be loaded in the BioLayout *Express*<sup>3D</sup> software. The readable output was tweaked from the `merge2clusters.R` scripts to get the output that could be imported into the BioLayout. Nonetheless, the GraphNGS pipeline possesses a few limitations. This includes the limitation to visualise and explore alternative splicing or isoform divergence within a single cell or across tissues. Therefore, in the next section, I described a network-based visualisation pipeline which can analyse a comprehensive gene transcript using BioLayout *Express*<sup>3D</sup>.

## 2.3 Network-based visualisation pipeline: an alternative solution

Since the GraphNGS pipeline output was not able to produce a complete analysis to visualise the DNA assembly network on BioLayout *Express*<sup>3D</sup> as well as the limitation of SeqGrapheR on visualising the network without any exon or isoform annotation, a seamless pipeline was needed to visualise, explore and determine a DNA assembly network of a gene. The idea to develop this pipeline is based on a set of linked Bash and Python scripts that perform the following tasks. **Figure 2.5** shows an overview of the data processing pipeline for analysis of RNA-seq data developed in this chapter for the preparation of network files for transcript visualisation. The initial steps in this pipeline i.e. data QC and mapping of reads to a reference genome are described elsewhere (Garber et al., 2011; Trapnell et al., 2012). Described here are the stages from the output of an alignment analysis (BAM file) to the generation of a text file suitable for network visualisation using BioLayout *Express*<sup>3D</sup>.

To be able to run this pipeline, a user requires the short sequence read files generated by an NGS platform i.e. FASTQ file. The size of short sequence reads vary from platform to platform and in this chapter 100 bp paired-end reads generated from Illumina sequencer were used. Normally short read sequences produced from a sequencer will produce high-quality sequence reads. If data is of a poor quality then downstream sequence analysis is compromised by low-quality sequences, sequence contamination, and sequence artefacts, ultimately leading to misassembly and potentially leading to erroneous conclusions. Such data requires enhanced tools for preprocessing and quality control of sequence datasets.

The quality assessment for most NGS data sets includes analysis of sequence length, quality score, GC content and sequence complexity distributions (i.e. sequence duplication, artefacts, contamination and number of ambiguous bases). In the pre-processing stage, unwanted sequence resulting from adaptors or poor-quality sequence should be trimmed and filtered, respectively. The Kraken tool (Davis et al., 2013) is recommended to perform this quality control step and is it necessary to

perform subsequent steps in generating the network-based analyses (**Figure 2.5 – Step 1**).

After quality control, the next step is to align the FASTQ files to a reference genome using tools such as TopHat (Trapnell et al., 2009) or STAR (Dobin et al., 2013) (**Figure 2.5 – Step 2**). The output from these tools is a sequence alignment in a binary format (BAM – Binary Alignment Mapping file). This sequence alignment will then be input into the NGS Graph Generator pipeline. The NGS Graph Generator pipeline comprises four core components; SAMtools, *GenomicRanges* (Lawrence et al., 2013b), MegaBLAST and Uniquification (**Figure 2.5 – Step 3 to 6**). The input for the pipeline is a BAM, GTF and a chromosome length file. The chromosome length file is a tab-delimited text file that contains the chromosomal name and the chromosome length in bases. If the alignment is an unsorted BAM file, the file will be sorted based on the genome location using the SAMtools sort function (Li et al., 2009a) (**Figure 2.5– Step 3**).

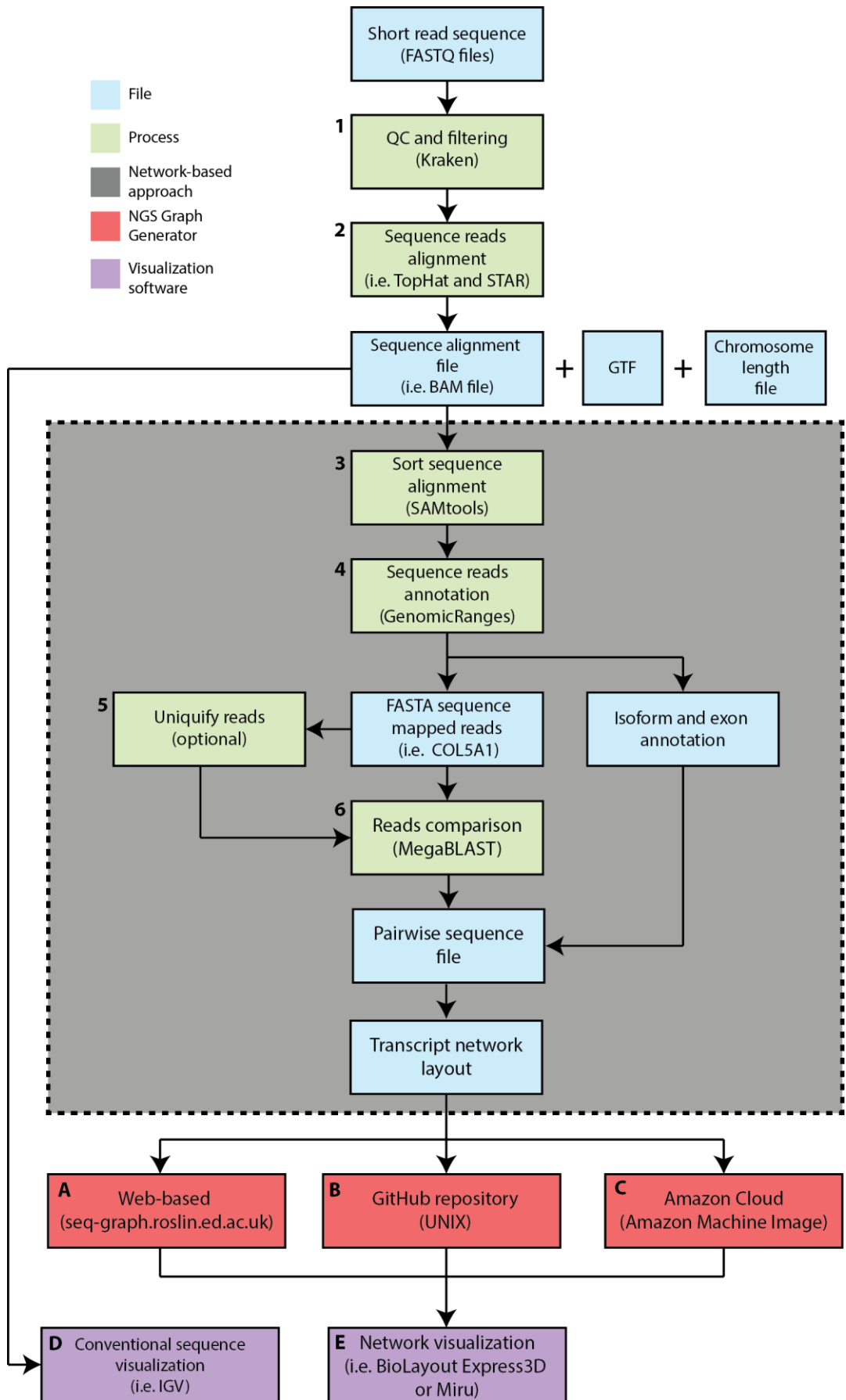
The R package *GenomicRanges* (**Figure 2.5 – Step 4**) is then used to create the file, a GTF (gene transfer format) file is required for annotating nodes in the network (representing sequence reads). A GTF file holds information about gene structure (as defined by Ensembl GRCh37.71) and is used by *GenomicRanges* to obtain isoform and exon information. The output from this step is a tab-delimited file containing information that includes Ensembl transcript IDs and exon number. This information is extracted from the BAM file and can be overlaid onto networks using the ‘class sets’ function within BioLayout *Express*<sup>3D</sup>. Upon selection of the Ensembl transcript ID, nodes representing reads that map to this transcript model will be coloured arbitrarily according to the exon number. The sequence mapping to a selected gene (e.g. *COL5A1*) will be created as a FASTA file.

Depending on the target gene’s length and coverage, it will take longer to generate a network of a highly expressed gene than moderate or low expressed genes. One option is to ‘uniquify’ (discard redundant reads) thereby reducing the computational time to calculate the read-to-read similarity matrix (**Figure 2.5 – Step 5**).

The next step is to define the similarity (sequence identity) between reads mapping to a gene of interest (**Figure 2.5 – Step 6**). A FASTA file containing all the sequences that have been mapped to a particular gene is extracted and the supporting information used for the visualisation of transcript isoforms in the context of the resultant network. For read-to-read comparison, MegaBLAST (Zhang et al., 2000) is used to generate a similarity matrix with edge weights derived from the MegaBLAST bit score. The better the alignment between a pair of reads the higher the bit score.

Parameterisation of this step i.e. defining the threshold for percentage sequence similarity ( $p$ ) and length ( $l$ ) coverage for which two sequences must be similar for an edge to be drawn between them is of importance. Ideally, a network should contain the maximum number of reads (nodes) connected by a minimum number of edges and where possible give rise to a single network component i.e. a single group of connected nodes that together represent the mRNA species of interest. These parameters,  $p$ , and  $l$  can be varied. If the read depth is high, the parameters can be more stringent, the opposite being true when coverage read is low.

The NGS Graph Generator pipeline was implemented as three different platforms to reach different audiences. First, a DNA assembly layout of a gene can be generated from a website (**Figure 2.5A**). Secondly, the source's code was deposited in GitHub repository (**Figure 2.5B**) and finally the pipeline is implemented in as an Amazon Machine Image (AMI) at Amazon Web Services (**Figure 2.5C**). The layout file is the product of the NGS Graph Generator pipeline. This file is saved as a text file that contains a list of weighted edges between pairs of reads, together with class sets information that defines reads as belonging to Ensembl transcript models. This network description file is called a .layout file and can be opened using BioLayout *Express*<sup>3D</sup> or Miru (**Figure 2.5E**). To compare and analyse the network layout, a BAM file can be visualised using IGV software (**Figure 2.5D**).





**Figure 2.5: Pipeline for network-based visualisation of RNA-seq data.** The analysis pipeline is shown for building networks of RNA-seq data from raw sequencing FASTQ files through a series of analysis steps (green box), generating several files (blue box) leading to the production of a file for network visualisation layout in the BioLayout *Express*<sup>3D</sup> software. Short read sequence (i.e. FASTQ) files from the NGS sequencer are used as the input for this pipeline, followed by quality control (1) for which the Kraken pipeline is recommended. Then the FASTQ files are aligned to a reference genome for the sequence alignment step using TopHat or STAR (2). The output from the aligners is a BAM file which is used as an input for the NGS Graph Generator (grey box) pipeline together with GTF and the chromosome length file. The BAM file will subsequently be sorted based on the genome coordinates using SAMtools (3). This is followed by sequence read annotation using *GenomicRanges* (4). The output of this step is FASTA sequence of mapped reads and the node class file of exon and isoform annotation for each sequence read. Depending on the depth of sequencing (gene expression level), an option to remove redundancy of sequence reads can be used (5). The next step is to perform a read-to-read comparison of all short sequence reads of selected gene/locus using MegaBLAST (6) producing a pairwise sequence file. The pairwise sequence and node class file merge to produce a transcript network layout. This NGS Graph Generator (grey box) is incorporated into three different platforms, (A) web-based application, (B) GitHub (UNIX) and (C) Amazon Cloud. To view the RNA-seq data using a conventional visualisation platform, IGV viewer should be used, and the results can be compared with the network-based visualisation approach using BioLayout *Express*<sup>3D</sup>.

This chapter presents a network-based visualisation pipeline to generate networks from RNA-seq datasets. This pipeline has provided the results described in the subsequent chapters. The pipeline was written in Bash script which includes Python, Bash, and R to analyse short-read NGS data from RNA-seq data of the cell cycle of human fibroblasts. The package provides a layout file to help scientists to visualise the result of a gene or transcript.

All the pipeline design and implementation work was my own apart from the corrected GraphNGS pipeline and adaptation of the pipeline for its use as web application and the preparation of the software, which were done in collaboration with Tim Angus, The Roslin Institute, and Anton Enright from the EMBL European Bioinformatics Institute (EBI). The implementation of the pipeline for its use in the

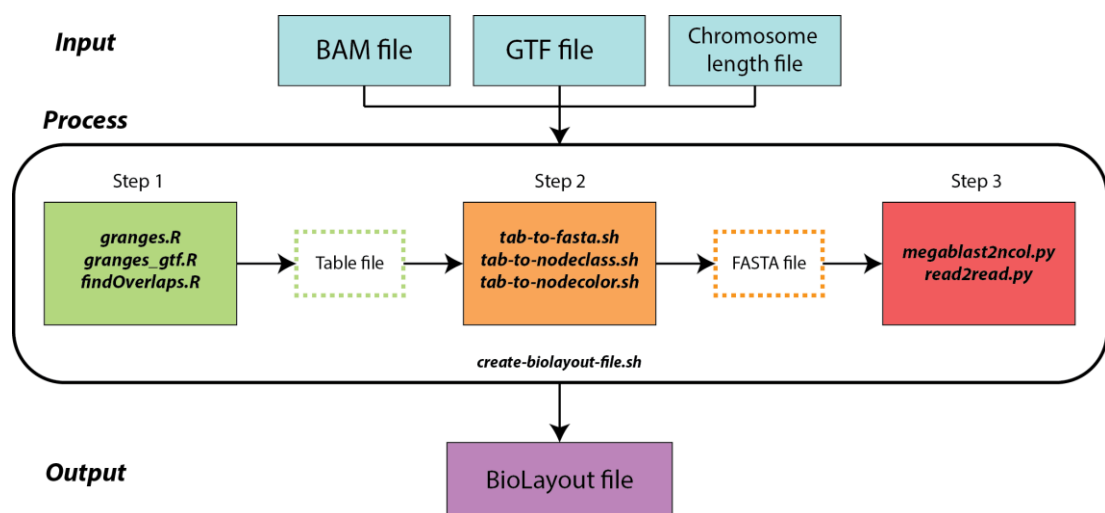
Amazon Machine Image which was done together with Professor Mick Watson from Edinburgh Genomics, The Roslin Institute.

This pipeline processes data from a BAM file to a layout file which can be visualised using network analysis software, BioLayout *Express*<sup>3D</sup>. A Bash file, *create-biolayout-file.sh*, written in a Bash script, contains a series of commands to execute different scripts including R, Python, and shell script. It takes mapping data, preferably an unsorted BAM file through all the processes to produce a network and visualise it using network visualisation software. The NGS Graph Generator is an open source and can be run on Linux systems.

Documentation and the full source code are in the NGS Graph Generator package and can be downloaded from GitHub allowing users to analyse their own data using this approach. BAM and GTF files are required to run this pipeline. In order to demonstrate this pipeline without the need to run it on in-house servers, I have developed a front-end interface that allows the scripts to be run from a website. This front-end is called ‘NGS Graph Generator’ and can be accessed at <http://seq-graph.roslin.ed.ac.uk>. Using this website, a user can select BAM files of RNA-seq time-course samples from human fibroblasts or example data from the human tissue atlas. Other default settings are required, such as the chromosome length file, the GTF file and the name of the gene of interest. A user can adjust the desired percentage similarity,  $p$  and percentage sequence length overlaps,  $l$ , to use with MegaBLAST and choose whether to include an option `-u` to uniquify (discard redundant reads). The user must provide their email address, and they will be informed through email once the job has finished. The results section shows the job id, owner of the job, arguments on the Linux system; time queued, processing time on the job and result. When the job has finished, the network layout file can be accessed, in the results section. When selected (clicked) BioLayout *Express*<sup>3D</sup> or Miru will automatically open with the network displayed (if installed on the computer).

## 2.4 Component of NGS Graph Generator

Described in the previous section is the principle of each step in building a network of a transcript from RNA-seq data and visualisation of the resultant network. Here, I describe the development of the framework and getting it into a pipeline that can be used by others. The framework in **Figure 2.6** is a shell script called *create-biayout-file.sh* which was written in conjunction with Tim Angus and contains all scripts necessary to build a layout file. It also has dependencies on R packages, SAMtools and MegaBLAST. The input needs a BAM file, a GTF file (the same GTF file use in the mapping process) and a chromosome length file which is used to annotate sequence reads. In this network-based visualisation, a gene or locus of interest is used to generate a layout file.



**Figure 2.6: Framework of NGS Graph Generator.** The overall scripts used in the framework are written in R, Python, and Bash. Inputs of BAM file, GTF and chromosome length file are fed into the framework. These inputs are processed through a series of scripts and eventually produce a layout file which can be visualised using BioLayout *Express*<sup>3D</sup>. **Step 1 – Extracting mapped reads of a specific gene.** R scripts of *granges.R*, *granges\_gtf.R* and *findOverlaps.R* are used to manipulate the BAM file and extract reads mapped to a specific gene. These processes produce a table file containing all important information such as read ID, sequence, isoform ID and exon. **Step 2 – Creating FASTA, node class, and node colour files.** In this step, a FASTA file of reads mapped to a specific gene or locus is created using a Bash script *tab-to-fasta.sh*, whilst to annotate all reads to their isoforms and exons in a later layout file, other Bash scripts *tab-to-nodeclass.sh* and *tab-to-nodecolor.sh* are used. **Step 3 – Read to read comparison.** The last step is a read-to-read comparison. All reads mapped to a specific gene or loci are compared

using a Python script *read2read.py*. The output produces a large MegaBLAST file and another Python script *megablast2ncol.py* is used to cut off to a certain edge weight. The parameter percentage sequence similarity over percentage length coverage is used to determine edge weight between two reads.

This task is accomplished using *grangesscript.R* and *grangesscript\_gtf.R* as shown in **Figure 2.6 – Step 1**. These two R scripts utilise the GTF file and classified reads mapped to associated isoforms. The output of these scripts is a table consisting of several pieces of information, including, read IDs, sequence read, exon number, gene ID, transcript ID, and isoform ID.

The next step is to compare all reads mapped to a gene above a certain threshold to build a network. Therefore, to build a network, all reads mapped to a given gene will be extracted out using an R script, *findOverlaps.R*. At this stage, all reads from all isoforms based on the GTF file are mixed up. The origin of all reads from each isoform in each gene needs to be specified. To extract reads mapped to the gene, a script *tab-to-fastq.sh* is used to get a FASTA file (**Figure 2.6 – Step 2**). If a gene is highly expressed, it will contain a lot of identical reads and use the option *-u* is highly recommended. It will take a lot of computational time if this option is not used and eventually the network will not be able to be visualised on a desktop computer due to the massive size of the layout file.

In the last step (**Figure 2.6 – Step 3**), comparison of all reads is needed to build up a network. To handle this task, a Python script *read2read.py* is used. This script takes the reads in FASTA format to create a database for the specific gene using *formatdb*. After that, a MegaBLAST program is executed to perform a read-to-read comparison and generate the output as a MegaBLAST file. In this output, a bit score of each alignment is used as an edge to build a network. Edges need to be passed at a certain threshold to build a minimal size network using another Python script *megablast2ncol.py*. This script will transform MegaBLAST output in *n* column (*ncol*) format and filter out edges below the given threshold. The script reads the output of MegaBLAST (*-D 3* option for output format). It takes a lot of memory if the MegaBLAST output file is big. To overcome this problem, I parallelised the

MegaBLAST based on the available server memory. The output of this script is a basic layout file without any annotation. This file can be visualised using network analysis software, but it has no information on the origin of reads (nodes). To annotate the network, other scripts *tab-to-nodeclass.sh* and *tab-to-nodecolor.sh* are used. They take the output MegaBLAST and annotate each read as belonging to known transcript isoforms. Finally, the layout file, node class, and node colour files are merged to produce the final layout file which can be visualised with BioLayout *Express*<sup>3D</sup> or Miru.

## 2.5 Implementation network-based visualisation pipeline

### 2.5.1 Web-based application

NGS Graph Generator is a web-based application for visualisation of RNA-seq data assemblies as a network. A user can generate a layout file from a website or provides access to a GitHub account where the software can be downloaded for local use. The website version can only use pre-loaded RNA-seq data from our analysis, whilst the local version allows users to process their own RNA-seq data. The front-end of NGS Graph Generator was written in PHP and HTML, together with a Python daemon whose responsibility it is to execute the jobs as created on the website. The system requires a MySQL database to which access is configured through *dbSettings.json* file. Executing daemon will create tables if they do not already exist, but further supplementary settings must be provided to suit the installation environment. These settings are made via a MySQL command line or an SQL script such as the supplied *init.sql*. The prerequisites needed before running this front-end are Python, PHP, and MySQL.

### 2.5.2 Public repository

The pipeline is uploaded on the public repository GitHub <https://github.com/systems-immunology-roslin-institute/ngs-graph-generator> to allow a user to download and run the pipeline on a UNIX server system.

### 2.5.3 Amazon machine image (AMI)

Cloud computing has become a powerful technology platform to perform large-scale and complex computations. It eradicates the need to sustain high-priced computing hardware, software and dedicated space. A massive growth in the amount of data generated through cloud computing has been reviewed (Hashem et al., 2015). Cloud-based storage and analysis are becoming a popular alternative for NGS data processing (Stein, 2010), due to the relative flexibility, scalability, and affordability (Baker, 2010; Dudley and Butte, 2010; Dudley et al., 2010; Marx, 2013). Genomics cloud computing providers (i.e. Google Genomics and Amazon Web Services), offer services using various models and pipelines. Further detailed cloud-based bioinformatics workflow platforms offer additional capabilities.

Cloud computing offers an effective solution to the physical infrastructure of any software application. The ability to promptly acquire, setup, and scale physical resources is a vital feature provided by Amazon's Cloud through the Infrastructure-as-a-Service or IaaS layer. A part of this chapter uses the IaaS layer through a Software-as-a-Service (SaaS) web application to create a virtual network-based pipeline for users. This will allow users to create, connect and terminate instances, hence providing an immediate service to hold multiple instances. It exploits a flexible architecture to effortlessly scale the resource consumption based on the data usage. The Amazon Web Service (AWS) APIs are used to connect with the IaaS layer of cloud and add computational demands based on a pay-per-use basis. The Amazon Cloud provides the computation using the Elastic Cloud Compute (EC2) product. It offers several machine images to create virtual computational instances ranging from micro to High-CPU. This instance supports several operating systems (OS) such as Windows and Linux. Therefore, the Amazon Machine Images (AMI) format of this pipeline was created using the Amazon Web Services (AWS) to reach a wider audience of users.

## 2.6 Discussion

RNA-seq technology has rapidly evolved into standard methodologies for the identification of isoform variation in various research. This has resulted in the rapid development of bioinformatics tools to visualise splice variation and identification of isoform expression of these data. However, the visualisation for the analysis of transcriptomes data still represents an unresolved challenge for researchers looking for alternative visualisation aid. The data generated is large and the tools for its assembly, analysis, and visualisation are still under development. Furthermore, while a few have developed a fundamental work of network analysis using repetitive DNA sequence (Novák et al., 2010), there has been no emphasis on approaches for network-based visualisation of RNA-seq data. The aim of this chapter is to describe the development of network-based visualisation pipeline for analysis RNA-seq data. This pipeline, NGS Graph Generator pipeline provides a new perspective for the visualisation and analysis of RNA-seq data.

SeqGrapheR (Novák et al., 2010) is a graph-based visualisation tool of the DNA sequence cluster, however, inability to overlay additional information for visualising RNA-seq data such as isoform and exon information. Tools such IGV (Thorvaldsdóttir et al., 2013), UCSC genome browser (Kent et al., 2002), GBrowse (Donlin, 2002) and Ensembl (Yates et al., 2016) allow for the visualisation of sequencing reads mapped to a reference genome, mutations i.e. SNP and characteristics profile i.e. ChIP-seq and DNA methylation. While constructive for various applications, ability to visualise splice variation is basically limited by the representation of junctions read to identify isoform expression i.e. Sashimi plots (Katz et al., 2013). Overall, whereas all the tools described above can be practical for genome-wide analysis, a pipeline that allows RNA-seq data to be explored using network approaches is still lacking.

Network-based visualisation tools such as Cytoscape (Shannon et al., 2003) or BioLayout *Express*<sup>3D</sup> (Freeman et al., 2007; Theocharidis et al., 2009) allow one to visualise a 2-dimensional or 3-dimensional data, respectively with the representation of correlation shown as nodes connected by edges. These tools offer more flexibility

on how the data is visualised and represented which typically overlay gene expression data. This has become interesting for us to exploit the function in these tools, especially to visualise RNA-seq which overlay to isoform and exon annotation data. Therefore, I have developed a network-based visualisation of RNA-seq data as a tool to explore the data and aid researchers with the exploration and identification of their RNA-seq data, especially identification of splice variants. This approach allows one to explore data, the user to enter specific gene names to construct gene network transcript. A gene of interest can be visualised after the data has been processed through this pipeline. The network can help one to determine whether a transcript exists as a single linear form or is expressed as multiple isoforms. However, as the pipeline is not able to concurrently visualise multiple genes or loci, the functionality can impose limits to visualise data at a genome-wide level. The NGS Graph Generator is designed to allow the user to remove redundant sequence reads of the highly expressed genes, i.e. *TUBA1C* and *GAPDH* (see Chapter 3, Section 3.3.3.1) while still allowing users to visualise the network in different size of nodes relative to sequence coverage.

Through my use case to optimise network structure, I show that network-based visualisation can help to address this need. While the major aim of the development of network-based visualisation of RNA-seq was to provide a seamless pipeline for preparing data, it also allows the user to set a parameter prior to network construction. It also will allow users to generate layout file of their RNA-seq data. Given incredible amounts of transcriptomic data being generated now, it is crucial that researchers have an alternative approach to explore their RNA-seq data especially in the identification of splice variation. Therefore, cell cycle sample of human fibroblast (Freeman et al, unpublished) and human tissue (Fagerberg et al., 2013) data types are provided from the NGS Graph Generator web-based application. The network transcript (i.e. *COL5A1*) can easily be visualised in BioLayout *Express*<sup>3D</sup>. From the node class option, users can select to display data from other isoforms (e.g. alternative splicing isoform). Moreover, the node class of BioLayout *Express*<sup>3D</sup> could be added to incorporate other information such as regulatory relationships or additional types of data e.g. SNP, DNA methylation. By maintaining the NGS Graph Generator code in a public repository (GitHub), this will



facilitate the implementation of new functionality in the future release versions. This pipeline has been used for the exploration of transcript network in subsequent Chapter 3 (human fibroblasts) and Chapter 4 (human tissues). In summary, although no single tool reveals the splice variation, through a network approach will allow users to visualise and identify alternative splicing of their data, making it easier to potentially understand the biological to which it may be relevant.

## Chapter 3 - Network-based visualisation and analysis of RNA-seq data

### 3.1 Introduction

Networks are increasingly used in biological research, in particular for plotting experimentally or computationally derived relationships between genes and proteins. Networks consist of nodes connected by edges (lines), where nodes usually represent an entity and edge a relationship between them (Miller et al., 2010). Networks are now employed widely in biological research as a means to analyse a wide variety of complex data types, to infer functional associations based on neighbourhood analyses or clustering, and to model pathways (Barabasi and Oltvai, 2004).

Network visualisation of DNA sequence data has been little explored. Novák et al. (2010) first introduced the idea of using a network-based method to visualise DNA assemblies. As with the approach described here, in their studies nodes represent individual reads of DNA sequence, whilst edges denote a sequence similarity i.e. homology between reads above a defined threshold. Overlaps must be pre-computed by a series of computationally intensive pair-wise sequence alignments (Pop, 2009), and this represents the first step in building a network from such data. After generating this matrix of similarity scores from an all-versus-all read comparison, read similarities exceeding a specified threshold are used to define network edges. Their study focused on the characterisation of the repetitive regions of plant genomes (pea and soybean). It was argued that the complex topology and diversity of the networks produced could be used to better analyse the variability and evolutionary divergence of repeat families, as well as to discover and characterise novel elements. Network visualisation, however, was in the form of generating a PDF file limiting the opportunity for data exploration. Here, network-based visualisations of sequencing data are further explored, applying the fundamental principles first described by Novak et al. (2010) to RNA-seq data. The aim of this work has been to develop a complementary approach to understanding differences between RNA assembly algorithms as well as to better understand transcript structure and splice

variation. In this chapter, a novel method for the visualisation of RNA-seq data using the network analysis tool BioLayout *Express*<sup>3D</sup> was developed. In so doing, a platform that supports the improved interpretation of complex transcript isoforms is presented. This approach will be useful in the exploration and discovery of new biological insights from sequence data.

The objectives of this chapter were:

- a) To process the newly generated RNA-sequencing data derived from human fibroblasts at different points following serum starvation synchronisation (map to the genome, assemble transcripts and generate normalised read counts).
- b) To optimise parameters for read overlap to minimise the number of edges and better display features. To explore different graph layout algorithms.
- c) To explore various cell cycle gene networks by generating a network-based layout using RNA-seq data and identify the structure of networks across samples.
- d) To compare network assemblies of transcripts with those using conventional assembly methods highlighting the advantages or disadvantages of the approach.

## 3.2 Methods

### 3.2.1 RNA-seq data used for these studies

Four samples of RNA-seq data were generated from serum-starved human fibroblasts (NHDF) (0 h) and three-time points (12, 18 and 24 h) following serum refeeding during the cultures partially synchronised entry into the cell cycle. This work was performed by Dr Mark Barnett and David Chen from Freeman Lab. RNA sequencing was performed on the Illumina HiSeq 2500 platform (Illumina, San Diego, California, USA) with 100 bp paired-end sequencing in a rapid mode according to the manufacturer's recommendations. RNA sequencing was carried out by Ark Genomics, at The Roslin Institute using the TruSeq™ RNA Sample Prep Kit (Illumina). Briefly, poly-(A) RNA was isolated from total RNA using oligo d(T) coupled to magnetic beads and fragmented using divalent cations to produce fragments of an average 180 bases in length. Fragmented RNA was reverse transcribed using a random primer and Superscript II enzyme (Invitrogen, Carlsbad, California, USA). A single-stranded DNA template was used to generate double-strand cDNA using RNase H and DNA polymerase. The resulting double-stranded cDNA was blunt-ended using T4 DNA polymerase prior to the addition of an adenosine base to assist ligation of the sequencing adapters. Flowcell preparation was carried out according to Illumina protocols; the libraries were denatured and diluted to a concentration of 15 pM for loading into the flow cells.

### 3.2.2 RNA-seq data processing

RNA-seq data were processed using Kraken, a set of tools for quality control and analysis of high-throughput sequence data developed by Davis et al. (2013) (Davis et al., 2013). This FASTQ format can directly use the Kraken pipeline which includes reaper, filter and annotates the data. The Kraken package downloaded from (<http://www.ebi.ac.uk/research/enright/software/kraken>) version 13-274 (11 October 2013). The analysis using Kraken has been incorporated in the SequenceImp pipeline which contains all the tools (Reaper, tally, BowTie and various R BioConductor packages). The filtered RNA-seq data were processed using packages in the TopHat and Cufflinks. The data were aligned to the Homo sapiens reference genome (GRCh37.71) (Flicek et al., 2014) using Bowtie with Ensembl annotation

file (gtf file). Read alignment processed within TopHat v2.0.9 (Trapnell et al., 2009) to identify loci and splice junctions. Then, Cufflinks v2.1.0 (Langmead et al., 2009) was run with the Ensembl annotation file to estimate the relative abundance of the transcripts in the data. The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) metrics at the gene, and transcript level was used for subsequent analysis of differential expression and sample variation. The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. Primary visualisation of the data was performed using IGV to visualise the reads mapped onto the reference genome in a certain locus or gene across samples.

### 3.2.3 Network layout

BioLayout *Express*<sup>3D</sup> had for a long time used a version of the Fruchterman-Reingold (F-R) algorithm (Fruchterman and Reingold, 1991) for the layout of large networks derived from biological data, such as protein interaction/similarity networks and correlation networks derived from genome-wide expression data. Whilst implementation of the F-R algorithm was capable of producing layouts for many large networks the unusual topology of DNA/RNA sequence networks necessitated a new layout approach. The Fast Multipole Multilevel Method (FMMM) (Hachul and Jünger, 2005) algorithm was examined and shown to be well suited to the layout of these types of network. The FMMM algorithm was re-implemented in Java from the Open Graph Drawing Framework (OGDF), including the novel ability to perform network layout in 3D space and incorporated into the BioLayout code base. This work carried out by Tim Angus from Freeman Lab.

Initially, a ‘perfect’ overlap network was generated by assuming 100 ordered reads (nodes) where each read overlaps the previous read by 95% of its length (i.e. edge weight between adjacent nodes is 0.95). In this paradigm, the first read would overlap successive reads by 5% less each time, sharing only 5% similarity to read 20 and no similarity to subsequent reads. The second file of this type was prepared to represent two splice variants, where one variant was identical to the first network but a second variant included a 50 read addition where similarities started to branch off after the first 50 reads from the first. Following examination of these synthetic

‘transcript’ graphs, a series of tests were performed using data mapping to collagen type 1 V alpha 1 (*COL5A1*), a long (8.5 kb) and highly expressed genes in human fibroblasts encoded by a single transcript species.

### 3.2.4 Optimisation of read comparison parameters

In the approach described here, quality of network visualisation is primarily determined by the depth of sequencing of the target gene, which in turn is based on its relative expression level. Therefore, I set out to determine generalised default parameters for network visualisation exploring the threshold settings for percentage similarity ( $p$ ) and percentage length coverage ( $l$ ) between reads.

Briefly, there were two types of data used in these analyses; ‘synthetic’ and ‘real’ RNA-seq data. In the first instance, a synthetic RNA-seq dataset was generated from a 5-kb sequence of *COL5A1* mRNA. A read length of 100 bp was used with reads being spaced at 1, 5, 10 or 20 bp intervals. This mimics real data where read-depth varies. These represent perfect data where the spacing of reads is regular, and sequence reads do not contain sequencing errors or SNPs. To explore the threshold settings, for each set of reads, a read-to-read comparison with various ranges of parameter percentage length coverage,  $l=10-100$  (percentage similarity was set fixed at 100% as there were no sequencing errors or SNPs in these data) were performed.

Whilst for ‘real’ data, four transcripts with anomalous network structures were identified. These networks were selected because they either possessed a splice variant (*CASC5* and *CENPE*), internal homology (*SGOL2*) or an issue with assembly due to the presence of an overlapping gene on the opposite strand (*CENPO*). All the reads associated with these genes were placed in a single read bin and used to determine the generalised optimal setting for network construction by performing a read-to-read comparison at various settings of the  $p$ , and  $l$ , parameters. At each setting for both data sets, the number of nodes, edges, and components were recorded for the resultant networks.

### 3.2.5 Collapsing of redundant reads

Highly expressed genes are represented by a high number of reads. In these instances, there can be a high degree of redundancy in read coverage i.e. exactly the same sequence may be present in the data many times. This makes the read-to-read comparison step unnecessarily time-consuming and the resultant network difficult or impossible to visualise, whilst adding little or nothing to the network's readout. Using *Tally* from the Kraken package, multiple identical reads are mapped to a single identifier a FASTA file being produced where the identifier is linked to the number of occurrences of that specific sequence. In the visualisation of collapsed node networks, a single node is used to multiple reads and the diameter of a node is proportional to the original number of occurrences of the reads it represents.

### 3.2.6 Analysis of the network structure

Initially, a set of 550 genes whose expression was up-regulated as fibroblasts entered into S, G2 and M phase of the cell cycle (0, 12, 18 and 24 h after being refed serum) were chosen to be examined. A network derived from the 24 h data was plotted for each gene in each case using generalised MegaBLAST parameters ( $p=98$ ,  $l=31$ ). Where the topology of a given gene network is relatively simple an explanation of its structure required only the overlay of individual transcript exon information in order to identify splice variant(s) represented. In other cases, more detailed analyses are required.

### 3.2.7 Validation of splice variant using RT-PCR

To validate the existence of splice variants predicted by network analyses, RT-PCR of candidate splice variants was performed. Total RNA from human fibroblasts used for the RNA-seq experiment was reverse transcribed using SuperScript III (Invitrogen, Carlsbad, CA) as follows; 50  $\mu$ M of oligo (dT)<sub>20</sub>, 10 mM dNTPs: Heat the mixture to 65°C for 5 minutes and incubated in ice at least 1 minute. 5X buffer, 0.1 M RNaseOut (Invitrogen), 1  $\mu$ g of total RNA were added. Incubate the mixture for 2 minutes at 42°C, and then SuperScript III was added, followed by 60 min at 50°C and inactivated the reaction at 75°C for 15 min. Primers were designed using Primer3 software<sup>33</sup> to amplify the region for validation of the splice variant. For

*LRR1*, a pair of primers was designed to amplify three splice variants as suggested from the network visualisation while for *PCMI* two pairs of primers were designed across two different splice variant locations. For *LRR1*: Forward primer 5'-TGTTGAGCCTCTGTCAGCAG-3' and reverse 5'-GTGTGGGCAACAGAATGCAG-3' that span exon 3, 4 and 5. *PCMI* set 1: Forward primer 5'-TCTGCTAATGTTGAGCGCCT-3' and reverse 5'-TGCAGAGCTAGAAGTGCAGC-3' and *PCMI* span exon 7, set 2: Forward 5'-ACGGAAGAAGACGCCAGTTT-3' and reverse 5'-AGCTGCAGCTCATGGAAGAG-3' span exon 24. PCR was carried out by preparing a master mix that containing 34.6% sucrose,  $\beta$ -mercaptoethanol (1:10 in 0.1x Tris-ETA buffer), 10 mM dNTPs, 12  $\mu$ M pre-mixed each primer pair, 10X reaction buffer (1 M Tris-HCl pH 8.8, cresol red solution, 1 M MgCl<sub>2</sub>, (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>), dilution buffer (T0.1E, cresol red solution and 4M NaOH), Taq Polymerase (Invitrogen) in PCR thermocycler machine. The products were amplified through 35 cycles (92°C, 30 secs; 60°C, 90 secs; 72°C, 60 sec) and run on a 2% agarose gel. Gels visualised by UV on a Syngene transilluminator and recorded using the GeneSnap acquisition software (Syngene, Synoptics, Frederick, MD).

### 3.3 Results

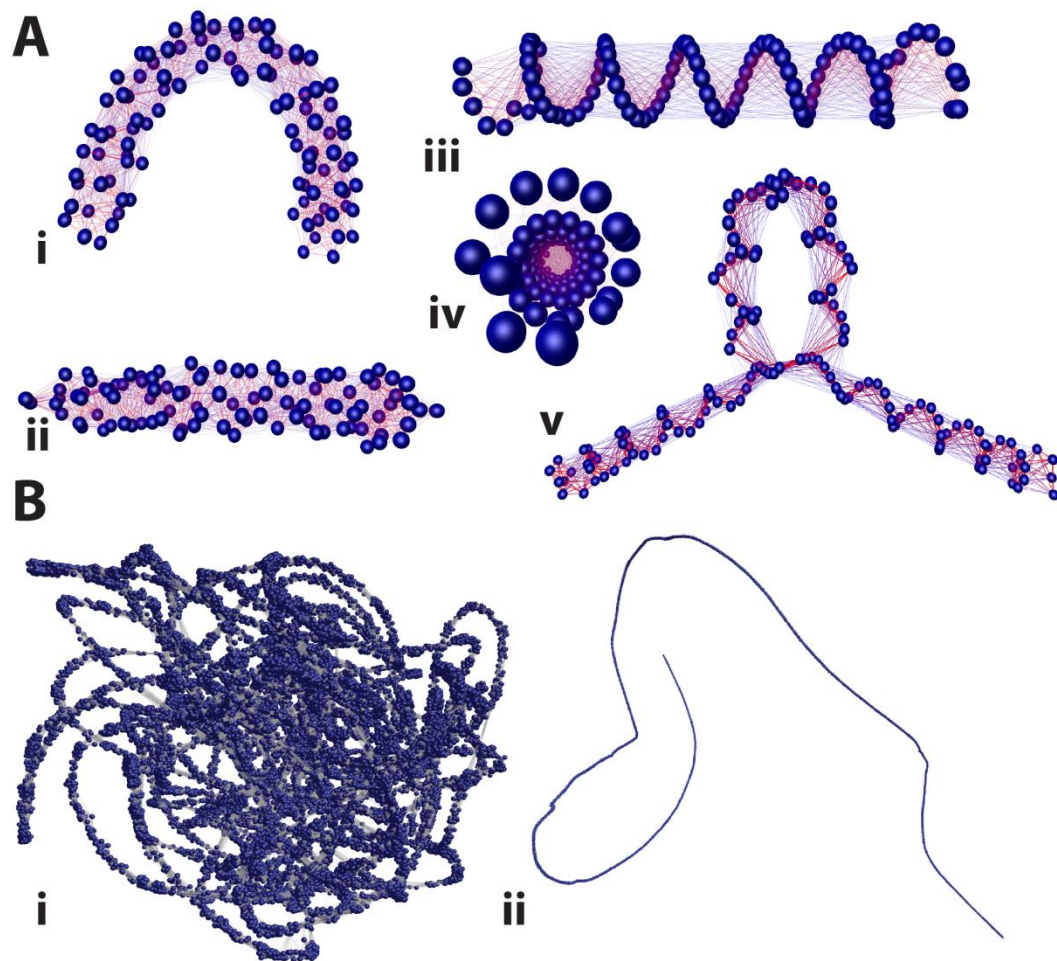
The principle of rendering sequence data as a network has been discussed and illustrated in Chapter 2. A pipeline to create such networks from RNA-seq data was developed, where the outputs can be visualised in IGV or as a network within BioLayout *Express*<sup>3D</sup>. In this chapter, RNA-seq data from four human fibroblast samples that were sequenced following serum starvation when the cells undergoing partially synchronised cell division were generated. Paired-end cDNA libraries for each RNA sample were prepared and sequenced. A summary of RNA-seq sequencing data after filtering using the Kraken pipeline and alignment using TopHat is provided.

#### 3.3.1 Optimisation of network visualisation 'perfect' overlap data

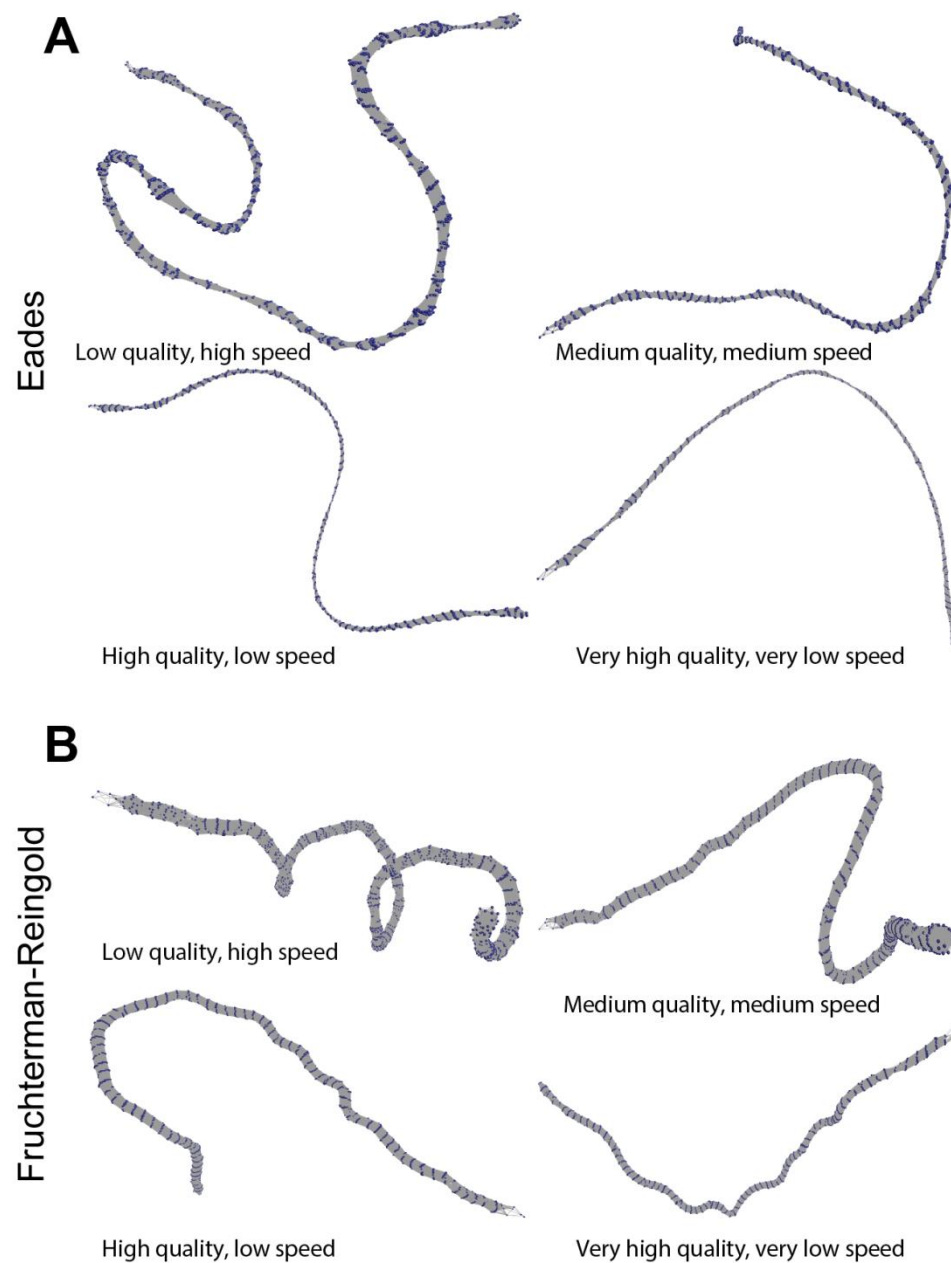
An optimal layout is crucial to the interpretation of network structure. Initially, the layout of RNA-seq data (NHDF 24 h post-serum refeeding) for *COL5A1* was



explored; *COL5A1* has 66 exons, 8,471 bp long transcript that is highly expressed by fibroblasts (40,170 reads mapped to this gene in this sample). Initial studies of RNA-seq network visualisation for *COL5A1* within BioLayout *Express*<sup>3D</sup> demonstrated the Fruchterman-Reingold (F-R) layout algorithm implemented within the tool performed poorly on these types of networks producing difficult to interpret knot-like structures (**Figure 3.1B**). It was clear that if this approach was to provide interpretable results, an improved network layout would be required. Following examination of the available algorithms for network layout, the FMMM (Hachul and Jünger, 2005) algorithm was incorporated into the tool's code base, enabling layout in a 3D environment in the process. Implementation of the FMMM provides an interface where the force model can be selected (F-R or Eades) and includes various settings that offset layout quality versus the speed of network layout. The higher the quality i.e. the 'straighter' the network of a linear sequence becomes, but the more time the algorithm will take to run and vice versa (**Figure 3.2A and B**). Network visualisation of a theoretical matrix of 100 reads where consecutive reads overlap by 95% demonstrates that they show a corkscrew-like appearance at a local level (**Figures 3.1C and D**). Indeed, the more edges present (defined by the stringency of similarity cut off) the tighter a network is coiled. This unique feature of overlap networks can also be observed in networks generated from RNA-seq data, particularly when the depth of sequencing is high. When a splice variant is introduced the alternatively spliced exon is seen to loop out (**Figure 3.1E**) as reads at alternative splice junctions are connected but pulled in different directions.



**Figure 3.1: Optimisation of the network layout.** (A) The layout of a perfect overlap matrix consisting of 100 ‘reads’ where consecutive reads overlap by 95% and a  $\geq 5\%$  similarity has been used as the minimum threshold for defining edges. In example **Ai** a modified Fruchterman-Reingold layout was used to the layout of the network, in **Aii** the FMMM algorithm using the Eades force model and a ‘Low Quality, High Speed’ setting and **Aiii** is the same as **Aii** but the ‘High-Quality, Low-Speed setting’ was used. **Aiv** is an end in view of **Aiii** illustrating the corkscrew structure of the graph. **Av** is a visualisation of a synthetic alternatively spliced transcript. (B) Network visualisation of *COL5A1*, which is highly expressed gene in human fibroblasts. **Bi** network layout using the Fruchterman-Reingold algorithm as originally implemented within BioLayout *Express*<sup>3D</sup> and **Bii** layout of *COL5A1* using FMMM algorithm (Eades force model and the High-Quality, Low-Speed setting).

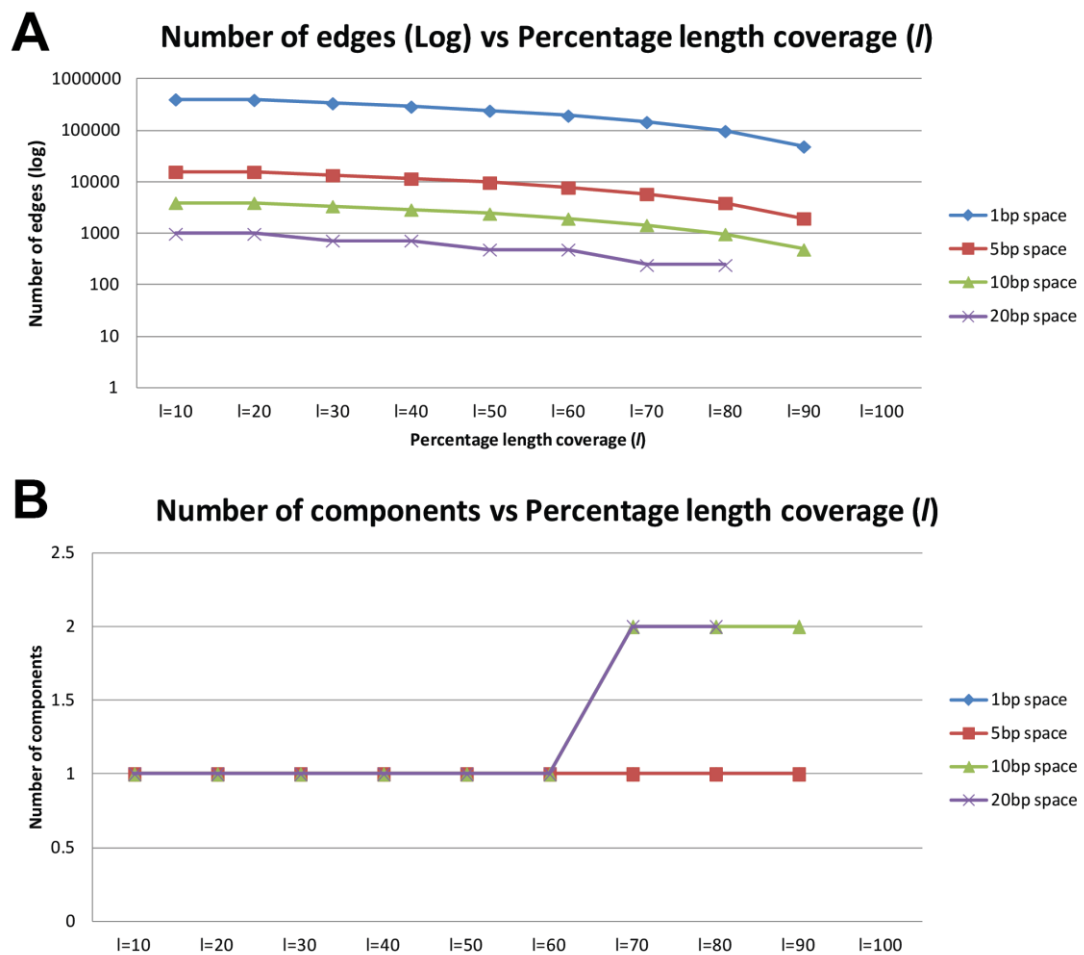


**Figure 3.2: Network layout - quality vs. speed.** Network layout of *BUB1* (A) force model Eades and (B) force model Fruchterman-Reingold in FMMM algorithm utilising different settings implemented within BioLayout *Express*<sup>3D</sup>. ‘Very high quality, very low speed’ is the most linear while in the mode of ‘Low quality, high speed’ is the less linear network were obtained for both force models. For model Eades ‘Very high quality, very low speed’, it took 8.93s to layout the network layout on a machine 3.2 GHz 32.0 GB RAM 64-bit Windows OS. For ‘High quality, low speed’ (3.28s), ‘Medium quality, medium speed’ (2.25s) and ‘Low quality, high-speed’ (1.91s). Whilst for Fruchterman-Reingold force model which took 1.90s for ‘Very high quality, very low speed’ setting, 1.83s, 1.45s and 1.71s for ‘High quality, low speed’, ‘Medium quality, medium speed’ and ‘Low quality, high speed’ respectively.

### 3.3.2 Optimisation of read-to-read comparison similarity score threshold

Another factor in defining the network structure is the method used to define an edge i.e. the similarity score threshold between a pair of reads. Two MegaBLAST parameters are adjustable, the length over which a similarity search is performed ( $l$ ) and the percent similarity required over this length ( $p$ ). If the thresholds for these settings are too high networks may fragment, and fine structure will be lost, if too low the underlying structure may be obscured and it will greatly increase the memory footprint of a network and layout time due to an excess number of edges.

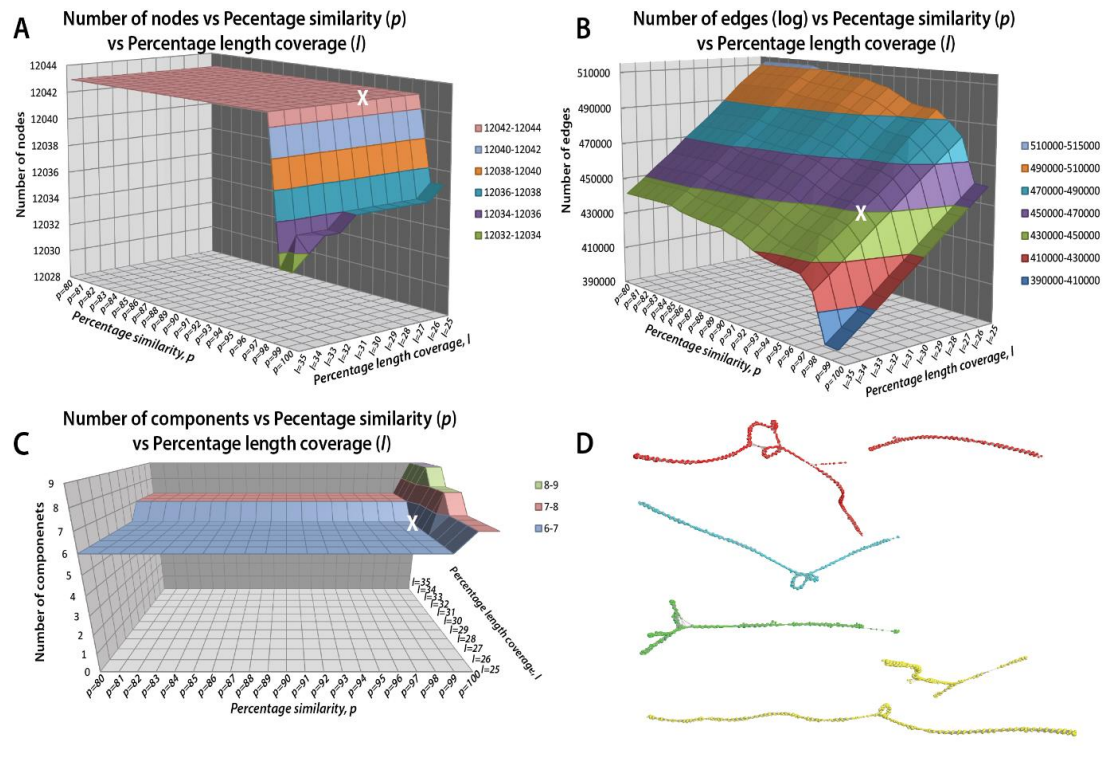
In order to perform this analysis, two different sets of data were examined; ‘synthetic’ and ‘real’ data. 5 Kb of *COL5A1* gene was selected from the human fibroblast data on the fact it produces a linear transcript without any anomaly, and it is highly expressed thus produces a convincing resultant network. After observing the ‘synthetic’ data, the number of edges steadily decreased with increasing value of  $l$ . Hence, this result was insufficient to resolve the best parameter for RNA-seq data. While for the number of components reached the optimum number of all sets of sequences at  $l=60$  before it was necessary to break the graph into more than one component (**Figure 3.3A**). This experimentation is significant to find the optimal parameters for ‘synthetic’ data and to infer the pattern of nodes, edges, and components. For ‘synthetic’ data, the percentage length coverage,  $l$ , should be less than 60 to be optimised otherwise the graph will be fragmented into more than one component to determine to build a network-based on the fact that the network should be retained as a linear network to be an optimum parameter (**Figure 3.3B**).



**Figure 3.3: Optimisation plot of ‘synthetic’ data of *COL5A1*.** Read-to-read comparison was performed using MegaBLAST with two parameters; a fixed percentage length similarity ( $p$ ) and percentage length coverage ( $l$ ). **(A)** Log number of edges for each set of sequences steadily decrease over the percentage length coverage ( $l$ ). Three sets of sequences (1, 5 and 10 bp space) show the lowest number of edges at  $l=90$  while a set of sequence of 20 bp space shows the lowest at  $l=80$ . **(B)** The number of components for all set of sequences is optimum i.e. 1 from  $l=10$  to  $l=60$  and further breaks into two components when it reaches  $l > 60$ .

In order to get the generalised value of these parameters, four transcripts were selected from the human fibroblast data based on the fact that they represented a range of sequencing depths and complex network topologies. After exploring the two variables for thresholding the read-to-read similarity score, a percentage length similarity  $p=98$  and percentage length coverage  $l=31$  were chosen, as reasonable generalised values for these variables (**Figure 3.4A, B, and C**). For these values, the networks for *CENPO* and *SGOL2* have just one component, but retain their higher

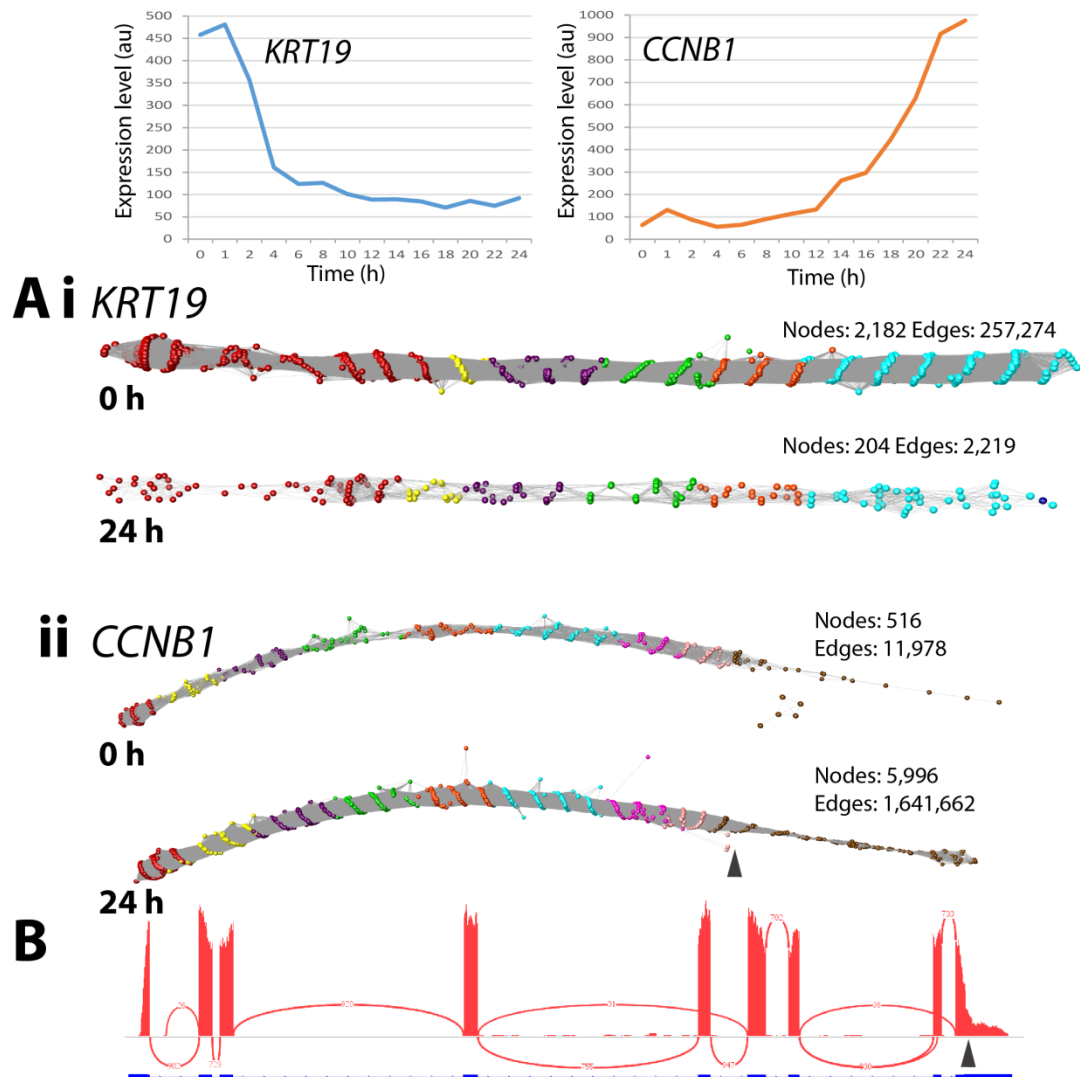
order structure, while networks for *CASC5* and *CENPE* have two components (as they do have at even lower stringency threshold values) but again retain their higher order structure (**Figure 3.4D**).



**Figure 3.4: Optimisation plot of ‘real’ data of four different complex genes (*CENPO*, *SGOL2*, *CASC5*, and *CENPE*).** This optimisation was performed at the same previous experiment for ‘synthetic’ data. Mark ‘X’ in each plot (A), (B) and (C) shows optimum parameter for network-based visualisation. The optimum parameter is a percentage similarity ( $p=98$ ) and percentage length coverage ( $l=31$ ) with maximum thresholds that maintained node numbers and interesting network structures, whilst where possible keeping the networks as single network components. (D) Network-based visualisation of four complex gene structures was generated using the optimum parameter ( $p=98$  and  $l=31$ ). At the optimum parameter, *CENPE* (red) and *CASC5* (yellow) generate two component networks while *SGOL2* (blue) and *CENPO* (green) generate one component network. Using these parameters, interesting features of the network can be retained without losing any information.

### 3.3.3 Network visualisation of transcripts

Having default general settings for network construction, 550 gene networks from the NHDF data of genes upregulated as cells undergo mitosis were manually examined. In most of cases, the networks of these genes suggest that a single transcript was expressed by fibroblasts. Similarly, to *COL5A1* the networks were single linear strings comprised of a single or multiple disconnected component depending on coverage (**Figures 3.5Ai** and **3.5Aii**). In these cases, no obvious secondary structure was present and arguably little had been learned by visualising these data as networks. Even with linear graphs transcript variance can be observed. The single linear network representing *CCNBI* there is clear evidence that two transcript isoforms were expressed by the fibroblasts, one of which was truncated at the 3' end. This was manifest by the fact that there were fewer reads present at the 3' end of the network, the fall off in read density occurring at the point where a known variant occurs (ENST00000505500), and the coiled structure broke down beyond that point. This decrease in reads at the 3' end of *CCNBI* is also visible in the standard IGV (Sashimi plots) (**Figure 3.5B**).

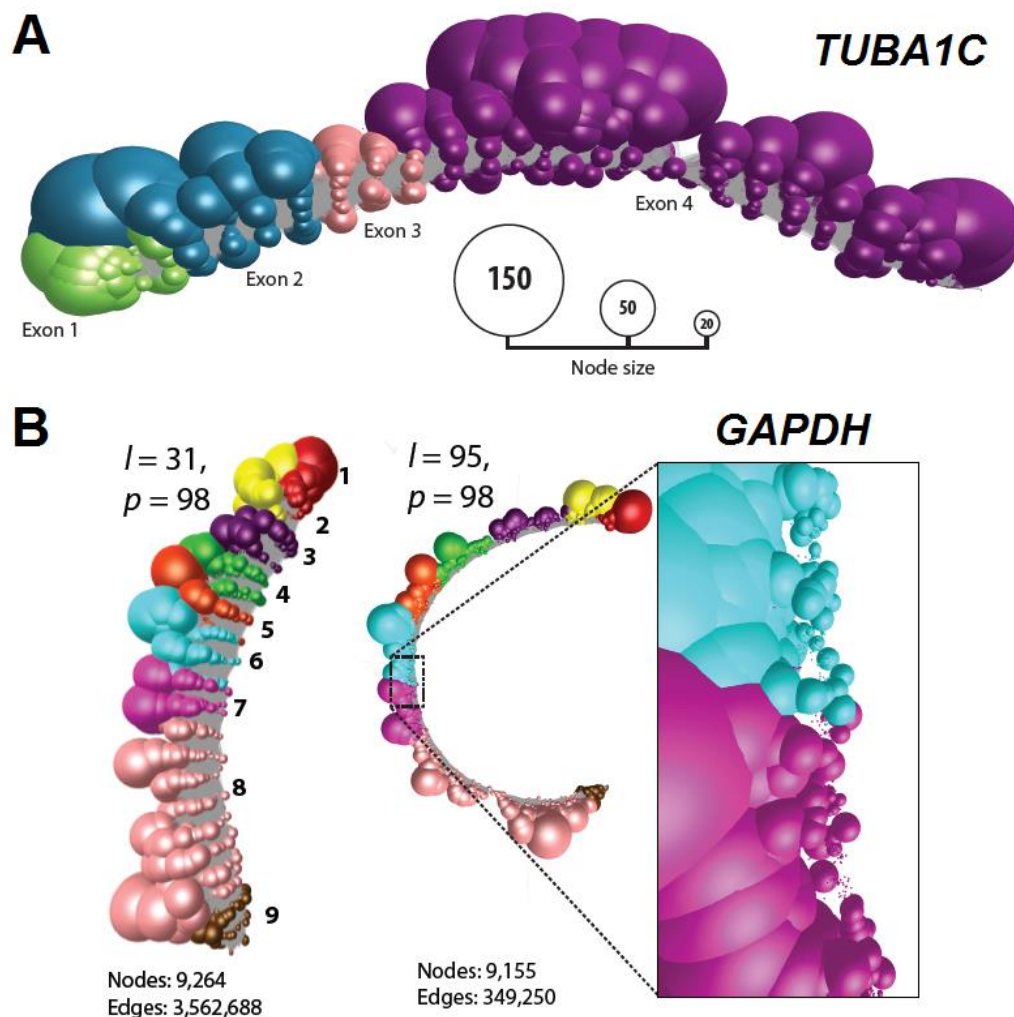


**Figure 3.5: Typical networks of RNA-seq data derived from linear transcripts.** (A) Examination of a wide range of RNA-seq assembly networks derived from the human fibroblast expressed genes reveals most networks are linear unbranched structures. Shown here are two such transcripts for (Ai) *KRT19* and (Aii) *CCNB1* (overlay node colour derived from ENST00000361566 and ENST00000256442, respectively). Top: expression profile of the two genes as measured by microarray analysis of the time-course of transcriptional events following serum refeeding (data not shown). Expression of *KRT19* is rapidly down-regulated and whilst *CCNB1* is up-regulated. This differential expression is evident from the networks with the number of nodes decreasing or increasing by approximately 10-fold in the 0 h derived vs. the 24 h derived RNA-seq data. It is interesting to note that in the *CCNB1* networks there is a rapid decrease in the density of nodes within exon 9 at both time points (marked by arrow). This corresponds to where the IGV view in (B) also shows a decrease in the density of reads and corresponds to *CCNB1* transcript (ENST00000505500) that exhibits a truncated exon 9 at this position.



### 3.3.3.1 Network reduction

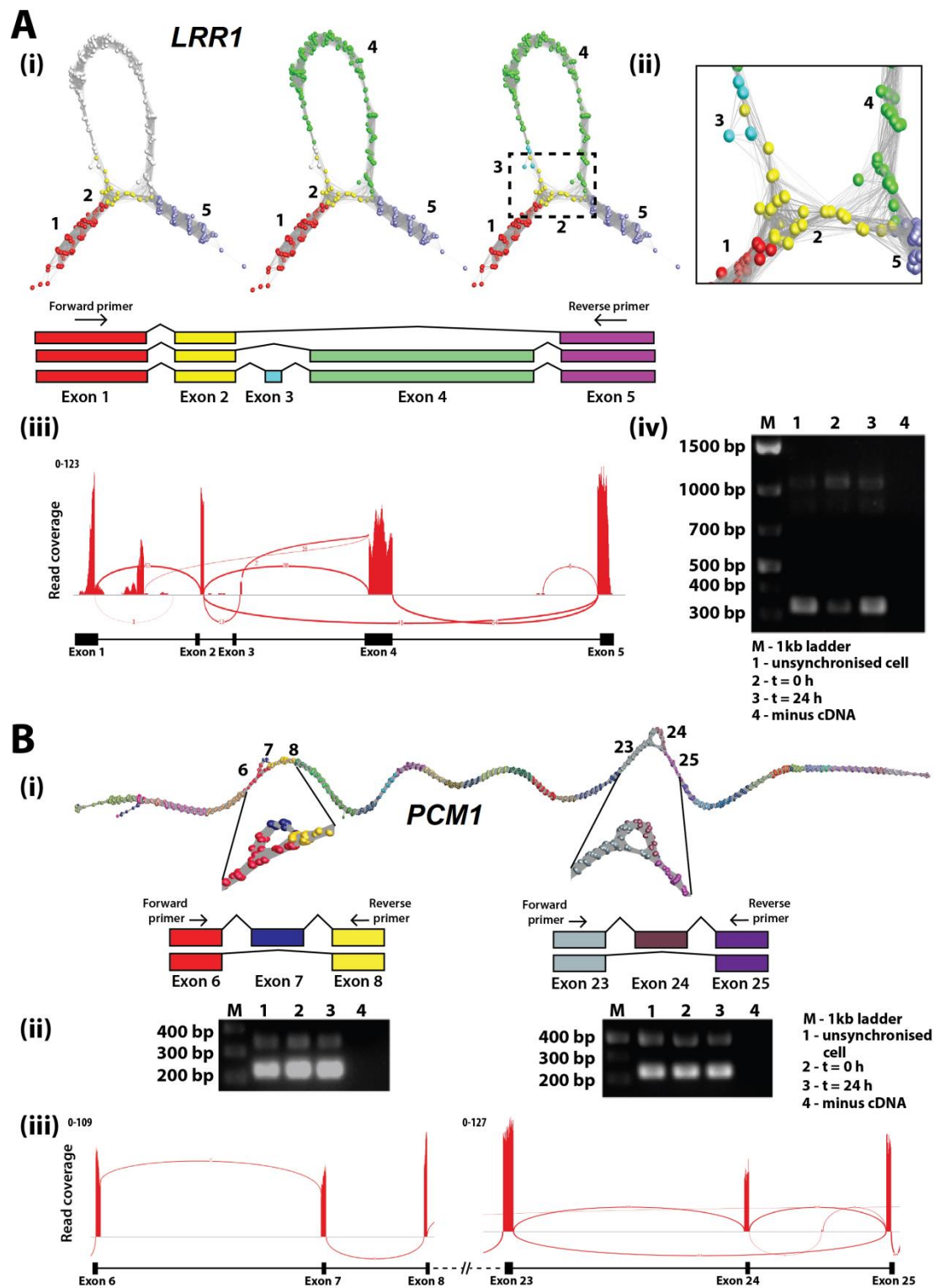
In some instances, large numbers of reads mapping to a gene mean that network visualisation is not possible due to the sheer number of nodes and edges need to represent the data as described above. For instance, in the 24 h serum, re-fed fibroblast samples the highly expressed genes *TUBA1C* and *GAPDH* had 38,294 and 59,998 reads mapping to them respectively. Network reduction is a process whereby identical reads are collapsed down to and represented by a single node; the size of the node being proportional to the number of nodes it represents. In the case of *TUBA1C*, this reduces the number of nodes from 38,294 to 6,511 nodes (**Figure 3.6A**), whilst the number of edges is reduced from 90,340,179 to 1,779,069. In the case of another gene *GAPDH*, the reduction in nodes is from 59,998 to 9,264, whilst the number of edges is reduced from 208,221,932 to 3,562,688 (**Figure 3.6B**). The reduced network for *GAPDH* is also shown generated at two different BLAST threshold settings. On the left, the network is generated at the default BLAST settings  $p=98$ ,  $l=31$ , the second using more stringent settings. Such is the depth of sequencing of this gene that even using BLAST setting of 98% similarity over 95 bp of length the network still forms one component where the number of edges is reduced by approximately 90% but the number of nodes by less than approximately 0.1%. At this higher stringency BLAST setting, the network uncoils exposing small nodes representing unique reads due to sequencing errors.



**Figure 3.6: Read unification of highly expressed genes *TUBA1C* and *GAPDH* in human fibroblasts.** (A) Network representation of *TUBA1C* after read unification. The number of reads mapping to *TUBA1C* was 38,294 in the 24 h serum-refed fibroblast sample. After read unification, this was reduced to 6,511 unique reads. When networks are collapsed down to unique sequences, node size is proportional to the number of individual reads represented, and nodes have been coloured according to the exon onto which they map. The *TUBA1C* transcript model here matching the network being ENST00000301072, a 3,001 bp transcript encoding a 449-amino acid protein. (B) Unification of *GAPDH* after read unification as described above but shown at two different thresholds of BLAST scores. On the left using the ‘standard’ threshold of 98% similarity over 31 bp, on the right 98% similarity over 95 bp, increasing the edge threshold dramatically reduces the number of edges, whilst the number of nodes is barely affected. It also “opens up the structure” and when edges are removed a large number of tiny nodes representing unique reads due to sequencing errors can be observed (inset). The *GAPDH* transcript model here matching the network being ENST00000396856, a 1,266 bp transcript encoding a 260-amino acid protein.

### 3.3.3.2 Splice variant network structure

In normal human dermal fibroblast (NHDF) networks approximately 5% of the transcripts studied exhibited complex topologies. In these instances, the underlying reasons for these unusual structures were investigated. LRR1 (leucine-rich repeat protein 1) interacts with TNFRSF9, a member of the tumour necrosis factor receptor (TNFR) superfamily. In the case of the *LRR1* network, a single loop was observed corresponding to the two known transcript isoforms for this gene. The first transcript (ENST00000298288) contains four exons while the other transcript (ENST00000318317) present in these data has only three exons and skips exon 3 (**Figure 3.7A**). There was also evidence for the presence of a nonsense-mediated decay product (ENST00000554869) as a small number of reads mapped to exon 3 specific to this transcript. The network transcript for *PCMI* (pericentriolar material 1), a 6,075 nucleotides gene containing 36 exons with seven known protein-coding variants, provided evidence for two splicing events when expressed in fibroblasts. One loop was indicative of the splicing out of exon 7 and the other exon 24 suggesting the presence of transcripts ENST00000517730 and ENST00000522275 respectively, in addition to the main isoform of this gene (ENST00000325083). RT-PCR confirmed the splicing events for *LRR1* and *PCMI* genes predicted by the network-based analysis. The visualisation of splice variants was also supported in the Sashimi plots for these genes but even in these relatively simple examples of splice variation, the plots can be challenging to interpret especially in the case of *LRR1* (**Figure 3.7B**).



**Figure 3.7: Splice variant visualisation and confirmation (A) Confirmation of splice variant *LRR1* using RT-PCR assay. (i) Loop in the *LRR1* network suggested multiple transcript isoforms expressed in the sample. Three isoforms are likely expressed in fibroblasts these are shown overlaid on the network together with a schematic representation of each isoform. (ii) Close-up view of exon skipping the**

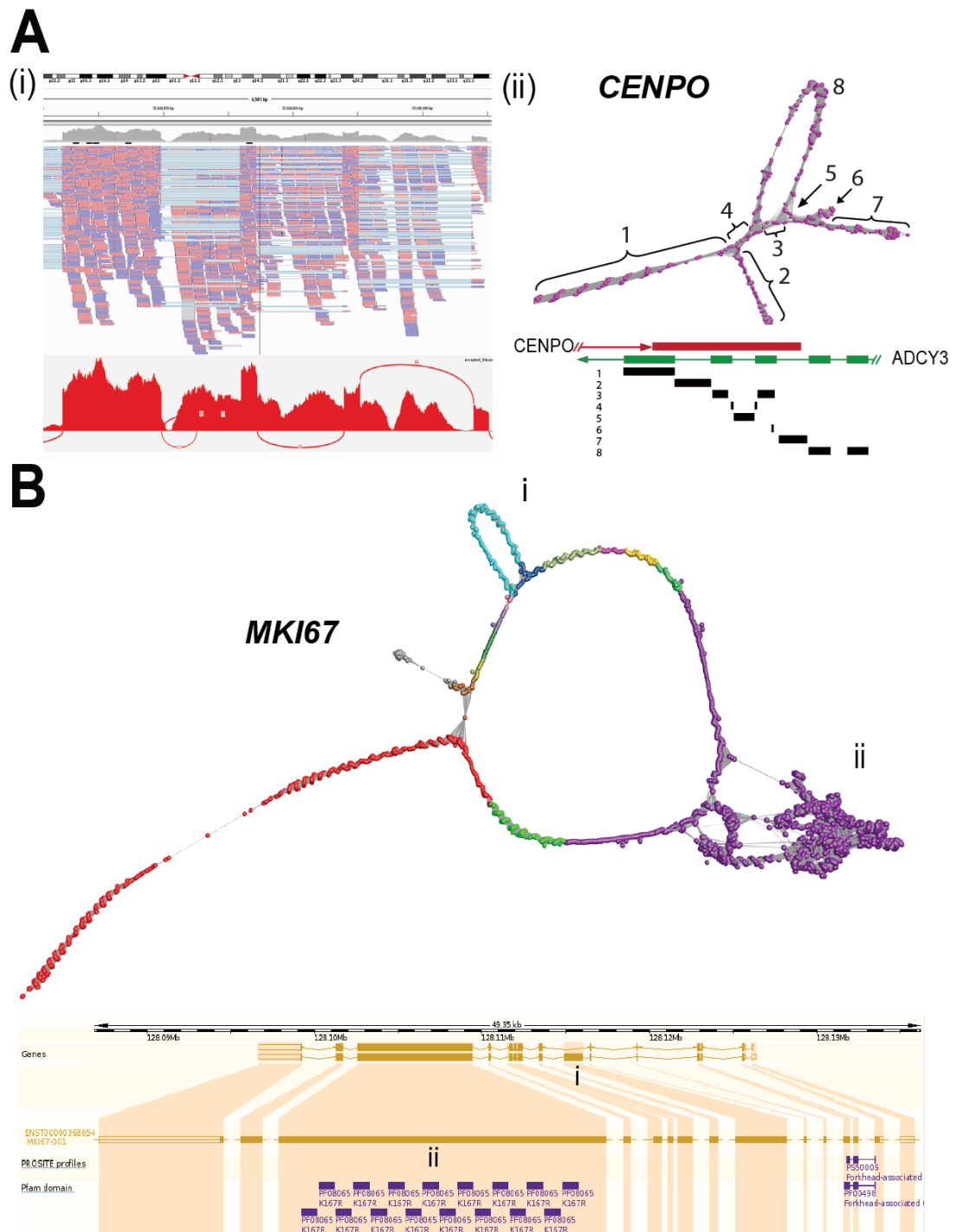
event in *LRR1* shows by the connection exon 2 (yellow nodes) to exon 5 (light purple nodes) skipping exons 3 and 4 (blue/green nodes). **(iii)** Sashimi plot generated in IGV showing RNA-seq reads mapping to *LRR1* locus of human fibroblast sample. Splice junctions are displayed as arcs connecting exons. The number of reads observed for each junction is indicated within segments, and y-axis ranges for the number of reads per exon base are shown. Arcs connecting a pair of exons indicate junctions. **(iv)** The result of RT-PCR of *LRR1* using time-course human fibroblast RNA. Three bands in the results represent alternatively spliced products due to either exon 3/4 skipping, or exon 3 skipping. The PCR band sizes were 1130 bp, 1022 bp, and 310 bp indicates the short, medium and longest isoform of *LRR1* gene. **(B) Confirmation of splice variant *PCMI* using RT-PCR assay.** **(i)** Two different splicing events for *PCMI* were evident from the network visualisation of this gene. **(ii)** Result for RT-PCR of *PCMI* using time-course human fibroblast RNA for two different locations of splice variant. The two bands observed in each assay represent alternatively spliced product caused by exon 7 and 25 skipping, respectively. The PCR band sizes were 339 bp and 223 bp shows the skipping of exon 7, and 496 bp and 331 bp for the skipping of exon 25 of the *PCMI* gene. **(iii)** Representative Sashimi coverage plot generated in IGV showing RNA-seq reads mapping to the *PCMI* locus of human fibroblast sample.

### 3.3.3.3 Issues with assembly and internal repeats

*CENPO* encodes the centromere O protein which is a component of the Interphase Centromere Complex (ICEN) components. It is localised at the centromere throughout the cell cycle (Saito et al.) and required for bipolar spindle assembly, chromosome segregation, and checkpoint signalling during mitosis. When the network assembly of *CENPO* was visualised it showed a complex topology within its final 3' exon (**Figure 3.8A**). In principle network elements representing exons should form linear graphs, bifurcation of network structure only occurring at exon junctions. In order to explain the observed anomalies in the network structure, the location of those reads in the network giving rise to the looped structures was investigated. The genomic origin reads in the network mapping to junctions within the exon were examined using BLAST. It transpires that adenylate cyclase 3 (*ADCY3*) that encodes a membrane-associated enzyme and is located on the opposite strand of chromosome 2. There are a few its 5' exons that overlap with the final 3' exon of *CENPO*. The RNA-seq libraries were non-directional in nature, and due to an error in read mapping, there were reads in the assembly *CENPO* that were derived from *ADCY3*

and the exon boundaries of this transcript. These gave rise to the observed alternative splicing like the structure of the final portion of the *CENPO* network. These anomalies are difficult to observe using conventional visualisation tools such as IGV and even with the Sashimi plots, it is not easy to distinguish that reads are derived from two overlapping genes.

*MKI67* encodes the antigen Ki-67 is a well-established cell proliferation marker. The RNA-seq mapping network of *MKI67* contains two complex features; a loop representing a splice variant and knotted structure due to internal repeat sequences. In the case of the former, exon 7 is spliced out in transcript ENST00000368653 as compared to ENST00000368654, both isoforms being expressed within fibroblasts, and within their exon 14 are 13 repeats of a K167/chmadrin domain leading to the formation of internal homology loops (**Figure 3.8B**).



**Figure 3.8: Complex gene network structure. (A) Miss-assembly of reads from overlapping gene. (i)** IGV visualisation of *CENPO* exon 8 together with corresponding Sashimi plot. *ADCY3* overlaps with *CENPO* on the opposite strand of DNA. **(ii)** In this case, reads derived from *ADCY3* mRNA are being wrongly mapped to *CENPO* resulting in the complex network structure observed in exon 8. The schematic diagram of overlapping genes *CENPO* and *ADCY3* is shown above and regions of the network mapped back to it. The loops in exon 8 of *CENPO* are being

formed by junction reads derived from *ADCY3* encoded on the opposite strand. **(B) Repeat sequence causes perturbation in the network structure.** Network-based visualisation of *MKI67* (ENST00000368654). In this network, there are two structures **(i)** an alternatively spliced exon and **(ii)** internal duplication. Skipping of exon 6 giving rise to ENST00000368653 can be observed as the loop structure, whilst the knotted structure is formed due to the presence of 16 K167/Chmadrin repeat domains within exon 12.

### 3.3.3.4 Highly expressed gene network analysis

In order to explore transcript variation within a single gene, one gene *TPMI* that encodes for the muscle/cytoskeletal protein tropomyosin was focused. This gene was selected because it is widely expressed across human tissues but at very variable levels, being particularly strongly expressed in muscle (<https://www.ebi.ac.uk/gxa/>). It has 10 exons and a large number of potential isoforms. There are 19 protein-coding transcript isoforms, with a further 14 non-coding isoforms i.e. with a retained intron or the products of nonsense-mediated decay are reported in Ensembl. Using a network approach, a network of *TPMI* generated from in-house RNA-seq data (9,702 reads) was analysed (**Figure 3.9A**).

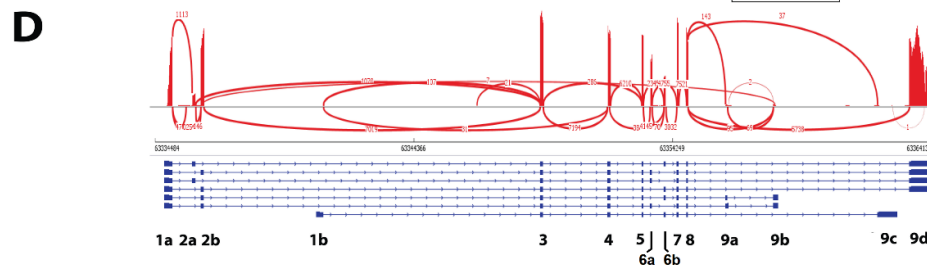
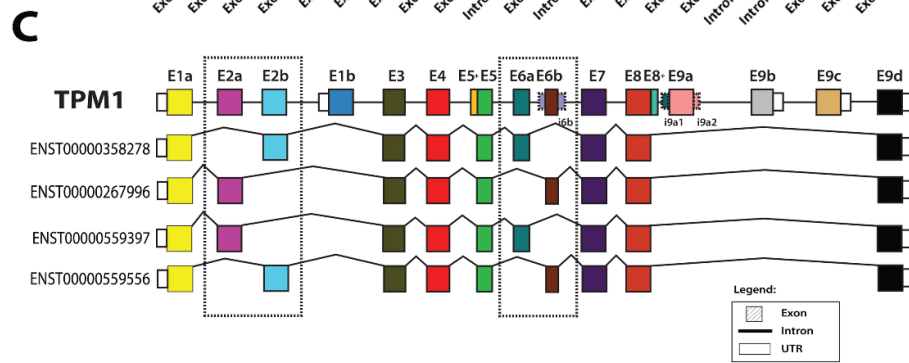
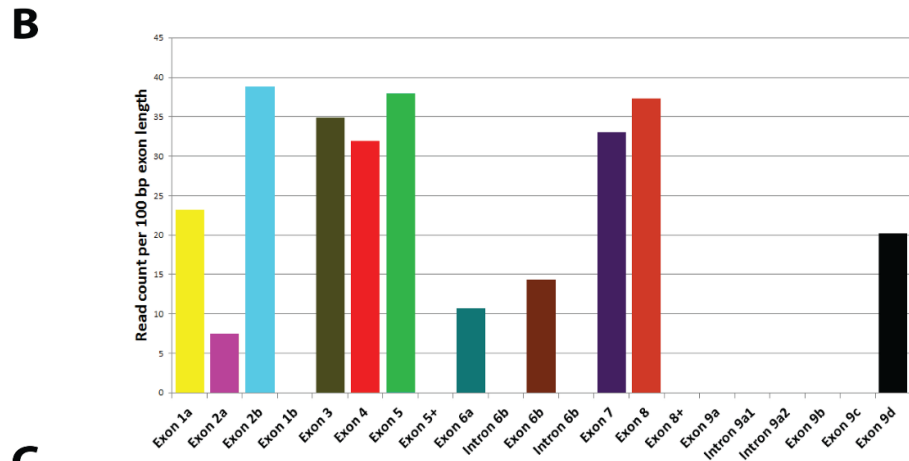
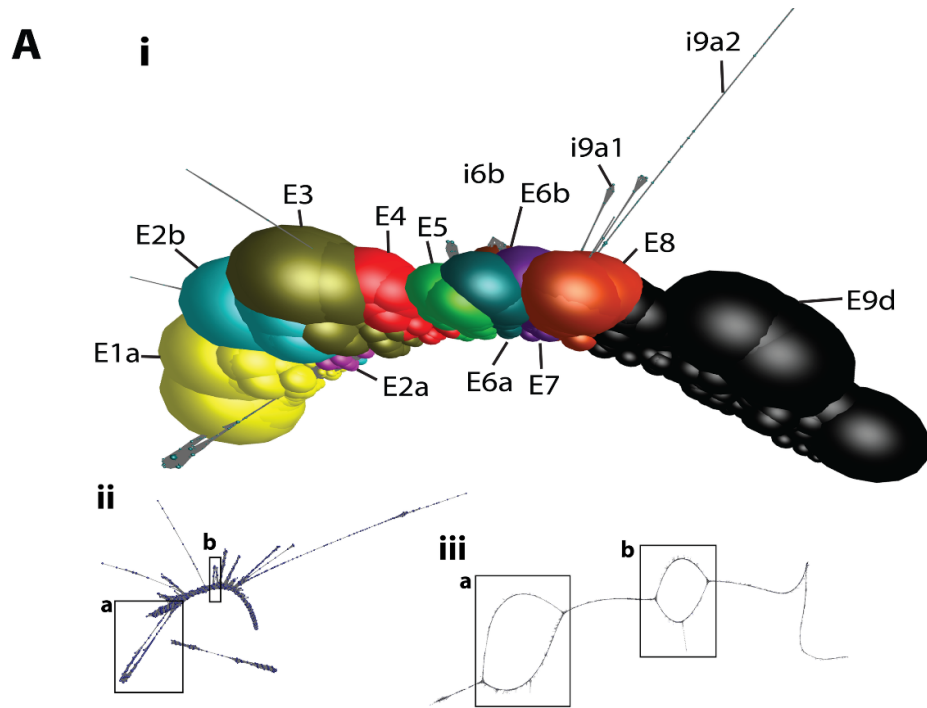
Network visualisations provide a clear indication of underlying transcript complexity. Within BioLayout *Express*<sup>3D</sup> the ability to overlay different transcript models onto the network greatly assists in working out where network elements map back to on the genome. Where networks really help us in providing a visual representation of complexity, which in turn helps define the branch points in transcript diversity.

Network-based analysis of the *TPMI* gene expression in the human fibroblast, revealed that there was possibility of four major isoforms expressed in this cell (ENST00000348278, ENST00000267996, ENST00000559397 and ENST00000559556) containing exons 1a, 2a/2b, 3, 4, 5, 6a/6b, 7, 8 and 9d, the deep coverage of which can be inferred from the size of nodes. It appears that there are two mutually exclusive exons (MXE) sites, in this case, which is exon 2 and exon 6.



From the network and histogram read count per 100 bp, the tendency selection of exon 2b is higher compared to exon 2a was observed.

Whilst for exon 6, there are quite a similar number of reads and size of the node between exon 6a and 6b. However, there appeared to be evidence of other minor transcript isoforms being expressed in this tissue as visualised from the small branches emerging out from of the dominant network structure. Mapping the reads from these minor structures to exon uniquely present in known *TPM1* transcript isoforms and their branch points suggested the presence up to four other transcript isoforms in the fibroblast data. Network-based analyses of *TPM1* was compared as summarised in **Figure 3.9C** with the corresponding Sashimi plots (**Figure 3.9D**).



**Figure 3.9: Network-based visualisation of *TPMI* of human fibroblasts at 24 after serum refeeding.** (Ai) Network-based visualisation of human fibroblast where the colour nodes represent exon, nodes represents a sequence read and edges similarity score between reads above a defined threshold. In these graphs, node size reflects the relative number of reads depth. In this graph, it shows one major isoform expressed while another three isoforms are minor isoforms which can be visualised from the fewer nodes branch out from the graph. The mutually exclusive exon 6a and 6b can hardly be seen from the graph while the retention intron of 6b can be visualised from a few nodes branching from the graph. The branch out nodes are supported by the gene model of human reference genome shows that protein-coding and retained intron which is an alternatively spliced transcript that contains an intronic sequence. (Aii) *TPMI* transcript network lacking 'Node size' class. The 'loop' and bifurcation of the network can be seen clearly in this network. (Aiii) Whilst, the same layout file generated using GraphTool with filtering edges of similarity Mega BLAST score less than 170 reveals the obvious two MXE sites in the network. This high stringency of network parameter shows this splice variant site is expressed in real data. The network shows here reveals the alternative splicing sites which occur at two mutually exclusive (MXE) sites which are (Aiiia & Aiiia) exon 2a or 2b and (Aiiib & Aiiib) exon 6a or 6b. (B) Histogram of *TPMI* human fibroblasts. Histogram number of reads per exon per sample in each tissue is shown in the graph. The coloured histogram represents exon in the graph visualisation. (C) Schematic gene representation of *TPMI*. All isoforms expressed shows at the top of the isoforms gene representation and expressed in human fibroblasts. (D) Representative sashimi coverage plot generated in IGV showing RNA-seq reads mapping to *TPMI* locus from human fibroblasts at 24 h after serum refeeding. The height of the bars represents overall read coverage. Splice junctions are displayed as loops. The number of reads observed for each junction is indicated within segments, and y-axis ranges for the number of reads per exon base are shown (read coverage, left). The plot suggests different isoforms expressed in the sample shows by the arc connecting a pair of exons.

The network-based analyses were continued to compare non-collapsed network of *TPMI* gene of human fibroblasts. In this network, it shows two alternative splice sites represent 'loop' in the network. Whilst intron could be seen from the branching nodes out of the network (**Figure 3.9Aii**), the other loop can be seen from the network was not convincing splice sites and could be a sequence homology in the network and intron sequences. The mutually exclusive exons (MXE) sites in the network could be seen when the same layout file was laid out using GraphTool by increasing stringency of the network (**Figure 3.9Aiii**). By filtering the number of

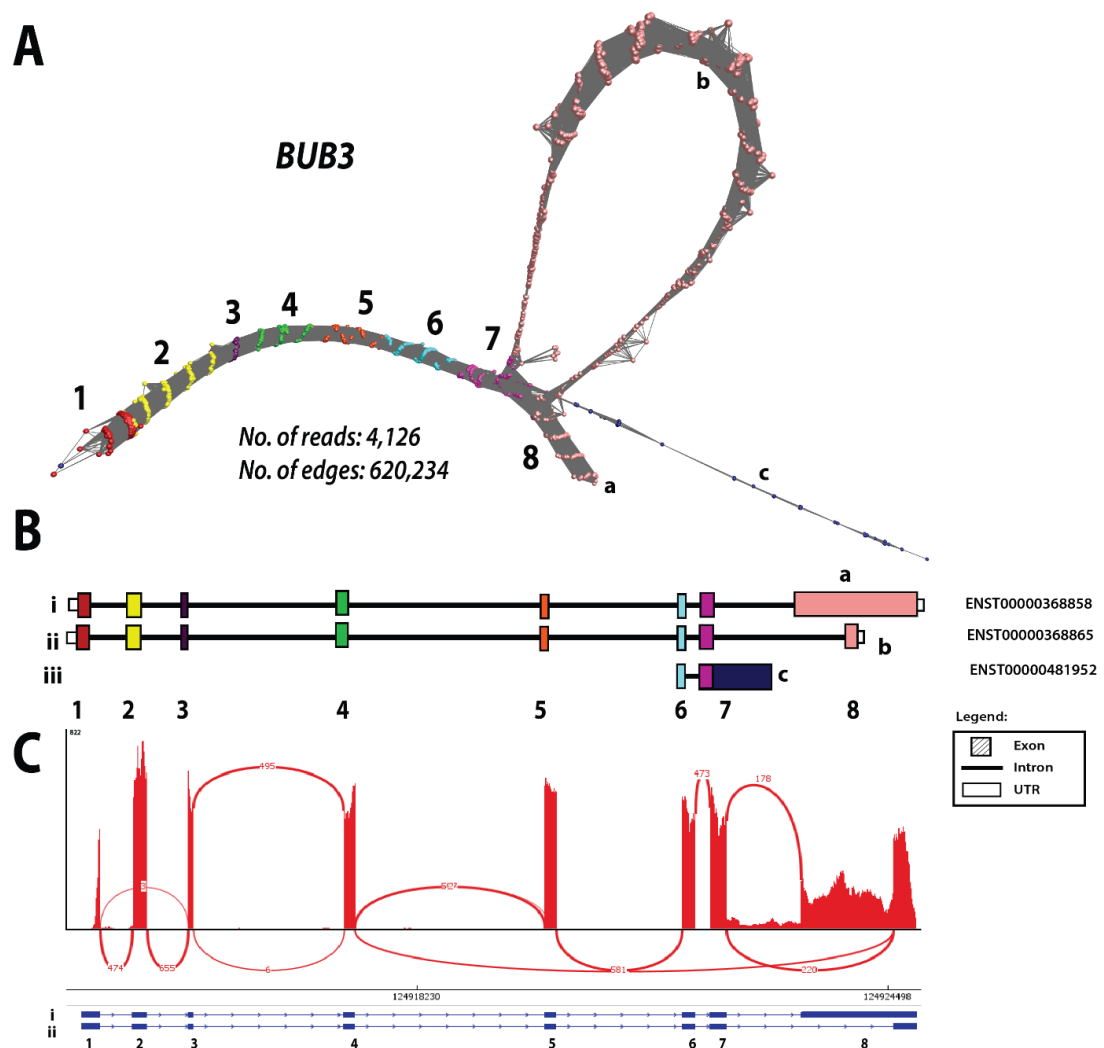
edges MegaBLAST score between nodes below 170, shows the strong evidence the existence of four possibilities of the expressed transcript.

### 3.3.3.5 Internal splicing network structure

In order to further explore network-based visualisation of human fibroblasts, three more genes have been examined which are *BUB3*, *FAM64A*, and *NRM*. In the first case, *BUB3* encodes mitotic checkpoint protein, which it has a dual function in spindle-assembly checkpoint signalling and promotes the establishment of proper kinetochore-microtubule (K-MT) attachments (Logarinho et al., 2008; Tang et al., 2004). It has 8 exons and a number of potential isoforms. Ensembl reports there to be 4 protein-coding transcript isoforms and only one non-coding isoform i.e. with a retained intron or the products of nonsense-mediated decay. Using network approach, a network of *BUB3* generated from human fibroblast RNA-seq after 24h serum refeeding data was analysed.

In this *BUB3* network, two major isoforms can be visualised from a loop that caused by the different size of same exon 8 generate two transcripts variant encoding different isoforms whilst nodes emerge from major network structure was identified as a processed transcript. Both of the major isoforms have 8 exons while minor isoform isoforms have only 2 exons (**Figure 3.10A**).

Network-based analysis of *BUB3* transcript co-expressed in the human fibroblast revealed that there were essentially two major isoforms expressed. Due to internal splicing, the shorter exon of second isoform (ENST00000368865) causes a loop that emerges from the same location of a longer exon of first isoform (ENST00000368858). The small connecting nodes emerge from the major network identified as a processed transcript (ENST00000481952) (**Figure 3.10B**).



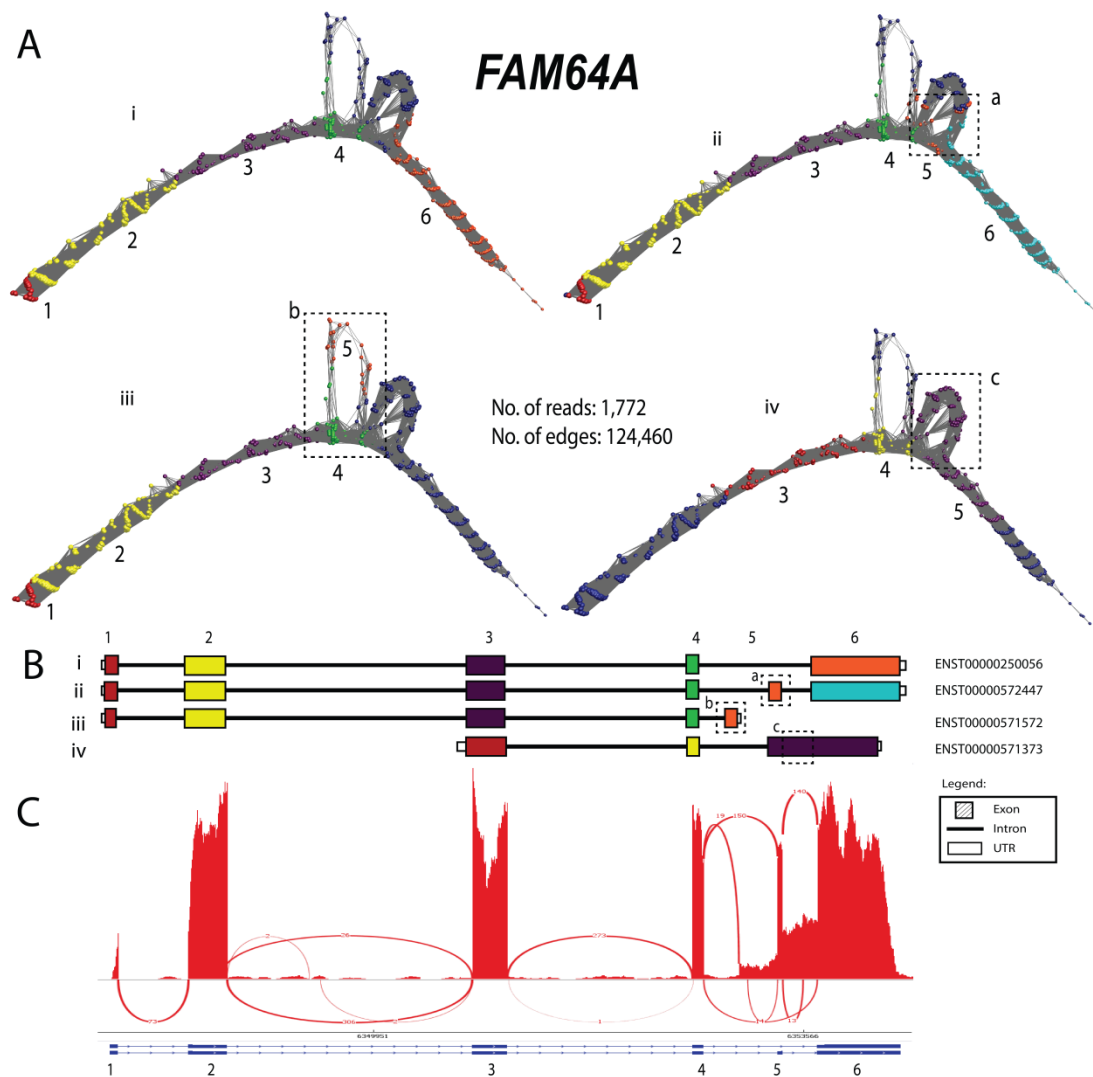
**Figure 3.10: Network-based visualisation of *BUB3* of human fibroblasts. (A)** In this network, it shows two major isoforms expressed which can be visualised from the loop branch out from the graph caused by the last exon of an isoform. A minor isoform (processed transcript) can be seen from single linear nodes emerge the graph. **(B)** Schematic gene representation of *BUB3*. **(C)** Sashimi plots.

### 3.3.3.6 Three loops network structure

In the second case, network transcript of *FAM64A* was examined. *FAM64A* (Family with Sequence Similarity 64, Member A) encodes protein FAM64A or other names of regulator of chromosome segregation protein. It may play a role in the control of metaphase-to-anaphase transition during mitosis. *FAM64A* was selected because of its network structure has two loops features. It has 6 exons and a number of isoforms – 9 in total. The network of *FAM64A* contains three loops which representing splice

variants due to a different transcript expressed. A structure of the *FAM64A* transcript in human fibroblast network has 1,772 reads and 124,460 edges.

Network-based analysis of *FAM64A* transcript network in the human fibroblasts (**Figure 3.11A**), revealed that there were at least three major isoforms expressed in this tissue ENST00000250056 contains exon 1, 2, 3, 4 and 5 (**Figure 3.11Ai**), ENST00000572447 contains exon 1, 2, 3, 4, 5 and 6 and ENST00000571373 contains exon 3, 4 and 6/7. These transcripts can be seen from small nodes emerge inside a loop at exon 5 (**Figure 3.11Aii-a**) and a splice out at exon 5 (**Figure 3.11Aiv-c**). However, there appeared to be evidence of another minor isoform (ENST00000571572). This minor isoform can be visualised from splice out of exon 4 (**Figure 3.11Aiii-b**). In **Figure 3.11B**, schematic gene representation of *FAM64A* which was believed to be expressed in the fibroblast. The Sashimi plots show the quantitative visualisation of the RNA-seq read alignment of the liver (**Figure 3.11C**). In this view, the isoforms detected in fibroblast are only two rather than five isoforms detected using network-based analysis.



**Figure 3.11: Network-based visualisation of *FAM64A* of human fibroblasts. (A)** In this network, it shows three major isoforms expressed and one isoform is a minor isoform. The major isoform can be visualised from the linear graph (**Ai**), the large and small loop emerge from the network (**Aii-a and Aiv-c**). While the minor isoform (processed transcript) can be visualised from the nodes looped out the network. (**B**) Schematic gene representation of *FAM64A*. (**C**) Sashimi plots.

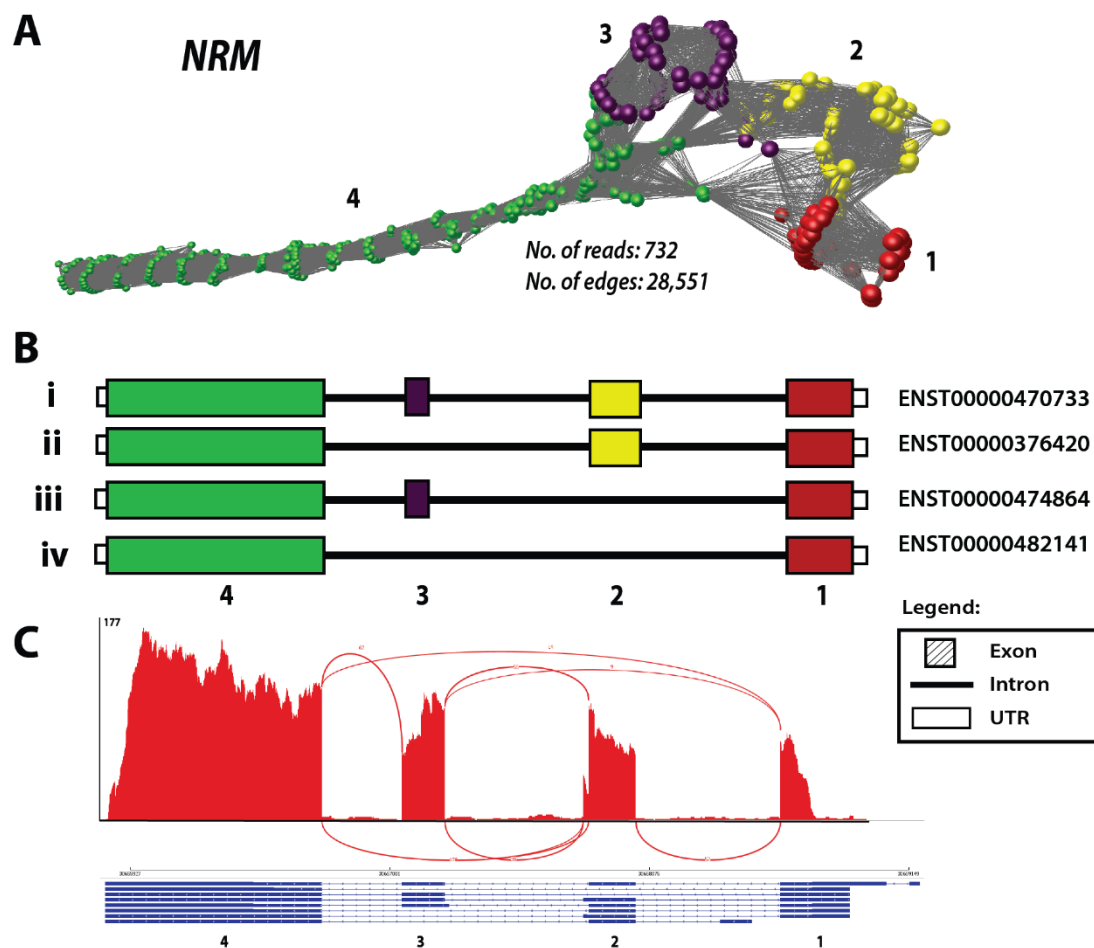
### 3.3.3.7 Alternative splice network structure

In the last case, the network transcript of *NRM* was examined. *NRM* encodes nuclear envelope membrane protein. The protein encoded by this gene contains transmembrane domains and resides within the inner nuclear membrane, where it is tightly associated with the nucleus. It shares homology with isoprenylcysteine

carboxyl methyltransferase (ICMT) enzymes. Alternative splicing results in multiple transcript variants encode different protein isoforms. *NRM* has 4 exons and a number of isoforms which four protein-coding isoforms and five processed transcripts – with a total number of nine different isoforms. The network transcript of *NRM* contains three loops; two big loops which can be seen from the network whilst a small loop can be seen from the small nodes loop inside the bigger loop. The network consisted of exon 1-4, with alternative splicing occurred at exon 2 or 3 while another isoform skips both exon 2 and 3. This network has 4,301 reads.

Network-based analysis of *NRM* in the human fibroblasts (**Figure 3.12A**) revealed that there were four major isoforms expressed in this tissue, ENST00000470733 contains exon 1, 2, 3 and 4 (**Figure 3.12Bi**), ENST00000376420 contains exon 1, 2 and 4. This transcript can be seen spliced out at exon 3 (**Figure 3.12Bii**), ENST00000474864 contains exon 1, 3 and 4 (**Figure 3.12Biii**) and final isoform ENST00000482141 contains only exon 1 and 4 (**Figure 3.12Biv**). In **figure 3.12B**, the schematic gene representation of *NRM* gene is shown, which is believed to be expressed in fibroblasts. The Sashimi plot shows only two isoforms rather than four isoforms detected using network-based analysis.





**Figure 3.12: Network-based visualisation of *NRM* of human fibroblasts. (A)** In this network, it shows four possible isoforms expressed which can be visualised from the two loops branch out from the network. **(B)** Schematic gene representation of *NRM*. All isoforms expressed shows at the top of the isoforms gene representation and expressed in human fibroblasts. **(C)** Sashimi plot.

### 3.4 Discussion

As with microarrays RNA sequencing has the potential to not only measure transcript abundance but also offers a platform to explore transcript diversity within and between cells and tissues and to be able to analyse expression from unsequenced genomes. Sequencing platforms that produce short-read data (50-250 bp) currently dominate the field. Many tools and analysis pipelines already exist to process these data from the DNA sequencer, through mapping onto a genome or *de novo* assembly and summarise these data down to read counts per gene/transcript. These data are

then ready for differential expression or cluster-based analyses. It is also routine practice to port data into tools such as IGV, where it can be visualised in the context of the reference genome (where available). Reads are shown stacked onto the point from which they have been determined to originate. In instances where a single RNA species are transcribed from a specific locus, existing visualisations are sufficient for most needs. However, where multiple transcripts are produced from a given loci, deconvolution of that assembly into the component transcripts can be challenging. Tools such as Sashimi plots use information derived from exon boundaries and paired-end reads to display connections between exons, the thickness of the joining line indicating the number of reads derived from a given exon-exon boundary. When transcript diversity is relatively simple, these views provide a good and sufficient representation of events, when transcript variance is complex they can be difficult to interpret.

This work describes a new and complementary approach to the analysis of RNA-seq data that is based upon the construction and visualisation of RNA assembly networks. In this method, RNA-seq reads mapping to specific loci are directly compared with each other by calculating an all-vs-all similarity matrix. In the context of a network visualisation of these data, nodes represent individual reads or collections of identical reads, whilst edges represent similarity scores between them above a given threshold. Information about a read can be mapped onto them and used to annotate nodes. In this manner, different transcript models can be overlaid onto the network assemblies, reads derived from a given exon sharing the same colour. This provides a way to quickly visualise how well a given transcript model matches the assembly.

A fundamental challenge in implementing this approach is the ability to layout and display network assemblies of data, such that the underlying structure of the networks can be interpreted. BioLayout *Express*<sup>3D</sup> was originally developed for the visualisation and the analysis of expression data as correlation networks, a purpose for which it still provides a powerful solution. In this paradigm, networks can consist of 10's of thousands of nodes (each representing a gene or transcript) connected by

millions of edges based on calculated similarity measures, i.e. correlation coefficient between expression profiles. For this reason, the tool has been designed to support the visualisation and exploration of big network diagrams up to 30,000 nodes (Freeman et al., 2007; Theocharidis et al., 2009) e.g. In a gene expression atlas of the domestic pig study by Freeman et al. (2012), a graph consisted of 20,355 nodes and 1,251,575 edges was generated while a study by Mabbott et al. (2013) on primary human cell, they generated a graph consisted of 24,808 nodes connected by 1,476,632 edges. With such networks, 3D visualisation offers distinct advantages. The use of OpenGL to render networks means the size of network supported can be larger (and are arguably more beautiful) and interpretation of their layout and navigation around them. While the topologies of correlation networks are usually complex, due to multiple cliques (areas of high connectivity) formed by groups of co-expressed genes, their layout using a Fruchterman-Reingold (F-R)-based algorithm is sufficient to allow comprehension of the topology of these networks. However, RNA-seq assembly networks have a fundamentally different type of structure. Reads only share edges with others derived from up or downstream positions in the genome, thereby giving rise to a string-like overlap network. The F-R algorithm is effective in separating nodes on a local scale, but our early studies showed it to perform poorly in separating more distant interactions resulting in twists in the 'string'. The FMMM algorithm is based on a combination of an efficient multilevel scheme and a strategy for approximating the repulsive forces in the system by rapidly evaluating potential fields (Hachul and Jünger, 2005). When used to layout linear RNA assembly networks (at its highest quality setting) they straighten out allowing clear visualisation of the underlying structure. At a local level when read density provides sufficient high and even coverage, nodes are arranged in a spiral structure by the FMMM. Lower quality but higher speed settings of the algorithm are often sufficient to appreciate higher level network structure with greatly reduced layout times.

Here I explore the power of good network visualisations of RNA-seq read assemblies to interpret transcript diversity. When a sole transcript is present in the data, a linear string-like network is generated. If the coverage of the gene is low, the network may

appear as a number of isolated components, at higher levels of coverage a single network component is formed. As coverage increases, the network started showing a characteristic of spiral appearance as a ‘perfect’ overlap network is achieved. At a point where every base position along a transcript has one read mapping to it, in principle, further reads add nothing to the network. However, additional reads add greatly to the computation time taken to calculate a read similarity matrix, and add nodes and edges have to be rendered. Collapsing redundant reads to a unique sequence/node, therefore, speeds up all aspects of visualisation. Currently, in this analysis pipeline only maps down to unique reads, so reads that map to a specific haplotype, i.e. they cover contain SNP or reads containing sequencing errors are represented in the network. In principle, networks could be collapsed further if a small number of within sequence variations were ignored.

In order to differentiate nodes representing many reads from those that represent unique reads in lower numbers, I visually encoded this information using node size. Abundant identical reads being represented by large nodes, the size of the node decreasing with read depth. The graphs are shown here of *GAPDH* and *TUBA1C* illustrate results of this approach to collapsing complexity of RNA-seq assemblies. These graphs are of single transcripts sequenced at high depth; the many small nodes are an interesting visualisation of sequencing ‘noise’. They are almost impossible to visualise when this collapsing strategy is not applied. The size of networks, in particular, the number of edges, is largely determined by the set threshold for similarity; lower thresholds are giving rise to an increased number of edges. In order to minimise layout times and improve the fluidity of network rendering, a similarity threshold should ideally be selected that allows the construction of a single component network with a maximum number of nodes and a minimum number of edges. This threshold value is dependent on the depth of sequencing, i.e. number of reads per unit length of DNA. From the experiments exploring the ‘generalised’, a  $p$ -value (% length similarity) of 98 and  $l$ -value (% length coverage) of 31 as default MegaBLAST settings were defined, to be made more stringent when read coverage is a high potential.

Whenever sequences diverge or contain homologous domains, the RNA-seq assembly networks take on 'loop' structure. Where alternatively two spliced transcripts deviate in sequence, forks in the network are observed starting and finishing at exon boundaries. *LRR1*, as expressed in human fibroblasts, is a relatively simple example of an alternatively spliced transcript. In one version of the *LRR1* transcript expressed by these cells, exon 3 and 4 are spliced out, and a large loop is observed in the network. The network of *PCMI* possesses two loops corresponding to known splice variants at exon 7 and 24. The splicing events are immediately obvious from the network visualisations; debatable they are less easy to understand from the corresponding sashimi plot.

In certain gene networks, I observed different structure within and between nodes position within an exon. Within exon 8 of *CENPO*, I observed complex network topology. In this case, the analysis showed that it was due to reads produced from the transcription of *ADCY3* whose terminal 5' exon overlaps on the opposite strand. Exon boundary reads mis-mapped from the transcripts of *ADCY3* causing loops within network representing exon of *CENPO*. The inability to correctly map reads from overlapping transcribed exons is one of the reasons the majority RNA-seq analyses are now generated from directional cDNA libraries. In the case of *MKI67*, a series of 14 K167/Chmadrin domains within exon 14 of the gene cause a knotted structure in that portion of the network representation. An alternative splice variant missing exon 6 is also apparent in the network visualisation of this gene.

Next, I wanted to test the potential of network visualisation to analyse transcript complexity of highly expressed genes. In the case shown here, *TPMI* transcript diversity in RNA-seq data derived from human fibroblast 24 hours after serum refeeding was examined. Tropomyosin 1 is most heavily expressed most tissues including human fibroblast where it functions as an actin-binding protein involved in the contractile system of muscle. A dominant and possible sole functional transcript is expressed to fibroblast expressed isoform of the protein. Also, a relatively small number of reads mapped to exon 2a and terminal intron sequences, suggesting the presence of a very low number of other transcript isoforms. Whether these represent

the presence of transcriptional noise or transcription of these isoforms by cell types present in a low abundance, it is not clear. Through studying these networks and mapping this information back to the Ensembl transcript models for this gene, up to four transcript isoforms are estimated to be expressed in human fibroblasts. In this network, there are two mutually exclusive exons (MXE) sites which are either exon 2a or 2b and either exon 6a or 6b. Based on the histogram analysis, I could consider that exon 2b is more highly expressed than 2a while for exon 6 seem to be relatively as the same level of expression. Therefore, these two transcripts were possible to be expressed in human fibroblasts with only two transcripts which have exon 2b, and either exon 6a and 6b are likely to be expressed. This is largely based on the presence of the data of reads mapping back transcript-specific exons. Even with the availability of network visualisations and other visualisation tools, interpreting these data is difficult. Transcript assemblies such as these are inherently complex.

Here I present a new and complementary approach to aid the analysis of RNA-seq data. In this chapter, I describe an analysis pipeline whereby reads mapping to given loci can be compared, and then assemblies visualised as a network. In this environment, information can be overlaid onto the network regarding node colour and/or size (potentially shape), and the structure of the network can be explored in 3-dimensional space. The network structure is revealing the nature of the underlying sequence assembly and complexities therein. I demonstrate the ability to recognise splice variants in these networks, areas internal homology and issues with read mapping using this approach. I also provide an example of just how complex these assemblies can be and the strengths and limitations of networks and other approaches in these instances. Overall, I have attempted to show how the visual cues provided by this network visualisation can be used to explore the reasons behind observed complexity and complement other solutions to the analysis of these data.

# Chapter 4 – An analysis of transcript variation in human tissues

## 4.1 Introduction

Alternative splicing (AS) plays a fundamental role in the diversification of protein function, regulation and the main contributor to cellular diversity (Tazi et al., 2009). It is frequently being used to produce tissue-specific protein isoforms (Merkin et al., 2012). While the disruption of specific AS events and wrong splice sites usage have been associated with a number of human genetic diseases (Xiong et al., 2015). To date, the 20,000 or so protein-coding genes in the human genome have been shown to generate more than 140,000 different gene transcripts (Flicek et al., 2014). Furthermore, divergence in isoform-specific read coverage indicates that most AS, cleavage and polyadenylation events differ between tissues and vary between individuals (Yeo et al., 2004). AS may occur in a different situation: exon skipping, intron retention, mutually exclusive exons, alternative first and last exons, alternative 5' and 3' splice sites, and alternative 5' and 3' untranslated regions (UTRs) (Wagner and Berglund, 2014; Wang et al., 2008). However, the identification and quantification of differentially spliced transcripts in genome-wide transcript analysis are very important aspects (Conesa et al., 2016).

### 4.1.1 Alternative splicing analysis

RNA-seq has a wide variety of application, but no single analysis pipeline can be used in all cases. Many computational methods have been developed for the past several years for RNA-seq analysis of alternative splicing (Shen et al., 2014). However, these methods have limitations and disadvantages of replicate RNA-seq data (Hooper, 2014). There are two major categories to analyse alternative splicing which is transcript-level and exon-level differential expression. In the analysis of transcript-level differential expression, it is able to detect changes in the expression level of transcript isoforms within the same gene. Furthermore, another tool such as Cufflinks (Trapnell et al., 2013) and DiffSplice (Hu et al., 2013) use the Jensen–Shannon divergence metric to infer differential isoform proportion while accounting

for variability between replicates. rSeqDiff employs a hierarchical likelihood ratio test to identify both differential gene and isoform expression (Shi and Jiang, 2013). Nevertheless, all these methods are mostly obstructed by the limitations of short-read sequencing for accurate identification at the isoform level (Xie et al., 2014). Cufflinks consider the estimation uncertainty, nonetheless, the test statistic unable to distinguish the contributions from replicates with high or low degrees of estimation uncertainty (Trapnell et al., 2013). ALEXA-seq (Griffith et al., 2010), MISO (Katz et al., 2010), rSeqDiff (Shi and Jiang, 2013), and SpliceTrap (Wu et al., 2011) is designed for two-sample comparison, however, unable to handle replicates samples.

On the other hand, the second category is the exon-based approach. The signal of alternative splicing can be detected by comparing the distributions of reads on exons and junctions of the genes between the compared samples (Anders et al., 2013). This approach is based on the theory that differences in isoform expression can be detected through the signals of exons and its junctions. DEXSeq (Anders et al., 2013) and DSGSeq (Wang et al., 2013), test for significant differences in read counts on exons and junctions of the genes to detect differentially spliced genes. rMATS infers differential usage of exons by comparing exon-inclusion levels defined with junction reads (Shen et al., 2014). Furthermore, another tool called rDiff (Drewe et al., 2013) compares read counts on alternative regions of the gene in the presence of annotated alternative isoforms in order infer differential isoform expression (Conesa et al., 2016). The better accuracy of an exon or junction methods in identifying individual alternative splicing events is the advantage of this method (Conesa et al. 2016). While to study on the inclusion and exclusion of specific exons, functional protein domains are suitable to use exon-based methods. However, no existing method handles paired replicate data and this is important to analyse and determine splice variant in this chapter. Therefore, by using rMATS which were developed because of the need for robust analytic tools to detect alternative splicing changes from replicate samples as well as can handle different types of replicate study design, i.e. unpaired or paired (Shen et al., 2014).



### 4.1.2 rMATS – one of the best tool to detect alternative splicing

A number of methods to detect and visualise AS have been described. Shen et al. (2014) designed a new statistical model and developed a computer program called replicate Multivariate Analysis of Transcript Splicing (rMATS) to detect differential AS from replicate RNA-seq data. It employs a hierarchical model to simultaneously account for sampling uncertainty in individual replicates and variability among replicates. In addition to the analysis of unpaired replicates, rMATS also includes a model specifically designed for paired replicates between sample groups. The hypothesis-testing framework of rMATS is flexible and can measure the statistical significance over any magnitude of splicing change.

Shen et al. (2014) described rMATS has several key features. The first key feature is that rMATS uses a hierarchical framework to model exon inclusion levels denoted as  $\Psi$  (PSI –percent spliced in), which simultaneously accounts for estimation uncertainty in individual replicates and variability among replicates. The second important feature is, rMATS includes a model specifically designed for paired replicates as well as unpaired replicate data. Bivariate normal distribution is introduced with a correlation parameter to model the correlation between matched pairs. Importantly, the use of paired-read information in paired replicate data eventually improves the statistical power. The third key feature of rMATS, a user can define either null or alternative hypotheses in the hypothesis-testing framework for differential alternative splicing. rMATS employs a likelihood-ratio test to compute the  $P$ -value that the difference in the mean  $\psi$  values between two sample groups exceeds a given threshold. From this framework, rMATS can measure the statistical significance based on the user-defined magnitude of splicing change. Moreover, the use of the likelihood-ratio test in rMATS significantly improves the speed of the computation compared to the sampling-based  $p$ -value calculation in previously in MATS (Shen et al., 2014). The last key feature, rMATS able to analyse all major types of AS patterns and use RNA-seq reads mapped to both exons and splice junctions because it uses a statistical model that normalises the lengths of individual splice variants. In their studies, rMATS outperformed two existing methods, Cuffdiff

(Trapnell et al., 2012) and DEXSeq (Anders et al., 2012) for replicate RNA-seq data in all simulation settings.

### 4.1.3 Visualising alternative splicing

Visualising the complexity of AS is an important aspect of the analysis. Visualisation for RNA-seq needs dedicated tools that can efficiently process a huge amount of data from several different samples. This has triggered the development of tools to visualise alternative isoforms and events from RNA-seq data. One of the simplest ways to visualise isoforms and events is to produce track files and upload into genome browser (Wang et al., 2015). For instance, TopHat or STAR produces BAM files that can be viewed in the Integrative Genomics Viewer (IGV). Likewise in SpliceGrapher (Rogers et al., 2012) and DiffSplice (Hu et al., 2013) generate files in GFF formats and can be uploaded into the IGV. Nonetheless, tools such as SpliceGrapher (Rogers et al., 2012) and Alexa-Seq (Griffith et al., 2010) have their own visualisation utilities. For comparative study between network approaches, the Sashimi plots are generated using alignments file which stored in the SAM/BAM format and gene model annotations. Furthermore, Sashimi plots can be generated to quickly scan the differentially spliced exons along genomic regions of interest. Finally, there are standalone tools that provide visualisation of results together with additional information of mapping reads, quantifying events and differential splicing such as SpliceSeq (Ryan et al., 2012) and SplicingViewer (Liu et al., 2012). Nevertheless, these visualisation tools employ read stack to the genome as well as a linear line connecting over exon. This is rather complicated to visualise and determine the splice variant.

In Chapter 3, explorations of splice variants using network analysis were limited to transcripts expressed in a single cell type (human fibroblast). However, in this chapter, the use of network-based visualisations of RNA-seq data is explored further by an investigation of splice-variation between different human tissues. 95 individuals of RNA-seq data derived from 27 human tissues were subjected to quality control using network analysis. A correlation network was constructed using only sample-to-sample relationships and samples with a low correlation to samples

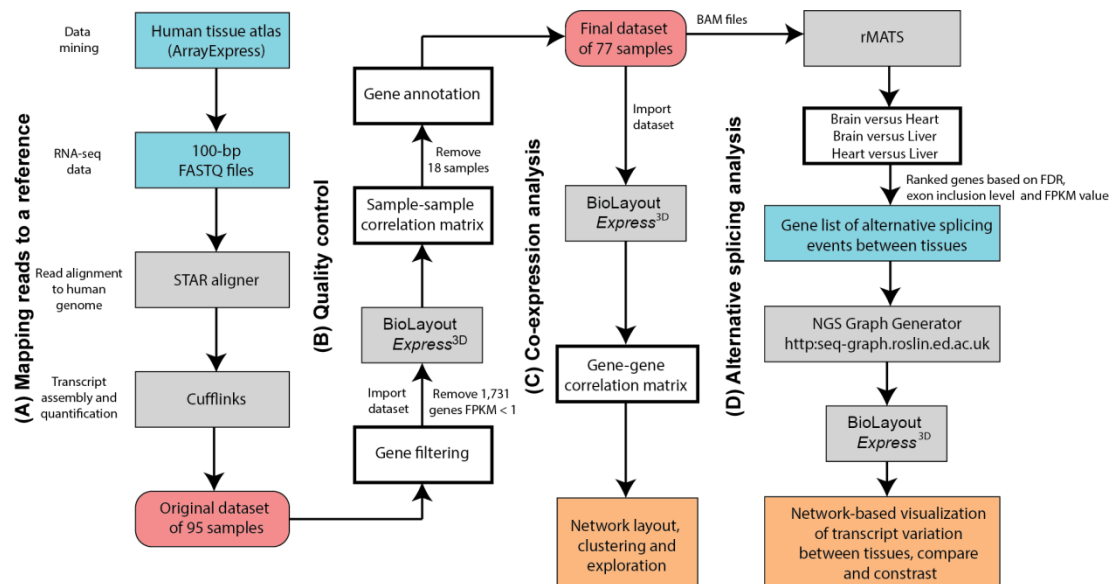
from the same tissue were excluded from subsequent analyses. Subsequently, all 27 human tissues were aligned individually to the human genome, followed by splicing analysis to infer patterns of AS in human genes. The usefulness of the network-based visualisation approach in determining AS of a gene across different tissues is presented here. In so doing, these analyses further validate the utility of network visualisation to explore AS events.

Therefore, the objectives of this chapter were:

- 1) To work up a human tissue atlas of RNA-seq data using network analysis.
- 2) To identify and analyse AS events between different human tissues using the rMATS method.
- 3) To validate the AS of genes detected by rMATS through a network analysis.

## 4.2 Methods

Shown in **Figure 4.1** are the data analysis workflow from quality control to data visualisation and exploration of AS across different tissues. In this workflow, there are four different stages: **(A)** mapping reads to a reference, **(B)** quality control, **(C)** co-expression analysis, and **(D)** alternative splicing analysis. For stage **(A)** the data were downloaded and aligned to the reference human genome using STAR aligner and using Cufflinks for transcript assembly and quantification. For stage **(B)** and **(C)**, BioLayout *Express<sup>3D</sup>/Miru* was used for the quality control and co-expression analysis. In stage **(D)**, rMATS was used for the identification of alternative splicing events. BioLayout *Express<sup>3D</sup>/Miru* software was used for QC, gene clustering and network visualisation in stages **(B)**, **(C)** and **(D)**.



**Figure 4.1: Data analysis workflow.** Data analysis pipeline for RNA-seq data from ArrayExpress, including read alignment to the human genome, transcript quantification, quality control, annotation and network analysis. Further analyses were conducted to detect AS in the dataset using the rMATS tool which produces a list of genes likely to be AS between tissues providing a readout that indicates whether there is a skipped exon, alternative 5' splice sites, alternative 3' splice sites, intron retention and/or mutually exclusive exons. The lists of genes from each tissue comparison using rMATS were then sorted by maximum FPKM fold change and FDR. Network layouts of four genes were generated using NGS Graph Generator, and the resulting networks explored.

#### 4.2.1 Datasets

FASTQ files of RNA-seq atlas of human tissues were downloaded from ArrayExpress (Ac.No.: E-MTAB-1733). The data consisted of 95 samples derived from 27 tissues consisting of between two and seven biological replicates per tissue. The samples were sequenced multiplexed 15 per lane, producing an average of 18 million mappable read pairs per sample (Fagerberg et al., 2014).

#### 4.2.2 Read alignment and quantification

100 bp paired-end reads for each tissue were individually mapped to the human genome (Ensembl GRCh37.71) (Flicek et al., 2014) with STAR v2.4.1c (Dobin et al., 2013) and transcripts assembly using Cufflinks v2.2.1 (Trapnell et al., 2010). Moreover, Cufflinks was used to calculate for each sample the number of fragments

per kilobase of exon per million fragments mapped (FPKM) for downstream analyses. A gene was determined as expressed if the FPKM value was more than 1. This threshold has been recommended as the minimum expression required for transcript and protein detection (Fagerberg et al., 2013; Vogel and Marcotte, 2012). A numerical matrix of FPKM values for each gene in all 95 human tissue samples was produced.

### 4.2.3 Quality control and data analysis

To examine this dataset, a sample-sample correlation network was constructed where nodes represent samples and edges represent the Pearson correlation value between samples above the cutoff value. In this dataset, one tissue has seven biological replicates (testis), three tissues have five replicates (lymph node, lung, and colon), eight tissues have four replicates (prostate, placenta, kidney, bone marrow, spleen, heart, small intestine, thyroid), eleven tissues have three replicates (skin, fat, endometrium, gallbladder, liver, appendix, stomach, esophagus, salivary gland, adrenal, and brain), and four tissues have only two replicates (ovary, urinary bladder, pancreas, and duodenum).

Details of this data set are summarised in **Supplementary Table 4.1**. All non-protein coding genes and coding genes with  $FPKM \leq 1$  in all samples were excluded, resulting in expression levels for 18,319 protein-coding genes. In order to produce a more balanced dataset with only three replicates per tissue (where possible), samples with a low correlation between the same tissue type or samples with the lowest read count in a case where all samples had highly similar (correlated value), e.g. testis, were removed for subsequent analysis. Ideally, when visualised using a sample-sample correlation network, samples from the same tissue should cluster by tissue type.

After QC, the final data set consisted of 77 samples derived from 27 tissues which were ordered according to tissue type and saved as an '.expression' file. The data were then loaded into BioLayout *Express*<sup>3D</sup>. A pairwise Pearson correlation matrix was calculated for each protein-coding genes (18,319 genes) represented in the

RNA-seq data as a measure of similarity between the FPKM values derived from different genes. All Pearson correlations ( $r \geq 0.7$ ) were saved to a '.pearson' file and a correlation cutoff of  $r = 0.85$  were used to construct a network. At this threshold correlation cutoff, the data set is relatively large and diverse in the range of biology it represents.

Network layout was performed using the Fast Multipole Multilevel Method (FMML) algorithm (Hachul and Jünger, 2005) and Fruchterman-Reingold force model (Fruchterman and Reingold, 1991). Markov clustering (MCL) algorithm (van Dongen, 2000), which has proven to be one of the most efficient graph-based clustering algorithms (Brohée and Helden, 2006), was used within the BioLayout *Express*<sup>3D</sup> tool to define co-expression clusters of genes. The MCL inflation value, which controls the granularity of clustering, was set to 2.2 to as this value has been demonstrated to be optimal when working with highly structured expression graphs (Freeman et al., 2007).

#### **4.2.4 Functional annotation**

The expression profile and gene content of each cluster were examined to mine genes for overrepresentation of classes of tissue. This approach is broadly used in research to explore gene lists to query for overrepresentation of certain classes of genes usually relating to gene function, i.e. GO terms. A gene ontology web application (<http://www.geneontology.org/>) was used to assess overrepresentation of GO categories of overlapped regulated DEGs and collaboratively regulated genes in the biological process with FDR < 0.05 were chosen as the cut-off criterion.

#### **4.2.5 Differential splicing**

RNA-seq reads were mapped to the human genome (Ensembl GRCh37.71) using STAR v2.4.1c (Dobin et al., 2013b). Differential AS events was identified between pairs of samples, for a group of three human tissues, e.g. brain versus heart, brain versus liver and heart versus liver using rMATS v3.2.2 (<http://rnaseq-mats.sourceforge.net>). Only three comparisons were used in this study to determine the effectiveness of network analysis. This will be the foundation of future works to

transcriptome-wide network analysis. Furthermore, these tissues were chosen as they are known to be very different in biology due to the different gene expression level. This is an important aspect to explore the different alternative splicing pattern in network analysis. Using rMATS, five major types of AS events from RNA-seq data with replicates can be detected (Dobin et al., 2013). They are alternative 5' splice site (A5SS), 3' splice sites (A3SS), skipped exons (SE), mutually exclusive exons (MXE) and retained introns (RI). In each rMATS run, all replicates from the first group were compared to all replicates from the second group to identify differential splicing events with an associated change in the percent spliced in (PSI -  $\Psi$ ) of these events. The inclusion level of each candidate splicing event was calculated using reads mapping to the body of exons as well as splice junctions from three human tissue samples. Differentially spliced events were required to have a complete difference in inclusion ( $\Psi$ ) level greater than 50% and a false discovery rate (FDR) less than 1%. These settings to ensure the list of genes generated from rMATS was contained alternative splicing and useful for network analysis.

#### 4.2.6 Analysis of the network structure

Genes with highest significance score for AS events as detected by rMATS were chosen for examination. Each of the genes expressed in the brain, heart, and liver data were subjected to network visualisation using default MegaBLAST  $p=98$  and  $l=31$ . For highly expressed genes, e.g. *TPM1* uniquification was applied to remove redundant reads for network construction. After that, the networks were examined and compared to the result of Sashimi plots and Vials. Where topology of a given gene, network is relatively simple an explanation of its structure required the only overlay of individual transcript exon information in order to identify splice variant(s) represented. In other cases, more detailed analyses were required. Determining transcript isoform expression from these networks requires comparing each network's structure to the Ensembl database. By calculation, the number of reads mapping to each exon, a histogram of the number of reads per exon per sample for each tissue was generated for comparison the transcript assembly network. Sashimi plots were built from IGV to compare and determine isoforms generated from the network. Recently, Strobelt et al. (2016) developed a visualisation of splice variant

tool called Vials. It was also used to compare and contrast transcript expression with the network visualisations as described here (refer to Chapter 1 Figure 1.12).

## 4.3 Results

### 4.3.1 Quality control of human tissue atlas RNA-seq

The layout of a sample-sample correlation network of 95 human tissue samples showed them to cluster according to tissue type, except one of the lung samples (**Figure 4.2A**). This sample was a clear outlier and removed from the dataset. To avoid a potential problem of insignificant or noisy genes, the outlier samples were removed in a sample-sample network correlation. Colon, testis and small intestine had more than three samples, and the outlier samples were removed. 18 samples were removed from the original dataset to leave 77 samples shown in **Table 4.1** were clustered again using BioLayout *Express*<sup>3D</sup> (**Figure 4.2B**). The reasons of these samples were removed when either a low correlation between samples or low sequencing reads count. For an instant, the correlation value of placenta is 0.68 was removed from the original set. While in the colon, even though the correlation value is 0.89 is considered high, it was removed to limit up to three samples per tissue.

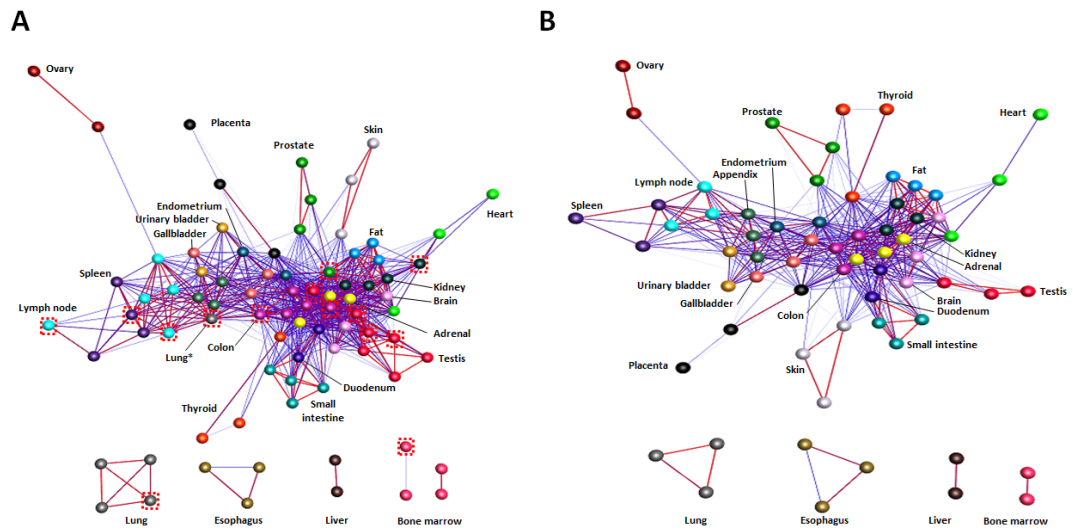
### 4.3.2 Network construction and layout

The BioLayout *Express*<sup>3D</sup>/MIRU software reads a text file containing FPKM data in tabular format. Each row of this file starts with a unique gene symbol, followed by a number of columns of annotation tissue data specific to that gene. The final data columns represent FPKM values for that gene across different tissues. The tool then performs all-versus-all Pearson correlation calculations for all genes. This step is highly optimised and performed in memory because the number of calculations required is very large. Pairs of genes whose Pearson correlation is greater than a threshold ( $p \geq 0.85$ ) are calculated and subsequently generated a network graph. The network consists of genes (nodes) connected by expression (FPKM) correlations above a threshold value (edges). Nodes are connected by weighted lines, which represent correlations between similar expression profiles. Nodes are connected with each other if the Pearson correlation coefficient between them exceeds 0.85. A high correlation threshold of 0.85 was chosen in order to restrict the network analysis to



relationships between very similar expressions. This threshold can impact on the network characteristics which provide a more in-depth understanding of the gene expression profiles of the tissues.

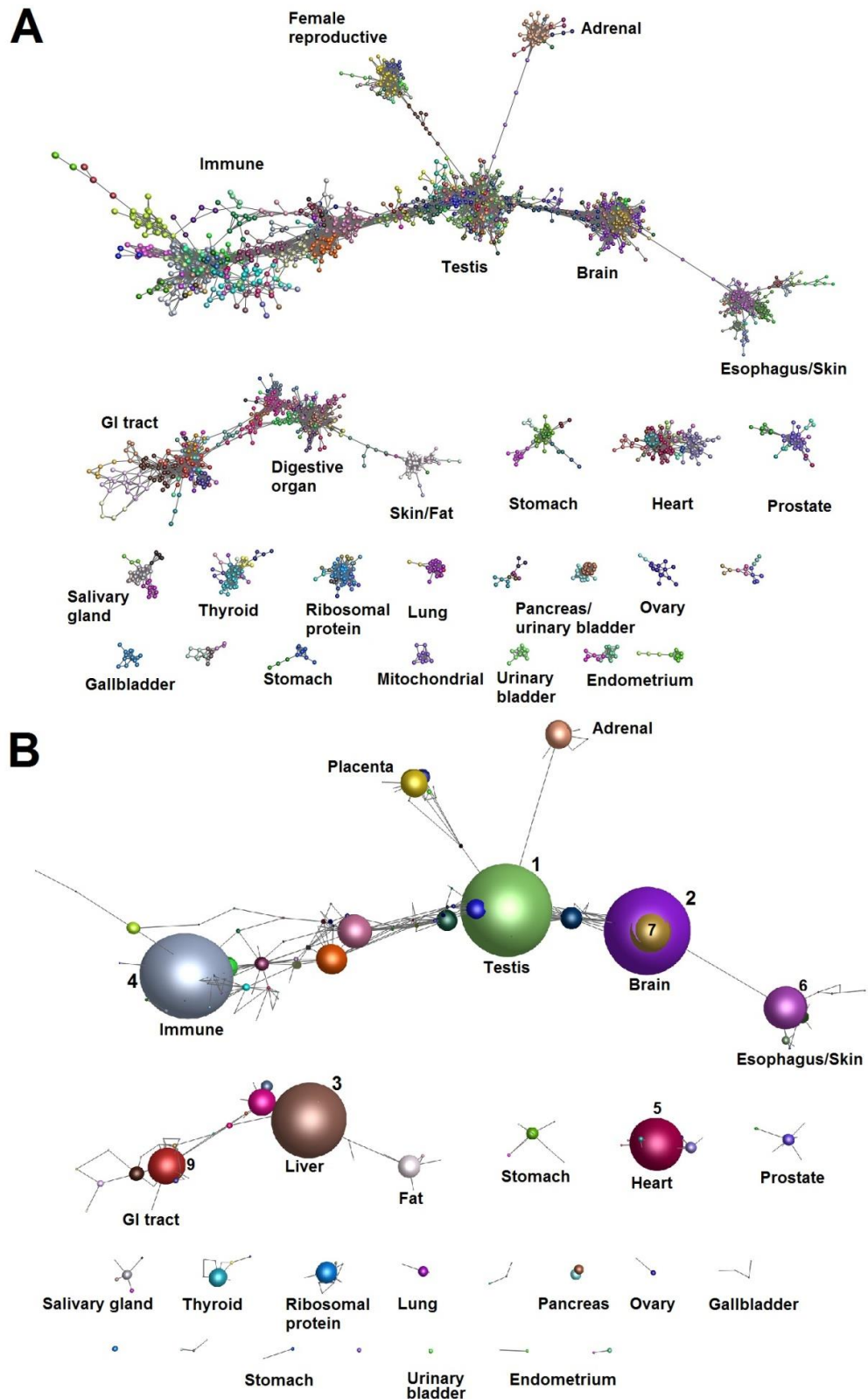
A network graph was constructed using a correlation threshold of  $r \geq 0.85$ , whereby nodes represent genes and edges represent correlation greater than the threshold. This threshold allows high correlated nodes included in the network. Increasing threshold will lose nodes or break the network while lowering threshold will tighten structure of the network. The network comprises 6,109 nodes (genes) and 1,091,477 edges (correlations) and network clustering performed using MCL algorithm within BioLayout *Express*<sup>3D</sup> at an inflation value of 2.2. Inflation affects the granularity of the resulting clustering. Lowering the inflation value will obtain coarser clusters while increasing it will produce more granular clusters. The MCL algorithm simulates stochastic flow through the network, iteratively enhancing flow to well-connected nodes at the expense of poorly connected nodes until a stable state occurs in which inherent network structure is revealed. An MCL inflation value of 2.2 was used as the basis for determining the granularity of network, i.e. the inflation value determines the size of individual clusters. This generated 20 clusters containing more than ten genes. Genes in a component smaller or less than ten were not included in this network. **Figure 4.3A** shows the network graph produced. Clustering of the network using Markov clustering algorithm inflation value of 2.2 resulted in 64 components. The largest component consisted of 4,498 genes, and it comprises of 240 clusters defined by MCL algorithm. The remainder of the network was a sparse topology and divided into various sizes of small clusters. While collapsed cluster diagram is another way of visualising at such transcriptional networks which one of the functions in BioLayout *Express*<sup>3D</sup>. An important feature of network graph is that clusters with similar expression profiles tend to form neighbourhoods. **Figure 4.3B** shows the relationship between cluster size and position of the cluster within the network.



**Figure 4.2: Clustering of the human tissue data sample.** A Pearson correlation matrix was prepared to compare data derived from all samples from human tissue atlas. A network was then constructed using only those sample-to-sample relationships where  $r \geq 0.85$ . Nodes represent samples and edges are coloured according to the strength of correlation (red = high correlation, blue = low correlation). Samples are coloured based on tissue type. Networks are shown (A) before and (B) after filtering outlier, samples which have lowest correlation within the tissue (thyroid gland, small intestine, colon, heart, spleen, lymph node, bone marrow, kidney, placenta, prostate, lung, testis). While the lowest read count within the tissue (colon and testis). Samples removed are marked with a red box. Samples with no correlations  $r < 0.85$  were not included in these networks.

**Table 4.1: List of samples removed.** Eighteen samples were removed based on a low correlation value or lowest read count within the same tissue.

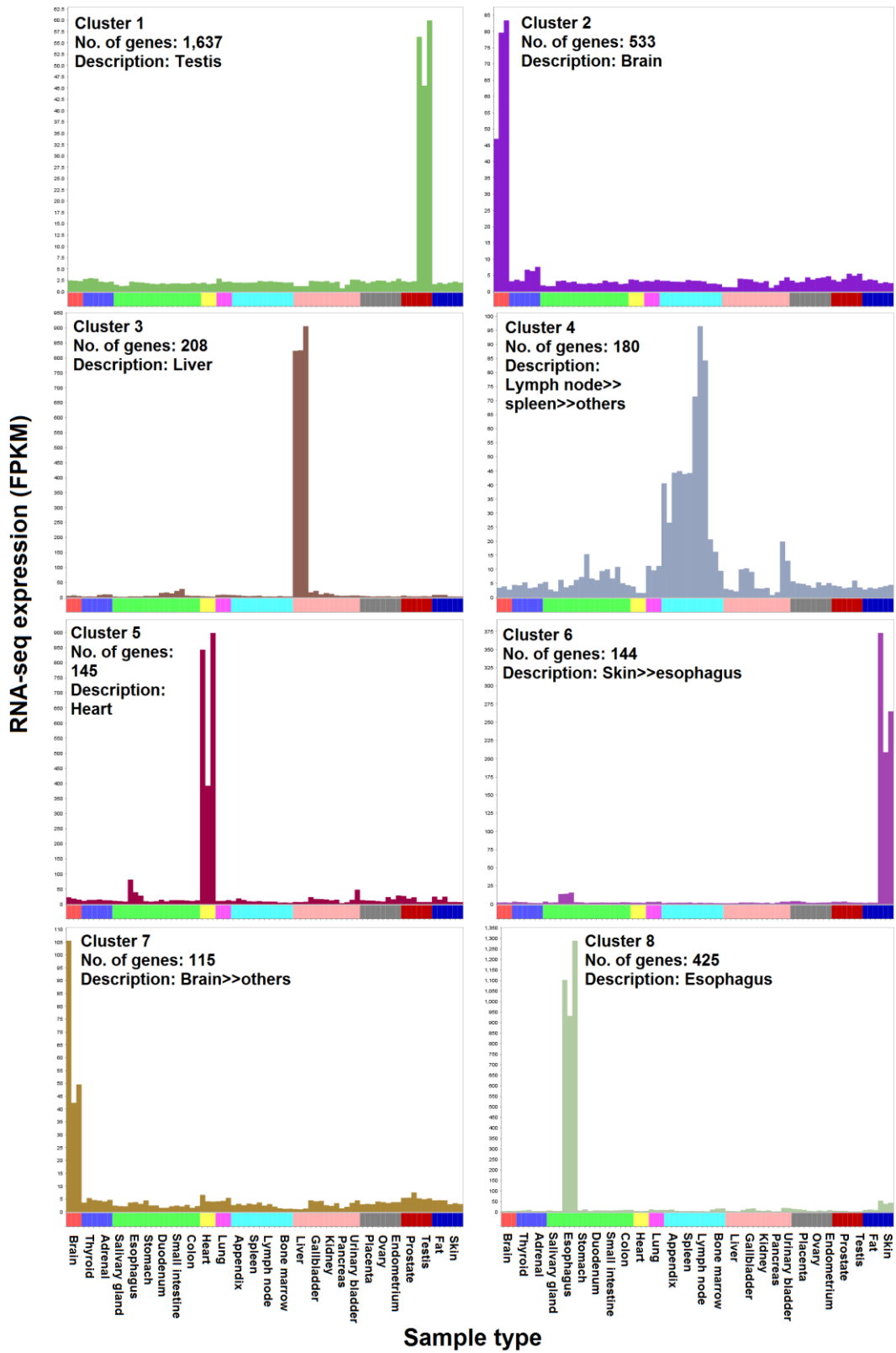
| Tissue type     | Sample ID              | Read count | Reason sample removed                   |
|-----------------|------------------------|------------|---|
| Thyroid gland   | thyroid_5a.V196        | 12,190,400 | Lowest correlation within tissue (0.65) |
| Small intestine | smallintestine_4a.V151 | 4,999,121  | Lowest correlation within tissue        |
| Colon           | colon_c.V14            | 20,278,760 | Lowest correlation within tissue (0.89) |
| Colon           | colon_f.V22            | 9,356,516  | Lowest read count                       |
| Heart           | heart_6a.V235          | 12,099,974 | Lowest correlation within tissue (0.73) |
| Lung            | lung_4d.V133           | 7,844,042  | Lowest correlation within tissue (0.93) |
| Lung            | lung_4a.V130           | 4,950,061  | Lowest read count                       |
| Spleen          | spleen_3b.V83          | 7,004,952  | Lowest read count                       |
| Lymph node      | lymphnode_5a.V190      | 9,987,765  | Lowest correlation within tissue (0.85) |
| Lymph node      | lymphnode_5b.V192      | 7,930,958  | Lowest correlation within tissue (0.85) |
| Bone marrow     | bonemarrow_5a.V230     | 8,885,586  | Lowest correlation within tissue (0.68) |
| Kidney          | kidney_b.V6            | 16,980,044 | Lowest correlation within tissue (0.87) |
| Placenta        | placenta_3a.V76        | 18,517,742 | Lowest correlation within tissue (0.61) |
| Prostate        | prostate_a.V12         | 11,162,655 | Lowest correlation within tissue (0.88) |
| Testis          | testis_7c.V257         | 22,389,166 | Lowest correlation within tissue (0.85) |
| Testis          | testis_7b.V256         | 20,609,541 | Lowest correlation within tissue (0.85) |
| Testis          | testis_7a.V255         | 12,174,081 | Lowest correlation within tissue (0.87) |
| Testis          | testis_4a.V134         | 5,183,071  | Lowest read count                       |

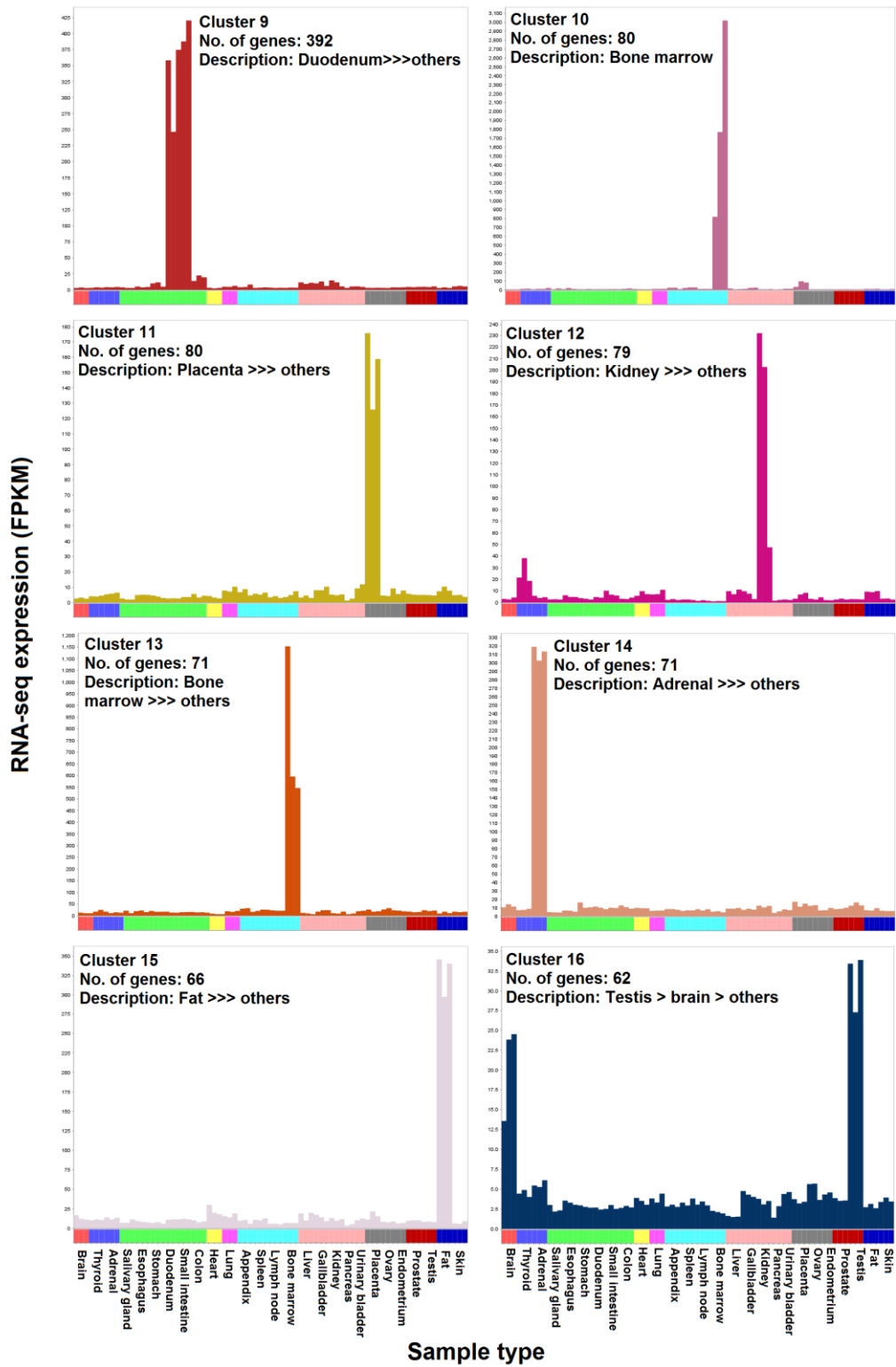


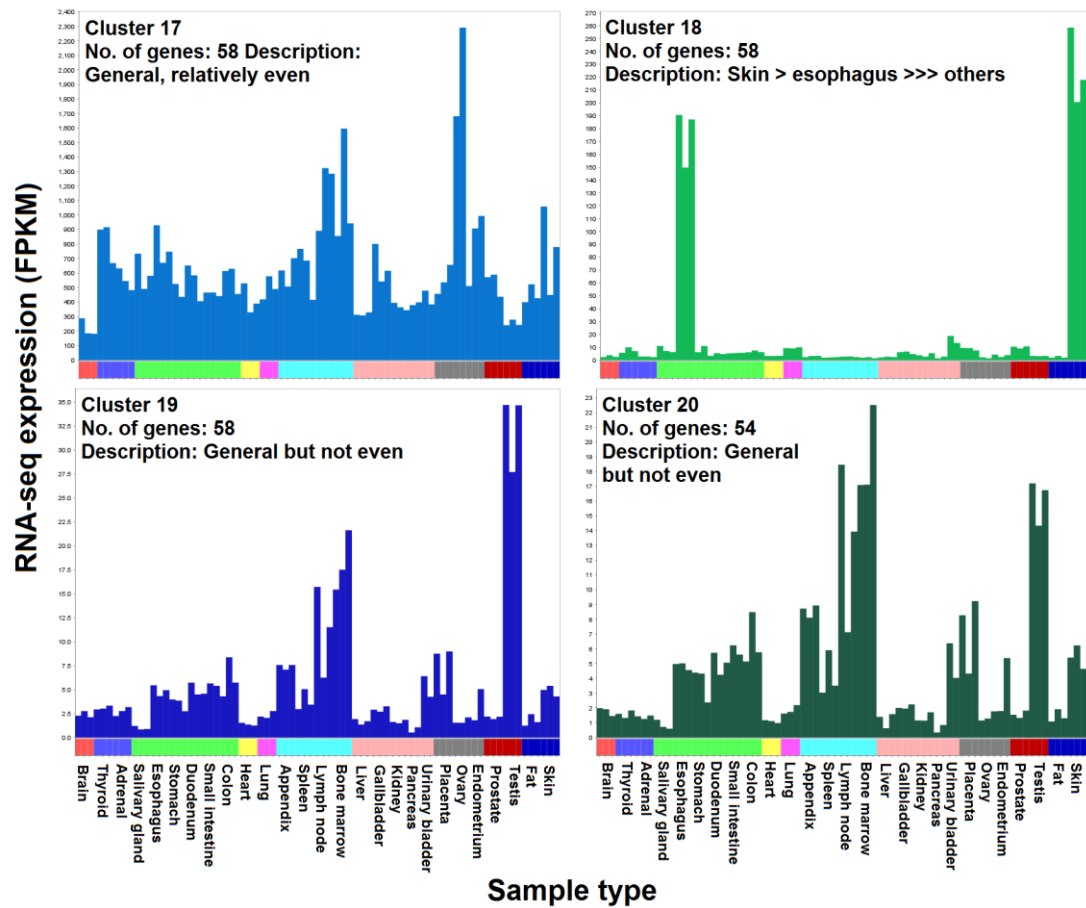
**Figure 4.3: (A) Network visualisation and clustering of the human tissue atlas RNA-seq data.** This is a 3D visualisation of a Pearson correlation graph of data generated from analysis of the human tissues. Each node (sphere) in the graph

represents a gene and edges (lines) between nodes represent correlations between individual measurements  $r \geq 0.85$ . This network comprises 6,109 nodes (genes) and 1,091,477 edges (correlations). This highly structured network is a result of co-expressed genes forming groups of a highly connected node within the network. MCL (Markov Cluster) algorithm value of 2.2 within BioLayout *Express*<sup>3D</sup> was used to assign and cluster co-expressed genes in the same group. **(B) Collapsed cluster diagram of human tissue expression atlas.** This is a simplified version of the network shown in **(A)**. Each node represents one of the 79 largest clusters of genes, the size of the node is proportional to the number of individual nodes (genes) within that cluster while edges represent correlations  $r \geq 0.85$  between clusters whereby nodes in one cluster share edges with nodes in another.

The expression profile and gene content of each cluster were examined in details, and top 20 largest clusters are shown in **Figure 4.4**. Some of these clusters represent genes co-expressed in a tissue-specific manner while others not. The network structure is derived from clustering of genes which are co-expressed and connected by a number of edges forming groups within the network. Gene set enrichment analysis was performed on clusters using DAVID (<http://david.abcc.ncifcrf.gov/>) analysis tools. Analysis using DAVID used the functional annotation clustering tool to identify and classify the clusters. The 20 largest clusters are shown in **Table 4.2**.







**Figure 4.4: Expression profile of top 20 clusters from the human tissue atlas RNA-seq data.** These expression profiles can be seen from a class viewer by selecting cluster in BioLayout *Express*<sup>3D</sup>. Histograms are an average expression level (FPKM) of all genes in the cluster.



**Table 4.2: List of 20 largest gene clusters.** Listed in the table are the 20 largest clusters of genes originating from the analysis of the human tissue atlas. Clusters are numbered according to their size (the largest is designated as cluster 1). The first two columns give cluster ID, and a number of genes present in the cluster and subsequent column describe an average expression profile of all genes within the cluster. The next column aims to group clusters according to the class of tissues and which these genes are predominately expressed and the tissues they represent. The last three columns aim to present the top three ontologies (GO) terms for each cluster together with their accession number and a *p*-value of enrichment.

| Cluster ID | Number of Genes | Profile Description          | Class of tissue   | GO terms                            | GO term accession number | <i>P</i> -value |
|------------|-----------------|------------------------------|-------------------|-------------------------------------|--------------------------|-----------------|
| 1          | 1887            | Testis                       | Male reproductive | Sexual reproduction                 | GO:0019953               | 2.49E-67        |
|            |                 |                              |                   | Spermatogenesis                     | GO:0007283               | 4.09E-66        |
|            |                 |                              |                   | Male gamete generation              | GO:0048232               | 5.63E-66        |
| 2          | 782             | Brain                        | CNS               | Nervous system development          | GO:0007399               | 5.27E-81        |
|            |                 |                              |                   | Chemical synaptic transmission      | GO:0007268               | 4.11E-58        |
|            |                 |                              |                   | Trans-synaptic signalling           | GO:0099537               | 4.11E-58        |
| 3          | 262             | Liver                        | Digestive         | Immune system process               | GO:0002376               | 7.06E-86        |
|            |                 |                              |                   | Immune response                     | GO:0006955               | 2.75E-73        |
|            |                 |                              |                   | Regulation of immune system process | GO:0002682               | 1.58E-71        |
| 4          | 262             | Lymph node>>spleen>>appendix | Immune            | Digestion                           | GO:0007586               | 5.69E-19        |
|            |                 |                              |                   | Digestive system process            | GO:0022600               | 6.75E-10        |
|            |                 |                              |                   | Xenobiotic metabolic process        | GO:0006805               | 1.82E-08        |
| 5          | 260             | Heart                        | Circulatory       | Protein activation cascade          | GO:0072376               | 4.99E-34        |
|            |                 |                              |                   | Organic acid metabolic process      | GO:0006082               | 8.75E-32        |
|            |                 |                              |                   | Small molecule metabolic process    | GO:0044281               | 2.72E-30        |
| 6          | 253             | Skin>>esophagus              | Skin              | Epidermis development               | GO:0008544               | 1.73E-54        |
|            |                 |                              |                   | Skin development                    | GO:0043588               | 2.20E-43        |
|            |                 |                              |                   | Keratinocyte differentiation        | GO:0030216               | 1.53E-40        |
| 7          | 227             | Brain>>others                | CNS               | Immune system process               | GO:0002376               | 1.36E-17        |
|            |                 |                              |                   | Defence response                    | GO:0006952               | 4.37E-13        |
|            |                 |                              |                   | Response to another organism        | GO:0051707               | 2.12E-11        |






| Cluster ID | Number of Genes | Profile Description      | Class of tissue         | GO terms  | GO term accession number | P-value   |
|------------|-----------------|--------------------------|-------------------------|---|--------------------------|-----------|
| 8          | 210             | Esophagus                | GI tract                | Cell cycle process  | GO:0022402               | 4.77E-95  |
|            |                 |                          |                         | Cell cycle  | GO:0007049               | 4.48E-93  |
|            |                 |                          |                         | Mitotic cell cycle process                                  | GO:1903047               | 5.38E-82  |
| 9          | 173             | Duodenum                 | GI tract                | Muscle structure development                                | GO:0061061               | 6.29E-36  |
|            |                 |                          |                         | Muscle contraction  | GO:0006936               | 1.13E-34  |
|            |                 |                          |                         | Muscle system process                                       | GO:0003012               | 2.96E-34  |
| 10         | 168             | Bone marrow              | Immune                  | Organic acid metabolic process                              | GO:0006082               | 6.86E-19  |
|            |                 |                          |                         | Small molecule metabolic process                            | GO:0044281               | 2.16E-17  |
|            |                 |                          |                         | Anion   | GO:0006820               | 4.54E-17  |
| 11         | 164             | Placenta                 | Dermal                  | Female pregnancy  | GO:0007565               | 3.26E-09  |
|            |                 |                          |                         | Multi-multicellular organism process                        | GO:0044706               | 3.11E-08  |
|            |                 |                          |                         | Reproductive process  | GO:0022414               | 4.55E-07  |
| 12         | 111             | Kidney                   | GI tract                | Keratinization  | GO:0031424               | 3.44E-16  |
|            |                 |                          |                         | Epidermis development                                       | GO:0008544               | 3.59E-16  |
|            |                 |                          |                         | Keratinocyte differentiation                                | GO:0030216               | 4.87E-15  |
| 13         | 71              | Bone marrow>>others      | Immune                  | Defence response to bacterium                               | GO:0042742               | 1.16E-07  |
|            |                 |                          |                         | Nucleosome assembly   | GO:0006334               | 1.43E-06  |
|            |                 |                          |                         | Chromatin assembly  | GO:0031497               | 4.18E-06  |
| 14         | 71              | Adrenal>>others          | Endocrine               | Hormone biosynthetic process                                | GO:0042446               | 1.19E-13  |
|            |                 |                          |                         | C21-steroid hormone metabolic process                       | GO:0034754               | 2.61E-10  |
|            |                 |                          |                         | Cellular hormone metabolic process                          | GO:0008207               | 5.09E-10  |
| 15         | 66              | Fat                      | Adipose                 | Metabolic process   | GO:0008152               | 1.51E-08  |
|            |                 |                          |                         | Lipid metabolism process                                    | GO:0006629               | 1.69E-08  |
|            |                 |                          |                         | Cellular lipid metabolic process                            | GO:0044255               | 1.77E-07  |
| 16         | 62              | Testis>Brain             | Cell-cell communication | Unclassified  | -                        | -         |
| 17         | 58              | General, relatively even | Pathway                 | SRP-dependent cotranslational protein targeting to membrane | GO:0006614               | 6.31E-119 |
|            |                 |                          |                         | Protein targeting to ER                                     | GO:0045047               | 1.72E-    |

| Cluster ID | Number of Genes | Profile Description             | Class of tissue    | GO terms                                      | GO term accession number | P-value   |
|------------|-----------------|---------------------------------|--------------------|---|--------------------------|-----------|
|            |                 |                                 |                    |   |                          | 117       |
|            |                 |                                 |                    | Cotranslational protein targeting to membrane | GO:0006613               | 4.96E-117 |
| 18         | 58              | Skin>esophagus>>urinary bladder | Epidermis          | Epidermis development                         | GO:0008544               | 6.84E-07  |
|            |                 |                                 |                    | Desmosome organization                        | GO:0002934               | 1.76E-02  |
|            |                 |                                 |                    | -   | -                        | -         |
| 19         | 58              | General but not even            | Cell-cycle related | Cell cycle process                            | GO:0022402               | 1.72E-39  |
|            |                 |                                 |                    | Cell cycle                                    | GO:0007049               | 1.83E-37  |
|            |                 |                                 |                    | Mitotic cell cycle process                    | GO:1903047               | 4.85E-28  |
| 20         | 54              | General but not even            | Cell-cycle related | Cell cycle process                            | GO:0022402               | 2.75E-41  |
|            |                 |                                 |                    | Cell cycle                                    | GO:0007049               | 1.79E-39  |
|            |                 |                                 |                    | Mitotic cell cycle process                    | GO:1903047               | 3.40E-38  |

### 4.3.3 Analysis of alternative splicing between tissue using rMATS

rMATS v3.2.2 (Shen et al. 2014) was used to identify alternatively spliced genes through comparison of three human tissues (brain, heart, and liver). These tissues have very different biology. By using a cut-off false-discovery rate (FDR) < 0.01, comparison of brain versus heart, brain versus liver and heart versus liver detected 4992, 4804 and 3990 splicing events respectively. At FDR < 0.001, around 1 in 1,000 alternative spliced genes was expected is false discoveries, thus from this fewer than ten false discoveries can be expected in the list of alternatively spliced genes. Therefore, the difference required between two isoforms to be biologically significant enough to define as differential splicing. Shown in **Figure 4.5** is a summary of AS events with a total number of splicing event and number of significant AS events detected by rMATS for each group of comparison. Several criteria were taken into consideration to select a gene list for the network analysis such as the lowest FDR and FPKM value of the gene. The output from rMATS from each of pairwise analysis (brain versus heart, brain versus liver and heart versus liver) were ranked according to p-value and FDR. To confirm these differences, top 30 AS genes for network-based visualisation structure confirmation mainly on their

potential difference across tissues shown in **Table 4.3**. This table was filtered with  $FDR < 0.01$  and sorted according to largest FPKM fold change from each of tissue comparison.

| AS events  | Brain vs. Heart    |                                | Brain vs. Liver    |                                | Heart vs. Liver    |                                |
|--|--------------------|--------------------------------|--------------------|--------------------------------|--------------------|--------------------------------|
|  | Total of AS Events | No. of significance AS events* | Total of AS Events | No. of significance AS events* | Total of AS Events | No. of significance AS events* |
| SE    | 19483              | 2960                           | 17973              | 2866                           | 15766              | 2259                           |
| MXE   | 1409               | 292                            | 1296               | 302                            | 1234               | 236                            |
| A3SS  | 3888               | 505                            | 3647               | 469                            | 3262               | 437                            |
| A5SS  | 2427               | 352                            | 2226               | 331                            | 2010               | 302                            |
| RI    | 3678               | 883                            | 3541               | 836                            | 3385               | 756                            |
| <b>Total</b>   |                    | <b>4992</b>                    |                    | <b>4804</b>                    |                    | <b>3990</b>                    |

**Figure 4.5: Summary of different types of significant AS events.** This AS events identified from comparison of three human tissue atlas; brain versus heart, brain versus liver and heart versus liver. SE skipped exon; MXE, mutually exclusive exon; A5SS, alternative 5' splice site; A3SS, alternative 3' splice site; RI, retained intron.

**Table 4.3: Differential splicing events in human tissue atlas ranked by FDR value (brain versus heart; brain versus liver; heart versus liver).** The output of rMATS was filtered out with  $FDR > 0.01$  and inclusion level difference  $|\Delta\psi| > 0.5$ . The first five columns give the gene, gene description, chromosome, the location of exon start and end. The next column aims to include the FDR value of rMATS analysis. The next column is exon inclusion level difference. A negative number means more inclusion and a positive number more exclusion of the sequence in tissue comparison. The exon inclusion level difference is an absolute, rather than relative, change in the percentage of a specific splicing isoform in all mRNAs produced from the parent gene that follows the indicated splicing pattern. Event Types: 1) A3SS: alternative 3' splice site 2) A5SS: alternative 5' splice site 3) MXE: mutually exclusive exons 4) RI: retained intron and 5) SE: skipped exon. The last four columns give the sample examined and the expression level in FPKM value.

| Gene          | Description                        | Chr | Exon Start | Exon End  | FDR       | Inclusion Level Difference ( $\psi_1 - \psi_2$ ) | AS Event | Sample 1 | Sample 2 | FPKM Sample 1 | FPKM Sample 2 |
|---------------|------------------------------------|-----|------------|-----------|-----------|--|----------|----------|----------|---------------|---------------|
| <i>KLC1</i>   | kinesin light chain 1              | 14  | 104145720  | 104153548 | 5.57E-308 | 0.547  | SE       | Brain    | Heart    | 218.3         | 39.9          |
| <i>FUS</i>    | FUS RNA binding protein            | 16  | 31196259   | 31199678  | 1.22E-292 | 0.64   | RI       | Brain    | Liver    | 129.2         | 44.9          |
| <i>TPM1</i>   | tropomyosin 1 (alpha)              | 15  | 63353067   | 63354476  | 3.83E-272 | -0.771   | MXE      | Heart    | Liver    | 6863.5        | 33.8          |
|               |                                    |     | 63354774   | 63358292  | 1.32E-224 | 0.532  | SE       | Heart    | Liver    | 6863.5        | 33.8          |
|               |                                    |     | 63353067   | 63353987  | 6.12E-116 | -0.732   | SE       | Heart    | Liver    | 6863.5        | 33.8          |
|               |                                    |     | 63353067   | 63354476  | 1.48E-43  | -0.592   | MXE      | Brain    | Liver    | 49.1          | 33.8          |
|               |                                    |     | 63353396   | 63354476  | 4.73E-38  | 0.572  | RI       | Heart    | Liver    | 6863.5        | 33.8          |
|               |                                    |     | 63353396   | 63354476  | 1.93E-28  | 0.559  | SE       | Brain    | Liver    | 49.1          | 33.8          |
|               |                                    |     | 63353067   | 63353987  | 8.90E-28  | -0.587   | SE       | Brain    | Liver    | 49.1          | 33.8          |
| <i>SORBS2</i> | sorbin and SH3 domain containing 2 | 4   | 186551702  | 186567936 | 2.94E-245 | 0.566  | SE       | Heart    | Liver    | 651.0         | 58.7          |
| <i>TPM3</i>   | tropomyosin 3                      | 1   | 154143124  | 154145454 | 1.23E-244 | -0.767   | MXE      | Heart    | Liver    | 75.7          | 58.1          |
| <i>GUK1</i>   | guanylate kinase 1                 | 1   | 228328018  | 228333325 | 2.88E-230 | 0.548  | SE       | Heart    | Liver    | 112.3         | 54.3          |

|                |   |    |           |           |           |        |     |       |       |        |       |
|----------------|---|----|-----------|-----------|-----------|--------|-----|-------|-------|--------|-------|
|                |   |    | 228327982 | 228333768 | 2.36E-19  | 0.558  | SE  | Heart | Liver | 112.3  | 54.3  |
|                |   |    | 228328018 | 228333325 | 1.74E-15  | 0.511  | SE  | Brain | Liver | 122.4  | 54.3  |
| <i>APP</i>     | amyloid beta (A4) precursor protein   | 21 | 27354656  | 27394358  | 1.21E-218 | -0.873 | SE  | Brain | Liver | 564.9  | 133.0 |
| <i>PDLIM5</i>  | PDZ and LIM domain 5  | 4  | 95497093  | 95506888  | 2.64E-197 | 0.944  | SE  | Heart | Liver | 1081.3 | 49.1  |
| <i>SLC25A3</i> | solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3 | 12 | 98987756  | 98991813  | 1.27E-189 | -0.659 | MXE | Brain | Heart | 145.4  | 467.9 |
| <i>TMED2</i>   | transmembrane emp24 domain trafficking protein 2                              | 12 | 124071293 | 124074993 | 2.04E-184 | 0.568  | SE  | Heart | Liver | 64.0   | 115.3 |
| <i>CAMK2D</i>  | calcium/calmodulin-dependent protein kinase II delta                          | 4  | 114421618 | 114430831 | 8.31E-143 | -0.677 | MXE | Brain | Heart | 42.8   | 90.6  |
|                |   |    | 114372187 | 114378719 | 1.01E-130 | -0.619 | SE  | Brain | Heart | 42.8   | 90.6  |
|                |   |    | 114421618 | 114429424 | 6.75E-79  | 0.722  | SE  | Brain | Heart | 42.8   | 90.6  |
| <i>GNAS</i>    | GNAS complex locus  | 20 | 57470666  | 57478640  | 1.43E-135 | 0.564  | SE  | Brain | Liver | 723.1  | 196.4 |
| <i>CLTB</i>    | clathrin, light chain B   | 5  | 175819455 | 175824719 | 1.62E-123 | 0.66   | SE  | Brain | Heart | 44.6   | 87.3  |
| <i>NDRG4</i>   | NDRG family member 4  | 16 | 58528867  | 58537807  | 6.66E-111 | 0.584  | SE  | Brain | Heart | 306.6  | 204.1 |
| <i>SEC31A</i>  | SEC31 homolog A ( <i>S. cerevisiae</i> )                                      | 4  | 83778841  | 83784545  | 1.64E-101 | 0.587  | SE  | Heart | Liver | 49.7   | 46.8  |
| <i>UGP2</i>    | UDP-glucose pyrophosphorylase 2   | 2  | 64068087  | 64083567  | 2.98E-96  | -0.618 | SE  | Brain | Liver | 59.1   | 243.6 |
| <i>RBM3</i>    | RNA binding motif (RNP1, RRM) protein 3                                       | X  | 48433948  | 48434471  | 2.72E-84  | 0.74   | RI  | Brain | Liver | 64.3   | 37.2  |
|                |   |    | 48433948  | 48434471  | 3.58E-15  | 0.686  | RI  | Heart | Liver | 91.7   | 37.2  |
|                |   |    | 48433948  | 48434807  | 9.94E-15  | 0.572  | RI  | Heart | Liver | 91.7   | 37.2  |
| <i>ABLIM1</i>  | actin binding LIM protein 1   | 10 | 116233637 | 116247775 | 6.38E-84  | -0.653 | SE  | Brain | Heart | 38.1   | 129.9 |

|                 |  |    |           |           |          |        |      |       |       |       |       |
|-----------------|--|----|-----------|-----------|----------|--------|------|-------|-------|-------|-------|
| <i>DST</i>      | dystonin   | 6  | 56328362  | 56330993  | 7.94E-81 | 0.523  | SE   | Brain | Liver | 64.1  | 36.4  |
|                 |  |    | 56328362  | 56330993  | 4.53E-45 | 0.701  | SE   | Heart | Liver | 46.7  | 36.4  |
|                 |  |    | 56393638  | 56394931  | 2.68E-25 | -0.563 | SE   | Brain | Heart | 64.1  | 46.7  |
| <i>KIAA1191</i> | KIAA1191   | 5  | 175782573 | 175788742 | 3.50E-79 | -0.514 | SE   | Brain | Heart | 59.3  | 46.6  |
| <i>CLTA</i>     | clathrin, light chain A                                    | 9  | 36204064  | 36210657  | 5.53E-78 | 0.849  | SE   | Brain | Heart | 81.0  | 37.7  |
| <i>ACTN4</i>    | actinin, alpha 4   | 19 | 39200034  | 39205201  | 7.03E-76 | -0.638 | MXE  | Brain | Liver | 93.9  | 75.8  |
| <i>TPD52L1</i>  | tumor protein D52-like 1                                   | 6  | 125574862 | 125584208 | 4.44E-74 | -0.756 | SE   | Heart | Liver | 71.4  | 33.8  |
|                 |  |    | 125574862 | 125584372 | 9.67E-17 | 0.664  | SE   | Brain | Heart | 43.4  | 71.4  |
|                 |  |    | 125574862 | 125584208 | 5.86E-14 | 0.778  | SE   | Brain | Heart | 43.4  | 71.4  |
| <i>DCAF6</i>    | DDB1 and CUL4 associated factor 6                          | 1  | 167973770 | 168007726 | 1.10E-72 | -0.687 | SE   | Brain | Heart | 33.3  | 71.1  |
| <i>MACF1</i>    | microtubule-actin crosslinking factor 1                    | 1  | 39715685  | 39720047  | 1.11E-70 | -0.748 | SE   | Brain | Heart | 41.6  | 33.1  |
| <i>QKI</i>      | QKI, KH domain containing, RNA binding                     | 6  | 163987752 | 163984751 | 4.64E-69 | 0.647  | A3SS | Brain | Heart | 199.4 | 140.3 |
| <i>ANK2</i>     | ankyrin 2, neuronal  | 4  | 114294514 | 114302672 | 5.77E-69 | -0.504 | SE   | Brain | Heart | 79.9  | 60.7  |
|                 |  |    | 114294472 | 114304888 | 5.01E-50 | -0.554 | SE   | Brain | Heart | 79.9  | 60.7  |
| <i>MFF</i>      | mitochondrial fission factor                               | 2  | 228205007 | 228212100 | 1.80E-65 | -0.543 | SE   | Brain | Heart | 44.8  | 30.5  |
|                 |  |    | 228205007 | 228220477 | 6.95E-34 | -0.571 | MXE  | Brain | Heart | 44.8  | 30.5  |
| <i>PKIG</i>     | protein kinase (cAMP-dependent, catalytic) inhibitor gamma | 20 | 43160425  | 43218507  | 1.62E-61 | -0.62  | SE   | Brain | Heart | 56.4  | 172.7 |
| <i>CDK5RAP3</i> | CDK5 regulatory subunit associated protein 3               | 17 | 46050884  | 46051397  | 6.29E-61 | -0.573 | RI   | Brain | Heart | 49.6  | 31.1  |

### 4.3.4 Network analysis of alternative splicing transcripts of human tissue atlas

At threshold  $|\Delta\psi| > 50\%$ , false discovery rate (FDR) of 0.01 and FPKM  $\geq 30$ , rMATS (see Chapter 1 Section 1.5.8) identified 78 differential AS events in three pairwise comparisons; brain vs heart, brain vs liver and heart vs liver, using both splice junction counts and exon body counts as an input for rMATS. To demonstrate the utility of network-based visualisation, a network of four genes were examined using NGS Graph Generator (as described in Chapter 2) using default parameters ( $p=98$   $l=31$ ). All genes in **Table 4.3** have been laid out using BioLayout *Express*<sup>3D</sup>, however, only four networks (*KLC1*, *GUK1*, *SORBS2*, and *TPM1*) were selected to be studied due to the interesting structure of network, and they represent different type of AS and features, e.g. skipped exon (SE), mutually exclusive exons (MXE), alternative 5' splice site (A5SS) and length of the gene. In certain cases, unquify option was used to remove redundant reads of highly expressed genes, e.g. *TPM1* before network construction. In order to compare and contrast the network-based visualisation to other methods, Sashimi plots and Vials diagrams are presented.

#### 4.3.4.1 Analysis of *KLC1*

*KLC1* was selected because it was reported as the most significant alternatively spliced exon detected using rMATS tools when comparing between the brain and heart tissue. rMATS analysis for *KLC1* shows exon 15 to be skipped with the lowest FDR value of  $5.57E-308$ . In the tissue comparison of the brain vs heart, the exon inclusion level of exon 15 is 0.9 in the brain compares only 0.35 in the heart (**Table 4.3**). **Figure 4.6A** is an expression profile of *KLC1* across 27 human tissues demonstrating that *KLC1* is highly expressed in the brain, with relatively lower expression in other tissues. This observation is consistent with protein expression data (Uhlen et al., 2010). *KLC1* has been shown to play a crucial role in organelle transport in the brain. Genetic variation in the transport of protein KLC1 may contribute to the risk of Parkinson's disease (PD), Alzheimer's disease (AD) and cataract (Andersson et al., 2007; Dhaenens et al., 2004; von Otter et al., 2010). Furthermore, *KLC1* transcript variant expression might also function in the development of axoplasmic transport defects in AD (Morel et al., 2011).



In order to verify the result from rMATS, each transcript network from brain and heart are presented. *KLC1* has 21 exons, and a large number of potential isoforms – 27 in total, 20 protein-coding transcript isoforms, with a further seven transcript non-coding isoforms are recorded in Ensembl. Underneath each network, schematic gene models of *KLC1* in the brain and heart are presented. The number of reads mapped to *KLC1* in the brain and heart is 8,726 reads and 1,105 reads, respectively. These two transcript networks consist of a start exon 2, exon 3-14. At the c-terminus, the last exon of each the transcript has the choice to terminate at exon 14, 15, 16, 17 or 21. To further illustrate *KLC1* splicing across different tissues, the transcript network from liver was included.

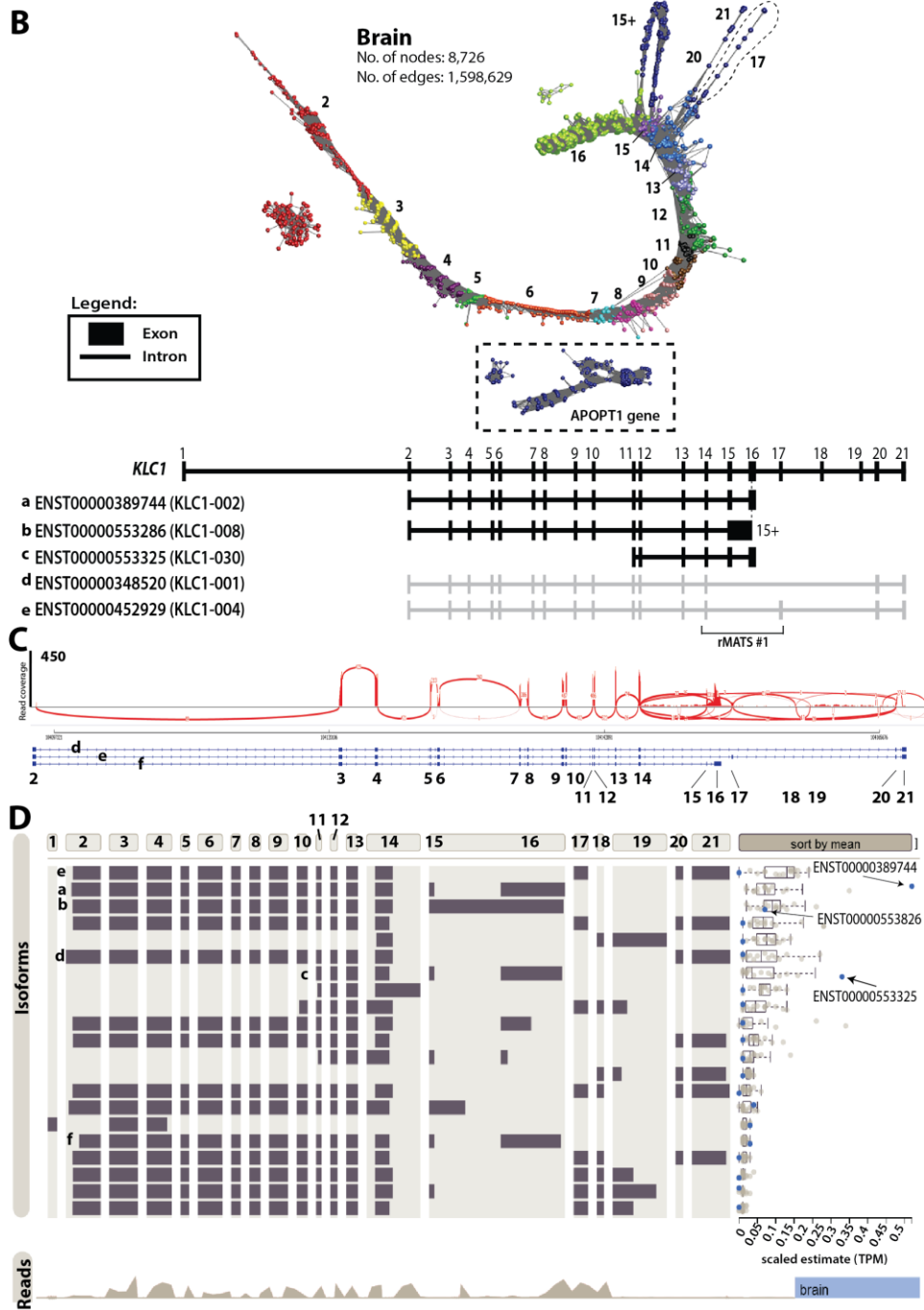
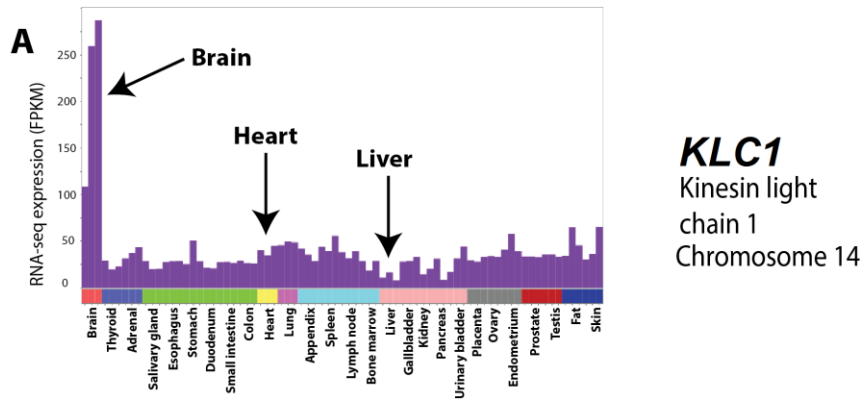
Network-based analysis of the *KLC1* transcripts co-expressed in the brain (**Figure 4.6B**) shows there were three major and two minor isoforms expressed. The first major isoform is indicated by the structure from exon 2 to 16 (ENST00000389744 – designated as **a** on the gene model below the brain DNA mapping transcript network). The second major isoform can be observed from the loop of the exon 15+ (ENST00000553286 - **b**) which is not the alternatively spliced but the larger size of the exon 14 and overlaps to the middle of exon 16. The last major isoform is believed to be expressed (ENST00000552325 - **c**).

The minor isoforms can be observed from a small number of nodes emerging out from the major network. From the bifurcation of the network, it was identified as minor isoforms of *KLC1* in the brain tissue which are ENST00000348520 (exons 2 to 14 and the last exon of 17 ENST00000348520 – **d**) and ENST00000347839 (exon 2-14 and 20, 21 – **e**). These were supported by the brain RNA-seq mapping transcript network structure. Another gene *APOPT1* was present in this network.

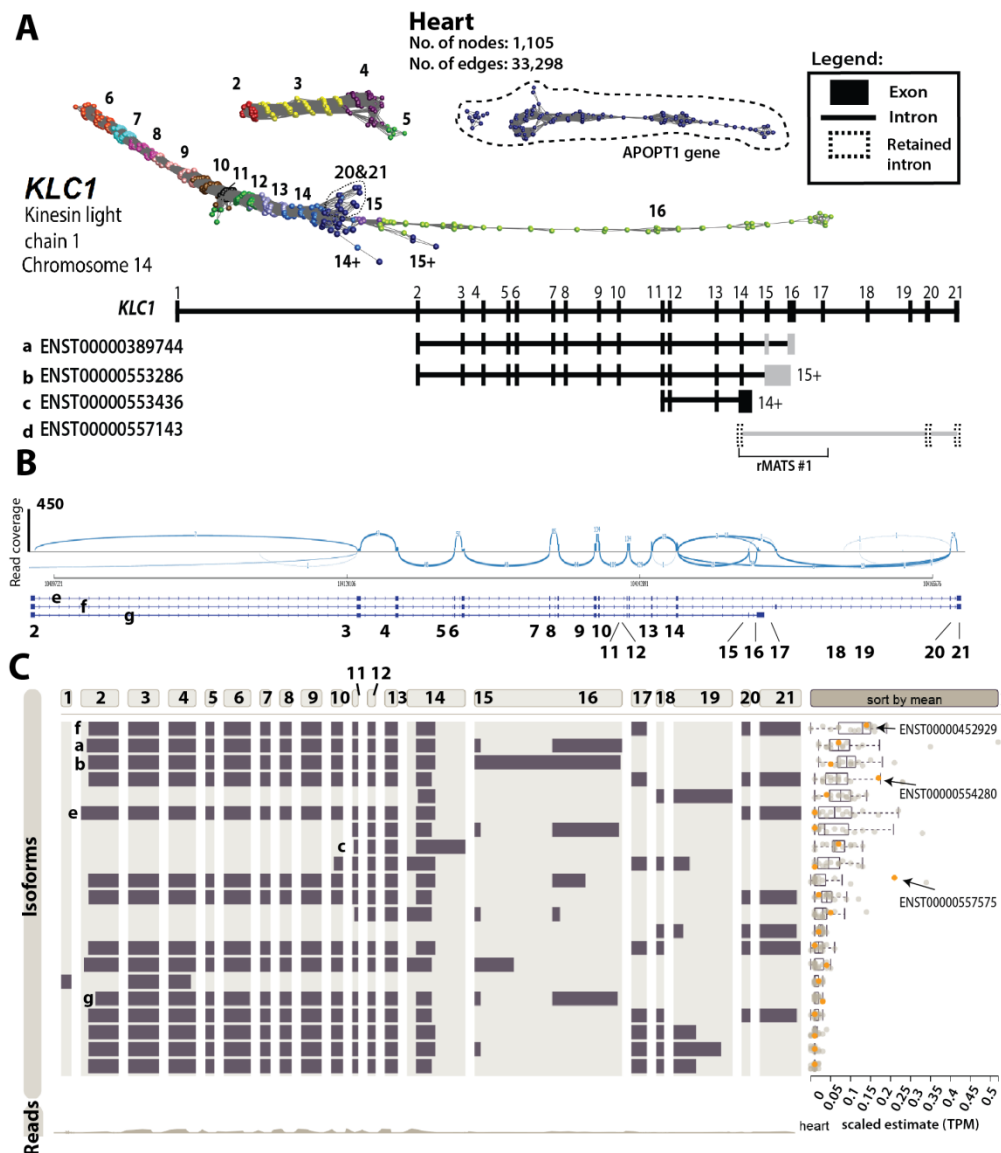
**Figure 4.6C** is a Sashimi plot generated using IGV showing exon expression and junction support of the brain and heart RNA-seq data. The visualisation of isoform expression with the gene model at the bottom of the plot can be visualised. There are only three isoforms reported in IGV which are ENST00000348520 (**d**), ENST00000452929 (**e**) and ENST00000389744 (**f**). Junctions are shown by the arcs

join from one exon to another exon is barely perceivable especially the edges connecting in the high-density area, e.g. the edges connecting in the region of exon 15 are difficult to explain. Isoform **d** and **e** are only observed in the brain RNA-seq mapping transcript network. It is demonstrated that IGV did not report the major isoform.

**Figure 4.6D** shows the Vials visualisation *KLC1* of the human transcriptome of the brain from Illumina BodyMap 2.0 data. The visualisation of all isoforms, isoform abundance, expression and junction views are shown for brain tissue. In this view, the expression level displays as transcript per million (TPM) and the highest isoforms detected in the brain (blue) is **(a)** ENST00000389744 (TPM=0.47) follow by **(b)** ENST00000553826 (TPM=0.32) and **(c)** ENST00000553325 (TPM=0.14). It is interesting to note that the agreement of RNA-seq expression between a network analysis and Vials. It indicates that the major isoforms expressed visualised in network analysis **(a, b, and c)** displays as the highest expression in Vials. However, isoforms **d, e, and f** reported in IGV displays low expression in Vials and minor isoforms in network visualisation.



**Figure 4.6: Visualisation of *KLC1* transcript in the human brain.** (A) **RNA-seq gene expression profile across 27 different tissues.** *KLC1* is highly expressed in the brain and lowly expressed in the heart. (B) **Network-based visualisation of *KLC1* in the brain.** All exons showed at the top of the isoforms gene model and expressed in the brain. In this network, it shows three major isoforms expressed in the brain tissue (gene model in black). Other isoforms are minor isoforms (gene model in grey) which can be visualised from the fewer nodes branching out of the major network. ‘Loop’ network structure that separates from the major network essentially is *APOPT1* gene (dashed box). (C) **Sashimi plots.** Representative Sashimi coverage plots generated in IGV showing RNA-seq reads mapping to *KLC1* locus from human brain (red) and heart (blue). IGV reports three isoforms in this view. The height of the bars represents overall read coverage. Splice junctions are displayed as loops. The number of reads observed for each junction is indicated by segments, and y-axis (450) ranges for the number of reads per exon base are shown (read coverage, left). The plot suggests different isoforms expressed in the sample is indicated by the arc connecting a pair of exons. (D) **Vials – visualising AS of genes.** Data shown here are from the Illumina BodyMap 2.0. There are two views which are isoform abundance and expression view. For each row in the isoform abundance view represent a particular isoform. Dot plots indicate abundance for brain isoforms. The tissue of brain was selected and multiple dot plots (blue) shown to allow comparison between isoforms. Two highest isoforms expression in the brain are ENST00000389744 (TPM=0.47) and ENST00000553325 (TPM=0.28). TPM, transcript per million.



**Figure 4.7: Visualisation of AS gene of *KLC1* in the human heart. (A) Network-based visualisation of gene *KLC1* in the heart tissues.** In this network, it indicates three isoforms expressed (gene model in black) in the heart tissue. Other isoforms are retained intron isoforms which can be visualised from the fewer nodes branch out of the major network. None of these isoforms is considered as major isoform except the exon 2 to 14, and there is no isoform of these exons currently available. All the bifurcation with a low number of nodes immediately after exon 14 and they are considered as minor reads. Splice variant network structure that separates from the major network essentially is *APOPT1* gene (dashed box). In agreement with rMATS analysis, the network indicates exon 15 is a low number of nodes in the heart. **(C) Sashimi plot.** Only three isoforms reported by IGV from this view. **(D) Vials – visualising AS of genes.** Three highest isoforms expression in the heart are ENST00000557575 (TPM=0.21), ENST00000554280 (TPM=0.17) and ENST00000452929 (TPM=0.14). TPM, transcript per million.

The analysis of *KLC1* transcript network of the heart, four isoforms expressed is shown in **Figure 4.7**. All these isoforms have different c-terminus of exon 14, 15 or 16 which can be observed from split branches at the end of the network. The first isoform (ENST00000389744 – **a**) is believed to be expressed can be explained by the last exon 16 emerge from the major network. This isoform contains exon 2 to 16. The second and third isoform (ENST00000553286 – **b** and ENST00000553436 – **c**) can be explained by small bifurcation of exon 15+ and 14+, respectively. However, the lower number of nodes of the last exon (grey exon) of all isoforms except from exon 2 to 14, none of them is considered as major isoforms. There is no isoform contains exon 2 to 14 can be retrieved from Ensembl database. A retain intron (ENST00000557143 – **d**) can be visualised from the small branch out of exon 20 and 21. As in the brain, an *APOPT1* was also present in this network which will be explained in **Figure 4.9**.

**Figure 4.7B** is Sashimi plots generated using IGV displays exon expression and junction support of the heart can be visualised. Three isoforms reported in IGV which are ENST00000348520 (**e**), ENST00000452929 (**f**) and ENST00000389744 (**g**) indicate similar expression with the brain. The junctions are shown by the arcs align from exon 2 to 14 are apparent and support the network analysis. This can be observed from the spiral nodes from exon 2 to 14. The junction following exon 14 apparent to distinguish compared to the brain due to lower of *KLC1* expression in the heart. None of the isoforms identified in the heart RNA-seq mapping transcript network is reported in the IGV.

**Figure 4.7C** shows the Vials visualisation *KLC1* of the human transcriptome of heart from Illumina BodyMap 2.0 data. The visualisation of isoforms, isoform abundance, isoform expression and junction views are shown for brain tissue. In this view, the expression level displays as transcript per million (TPM) and the highest isoforms detected in the heart (orange) are ENST00000557575 (TPM=0.21) follow by ENST00000554280 (TPM=0.17) and ENST00000452929 (TPM=0.14). It is interesting that there is disagreement between network analysis and Vials visualisation where three highest RNA-seq expressions in Vials are not observed in

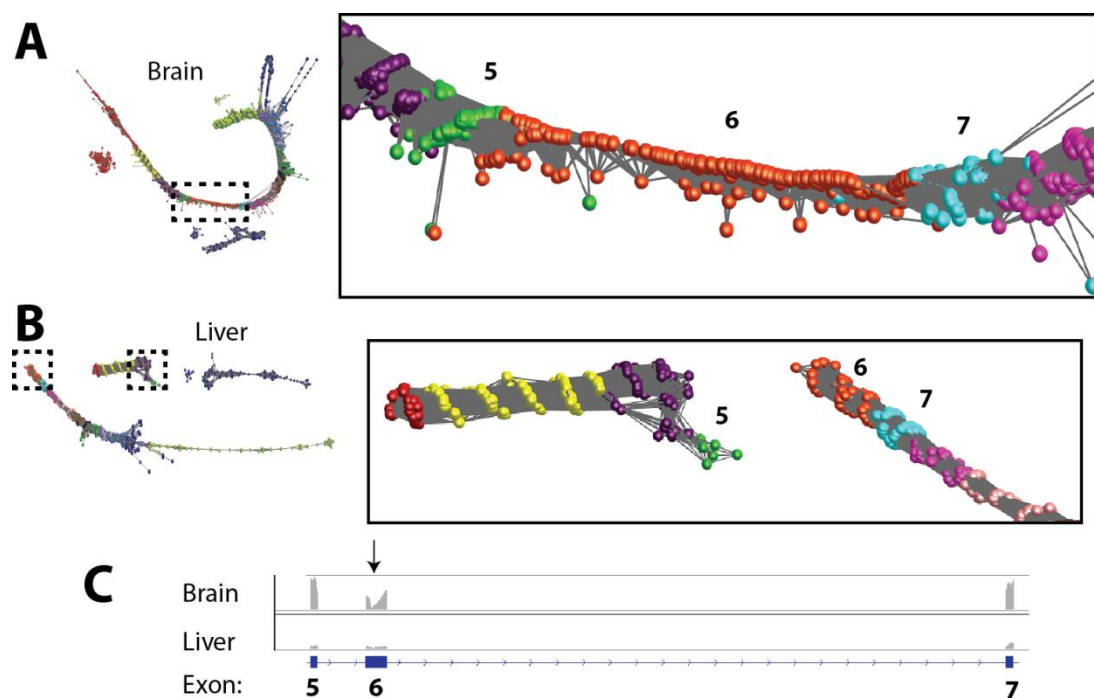
the network. The highest expression in Vials shows by the ENST00000557575 contains a truncated exon of 16. However, there is no evidence of skipping exon 15 in the network. The second and third highest isoform (ENST00000554280 and ENST00000452929 – **f**) expressed contain exon 17. However, there is no exon 17 observed in the network. While the agreement between Vials and IGV on isoform (**a**), except the lower number of reads on this exon, caused the network does not support the mapping on exon 17.

**Table 4.4** shows a tissue comparison of *KLC1* between brain and heart in two different visualisation approaches. Network-based analysis of the *KLC1* gene transcripts co-expressed in the brain revealed that there was one major isoform expressed in this tissue disagree with Sashimi plot. However, another one major isoform of network agrees with three isoforms in Sashimi plot. Furthermore, there are two isoforms cannot be seen from the network not shown in Sashimi plots.

**Table 4.4: Summary of visualisation analysis of *KLC1* between the brain and heart tissue.**

| Visualisation approach/Tissue | Brain   | Heart   |
|-------------------------------|---|---|
| Network                       | <p><b>Major</b><br/>(3 major isoforms)<br/>a - ENST00000389744<br/>b - ENST00000553286<br/>c - ENST00000553325</p> <p><b>Minor</b><br/>(2 minor isoforms)<br/>d - ENST00000348520<br/>e - ENST00000452929</p> | <p><b>Major</b><br/>(3 major isoforms)<br/>a - ENST00000389744<br/>b - ENST00000553286<br/>c - ENST00000553436</p> <p><b>Minor</b><br/>(1 minor isoforms)<br/>d - ENST00000557143</p> |
| Sashimi plots                 | <p>- Two isoforms (d, e) – shown in the network</p> <p>- One isoform (f) – not shown in the network</p>   | <p>- Three isoforms (e, f, g) – none show same with network</p> <p>- only agree with isoform (a) in Sashimi plot</p>  |

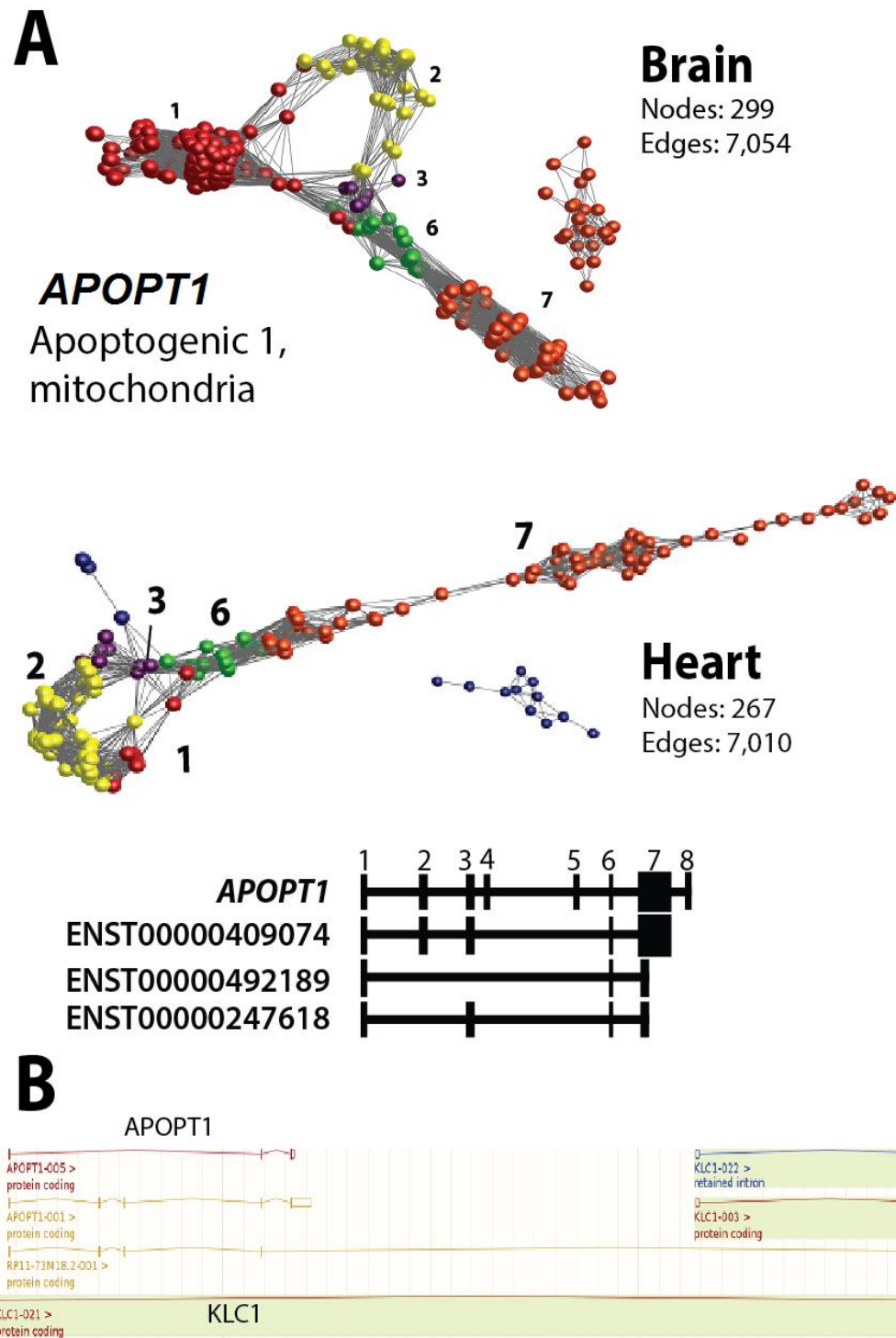
There are issues with network RNA-seq assembly observed in both brain (**Figure 4.8A**) and liver (**Figure 4.8B**) RNA-seq assembly transcript network. In the brain network, a ‘thinning’ structure appears at the exon 6 and no ‘corkscrew’ structure was observed. The lower number of reads of this exon exhibits this structure, and it appeared to be an immediate drop in sequence coverage (marked by arrow) of exon 6 shows in IGV (**Figure 4.8C**). Another issue in the liver RNA-seq mapping transcript network is the structure disconnect between exon 5 and 6. There is evidence that the low sequence coverage at the exon 5 and 6 as seen in IGV (**Figure 4.8C**).



**Figure 4.8: Issue with DNA read-assembly. (A) Network-based visualisation of *KLC1* transcript in the brain tissue.** The network of exon 6 shows ‘thinning’ caused by the immediate reduction of sequence coverage in the middle of exon 6 observed in IGV (C) (marked by arrow). **(B) Network-based visualisation of *KLC1* transcript in the liver.** Network breaks between exon 5 and 6 in the heart RNA-seq mapping transcript network due to the low sequence coverage in exon 5 and 6 also can be observed in IGV (C).



When the network assembly of *KLC1* was visualised in the brain and heart, it indicates a splicing structure (dashed box in **Figure 4.7A** and **Figure 4.8A**). However, none of the *KLC1* isoforms overlaid to this structure of both networks. In order to explain the observed separate graph in the network structure of *KLC1*, the location of those reads was investigated through visualising the genome location in Ensembl. It turns out that Apoptogenic 1, mitochondrial (*APOPT1*) gene that encodes a mitochondrial protein and is located within the *KLC1* locus of chromosome 14 (**Figure 4.9A**). When the process of extracting mapped read of *KLC1* using NGS Graph Generator, the reads of *APOPT1* were included for building network-based on the location of the gene. The location of *APOPT1* just 69 bp downstream of the *KLC1*, hence reads mapped to this location including *APOPT1* were included (**Figure 4.9B**). These gave rise to the separated splicing structure of *KLC1* network. The ‘loop’ of *APOPT1* network indicates an AS of exon 2. There were three isoforms expressed in both brain and heart tissue.



**Figure 4.9: Network-based visualisation of *APOPT1* transcript in the brain and heart.** The *APOPT1* loci are located within the *KLC1* locus resulting the read assembly of *APOPT1* includes in the *KLC1* network. The start exon of *APOPT1* is 69 bp downstream the first exon of *KLC1*. The bifurcation of the networks is mostly derived from the processed transcript.

#### 4.3.4.2 Analysis of *GUK1*

*GUK1* appeared to be alternatively spliced at exon 2 in two different tissue comparisons which were; heart vs liver and brain vs liver by rMATS analysis (**Table 4.3**). Therefore, the isoforms expressed in three tissues; brain, heart, and liver from network analysis are presented. The inclusion level in the heart is 0.9 and 0.35 in the liver, for tissue comparison of heart vs liver; whereas the inclusion level in the brain is 0.83 and 0.28 in the liver, for tissue comparison of brain vs liver. **Figure 4.10A** is an expression profile of 27 human tissues demonstrates *GUK1* is expressed in all tissues. This observation is consistent with protein expression data (Uhlen et al., 2010). Also, protein *GUK1* has been shown to play a vital role to catalyse the transfer of a phosphate group from ATP to guanosine monophosphate (GMP), to form guanosine diphosphate (GDP). AS of *GUK1* results in multiple transcript variants with at least one of which encodes a guanylate kinase protein to produce the mature protein. The encoded protein is assumed to be a good target for cancer chemotherapy (Young et al., 2008). Several transcript variants encoding different isoforms have been found for this gene. A study by Joehanes et al. (2013), discovered that *GUK1* is associated with coronary heart disease (CHD). In their study, to identify transcriptomic biomarkers of CHD in 188 cases with CHD, they found a total of 35 genes were differentially expressed in cases with CHD versus controls including *GUK1*.

*GUK1* has 9 exons and a large number of potential isoforms. Hence, 40 in total; 13 protein-coding transcript isoforms with a further 27 processed transcripts are recorded in Ensembl. Three network structures of the *GUK1* in three human tissues (brain, heart, and liver) were analysed through the public human RNA-seq data, using a network-based approach. The number of nodes in each network of brain, heart, and liver is 1,970 reads, 804 reads, and 207 reads respectively. All these networks consisted of an alternative promoter start exon 1a, 1b, 1c or 1d and common exon 2-9.

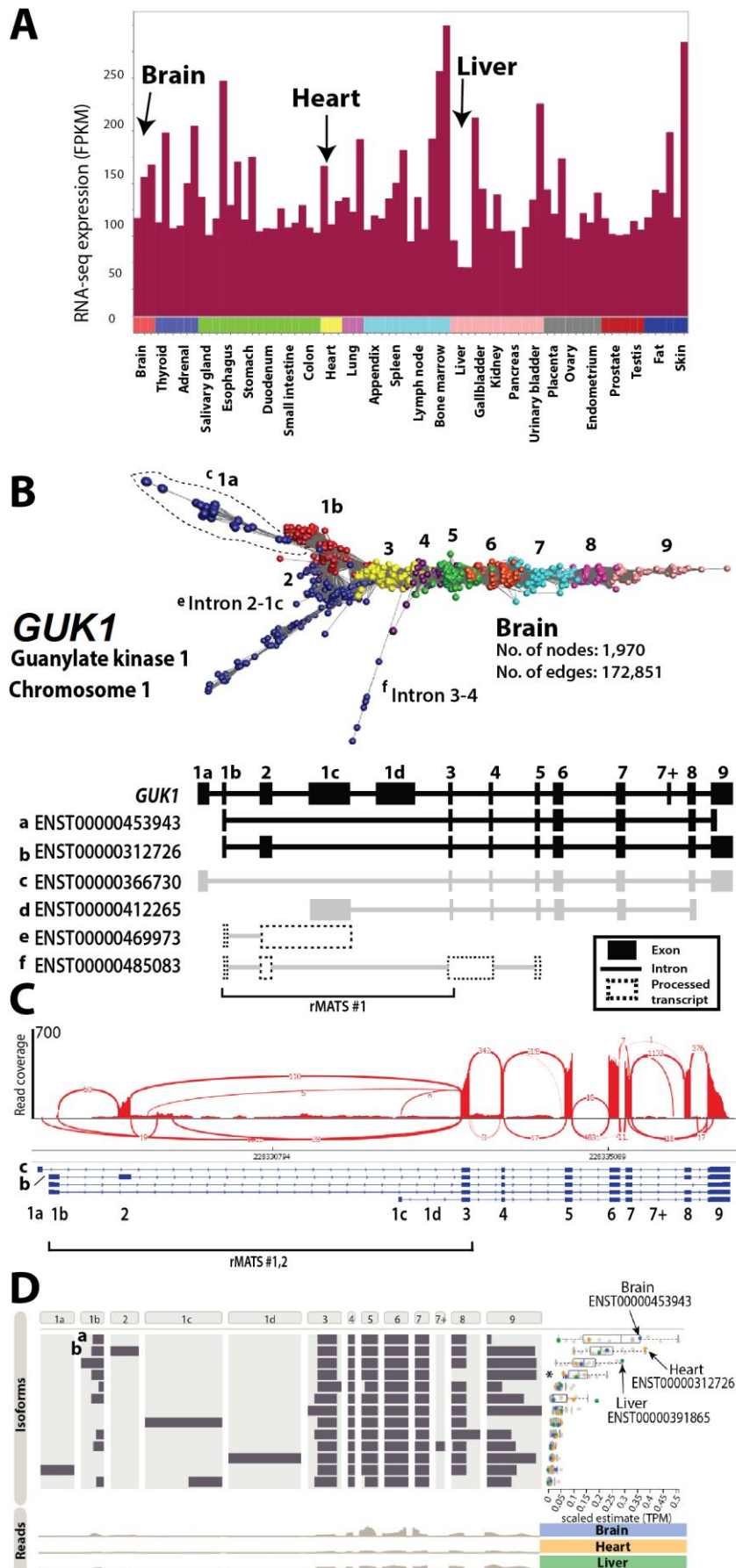
Network-based analysis of the *GUK1* transcripts co-expressed in the brain, two major isoforms (protein-coding) and four minor isoforms (3 protein-coding and 2

processed transcripts) are believed to be expressed in the brain (**Figure 4.10B**). *GUK1* in the brain tissue contains exon 1a, 1b, 1c, 2, 3, 4, 5, 6, 7, 8 and 9, and it supports the brain RNA-seq mapping transcript network structure. In this network, the truncated exon 9 can be visualised from the thin nodes at the end of the network.

The network revealed that there were two major isoforms expressed in this tissue. There are ENST00000453943 (**a**) contains exon 1b, 3, 4, 5, 6, 7, 8 and 9 (truncated) and ENST00000312726 (**b**) contains exon 1b, 2, 3, 4, 5, 6, 7, 8 and 9. However, there appeared to be evidence of four minor isoforms which are two proteins coding (ENST00000366730 – **c** and ENST00000412265 – **d**) and two processed transcripts (ENST00000469973 – **e** and ENST00000498092 – **f**) expressed in this tissue. All these four minor isoforms can be visualised from the small branch nodes emerge from the network.

The Sashimi plots indicate a quantitative visualisation of the RNA-seq read alignment of the brain, heart, and liver together with expression in FPKM value (**Figure 4.10C**). The junctions indicated by the arcs that align from one exon to another exon. The visualisation of read alignment with the gene model on the bottom of the plot can be visualised. Only isoform **c**, **d**, and **e** indicate in the Sashimi plot while another two models show disagreement with network structure.

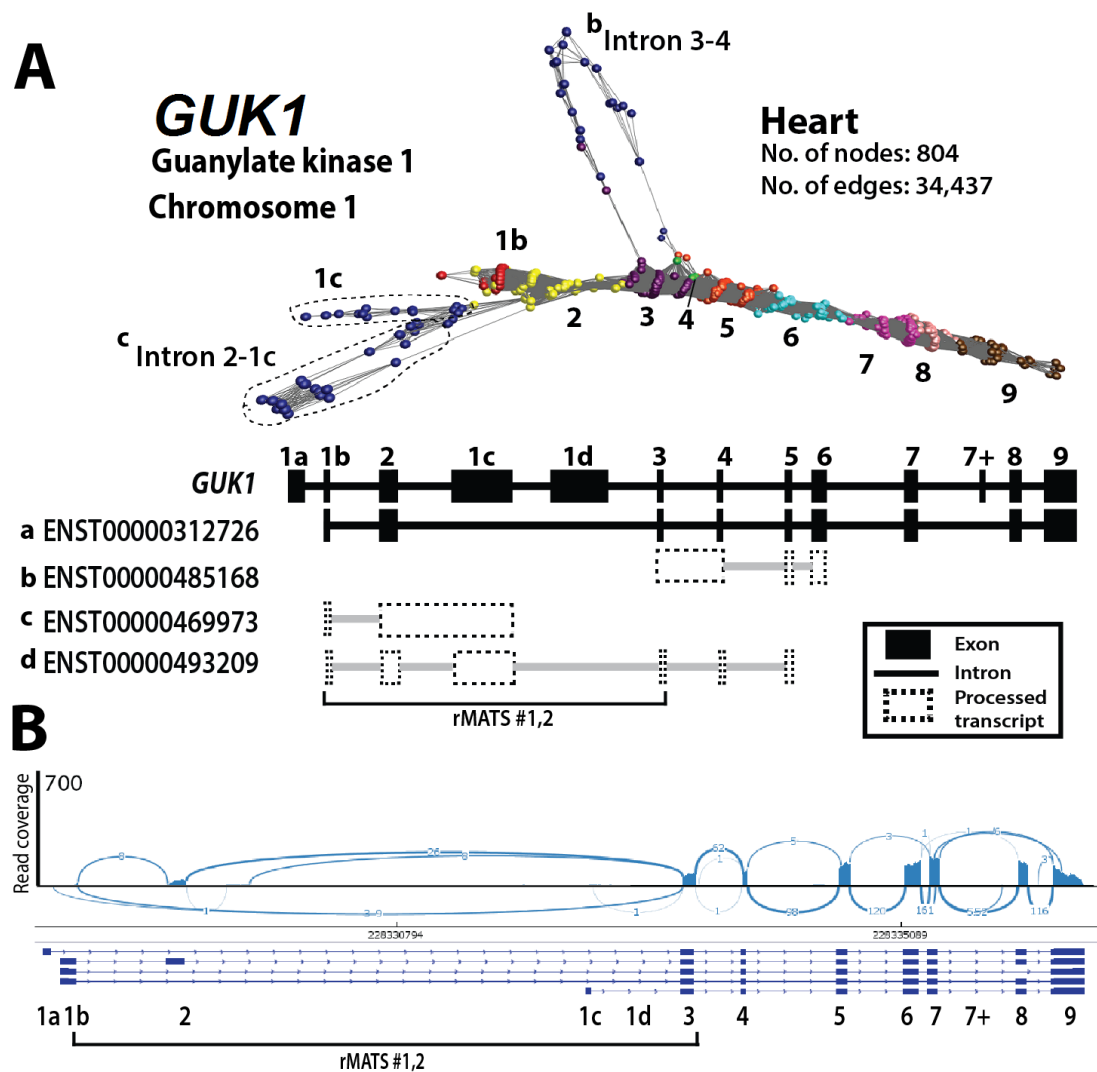
**Figure 4.10D** shows a Vials visualisation of *GUK1* of brain, heart, and liver. The data is from the human transcriptome Illumina BodyMap 2.0. In this view, the expression level shows as transcript per million (TPM) and the highest isoforms expressed in the brain (blue) is ENST00000453943, heart (orange) is ENST00000312726 and liver (green) is ENST00000391865. All these highly expressed tissues indicate agreement with network structure.



**Figure 4.10: Visualisation of *GUK1* transcript in the human brain. (A) RNA-seq gene expression profile across 27 different tissues.** *GUK1* is ubiquitously expressed in all tissues **(B) Network-based visualisation of *GUK1* transcript in the brain.** In this network, it indicates one isoform expressed in each tissue (ENST00000453943 – brain) while others are minor isoforms which can be visualised from the fewer nodes branch out of the major network. The ‘loop’ and bifurcation of the networks are mostly derived from the processed transcript. In agreement with two rMATS analysis; in both tissue comparisons brain vs liver and heart vs liver, exon 2 is skipped in the liver. It is interesting to note that in the *GUK1* networks, there is a rapid decrease in the density of nodes within exon 9 in the brain transcript network. This corresponds to where the IGV view in **(C)** also indicates a decrease in the density of reads and it corresponds to *GUK1* transcript (ENST00000453943) that exhibits a truncated exon 9 at this position. **(B) Sashimi plot.** Representative Sashimi coverage plot generated in IGV indicating RNA-seq reads mapping to *GUK1* locus from human brain (red), heart (blue) and liver (green). **(D) Vials – visualising AS of genes.** The tissue of heart (orange), liver (green) and brain (blue) were selected. Tissues indicate different highest isoform expression; brain (ENST00000453943, TPM=0.36), heart (ENST00000312726, TPM=0.38) and liver (ENST00000391865, TPM=0.29). TPM, transcript per million.

The analysis of the *GUK1* transcript network of from heart, there were four isoforms expressed (one major isoform and three minor isoforms) **(Figure 4.11A)**. One major isoform expressed (ENST00000312726) contains exon 1b, 2, 3, 4, 5, 6, 7, 8 and 9. While three minor isoforms expressed in this tissue are processed transcripts. The ‘loop’ (<sup>b</sup>Intron 3-4 - ENST00000485168) contains an intronic sequence of 3-4, exon 5 and final exon 6. The branch out nodes (<sup>c</sup>Intron 2-1c - ENST00000469973) appears at the beginning of the network. This isoform contains exon 1b and intronic sequence of 2 to 1c. The last minor isoform is ENST00000493209 contains exon 1b, 2, 1c, 3, 4 and 5. All these minor isoforms are processed transcript.

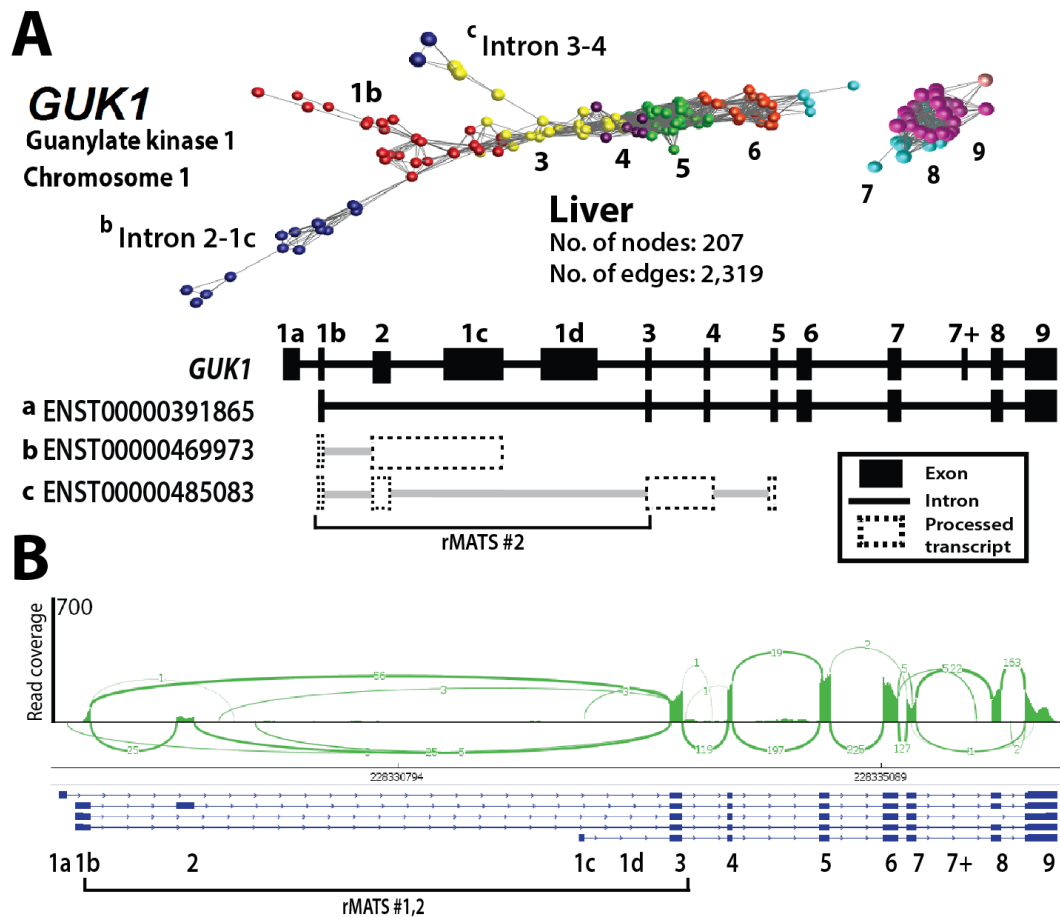
The Sashimi plots show a quantitative visualisation of the RNA-seq read alignment of the brain, heart, and liver together with expression shows in FPKM value **(Figure 4.11B)**. The junctions are shown by the arcs that align from one exon to another exon. The visualisation of read alignment with the gene model on the bottom of the plot can be visualised.



**Figure 4.11: Visualisation of *GUK1* transcript in the heart. (A) Network-based visualisation of gene *GUK1* in the heart.** In this network, it shows one major isoform expressed (ENST00000312726 – a) while other isoforms are minor isoforms which can be visualised from the fewer nodes branch out of the major network. The ‘loop’ (ENST00000485168 – b) is derived from intronic sequence, and bifurcation of the networks (ENST00000469973 – c and ENST00000493209 – d) is derived from the processed transcript. **(B) Sashimi plot.** IGV reports four isoforms in this view.

Perhaps, the simplest network analysis of *GUK1* was derived from the liver due to a low expression level of this gene (**Figure 4.12A**). From the observation of this network, three isoforms were believed to be expressed in the liver tissue. There was only one major isoform, and the other two were a processed transcript. The branch out nodes caused by the alternatively start sites of exon 1b of major isoform

(ENST00000391865) and intronic sequence at exon 2 (ENST00000469973) and exon 3 (ENST00000485082). The intronic sequence (<sup>b</sup>Intron 2-1c and <sup>c</sup>Intron 3-4) can be visualised from the nodes branch out from the network. These isoforms are processed transcript.



**Figure 4.12: Visualisation of *GUK1* transcript in the liver. (A) Network-based visualisation of gene *GUK1* in the liver.** In this network, it shows one major isoform expressed (ENST00000391865 - a) while others are minor isoforms which can be visualised from the fewer nodes branch out of the major network. The bifurcation of the network is derived from the processed transcript (ENST00000469973 – b and ENST00000485083 – c). **(B) Sashimi plot.** IGV reports five isoforms in this view.

**Table 4.5** shows a tissue comparison of *GUK1* between brain, heart, and liver in three different visualisation approaches. Network-based analysis of the *GUK1* gene transcripts co-expressed in the brain revealed that there was one major isoform



expressed in this tissue disagree with Sashimi plots. However, another one major isoform of network agrees with three isoforms in Sashimi plot. Furthermore, there are two isoforms was not observed from the network not shown in Sashimi plot. Interestingly, all major isoforms in each tissue indicate agreement between network and Vials.

**Table 4.5: Summary of visualisation analysis of *GUK1* between the brain, heart, and liver tissue.**

| Visualisation approach/Tissue | Brain   | Heart  | Liver  |
|-------------------------------|---|--|--|
| Network                       | <p><b>Major</b><br/>(2 major isoforms)<br/>a – ENST00000453943<br/>b – ENST00000312726</p> <p><b>Minor</b><br/>(4 minor isoforms)<br/>c – ENST00000366730<br/>d – ENST00000412265<br/>e – ENST00000469973<br/>f – ENST00000485083</p> | <p><b>Major</b><br/>(1 major isoform)<br/>a – ENST00000312726</p> <p><b>Minor</b><br/>(3 minor isoforms)<br/>c – ENST00000485168<br/>d – ENST00000469973<br/>e – ENST00000493209</p> | <p><b>Major</b><br/>(1 major isoform)<br/>a – ENST00000391865</p> <p><b>Minor</b><br/>(2 minor isoforms)<br/>b – ENST00000469973<br/>c – ENST00000485083</p> |
| Sashimi plots                 | - 3 isoforms only<br>(b, c, d)  | - 1 isoforms only (a)  | - 1 isoform only (a)   |

#### 4.3.4.3 Analysis of *SORBS2*

Large gene is more likely to give a difficulty for most visualisation tools to visualise RNA-seq data. The setbacks are including determining isoform expressed to distinguish AS event in such gene. *SORBS2* possesses a large of gene size of 371,273 bp located on chromosome 4. It gives the challenge to visualise a large gene. This includes the IGV and Vials, however, these two tools have their capability and network analysis gives an alternative way in visualising AS. In this case, *SORBS2* was selected because it was reported as one of the most significance alternatively spliced exon detected using rMATS tools in tissue comparison between heart and liver. From the rMATS analysis for *SORBS2* indicates significantly alternative spliced of skipped exon 27 with the FDR value of 5.57E-308. In the tissue

comparison of the heart vs liver, where the exon inclusion level of exon 27 is 0.9 in the heart and 0.35 in the liver (**Table 4.3**).

The *SORBS2* expression is strongly expressed in the heart tissue mainly in the nucleus. This gene is expressed in all tissues and enriched in the heart, thyroid, adrenal and urinary bladder. It has 39 exons and a large number of potential isoforms – 64 in total, 40 protein-coding transcript isoforms, with a further 24 transcript non-coding isoforms which include retained intron and process transcript. To further illustrate the *SORBS2* gene across tissues, brain and thyroid network were also included. The number of reads in each network of the heart and liver is 8,766 reads and 604 reads respectively. While for the brain and thyroid network consists of 1,747 reads and 10,919 reads, respectively.

**Figure 4.13A** is an expression profiles of 27 human tissues demonstrate that the *SORBS2* was highly expressed in the heart, ubiquitously low expressed in other tissues except in thyroid, adrenal, urinary bladder, and oesophagus. Moreover, *SORBS2* has been shown to play an important role as an adapter protein to assemble signalling complexes in stress fibres in the heart.

An RNA-seq assembly network of the heart was examined to determine isoform expressed in this tissue (**Figure 4.13B**). The structure of this network is complex and important to examine it carefully. The analysis of the *SORBS2* of heart, there were four major and three minor isoforms. The first major isoform (ENST00000393528 – **a**) is explained by the structure from exon 4 to 39. This isoform can be observed as a linear form where the nodes colour along the structure. It contains exon 4, 8, 13, 15, 16, 19, 20, 22-25, 27-30, 32 to 39 (truncated exon 39).

The second major isoform can be observed from the alternative 5' start site of exon 21 (ENST00000418609 - **b**) to the final exon 39+. The broken structure of exon 39 is an exon 39 (designated as 39+). From the network structure, exon 30 is directly connected to exon 32, and there is no indication of split or loop structures were observed. Therefore, exon 31 (white box) is completely skipped, and it is believed to

be spliced out during mRNA processes. When referring the gene model to Ensembl, there is no other isoform exactly match to this model. This would suggest the novel isoform was discovered but further investigation needed to confirm the novelty of this isoform.

The third major isoform (ENST00000448662 – **c**) can be explained from the c-terminus exon 39+ with skipping exon 31. However, the only differences between this gene model from Ensembl are lack of start exon of 6 in the network.

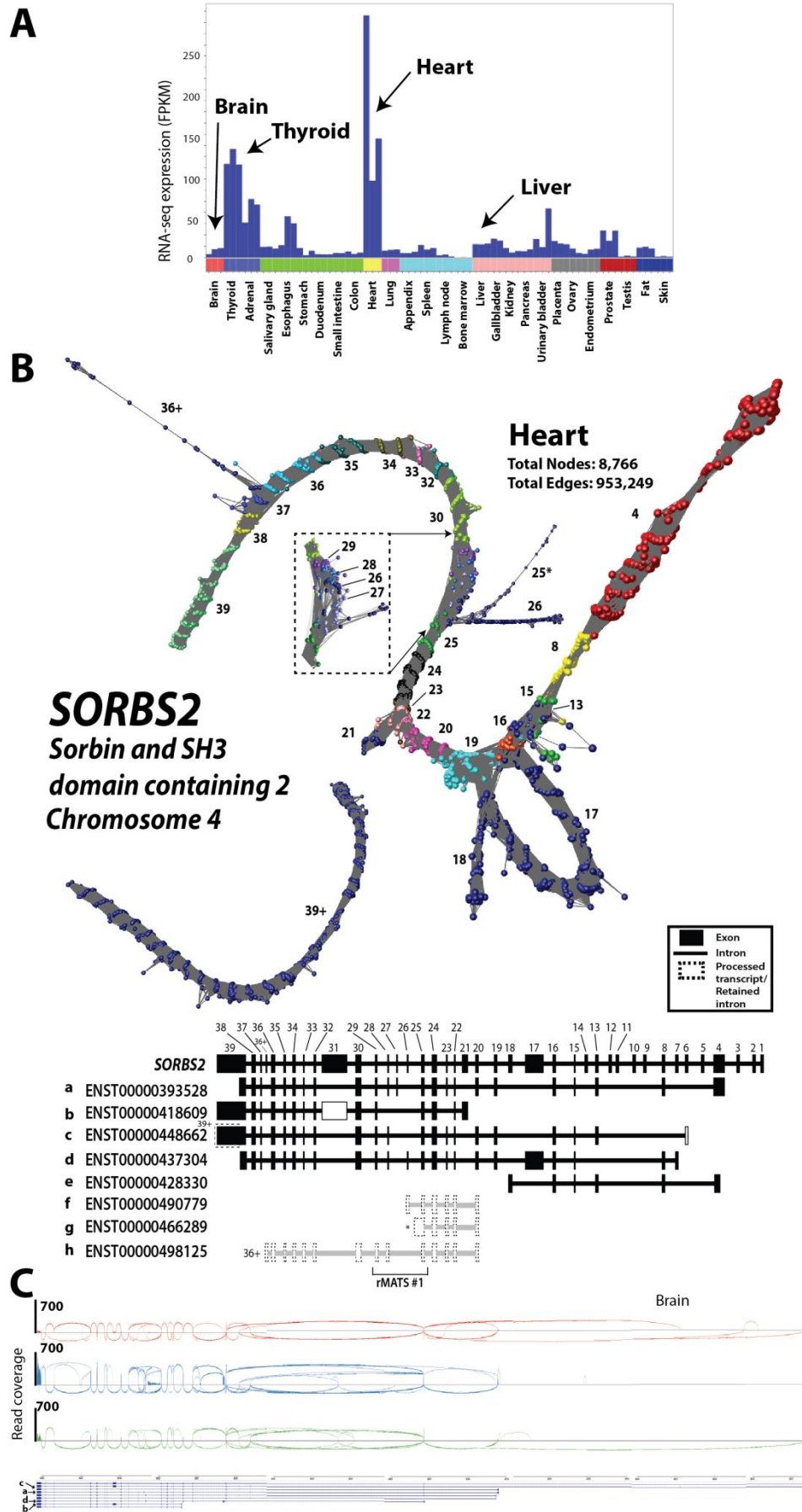
The fourth major isoform can be explained from the big loop of exon 17 which is an isoform of (ENST00000437304 – **d**). It matches to this isoform in all exons where it is the only isoform had the full-length of exon 17 and reported as the second highest expression in the heart. Except for the first exon of this isoform where it starts at exon 8. Based on the Ensembl, start exon of this isoform is exon 7. However, from the observation of the network structure shows there is no evidence the existence of exon 7. This could be the possibility of novel isoform as found in the previous isoform and this needs further investigation to confirm the novelty of this isoform.

The fifth and last major isoform is believed to be expressed in the heart is ENST0000428330 (**e**). This isoform can be observed from the branch nodes emerge from the network which is the alternative 3' end of exon 18. This isoform contains exon 4, 8, 13-16 and last exon 18. All these five major isoforms demonstrate by the heavily 'corkscrew' structure.

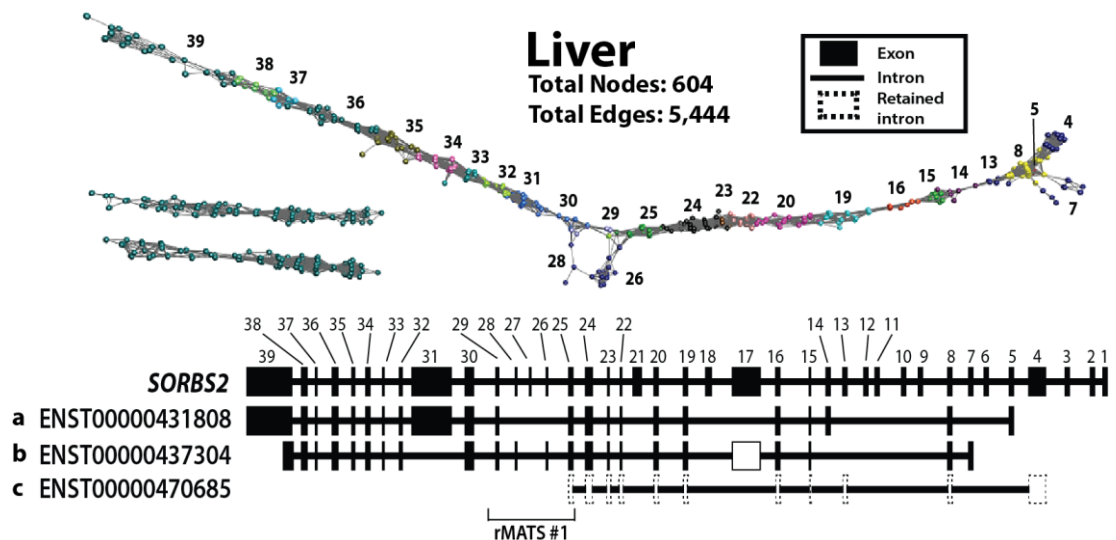
There are a small number of nodes emerge from the major network in this heart RNA-seq mapping transcript network. These structures are considered as minor isoforms. According to Ensembl, these minor isoforms are identified as retained intron and processed transcript. It is obvious from this network that the bifurcation was identified as minor isoforms of *SORBS2* in the heart. These isoforms are ENST00000490779 (**f**) contains exon 20, 22-26, ENST00000466289 (**g**) contains exon 220, 21-24 and extended of last exon 25, and the last minor isoform is ENST00000498125 (**h**) contains exon 20, 21-25, 28-30, and 32 to 36+. All these

minor isoforms are supported by the heart RNA-seq mapping transcript network structure.

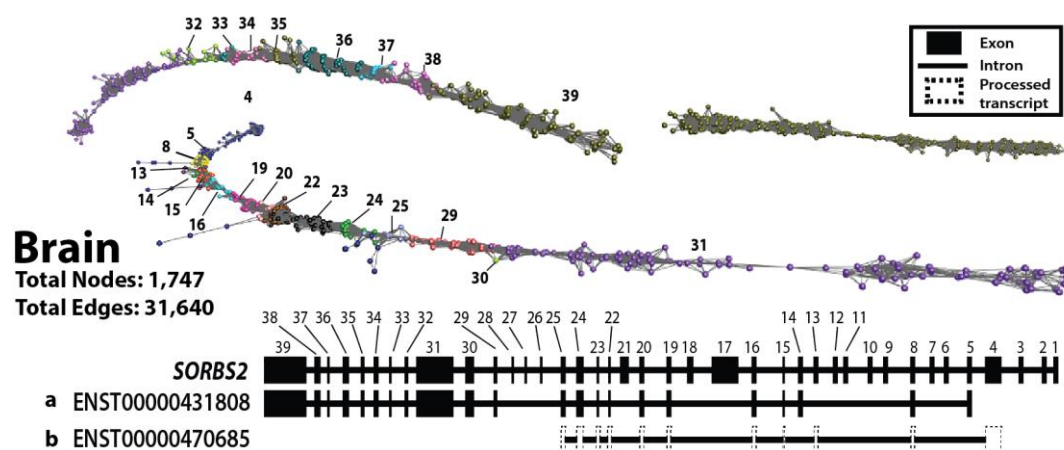
A Sashimi plot was used to compare with network-based analysis. However, the plot of *SORBS2* generated using IGV is lengthy and cannot be fit in the figure. The reason is that the size of the gene is large hence the IGV cannot scale it down. **Figure 4.13C** display the Sashimi plot with exon expression and junction support of the brain and heart. There are nine isoforms showed in this plot which are ENST00000284776, ENST00000355634, ENST00000449407, ENST00000393528 (**a**), ENST00000319471, ENST00000448662 (**c**), ENST00000437304 (**b**) and ENST00000418609. Only isoform **a**, **b** and **c** are observed in the heart RNA-seq mapping transcript network.



**Figure 4.13: Visualisation of the *SORBS2* transcript in the human heart.** (A) RNA-seq gene expression profile across 27 different tissues. *SORBS2* is highly expressed in the heart and ubiquitously expressed in other tissues except in thyroid, adrenal and urinary bladder. (B) Network-based visualisation of *SORBS2*. In this network, it indicates five major isoforms expressed in the heart tissue. Three minor isoforms can be visualised from the fewer nodes emerge from the major network. The bifurcation of the networks is derived from the processed transcript. (C) Sashimi plot. There are nine isoforms are reported by IGV in this view.



**Figure 4.14: Network-based visualisation of the *SORBS2* transcript in the liver.** In this network, it shows three isoforms expressed (ENST00000431808 – a, ENST00000437304 – b, and ENST00000470685 - c).



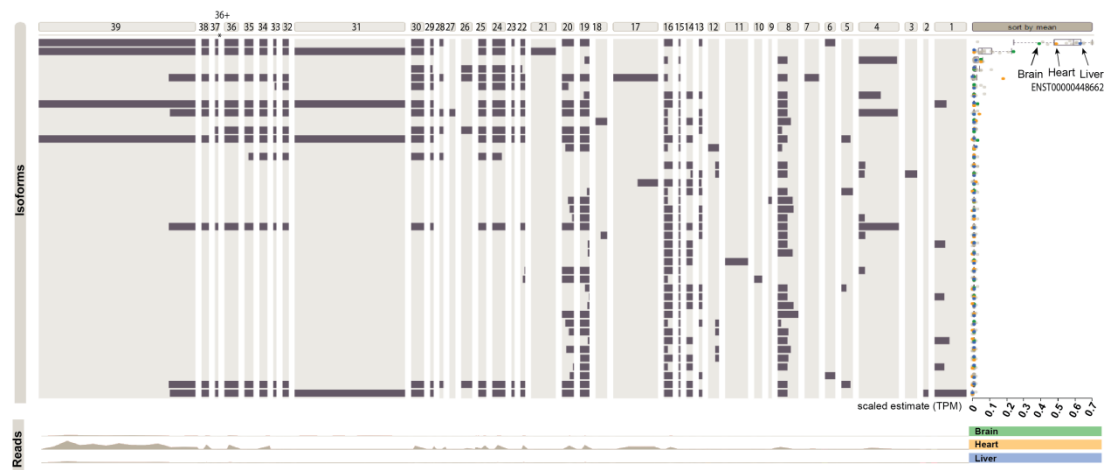
**Figure 4.15: Network-based visualisation of the *SORBS2* transcript in the brain.** In this network, it shows two isoforms expressed (ENST00000431808 – a, and ENST00000470685 – b).

For liver and brain tissue networks (**Figure 4.14** and **4.15**), there is no major or minor isoform can be determined due to the low expression of *SORBS2* in these tissues. Liver DNA mapping transcript network (**Figure 4.14**) shows a linear structure with an alternative start site and AS of exon 27. There are three possible isoforms are believed to be expressed in this tissue. The first isoform is ENST00000431808 (**a**) start with exon 5, 8, 14-16, 19, 20, 22-25, 29-38 and final exon of 39. This isoform also can be observed from spliced out of exon 27. The second isoform can be explained from the branch out nodes of exon 7. This isoform is ENST00000437304 (**b**) contains start exon of 7, 8, 15, 19, 20, 22-26, 28-30, 32-37 and last exon of truncated exon 39. The only isoform contains start exon of 7 is ENST00000437304 which also contain exon 17, however, in this network, there is no evidence of exon 17, and it skipped in the network. Therefore, the exon 17 in the gene model is designated as a white box. The last isoform is believed to be expressed in this tissue is ENST00000470685 (**c**) (retained intron) which can be explained from the start exon 4.

From the network analysis of heart and liver, it is indicated that the agreement between network analysis and rMATS analysis where exon 27 is skipped in the liver tissue. However, the analysis performed here is not only validating the rMATS result but also found evidence of new isoforms in both tissues. With the comparison between network analysis and Ensembl, one isoform from each of heart (ENST00000418609) and liver (ENST0000043304) networks indicate the novelty however further investigation is needed.

rMATS was used to analyse the differential splicing between three different tissue; brain, heart, and liver. However, differential splicing of *SORBS2* was only between heart and liver. Therefore, it is worth to visualise the brain RNA-seq mapping transcript network as well. The brain tissue network of *SORBS2* (**Figure 4.15**) shows a linear structure, broken at exon 39, with a total number of nodes is 1,747 and number of edges are 31,640. This reason is that the low expression of *SORBS2* in the brain. Two isoforms are believed to be expressed in this tissue; ENST00000431808 (**a**) and ENST00000470685 (**b**). Both isoforms can be explained from the two splits of alternative 5' end.

**Figure 4.16** shows Vials visualisation of *SORBS2* from Illumina BodyMap 2.0 data. Four tissues, heart, liver, and brain, were selected in this view. All these three different tissues share common highest isoform expression which is (ENST00000448662 – c). The expression of this isoform in the heart, liver, and brain estimated in transcript per million (TPM) is 0.49, 0.63, 0.39 and 0.58 respectively. It is interesting the agreement between a network analysis and Vials where this isoform (ENST00000448662) can be observed in the heart and thyroid network transcript analysis.



**Figure 4.16: Vials – visualizing AS of genes.** Three tissues of heart, liver, and brain were selected, and multiple dot plots are shown to allow comparison between these tissues. All four tissues indicate the same highest isoform expression which is ENST00000448662. TPM, Transcript per million.

**Table 4.6** shows a tissue comparison of *SORBS2* between heart, liver, and brain in three different visualisation approaches. Network-based analysis of the *SORBS2* gene transcripts co-expressed in the heart revealed that there were five major isoforms expressed in this tissue while only three and two major isoforms expressed in the liver and brain. One isoform indicates agreement between heart and liver (ENST00000437304), two isoforms indicate agreement between liver and brain (ENST00000431808 and ENST00000470685). However, there is no isoform indicate agreement between heart and brain. Furthermore, there are seven isoforms only indicate expression in the heart (four major, three minor). Nonetheless, there were



two possible of novel isoform which ENST00000418609 (exclude exon 31) and ENST00000437304 (include exon 17) can be visualised in the heart using network approach. Nonetheless, there was difficulty in visualising splicing event due to the vast length of the Sashimi plots of this gene.

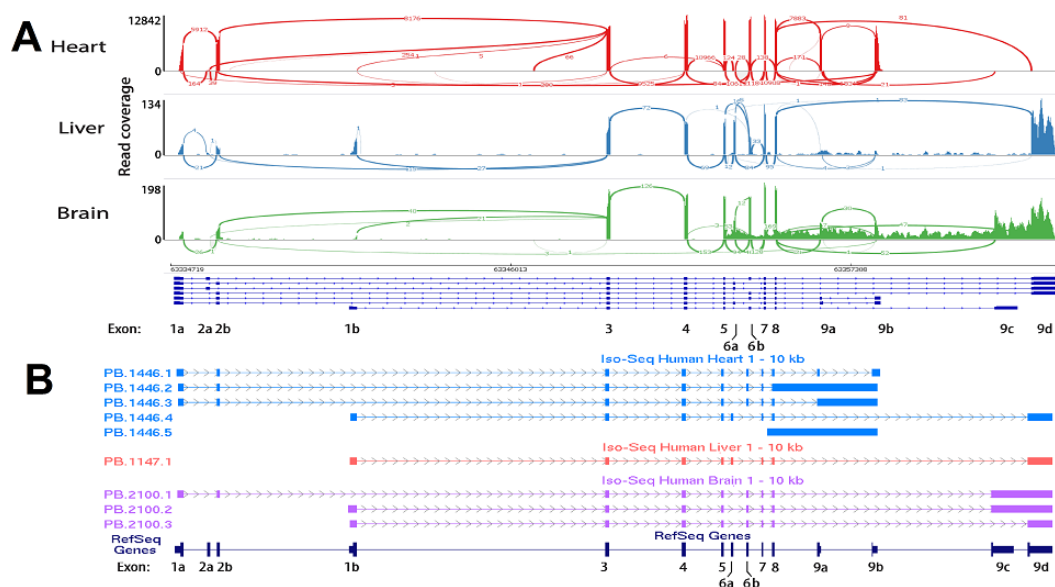
**Table 4.6: Summary of visualisation analysis of *SORBS2* between the heart, liver, and brain tissue. Asterisk mark (\*) is a novel isoform detected using network approach.**

| Visualisation approach/Tissue | Heart  | Liver  | Brain                                      |
|-------------------------------|--|--|--|
| Network                       | <p><b>Major</b><br/>           a - ENST00000393528<br/>           b - ENST00000418609*<br/>           c - ENST00000448662<br/>           d - ENST00000437304*<br/>           e - ENST00000428330</p> <p><b>Minor</b><br/>           f - ENST00000490779<br/>           g - ENST00000466289</p> | a - ENST00000431808<br>b - ENST00000437304*<br>c - ENST00000470685 | a - ENST00000431808<br>b - ENST00000470685 |
| Sashimi plots                 | - a, b, c and d only   | - a, b, c and d only   | - a, b, c and d only                       |

#### 4.3.4.4 Analysis of *TPM1*

In order to explore the issues associated with the network-based analysis of transcript variation between tissues here, gene *TPM1* was focused and examined. *TPM1* was selected because it is widely expressed across tissues, but its expression varies considerably in different tissues, being particularly strongly expressed in muscle (**Table 4.3**). It has 15 exons and a large number of potential isoforms – 33 in total, 19 protein-coding transcript isoforms, with a further 14 transcript non-coding isoforms, e.g. with a retained intron or the product of nonsense-mediated decay, are recorded in Ensembl. Three structures of the *TPM1* transcript in three human tissues (heart, liver, brain) networks whereas analysed using the public human RNA-seq data using a network-based approach.

The Sashimi plots show the quantitative visualisation of the RNA-seq read alignment of heart, liver, and brain (**Figure 4.17A**). The visualisation of read alignment with the gene model on the bottom of the plot can be visualised. **Figure 4.17B** shows the UCSC Genome Browser visualisation *TPMI* of human transcriptome from the brain, heart, and liver platform from PacBio sequencing (Pacific, 2014). The visualisation is a polished, non-redundant, full-length transcript sequences are indicated for each tissue. In this view, the isoforms detected in the heart (blue), liver (orange) and brain (purple) are five, one and three respectively. It indicates partial agreement between PacBio and network visualisation. The disagreement is included a few transcripts indicate a combination exon and intron as one exon in the gene model, e.g. in the heart, PB.1446.2 (from exon 8 to 9b), PB.1446.3 (from Exon 9a to 9b) and PB.1446.5 (from exon 8 to 9b). Furthermore, in the brain which is PB.2100.1 and PB.2100.2 (from exon 9c to 9d).



**Figure 4.17: Visualisation of RNA-seq data of *TPMI*.** (A) Representative Sashimi coverage plot generated in IGV indicating RNA-seq reads mapping to *TPMI* locus from different tissue. The height of the bars represents overall read coverage. Splice junction is displayed as loops. The number of reads observed for each junction is indicated by segments, and y-axis ranges for the number of reads per exon base are shown (read coverage, left). The plot suggests different isoforms expressed in the sample is indicated by the arc connecting a pair of exons. (B) UCSC Genome Browser browser sequence visualisation of *TPMI* from a different study of the whole human transcriptome of brain, heart, and liver using the PacBio platform. The number of isoforms detected by PacBio relatively different from the network-based

based visualisation. PacBio detected five isoforms in the heart (blue), one in the liver (orange) and three in the brain (purple).

The number of reads in each network of heart, liver, and brain is 10,347 reads, 431 reads, and 1,006 reads respectively. All these three networks consisted alternative promoter start exon 1a or 1b, mutually exclusive exon 2a or 2b, exon 3-9 with AS occurred at exon 6 with the mutually exclusive exon 6a or 6b. At the c-terminus, the transcript is spliced again at exon 9, with the choice of exon 9a, 9b, 9c, or 9d.

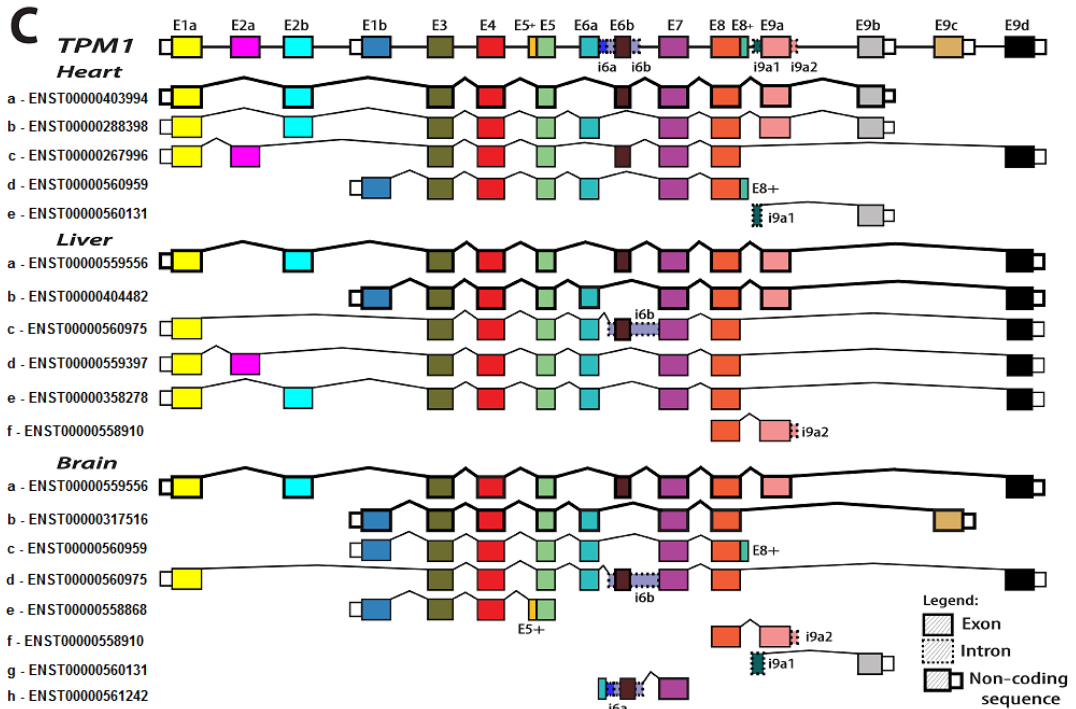
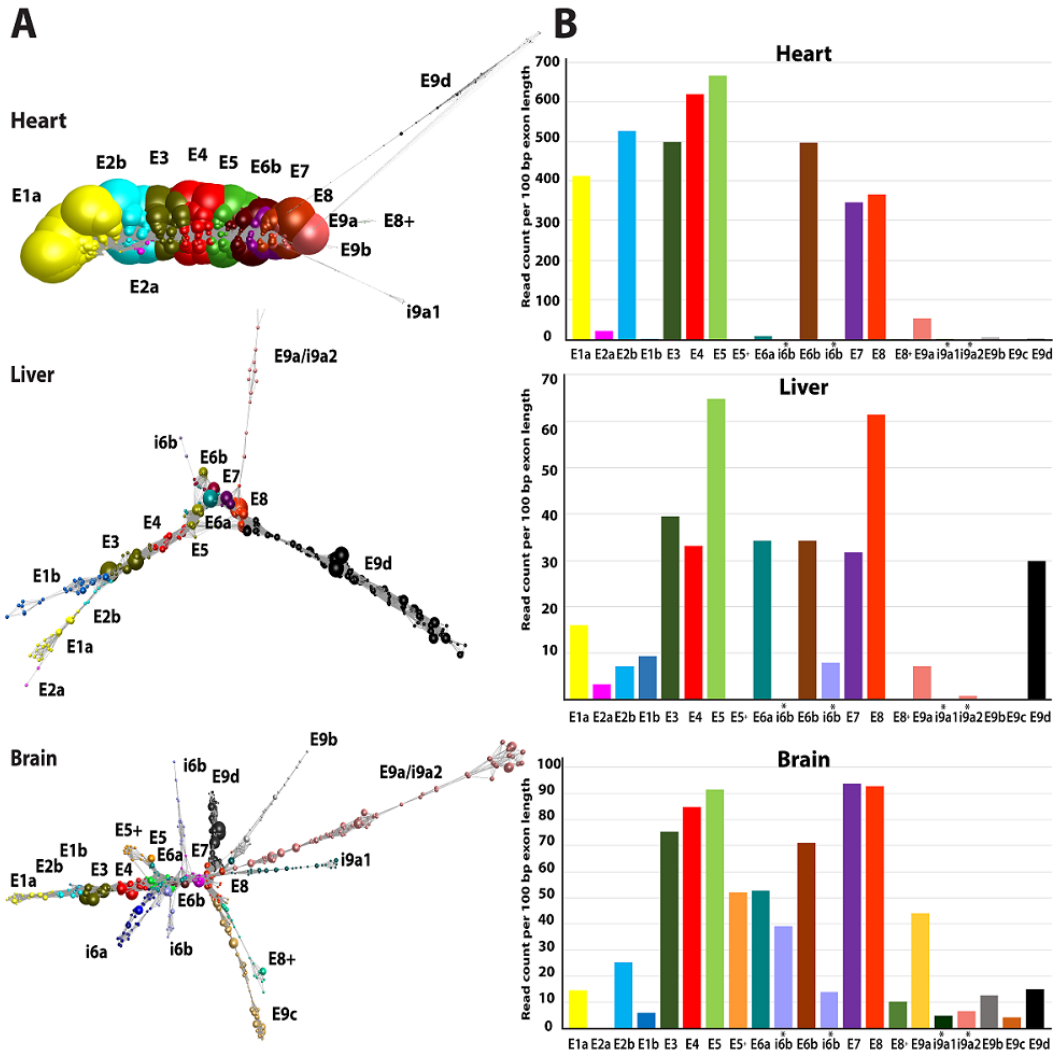
Network-based analysis of the *TPMI* gene transcripts co-expressed in the heart (**Figure 4.18A**), revealed that there was only one major isoform expressed in this tissue (ENST00000403994 - **a**) contains exon 1a, 2b, 3, 4, 5, 6b, 7, 8 and 9a/b. However, there appeared to be evidence of another four minor isoforms (ENST00000288398 - **b**, ENST00000267996 - **c**, ENST00000560959 - **d**, and ENST00000560131 - **d**). The minor isoforms can be visualised from the small branches nodes emerge from of the major network. *TPMI* gene in the heart muscle tissue contains exon 1a, 2b, 3, 4, 5, 6b, 7, 8 and 9a/b, and it supports heart RNA-seq mapping transcript network structure. In the network, the isoform which uses these exons is deep in coverage, which can be visualised from the size of nodes.

The analysis of *TPMI* transcript network of the liver, there were six isoforms expressed. There are two major isoforms where the alternative 5' splicing can be observed from the split branch at the beginning of the network (ENST00000559556 - **a**, and ENST00000404484 - **b**). The alternative splice of exon 6a and 6b (ENST00000559397 - **d**, ENST00000358278 - **e**) also can be visualised from the network even though it was not an obvious splice event. The small branch out two minor isoforms can be visualised from the fewer nodes emerge from intron 6 and intron 9 (ENST00000560975 - **c**, and ENST00000558910 - **f**).

The most complex network of this *TPMI* was derived from the brain. In this network, eight isoforms are believed to be expressed in this tissue. There were three major isoforms, and the rest were retained introns. The intronic sequence can be

visualised from the nodes emerge from the network, for instance, intron 5, 6 and 8 (ENST00000560975 – **d**, ENST00000558868 – **e**, and ENST00000561242 – **h**). In this network, the major isoform splice at 3' end, which can be visualised in exon 9c and 9d (ENST00000559556 – **a**, and ENST00000317516 – **b**). Another isoform (ENST00000560959 – **c**) is indicated by the branching out of nodes from the network which contain a longer size of exon 8. Two isoforms which contains intronic sequence 9a1 (i9a1) (ENST00000560131 – **g**) and exon 9a2 (i9a2) (ENST00000558910 – **f**) were another retained intron expressed in this tissue.

Histogram read count per 100 bp exon length represents a relative number of reads in the networks (**Figure 4.18B**). In the heart histogram, the major isoform uses exon 2b rather than exon 2a, exon 6b rather than exon 6a and exon 9a. Fewer numbers of read indicate the minor isoforms. In the liver histogram, the alternative promoter indicates two different primary isoforms from different start site which are exon 1a and exon 1b. However, the alternative splice exon for exon 6a and 6b show the relatively same number of reads. Furthermore, alternative c-terminus of exon 9d only shows expression in the liver and brain. In the brain histogram, read count for exon 1a is more than exon 1b and the choices for exon 2b more over 2a. In this histogram, the uses an intronic sequence of intron 6b (i6b) is higher compared to another intronic sequence of intron 9a1 (i9a1) and 9a2 (i9a2). For the isoform that consists longer for the last exon 8 is designated as exon 8+ relatively indicate a similar expression with the intronic sequence of exon 9a (i9a). From these histograms, the level of read for exon 3, 4, 5, 7 and 8 is relatively higher indicating a highly conserved exon of *TPM1*. In **Figure 4.18C**, schematic gene representation of *TPM1* gene which was believed to be expressed in the heart, liver, and brain.



**Figure 4.18: Network-based visualisation of *TPMI* transcripts from a different tissue.** (A) Network-based visualisation of a heart, liver, and brain. In these networks, node size reflects the relative number of reads depth. For the network of heart, it shows one major isoform expressed while another four isoforms are minor isoforms which can be visualised from the fewer nodes branch out from the network. In the network of the liver, all isoforms are expressed almost at the same level. Alternative transcript initiation can be visualised from the first two branches of the network. The mutually exclusive exon 6a and 6b can be seen from the network while the retention intron of 6b and 9a can be visualised from a few nodes branching from the network. In the network of the brain, multiple isoforms expressed is indicated by the tangled branches of the network from different locations. The branch out nodes are supported by the gene model of human reference genome indicates that protein-coding and retained intron which is an alternatively spliced transcript that contains an intronic sequence. (B) Histogram number of reads per exon per sample in each tissue is indicated in the network. The coloured histogram represents exon in the network visualisation. These isoforms are believed to be expressed in each tissue sample based on the network-based visualisation. (C) Schematic gene representation of *TPMI*. All isoforms expressed indicates at the top of the isoforms gene representation and expressed in each of the tissues; heart, liver, and brain.

**Table 4.7** shows a tissue comparison of *TPMI* between heart, liver, and brain in three different visualisation approaches. Network-based analysis of the *TPMI* gene transcripts co-expressed in the heart, liver, and brain revealed that there was one major isoform expressed in these tissues. The major isoforms expressed in the liver and brain shared the same isoform but not in the heart.

**Table 4.7: Summary of visualisation analysis of *TPMI* between heart, liver, and brain tissue.**

| Visualisation approach/Tissue | Heart   | Liver   | Brain  |
|-------------------------------|---|---|--|
| Network                       | <b>Major</b><br>(1 major isoforms)<br>a – ENST00000403994<br><br><b>Minor</b><br>(4 minor isoforms)<br>b – ENST00000288398<br>c – ENST00000267996<br>d – ENST00000560959<br>e – ENST00000560131 | <b>Major</b><br>(2 major isoforms)<br>a – ENST00000559556<br>b – ENST00000404484<br><br><b>Minor</b><br>(4 minor isoforms)<br>c – ENST00000560975<br>d – ENST000005599397<br>e – ENST00000358278<br>f – ENST00000558910 | <b>Major</b><br>(3 major isoforms)<br>a – ENST00000559556<br>b – ENST00000317516<br>c – ENST00000560959<br><br><b>Minor</b><br>d – ENST00000560975<br>e – ENST00000558868<br>f – ENST00000558910<br>g – ENST00000560131<br>h – ENST00000561242 |
| Sashimi plots                 | 7 isoforms  | 7 isoforms  | 7 isoforms   |
| PacBio                        | <b>5 isoforms</b><br>PBB.1446.1<br>PBB.1446.2<br>PBB.1446.3<br>PBB.1446.4<br>PBB.1446.5   | <b>1 isoform</b><br>PB.1147.1   | <b>3 isoforms</b><br>PB.2100.1<br>PB.2100.2<br>PB.2100.3   |

### 4.3.5 Comparing visualisation approaches

A critical review on this visualisation comparison between network approach and Sashimi plots, i.e. what is agreed disagree and what is novel are presented. Four genes were analysed to compare between these approaches; *KLC1*, *GUK1*, *SORBS2*, and *TPMI*. All these genes have been subjected to the analysis to find out the agreement/disagreement and to discover a novelty of this approach.

First, the comparison of *KLC1* gene between brain and heart tissue are presented. Two major isoforms agreed in both brain and heart network, and one disagreed. However, two minor isoforms from both brain and heart are different. Regarding a comparison between network and Sashimi plots in the brain, there are only two isoforms agreed while the other four isoforms disagreed. Nonetheless, in the heart, all isoforms disagreed between network and Sashimi plot except one isoform.

Whereas for the *GUK1* gene, three isoforms from Sashimi plot are agreed with the network while the other three isoforms disagreed. It shows only one major isoform, and two minor isoforms from Sashimi plots agreed with the network. In the heart, only one isoform agreed while other three disagreed. The disagreement shows by the minor isoforms in network approaches. In the liver, it shows a consistency as in the heart whereas only one isoform agreed with the network while other two disagreed. The disagreement shows by two minor isoforms in the network approaches as well.

In the comparison in *SORBS2* of network and Sashimi plot approaches, agreement, disagreement and novel isoform were found. In the heart, four isoforms in Sashimi plot agreed with the network while three isoforms disagreed. However, it appears that two isoforms are novel found in the network compared to the plot. Likewise, in the liver, only one isoform disagreed shows only in Sashimi plot while one isoform found to be a novel isoform.

Finally, for *TPM1* gene, there are two isoforms in the Sashimi plot disagreed with the network while only one isoform disagreed in the liver. For the brain, it all agreed between network and Sashimi plots. However, none of the isoforms found in the PacBio disagreed with the network.



## 4.4 Discussion

Our understanding of gene expression has been revolutionised considerably over the last decade, mainly because of technological improvements. High-throughput sequencing of cDNA (RNA-seq) has generated enormous amounts of gene expression data that are deposited in public repositories, e.g. ArrayExpress. These data are accessible to every biologist to reuse and further analyse the data. This includes when it comes to analysing human transcriptome which comprises a large number of tissues to be sequenced. Therefore, public RNA-seq data set becomes a choice to explore network-based visualisation across human tissues. Many tools for quality control exist to process these public data which typically come as a ‘raw’ FASTQ files. These data are then set for mapping to the genome or transcriptome data, and subsequently, ready for differential expression and alternatively spliced analyses. These data can be visualised using a tool such as IGV, and splice variation across multiple samples can be inspected using Sashimi plot. The data are usually summarised as read densities while junction reads are collapsed into arcs whose width is proportional to the number of reads spanning the exons connected by the arc. However, when the multiple transcripts have been expressed for such gene, the arcs that display the junction connection is difficult to explain as discussed in Chapter 3.

The first half of this chapter aims to perform a quality control of human tissue atlas RNA-seq data using a network-based approach. But this time, it would be based on the correlation in expression between genes and samples. In this method, the RNA-seq data are performed as sample-to-sample clustering expression profile as a mean for quality control. In this context, nodes represent samples and edges are denoted as the correlation between samples. In this situation, samples with low correlation within the same tissue type are sparse or separated from the main cluster. This provides a way to identify the low quality of these samples. Furthermore, lowest correlation values within a group of tissues where all tissue samples had high inter-tissue correlation values; the read count was used as the basis to remove samples.

Top 20 clusters were subjected to functional annotation with GO terms and its p-value. This analysis is based largely upon the human GO terms content. Moreover, the importance of such co-expression information is evident in the analysis of genetic data. It can be deduced the possible phenotype of a mutation in any specific gene from its pattern of expression. There are a number of GO terms that were not expected in the list, e.g. Cluster 4 with profile description of lymph node, spleen, and appendix. This cluster has GO terms of digestive, digestion system process and xenobiotic metabolic process. The reason of this case that these analyses examine the gene list for the event of GO terms that are more ubiquitous in the query gene list than expected by chance (Yon Rhee et al., 2008). Therefore, over-represented terms may preferentially and differentially regulate in such cluster. A feature of GO that is both a strength and a limitation is its hierarchical structure. Even though efforts have been made to explain this structure in GO enrichment analysis (Jantzen et al., 2011), it can still be hard to resolve which level of the hierarchy is most liable for the statistical enrichment. Usually, the most enriched terms often are broad functional categories which can be of limited use to inform new functional insight.

BioLayout *Express*<sup>3D</sup> (Freeman et al., 2007; Theocharidis et al., 2009) is a tool for the analysis of large complex expression datasets, e.g. microarray and RNA-seq. The principle of co-expression is that a gene-to-gene comparison of expression of value across human tissue is performed by calculation of a Pearson correlation matrix, which is the main statistical measure in the tool. Therefore, with given any gene comparison, the Pearson value can range from +1 (perfect correlation) to -1 (perfect anti-correlation). A graph derived from any correlation cut-off value includes genes that are related to the expression of others above the selected threshold. Consequently, a decrease in this value results in the production of a more complex graph, while an increase in this value results in a less complex graph.

A limiting factor in network analysis is the current inability to know upfront what portions of the data might be worth visualising as a DNA mapping network and to examine more than one network at a time (see Chapter 3). Many existing tools are used to detect splice variants such as DEXSeq (Anders et al., 2012), or rMATS

(Shen et al., 2014) before network construction. This way, time is not wasted creating large uninformative visualisations of linear transcripts. For this reason, a tool for detecting alternative splicing (AS) is needed. Therefore, the second half of this chapter aimed to explore, analyse, and visualise network DNA sequence for examining differential splicing using rMATS packages that have already been developed and widely used for differential splicing analysis. The rMATS package provides a framework for analysing experiments with multiple sample groups; also, it provides robust statistics using network graph to identify and compare the AS analysis with Sashimi plot and Vials. However, the main issue with the rMATS approach is able to perform only pairwise analysis and prevent genome-wide analysis of human AS. The expression of an alternatively spliced gene is essential to produce a network adequately. Thus, a minimum threshold of expression, the inclusion of exon level, and a  $p$ -value of an alternatively spliced gene were applied to rMATS result to produce a gene list.

Here, I explore the potential of network visualisation to compare an alternatively spliced exon reported from rMATS tool between tissues, and to interpret isoform expression. In principle, when the exon is being skipped, the number of junction reads of that exon is low. The resultant network of that skipped exon can be easily visualised between tissues. The network is shown here of *KLC1*, *GUK1*, *SORBS2*, and *TPM1* illustrate results of rMATS. In the analysis, not only the network of a skipped exon can be visualised, but also multiple isoform expression can be determined. In the previous chapter (see Chapter 3 section 3.3.3.3), I observed an issue with a secondary structure within an exon of human fibroblast cell (*CENPO* and *ADCY3*). While in the case of *KLC1*, another structure appeared on the network and was observed as a separate structure. In my analysis, it showed that it was an *APOPT1* whose gene location within *KLC1* locus. This issue is most probably related to my pipeline where *GenomicRanges* tool was unable to extract read mapped based on gene name and not gene location, i.e. artefact. Transcript network of *KLC1* in the brain possesses a ‘thinning’ structure which corresponds to an immediate reduction of sequence reads within exon 6. The possible reason is that when an aligner does not align any reads with intermediate indel, there is a significant coverage drop around

the indel region. It may affect downstream differential expression analysis between samples with and without the indel. The significant drop may also potentially lead to a false alternative splicing event at this exon (Sun et al., 2016). An alternative splice variant skipped exon 15 is apparent in the network visualisation of *KLC1*.

Perhaps, the simplest network graph comparison between tissues is *GUK1*. The transcript diversity of *GUK1* in the brain, heart, and liver was analysed. *GUK1* is ubiquitously expressed in all tissues and mainly function as a housekeeping gene, which is related to metabolism and pathway. The *GUK1* role in catalysing the transfer of a phosphate group is essential in these tissues. Meanwhile, network analysis of these tissues shows agreement with rMATS analysis with skipped exon 2 in the brain and liver. The different role of exon 2 is mainly functioning in the heart and study by Joehanes et al. (2013), discovered that *GUK1* is associated with coronary heart disease (CHD). Furthermore, the highest isoform expression in the Illumina BodyMap 2.0 agreed with the network analysis which suggested that network analysis has a potential to deduce an isoform expression in RNA-seq data.

In the case shown here, the *SORBS2* transcript diversity from four different human tissue, heart, liver, brain, and thyroid was examined. *SORBS2* is highly expressed in the heart where it functions as an adapter protein to assemble signalling complexes in stress fibres. The skipped exon 31 event in the DNA network transcript of *SORBS2* in the heart and 17 in the liver suggest new isoforms expressed in these tissues; and if these new isoforms are still not clear, then further investigation is needed. Also, in the heart RNA-seq mapping transcript network, no evidence of exon 6 structure suggesting there is no transcriptional start site from this exon, but this isoform is the only one that supports the skipped exon of 31 and full-length exon 39. This isoform also appears in the thyroid network and suggests the same protein function in the heart and thyroid tissue. In the liver RNA-seq mapping transcript network, there also appears a skipped exon 17; but based on the Ensembl gene model, only this isoform supports the start exon 7. However, this would not be convincing as the expression level of *SORBS2* is low in the liver tissue. The expression level of *SORBS2* in the liver and brain is approximately ten times lower than in the heart, which suggests the

different role in signalling complexes. It is interesting that the highest isoform expression of *SORBS2* in Illumina BodyMap 2.0 of these four tissues is only found in the heart and thyroid; hence, this would be based on the expression of tissue-specific isoform. This is the reason why this isoform is not being expressed in the liver and brain. Another interesting note is that using network analysis; non-coding transcript can be detected and comparison to Sashimi plot is much clearer to identify the transcripts.

In the case shown here, *TPM1* transcript diversity in RNA-seq data derived from three human tissues, heart, liver, and brain was examined. Tropomyosin 1 is most heavily expressed in the heart (and other muscles) where it functions as an actin-binding protein involved in the contractile system of muscles. A dominant and possibly sole functional transcript isoform is expressed corresponding to muscle expressed isoform of the protein. Also, a relatively small number of reads mapped to exon 2a and terminal intron sequences suggests the presence of a low number of other transcript isoforms. However, it is not clear whether these represent transcriptional noise or transcription of these isoforms by cell types present in low abundance. Expression levels of *TPM1* in the liver and brain are approximately ten times lower than in the heart. In these tissues, tropomyosin 1 is thought to play different roles in a cytoskeletal organisation. The two RNA assembly networks generated for *TPM1* exhibited complex topologies. Through studying these networks and mapping this information back to the Ensembl transcript models for this gene, up to 6 transcript isoforms to be expressed in the liver, 10 in the brain were expressed. This is largely based on the presence of the data of reads mapping back transcript-specific exons. Despite the availability of network visualisations and other visualisation tools, interpreting these data is difficult. These types of transcript assemblies are inherently complex. It is interesting to note that the publicly available PacBio analyses of these three tissues were only in partial agreement with my analysis. Many transcript isoforms suggested by my analysis were not reported in the PacBio data. However, it also appeared that in our analyses, the use of only reads that mapped to exons for network construction failed to represent transcribed intronic regions observed in the PacBio data.

# Chapter 5 – Evaluating the usability of network-based visualisation approach using NGS Graph Generator & BioLayout *Express*<sup>3D</sup>

## 5.1 Introduction

The use of visualisation tools for RNA-seq data analysis and exploration of alternative splicing (AS) plays an essential role in research using next-generation sequencing (NGS). Unfortunately, many of existing visualisation tools are still showing read stack to the reference genome. Therefore, a network-based visualisation pipeline was developed to provide an alternative way to visualise splice variation. This pipeline is implemented as a web-based application named – NGS Graph Generator. Using this application, one can produce a network of a transcript of RNA-seq data and visualise it using the BioLayout *Express*<sup>3D</sup> software. However, to ensure that the network-based visualisation approaches useful, a number of participants were asked to evaluate and give feedback.

One of the best and most popular methods to find issues with any new visualisation approach is usability testing (Nielsen, 2012; Rubin et al., 2008). Usability testing is a cost-effective approach to study how users interact with a new tool. Usability testing asks participants to perform representative tasks using the tool and observes what they do, where they have difficulties, and where they succeed (Nielsen, 2012). The usability test was carried out by the network-based visualisation pipeline (NGS Graph Generator) described in Chapter 2. The work described in this chapter aimed to gain user feedback to understand the usability and interpretability of network-based visualisations of RNA-seq data. Understanding the user experience gives an opportunity to improve the pipeline and the application. By improving the web application based on participant needs, it can help and aid in the visualising RNA-seq assemblies as a network better. This test focuses on two aspects. First is how usability testing will be used to improve the application and draws implications for

analysis. Second, it focuses on the survey responses based on network-based approach.

Therefore, the objectives of this usability test were:

- 1) To conduct a usability test involving a small group of participants to evaluate ease of visualising, ease of application use, and participant satisfaction.
- 2) To focus on getting to know the design context and identifying the task participants experienced to be most challenging.
- 3) To focus on how the workflow of the application should be designed to support the participants work best.

## **5.2 Method**

### **5.2.1 Test metrics**

Usability is measured using a number of observable and quantifiable metrics that overcome the need to rely on simple intuition. A few of the metrics that were used in this usability test were referred to the website ([www.usability.gov](http://www.usability.gov)) resource for user experience (UX) best practices and guidelines. The test metrics were as follows:

#### **5.2.1.1 Successful task completion**

Each situation needs the participant to attain specific data that would be used in a typical task. The situation is completed when the participants indicated they had found the answer or completed the task goal.

#### **5.2.1.2 Critical errors**

Critical errors are differences at completion from the goals of the task. For instance, reporting the wrong figure in the gallery due to the participant's behaviour such as they were not concentrating on the task given. Ultimately the participant will not be able to complete the task. A participant may or may not be aware that the task goal is wrong or incomplete.

### 5.2.1.3 Non-critical errors

Non-critical errors are errors that do not interfere with a participant's ability to complete the task but result in the task being completed less efficiently. For instance, an exploratory action such as opening the wrong navigation menu or using a control incorrectly is non-critical errors.

### 5.2.1.4 Likes, dislikes, and future recommendations

Participants give what they liked most or what they liked least about the application, and provide future recommendations for improving it.

## 5.2.2 Usability test

The test was conducted by allowing the participant to complete several tasks. This includes generating a layout file from the web application NGS Graph Generator, downloading the file and visualising the layout file using the BioLayout *Express*<sup>3D</sup> software. The tasks were read to them and their activities observed. This process will allow determining any potential problems or mistakes during the test.

### 5.2.2.1 Session introduction

Five participants were given an introduction verbally. The introduction was read to the participants to ensure that all participants receive consistent information. The participants were briefed on the expectation from the test. Then, they were given a chance to raise any questions or concerns regarding the application. The main ideas of the brief introduction were to ensure the participant feel relaxing and know what they are going to do in the test. Another three participants were completing the task by themselves on their computer, and returning a form with comments via email.

### 5.2.2.2 Pre-test briefing

A short prerequisite training was conducted before giving the participants the tasks to ensure that all participants had a basic knowledge of BioLayout *Express*<sup>3D</sup>. It involved explaining how a network-based visualisation in our case typically looks like and how it works. I also demonstrated a bit of the application, basic operations such as selecting "Gene", "Parameters" and "Layout" file and how this correlated to



the determining alternative splicing. Participants from the outside of the Institute had familiarised themselves with a tutorial page on the website (<http://www.seq-graph.roslin.ed.ac.uk/tutorial>).

### 5.2.2.3 Tasks

Six tasks were created for the usability participants to try completely in a half hour session. The first three tasks were relatively simple to ensure the participant more comfortable and get a sense of usability testing. The last three tasks were a bit difficult than the first three tasks. The tasks were designed by looking at the context of the application and how it can be used by the real participants, i.e. biologist. By understanding the context and having developed the application with my colleagues, a list of tasks that emulated real-world practice was created. There were six tasks were laid out for the participant to perform this usability test (**Table 5.1**).

**Table 5.1: Usability test task.** Six tasks were given for participants to perform in this test.

| No. | Task description  |
|-----|---|
| 1.  | Can you show a figure of <i>CERS5</i> gene in the gallery?  |
| 2.  | Can you view the Ensembl browser in that gene?  |
| 3.  | Can you download and open <i>CERS5</i> gene in BioLayout <i>Express</i> <sup>3D</sup> /Miru software?                                       |
| 4.  | Can you generate an <i>LRR1</i> gene of NDHF (24h) layout file from the application using default setting without removing redundant reads? |
| 5.  | Can you open the file and explain about the network of <i>LRR1</i> ?  |
| 6.  | Can you show the ENST00000627738 isoform?   |

### 5.2.2.4 Questionnaire survey

A questionnaire survey was sent out soon after the usability test to find out what they experience with the application. In this questionnaire, they were allowed to choose more than one answer.

### 5.2.2.5 Participants

The NGS Graph Generator pipeline was developed for the biologist who lacks bioinformatics background but has an interest in splice variation. The participants were first recruited by sending an email as shown in Appendix A (Supplementary Material Chapter 5). The time and location were scheduled based on the availability of the students and staff.

Eight participants participated for the usability survey. Five participants were from The Roslin Institute, while three were from the outside of Institute. Six people tested were familiar with the BioLayout *Express*<sup>3D</sup>/Miru software, while others were not familiar (**Table 5.2**). Some had even used the software and worked on expression data and RNA-seq but used different approaches and methodologies to its analysis such as R and other open-source software tools. They also worked with computers on a daily basis such to perform bioinformatics analysis. There were four PhD students, two postdoctoral researchers, one bioinformatician and one programmer took part of this usability and questionnaire survey. However, none of them was familiar with the network-based approach to visualising RNA-seq data as a network.

**Table 5.2:** Breakdown of participants who participated in the usability test and questionnaire survey with the familiarity of BioLayout *Express*<sup>3D</sup>/Miru software.

| Aspect / Location  | Roslin | Outside Roslin |
|--|--------|----------------|
| Usability test   | 5      | 3              |
| Familiarity with BioLayout<br><i>Express</i> <sup>3D</sup> /Miru | 5      | 1              |

### 5.2.2.6 Location and usability setup

Five participants from the Institute took the test in the PhD Thesis Writing's room at Alexander Robson Building, The Roslin Institute. It was a basic setup comprising of a desk with a computer fitted with a high-spec graphics card. Three participants from

the outside of Institute were performing the test by themselves and returning a task form with comments via email.

## 5.3 Results

### 5.3.1 Successful task completion

This section shows task completion results. The participants were measured by determining if they passed or failed each task. The summary of the results (**Figure 5.1**) shows how many participants that passed or failed each task. In the results also shows that only one participant was unable to find a network gene in the gallery in task 1. Three participants failed in task 2 where they failed to open the Ensembl view of the selected gene. It also shows that three participants failed to download and open the layout file using BioLayout *Express*<sup>3D</sup>/Miru. Half of the participants failed in task 4 where they failed to generate network assembly from the web application. Lastly, the results show that at least five participants failed to complete task 5 and 6. Task 5 involved opening and understanding how information the layout file and task 6 showing a specific isoform from the nodes class button. Figure 5.1 shows the number of participants that successfully finished tasks 1-6.

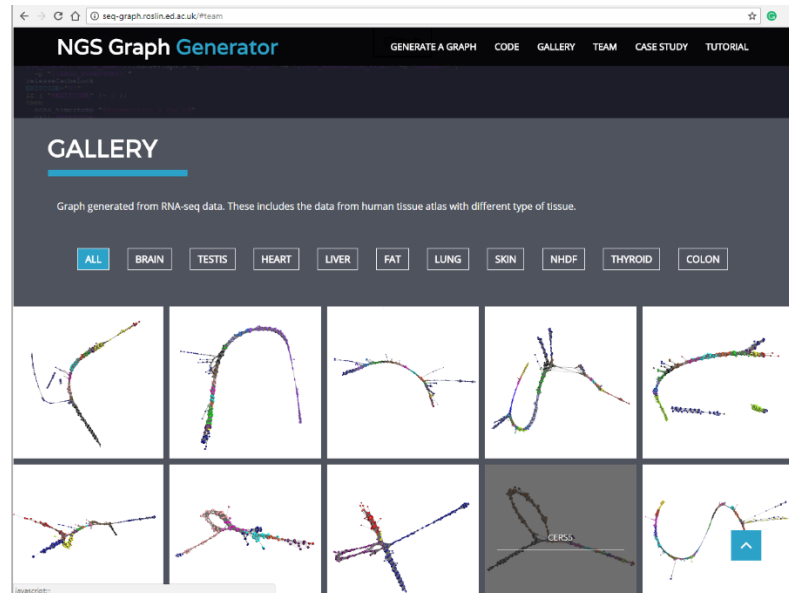


**Figure 5.1:** Amount of people who completed each task.

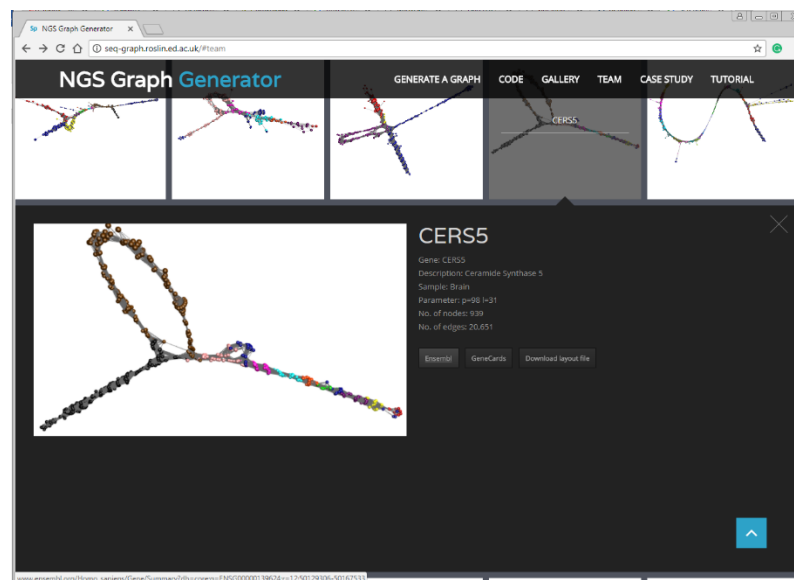
### 5.3.2 Finding a network figure of a gene

Only one participant was unable to locate a *CERS5* gene in the gallery. When asked to find the gene, the participant struggled to locate the gene in the gallery and tried to find it using a control+f function. While the other participants were able to find the

gene even though they were having difficulties to find them at first. They finally discovered that they needed to hover a mouse pointer on the network of *CERS5* gene image (Figure 5.2). On the other hand, two participants were unable to view the Ensembl record for *CERS5* gene or download the layout file (Figure 5.3).



**Figure 5.2: Finding network assemblies of a gene.** Participants need to hover on the network image box to find a gene name. A grey transparent box showed *CERS5* gene when a mouse hovered on this box.



**Figure 5.3: *CERS5* gene information.** The gene information box opened when the box is clicked. The information includes the gene name, description, sample, parameter, the number of nodes and number of edges. Three different buttons appear on this gene which is a set of a link to Ensembl, GeneCard and a button to download a pre-set layout file of this gene.

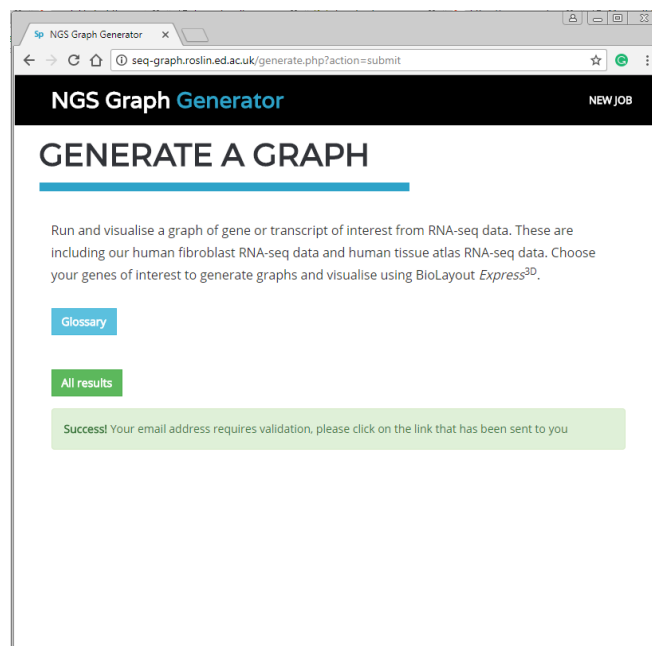
### 5.3.3 Generating a layout file

Half of the participants did not understand how a network visualisation of RNA-seq worked. When asked to set up generate a layout file of the specified gene they thought they were supposed to download the network file in the gallery without having created the file from “Generate A File” tab (**Figure 5.4**). The participants all swiftly recovered from this error when they saw that they had no file to select from the gallery. This problem can be improved by placing the “Generate A File” in front of the web application rather than on the separate page. This separation is to ensure that the application of NGS Graph Generator as the main on the website.

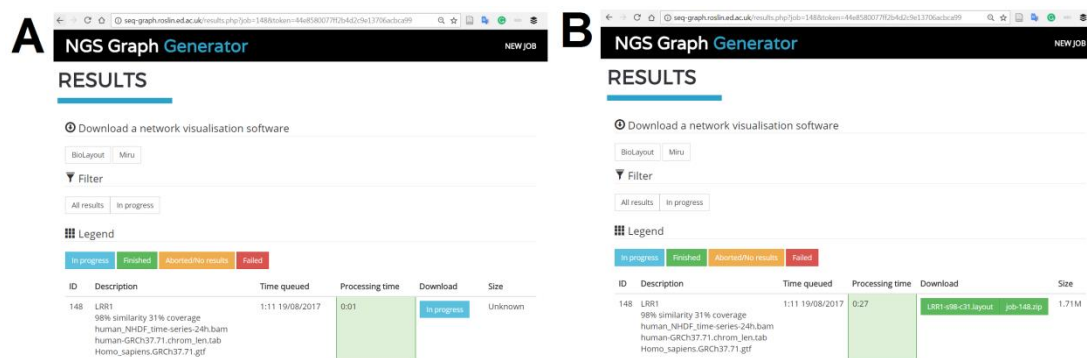
**Figure 5.4: A network-based visualisation pipeline page.** A network is created from ‘Generate A Graph’ page. In this page, a participant will select a suitable parameter to generate a layout file.

Another problem occurred to most of the participants after submitting a job to generate a layout file on this page shown in **Figure 5.5**. Participants confused with the information and no result shows on this page. However, after a while, all the participants were able to click on ‘All results’ button to go to the result page (**Figure 5.6**). On this page, the processing time froze, but most of the participants were

eagerly waiting for the result to appear. Ultimately, only two participants could figure out by clicking the refresh button on the browser, and the results appeared on the page.



**Figure 5.5: Job sent confirmation page.** On this page shows a successful job schedule from the previous page. A confirmation email was sent to a participant for validation. In this page contains a button link to glossary and results of a running job.



**Figure 5.6: Result page.** A result page appears if a button ‘All results’ is clicked from the previous page. It shows an ID, description of gene and the parameters used, time, processing time and the status of the process. However, the processing time remains freeze (A) unless the participant clicks the refresh button on the browser. The time and the layout file are ready to be downloaded shown in (B).

### 5.3.4 Opening BioLayout *Express*<sup>3D</sup>

One participant failed to open the BioLayout *Express*<sup>3D</sup> software to visualise RNA-seq as a network. The participant instead tried to open the layout file itself rather than imported into the software.

### 5.3.5 Visualising network using BioLayout *Express*<sup>3D</sup>

Four of eight participants failed to find the node class for isoforms information in BioLayout *Express*<sup>3D</sup>. The main problem of the participants was finding where to run the application form and where to select isoforms from the class. This could be considered a major problem if they were not realised what the network of RNA-seq meant. Another major concern by the participants was that the lacking information in the network which was a great disappointment such as sequence coverage, chromosome information and information of comparison network and another method.

### 5.3.6 Determining alternative splicing in the network

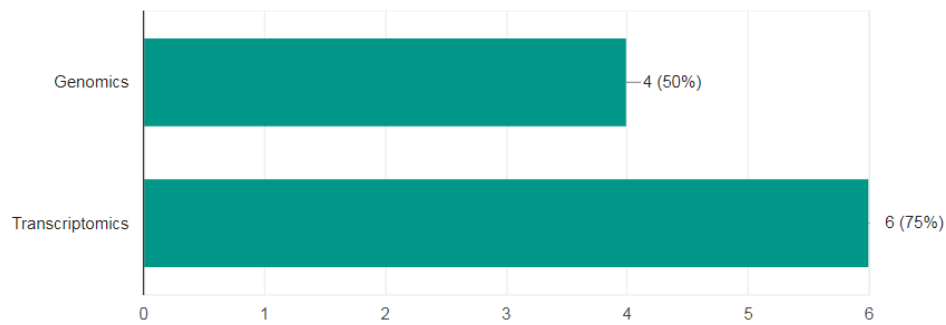
All participants could find and explain alternative splicing isoform in the network. However, this is not straightforward visualising as this need careful viewing on the software.

### 5.3.7 Questionnaire survey

A questionnaire was sent to the participants by email, requesting them to answer after usability test. This questionnaire was answered by all eight participants, and they can choose more than one answer.

#### 5.3.7.1 Question 1: What field are you working on?

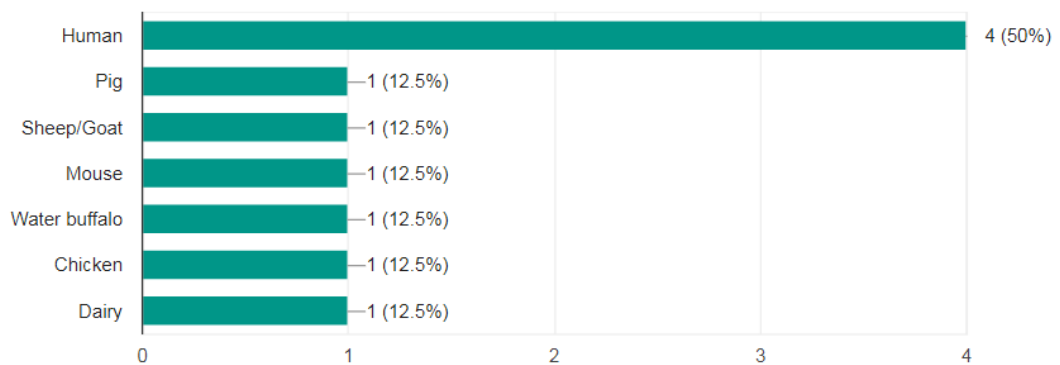
In the first question, most of the participants are working on both genomics and transcriptomics field (**Figure 5.7**).



**Figure 5.7: Question 1: What field are you working on?**

### 5.3.7.2 Question 2: What organism are you working on?

In the second question, half of the participants are working with a human while others are working with pig, sheep, mouse, chicken, dairy and buffalo (**Figure 5.8**).

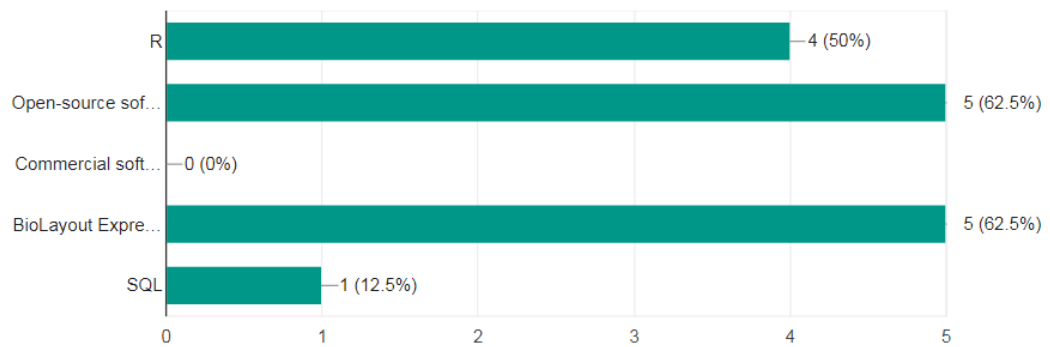


**Figure 5.8: Question 2 - What organism are you working on?**

### 5.3.7.3 Question 3: How do you analyse your RNA-seq data?

In the third question, five participants responded they used open source software and BioLayout *Express*<sup>3D</sup> to analyse their RNA-seq data. While four participants used R package and one participant analysed using SQL (**Figure 5.9**).

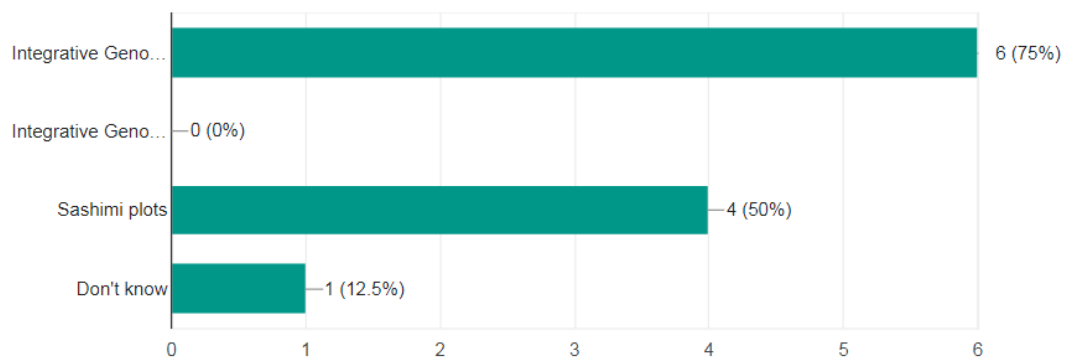




**Figure 5.9: Question 3: How do you analyse your RNA-seq data?**

#### 5.3.7.4 Question 4: How do you visualise your RNA-seq data?

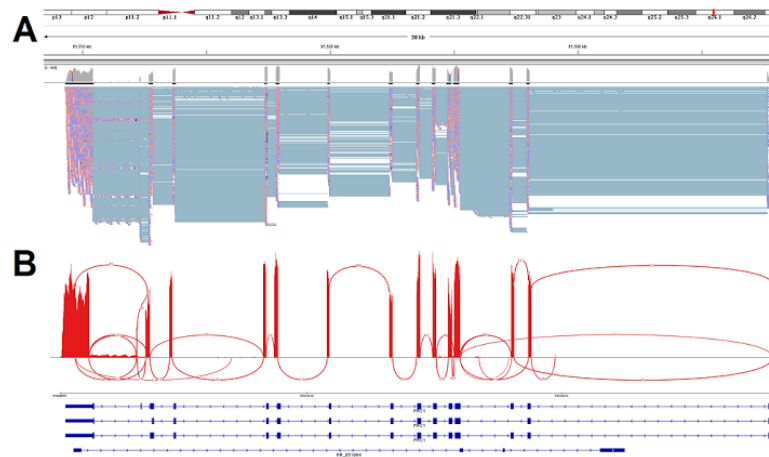
In the fourth question, the participants were asked about the software or tools they are using to visualise their RNA-seq data. The response shows most of them (90%) are using Integrative Genomic Viewer (IGV) to visualise RNA-seq data followed by Sashimi plots (70%) and Integrative Genomic Browser (IGB) (20%) (**Figure 5.10**). These tools are very common among researchers who are working on genomics and transcriptomics study. However, only one participant is using Artemis to visualise the data and one participant unable to answer the question.



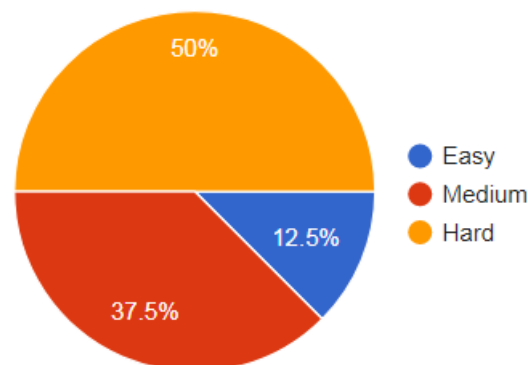
**Figure 5.10: Question 4: How do you visualise your RNA-seq data?**

### 5.3.7.5 Question 5: How do you find visualising your data using IGV/Sashimi plot?

In the fifth question, the participants have shown a figure of RNA-seq data using IGV viewer and Sashimi plot (**Figure 5.11**). Most of them (87.5%) responded that visualising the data using IGV and Sashimi plot was hard or medium while only 12.5% of them responded easy (**Figure 5.12**).



**Figure 5.11:** Visualisation of RNA-seq data using (A) IGV and (B) Sashimi plots.

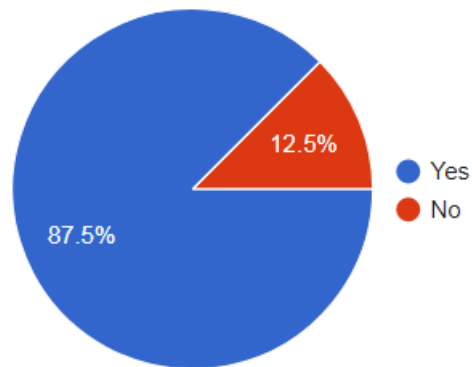


**Figure 5.12:** Question 5: How do you find visualising your data using IGV/Sashimi plots?

### 5.3.7.6 Question 6: Have you ever used BioLayout Express<sup>3D</sup>/Miru for visualisation?

In the sixth question, the participants were asked whether they had experienced using BioLayout Express<sup>3D</sup>/Miru for visualisation purposes. The responses show that seven

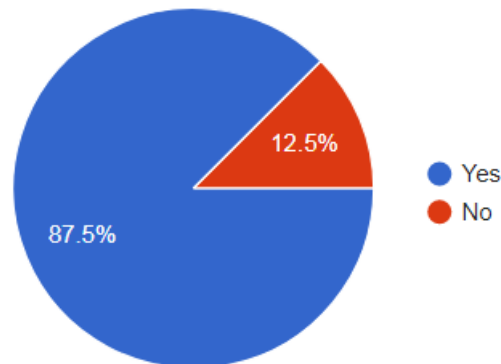
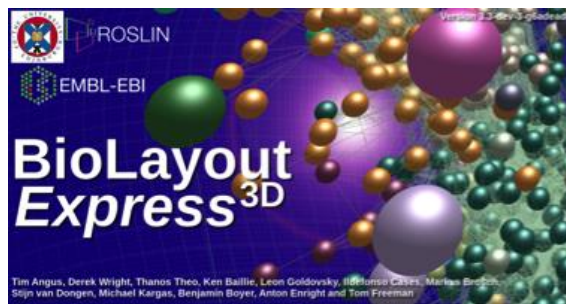
participants have used this software while only one participant has not used this software (**Figure 5.13**).



**Figure 5.13: Question 6:** Have you ever used BioLayout *Express*<sup>3D</sup>/Miru for visualisation?

**5.3.7.7 Question 7:** Do you know that we can visualise RNA-seq assemblies of a gene as a network in BioLayout *Express*<sup>3D</sup>/Miru?

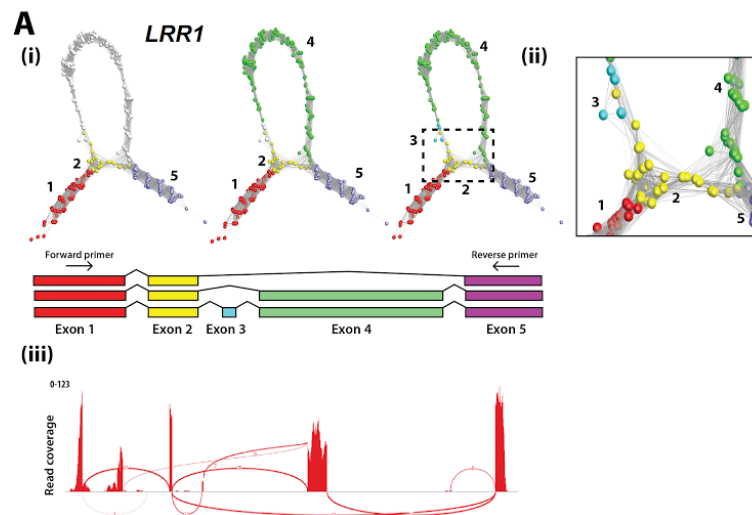
The response was same from the previous question presuming they were unaware of this software for visualising purposes. 87.5% response they did not know while 12.5% response they knew you could look at the RNA-seq assemblies (**Figure 5.14**).



**Figure 5.14: Question 7:** Do you know that we can visualise RNA-seq data of a gene as a network in BioLayout *Express*<sup>3D</sup>/Miru?

**5.3.7.8 Question 8:** Here is an example of a RNA-seq data of *LRR1* gene using network-based visualisation. How do you find this network visualisation (i and ii) of *LRR1* gene compare to Sashimi plot (iii) in terms of splice variant?

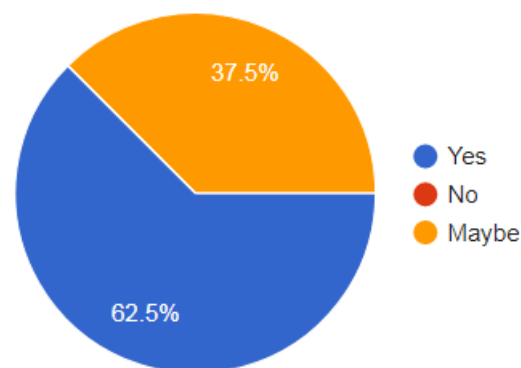
In this question, a figure showing two different approaches visualising RNA-seq data; network approaches using BioLayout *Express*<sup>3D</sup> (i and ii) and Sashimi plots (iii) (**Figure 5.15**). For this participant mostly responded that splice variant from network visualisation better than Sashimi plot in terms of distinguishing between variant. The looping of *LRR1* indicates the alternative splicing exist in the data. Some of them responded it is more intuitive and illustrative to understand the splice variant compare to Sashimi plot. One of them responded that the colour-coded of an exon was helped to determine splice variant.



**Figure 5.15: Network visualisation of *LRR1*.** (i) Network-based visualisation of *LRR1* gene (ii) Zoom-in splice variant of *LRR1*, (iii) Sashimi plot of *LRR1*.

### 5.3.7.9 Question 9: Will you use NGS Graph Generator in the future?

In this question, most of them answered ‘Yes’ with 62.5% of them and 37.5% answered ‘Maybe’. However, none of them answered ‘No’ (**Figure 5.16**).



**Figure 5.16: Question 9: Will you use NGS Graph Generator in the future?**

### 5.3.7.10 Question 10: What do you like about NGS Graph Generator?

Explain.

In the tenth question, the participants were asked about what they like about NGS Graph Generator and its approaches. Most of them like the way of its visualisation which gives them an alternative way to visualise NGS data that complements with

other methods. Easier to generate a network of gene interest and much better than Sashimi plot when visualising alternative splicing. The networks are very intuitive and aesthetically pleasing network graph generated from this approach. Some of them responded that the network approaches different from existing approach. One of them responded graph network is nice, distinguishable exon by colour-coded and easy to understand.

#### **5.3.7.11 Question 11:** What do you dislike about NGS Graph Generator?

Explain.

In this question, the participants were also asked about on what they dislike about NGS Graph Generator. It is important to improve web application in the future. In the response, some of the participants responded it could be difficult to visualise the splice variant loops in 3D as other parts of the graph may hide them. General difficulty in interpreting what the graph structure signifies apart from splice variant loops. Some of them responded the network could be confused to analyse. A major problem for them is to wait quite a long time to generate a large gene. When they were laying out the network, and their computer crashed.

The most important feedback on network approaches lack information such as sequence coverage, chromosome location and comparison to the reference sequence (genome) can be hard to the participant. Other than that, the network can be confusing to identify the exon location. They also find hard to determine which isoform and alternative splicing. Several participants responded it was hard to determine which isoform that has been spliced and they need to be more careful in analysing and visualising the data. The major problem is that they confused about the next things to do after generating this network.

#### **5.3.7.12 Question 12:** Can you give overall feedback, suggestions or recommendation for this application, NGS Graph Generator?

In the final question, the participants were asked to give overall feedback to improve this NGS Graph Generator. Some of the participants responded it is a good web application for NGS data analysis but need to put more effort to improve the pipeline as well as the participant experience on analysing the data. This feedback would

change the overall network approach. They suggest that this would be a complementary to current existing approach. It will be more popular if it becomes an interactive platform.

One of the participants suggests that the NGS Graph Generator should work solely without needing to export the map into BioLayout *Express*<sup>3D</sup>/Miru for better visualisation. It might output a report document with information and graph statistics, e.g. transcript list and links to Ensembl database. The NGS Graph Generator should allow the participant to create an account on the website to save results.

Some participants concerned about the time for generating the network. They suggested this need to be improved especially when laying out time and if it is possible to faster the process of generating the network. One of the participants suggests that we should simplify graph and make a comparison output with Sashimi plot. This would be a better comparison with another method. This simplified graph would make faster to layout a graph.

While in the aspect of network approaches, the respondents mentioned that the network graph is a valuable resource for the analysis of NGS data especially in analysing alternative splicing of genes. They satisfied with the approaches because of easy-to-understand data. The ways to visualise splice variants is easy to use the tool, and it provides an interesting option to visualise transcript isoforms.

One of them gives feedback that this approach is a more intuitive and aesthetically pleasing method for analysing transcript variants than the current standard of Sashimi plots. This would be most of the participants suggest adding more information to the graph so then will be easy to understand the graph and determine the splice variant. Improve by adding more information such as sequence coverage, chromosome information, type of splice variant and possibility of splice variant graph. They suggest if it had information prior generating the graph before creating the graph.

## 5.4 Discussion and future recommendation works

There are some proposed directions and future work based on the results of this usability test and questionnaire survey. The results are important to evaluate and improve the usability and effectiveness for more various and complex analysis problems. In addition to making some of these features available in a future release of NGS Graph Generator, this has helped to improve the process of developing new features.

First of all, the user interface (UI) can be quite a problem a first-time participant specially to generate a graph. The UI is not straightforward; however, it can be improvised to enable the participant to generate a graph directly from the website easily. A search tab feature can be included to ensure the accessible and easy when searching a network of a gene in the gallery. Several features are still required to support the usability of the web application.

Another feature can be introduced in the web-application is to include a ‘compare and contrast’ between network-based visualisation and another visualisation tool output, i.e. Sashimi plot. Providing a Sashimi plots side-by-side to the network would be better to compare and contrast between two approaches. This would be beneficial to a participant to confirm a splicing variant from two different tools.

Lacking information on the network such as read depths was one of the major complaints of the network approach. To improve this, the pipeline should process data that include read coverage information on the nodes to ensure the analysis using network-based which is similarly offered by tools such as IGV/Sashimi plots. Furthermore, parameters to generate a network, such as a graphical summary of network analysis of multiple sample/tissue could be added hence improve the usability of the tool. A participant voiced his minor dislike for a gallery on the website. In particular, the gene name of each network in the gallery needs to hover over the images to find the gene name. This process will be easier if the website has the search bar to find the desired gene.



All participants struggled to find the class set options, and they had to do some exploration on the web application to find them. An example of this was where they had to change the isoform information. Most of them thought it was automatically open the isoform information when they open up the layout. While most of them are impressed the way they visualise the network of RNA-seq, some of them found it difficult to know what they should do next and without the gene model or Ensembl browser to compare.

A long waiting time to generate a graph would be a major problem most of the participants. However, this could be improved by providing an estimation time for the application to finish generating a graph. The participant could be warned if the gene of interest is large and the uniqueness feature should be used. This test shows that this application which is already online could improve regarding their usability. The deficiencies found, lead to longer analysis time and confusion with a high potential of dislike to participants. Therefore, it seems to be necessary for the new algorithm to be implemented for the visualising of the big and large network. From the output of the test, it can be assumed, that mixed satisfying and unsatisfying usability bears a high potential for improvement in the future. Therefore, it is necessary to improve the user experience regarding usability and to implement these findings into the application improvements.

## 5.5 Conclusions

The focus during the development of NGS Graph Generator was to develop an alternative tool for visualising RNA sequencing data. A normal extension is toward a general statistical analysis such as the number of read coverage in the graph compared to other methods. My vision is to eventually encourage the biologist also to generate and analyse their RNA-seq data themselves and make them accessible through a personal account. Although most biologists would like to visualise their data better, the platform and visualisation tool must first be adopted by them. Hence, for now, this web-based application needs to continue including a different feature to the existing pipeline.

## Chapter 6 – General discussion and conclusions

High-throughput sequencing of RNA (RNA-seq) for the first time allows concurrent measurement of sequence and expression of RNAs. However, analysis of RNA-seq data remains a challenge for many biologists. Data generated by these platforms are large and complex, and the need to analyse these data has necessitated the development of novel bioinformatics approaches for mapping, analysis, and visualisation. A number of NGS visualisation tools can be used to examine the data including GBrowse (Donlin, 2002), BamView (Carver et al., 2010), UCSC Browser (Kent et al., 2002) and perhaps the most popular, Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013). In these cases, visualisation of RNA-seq data involves showing reads stacked onto a reference genome. Furthermore, these visualisation tools offer numerous useful features including flexibility to display read abundances on exons, transcripts, and junctions. Even though the Sashimi plots have prominent features, it also does have its limitations when it comes to a complex genome or transcript assemblies. The constraints include difficult to spot errors in assembly, to visualise splicing event and to deduce isoform expression.

Data visualisation is increasingly recognised as a key element of genomic and transcriptomic data analysis since it allows large and complex datasets to be better understood. However, interpretation of the data from visualisation tool is critical and essential to incorporate into findings. There are many ways to overcome the limitation, but an approach that has been gaining traction in biological research is derived from an application of network visualisation and analysis methods. Networks consist of nodes which usually represent an entity, e.g. genes, transcripts or proteins connected by edges i.e. lines where edges are experimentally or computationally derived relationships between them.

This thesis presents four stages of a study examining a novel means of analysing RNA-seq data through its visualisation as RNA-seq assembly networks. I have

developed a novel data processing pipeline to generate files for network construction. Furthermore, I have focused on the application of this method to study human transcript diversity in data derived from human cells and tissues where visualisation of resulting network allows for a deeper understanding of transcript structure, splice variation and issues with sequence assembly. Finally, I focused on usability test for NGS Graph Generator and BioLayout *Express*<sup>3D</sup> application.

My first task was to develop a pipeline to process RNA-seq read mapping ready for network visualisation which had been described in Chapter 2. The pipeline consists of a combination of different scripts was written in various programming languages and statistical packages such as Python, Bash, Perl, and R. However, this could be enhanced to enable a seamless integration of pipeline by using a single type of language rather than multiple languages. For instance, Ruby which is described as more intuitive and easier programming language to implement in this pipeline developed here, even though it is a less popular programming language compared to C++, Python, or Perl. This pipeline can also be implemented using web application framework such as Ruby on Rails (ROR) which allows multiple users to generate multiple data. Ultimately, the framework will be able to provide a seamless way to analyse and visualise RNA-seq data. With the revolution of the NGS sequencing machine and the affordability of sequencing more samples nowadays, this pipeline can evolve to become a better tool for network visualisation of NGS data which can be a method choice in the field of biology.

The pipeline has been devised to go from raw RNA-seq data mapping file to a file format which supports data visualisation as a 'cDNA assembly graph'. In cDNA assembly graphs, nodes represent sequence reads while edges denote a sequence homology between reads, above a defined threshold. Following the mapping of reads to a reference genome and defining which reads a map to a given locus, pairwise sequence alignments are performed between reads using MegaBLAST. It produces a matrix of weighted similarity scores that are used to define edges between reads. Visualisation of the resulting networks is then carried out using BioLayout

*Express*<sup>3D</sup>/Miru that can render large networks in 3-D, thereby allowing a better appreciation of the often-complex network structure.

The NGS Graph Generator is a web-based application for visualising RNA-seq data assemblies (see Chapter 2). While NGS Graph Generator is an underpinning technique, there is still room for improvements. I believe the individual transcript of network visualisation and the cases I have discussed will be informative to other data sets. For instance, it would be advantageous to determine if the RNA-seq network transcript data can be associated and link with other pathways such as cancer data or protein atlas data. Furthermore, this pipeline would serve as a foundation for future works such as pipeline integration with additional functionality to visualise and identify SNP marker using network approach.

The NGS Graph Generator is the first network-based visualisation approach that allows biologists to explore transcripts and alternative isoforms within or across different samples. It can also overcome the limitations of SeqGrpheR and GraphNGS (see Chapter 2, Section 2.2) as it underlines the use of PDF and less efficient visualisation performance. Furthermore, in several cases, I have shown that network visualisation outperformed Sashimi plots where the determination of isoform expression and splice junction can sometimes be complicated in the Sashimi plots. For instance, in the case of network transcript of *SORBS2* (see Chapter 4, Figure 4.14) and *TPMI* (see Chapter 4, Figure 4.20) where the visualisation of splice junction was very difficult to visualise and hard to determine the isoform expression. Even though network analyses performed better than Sashimi plots, however, there are also a few setbacks of network analysis approaches. Currently, the NGS Graph Generator can process only a gene/transcript per analysis. It would be impossible if we are investigating into the genome-wide analysis. However, there are few ways to overcome this limitation which is including the implementation of a different network RNA-seq assembly drawing algorithm to reduce the computation of network layout processes. Looking further into the future, I foresee network tools can visualise big network, smoother visualisation of isoform expression and edge thresholds can be filtered dynamically in more detail analysis of network structure.

Side-by-side comparison with existing approaches such as Sashimi plot or Vials would give a better comparison and details analysis on the analysis of selected genes. While visualising a network of the gene with additional information such as exon information and read coverage when a user hovers on the network would give the great advantage of this approaches.

While a process of read-to-read comparison using MegaBLAST can be computationally expensive for a highly expressed gene, e.g. *TUBA1C* or *GAPDH*. However, a recently published tool called HS-BLASTN (Chen et al., 2015) can be implemented to improve the speed of reads comparison. The computational speed is greatly faster than MegaBLAST used in this study. The HS-BLASTN is 22 times faster than MegaBLAST, and it demonstrates better parallel performance. This implementation will improve the overall performance of NGS Graph Generator which mainly reaches the goal to investigate genome-wide transcriptome using network analysis.

The NGS Graph Generator integrates information needed for isoform analysis, such as isoform transcript ID and exons but it needs more improvement. For instance, more useful information such as read coverage for each exon is essential and will be eventually transformed to bar plots to illustrate the isoform expression of such gene that can be added. It will not only provide an overview of isoforms expression but also enables one to exploit the data associated with individual isoform. I have shown in several cases that the network-based technique is compatible in determining alternative splice isoform using this plot which I manually generated for *TPM1*.

This pipeline was turned into a web-based application for RNA-seq network visualisation analysis. Another way to utilise this pipeline is by downloading from GitHub or running it as an Amazon Machine Image (AMI). This pipeline has formed the basis for my subsequent work on the exploring and analysing alternative splicing in human RNA-seq data. A platform for community curation of network transcript visualisation can be developed which will benefit others in this field.

Chapter 3 describes my initial explorations on the potential visualising of DNA assembly networks for the analysis of transcript diversity in short read data. The work aimed to understand better fundamental and challenges related to the network visualisation of RNA-seq data particularly on how it could be used to visualise transcript structure, isoform divergence, and splice variation. In order to start the network exploration, these analyses were performed on RNA-seq data produced from four samples of human fibroblasts which were taken at a different stage of the human cell cycle. A cleanup and quality control were performed on all dataset using Kraken pipeline to remove low-quality data before constructing network graph using NGS Graph Generator.

Nonetheless, the first challenges I encountered was the fact that existing network layout algorithm (Fruchterman-Reingold) implemented within BioLayout *Express*<sup>3D</sup> did not produce an optimal layout of unusual network structures produced in these analyses. Initial visualisations of transcript networks, e.g. *COL5A1* were poorly laid out, twisted and knotted structure of the networks which was making them impossible to interpret. However, following implementation of an advanced layout algorithm FMMM in the BioLayout *Express*<sup>3D</sup>, visualising network structure could be far better appreciated. After that, work began on optimising the best network visualisation conditions with a series of different datasets. This includes using ‘real’ and ‘synthetic’ sequence data to find a generalised parameter. The default parameter settings to build up networks are varied for each gene, even though I discovered that  $p=98$  and  $l=31$  would be the best settings. Nevertheless, the parameters should be less stringent when working with a highly expressed gene or vice versa. After using default parameter for transcript network construction, I observed the majority of genes sequenced to a sufficient depth and assembled into networks with a linear ‘corkscrew’ structure and when representing single isoform transcripts, add little to existing views of these data. The most reasonable evidence of such network is because of only one major isoform expressed in the cell cycle genes.

However, in a few number of cases (~5%), the RNA-seq assembly transcript networks in human fibroblasts possess more complex structures with ‘loops,’ ‘knots,’

and multiple ends being observed. In most of cases examined, these loops were associated with alternative splicing events, a fact confirmed by RT-PCR analyses. Instrumental for this work was a different network structure. For this, I have used BLAST and Ensembl browser for confirmation network structure and gene model for such gene. This includes genes such as *MKI67* that exhibit knot-like structures, which was found to be due to the presence of repetitive sequence within an exon of the gene. In this situation, it can be an extra feature of network analysis to identify such repetitive sequence, especially when dealing with another organism such as animal and plant. It is well known that repetitive DNA sequences are mostly in plant and animal genome. This network analysis can be a potential approach to analyse such data by integrating a repetitive sequence database, e.g. RepeatMasker or RepBase.

In another case of *CENPO*, unusual structure observed was due to reads derived from an overlapping gene of the *ADCY3* present on the opposite strand with reads being wrongly mapped to *CENPO*. It is important to have a stranded rather than unstranded sequencing for better analyses and accurate interpretation of the data, e.g. which strand of the RNA is transcribed. When I analysed the human fibroblasts samples, I discovered that most networks of DNA assembly are linear and some are complex. For instance, it depends on the sequence reads mapped to such gene and it sometimes does not simply turn into a theoretical model. However, to build up an RNA-seq assembly transcript, the network needs to consider the amount of read mapping of such gene.

Therefore, I explored the use of a network reduction strategy as an approach to visualising highly expressed genes such as *GAPDH* and *TUBA1C*. It is near impossible to layout such highly expressed gene that contains a huge number of redundant sequences (deep coverage). This observation triggered the development of the reduction of identical sequence reads of these genes. While redundancy of sequence reads is being represented by the size of nodes, very small nodes observed denote sequencing error in the data. This can be improved by lowering threshold to remove small nodes in the network assembly. Another insight that can be suggested

is that simplification of the RNA-seq network assembly in determining the network structure prior layout visualisation, but it requires a different algorithm to process the data. The strategy of having information on network structure will give more advantages on genome-wide analysis. Furthermore, successfully demonstrates the utility of networks in analysing transcript isoforms of data derived from a single cell type which I have set out to explore its utility in analysing transcript variation in tissue data where multiple isoforms expressed by different cells within the tissue might be present in each sample.

In Chapter 4, I used the same experimental pipeline and methodology with additional tools to explore isoform divergence in an RNA-seq dataset derived from human tissues. There are two parts in this chapter which are quality control and detection of a splicing event in the data. In the first half, quality control of these data was performed using a network-based approach based on co-expressed between genes and samples. When a sample-to-sample correlation network analysis was employed with edges which represented the Pearson correlation value and nodes represented samples, I found a number of replicates which apparently were not grouped with similar samples and these samples then were removed for subsequent analysis. Only 77 out of 95 samples derived from human tissues passed the quality control. The most likely errors happened when collecting or processing these samples. However, this shows one of the advantages of network analysis, which has been successfully detected the lowest tissues correlation of samples in the datasets. Hence, a network was constructed using a correlation threshold of  $r \geq 0.85$ , which comprises of 6,109 nodes (genes) and 1,091,477 edges (correlations) and clustered using MCL algorithm. Subsequently, the profile and gene content of each cluster was examined, and enrichment of GO terms was analysed.

In the second half of this chapter, the aims were to detect and analyse alternative splicing events between different tissues using splice variant detector tool. For this purpose, I assessed the alternative splicing events detected using rMATS tools. The tools reported 4,992 splicing events in the tissue comparisons of brain vs heart, 4,804 events in the brain vs liver and 3,990 events in the heart vs liver based on the false-



discovery rate (FDR) cut-off of  $< 0.01$ . The further threshold was applied to select the best candidates for network analysis of the gene with more than 50% of exon inclusion level and expression level more than FPKM 30. It ended up producing a list of 78 splicing events from 52 genes.

However, not all the gene transcripts can be constructed as network due to the low expression level and eventually only a few genes with a fair amount of expression were selected for further analysis. This includes the complex network structure of transcripts diversity derived from the tissue, and cDNA assembly networks for *KLC1*, *GUK1*, *SORBS2*, and *TPM1* were explored. Each of these networks exhibited different types of alternative splicing events, and it was sometimes difficult to determine the isoforms expressed between tissues using other approaches. I discovered that isoform expression can be determined and has additional advantages over other 'read-stack to reference genome' visualisation platform. For instance, there is a setback when I want to visualise the assembly of long genes such as *KLC1* and *SORBS2* in Sashimi plot or even Vials. This setback is just because of the number of exons and the size of their genomic loci.

The network reconstruction of these transcripts, however, easily identifies specific isoforms and exons expressed in a tissue. Especially when working on *SORBS2* network where novel isoforms were detected based on the evidence from the network and compared with existing Ensembl data. The only isoform that supports start at exon 21 in the RNA-seq network assembly of *SORBS2* in heart missed exon 31 is likely to produce novel isoform. Therefore, further investigation is needed. However, such evidence could potentially show that network analysis could detect not only the alternative splicing but also identify novel isoform. The way to identify novel isoform is based on the algorithm and statistical analysis of such tool as Cufflinks and TopHat. Undoubtedly, this is impossible when it comes to visualising data and to identify novel isoform if we depend only on Sashimi plot or another visualisation platform. Hence, network analysis has the potential in identifying novel isoform.

Arguably, the most complex analysis is the RNA-seq mapping transcript network of *TPMI* where the uniquification step was employed for this highly expressed gene of heart. However, the other two network structures were perhaps simpler but need a careful approach in identifying the isoform expression. While I believe that network-based visualisation of RNA-seq is a new and useful alternative splicing visualisation technique, it also has some limitations. The limitations include the visualisation of RNA-seq data that are limited to a single gene or locus only compared to Sashimi plot where all information related to one or more sample is in a single tool. It is more difficult in network-based approach to visualising vast network structure in a standard machine as they require a high-quality graphics card for rendering 3-D network structure. However, in these two visualisation tools, Sashimi plots and Vials have different ultimate goals where Vials is developed as an exploratory data analysis tool for visualisation of lots of samples; while Sashimi plots are fit for visualisation of individual samples (Strobelt et al., 2016).

In chapter 5, I conducted a usability test for NGS Graph Generator. This test was important to ensure that the application is well received and utilised by the user. Almost all participants of this usability test agree that this application would encourage biologists to visualise and understand the alternative splicing together with existing tools. The participants agreed that Sashimi plot rather difficult to view and visualise and perhaps would lose something interesting features. However, there were also reviews of this application that need improvements such as the capability to analyse big network in a short time, side-by-side analysis of network with Sashimi plot and Ensembl. Additional information of the network would be necessary to improve the understanding of the alternative splicing.

While working on these sets of data, there are few limitations which can be observed from the network analysis. In early of this work, I expected to analyse a genome-wide analysis of isoform expression and a well-defined mixture of a set of different isoforms. For instance, RNA-seq sequence assemblies of isoforms from different tissue can be distinguished better. However, it turns out to be impossible to lay out a multiple or combination of more than one sample which is not only difficult in

interpreting its data but also adding the computational setback. The network analysis of RNA-seq data has proven the advantages over other visualisation approaches, though it exhibits some limitation which can be improved. Furthermore, many NGS applications such as genomic, metagenomic and epigenomic have yet to be explored using network-based approaches. I believe that these areas have something to offer and provide a better solution to understanding the biological question.

The development of a new approach to network visualisation of NGS data using network visualisation platform is providing a detailed enhancing characterisation of the splicing events characterisation. In the last two chapters, I have demonstrated how graph network using RNA-seq data enables the elucidation of understanding the gene expression study by visualising the pattern of the network.

Next-generation sequencing technologies are growing swiftly, and it is expected that RNA-seq will become routine for many laboratories in the next few years. The area of alternative splicing variation analysis using RNA-seq data is still in its infancy and would benefit from new approaches and strategies. An extensive assessment and comparison of the existing methods would be advantageous, and until now, there is no common agreement regarding the method that performs best under particular conditions and situations. We are anticipating seeing the novel visualising methods to be developed and discovered in this growing field shortly.

The advancement of sequencing technology and computational analyses have significantly increased our knowledge of gene transcription and its regulation. However, many challenges remain to be addressed. Difficulties in identifying and visualising the isoform expressed of the human gene exist at a different level, whether within a cell or tissue, significant amounts of intervening (noncoding) sequences, and the development of computational analysis tools steadily increased. For instance, using network-based approach, these issues can be solved, and the isoform can be determined. It will be impossible to visualise multiple genes using this method. However, the ultimate solution of annotation lies in developing a simple

yet constructive approach to reducing the amount of computational time when performing a read comparison and speed up the drawing network structure.

I believe the data pipeline and tools presented here will provide an analytical platform that will be a useful addition to the available tools for the analysis of the huge amounts of complex but information-rich data produced by modern DNA sequencing machines. This new approach of visualising RNA-seq data is crucial to discover isoform expression, especially related to cancer and diseases. The insight from alternative splicing isoform using network-based visualisation not only better our understanding about splicing expression regulation but leads to advance scientific knowledge of the biological processes of the disease.

In conclusion, this work demonstrates the utility of network visualisation of RNA-seq data where the unusual structure of these networks can be used to identify issues in assembly, repetitive sequences within transcripts and splice variation. As such, this approach has the potential to improve our understanding of transcript complexity significantly. In summary, this thesis demonstrates that network-based visualisation provides a new and complementary approach to characterise alternative splicing of RNA-seq data and has the potential to be useful for the analysis and interpretation of other kinds of sequencing data.

## References

- Alberts, B., Johnson, A., and Lewis, J. (2002). *From DNA to RNA* (New York: Garland Science).
- Alekseyenko, A.V., Kim, N., and Lee, C.J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA* *13*, 661–670.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* *215*, 403–410.
- Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.* *22*, 2008–2017.
- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* *31*, 166–169.
- Andersson, M.E., Sjölander, A., Andreassen, N., Minthon, L., Hansson, O., Bogdanovic, N., Jern, C., Jood, K., Wallin, A., Blennow, K., et al. (2007). Kinesin gene variability may affect tau phosphorylation in early Alzheimer’s disease. *Int. J. Mol. Med.* *20*, 233–239.
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Ansorge, W., Sproat, B., Stegemann, J., Schwager, C., and Zenke, M. (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* *15*, 4593–4602.
- Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziusko, O., Vitezic, M., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* *347*, 1010–1014.
- Ast, G. (2004). How did alternative splicing evolve? *Nat. Rev. Genet.* *5*, 773–782.
- Bader, G., and Enright, A. (2005). Biomolecular interactions and biological pathways. In: Baxevanis AD, editor. *Bioinformatics: A practical analysis of genes and proteins*. (John Wiley New York).
- Baer, R., Bankier, A.T., Biggin, M.D., Deininger, P.L., Farrell, P.J., Gibson, T.J., Hatfull, G., Hudson, G.S., Satchwell, S.C., Séguin, C., et al. (1984). DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature* *310*, 207–211.
- Bah, A., and Forman-Kay, J.D. (2016). Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *J. Biol. Chem.* *291*, 6696–6705.

- Baker, M. (2010). Next-generation sequencing: adjusting to data overload. *Nat. Methods* 7, 495–499.
- Batut, P., Dobin, A., Plessy, C., Carninci, P., and Gingeras, T.R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* 23, 169–180.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.
- Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40, e72.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.
- Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C., et al. (2010). Integrative analysis of the melanoma transcriptome. *Genome Res.* 20, 413–427.
- Bhasi, A., Philip, P., Sreedharan, V.T., and Senapathy, P. (2009). AspAlt: A tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. *Genomics* 94, 48–54.
- Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. *Science* 277, 1453–1462.
- Blencowe, B.J., Ahmad, S., and Lee, L.J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.* 23, 1379–1386.
- Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527.
- Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7, 488.
- Buermans, H.P.J., and den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1842, 1932–1941.
- Burgess, D.J. (2014). Alternative splicing: Retaining introns to sculpt gene expression. *Nat. Rev. Genet.* 15, 707–707.

- Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* *40*, 722–729.
- Carrara, M., Lum, J., Cordero, F., Beccuti, M., Poidinger, M., Donatelli, S., Calogero, R.A., and Zolezzi, F. (2015). Alternative splicing detection workflow needs a careful combination of sample prep and bioinformatics analysis. *BMC Bioinformatics* *16*, S2.
- Carver, T., Böhme, U., Otto, T.D., Parkhill, J., and Berriman, M. (2010). BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* *26*, 676–677.
- Carver, T., Harris, S.R., Berriman, M., Parkhill, J., and McQuillan, J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* *28*, 464–469.
- Cech, T.R., and Steitz, J.A. (2014). The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* *157*, 77–94.
- Chen, L., Tovar-Corona, J.M., and Urrutia, A.O. (2012). Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. *Int. J. Evol. Biol.* *2012*.
- Chen, Y., Ye, W., Zhang, Y., and Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.* gkv784.
- Chimani, M. (2007). The open graph drawing framework. (15th International Symposium on Graph Drawing), p.
- Chu, Y., and Corey, D.R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* *22*, 271–274.
- Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* *38*, 1767–1771.
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* *393*, 537–544.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* *17*.
- Cooper, G. (2000). *The Complexity of Eukaryotic Genomes - The Cell - NCBI Bookshelf* (Sunderland (MA): Sinauer Associates).

- Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* *16*, 1–10.
- Darnell, J.E. (2013). Reflections on the history of pre-mRNA processing and highlights of current knowledge: A unified picture. *RNA* *19*, 443–460.
- Das, S., and Vikalo, H. (2013). Base calling for high-throughput short-read sequencing: dynamic programming solutions. *BMC Bioinformatics* *14*, 129.
- Davidson, R., and Harel, D. (1996). Drawing graphs nicely using simulated annealing. *ACM Trans Graph* *15*, 301–331.
- Davis, M.P.A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A.J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods San Diego Calif* *63*, 41–49.
- De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* *4*, 49–60.
- Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE* *8*, e85024.
- Denti, M.A., Viero, G., Provenzani, A., Quattrone, A., and Macchi, P. (2013). mRNA fate: Life and death of the mRNA in the cytoplasm. *RNA Biol.* *10*, 360–366.
- Dhaenens, C.-M., Van Brussel, E., Schraen-Maschke, S., Pasquier, F., Delacourte, A., and Sablonnière, B. (2004). Association study of three polymorphisms of kinesin light-chain 1 gene with Alzheimer's disease. *Neurosci. Lett.* *368*, 290–292.
- van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* *30*, 418–426.
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* *14*, 671–683.
- Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A.M., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. *Nature* *489*, 101–108.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013a). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15–21.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013b). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15–21.



- van Dongen, S. (2000). Graph clustering by flow simulation. PD Thesis. University of Utrecht.
- Donlin, M.J. (2002). Using the Generic Genome Browser (GBrowse). In *Current Protocols in Bioinformatics*, (John Wiley & Sons, Inc.), p.
- Dozmorov, M.G., Adrianto, I., Giles, C.B., Glass, E., Glenn, S.B., Montgomery, C., Sivils, K.L., Olson, L.E., Iwayama, T., Freeman, W.M., et al. (2015). Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. *BMC Bioinformatics* *16*, S10.
- Dudley, J.T., and Butte, A.J. (2010). In silico research in the era of cloud computing. *Nat. Biotechnol.* *28*, 1181–1185.
- Dudley, J.T., Pouliot, Y., Chen, R., Morgan, A.A., and Butte, A.J. (2010). Translational bioinformatics in the cloud: an affordable alternative. *Genome Med.* *2*, 51.
- Eades, P. (1984). A heuristic for graph drawing. *Congr. Numerantium* *42*, 149–160.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* *323*, 133–138.
- Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* *14*, 496–506.
- Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., The RGASP Consortium, Räsch, G., Goldman, N., Hubbard, T.J., Harrow, J., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* *10*, 1185–1191.
- Enright, A.J., and Ouzounis, C.A. (2001). BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics* *17*, 853–854.
- Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* *30*, 1575–1584.
- Enright, A.J., Kunin, V., and Ouzounis, C.A. (2003). Protein families and Tribes in genome sequence space. *Nucleic Acids Res.* *31*, 4632–4638.
- Eswaran, J., Horvath, A., Godbole, S., Reddy, S.D., Mudvari, P., Ohshiro, K., Cyanam, D., Nair, S., Fuqua, S.A.W., Polyak, K., et al. (2013). RNA sequencing of cancer reveals novel splicing alterations. *Sci. Rep.* *3*.
- Fagerberg, L., Oksvold, P., Skogs, M., Algenäs, C., Lundberg, E., Pontén, F., Sivertsson, A., Odeberg, J., Klevebring, D., Kampf, C., et al. (2013). Contribution of antibody-based protein profiling to the human Chromosome-centric Proteome Project (C-HPP). *J. Proteome Res.* *12*, 2439–2448.

- Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpour, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Mol. Cell. Proteomics* *13*, 397–406.
- Filloux, C., Cédric, M., Romain, P., Lionel, F., Christophe, K., Dominique, R., Abderrahman, M., and Daniel, P. (2014). An integrative method to normalize RNA-Seq data. *BMC Bioinformatics* *15*, 188.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., and al, et (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* *269*, 496.
- Flicek, P., and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* *6*, S6–S12.
- Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* *42*, D749–D755.
- Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* *28*, 3169–3177.
- Frankish, A., Uszczyńska, B., Ritchie, G.R., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R., et al. (2015). Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* *16*, 1–11.
- Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S., Mazière, P., Grocock, R.J., Freilich, S., Thornton, J., and Enright, A.J. (2007). Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data. *PLoS Comput Biol* *3*, e206.
- Freeman, T.C., Ivens, A., Baillie, J.K., Beraldi, D., Barnett, M.W., Dorward, D., Downing, A., Fairbairn, L., Kapetanovic, R., Raza, S., et al. (2012). A gene expression atlas of the domestic pig. *BMC Biol.* *10*, 90.
- Fruchterman, T.M.J., and Reingold, E.M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exp.* *21*, 1129–1164.
- Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., et al. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* *10*, 161.
- Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* *461*, 186–192.

- Gajer, P., Goodrich, M.T., and Kobourov, S.G. (2001). A Multi-dimensional Approach to Force-Directed Layouts of Large Graphs. In *Graph Drawing*, J. Marks, ed. (Springer Berlin Heidelberg), pp. 211–221.
- Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477.
- Gelinas, R.E., and Roberts, R.J. (1977). One predominant 5'-undecanucleotide in adenovirus 2 late messenger RNAs. *Cell* 11, 533–544.
- Godiyal, A., Hoberock, J., Garland, M., and Hart, J.C. (2009). Rapid Multipole Graph Drawing on the GPU. In *Graph Drawing*, I.G. Tollis, and M. Patrignani, eds. (Springer Berlin Heidelberg), pp. 90–101.
- Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D.C., and Shyr, Y. (2013). Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. *PLOS ONE* 8, e71462.
- Hachul, S., and Jünger, M. (2004). Drawing Large Graphs with a Potential-Field-Based Multilevel Algorithm. In *Graph Drawing*, J. Pach, ed. (Springer Berlin Heidelberg), pp. 285–295.
- Hachul, S., and Jünger, M. (2005). Drawing Large Graphs with a Potential-Field-Based Multilevel Algorithm. In *Graph Drawing*, J. Pach, ed. (Springer Berlin Heidelberg), pp. 285–295.
- Hachul, S., and Jünger, M. (2007). Large-Graph Layout Algorithms at Work: An Experimental Study. *J. Graph Algorithms Appl.* 11, 345–369.
- Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced Applications of RNA Sequencing and Challenges. *Bioinforma. Biol. Insights* 9, 29–46.
- Harel, D., and Koren, Y. (2001). A Fast Multi-scale Method for Drawing Large Graphs. In *Graph Drawing*, J. Marks, ed. (Springer Berlin Heidelberg), pp. 183–196.
- Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774.
- Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Inf. Syst.* 47, 98–115.
- Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N., and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *BioTechniques* 29, 548–550, 552–554, 556 passim.

- Herman, I., Melancon, G., and Marshall, M.S. (2000). Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE Trans. Vis. Comput. Graph.* 6, 24–43.
- Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W., et al. (2008). Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* 5, 183–188.
- Holt, R.A., and Jones, S.J.M. (2008). The new paradigm of flow cell sequencing. *Genome Res.* 18, 839–846.
- Hooper, J.E. (2014). A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum. Genomics* 8, 3.
- Hoskins, A.A., and Moore, M.J. (2012). The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem. Sci.* 37, 179–188.
- Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.-F., Hammond, S.M., Makowski, L., et al. (2013). DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* 41, e39.
- Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9, 868–877.
- International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.
- Irimia, M., and Roy, S.W. (2014). Origin of Spliceosomal Introns and Alternative Splicing. *Cold Spring Harb. Perspect. Biol.* 6, a016071.
- Jalkanen, A.L., Coleman, S.J., and Wilusz, J. (2014). Determinants and Implications of mRNA Poly(A) Tail Size - Does this Protein Make My Tail Look Big? *Semin. Cell Dev. Biol.* 0, 24–32.
- Jantzen, S.G., Sutherland, B.J., Minkley, D.R., and Koop, B.F. (2011). GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets. *BMC Res. Notes* 4, 267.
- Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16, 1370–1386.
- Joehanes, R., Ying, S., Huan, T., Johnson, A.D., Raghavachari, N., Wang, R., Liu, P., Woodhouse, K.A., Sen, S.K., Tanriverdi, K., et al. (2013). Gene Expression Signatures of Coronary Heart Disease. *Arterioscler. Thromb. Vasc. Biol.* 33, 1418–1426.
- Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729.

- Kamada, T., and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* *31*, 7–15.
- Karamysheva, Z., Díaz-Martínez, L.A., Warrington, R., and Yu, H. (2015). Graded requirement for the spliceosome in cell cycle progression. *Cell Cycle* *14*, 1873–1883.
- Katz, Y., Wang, E.T., Airoidi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* *7*, 1009–1015.
- Katz, Y., Wang, E.T., Silterra, J., Schwartz, S., Wong, B., Mesirov, J.P., Airoidi, E.M., and Burge, C.B. (2013). Sashimi plots: Quantitative visualization of RNA sequencing read alignments. *ArXiv13063466 Q-Bio*.
- Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. *Gene* *514*, 1–30.
- Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Res.* *12*, 656–664.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Res.* *12*, 996–1006.
- Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* *11*, 345–355.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* *14*, R36.
- Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., et al. (2006). Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* *16*, 55–65.
- Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* *390*, 364–370.
- de Klerk, E., and ‘t Hoen, P.A.C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* *31*, 128–139.
- Kobourov, S.G. (2012). Spring Embedders and Force Directed Graph Drawing Algorithms.

- Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* *14*, 153–165.
- Kosmyna, B., and Query, C.C. (2016). Structural biology: Catalytic spliceosome captured. *Nature advance online publication*.
- Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. *Nat. Rev. Mol. Cell Biol.* *12*, 283–294.
- Kwak, H., and Lis, J.T. (2013). Control of Transcriptional Elongation. *Annu. Rev. Genet.* *47*, 483–508.
- Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* *2013*, 463786.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* *409*, 860–921.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* *10*, R25.
- Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* *9*, e1003118.
- Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* *7*, 709–715.
- Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* *5*.
- Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* *12*, 323.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* *25*, 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* *25*, 2078–2079.
- Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* *13*, 2178–2189.

- Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967.
- Li, S., Łabaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, L., Wang, M., Wang, C., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32, 888–895.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* 2012.
- Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular Cell Biology* (W. H. Freeman).
- Logarinho, E., Resende, T., Torres, C., and Bousbaa, H. (2008). The human spindle assembly checkpoint protein Bub3 is required for the establishment of efficient kinetochore-microtubule attachments. *Mol. Biol. Cell* 19, 1798–1813.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
- Mabbott, N.A., Baillie, J.K., Brown, H., Freeman, T.C., and Hume, D.A. (2013). An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* 14, 632.
- Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., and Chinnaiyan, A.M. (2009). Transcriptome Sequencing to Detect Gene Fusions in Cancer. *Nature* 458, 97–101.
- Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9, 34.
- Mardis, E.R. (2013). Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* 6, 287–303.
- Marguerat, S., and Bähler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci. CMLS* 67, 569–579.
- Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.
- Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682.
- Marx, V. (2013). Genomics in the clouds. *Nat. Methods* 10, 941–945.
- Matera, A.G., and Wang, Z. (2014). A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* 15, 108–121.

- Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* *74*, 560–564.
- Melamud, E., and Moulton, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Res.* *37*, 4873–4886.
- Merkhofer, E.C., Hu, P., and Johnson, T.L. (2014). Introduction to Cotranscriptional RNA Splicing. *Methods Mol. Biol. Clifton NJ* *1126*, 83–96.
- Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* *338*, 1593–1599.
- Metzker, M.L. (2010). Sequencing technologies — the next generation. *Nat. Rev. Genet.* *11*, 31–46.
- Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* *95*, 315–327.
- Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet—next generation sequence assembly visualization. *Bioinformatics* *26*, 401–402.
- Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* *12*, R112.
- Morel, M., Héraud, C., Nicaise, C., Suain, V., and Brion, J.-P. (2011). Levels of kinesin light chain and dynein intermediate chain are reduced in the frontal cortex in Alzheimer’s disease: implications for axoplasmic transport. *Acta Neuropathol. (Berl.)* *123*, 71–84.
- Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., and Schäffer, A.A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics* *24*, 1757–1764.
- Morozova, O., and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* *92*, 255–264.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* *5*, 621–628.
- Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* *420*, 520–562.



- Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P., et al. (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* *10*, 354–360.
- Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., et al. (2000). A Whole-Genome Assembly of *Drosophila*. *Science* *287*, 2196–2204.
- Naftelberg, S., Schor, I.E., Ast, G., and Kornblihtt, A.R. (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* *84*, 165–198.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* *320*, 1344–1349.
- Nagalakshmi, U., Waern, K., and Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Curr. Protoc. Mol. Biol.* Ed. Frederick M Ausubel *Chapter 4*, Unit 4.11.1-13.
- Nielsen, J. (2012). Usability 101: Introduction to Usability.
- Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* *11*, 378.
- von Otter, M., Landgren, S., Nilsson, S., Lundvall, C., Minthon, L., Bogdanovic, N., Andreasen, N., Gustafson, D.R., Skoog, I., Wallin, A., et al. (2010). Kinesin light chain 1 gene haplotypes in three conformational diseases. *Neuromolecular Med.* *12*, 229–236.
- Pacific, B. (2014). Data Release: Whole Human Transcriptome from Brain, Heart, and Liver.
- Padgett, R.A. (2012). New connections between splicing and human disease. *Trends Genet.* *28*, 147–154.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* *40*, 1413–1415.
- Pariset, L., Chillemi, G., Bongiorno, S., Romano Spica, V., and Valentini, A. (2009). Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *New Biotechnol.* *25*, 272–279.
- Paule, M.R., and White, R.J. (2000). Transcription by RNA polymerases I and III. *Nucleic Acids Res.* *28*, 1283–1298.

- Peng, Z., Cheng, Y., Tan, B.C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* *30*, 253–260.
- Pérez-Ortín, J.E., Alepuz, P., Chávez, S., and Choder, M. (2013). Eukaryotic mRNA Decay: Methodologies, Pathways, and Links to Other Stages of Gene Expression. *J. Mol. Biol.* *425*, 3750–3775.
- Pimentel, H., Parra, M., Gee, S., Ghanem, D., An, X., Li, J., Mohandas, N., Pachter, L., and Conboy, J.G. (2014). A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res.* *42*, 4031–4042.
- Pirim, H., Ekşioğlu, B., Perkins, A., and Yüceer, Ç. (2012). Clustering of High Throughput Gene Expression Data. *Comput. Oper. Res.* *39*, 3046–3061.
- Pohl, M., Bortfeldt, R.H., Grützmann, K., and Schuster, S. (2013). Alternative splicing of mutually exclusive exons—A review. *Biosystems* *114*, 31–38.
- Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* *10*, 354–366.
- Porrua, O., and Libri, D. (2013). RNA quality control in the nucleus: The Angels' share of RNA. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* *1829*, 604–611.
- Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* *42*, D756–D763.
- Quigley, A., and Eades, P. (2001). FADE: Graph Drawing, Clustering, and Visual Abstraction. In *Graph Drawing*, J. Marks, ed. (Springer Berlin Heidelberg), pp. 197–210.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* *26*, 841–842.
- Ramanathan, A., Robb, G.B., and Chan, S.-H. (2016). mRNA capping: biological functions and applications. *Nucleic Acids Res.* gkw551.
- Ramsay, G. (1998). DNA chips: state-of-the art. *Nat. Biotechnol.* *16*, 40–44.
- Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P.M., and Thompson, J.F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PloS One* *6*, e19287.
- Robertson, M.P., and Joyce, G.F. (2012). The Origins of the RNA World. *Cold Spring Harb. Perspect. Biol.* *4*.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nat. Biotechnol.* *29*, 24–26.

- Rogers, M.F., Thomas, J., Reddy, A.S., and Ben-Hur, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* *13*, R4.
- Rosenberg, R.N., and Pascual, J.M. (2014). *Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease: Fifth Edition* (Elsevier).
- Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* *475*, 348–352.
- Rubel, O., Weber, G.H., Huang, M.Y., Bethel, E.W., Biggin, M.D., Fowlkes, C.C., Hendriks, C.L.L., Keranen, S.V.E., Eisen, M.B., Knowles, D.W., et al. (2010). Integrating Data Clustering and Visualization for the Analysis of 3D Gene Expression Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* *7*, 64–79.
- Rubin, J., Chisnell, D., and Spool, J. (2008). *Wiley: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd Edition.
- Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* *239*, 487–491.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977a). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* *74*, 5463–5467.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977b). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* *265*, 687–695.
- Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Res.* *20*, 1165–1173.
- Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* *270*, 467–470.
- Schoenberg, D.R., and Maquat, L.E. (2012). Regulation of cytoplasmic mRNA decay. *Nat. Rev. Genet.* *13*, 246–259.
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* *473*, 337–342.
- Sevimoglu, T., and Arga, K.Y. (2014). The role of protein interaction networks in systems biomedicine. *Comput. Struct. Biotechnol. J.* *11*, 22–27.
- Shah, N.H., and Muir, T.W. (2014). Inteins: nature's gift to protein chemists. *Chem Sci* *5*, 446–461.

- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* *13*, 2498–2504.
- Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z., Zhou, Q., Carstens, R.P., and Xing, Y. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* gkr1291.
- Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* *111*, E5593-5601.
- Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* *26*, 1135–1145.
- Shi, Y., and Jiang, H. (2013). rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. *PloS One* *8*, e79448.
- Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci.* *109*, 1347–1352.
- Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., et al. (2016). The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* *44*, D717–D725.
- Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., The RGASP Consortium, Hubbard, T.J., Guigó, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* *10*, 1177–1184.
- Stein, L.D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* *11*, 207.
- Strobelt, H., Alsallakh, B., Botros, J., Peterson, B., Borowsky, M., Pfister, H., and Lex, A. (2016). Vials: Visualizing Alternative Splicing of Genes. *IEEE Trans. Vis. Comput. Graph.* *22*, 399–408.
- Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., and Kocher, J.-P.A. (2016). Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* bbw069.
- Supper, J., Gugenmus, C., Wollnik, J., Druke, T., Scherf, M., Hahn, A., Grote, K., Bretschneider, N., Klocke, B., Zinser, C., et al. (2013). Detecting and visualizing gene fusions. *Methods* *59*, S24–S28.
- Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W.S. (2012). Alternative splicing in plants – coming of age. *Trends Plant Sci.* *17*, 616–623.

- Tang, Z., Shu, H., Oncel, D., Chen, S., and Yu, H. (2004). Phosphorylation of Cdc20 by Bub1 provides a catalytic mechanism for APC/C inhibition by the spindle checkpoint. *Mol. Cell* 16, 387–397.
- Tariq, M.A., Kim, H.J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res.* 39, e120.
- Tazi, J., Bakkour, N., and Stamm, S. (2009). Alternative splicing and disease. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1792, 14–26.
- The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.
- Theocharidis, A., van Dongen, S., Enright, A.J., and Freeman, T.C. (2009). Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat. Protoc.* 4, 1535–1550.
- Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192.
- Tian, B., and Manley, J.L. (2013). Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem. Sci.* 38, 312–320.
- Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578.
- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 31, 46–53.
- Tress, M.L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol.* 9, R162.
- Trivedi, U.H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., and Gharbi, K. (2014). Quality control of next-generation sequencing data without a reference. *Front. Genet.* 5.

- Tunkelang, D. (1998). JIGGLE: Java Interactive Graph Layout Environment. In Graph Drawing, S.H. Whitesides, ed. (Springer Berlin Heidelberg), pp. 413–422.
- Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L.J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* *12*, R13.
- Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* *28*, 1248–1250.
- Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* *10*, 252–263.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. *Science* *291*, 1304–1351.
- Villalba, A., Coll, O., and Gebauer, F. (2011). Cytoplasmic polyadenylation and translational control. *Curr. Opin. Genet. Dev.* *21*, 452–457.
- van Vliet, A.H.M. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol. Lett.* *302*, 1–7.
- Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* *13*, 227–232.
- Wagner, S.D., and Berglund, J.A. (2014). Alternative pre-mRNA splicing. *Methods Mol. Biol. Clifton NJ* *1126*, 45–54.
- Walshaw, C. (2003). A Multilevel Algorithm for Force-Directed Graph-Drawing. (Springer-Verlag), pp. 171–182.
- Wang, X. (2016). Next-Generation Sequencing Data Analysis (CRC Press).
- Wang, Z., and Burge, C.B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* *14*, 802–813.
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* *456*, 470–476.
- Wang, J., Vesterlund, L., Kere, J., and Jiao, H. (2016). Identification of Novel Transcribed Regions in Zebrafish ( *Danio rerio* ) Using RNA-Sequencing. *PLOS ONE* *11*, e0160197.

- Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* *38*, e178.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* *10*, 57–63.
- Watson, J.D., and Crick, F.H.C. (1953). The Structure of Dna. *Cold Spring Harb. Symp. Quant. Biol.* *18*, 123–131.
- Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* *452*, 872–876.
- Will, C.L., and Lührmann, R. (2011). Spliceosome Structure and Function. *Cold Spring Harb. Perspect. Biol.* *3*, a003707.
- Wong, K.-L., But, P.P.-H., and Shaw, P.-C. (2013). Evaluation of seven DNA barcodes for differentiating closely related medicinal *Gentiana* species and their adulterants. *Chin. Med.* *8*, 16.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinforma. Oxf. Engl.* *30*, 1660–1666.
- Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* *347*, 1254806.
- Xiong, J., Lu, X., Zhou, Z., Chang, Y., Yuan, D., Tian, M., Zhou, Z., Wang, L., Fu, C., Orias, E., et al. (2012). Transcriptome Analysis of the Model Protozoan, *Tetrahymena thermophila*, Using Deep RNA Sequencing. *PLOS ONE* *7*, e30630.
- Xue, J., Schmidt, S.V., Sander, J., Draffehn, A., Krebs, W., Quester, I., De Nardo, D., Gohel, T.D., Emde, M., Schmidleithner, L., et al. (2014). Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation. *Immunity* *40*, 274–288.
- Yandell, M., and Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.* *13*, 329–342.
- Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. *Nucleic Acids Res.* *44*, D710–D716.
- Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004). Variation in alternative splicing across human tissues. *Genome Biol.* *5*, R74.

- Yon Rhee, S., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515.
- Young, P., Ebner, R., Weaver, Z., Strovel, J., Horrigan, S., Shea, M., Weigle, B., Rieger, M., Rick, J., and Cain, C. (2008). Cancer-linked genes as targets for chemotherapy.
- Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829.
- Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 7, 203–214.
- Zhao, S., and Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* 16.
- Zhao, C., Waalwijk, C., de Wit, P.J.G.M., Tang, D., and van der Lee, T. (2013). RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. *BMC Genomics* 14, 21.
- Zhuang, F., Fuchs, R.T., Sun, Z., Zheng, Y., and Robb, G.B. (2012). Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res.* 40, e54.



## Supplementary Materials

### Appendix – Supplementary Chapter 4

**Supplementary Table 4.1: Details list of 95 samples from human tissue atlas.** In this table shows the list of all 95 samples. By analysing sample-sample correlation, tissues that have the lowest correlation within tissues and have lowest read count were discarded for subsequence analysis.

| Tissue class | Tissue type     | Sample ID             | Ensembl gene ID | Read count | Sample removed | No. of gene expressed (FPKM > 1) |
|--------------|-----------------|-----------------------|-----------------|------------|----------------|----------------------------------|
| Brain        | Cerebral cortex | brain_3b.V102         | ERR315477       | 17,870,547 |                | 13,300                           |
|              | Cerebral cortex | brain_3c.V103         | ERR315455       | 15,948,377 |                | 13,594                           |
|              | Cerebral cortex | brain_a.V29           | ERR315432       | 12,040,330 |                | 13,397                           |
| Glandular    | Thyroid gland   | thyroid_5b.V197       | ERR315412       | 16,396,512 |                | 12,845                           |
|              | Thyroid gland   | thyroid_5a.V196       | ERR315358       | 12,190,400 | Yes            | 12,224                           |
|              | Thyroid gland   | thyroid_5d.V199       | ERR315397       | 9,648,231  |                | 12,909                           |
|              | Thyroid gland   | thyroid_5c.V198       | ERR315363       | 8,812,675  |                | 12,293                           |
|              | Adrenal gland   | adrenal_4d.V122       | ERR315417       | 6,911,906  |                | 13,288                           |
|              | Adrenal gland   | adrenal_4a.V119       | ERR315452       | 6,005,653  |                | 13,112                           |
|              | Adrenal gland   | adrenal_4c.V121       | ERR315450       | 5,826,516  |                | 13,041                           |
|              | Salivary gland  | salivarygland_6b.V239 | ERR315459       | 9,591,599  |                | 11,965                           |
|              | Salivary gland  | salivarygland_6a.V238 | ERR315382       | 9,375,978  |                | 12,681                           |

|             |                 |                            |             |            |            |        |        |
|-------------|-----------------|----------------------------|-------------|------------|------------|--------|--------|
|             | Salivary gland  | salivarygland_6c.V2<br>40  | ERR315418   | 5,938,808  |            | 11,929 |        |
| GI tract    | Esophagus       | esophagus_5c.V185          | ERR315398   | 13,330,622 |            | 12,883 |        |
|             | Esophagus       | esophagus_5b.V184          | ERR315411   | 11,747,261 |            | 13,012 |        |
|             | Esophagus       | esophagus_5a.V183          | ERR315489   | 9,988,919  |            | 13,436 |        |
|             | Stomach         | stomach_3a.V90             | ERR315379   | 13,080,365 |            | 13,410 |        |
|             | Stomach         | stomach_a.V18              | ERR315467   | 11,072,372 | Yes        | 13,144 |        |
|             | Stomach         | stomach_3b.V91             | ERR315485   | 8,534,846  |            | 12,929 |        |
|             | Duodenum        | duodenum_4b.V145           | ERR315461   | 5,318,405  |            | 13,238 |        |
|             | Duodenum        | duodenum_4c.V150           | ERR315442   | 4,617,568  |            | 13,144 |        |
|             | Small intestine | smallintestine_4c.V1<br>53 | ERR315381   | 6,857,948  |            | 13,611 |        |
|             | Small intestine | smallintestine_4b.V1<br>52 | ERR315408   | 6,196,966  |            | 13,278 |        |
|             | Small intestine | smallintestine_4d.V1<br>56 | ERR315409   | 5,635,651  |            | 13,325 |        |
|             | Small intestine | smallintestine_4a.V1<br>51 | ERR315344   | 4,999,121  | Yes        | 13,106 |        |
|             |                 | Colon                      | colon_b.V11 | ERR315357  | 27,738,457 |        | 13,413 |
|             |                 | Colon                      | colon_c.V14 | ERR315484  | 20,278,760 |        | 13,281 |
|             | Colon           | colon_d.V15                | ERR315400   | 17,425,564 |            | 12,760 |        |
|             | Colon           | colon_d.V10                | ERR315348   | 11,000,633 | Yes        | 13,202 |        |
|             | Colon           | colon_f.V22                | ERR315462   | 9,356,516  | Yes        | 12,932 |        |
| Circulatory | Heart muscle    | heart_5b.V195              | ERR315384   | 12,449,232 |            | 12,264 |        |
|             | Heart muscle    | heart_6a.V235              | ERR315356   | 12,099,974 |            | 12,178 |        |
|             | Heart muscle    | heart_6b.V237              | ERR315367   | 10,542,138 |            | 12,016 |        |
|             | Heart muscle    | heart_5a.V191              | ERR315328   | 9,362,692  | Yes        | 11,652 |        |

|             |                    |                    |            |            |        |        |
|-------------|--------------------|--------------------|------------|------------|--------|--------|
| Respiratory | Lung               | lung_3f.V81        | ERR315341  | 18,601,187 |        | 13,463 |
|             | Lung               | lung_3e.V80        | ERR315346  | 15,539,350 |        | 13,717 |
|             | Lung               | lung_4d.V133       | ERR315487  | 7,844,042  |        | 13,532 |
|             | Lung               | lung_4a.V130       | ERR315424  | 4,950,061  | Yes    | 13,479 |
|             | Lung               | lung_4b.V131       | ERR315444  | 3,011,499  | Yes    | 13,393 |
| Immune      | Appendix           | appendix_4c.V160   | ERR315481  | 5,288,209  |        | 13,531 |
|             | Appendix           | appendix_4a.V154   | ERR315465  | 5,102,962  |        | 13,277 |
|             | Appendix           | appendix_4b.V155   | ERR315366  | 5,062,707  |        | 13,230 |
|             | Spleen             | spleen_3d.V85      | ERR315448  | 16,957,358 |        | 13,149 |
|             | Spleen             | spleen_3a.V82      | ERR315338  | 16,211,143 |        | 13,004 |
|             | Spleen             | spleen_3c.V84      | ERR315473  | 14,577,236 |        | 12,982 |
|             | Spleen             | spleen_3b.V83      | ERR315405  | 7,004,952  | Yes    | 13,015 |
|             | Lymph node         | lymphnode_5a.V190  | ERR315493  | 9,987,765  |        | 12,350 |
|             | Lymph node         | lymphnode_5c.V193  | ERR315329  | 9,549,668  |        | 12,791 |
|             | Lymph node         | lymphnode_5b.V192  | ERR315426  | 7,930,958  | Yes    | 12,611 |
|             | Lymph node         | lymphnode_4a.V157  | ERR315371  | 6,314,443  |        | 12,837 |
|             | Lymph node         | lymphnode_4b.V164  | ERR315488  | 6,144,172  | Yes    | 12,528 |
|             | Bone marrow        | bonemarrow_6c.V250 | ERR315333  | 11,006,285 |        | 11,164 |
|             | Bone marrow        | bonemarrow_6b.V249 | ERR315406  | 10,964,022 |        | 11,253 |
|             | Bone marrow        | bonemarrow_6a.V248 | ERR315396  | 10,073,171 | Yes    | 10,959 |
| Bone marrow | bonemarrow_5a.V230 | ERR315469          | 8,885,586  |            | 10,659 |        |
| Liver       | liver_d.V111       | ERR315414          | 16,209,827 |            | 11,372 |        |
| Liver       | liver_a.V108       | ERR315463          | 11,368,772 |            | 11,554 |        |

|                 |                     |                        |                 |            |            |        |
|-----------------|---------------------|------------------------|-----------------|------------|------------|--------|
| Digestive organ | Liver               | liver_c.V110           | ERR315327       | 4,712,077  |            | 11,507 |
|                 | Gallbladder         | gallbladder_5a.V179    | ERR315474       | 11,892,467 |            | 13,875 |
|                 | Gallbladder         | gallbladder_5b.V182    | ERR315480       | 11,277,457 |            | 13,733 |
|                 | Gallbladder         | gallbladder_5c.V186    | ERR315360       | 9,850,735  |            | 13,451 |
|                 | Kidney              | kidney_b.V6            | ERR315443       | 16,980,044 |            | 12,865 |
|                 | Kidney              | kidney_d.V24           | ERR315383       | 11,972,258 |            | 12,860 |
|                 | Kidney              | kidney_a.V5            | ERR315494       | 11,516,286 |            | 13,609 |
|                 | Kidney              | kidney_c.V23           | ERR315468       | 7,717,404  | Yes        | 13,464 |
|                 | Pancreas            | pancreas_6a.V229       | ERR315466       | 9,844,424  |            | 10,562 |
|                 | Pancreas            | pancreas_6b.V232       | ERR315436       | 8,021,996  |            | 12,107 |
|                 | Urinary bladder     | urinarybladder_5c.V177 | ERR315355       | 11,291,669 |            | 13,693 |
|                 | Urinary bladder     | urinarybladder_5b.V176 | ERR315447       | 10,656,508 |            | 13,669 |
|                 | Female reproductive | Placenta               | placenta_3a.V76 | ERR315375  | 18,517,742 | Yes    |
| Placenta        |                     | placenta_6c.V224       | ERR315336       | 11,639,826 |            | 12,989 |
| Placenta        |                     | placenta_6a.V221       | ERR315374       | 10,632,558 |            | 12,952 |
| Placenta        |                     | placenta_6b.V223       | ERR315478       | 9,800,239  |            | 13,105 |
| Ovary           |                     | ovary_6b.V234          | ERR315458       | 10,813,749 |            | 12,294 |
| Ovary           |                     | ovary_6a.V233          | ERR315380       | 10,451,444 |            | 12,756 |
| Uterus          |                     | endometrium_5a.V200    | ERR315495       | 9,950,110  |            | 13,359 |
| Uterus          |                     | endometrium_4b.V165    | ERR315433       | 4,959,222  |            | 13,144 |
| Uterus          | endometrium_4a.V143 | ERR315438              | 3,866,425       |            | 13,341     |        |
| Prostate        | prostate_a.V12      | ERR315410              | 11,162,655      | Yes        | 13,651     |        |

|                   |                |                  |           |            |     |        |
|-------------------|----------------|------------------|-----------|------------|-----|--------|
|                   | Prostate       | prostate_4a.V127 | ERR315359 | 6,803,971  |     | 13,458 |
|                   | Prostate       | prostate_4b.V128 | ERR315407 | 6,013,636  |     | 13,364 |
|                   | Prostate       | prostate_4c.V129 | ERR315365 | 5,284,571  |     | 13,317 |
| Male reproductive | Testis         | testis_7f.V260   | ERR315492 | 28,722,714 |     | 15,143 |
|                   | Testis         | testis_7e.V259   | ERR315415 | 28,338,436 |     | 14,988 |
|                   | Testis         | testis_7c.V257   | ERR315391 | 22,389,166 |     | 14,977 |
|                   | Testis         | testis_7d.V258   | ERR315446 | 22,219,815 | Yes | 14,754 |
|                   | Testis         | testis_7b.V256   | ERR315456 | 20,609,541 | Yes | 14,878 |
|                   | Testis         | testis_7a.V255   | ERR315352 | 12,174,081 | Yes | 15,029 |
|                   | Testis         | testis_4a.V134   | ERR315350 | 5,183,071  | Yes | 15,002 |
| Adipose/skin      | Adipose tissue | fat_e.V20        | ERR315342 | 15,750,186 |     | 12,333 |
|                   | Adipose tissue | fat_a.V1         | ERR315332 | 13,469,445 |     | 12,576 |
|                   | Adipose tissue | fat_x1.V2        | ERR315431 | 2,830,606  |     | 12,661 |
|                   | Skin           | skin_5f.V247     | ERR315460 | 10,903,170 |     | 12,801 |
|                   | Skin           | skin_5e.V246     | ERR315401 | 9,874,830  |     | 12,745 |
|                   | Skin           | skin_6a.V245     | ERR315339 | 8,465,158  |     | 13,048 |

**Supplementary Table 4.2: Full list of differentially spliced events in human tissue atlas ranked by FDR value.** The output of rMATS was filtered out with  $FDR > 0.01$  and inclusion level differences  $|\Delta| > 0.5$ . The first five columns give the gene, gene description, chromosome, the location of exon start and end. The next column aims to include the FDR value of rMATS analysis. The next column is exon inclusion level difference. A negative number means more inclusion and a positive number more exclusion of the sequence in tissue comparison. The exon inclusion level difference is an absolute, rather than relative, change in the percentage of a specific splicing isoform in all mRNAs produced from the parent gene that follows the indicated splicing pattern. Event Types: 1) A3SS: alternative 3' splice site 2) A5SS: alternative 5' splice site 3) MXE: mutually exclusive exons 4) RI: retained intron and 5) SE: skipped exon. The last four columns give the sample examined and the expression level in FPKM value.

| Gene            | Description                        | Chr | Exon Start | Exon End  | FDR       | Inclusion Level Difference | AS Event | Sample 1 | Sample 2 | FPKM Sample 1 | FPKM Sample 2 |
|-----------------|------------------------------------|-----|------------|-----------|-----------|----------------------------|----------|----------|----------|---------------|---------------|
| <i>KLC1</i> *   | kinesin light chain 1              | 14  | 104145720  | 104153548 | 5.57E-308 | 0.547                      | SE       | Brain    | Heart    | 218.3         | 39.9          |
| <i>FUS</i>      | FUS RNA binding protein            | 16  | 31196259   | 31199678  | 1.22E-292 | 0.64                       | RI       | Brain    | Liver    | 129.2         | 44.9          |
| <i>TPM1</i> *   | tropomyosin 1 (alpha)              | 15  | 63353067   | 63354476  | 3.83E-272 | -0.771                     | MXE      | Heart    | Liver    | 6863.5        | 33.8          |
|                 |                                    |     | 63354774   | 63358292  | 1.32E-224 | 0.532                      | SE       | Heart    | Liver    | 6863.5        | 33.8          |
|                 |                                    |     | 63353067   | 63353987  | 6.12E-116 | -0.732                     | SE       | Heart    | Liver    | 6863.5        | 33.8          |
|                 |                                    |     | 63353067   | 63354476  | 1.48E-43  | -0.592                     | MXE      | Brain    | Liver    | 49.1          | 33.8          |
|                 |                                    |     | 63353396   | 63354476  | 4.73E-38  | 0.572                      | RI       | Heart    | Liver    | 6863.5        | 33.8          |
|                 |                                    |     | 63353396   | 63354476  | 1.93E-28  | 0.559                      | SE       | Brain    | Liver    | 49.1          | 33.8          |
|                 |                                    |     | 63353067   | 63353987  | 8.90E-28  | -0.587                     | SE       | Brain    | Liver    | 49.1          | 33.8          |
| <i>SORBS2</i> * | sorbin and SH3 domain containing 2 | 4   | 186551702  | 186567936 | 2.94E-245 | 0.566                      | SE       | Heart    | Liver    | 651.0         | 58.7          |
| <i>TPM3</i>     | tropomyosin 3                      | 1   | 154143124  | 154145454 | 1.23E-244 | -0.767                     | MXE      | Heart    | Liver    | 75.7          | 58.1          |
| <i>GUK1</i> *   | guanylate kinase 1                 | 1   | 228328018  | 228333325 | 2.88E-230 | 0.548                      | SE       | Heart    | Liver    | 112.3         | 54.3          |

|                |   |    |           |           |           |        |     |       |       |        |       |
|----------------|---|----|-----------|-----------|-----------|--------|-----|-------|-------|--------|-------|
|                |   |    | 228327982 | 228333768 | 2.36E-19  | 0.558  | SE  | Heart | Liver | 112.3  | 54.3  |
|                |   |    | 228328018 | 228333325 | 1.74E-15  | 0.511  | SE  | Brain | Liver | 122.4  | 54.3  |
| <i>APP</i>     | amyloid beta (A4) precursor protein   | 21 | 27354656  | 27394358  | 1.21E-218 | -0.873 | SE  | Brain | Liver | 564.9  | 133.0 |
| <i>PDLIM5</i>  | PDZ and LIM domain 5  | 4  | 95497093  | 95506888  | 2.64E-197 | 0.944  | SE  | Heart | Liver | 1081.3 | 49.1  |
| <i>SLC25A3</i> | solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3 | 12 | 98987756  | 98991813  | 1.27E-189 | -0.659 | MXE | Brain | Heart | 145.4  | 467.9 |
| <i>TMED2</i>   | transmembrane emp24 domain trafficking protein 2                              | 12 | 124071293 | 124074993 | 2.04E-184 | 0.568  | SE  | Heart | Liver | 64.0   | 115.3 |
| <i>CAMK2D</i>  | calcium/calmodulin-dependent protein kinase II delta                          | 4  | 114421618 | 114430831 | 8.31E-143 | -0.677 | MXE | Brain | Heart | 42.8   | 90.6  |
|                |   |    | 114372187 | 114378719 | 1.01E-130 | -0.619 | SE  | Brain | Heart | 42.8   | 90.6  |
|                |   |    | 114421618 | 114429424 | 6.75E-79  | 0.722  | SE  | Brain | Heart | 42.8   | 90.6  |
| <i>GNAS</i>    | GNAS complex locus  | 20 | 57470666  | 57478640  | 1.43E-135 | 0.564  | SE  | Brain | Liver | 723.1  | 196.4 |
| <i>CLTB</i>    | clathrin, light chain B   | 5  | 175819455 | 175824719 | 1.62E-123 | 0.66   | SE  | Brain | Heart | 44.6   | 87.3  |
| <i>NDRG4</i>   | NDRG family member 4  | 16 | 58528867  | 58537807  | 6.66E-111 | 0.584  | SE  | Brain | Heart | 306.6  | 204.1 |
| <i>SEC31A</i>  | SEC31 homolog A ( <i>S. cerevisiae</i> )                                      | 4  | 83778841  | 83784545  | 1.64E-101 | 0.587  | SE  | Heart | Liver | 49.7   | 46.8  |
| <i>UGP2</i>    | UDP-glucose pyrophosphorylase 2   | 2  | 64068087  | 64083567  | 2.98E-96  | -0.618 | SE  | Brain | Liver | 59.1   | 243.6 |
| <i>RBM3</i>    | RNA binding motif (RNP1, RRM) protein 3                                       | X  | 48433948  | 48434471  | 2.72E-84  | 0.74   | RI  | Brain | Liver | 64.3   | 37.2  |
|                |   |    | 48433948  | 48434471  | 3.58E-15  | 0.686  | RI  | Heart | Liver | 91.7   | 37.2  |
|                |   |    | 48433948  | 48434807  | 9.94E-15  | 0.572  | RI  | Heart | Liver | 91.7   | 37.2  |

|                 |  |    |           |           |          |        |      |       |       |       |       |
|-----------------|--|----|-----------|-----------|----------|--------|------|-------|-------|-------|-------|
| <i>ABLIM1</i>   | actin binding LIM protein 1                                | 10 | 116233637 | 116247775 | 6.38E-84 | -0.653 | SE   | Brain | Heart | 38.1  | 129.9 |
| <i>DST</i>      | dystonin   | 6  | 56328362  | 56330993  | 7.94E-81 | 0.523  | SE   | Brain | Liver | 64.1  | 36.4  |
|                 |  |    | 56328362  | 56330993  | 4.53E-45 | 0.701  | SE   | Heart | Liver | 46.7  | 36.4  |
|                 |  |    | 56393638  | 56394931  | 2.68E-25 | -0.563 | SE   | Brain | Heart | 64.1  | 46.7  |
| <i>KIAA1191</i> | KIAA1191   | 5  | 175782573 | 175788742 | 3.50E-79 | -0.514 | SE   | Brain | Heart | 59.3  | 46.6  |
| <i>CLTA</i>     | clathrin, light chain A                                    | 9  | 36204064  | 36210657  | 5.53E-78 | 0.849  | SE   | Brain | Heart | 81.0  | 37.7  |
| <i>ACTN4</i>    | actinin, alpha 4   | 19 | 39200034  | 39205201  | 7.03E-76 | -0.638 | MXE  | Brain | Liver | 93.9  | 75.8  |
| <i>TPD52L1</i>  | tumor protein D52-like 1                                   | 6  | 125574862 | 125584208 | 4.44E-74 | -0.756 | SE   | Heart | Liver | 71.4  | 33.8  |
|                 |  |    | 125574862 | 125584372 | 9.67E-17 | 0.664  | SE   | Brain | Heart | 43.4  | 71.4  |
|                 |  |    | 125574862 | 125584208 | 5.86E-14 | 0.778  | SE   | Brain | Heart | 43.4  | 71.4  |
| <i>DCAF6</i>    | DDB1 and CUL4 associated factor 6                          | 1  | 167973770 | 168007726 | 1.10E-72 | -0.687 | SE   | Brain | Heart | 33.3  | 71.1  |
| <i>MACF1</i>    | microtubule-actin crosslinking factor 1                    | 1  | 39715685  | 39720047  | 1.11E-70 | -0.748 | SE   | Brain | Heart | 41.6  | 33.1  |
| <i>QKI</i>      | QKI, KH domain containing, RNA binding                     | 6  | 163987752 | 163984751 | 4.64E-69 | 0.647  | A3SS | Brain | Heart | 199.4 | 140.3 |
| <i>ANK2</i>     | ankyrin 2, neuronal  | 4  | 114294514 | 114302672 | 5.77E-69 | -0.504 | SE   | Brain | Heart | 79.9  | 60.7  |
|                 |  |    | 114294472 | 114304888 | 5.01E-50 | -0.554 | SE   | Brain | Heart | 79.9  | 60.7  |
| <i>MFF</i>      | mitochondrial fission factor                               | 2  | 228205007 | 228212100 | 1.80E-65 | -0.543 | SE   | Brain | Heart | 44.8  | 30.5  |
|                 |  |    | 228205007 | 228220477 | 6.95E-34 | -0.571 | MXE  | Brain | Heart | 44.8  | 30.5  |
| <i>PKIG</i>     | protein kinase (cAMP-dependent, catalytic) inhibitor gamma | 20 | 43160425  | 43218507  | 1.62E-61 | -0.62  | SE   | Brain | Heart | 56.4  | 172.7 |
| <i>CDK5RAP3</i> | CDK5 regulatory subunit associated                         | 17 | 46050884  | 46051397  | 6.29E-61 | -0.573 | RI   | Brain | Heart | 49.6  | 31.1  |



|                |   |    |           |           |             |        |      |       |       |       |        |
|----------------|---|----|-----------|-----------|-------------|--------|------|-------|-------|-------|--------|
|                | protein 3   |    |           |           |             |        |      |       |       |       |        |
| <i>FXR1</i>    | fragile X mental retardation, autosomal homolog 1   | 3  | 180687945 | 180693192 | 1.55E-59    | -0.863 | SE   | Brain | Heart | 38.2  | 85.5   |
| <i>YPEL5</i>   | yippee-like 5 (Drosophila)                          | 2  | 30371110  | 30379658  | 4.96E-51    | -0.527 | SE   | Brain | Heart | 58.5  | 33.0   |
|                |   |    | 30371110  | 30379653  | 3.57E-54    | -0.577 | SE   | Brain | Heart | 58.5  | 33.0   |
| <i>SORBS1</i>  | sorbin and SH3 domain containing 1                  | 10 | 97081719  | 97099084  | 1.03E-51    | -0.569 | SE   | Brain | Heart | 37.6  | 174.2  |
|                |   |    | 97131082  | 97135813  | 1.84E-42    | -0.622 | SE   | Brain | Heart | 37.6  | 174.2  |
| <i>PPP2R5C</i> | protein phosphatase 2, regulatory subunit B', gamma | 14 | 102252354 | 102302769 | 2.45E-47    | -0.568 | SE   | Heart | Liver | 108.5 | 30.5   |
|                |   |    | 102252354 | 102302769 | 0.007130622 | -0.581 | SE   | Brain | Liver | 51.3  | 30.5   |
| <i>DTNA</i>    | dystrobrevin, alpha                                 | 18 | 32407556  | 32418135  | 1.12E-43    | 0.874  | SE   | Brain | Heart | 107.5 | 112.3  |
| <i>CAST</i>    | calpastatin   | 5  | 96058342  | 96063234  | 5.00E-42    | 0.617  | SE   | Heart | Liver | 175.4 | 37.0   |
| <i>TMBIM6</i>  | transmembrane BAX inhibitor motif containing 6      | 12 | 50135739  | 50146332  | 2.49E-38    | -0.568 | A5SS | Brain | Liver | 232.3 | 563.5  |
| <i>EWSR1</i>   | EWS RNA-binding protein 1                           | 22 | 29694722  | 29695270  | 7.63E-37    | 0.577  | RI   | Brain | Liver | 63.0  | 31.7   |
| <i>ANXA7</i>   | annexin A7  | 10 | 75148069  | 75156341  | 1.02E-30    | 0.559  | SE   | Brain | Liver | 46.5  | 47.0   |
| <i>SUN1</i>    | Sad1 and UNC84 domain containing 1                  | 7  | 882977    | 891119    | 1.05E-24    | -0.702 | SE   | Brain | Heart | 32.4  | 74.4   |
|                |   |    | 856916    | 872238    | 8.29E-05    | -0.538 | SE   | Brain | Heart | 32.4  | 74.4   |
| <i>APOC1</i>   | apolipoprotein C-I                                  | 19 | 45417503  | 45418206  | 3.54E-24    | -0.535 | SE   | Brain | Liver | 48.5  | 6823.0 |
| <i>MLIP</i>    | muscular LMNA-                                      | 6  | 54025164  | 54034370  | 5.28E-20    | 0.575  | SE   | Heart | Liver | 164.6 | 32.4   |

|               |   |    |           |           |             |        |      |       |       |       |       |
|---------------|---|----|-----------|-----------|-------------|--------|------|-------|-------|-------|-------|
|               | interacting protein                                   |    |           |           |             |        |      |       |       |       |       |
| <i>CIRBP</i>  | cold inducible RNA binding protein                    | 19 | 1271979   | 1274439   | 2.51E-15    | 0.599  | RI   | Heart | Liver | 173.2 | 83.8  |
| <i>NCAM1</i>  | neural cell adhesion molecule 1                       | 11 | 113105754 | 113126723 | 5.29E-14    | -0.686 | SE   | Brain | Heart | 205.2 | 82.9  |
| <i>KCNIP2</i> | Kv channel interacting protein 2                      | 10 | 103588831 | 103603677 | 4.03E-07    | 0.53   | SE   | Brain | Heart | 32.7  | 48.8  |
| <i>TACC1</i>  | transforming, acidic coiled-coil containing protein 1 | 8  | 38599868  | 38646337  | 8.63E-07    | -0.512 | MXE  | Brain | Heart | 44.6  | 36.7  |
| <i>PRR13</i>  | proline rich 13                                       | 12 | 53837462  | 53836517  | 4.80E-06    | -0.583 | A3SS | Brain | Liver | 45.5  | 51.6  |
| <i>PFDN5</i>  | prefoldin subunit 5                                   | 12 | 53689622  | 53691708  | 4.59E-05    | -0.5   | SE   | Brain | Liver | 154.5 | 172.7 |
| <i>PHLDB1</i> | pleckstrin homology-like domain, family B, member 1   | 11 | 118514789 | 118515409 | 0.00013919  | 0.61   | RI   | Brain | Heart | 60.8  | 44.0  |
| <i>CALM2</i>  | calmodulin 2 (phosphorylase kinase, delta)            | 2  | 47397872  | 47403650  | 0.000327377 | -0.755 | SE   | Heart | Liver | 217.9 | 159.6 |
| <i>MAN2C1</i> | mannosidase, alpha, class 2C, member 1                | 15 | 75659851  | 75660539  | 0.003159714 | 0.501  | A3SS | Brain | Heart | 59.6  | 51.5  |
| <i>HP1BP3</i> | heterochromatin protein 1, binding protein 3          | 1  | 21106304  | 21113124  | 0.02567602  | 0.5    | SE   | Brain | Heart | 67.6  | 42.6  |

## Supplementary Chapter 5

### Appendix A - Email invitation

Hello,

My name is Wan Fahmi and I am currently writing my PhD's thesis here at The Roslin Institute about network-based visualization and usability of user interfaces in NGS Graph Generator (<http://seq-graph.roslin.ed.ac.uk/>).

As a part of my thesis, I will be conducting a usability study on the NGS Graph Generator application. I am currently looking for people to take part in this usability study. Tim has told me that you are available for this.

#### *Prerequisites*

The prerequisite for participating is to have a basic knowledge of the BioLayout Express3D software. I am looking who have been involved with the splice variation/transcriptome/alternative splicing either using RNA-seq or microarray. You can try it yourself first and follow the tutorial section in NGS Graph Generator website.

#### *What will you be doing in a usability study?*

You will be asked to do several short tasks using the application and be asked to share your experience and perceptions of the application.

#### *How long is a session?*

30-60 min

#### *When and where?*

The plan is to do this usability test between April 24th to 29th during weekdays 10:00-17:00 and it will be done here at the Alexander Robson Building (PhD Thesis Writing's Room)

#### *Interested?*

Please reply to this email if you are interested with your name and when you are available for the study. The dates are not set in stone and are quite flexible and can be adjusted to your schedule. If you have any questions, please contact me at [wan.fahmi@roslin.ed.ac.uk](mailto:wan.fahmi@roslin.ed.ac.uk)

Thank you for reading,

Wan Fahmi

PhD Student

Systems Immunology Group, The Roslin Institute, University of Edinburgh