The Identification and Characterization of RNA-mediated Gene Fusions Across Primate Genomes

Ann M. Mc Cartney B.Sc. (Hons)

A thesis presented to Dublin City University for the Degree of
Doctor of Philosophy

Supervisors: Dr. Mary J. O'Connell & Dr. Tim Downing
School of Biotechnology
Dublin City University

November 2018

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:_____

(Candidate) ID No:_____

Date:_____

Table of Contents

Acknowledgements……...xi

Abbreviations……..xii

List of Figures ……xvi

List of Tables……...xxi

List of Code…….....xxv

Abstract ……xxvi

Thesis Aims……..xxvii

1. Chapter 1: Introduction……1

1.1. Molecular Evolutionary Theory……....2

1.1.1. Natural Selection, Variation and Drift ……...2

1.1.2. Pitfalls of Positive Selection Analyses……...5

1.1.3. Branch lengths and rates of change. ……..7

1.2. Mechanisms of new gene genesis in vertebrates……...8

1.2.1. The role of *de-novo* sequences in new gene formation across genomes…….....9

1.2.2. Gene duplication and its role in new gene generation in genomes……...10

1.2.3. New gene creation through retrotransposition mechanisms…....12

1.2.4. Exon/domain shuffling as a mechanism of new gene creation…13

1.2.5. The gene fusion/fission mechanism and it's role in new gene formation across genomes……16

1.3. Vertebrate Genome Architecture……23

1.3.1. A link between the dynamic nature of genome architecture and phenotype across vertebrates…….....25

1.3.2. Genome rearrangments and their impact on vertebrate evolution……..28

1.4. The significance of SD and phenotype diversity across vertebrate species……29

1.5. Transcription and methods of transcriptional control across vertebrates……..35

1.5.1. Mechanisms of eukaryotic transcription in eukaryotes……35
1.5.2. An overview of *cis* and *trans* factor involvement in transcriptional control……..48
1.5.2.1. Regulation of gene expression by transcription factors……..48
1.5.2.2. Splicing as a mechanism for transcriptional control…….....54
1.5.2.3. The role of histone modifications in transcriptional control…….59
1.5.2.4. Technological advancements in gene expression profiling……...61
1.5.2.5. Lineage-specific gene expression in vertebrates…….….67
1.5.2.6. The acquisition of expression profiles in new genes……68
1.6. The Mechanics of Translation …….……...70
1.6.1. A brief overview of translatome profiling technologies……...72
1.6.1.1. Polysomal Profiling…….…….72
1.6.1.2. Ribosomal Affinity Purification Techniques……....74
1.6.1.3. Ribo-tRNA sequencing.. …….75
1.6.1.4. Ribosome Profiling……...76
1.7. Application of network theory to understanding protein evolution……80
1.7.1. Graph theory and its use in biological data……..80
1.7.2. Identifying cliques and communities within graphs…….88
1.7.3. Sequence similarity networks in the identification of fusion genes…….89
1.7.4. FusedTriplets and MosaicFinder packages for fusion detection…….…..91
2. **Chapter 2: Identification and computational characterisation of RNA-mediated gene fusions across primate genomes…….....94**
2.1. Introduction……...……..95
2.2. Materials and Methods. …….…96

2.2.1. The identification of RMGF across primate and vertebrate genomes……...96
2.2.1.1. Primate dataset acquisition, cleaning and filtering……...96
2.2.1.2. Sequence similarity searches on dataset of primate protein coding genes…….…...99
2.2.1.3. Sequence similarity network generation to identify RMGFs……99
2.2.1.4. Validation of identified RMGF through pairwise alignment construction. ……100
2.2.1.5. An investigation of identified RMGF orthologs across vertebrate species……...102
2.2.2. To determine if RMGFs coincide with known human segmental duplication break points…….102
2.2.3. RMGF characterisation and comparison to non-fused protein coding genes……...105
2.2.3.1. A functional enrichment analysis across RMGFs and their parents……105
2.2.3.2. An investigation of human single nucleotide polymorphism and INDELs across human RMGFs……..105
2.2.3.3. An assessment of RMGF motif usage……..…….105
2.2.3.4. Codon usage analysis of human RMGF in comparison to non-fused human protein-coding genes…….106
2.2.3.5. An investigation of RMGF family location across the vertebrate tree……106
2.2.4. Rate heterogeneity and selective pressure heterogeneity analyses across candidate RMGF…….107
2.2.4.1. Selection, sequence acquisition and alignment of candidate RMGFs……..107
2.2.4.2. Phylogeny reconstruction of candidate RMGFs……110
2.2.4.3. Testing the selective pressures acting on RMGFs…….110
2.2.4.4. An assessment of the accuracy of the Ensembl Genome Browser's transcript annotation pipeline…….…..113
2.3. Results…….……..113

2.3.1. RMGFs detected in primate and vertebrate genomes……113
2.3.2. To determine if RMGF are overrepresented in regions of primate segmental duplication in comparison to non-fused protein coding genes……..…….120
2.3.4. A computational characterisation of RMGFs…….122
2.3.4.1.A Functional enrichment analysis across RMGFs and their parents……...122
2.3.4.2.An investigation SNPs and INDELs in human RMGFs…….127
2.3.4.3.A Motif Enrichment Analysis of RMGF……..127
2.3.4.4.Codon usage bias in RMGFs……128
2.3.4.5.RMGF family location across vertebrate species…….....131
2.3.5. Rate heterogeneity and selective pressure analyses across candidate RMGFs……...133
2.3.5.1.An investigation of evolutionary rates across RMGFs in comparison to non-fused protein coding genes of comparable length…..……..133
2.3.5.2.An investigation of the selective pressures acting on RMGFs…….139
2.3.5.2.1. Lineage-specific selective pressure heterogeneity……....139
2.3.5.2.2. Site-specific selective pressure variation of candidate RMGFs…….142
2.3.6. An assessment of the accuracy of the Ensembl Genome Browser's ortholog annotation pipeline……..144

2.4. Discussion…….146

3. Chapter 3: Transcriptomic and Translatomic Profiles of RNA-Mediated Gene Fusions……..149
3.1.Introduction…….150
3.2.Materials and Methods……152
3.2.1. An assessment of the RMGF transcriptomic profiles……152

3.2.1.1.Preparation and quality control of published RNAseq data (Brawand *et al*, 2011) …….152
3.2.1.2.Mapping of RNA sequence reads to reference genomes……....161
3.2.1.3.Differential Gene Expression Analysis……..……164
3.2.2. Wet-bench validation of RMGFs transcription profiles obtained through computational analysis…….....164
3.2.2.1.Quantitative RT-PCR to assess transcription of human RMGFs at 90 PI……164
3.2.2.2.RT-PCR analysis of two human RMGFs and their orthologs in gorilla and chimpanzee……170
3.2.3. An investigation of RMGFs with annotated alternative transcripts……..170
3.2.4. Uncovering translation profiles of detected RMGFs through ribosomal profiling…….171
3.3.Results…….180
3.3.1. Transcription profiles of RMGFs from analysis of RNA sequencing data……..……180
3.3.2. Alternative transcript frequency for human RMGFs and human ortholog containing primates……185
3.3.3. qRT-PCR analysis of 27 RMGFs using human as a representative of the Great Ape clade…….187
3.3.4. Differential expression analyses of primate and mouse RMGFs using the EdgeR package……..189
3.3.5. RT-PCR analysis of gorilla and chimpanzee tissue samples from the Barcelona Zoo……195
3.3.6. Analysis of ribosome profiles from limited datasets reveals signature of translation for small number of RMGFs……198

3.4. Discussion…….199

4. Chapter 4: Computational prediction of the regulation of expression in RMGFs……208
4.1.Introduction…….....209
4.2.Materials and Methods…….212

4.2.1. An investigation of activating and repressing signal profiles across RMGFs……...213
4.2.2. An analysis of splice factors across RMGFs……..215
4.2.2.1.A comparison of splice factor binding sites (SFBSs) across RMGFs and human non-fused protein coding genes…….....215
4.2.2.2.An assessment of splice factor transcriptional profiles across the ENCODE database (Harrow *et al.*, 2012) …….220
4.2.2.3.An assessment of SFBS frequency across the fusion breakpoint of RMGFs……...220
4.2.3. An analysis of histones across RMGFs …….....221
4.2.3.1.Data acquisition and histone abundance calculation across RMGFs ……221
4.2.3.2.Assessing the relationship between histone markers and splice factor usage in RMGFs…….....227
4.2.4. An assessment of annotated and experimentally validated transcription factor binding sites from the JASPER database across RMGFs……..……..……..……..……..……..……..…….....227
4.3.Results……..227
4.3.1. Computational characterisation of activator and repressor signals in human RMGFs across 8 tissues…….....227
4.3.2. Computational characterisation of splice factor binding site usage across RMGFs……...230
4.3.3. An investigation into the impact of SF presence on RMGF parent transcription profiles……..239
4.3.4. An exploration of histone binding sites across RMGFs….....250
4.3.4.1.An investigation of histone marker frequencies in RMGFs across a panel of human tissues…….....250
4.3.4.2.Linear regression analyses identify correlation between histone marker and specific splice binding site usage in RMGF…….……..……255
4.3.5. An assessment of TFBS usage and frequency comparison across RMGFs……..260
4.4.Discussion……....262

5. Chapter 5: An application of sequence similarity networks to understanding domain shuffling in vertebrates……...272
5.1.Introduction……..273
5.2.Materials and Methods…….....273
5.2.1. Creation of bipartite and unipartite networks from pFam-A domains...274
5.2.2. Domain co-occurrence network centrality…….....276
5.2.3. Comparison of degree distribution of the vertebrate gene network to a typical scale-free network…….....277
5.2.4. An investigation of domain usage across RMGFs and comparison to simulated dataset of human non-fused protein-coding genes…….277
5.2.5. A functional analysis of domains identified across RMGFs and human simulated data…….....280
5.3.Results……..283
5.3.1. Bipartite network of vertebrate protein coding genes and pFam domains……...283
5.3.2. Unipartite network creation in vertebrate protein coding gene dataset…….....285
5.3.3. Centrality of RMGF unipartite networks ……..287
5.3.4. An analysis of domains in RMGFs in comparison to non-fused human protein coding genes……...289
5.3.5. A GO term functional assessment of domains identified within RMGFs in comparison to human non-fused protein coding genes…..…….……...292
5.4.Discussion…….....295
6. **Chapter 6: Discussion and Conclusion**…….303
7. **Bibliography**…….31 3

Acknowledgements

David, Dad, Mam, Sonia, Jenny, Claire, Gerard, Mary, Tim, John, Damien and Fianna thank you for believing in me when I couldn't believe in myself and for lifting me up when I didn't have the strength.

This is for all of you.

Abbreviations List

A.A	Amino acid
Adh	Alcohol dehydrogenase
ALL	Acute lymphoblastic lymphoma
AR	Adaptive radiation model
ARM	AlkB-facilitated RNA methylation sequencing
AS	Alternative splicing
BAC	Bacterial artificial chromosome
BEB	Bayes empirical bayes
BS	Binding site
COS	Conservation of score
COS (WR)	Conservation of score (weighted rank) calculation
CHX	Cyclohexamide
CNL	Benign CML
CML	Chronic myeloid lymphoma
CMS	Clique minimal separators
CRM	cis-regulatory module
CT	Threshold cycle
CTD	C' terminal domain
DBD	DNA binding domain
DE	Differntial Expression
DMGF	DNA-mediated gene fusion
Dn	Non-Synonymous mutation
DNA-BP	DNA binding protein
DNAP	DNA polymerase
Ds	Synonymous mutation
E	Edge
EAC	Escape from adaptive conflict
eIF	Eukaryotic initiation factor
ESE	Exonic splice enhancer
ESS	Exon splice silencer
EST	Expressed sequence tagging

FDR	False discovery rate
G	Graph
GFP	Green fluorescent protein
GGI	Gene-gene interaction network
GO	Gene ontology
GTEX	Genotype-tissue expression database
GTFs	General Transcription Factors
HAR	Human accelerated region
HAT	Histone acetylase
HDAC	Histone deacetylase
HGT	Horizontal gene transfer
HMM	Hidden markov models
HPA	Human protein atlas database
IAD	Innovation, amplication and divergence model
IGS	Intergenic sequences
INDEL	Insertion deletion event
INR	Initiator
INSR	Insulin gene
ISE	Intron splice enhancer
ISS	Intron splice silencer
JTT	Jones-Taylor-Thornton model
Kb	Kilobases
LogCPM	Log counts per million
LogFC	Fold Change
LRT	Likelihood ratio test
MCMCMC	Metropolis-coupled Markov chain Monte Carlo
MMP	Maximal Mappable Prefix
MYA	Million years ago
N	Nodes
N_e	Effective population size
NEB	Naive empericel bayes
NMD	Non-sense mediated decay
NP	non-deterministic polynomial time problem

oc90	Otoconin 90
ORF	Open reading frame
p	Probability
PAP	Polyadenylate polymerase
PI	Percentage Identity
PIC	Promoter initiation complex
PIP5K1A	Phosphatidylinositol-4-phosphate 5-kinase
PSMD4	Proteasome 26S subunit, non-ATPase, 4
RBP	Ribosome binding protein
RMGF	RNA-mediated gene fusion
RNA	Ribonucleic Acid
RNAP	RNA polymerase
RSCU	Relative synonymous codon usage
RT	Retrotransposition
SAGE	Serial analysis of gene expression
SCPP	Secretory calcium binding phosphoprotein
SD	Segmental Duplication
SF	Splice factor
SFBS	Splice factor binding sites
SMRT	Small molecule real time
SNP	Single nucleotide polymorphism
SR	Serine Rich
SSN	Sequence similarity network
TAD	Trans-activating domain
TD	Tandem Duplication
TF	Transcription factor
TFBS	Transcription factor binding sites
TIC	Transcription initiation complex
TNNT	Troponin T
TRAP	Ribosomal affinity purification techniques
tRNA	Transfer RNA
TSS	Transcription start site
UTR	Untranslated region

V	Vertex
WGD	Whole Genome Duplication
WR	Weighted rank
ZMG	Zeromode wave guides
<Knn>	Nearest neighbour degree
ω	Omega

List of Figures

Figure 1.1: An illustration of exon/ domain shuffling……15
Figure. 1.2: Method of gene fission and subsequent sub-functionisation or neo-functionalisation……....18
Figure 1.3: Mechanisms of RNA-mediated and DNA-mediated gene fusion events.. ……20
Figure 1.4: Illustration of gene fusion creation through genomic rearrangments……...21
Figure 1.5: Duplication mechanisms behind large-scale genetic rearrangements……...27
Figure 1.6: Illustration of segmental duplication frequency across primate species.....32
Figure 1.7: Illustration of the mechanism behind creation of nested duplicon structure in genomes……..……34
Figure 1.8: An overview of transcription initiation in eukaryotic cells……39
Figure 1.9: Depiction of transcriptional elongation and 5' capping of transcript in eukaryotic cells……..41
Figure 1.10: RNAP1transcriptional termination in eukaryotes…….…….43
Figure 1.11: Illustration of the known mechanisms behind RNAP2 transcription termination in eukaryotes……..45
Figure 1.12: Torpedo transcriptional termination mechanism and polyA tail addition……;…….46
Figure 1.13: Transcription factor binding in eukaryotes……..50
Figure 1.14: Histone modifications and their role in the control of gene expression……52
Figure 1.15: The construction of *cis*-regulatory modules on eukaryotic promoters……53

Figure 1.16: Major and minor splice isoform expression profile characterisation……...56
Figure 1.17: An illustration of the *cis*-sequences required for accurate, controlled splicing in eukaryotes……..58
Figure 1.18: An illustration of first generation sequencing platforms……..62
Figure 1.19: A graphical representation of the 454 sequencing platform and the Ion torrent platform…….64
Figure 1.20: An illustration of Pacific Bioscience's third generation PacBio system……..66
Figure 1.21: A graphical representation of the ribosomal profiling mechanism……...78
Figure 1.22: An illustration of undirected and directed graphs……..……82
Figure 1.23: Calculating node degree in networks……..……83
Figure 1.24: Illustration of path length properties of graphs…….85
Figure 1.25: Sequence similarity networks based on global sequence similarity searches explained……...90
Figure 2.1: A phylogram illustrating selected high quality primate genomes for analysis……98
Figure 2.2: Alignment validation of SSN identified fusion genes……..101
Figure 2.3: A graphical representation of the methodology behind identifying RMGFs within regions of human segmental duplication…….104
Figure 2.4: Illustration of candidate RMGF selection process and sequence data acquisition…….109
Figure 2.5: The frequency of RMGF calculated across primates and mouse at 90%, 80%, 70%, and 50% identity thresholds…….116

Figure 2.6: The phylogenetic distribution of RMGF transcripts across a vertebrate database……..…….…….119
Figure 2.7: Location of RNMF parent gene orthologs across a selected high quality vertebrate database…….132
Figure 2.8: An assessment the Ensembl Genome Browser's orthology pipeline across human, chimpanzee and gorilla species…….....145
Figure 3.1: FASTqc quality assessment of Human RNAseq reads after adaptor trimming implementation……..…….154
Figure 3.2: Depiction of per base GC content analysis carried out by the FASTqc software package in chimpanzee RNA sequencing data…….…..155
Figure 3.3: FASTqc analysis of sequence length across chimpanzee RNA sequence reads……156
Figure 3.4: A sequence quality across chimpanzee RNA sequencing reads by FASTqc……..157
Figure 3.5: FASTqc analysis of chimpanzee RNA sequencing dataset post-trim for uncalled bases……...158
Figure 3.6: A post adaptor trim quality assessment of GC content per read of the chimpanzee dataset……159
Figure 3.7: Sequence duplication FASTqc result of the chimpanzee RNA sequence dataset post adaptor trim……160
Figure 3.8: Read counts taken after each step of the mapping protocol across each species transcriptome……163
Figure 3.9: FastQC relative enrichment quality check output of the Rooijer Dataset after adaptor trimming……...…172
Figure 3.10: An inspection of N content within the Rooijer Dataset after adaptor trimming by the FastQC package……...173
Figure 3.11: GC content quality check of the Rooijer Dataset after adaptor trimming……174
Figure 3.12: An overall quality score assessment across bases in reads from the Rooijer dataset……..……..175
Figure 3.13: FASTqc analysis result of GC distribution over all reads in the Rooijer dataset ……...176

Figure 3.14: Graphical display of the percentage of each nucleotide across each base in a read generated after adaptor trimming in the Rooijer dataset ……177
Figure 3.15: An analysis of the read length distribution within the Rooijer dataset.. …….178
Figure 3.16: An investigation of read frequency across all four ribosomal profiling datasets analysed…….179
Figure 3.17: RNA sequencing metadata analysis of RMGFs across a dataset of 6 primate species and mouse…….……182
Figure 3.18: Expression profiles for RMGF parents from RNAseq metadata analysis of 6 primate species and mouse across a panel of 6 tissues……..……...183
Figure 3.19: qRT-PCR results for 27 candidate human RMGFs identified at 90 PI across a panel of five human tissues……..188
Figure 3.20: Differential expression results of RMGFs across primate species and mouse…….192
Figure 3.21: RT-PCR polyacrylamide gel results of 2 human RMGFs and their corresponding orthologs in chimpanzee and gorilla……...197
Figure 4.1: An assessment of human RMGF activator and repressor signal profiles across 8 human tissues……..229
Figure 4.2: A comparison of SFmap COS (WR) scores calculated during an assessment of human RMGFs for experimentally validated SFBSs…….231
Figure 4.3: Expression profiles of Splice factors associated with SFBS motif's across a panel of 12 human tissues obtained from the ENCODE database……..234
Figure 4.4: Presence/absence of 16 SFBSs across 20 RMGFs and the tissue in which their corresponding splice factor is expressed at the highest level…….……...235
Figure 4.5: A weighted bipartite network constructed to analyse the relationship between the number of SFBSs and tissue expression profiles associated with splice factors across 18 RMGFs…….238
Figure 4.6: RMGF breakpoint analysis of SFBS results and location within their corresponding parents gene……241

Figure 4.7(a): Illustration of the gene expression profile for each RMGF parent……244
Figure 4.7 (b): Illustration of the gene expression profile for each RMGF parent…….245
Figure 4.8: An investigation of the distribution of SFBSs across human non-fused protein coding genes in comparison to human RMGF genes……..247
Figure 4.9: An illustration of the average number of histone binding sites in RMGFs across a panel of 9 human tissues……252
Figure 4.10: Histone marker frequency across RMGF transcripts across a panel of human tissues…….……...254
Figure 4.11: A network constructed based on statistically significant correlations between histone marker and splice site usage in a panel of human tissues across RMGFs……..257
Figure 4.11: An investigation of 15 gold standard TFBSs and their usage across human RMGFs……..261
Figure 5.1: Biological Process GO term hierarchical structure……281
Figure 5.2: GO term molecular function hierarchy……...282
Figure 5.3: A comparison of graph topology between vertebrate protein coding gene networks and randomly generated networks…….288
Figure 5.4: GO term analysis of identified domains across RMGFs and non-fused human protein coding genes.……293

List of Tables

Table 1.1: Genome Architecture comparison across vertebrates…….....26
Table 1.2: Inter and intra chromsomal SD clustering in across human chromosomes……...30
Table 1.3: A comparison of DNA forms in eukaryotic species……..37
Table 1.4: The 'Histone Code' and its effect on chromatin conformation…….…...60
Table 2.1: Likelihood ratio test calculations and number of parameters estimated for each model……..……..112
Table 2.2: MosaicFinder vs FusedTriplets comparative analysis of RMGF in humans at 90, 80, 70 and 50% identity threholds……...114
Table 2.3: MosaicFinder RMGF frequency results across our dataset ……117
Table 2.4: Human RMGF transcripts located in known regions of human segmental duplication…….121
Table 2.5: Functional enrichment results for human RMGF parents identified using at a soft threshold of 70% identity ……...123
Table 2.6: Functional enrichment results for parent RMGFs of fusions identified using MosaicFinder with a soft threshold of 80% identity……124
Table 2.7: Functional enrichment results for human RMGFs using MosaicFinder with 50 PI threshold……125
Table 2.8: Results of mouse RMGF parents functional enrichment analyses at a 70 PI threshold utilising the MosaicFinder software package……126
Table 2.9: GCUA Cumulative Codon Usage results for human RMGFs identified 90 PI threshold……...129
Table 2.10: Codon usage bias analyses of human non-fused protein coding gene transcripts……...130
Table 2.11: Candidate human RMGFs branch length analysis results……...…..134
Table 2.12: Branch length investigations of chimpanzee RMGFs compared to a panel of seleted vertebrates……135

Table 2.13: Investigation results of a branch length analysis ran across candidate mouse RMGFs compared to a dataset of vertebrates…….136
Table 2.14: An evolutionary rate analysis of candidate marmoset RMGFs compared to a panel of high quality vertebrate species genomes……..137
Table 2.15: Results of a RMGF branch length analysis in orangutan compared to a panel of vertebrates……..138

Table 2.16: Lineage-specific CodeML positive selection BEB results across candidate RMGF families……...……141

Table 2.17: Site-specific CodeML positive selection NEB results across candidate RMGF families……..143
Table 3.1: RT-PCR primer design for across-species expression analysis of 2 human RMGFs and their orthologs in gorilla and chimpanzee……...…….168
Table 3.2: Experimental design of RT-PCR cross-species human RMGF RNA-sequencing validation……..…..169
Table 3.3: The frequency of RMGFs expressed across 3 read mapping categories in a dataset of 5 primates and mouse……184
Table 3.4: An analysis of alternative transcripts across RMGFs identified at 90 PI and their orthologs across a panel of high quality vertebrate species…….……186
Table 3.5: Differential expression analysis of brain tissues between human RMGFs and their primate and mouse orthologs……194
Table 3.6: Expected expression level per RT-PCR experiment as determined by RNA sequencing metadata analysis (Section 3.3.1) …….196
Table 3.7: A summary of Ensembl's alternative isoform information present in the AltAsp database ……..….204
Table 4.1: Characteristics of mnemonics from the Roadmap Epigenomic Consortium (Roadmap Epigenomics Consortium *et al.*, 2015) ……214
Table 4.2: The panel of SFBSs represented in the SFmap software package (Paz *et al.*, 2010) ……..……..217
Table 4.3: Tissue samples utilised during the analysis of the h3k4me3 histone modification……222

Table 4.4: Tissue samples used during a h3k4me1 histone modification analysis of RMGFs……..223
Table 4.5: Tissue samples used during an analysis of the histone modification h3k9ac across RMGFs……...224
Table 4.6: Tissue panel obtained from the Roadmap Epigenomics Database for a h3k36me3 histone modification analysis……...225
Table 4.7: Tissue panel obtained from for a h3k27me3 histone modification analysis of RMGFs obtained from the Roadmap Epigenomics Database…….226
Table 4.8: The number of SFBSs identified in RMGFs as compared to human non-fused protein coding genes…….248
Table 4.9: A comparison of SFBS co-occurrence between human RMGFs and simulated human non-fused protein coding genes…….249
Table 4.10: An RNA expression profile analysis of SFBSs correlated with histone markers using Human Protein Atlas Data (Uhlén *et al.*, 2015) ……259
Table 5.1: Species names and database versions of the 30 vertebrate genomes used for domain usage assessment…….275
Table 5.2: Statistics for the 5 largest connected components in the bipartite network of vertebrate protein coding genes and pFam domains……284
Table 5.3: Top 3 largest unipartite networks based on bipartite human RMGFs and pFam domain database networks……....286
Table 5.4: pFam2go results of domains identified in domain usage analysis of human RMGFs……..290
Table 5.5: PANTHER investigation across RMGFs with identified pFam domains…….291
Table 5.6: Results of cellular component GO term analysis across the human simulated datasets…….294
Table 5.7: A statistical analysis of the current pFam Database (Finn *et al.*, 2014) across vertebrates…….296
Table 5.8: An analysis of the 20 largest identified domain families and promiscuous domains……..298

List of Code

Code_Box 1: Perl script to extract activator and repressor information for RMGFs from core15-state model mnemonic BED files…….215
Code_Box 2: Python script to generate human non-fused protein coding SFBS frequency distribution graphs prior to statistical significance testing……...219
Code_Box 3: "Dataset_randomgrabber.py" code used to generate 100 randomly sampled human non-fused protein coding genes……..…279

Abstract
Title: The identification and characterization of RNA-mediated gene fusions across primate genomes
Author: Ann Mc Cartney

New genes arise through gene duplication, retrotransposition, exon shuffling, gene fusion/fission, and *de-novo* genesis from noncoding DNA. Thus far, RNA-mediated gene fusion (RMGFs) has been shown to introduce functional novelty, divergent selective pressures, and divergent expression profiles when compared to unfused parent genes. However, the frequency and properties of these new genes remain largely unknown. Through the application of genome-wide networks to NGS data from Great Apes we aim to identify RMGFs, investigate their epigenetic profiles and analyse their potential mechanisms of generation, particularly through segmental duplications (SD). Subsequently, we aim to both computationally and experimentally investigate their expression and translation profiles and to characterise the cis-regulatory mechanisms behind RMGF transcription regulation. Finally, in order to enhance our understanding of the modular structure of RMGFs network based analyses were carried out to determine pFam domain usage patterns. 69 RMGFs were identified including 9 human-specific genes, their ancestry investigated across 32 high-quality vertebrate species and a significant enrichment in human SD shown. qRT-PCR and RNA-seq analyses reveal heterogeneous tissue expression with a bias towards testes specific expression in support of the 'out-of-testis' hypothesis. Moreover, *cis*-regulatory analyses of splice factor-binding sites, histone modifications and transcription factor binding sites support this profile of expression. Ribosomal profiling of human fibroblast cell lines has uncovered translation for 3 RMGFs and these genes remain functionally unannotated. RMGF domain usage pattern does not significantly differ from non-fused protein coding genes in human or indeed across vertebrates. Our genome-wide scan for RMGFs across primates has uncovered that their occurrence is frequent, they are enriched in regions of SD, their transcriptional output and cis-motifs support the 'out-of-testes' hypothesis and that their domain usage does not differ significantly to that of non-fused genes.

Thesis Aims

(1) To apply sequence similarity networks to identify RNA-mediated gene fusions across primate genomes (Chapter 2).

We wished to develop a streamlined pipeline using sequence similarity networks (SSNs) to generate a panel of putative RNA-mediated gene fusions across a dataset of 6 primate species and mouse. From here we wished to determine their phylogenetic distribution and the potential role for segmental duplication in driving the evolution of fused genes at breakpoints.

(2) To determine the transcriptional and translational profiles of RNA-mediated gene fusions across primate genomes (Chapter 3).

We wished to investigate if the identified RNA-mediated gene fusions had evidence for expression using both computational (RNAsequencing) and wet bench (RT-PCR and RT-qPCR) methods. Finally we wanted to know if there is evidence to suggest that these gene fusions produce viable protein products.

(3) To predict the potential role of *cis*-regulatory elements in controlling the expression of RMGFs (Chapter 4).

We wished to investigate which *cis*-elements had the potential to control transcription in the RNA-mediated gene fusions and we wished to assess how this profile compared to the characteristic pattern observed for non-fused protein coding genes across the human reference genome. Here we focussed on four regulatory elements: chromatin remodelling, splice factors, histone modifications and transcription factors.

(4) To determine how domain usage in RNA-mediated gene fusions compares to other non-fused vertebrate protein coding genes (Chapter 5)

We wished to assess if there is a different profile of fomain usage in RMGFs as compared to non-fused genes. We assessed the domain usage patterns (e.g. domain abundance and domain co-occurrence) in RMGFs differ to other protein coding genes from 30 high quality vertebrate genomes in an alternative application of the sequence similarity networks.

Chapter 1: Introduction

1.1) Molecular Evolutionary Theory

1.1.1) Natural Selection, Variation and Drift

In 1973, an essay written by Dobzhansky stated that "Nothing in Biology Makes sense Except in the Light of Evolution" (Dobzhansky, 1973). Evolutionary theory is based on two processes, namely Darwinian natural selection and non-adaptive processes such as mutational variation, recombination, biased gene conversion and genetic drift.

The theory of evolution by natural selection simply states that traits more likely to increase reproductive success are retained whilst traits with a negative effect are removed (Darwin, 1859). The modern synthesis saw the incorporation of molecular data into evolutionary theory and it is now understood that natural selection is driven by mutation. Mutations can produce deleterious effects and be removed by purifying selection, or they can produce advantageous effects and be retained by positive selection (Loughran *et al.*, 2012). They can either be silent in nature or synonymous (Ds) having no impact on the amino acid produced or change the amino acid produced, non-synonymous (Dn). The primary method for analysing selective pressures between species is by calculating the ratio of replacement: Dn/Ds to determine an ω value. This value has 3 potential outcomes: i) $\omega>1$ is indicative of positive selection, ii) $\omega<1$ is indicaftive of purifying selection and iii) is indicative of purifying selection. Positively selected alleles are commonly associated with the "hitch-hiking" effect whereby neutral, near-neutral or deleterious alleles linked with the advantageous allele increase in frequency (Maynard Smith and Haigh, 2008), (Sabeti *et al.*, 2006). This can lead to a reduction in variation around the positively selected site – a selective sweep (Andolfatto, 2001). Mutations acting on one allele can also have an advantageous effect on a heterozygote but a deleterious effect on homozygote and thus balancing selection is required to control an intermediate allele frequency (Simonsen, Churchill and Aquadro, 1995). Darwinian selection theory explains the evolution of mutations that confer a fitness effect on the organism however selection alone cannot explain genome evolution in isolation.

Many mutations have very little or no fitness effect i.e. neutral or near neutral, on an organism and in these situations Darwinian selection in isolation is insufficient in modelling evolution with Micheal Lynch deeming that such religious adherence to an adaptionist paradigm as being devoid of intellectual merit (Lynch, 2007). Evolutionary biologist's usage of natural selection as an omnipotent force to explain all of evolutionary theory had been recognised long before Lynch with Gould and Lewontin's labelling it the 'Panlgossian paradigm' in 1979 (Gould and Lewontin, 1979). With this in mind, non-adaptive processes require careful consideration perhaps even moreso than that of adaptive selection.

The importance of non-adaptive processes on evolution was initially proposed by Motoo Kimura whom recognised that the majority of evolutionary change within genomes was in fact driven by neutral or nearly neutral mutations. His hypothesis was based on the fact that mutations within protein-coding genes with less dramatic implications have a higher probability of being retained, that for the most part synonymous mutations occur at a higher than non-synonymous mutations and that non-coding regions evolve at a high rate similar to that of synonymous mutatations.

Neutral or nearly neutral alleles follow random genetic drift devoid of directional control ultimately leading to the fixation of some alleles and the complete removal of others (Kimura, 1989). Due to this fact some advantageous alleles are often lost and deleterious alleles are retained throughout a finite population (Ohta, 1992). Genetic drift is impacted by genomic processes such as biased gene conversion (Glémin *et al.*, 2015), the preferential inclusion of one allele over another during meiosis crossover as well as recombination (Barton and Otto, 2005). In order to understand the fixation rate of both neutral and nearly neutral mutations (Ohta, 1992) within genomes of finite diploid populations with an effective population size N_e, Ohta expanded upon Kimura's original theory (Kimura, 1989) to incorporate population genetics.

Equation 1:

$$P_x = \frac{1}{2N_e}$$

With this it became evident that the rate of fixation of neutral or near neutral mutations increases as N_e decreases, illustrating an inversely proportionate relationship.

Population genetics is crucial in the understanding of modern evolutionary biology with the previous statement by Dobzhansky (Dobzhansky, 1973) now updated to "Nothing in Evolution Makes Sense Except in Light of Population Genetics" (Lynch, 2007). Neutral and near neutral evolution have now become the basis of all selection based tests providing information on processes involved in genome functioning and has been shown to contribute to a species phenotypic and adaptive evolution (Zhang, 2018).

However, the simplistic nature of categorising mutations into advantageous, deleterious or neutral/ near-neutral is no longer sufficient. The effect of each mutation can be calculated by determining each mutation's fitness effect (Selection Co-efficient) and effective population size (Ne_s). When $Ne_S < 1$ the mutation is said to be under drift. It has become clear that phenotypes and genetic novelties are roots of non-adaptive processes thus understanding the evolution of these sequences is of vital importance when trying to understand the evolution of gene/genomic architecture, structure, developmental pathways, gene modularity and general organism complexity and evolvability (Lynch, 2007).

Selective pressures are assessed by comparing homologous sequences and can be measured for both protein coding (Webb *et al.*, 2015) and non-protein coding regions (McLean *et al.*, 2011). Although this thesis deals only with protein coding regions the importance of non-protein coding sequences and their effect on phenotype can not be underestimated with many cases highlighting its importance existing in the literature (McLean *et al.*, 2011; Enard, 2015; Franchini and Pollard, 2015). For example, in 2015 an enhancer sequence located in the human accelerated region (HAR) 5 was found to bind with the

promoter of the Fzd8 in human neural progenitor cells. This binding increased cell cycle rate within these cells resulting in a phenotype consequence shown by transgenic mice having a 12% increase in cortical surface area, more mid-layer neurons and a lateral expansion of the brain ventricular zone. Interestingly, the level of expression of the human HAR5 region in transgenic mice was 10-30 fold higher than it's syntenic region in chimpanzee and is therefore thought to contribute to the increased brain size of human (3 times larger) specifically (Boyd *et al.*, 2015).

Protein coding sequences are themselves or have regions that are highly conserved, sometimes across large evolutionary timescales e.g. (Consortium, 2002; Kellis *et al.*, 2003; Ureta-Vidal, Ettwiller and Birney, 2003; Richards *et al.*, 2005; Villar, Flicek and Odom, 2014; Jayaswal *et al.*, 2017) and mutations are rarely seen in these regions. This is due to the strong functional constraint on proteins (Fay and Wu, 2003). Therefore, the mutational space in protein-coding sequences is not as flexible as other regions in the genome (Margulies *et al.*, 2007).

Interspecies comparisons of positively selected sites are used as markers of functional discordance between species as they adapt to their surrounding environment (Tennessen, 2008). The reliance on this assumption has been extensively debated and it is important that follow-up analysis through rational mutagenesis and ancestral reconstruction be carried out (Loughran *et al.*, 2012), (Baldwin *et al.*, 2014). Previously our research group has performed the validation of positive selective forces acting on the myeloperoxidase enzyme revealing that positively selected residues resulted in the emergence of an novel activity, chlorination, in that enzyme (Loughran *et al.*, 2012). It remains one of a small number of examples in the literature. Although the mechanisms of validating predicted selective pressures are lacking, computational predictive software algorithms are increasing in accuracy, robustness and speed.

1.1.2) Pitfalls of Positive Selection Analyses

Each of the individual methods for measuring selective pressure variation have their own caveats, assumptions and pitfalls. However, the sequencing information currently available presents a major challenge as all of these algorithms and approaches rely on accurate sequence data. Therefore said data has to be high quality and high coverage.

Alignment error is one of the most common causes of false positive results in selective pressure analysis (Schneider *et al.*, 2010). To reduce alignment error it is important to test multiple alignment software packages and determine for each individual protein family what the best package is for that family in terms of significance of the resultant alignment (Redelings, 2014). Currently there are many sophisticated alignment software packages available (Edgar, 2004; Larkin *et al.*, 2007; Löytynoja and Goldman, 2010) there are also many packages available to assess the best alignment package for your input data, these include AQUA (Muller *et al.*, 2010), NorMD (Thompson *et al.*, 2001) and MetAl (Blackburne and Whelan, 2012).

Recombination is another factor to consider when carrying out a selective pressure analysis. Recombination affects the frequency and combination of alleles in a genome (Posada and Crandall, 2002), as well as altering codon usage in recombined sequences (Marais, Mouchiroud and Duret, 2001). Reports have in fact uncovered that a high rate of recombination within a sequence can consequently yield false positive calls (Anisimova, Nielsen and Yang, 2003). Not only this but signatures of biased gene conversion have been associated with recombination (Katzman *et al.*, 2011). Biased gene conversion is a neutral evolutionary process whereby GC content is elevated due to mismatch repair machinery favoring GC content at recombination breakpoints (Galtier and Duret, 2007). This bias in GC content can again affect ω and result in false positive results (Ratnakumar *et al.*, 2010).

Lastly exonic splice enhancers (ESEs) which are located close to exon boundaries are subject to purifying selection thus yeilding a decreased Ds value in these locations. It is currently unknown whether ESEs have any effect on

branch-site models. But it is known that low Ds values as opposed to increased Dn values in these regions can falsely give a Dn/Ds ratio indicative of positive selection (Parmley, Chamary and Hurst, 2006; Cáceres and Hurst, 2013). These limitations show that any claim of positive selection and functional shift in a protein or in a specific lineage could be as a result of a large number of data and algorithmic biases that require careful and expert navigation (Hurst and Pál, 2001).

1.1.3) Branch lengths and rates of change

Although informative, selective pressure in isolation does not provide the adequate information to completely understand biological diversity both within and between sequences. An indepth analysis of the pace of evolutionary change or more simply how fast or slow a given sequence is evolving provides an additional, vital layer of information contributing to our understanding of biological diversity. The pace, or rate of change can be assessed through phylogenomic based branch length calculation namely supertree generation (Wilkinson *et al.*, 2004), using genomic-scale data or by superalignment (Livak and Schmittgen, 2001) based on a concatonated multiple sequence alignment of sequences of interest in order to form a supermatrix. Both methods are computationally intense, particularly supertree analyses due to the size of the data under investigation. Here, a focus will be placed on superalignment phylogenomic reconstruction methods.

Branch length analyses have been used to answer many questions of molecular data for instance it has been used as a tool to investigate speciation events, to reconsiliate gene trees from species trees and to investigate the relative rates of evolution (Livak and Schmittgen, 2001). Unsurprisingly, due to the highly fluidic nature of genomic sequence data, the rate of evolution is not homogeneous across genomes, rate heterogeneity is affected by many factors including, mutation (Li, Tanimura and Sharp, 1987; Romiguier *et al.*, 2010) and biased gene conversion/GC content (Galtier and Duret, 2007). In order to account for the heterogeneic nature of genomes only high quality data can be used as well as sophisticated phylogenetic reconstruction programs e.g. P4 (Foster, 2004) and

Phylobayes (Lartillot *et al.*, 2013). Reconstruction using superalignment can be carried out through nucleotide or amino acid (A.A) alignment (Simmons, Ochoterena and Freudenstein, 2002). Nucleotide data can be directly assessed as only 4 character states are analysed, i.e. A, G, C and T. However A.As require a pre-processing step of converting each individual A.A to its corresponding Dayhoff category (a set of categories based on physio-chemical properties) (Hrdy *et al.*, 2004). This simplification reduces character number and computational load but information load is compromised. After pre-processing steps a phylogenetic tree is obtained, through phylogenetic estimation or fixed topology usage, along with a model of evolution to infer rates of change (Hall, 2013; Lartillot *et al.*, 2013). It is of huge importance that only a model that fits your data accurately while using the least amount of parameters, is selected. This requires carrying out posterior probability simulations on randomised datasets (Huelsenbeck and Rannala, 2004; Baele and Lemey, 2013). The rate of change can be extrapolated from composition vectors and rate matrices.

Phylogenomic reconstruction algorithms have limitations, particularly with erased evolutionary signal caused by mutational saturation at specific sites (Moreira and Philippe, 2000). It is also important that prior to carrying out phylogenomic methods both nucleotide and A.A alignments are considered as ignoring one method over another ignores the idiosyncracy of the data. Investigating selective pressures and comparing genomes through comparative phylogenomics is interesting especially when considering sequences that are dissimilar to that of "normal" protein-coding genes e.g. new genes.

1.2) Mechanisms of new gene genesis in vertebrates

Through positive selection acting on pre-existing protein coding regions, new genes with new functions can potentially be generated and there are many potential mechanisms behind their synthesis. However, associating a single genetic locus to a particular phenotype has proven a very slow process, and while there are a number of individual case studies that are well elucidated (Lamichhaney *et al.*, 2015) the complexity of *metazoan* genomes has meant that linking genotype to phenotype is not a straightforward process. Indeed, even to

identify corresponding genes between species of jawed vertebrates is challenging due to multiple whole genome duplication events – as outlined in the 2R hypothesis (Ohno, 1970).

At present only a handful of the morphological disparities that exist between human and our closest relatives have a genetic component fully elucidated (Varki and Altheide, 2005). Examples of genes with known direct impact on phenotype include the *FOXP2* gene that is associated with language acquisition, the *MYH16* gene that plays a role in mastication muscle usage, and the *HACNS1* gene that is associated with limb and digit specialisations (O'Bleness *et al.*, 2012). Improvements in sequencing technologies has resulted in a recent and remarkable growth in the volume of data available for comparative genomics (Alföldi and Lindblad-Toh, 2013). The accessability, ease of use, and improvement in read quality/length and cost, has meant that the number of questions that can be both asked and answered of any particular genome or population is increasing (Kaessmann, 2010; Buermans and den Dunnen, 2014) and with that an increased frequency of genome-phenome linkage is becoming more realistic (Rogers and Gibbs, 2014). Major questions that can now potentially be resolved include: how does novelty emerge, and how do novel proteins impact on the function and phenotype of an organism?

Genetic novelty through the generation of 'new genes' of novel function is a key contributor to evolutionary innovation within species (Ventura *et al.*, 2012). New genes are defined as those that have been created in a relatively recent evolutionary timescale (Long *et al.*, 2003). There are 5 main mechanisms known to drive the generation of new genes in genomes: *de-novo* genesis from non-coding DNA (Section 1.2.1) (Knowles and McLysaght, 2009); gene duplication (Section 1.2.2) (Ohno, 1970); retrotransposition (Section 1.2.3) (Boeke and Chapman, 1991); exon/domain shuffling (Section 1.2.4) (Kolkman and Stemmer, 2001), and gene fission/fusion events (Section 1.2.5) (Long, 2000), (Kaessmann, 2010). The five mechanisms are outlined below:

1.2.1) The role of *de-novo* sequences in new gene formation across genomes

De-novo genesis from non-coding DNA occurs when a non-coding DNA segment through spontaneous mutation generates either a promoter element or novel splice site allowing novel transcriptional activity (Knowles and McLysaght, 2009). Very few cases of *de-novo* genesis of non-coding DNA have been found across the phylogenetic tree of life with only a few isolated instances in yeast and *drosophila sp*. being identified (Levine *et al.*, 2006; Cai *et al.*, 2008; Schlötterer, 2015). More recently, three cases of spontaneous acquisition of transcriptional machinery by previously non-coding DNA were identified in human, these include: *CLLU1*, *C22orf45*, and *DNAH10OS* (Knowles and McLysaght, 2009) and are prime examples of accurate gene predictions using computational evolutionary biology approaches alone, *i.e.* transcriptional profiling or functional assays were not utilised during detection.

De-novo genes have a high death rate, however some survive and acquire functionality (Schlötterer, 2015). Function acquisition and fixation of a gene within a genome requires gaining regulatory machinery as well as integration into a pre-existing gene network (Schlötterer, 2015). Experiments comparing the intensity of purifying selection on *de-novo* generated genes against both coding and non-coding genes revealed that although the intensity of purifying selection is not as strong on *de-novo* generated genes in comparison to coding genes it is significantly higher than that of non-coding sequences (Schlötterer, 2015).

1.2.2) Gene duplication and its role in new gene generation in genomes

Gene duplication, the most predominant source of new gene acquisition, occurs when a DNA segment containing one or more genes is duplicated into a novel genetic environment (Ohno, 1970). Duplications are caused by one of three known mechanisms; unequal crossing over during meiosis, ectopic chromosomal duplication or by retrotransposition (Ohno, 1970). Ohno's work in the 1960s highlighted the impact of whole genome duplication on genome architecture and since our appreciation across vertebrates has only increased (Ohno, 1970). After a duplication occurs in a genome the new duplicate has the opportunity to acquire mutations which can either be lost or fixed - natural selection (Darwin, 1968). If fixed, the gene can undergo post fixation modifications leading to new

functions (neofunctionalization) e.g. *jingwei* (Long and Langley, 1993). They can also split the original gene's function between the original gene and the newly duplicated gene (sub-functionalisation) e.g *SIR3* and *ORC1* in *S.cerevisae* (Hickman and Rusche, 2010). Paralogous gene copies have been found to arise in eukaryotes at a rate of 0.01 paralogs per gene per million years (Lynch and Conery, 2003) and therefore many gene families of varying sizes exist across the tree of *metazoa* (De Grassi, Lanave and Saccone, 2008).

To explain how duplicates survive long enough to be acted upon by selective pressures there have been three models proposed (Long *et al.*, 2003). The first model is known as the adaptive radiation (AR) model. In this model it is proposed that a gene duplicate is positively selected for due to increased dosage having beneficial consequences on the organism (Gavrilets and Vose, 2005). The IAD (innovation, amplification, and divergence) model proposes that the original gene copy acquires a secondary advantageous function resulting in positive selection on duplicate copies to promote increase dosage of the secondary function (Bergthorsson, Andersson and Roth, 2007). Finally, the EAC model (escape from adaptive conflict) states that after the original gene acquires a secondary function it undergoes rapid adaptive selection in order to improve this function prior to duplication with the opportunity for further adaptive processes to occur post duplication (Deng *et al.*, 2010). Unsurprisingly, investigations also revealed that partial gene duplications are under stronger purifying selection than whole gene duplicates (Cardoso-Moreira *et al.*, 2016).

It has been difficult to investigate intra and inter species orthologs as the sheer abundance makes it difficult to extrapolate divergence times accurately. However, analyses have shown positively selected duplications between the human, macaque, mouse and rat genomes (Han *et al.*, 2009) with duplicates re-inserting onto a different chromosome having a higher probability of being under strong selective pressures (Han *et al.*, 2009). Finally, a large majority of positively selected duplicate families in human have increased copy numbers across all primates (O'Toole, Hurst and McLysaght, 2018) suggesting that

duplication has played a key role in genomic and perhaps functional complexity ascertainment across primate species.

1.2.3) New gene synthesis through retrotransposition mechanisms

Retrotransposition occurs when a gene is transcribed and reintegrated back into the genome at a novel location. Integrase is responsible for creating the initial DNA cleavage site while a reverse transcriptase converts, or reverse transcribes, the RNA to a DNA transcript facilitating it's integration into an alternative genetic locus (Boeke and Chapman, 1991). Both DNA and RNA-mediated retrotranspositions (RTs) have had a prominent impact on genomic structure, generating more than 40% of the human genome (Callinan, Batzer and Callinan, 2006) and have the potential to contain either full or partial gene fragments (Morgante *et al.*, 2005). Recent RT events are more likely to contain whole gene segments while more ancient RT events are more likely to consist of only partial fragments, as they have had more time to acquire sequence acquisition/loss (Cordaux and Batzer, 2009). RTs are usually transported around the genome by *L1* (Sassaman *et al.*, 1997) or *Alu* elements (Deininger, 2011) however, they do not contain a distribution pattern similar to either element e.g. the reintegration of new RTs is more likely to occur into transcriptionally inactive heterochromatic regions of the genome further supporting RTs biased distribution patterns within genomes (Wei *et al.*, 2013). This suggests that selection is constraining RTs distribution pattern (Graham and Boissinot, 2006).

Some models have suggested that this biased distribution and reintegration is caused by negative selective pressures placing new genes in regions of transcriptional inactivity in order to protect the host from any possible deleterious effects (Munoz-Lopez and Garcia-Perez, 2010); increased likelihood of ectopic recombination due to genomic instability caused by RT clustering (Crichton *et al.*, 2014), disruption of genetic regulation of essential genes (Mita and Boeke, 2016), or disrupting overall cell physiology by expression of new genes (Mita and Boeke, 2016). A second model suggests that RTs within heterochromatic regions can undergo positive selection (Pereira, 2004) with

studies of mouse RTs have identified a relaxation of selective pressures immediately after duplication potentially increasing mutation accumulation and promoting positive selection across newly generated RTs (Ho-Huu *et al.*, 2012).

1.2.4) Exon/domain shuffling as a mechanism of new gene creation

Domains are distinct small substructures within proteins (~150 residues in length) containing a hydrophobic core and have a distinct function or structure within a protein. Very similar domains can be found across genes of very different functionalities (Sowdhamini, Rufino and Blundell, 1996). Exon/domain shuffling occurs when a segment of coding DNA is duplicated and either partially or completely re-introduced into a new coding genetic environment resulting in a novel gene (Kolkman and Stemmer, 2001) (Figure 1.1). The transmission of domains/exons across genomes is aided by an increased level of regions with a high probability of recombination ('recombination hotspots') within intronic sequences that facilitate the integration of novel DNA segments into regions of transcriptional activity. Similarly to gene duplication exon/domain segments relocate across genomes predominantly through unequal crossover during meiosis (Kolkman and Stemmer, 2001) as well as through exon shuffling. The utilisation of exons allows the cleavage of domains from nucleotide sequences at exon boundaries, manipulating pre-existing cell machinery (Patthy, 2003; Keren, Lev-Maor and Ast, 2010). It was initially thought that the reintegration of domains into novel protein-coding environments would result in a phase shift rendering these novelties untranslated (Keren, Lev-Maor and Ast, 2010). However, it transpires that an excess of phase combinations exist at exon boundaries in multicellular eukaryotes that accommodate for this domain reintegration into a new protein-coding sequence environment. The overall result is an increase in the emergence of novel proteins and potentially novel functions in multicellular eukaryotes (Figure 1.1).

Bursts of domain shuffling have been correlated with instances of biological innovation (Patthy, 1999; Rajalingam, Parham and Abi-Rached, 2004) and modular proteins have been shown to contribute to the radiation of multicellular organisms through the expansion of extracellular matrix proteins and membrane

associated proteins, such as cell surface receptors and kinases (Babushok *et al.*, 2007). Domain shuffling has also played a role in hominoid evolution, for example the PIPSL gene created via an L1 retrotransposition of the PIP5K1A (phosphatidylinositol-4-phosphate 5-kinase) gene into chromosome 10 adjacent to the PSMD4 gene (proteasome 26S subunit, non-ATPase, 4) creating the PIPSL gene. Through positive selective pressure the PIPSL gene is fixed throughout all *hominoid* genomes and has phosphotase kinase activity (Babushok *et al*, 2007).

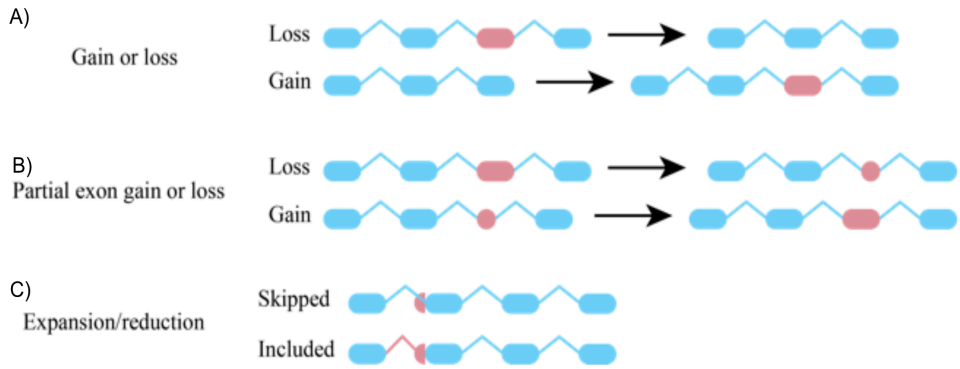


Figure 1.1: *An illustration of exon/ domain shuffling:* ***(a)*** *A gene can either gain or lose an exon subsequently altering transcriptional isoform output.* ***(b)*** *Genes can also acquire partial segments of protein coding DNA.* ***(c)*** *After gene duplication new splicing signals can be acquired from the reintegration into a new genic environment possibly causing the skipping or inclusion of new domains or exons into the gene.*

Quite like RTs, domain shuffling has a distinct distribution pattern around genomes. They are more likely to become inserted at the N' terminus of a gene as opposed to C' terminus or middle of a gene. This frequently results in fusion gene events as post domain duplication the duplicated copy becomes reinserted adjacent to a gene hitchhiking their regulatory machinery *e.g.* exons from the *TSG101* (member of a family of inactive ubiquitin-conjugating enzymes) duplicated and joined with the *UEVLD* gene (possible negative polyubiquitination regulator) (Buljan, Frankish and Bateman, 2010). Although many of these genes are transcribed they contain signatures for non-sense mediated decay (NMD). However, some shuffled genes can avoid NMD and remain within the genome allowing the adequate time neccesary to acquire mutations with the potential to result in positive selective if the mutation leads to a beneficial function acquisition (Keren, Lev-Maor and Ast, 2010). These selective pressures could lead to eventual fixation of the newly shuffled gene (Keren, Lev-Maor and Ast, 2010). Contrastingly, the reintegrated new gene could have a deleterious effect on host fitness and therefore undergo purifying selection (Long *et al.*, 2003). If the insertion does not impact the host's fitness or effect protein function/structure it has the potential to remain in the genome under neutral selective pressures (Long *et al.*, 2003).

1.2.5) The gene fusion/fission mechanism and it's role in new gene formation across genomes

Lastly, gene fusion and fission events are an under-studied source of new genes. Gene fission occurs when a gene is duplicated and split into one or more DNA segments with at least one becoming transcriptionally active (Doolittle, 1995). Typically, gene fission splits an ancestral 'parent' gene into two or more segments through genetic rearrangements *e.g.* recombination (Durrens, Nikolski and Sherman, 2008). Alternatively, a gene duplication event can occur and the resultant duplicated gene can accumulate mutations causing a shift in phase within the coding region of the gene resulting in a truncation event (Vogel and Morea, 2006) (Figure 1.2). Evidence for fission events have been identified in *Drosophila* (Wang, Yu and Long, 2004), fungi (Leonard and Richards, 2012),

bacterial (Pasek, Risler and Brézellec, 2006) and archaeal species (Snel, Bork and Huynen, 2000).

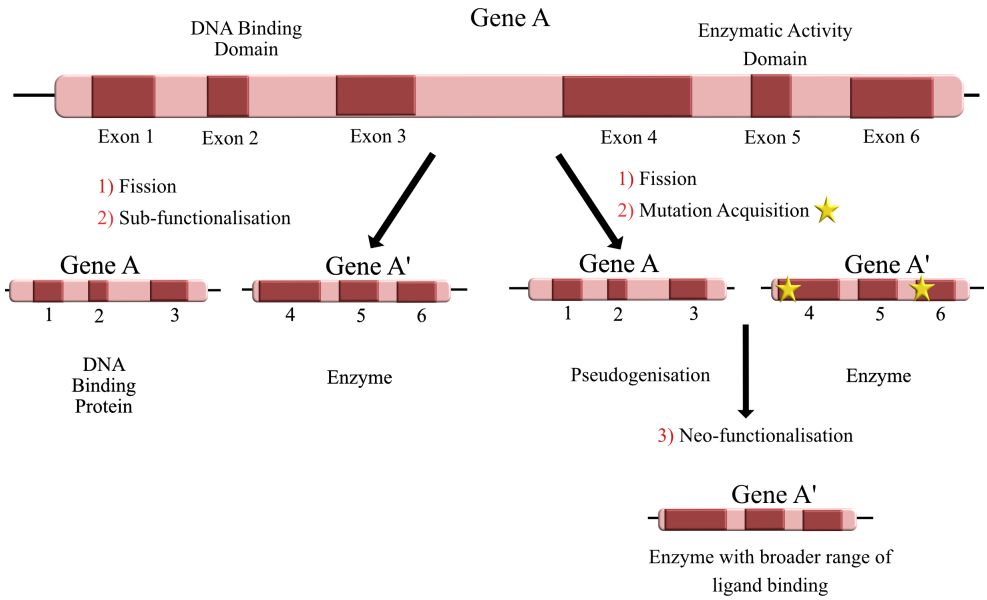


Figure. 1.2: *Method of gene fission and subsequent sub-functionisation or neo-functionalisation: Depicts a fission event of Gene A, whereby 2 genes are generated Gene A and Gene A'.* ***Left:*** *Illustrates subfunctionalisation post fission, generating 2 genes with 2 distinct functions, gene A forming a DNA binding protein, whilst Gene A' generates an enzymatic protein.* ***Right:*** *Illustrates a fission event across gene A whereby 2 genes are generated, Gene A retains original function and it's redundancy forces ts pseudogenisation whilst Gene A' acquires mutations allowing a increased range of ligand binding activity and thus is retained.*

There are two main mechanisms of gene fusion creation; 1) through the read-through of two adjacent genes to produce a novel transcript, i.e. RNA mediated gene fusion, or 2) through duplication of a gene and reinsertion of that copy into (or adjacent to) a coding sequence resulting in a fusion event (Kaessmann, 2010), i.e. DNA mediated gene fusion (Figure 1.3). The generation of gene fusions, perhaps unsurprisingly, correlates with regions of chromosomal instability caused by genomic rearrangement such as insertion/deletion, chromosomal inversion, and as translocation events (Edgren *et al.*, 2011) (Figure 1.4)

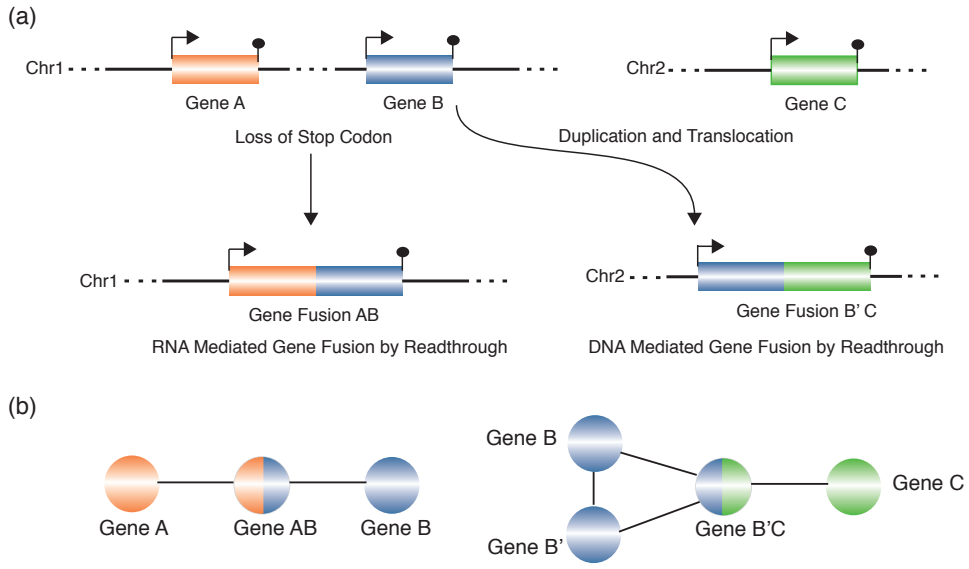


Figure 1.3: *Mechanisms of RNA-mediated and DNA-mediated gene fusion events:* ***a)*** *A depiction of the mechanisms behind both RNA-mediated and DNA-mediated gene fusion by readthrough***.** RNA mediated fusion occurs *through a loss of stop codon event between Gene A and B thus allowing readthrough transcription and the creation of the Gene AB. DNA-mediated gene fusion occurs after Gene B undergoes a translocation event from chr.1 to chr.2 adjacent to Gene C causing a gene fusion event and creation of Gene AC.* ***B)Left****: a graphical representation of the sequence similaity network generated by Gene AB.* ***Right****: a depiction of the sequence similarity network generated to illustrate the creation of Gene AC.*

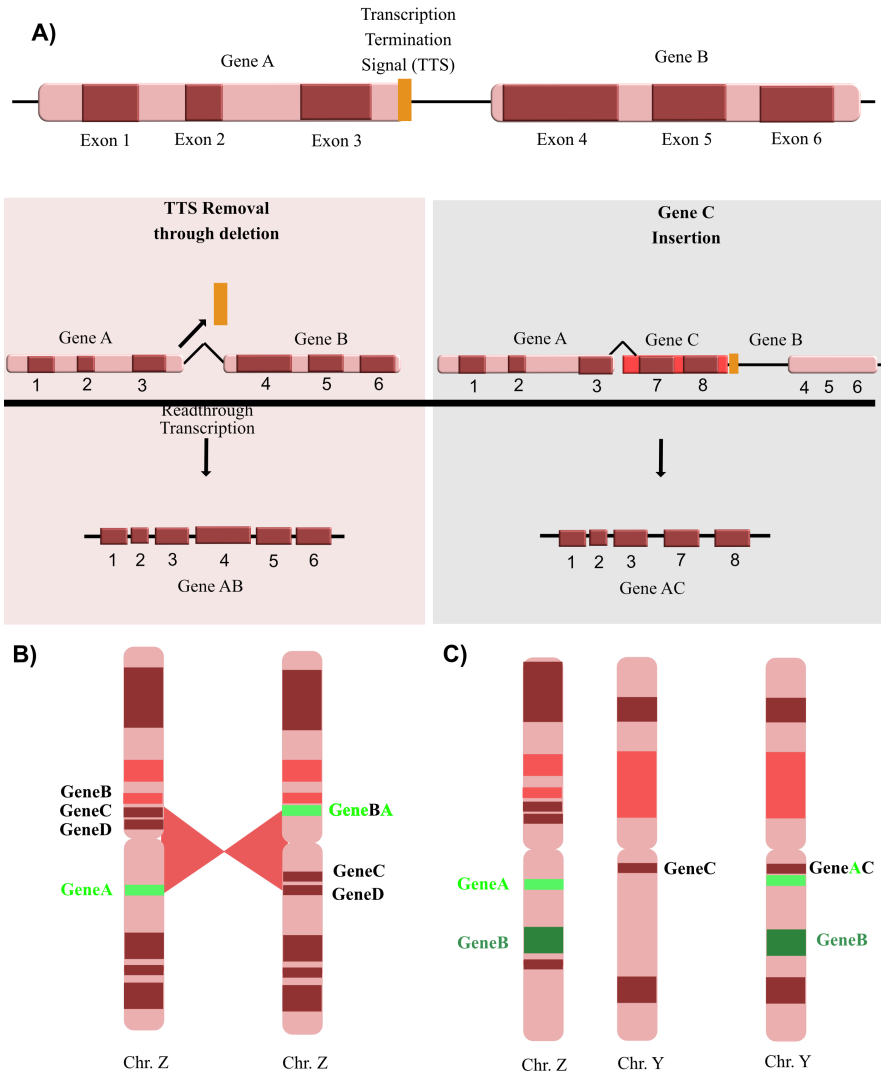


Figure 1.4: *Illustration of gene fusion creation through genomic rearrangments:* ***A)*** *Highlights fusion gene generation through INDEL with the pink box on the left hand side representing the creation of gene AB through the removal of a transcription termination signal through a deletion event allowing a readthrough event between gene A and gene B. The grey box highlights the creation of gene AC through an insertion of gene C between exon 3 of gene A and it's termination signal.* ***B)*** *Depicts a chromosomal inversion event on chromosome Z forcing gene A into a novel genetic loci adjacent to gene B forming gene AB.* ***C)*** *Diagram shows a translocation event of gene A and B on chromosome Z into a new chromosomal environment on chromosome Y forcing gene A into a novel genetic environment adjacent to gene C forming gene fusion AC.*

Genes near these unstable chromosomal locations are more promiscuous and tend to generate a larger number of fusion events in comparison to those located at more stable chromosomal regions. In general gene fusion was thought to be rare due to the probable deleterious effects caused by either truncation of the ancestral gene, possible frameshift mutations after fusion, or pseudogenisation (Kaessmann, 2010). Indeed, fused genes are often solely associated with cancerous growths due to their propensity for deleterious effects on the organism if transcribed and translated, and fused oncogenes have been associated with carcinomas (Teixeira, 2006), prostate cancers (Bartek, Hamerlik and Lukas, 2010), thyroid cancers (Ricarte-Filho *et al.*, 2009), epithelial tumors (French *et al.*, 2008), and lymphomas and leukemias (Sabir *et al.*, 2012) amongst numerous other cancers. However, individual cases of functional fused genes are known, e.g. *jingwei* in *drosophila* (Long *et al.*, 2003), the *kua-uev* fusion gene in human (Thomson *et al.*, 2000), *oc90* gene in early vertebrates (Wang *et al.*, 1998) and more recently a catalogue of fusion events in fungi (Leonard and Richards, 2012). In many cases these fused genes are not deleterious but rather they are selectively advantageous (Wang *et al.*, 1998; Thomson *et al.*, 2000; Long *et al.*, 2003).

The gene *jingwei* was discovered to display the classical property of a fused gene in that it had sequence similarity to two different genes/locations in the *Drosophila yakuba and Drosophila teissera* genomes. However, these regions have no sequence similarity to one another. Careful sequence comparison revealed that *jingwei* was produced from the fusion of a retrotransposed copy of the alcohol dehydrogenase (*Adh*) locus, into the third intron of a duplicate of the *yellow-emperor* gene, known as *yande* (Long and Langley, 1993). The jingwei protein has an increased substrate range in comparison to the original *Adh* gene allowing the breakdown of alternative diols, growth hormones and pheromones, thus providing it with a selective advantage and a critical role in development of the *Drosophila yakuba and Drosophila teissera* species.

Otoconin 90, *oc90*, is a fused gene that arose before the divergence of tetrapods and it is responsible for the production of otoconia crystals which provide a

species with a sense of orientation with respect to gravity (Wang *et al.*, 1998). In mammals otoconia crystals are composed of a lattice of otoconin proteins and calcium carbonate, whereas in amphibians and reptiles they are composed of otoconin proteins and aragonite, and fish possess the least stable lattice of otoconin proteins, vaterite (Wang *et al*, 1998). The *oc90* gene is thought to be a tripartite fused gene composed of phospholipaseA2/ phospholipase A2-like, HERV-H and HHAG-1 domains (Wang *et al.*, 1998).

The fused gene *kua-qev* was formed through loss of a stop codon and subsequent read-through of transcription from two adjacent genes (Thomson *et al.*, 2000). The resultant chimeric protein is a uniquitin transferase responsible for the elongation of non-canonical ubiquitin chains. Unlike it's parents, who locate to the nucleus, *kua-qev* locates to the cytoplasm.

The significant contribution of gene fusion/fission and exon/domain shuffling (referred to using as "gene remodelling" throughout this thesis) to the evolution of animals is clear from the multi-domain nature of the majority of animal proteins (Han *et al.*, 2007). Although many proteins contain multiple domains, a finite number of domains are responsible for the vast majority of the proteome and a mere 10,000 protein folds are known to exist amongst mammals (Koonin, Wolf and Karev, 2002). This pattern suggests that some protein coding regions (domains/exons) are promiscuous or prone to incorporation into many protein contexts, potentially causing many instances of new protein coding regions, whilst other regions are less compatible/promiscuous with others (Koonin, Wolf and Karev, 2002).

Surprisingly, new protein coding genes generated by gene remodeling has been under-studied and the contribution of gene remodelling to both genome evolution and speciation remains not entirely clear.

1.3) Vertebrate Genome Architecture

In 2009, Eugene Koonin defined genome architecture as "the totality of non-random arrangements of functional elements in the genome" (Koonin, 2009).

However, the non-random nature of genomes was first alluded to by Thomas *et al* (1971) who proposed the C-paradox that states an increase in genome size correlates with a more complex genome architecture (Thomas, 1971). This is evident across the tree of life but particularly when comparing the compact cistronic nature of prokaryote genomes (Jacob and Monod, 1961) to that of the large, complex, interleaved, intron containing nature of vertebrate genomes (Kapranov, Willingham and Gingeras, 2007; França *et al.*, 2017).

Paradoxically, despite the large size of vertebrate genomes and vast amount of intergenic DNA, genes tend to co-exist as part of high-density gene clusters (Tsochatzidou *et al.*, 2017). This clustering of coding content supports the non-random nature of genome architectures as clusters have a higher probability of sharing regulatory machinery/ expression level, chromatin structure/nucleosome accessibility and are likely to contain overlapping anti-sense transcripts that are responsible for regulating co-occurring genes (França *et al.*, 2017). The clustering of vertebrate genic material in specific genetic contexts highlights the importance of genome architecture and that specific genetic loci may have specific functional roles such as the antennapedia-like homeobox gene cluster that plays a crucial role in animal development (Schubert, Nieselt-Struwe and Gruss, 1993). The clustering of vertebrate olfactory receptor (Yoshihito Niimura, 2012) and immune genes also highlight this fact (Makino and McLysaght, 2008).

With advancements in population genetics theory it is now known that the framework of genome architecture is shaped by partially adaptive but mostly neutral selective pressures with increases in complexity caused by duplication and intron retention (to name but a few) only occurring at population bottlenecks (Koonin, 2009). In vertebrates specifically the small effective population size encourages complexity however, although containing highly structured regions such as protein coding gene clusters it paradoxically contains a wealth of highly mobile transposable elements creating highly disordered regions of the genome (Koonin, 2009). These data highlight although genome architecture is non-random it is not a perfect blueprint but rather under constant adjustment.

The process of fusion, is just one mechanism that has manipulated genomic architectural complexity and has subsequently played an important role in the evolution of vertebrate species through both chromosome and individual gene novel innovations. For instance, a fusion event of chromosomes 12 and 13 in chimpanzee is responsible for the origin of human chromosome 2 (Ventura *et al.*, 2012). This chromosome fusion was achieved by non-homologous recombination between two highly repetitive telomeric regions producing a speciation barrier between human and other closely related *hominid* species (Ventura *et al.*, 2012). Fusion has also had an impact on the single gene level across vertebrates *e.g.* the fatty acid synthase gene (McCarthy and Hardie, 1984), the glutamyl and prolyl synthase genes (Berthonneau and Mirande, 2000), and the *sp100rs* gene in mouse (Weichenhan *et al.*, 1998). These examples highlight the powerful nature of fusion genes in helping to shape vertebrate genome architecture and in creating inter-species differences, perhaps even on the inter-species phenome level.

1.3.1) A link between the dynamic nature of genome architecture and phenotype across vertebrates

Genome architecture has clearly played a vital role in the generation of genetic diversity amongst species, this is clear when comparing structural features across the vertebrate phylogenetic tree (Table 1.1) (Marques-Bonet, Girirajan and Eichler, 2009; Herrero *et al.*, 2016) Architectural genomic changes can be made at both a small and large scale. Here we will focus on the contribution of some of these gene and genomic architectural rearrangements and their contribution to phenotypic change. There are many sources of large-scale genetic rearrangements that contribute to vertebrate chromosomal and gene architectural changes with potential phenotypic consequences such as whole genome duplication (WGD) (Dehal and Boore, 2005), tandem duplication (TD) (Eicher *et al.*, 1976) and segmental duplication (SD) (Samonte and Eichler, 2002) (Figure 1.5).

Table 1.1: Genome Architecture comparison across vertebrates

Species (Release)	Genome Size (bp)	Protein-coding genes	SD percentage
Human (hg18)	3,609,003,417	20,376	5.32
Chimpanzee (panTro2)	3,385,800,935	23,534	3.78
Orangutan (panAbe2)	3,109,347,532	20,424	1.18
Macaque (rheMac2)	3,146,411,622	21,099	1.55
Mouse (mm8)	3,286,944,526	22,630	6.13
Rat (rn4)	3,042,355,753	22,250	1.60
Dog (canFam2)	2,392,715,236	19,856	0.82
Cow (bosTau4)	2,649,685,036	19,994	5.63
Opossom (monDom4)	3,501,660,279	21,329	2.70

Genome architectural comparisons across selected vertebrate species obtained from the Ensembl Genome Browser (Herrero et al., 2016). Comparisons include genome size, protein-coding gene quantity and segmental duplication (SD percentage across genome (Marques-Bonet, Girirajan and Eichler, 2009)

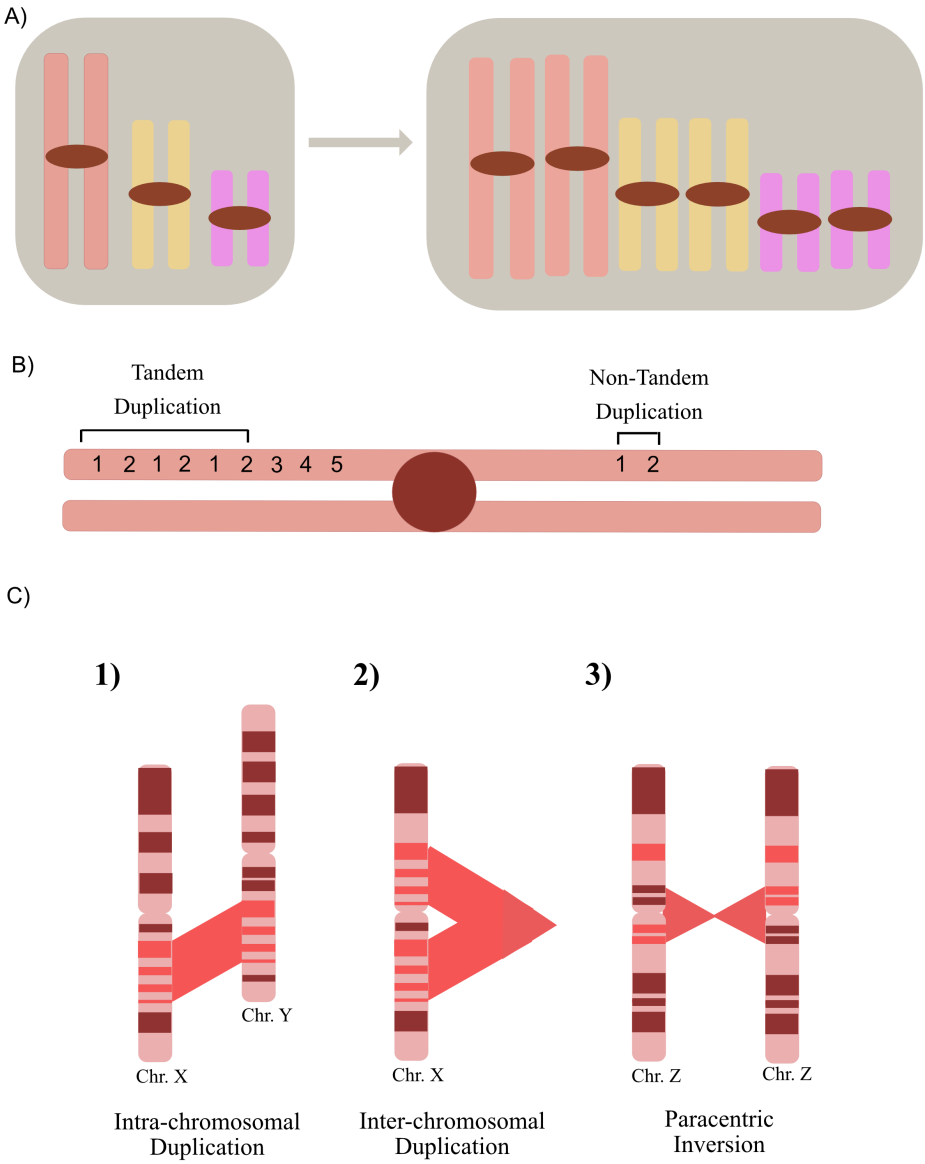


Figure 1.5: *Duplication mechanisms behind large-scale genetic rearrangements:* ***A)*** *Depiction of the WGD mechanism,* ***B)*** *illustration of tandem duplication (TD) genes with sequence 1-2 duplicating and re-inserting exactly adjacent to the original sequence.* ***C)*** *Diagram illustrating the 3 mechanisms of segmental duplication:* ***1)*** *highlights SD events between chromosome X and chromosome Y,* ***2)*** *shows an SD event from a genetic locus on chromosome X to a new genetic locus on the same chromosome and* ***3)*** ***a*** *paracentric inversion SD event whereby an SD flips orientation across the centromere of chromosome Z.*

1.3.2) Genome rearrangments and their impact on vertebrate evolution

WGD events across vertebrate genomes have been extensively studied and it is now known that two rounds of WGD are responsible for the diversification of vertebrates from their invertebrate ancestor. However, due to the 10-100million years of evolution, gene loss, mutation acquisition, and chromosomal rearrangements many polyploid genes have reverted back to diploid copies (rediploidisation) and therefore deciphering the contribution of WGD to the architectural diversification and evolution of vertebrate phenotypes remains challenging (MacKintosh and Ferrier, 2017). Although challenging, many cases of the contribution of WGD to vertebrate phenotype diversity exist in the literature, these include the expansion of the *sox* gene in teloest fish, a transcription factor controlling morphology, physiology and behaviour, whereby 11/19 *sox* genes within the present day cluster arose from WGDs thus facilitating phenotypic diversity (Voldoire *et al.*, 2017). The *Gli* gene family responsible for tissue shape during vertebrate development also expanded in vertebrate genomes through WGD (Ruiz I Altaba, 2011) as well as the PSD supercomplexes which are responsible for the control of cognitive abilities and emotions (Grant, 2016). More generally, WGD has been highlighted as a key contributor to vertebrate signal transduction elaborations from more simple transduction pathways to more complex, highly parallelized pathways (MacKintosh and Ferrier, 2017).

TD is a duplication mechanism known to play a role in the diversification of vertebrate genome architecture and phenotype. Cases of TDs contribution to phenotypic diversity currently exist in the literature. These include the secretory calcium binding phosphoprotein (*SCPP*) gene family that controls the mineralisation of dental tissue in modern vertebrates. The family originated from a WGD of the *SPARC* gene that created the *SPARCL1* gene, *SPARCL1* then underwent a subsequent TD during the split from ray finned to lobe finned fish creating two genes and two tissue types, surface and body (Kawasaki, Buchanan and Weiss, 2007). It has also been discovered that TDs are the main mechanism of gene reordering across vertebrate mitochondria, this is particularly evident in the WANCY mitochondrial region of vertebrate mitochondrial genomes (San Mauro *et al.*, 2006). Another mechanism of genomic rearrangement that has

contributed to architectural diversification across vertebrates causing phenotype shifts is segmental duplication.

1.4) The significance of SD and phenotype diversity across vertebrate species

Segmental duplicates (SDs) are 1-100kbp units that are typically >90% identical in sequence to that of the original sequence (Bailey and Eichler, 2006). Although the detailed mechanisms of SD generation remain disappointingly elusive there are 3 major processes known to play a role in the redistribution of SDs around genomes, they are (1) intra-chromosomal duplication (Zhang *et al.*, 2005), (2) inter-chromosomal duplication (Zhang *et al.*, 2005), and (3) paracentric inversions (Cuscó *et al.*, 2008), the mechanisms of which are outlined in Figure 1.5.

Some chromosomes are more prone to duplication than others, (Bailey and Eichler, 2006) and they cluster toward certain chromosomal regions, particularly telomeric and centromeric sequences (Bailey *et al.*, 2002; Zhang *et al.*, 2005). This is highlighted in an analysis of inter/intra chromosomal SDs abundance across human chromsomes (Table 1.2) (Zhang *et al.*, 2005).

Table 1.2: Inter and intra chromsomal SD clustering in across human chromosomes

Top 5 Elevated SD		Top 5 Reduced SD		Top 5 Pericentromeric SD		Top 5 Subtelomeric SD	
Chr	% SD	**Chr**	% SD	**Chr**	% SD	**Chr**	% SD
Y	14.4	**12**	1.9	**7**	7	**9**	14.8
16	9.2	**6**	1.7	**2**	6.9	**7**	11.7
17	7.9	**8**	1.5	**9**	5.7	**1**	9.9
7	7.4	**3**	1.3	**16**	4.9	**5**	7.6
22	7	**14**	1.1	**10**	4.7	**4**	6.2

Illustrates the distribution of both inter/ intra chromosomal SD events across human chromosomes. Column 1 depicts the human chromosomes with highest SD level and their % of the total chromosomal content. Column 2 illustrates the human chromosomes with the least SD content. Column 3 highlights the human chromosomes with the highest paracentromeric SD and column 4 represents human chromosomes with highest subtelomeric SD levels (Zhang et al., 2005).

Clustering of SDs in telomeric and centromeric regions has resulted in regions of genomic instability that increase the probability of further rearrangments in these locations creating highly complex nested regions of duplication (Bailey *et al*, 2006), (Cheng *et al.*, 2005). Clustering begins with a single duplication event into a founder recipient region - a founder event. This unstable founder location becomes a 'duplication hotspot' and further duplicates migrate into this recipient region leading to complex repetitive regions of nested SD (Symmons *et al*, 2008).

There is also a known shift in the chromosomal positioning of SD with an SDs transmitting interchromosomally before the divergence of Old World Monkeys to a subtelomeric and pericentromeric transmission after this divergence (Bailey and Eichler, 2006). Interestingly, this location switch corresponds to two major transitions in genome structure; 1) the shortening of telomere sequences, and 2) the introduction of alpha-satellites into centromeric sequences of primate species (Bailey and Eichler, 2006). Alpha-satellites are 171bp highly repetitive sequences found in primate centromere sequences, the clustering of these sequences have been associated with cell division due to it's role in kinetochore formation for microtubule attachment (Sullivan *et al,* 2017).

It has been shown that the Great Apes have undergone a surge of SD despite a co-occurring slowdown of all other genetic rearrangements - 'the hominid slowdown' (Figure 1.6) (Marques-Bonet, Girirajan and Eichler, 2009). This slowdown has been proposed to be a consequence of more effective proofreading machinery required to deal with increased genomic complexity (Yi, 2013). In *homininae*, SDs are not only found at an increased frequency but are also enriched in protein-coding genes, this is unlike any other *hominidae* and vertebrate species tested to date (Bailey *et al.*, 2002)

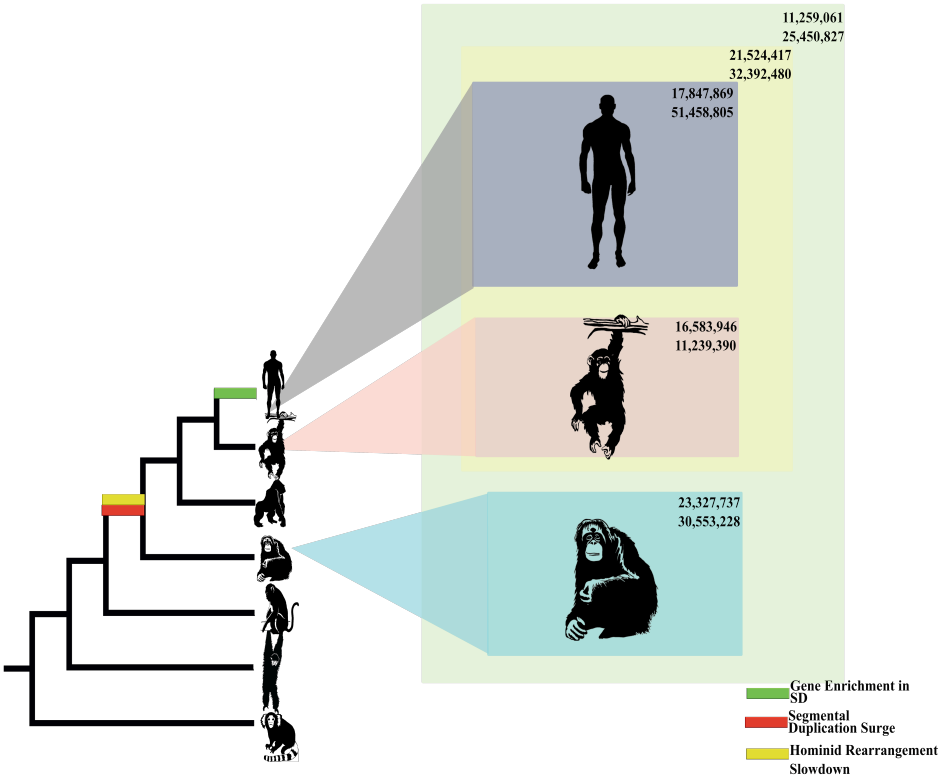


Figure 1.6: *Illustration of segmental duplication frequency across primate species: Phylogram indicating major genomic rearrangment transitions across a panel of primate species namely the hominid slowdown, SD surge and gene enrichment within regions of SD. Also highlighted within each species box;* ***Top:****Number of SDs identified with adjusted copy number variation and* ***Bottom:*** *Number of SDs identified with no adjusted copy number variation. The yellow box indicates SD numbers shared between human and chimpanzee species and the green box illustrates SD number across human, chimpanzee and orangutan.*

The human specific gene-rich nature of SD may be as a consequence of SD frequency within complex genomes in comparison to gene-poor duplication processes found across vertebrate species with less complex genomes. For instance, rodents contain a decreased frequency of SD (2-3% of overall genome content) in comparison to TDs (Bailey *et al.*, 2004), whilst primate genomes contain much higher SD content with the human genome containing ~5-10% of SD content and decreased TDs in comparison to mouse (Bailey *et al.*, 2004). Not only does this cause significant genome architectural differences across vertebrates but due to the gene-poor nature of rat and mouse TDs and SDs (mouse SD contain 0.56% annotated RefSeq genes) in comparison to the gene-rich nature of human SDs (human SD contain >6% annotated RefSeq genes and 177 human-specific genes) the gene frequency and gene family expansion frequency across vertebrates is effected (Cheng *et al.*, 2005; Tatusova *et al.*, 2015; Jiang and Ramachandran, 2016). Interestingly, SD breakpoints found in human with a single locus in mouse have been shown to have a higher probability of being correlated with disease phenotypes in human with 12 syndromes being localised to SD breakpoints (Lupski, 1998).

Human SDs are not only enriched for protein coding sequences but are also likely to contain at least 1/14 known 'duplicon' sequences (Lorente-Galdos *et al.*, 2013). 'Duplicons' are sequence elements that are commonly found throughout the genome, particularly in gene families. Their repetitive nature results in regions of genetic instability making locations containing duplicons susceptible to further rearrangement (Figure 1.7). SDs containing these 'core duplicons' have been found to have an increased level of protein coding sequences (Lorente-Galdos *et al.*, 2013). As genes within regions of SDs predominantly contain duplicon sequences, and duplicon sequences are commonly found in gene families this suggests that SD must play a role in the expansion of gene families.

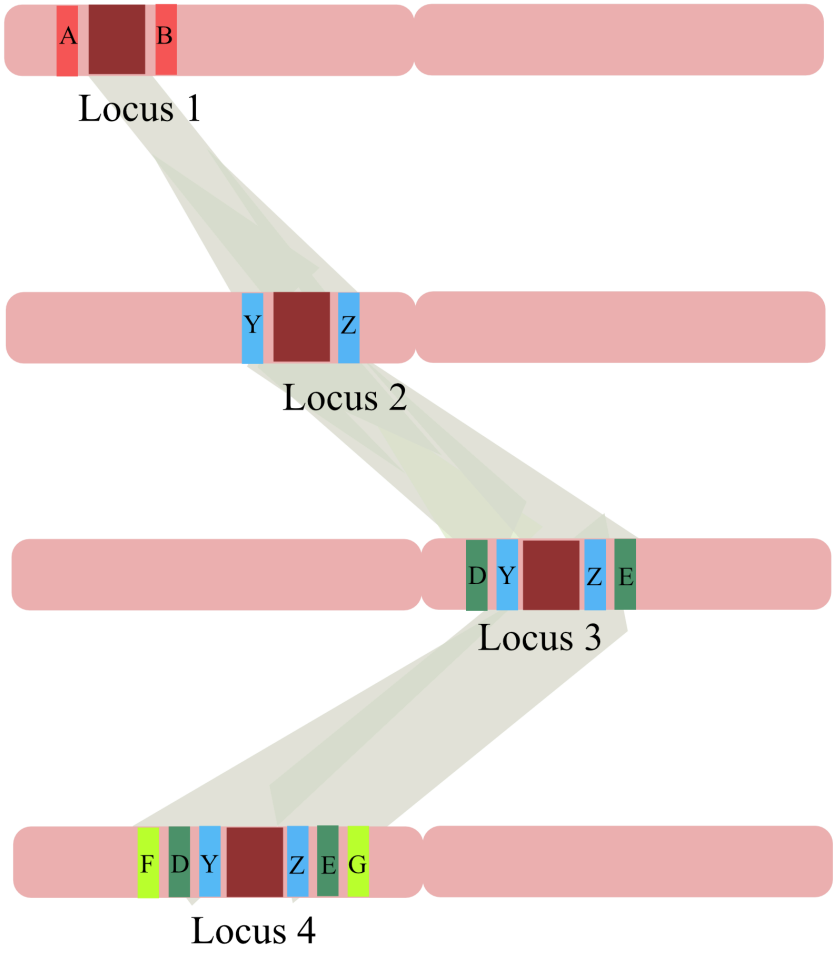


Figure 1.7: *Illustration of the mechanism behind the formation of nested duplicon structure in genomes: Illustration of the core duplicon expansion across regions of genomic instability (loci 1-4).*

The *LRRC37* gene (protein interaction motif) family expansion is an example of expansion due to a SD event. In mouse, two of this gene family have testis-specific expression however in macaque they are ubiquitiously expressed at low levels but are predominantly expressed in the testis (Giannuzzi *et al.*, 2013). Comparatively, human chromosome 17 has a total of 18 LRRC37 duplicates, 13 are transcribed ubiquitiously but at elevated levels in the cerebellum and thymus. Although these genes are highly conserved across primates, splice variants that allow relocalisation, new ligand binding, or increased affinities to original ligands are only found in Great Apes (Giannuzzi *et al.*, 2013). In short, SDs not only have the ability to generate novelty through the creation and expansion of gene families, but also through the generation of novel protein-coding genes by gene fusion. It is hypothesised that the increase in SD in vertebrates could therefore increase the probability of new gene genesis by gene remodeling (Stankiewicz *et al.*, 2004). Although the mechanisms behind the generation of new genes are largely understood not all new genes with a defined open reading frame (ORF) are actively transcribed as untranscribed RNA transcripts have been discovered with transcription regulatory roles. Therefore, an understanding of transcription and how it is controlled is essential to understanding the role of new genes in the generation of species diversity.

1.5) Transcription and methods of transcriptional control across vertebrates

1.5.1) Mechanisms of eukaryotic transcription in eukaryotes

Eukaryotic transcription mechanisms although similar across the tree of life have an increased complexity as a result of expansions in genome size and additional DNA compaction (Macadangdang *et al.*, 2014).

Transcription at the most basic level is the generation of a RNA transcript from a DNA template by a DNA-dependent RNA polymerase (RNAP) enzyme that are present across all of the branches of the tree of life (Werner and Grohmann, 2011). Although present, RNAP are different in both number and composition, with bacteria and archaea containing only one RNAP whilst eukaryotes contain

three RNAPs (Werner and Grohmann, 2011). Ancient/ancestral eubacteria such as *E. coli* contain a single RNAP protein composed of two β subunits (β and β') as well as three smaller subunits responsible for assembly and transcription regulation (Finn *et al.*, 2000). Cyanobacteria (the precursor of the modern day chloroplast) has a slightly more complex RNAP containing an additional γ subunit that arose from the β subunit gene undergoing a split (Gunnelius *et al.*, 2014) whilst mitochondrial RNAPs are more similar to that of bacteriophage containing α, β and β' subunits (Gaspari, Larsson and Gustafsson, 2004). As previously mentioned eukaryotes require three enzymes to cope with genomic complexity and transcriptomic diversity; RNAP1 is required for rRNA transcription, RNAP2 transcribes mRNA and some small non-coding DNA whilst RNAP3 is responsible for tRNA, small non-coding RNA and 5sRNA transcription (Korkhin *et al.*, 2009). Genome size playing a role in RNAP number is highlighted by plants (typically polyploidy) requiring an additional fourth RNAP for efficient transcription (Zhang *et al.*, 2007).

In eukaryotic cells RNAP2 is responsible for mRNA expression therefore enhanced enzymatic versatility while retaining tight control is of fundamental importance in order for mRNA levels to accurately meet both the spatial and temporal requirements of the gene in response to environmental stimuli. The importance of this enzyme cannot be underestimated as it underlies all of life's most basic processes. As previously mentioned DNA compaction amongst complex organism plays an important role in the initiation process and DNA binding proteins (DNA-BP) such as TFIID and SWI/SNF must unravel double helical structure prior to sequence motif binding, the most well studied of which being the helix-turn-helix motif (Aravind *et al.*, 2005). This occurs differently for different DNA forms. There are three known forms of DNA found in the cells of eukaryotes, namely Z, B, and A DNA forms (Potaman and Sinden, 2005) and the differences are highlighted in Table 1.3. Eukaryotic cells mainly rely on RNAP2 acting on B-form DNA for mRNA transcription using promoters with TATA boxes and INR (initiator) sequences.

Table 1.3: A comparison of DNA forms in eukaryotic species

	A form	B form	Zform
Helical sense	Right	Right	Left
Diameter	~26Å	~20 Å	~18 Å
Base pairs per turn	11	10.5	12
Helix rise per base pair	2.6 Å	3.4 Å	3.7 Å
Base tilt to helix	20^{o}	6^{o}	7^{o}
Abundance in cell	Rare	Frequent	Rare

Content of table compares the 3 DNA forms found in eukaryotic cells namely A, B and Z forms.

The process of transcription itself is a three step process; 1) Initiation (Krishnamurthy and Hampsey, 2009) 2) Elongation (Spencer and Groudine, 1990) and 3) Termination (Porrua *et al*, 2016) (Figure 1.8, Figure 1.9 and Figure 1.10 respectively**)**. Genes under RNAP2 control contain a core promoter sequence, enhancer elements and a TATA box (Figure 1.8) (Uzman *et al.*, 2000). Both general transcription factors (GTFs) and enhancer elements, specifically TFIID and SWI/SNF, are essential for DNA decompaction and the penetration of double helix structure surrounding the gene undergoing transcription. Prior to the initiation of transcription the assembly of a transcription initiation complex (PIC/TIC) and its recruitment of RNAP and 6 general GTFs are essential. The initiation process is highlighted in Figure 1.8.

When the appropriate RNAP has bound to the appropriate promoter through the appropriate accessory proteins this yields a closed initiator complex, followed by the melting of H+ bonds within the double helix around the transcription start site (TSS) forming an open initiator complex (Uzman *et al.*, 2000) (Figure 1.8). The chromatin modification factors, TFIIH and TFIIE, melt the DNA forming a single stranded RNA/DNA hybrid forming a bubble of ~25nts at the RNAP binding point allowing elongation to ensue (Figure 1.9) (Orphanides *et al*, 1996).

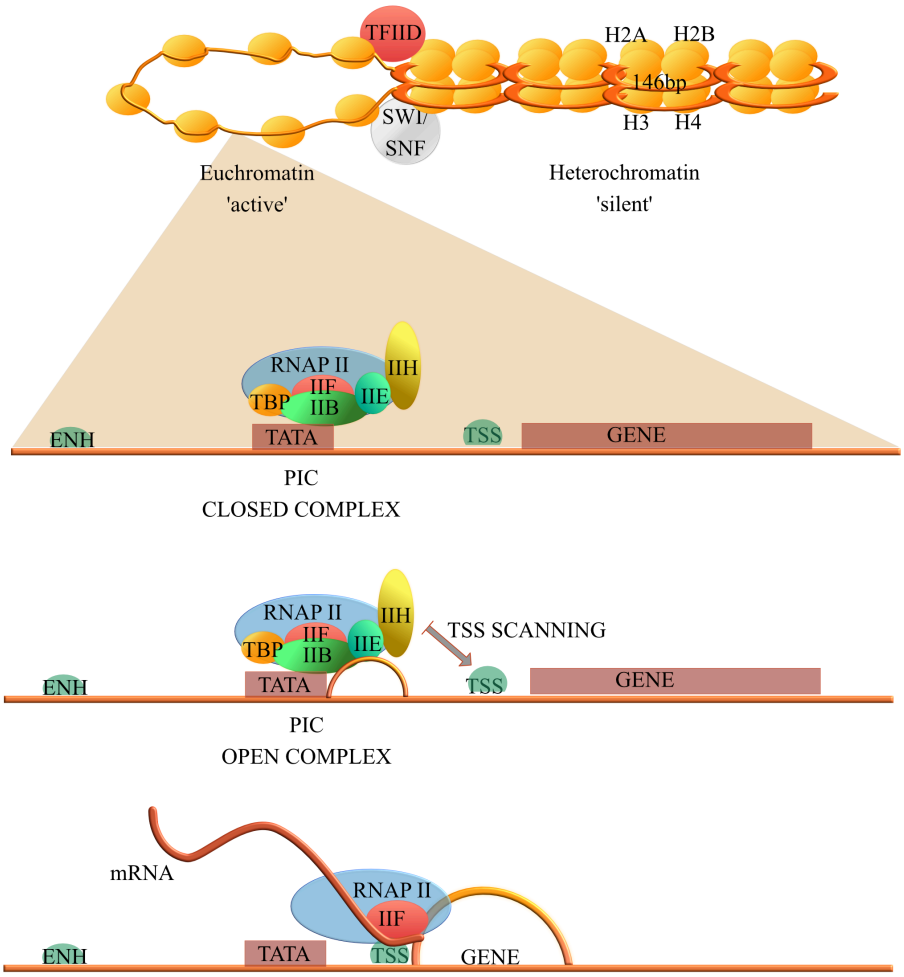


Figure 1.8: *An overview of transcription initiation in eukaryotic cells: Depiction of transcription initiation process in eukaryote cells starting with nucleosome decompaction by TFIID and SWI/SNF creating in a promoter initiation complex (PIC). Once this PIC has been formed hydrogen bonds melt across the gene of interest resulting in a bubble forming an open PIC complex. Finally the RNAP2 enzyme moves to the TSS so that transcription can ensue.*

Initiation is followed by elongation which is the addition of ribonucleic acids in a 5'-3' complementary manner to the target DNA template strand of a gene under active transcription (Spencer and Groudine, 1990) (Figure 1.9). In eukaryotes, ribonucleic acids (RNAs) are added at a rate of 22-25 NTPs per second. However, before elongation ensues three events occur 1) the PIC undergoes partial disassembly whereby some PIC subunits remain on the promoter to aid more efficient re-initiation of transcription (Panov, Friedrich and Zomerdijk, 2001). 2) Promoter clearance occurs due to the binding of two NTPs to the RNAP catalytic cleft/active site (ATP-dependent) (Luse, 2013) and 3) the traversing of RNAP along the nucleosome. After transcripts reach >25nts in length a stable elongation complex are formed (Pal and Luse, 2003) and the C' terminal domain (CTD) of RNAP is phosphorylated (ATP dependent) at Ser5 residues allowing elongation to ensue (Phatnani and Greenleaf, 2006) (Figure 1.9). This process is known as translocation. There are currently two main models of translocation 1) The power-state ATP dependent model requiring chemical to mechanical energy conversion and 2) The ATP independent Brownian Ratchet model that progresses by a series of condensation reactions (Imashimizu *et al.*, 2014).

Elongation is tightly linked to RNA processing and they both work more efficiently during in synchrony (Yamaguchi *et al*, 2013). The elongation factors *DSIF* and *NELF* control the 5' capping of the nascent transcript (Figure 1.9) through the recruitment of through three enzymes: 5' triphosphase, gaunyltransferase and methyltransferase. Splicing and PolyA tail addition are also both dependent on the phosphorylation of the CTD suggesting it acts as a binding dock for elongation factors (Yamaguchi *et al*, 2013).

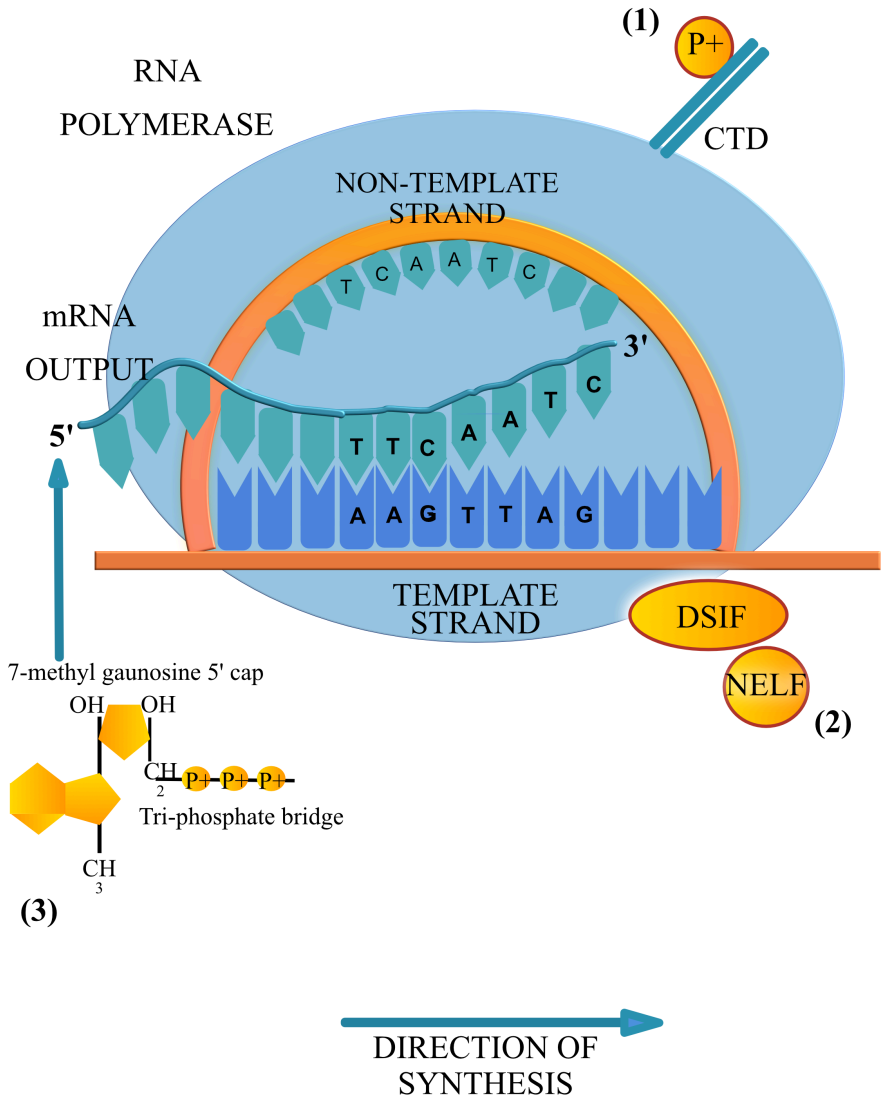


Figure 1.9: *Depiction of transcriptional elongation and 5' capping of transcript in eukaryotic cells: Shows the mechanism of transcription elongation post traversing the TSS in eukaryotic cells. Also highlighted is the synchronized 5'capping procedure with 1) demonstrating the phosphorylation of the CTD on the RNAP enzymes which is a prerequisite to 2) the binding of DSIF and NELF to the elongating RNA polymerase which recruits 3 enzymes resulting in 3) the addition of the 5'-7-methyl guanosine cap.*

Post-elongation, appropriate termination of nascent RNA transcripts is essential to protect the cell from aberrant transcript generation, alternative transcript production, adjacent promoter blockage, interference with downstream gene replication and expression through inappropriate RNAP and DNAP blockages.

In Eukarya, the mechanisms behind termination differ for each RNAP (RNAP1, 2, and 3). Detailed below are the mechanisms of termination behind each polymerase. As previously mentioned RNAP1 is responsible for the transcription of precursor rRNAs (18S, 5.8S, and 28S in humans). These genes are organised in a tandemly repeated fashion separated by intergenic sequences (IGS) and are actively transcribed in the nucleolus (Richard and Manley, 2009). The stepwise ATP dependent process of RNAP1 termination is highlighted in Figure 1.10

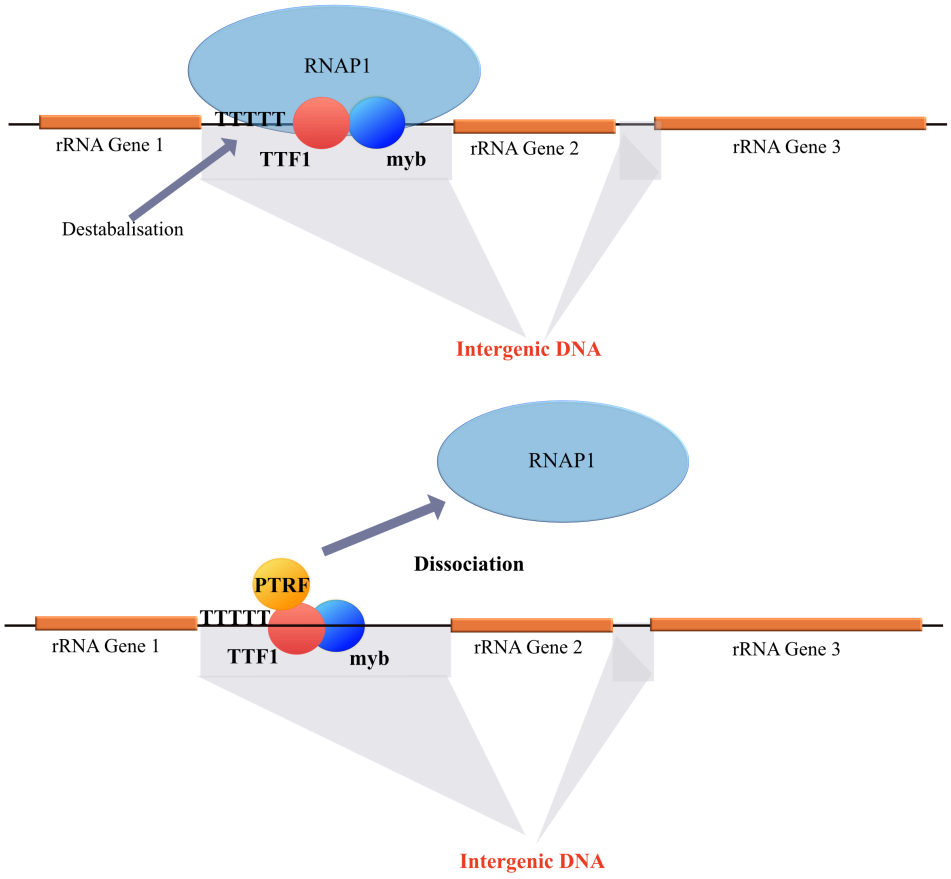


Figure 1.10: *RNAP1transcriptional termination in eukaryotes: An illustration of RNAP1 transcription termination across rRNA tandemly repeated units in eukaryotes. RNAP1 termination depends on 1) a destabilising event caused by a stretch of thymines (T's) upstream of the TTF1 binding motif, 2) binding of TTF1 and myb proteins to intergenic motifs and subsequent RNAP1 binding and 3) the binding of PRTF to TTF1.*

mRNA transcription is dependent on RNAP2 and thus this polymerase is the most abundant in eukaryotic cells used polymerase throughout eukaryotes. Therefore, it is of fundamental importance that the termination of RNAP2 transcription is under tight control with adequate back-up/redundancy termination and proofreading mechanisms of nascent transcripts (Richard and Manley, 2009). There are two major termination mechanisms in eukaryotes (Figure 1.11); 1) the NNS pathway and 2) the CTF pathway followed by torpedo termination (Figure 1.12).

The torpedo model follows the CTF termination pathway allowing the entry of a 5'-3' exonuclease, namely Xrn2 in human which not only causes RNAP2 dissociation but can also trigger nascent transcript degradation by the proteasome and stimulates polyA tail addition (Figure 1.12) (Richard and Manley, 2009).

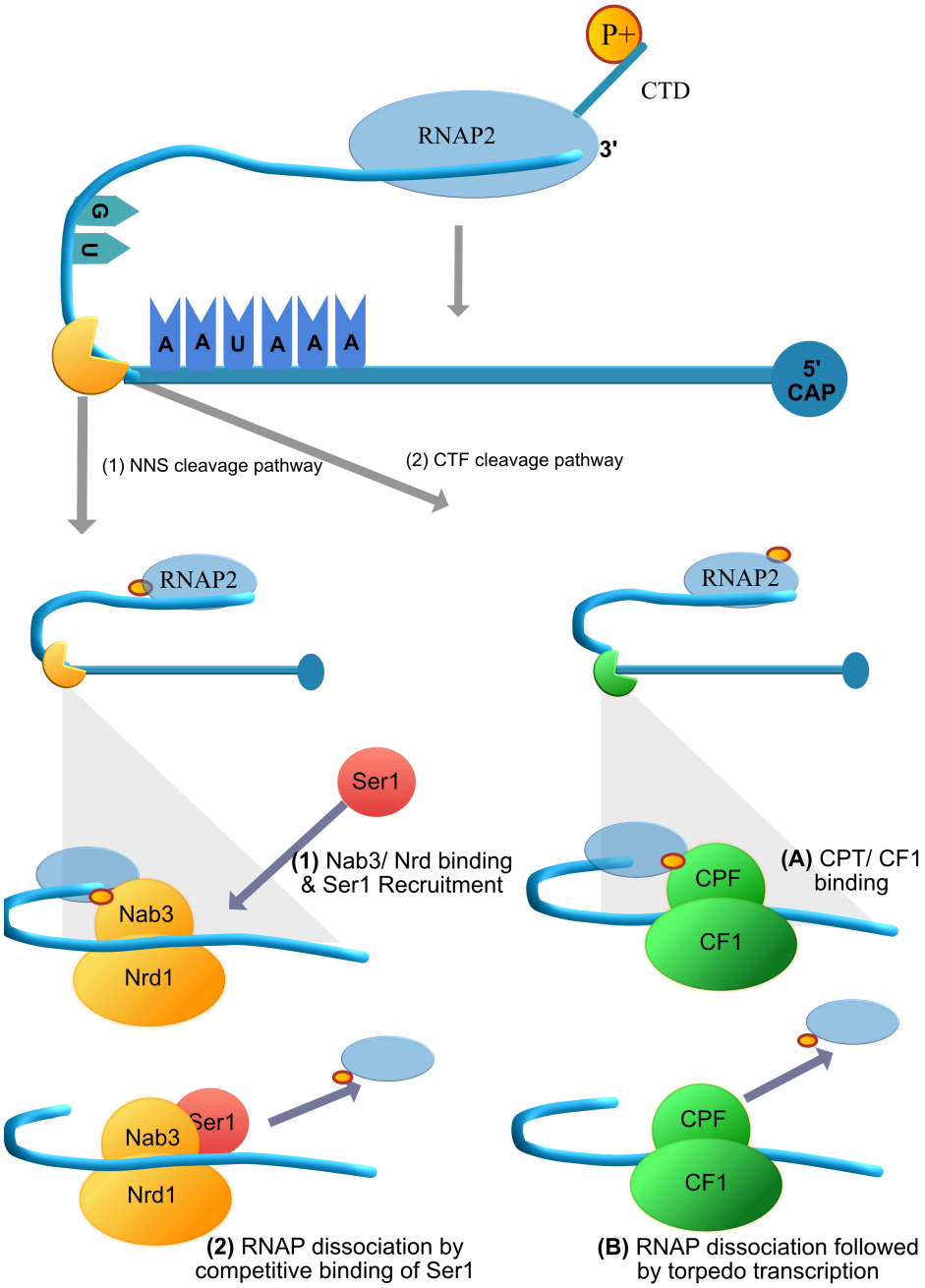


Figure 1.11: *Illustration of the known mechanisms behind RNAP2 transcription termination in eukaryotes: RNAP2 termination can occur by 2 mechanisms, the 1) the NNS pathway whereby Nab3/Nrd1 binding recruit the Ser1 protein causing termination and 2) the CPF/CF1 complex causes termination followed by torpedo termination*

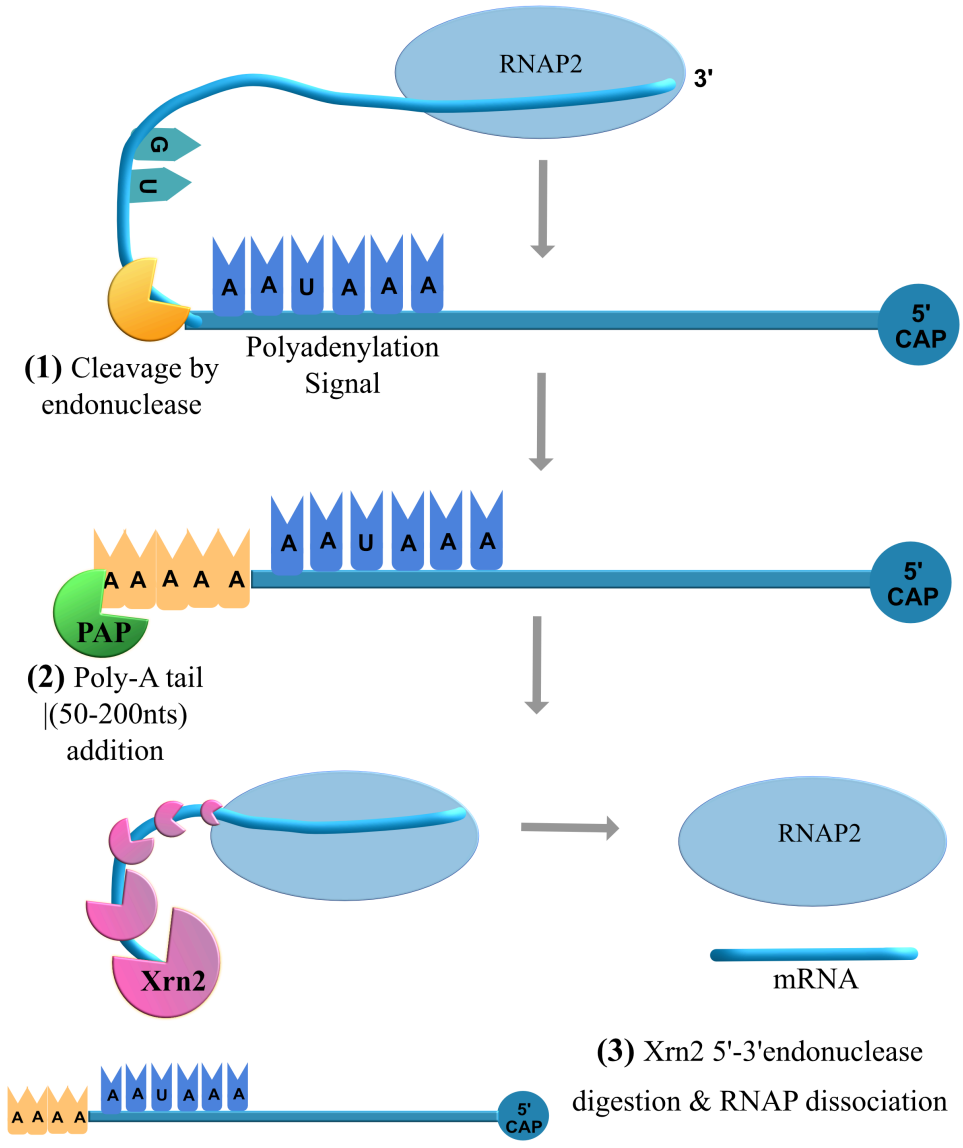


Figure 1.12: *Torpedo transcriptional termination mechanism and polyA tail addition:* *Depiction of transcription termination utilizing the CTF pathway and torpedo model of RNAP2 dissociation from the mRNA. PAP indicates the polyadenylate polymerase enzyme.*

Due to the importance of accurate termination particularly by RNAP2, fail-safe mechanics are present. For example a road-block mechanism exists that recruits the Reb1 transcription factor (TF) to RNAP2 causing ubiquitisation which triggers both nascent transcript and polymerase degradation (Richard and Manley, 2009).

Finally, the RNAP3 enzyme both efficiently and cyclically produces short, <400nts in length transcripts due to the pausing/reinitiating of the enzyme instead of it's complete removal from nascent RNA (Richard and Manley, 2009). Termination requires a stretch of thymine nucleotides, typically 4/5nts in mammals (Richard and Manley, 2009). Flanking up/downstream sequences have been shown to enhance termination but no conserved sequences have been identified to date. It has been proposed that the instability caused by the A-U RNA/DNA hybrid allows for efficient transcriptional termination. However, transcriptional pausing is only found after 4 thymine nucleotides have been transcribed, complete termination on the other hand requires transcription of a 5th thymine nucleotide (Richard and Manley, 2009).

RNAP2 is prone to active site conformational changes resulting in elongation pausing at specific DNA sequences due misalignment of RNAP2 and 3'OH end of the nascent transcript - backtracking. Backtracking is a proofreading mechanism and is useful in slowing elongation at exon/intron junction boundaries facilitating splicing (Voliotis *et al.*, 2008). However, if paused indefinitely DNA replication can become inhibited and cause cell toxicity (Voliotis *et al.*, 2008).

Understanding the mechanisms behind each RNAP entirely remains a challenge but must be completely defined in order to completely understand transcriptional control within cells. However to understand the mechanisms completely requires an understanding of *cis* and *trans* sequences involved in controlling transcription level.

1.5.2) An overview of *cis* and *trans* factor involvement in transcriptional control

Changes in gene expression can impact phenotypes, e.g. (Yi *et al.*, 2010; Ruzycki *et al.*, 2015) and species diversification (Jones *et al.*, 2012; Mack *et al,* 2016). Many *trans*-acting transcription regulatory factors have evolved, e.g., TFs (Vaquerizas *et al.*, 2009), RNA splice factors (SFs) (Busch and Hertel, 2012) and histones (Girton and Johansen, 2008). Their *cis*-acting counterparts are transcription factor binding sites (TFBS), splice factor binding sites (SFBS) and histone binding sites. .

1.5.2.1) Regulation of Gene expression by Transcription Factors

TFs are a set of DNA-BPs that bind specific *cis*-consensus regulatory sequences in order to either activate or repress gene expression in a spatial and temporal manner (Philips and Hoopes, 2008). TFs control gene expression on two levels. The first level is responsible for the control and maintenance of basal gene expression levels through gene promoter binding - a kind of housekeeping expression regulation (Kuhlman *et al.*, 1999). These TFs are usually part of the PIC/ TIC that binds the promoter of the gene (Kuhlman *et al.*, 1999). The second level of control is responsible for the enhancement or repression of genes already under basal transcription through enhancer/silencer sequence binding (Ishibashi *et al.*, 2014), (Griffiths *et al.*, 2000). TFs of this nature have the ability to bind many enhancer sequences across the same gene, however expression levels in target tissues may be influenced depending on what enhancer/silencer sequence is bound adding an additional layer of complexity in understanding TF impact on transcriptional control (Ko *et al.*, 2017). This additional level of TF control acts to respond to the organisms requirements during processes such as development (Spitz and Furlong, 2012), pathogenesis (Mahdi *et al.*, 2013), cell-cycle control (Bertoli *et al*., 2013), and even in response to external stimuli (Lorberbaum and Barolo, 2013). TFs have 2 key defining characteristics: 1) the presence of a DNA binding domain (DBD), and 2) a trans-activating domain (TAD) (Verrijzer *et al.*, 1990) and TF binding to their corresponding TFBSs through Van der Waals interactions is highlighted in Figure 1.13 (Inukai *et al*., 2017), (Farrel *et al*., 2016). Figure 1.13 also highlights that mutations in TFBSs are tolerated but not

preferable (Alberts *et al.*, 2002). On binding of the TF's DBD transcriptional activation or repression of the gene ensues. Examples of DBDs include: zinc fingers, helix turn helix, and basic leucine zippers (Alberts *et al.*, 2002). TFs generally do not function in isolation but recruit co-regulatory factors, *e.g.* p53, NFkB, and NFA, that provide additional genetic control and are recruited through the TF's TAD (Figure 1.13) (Blau *et al.*, 1996).

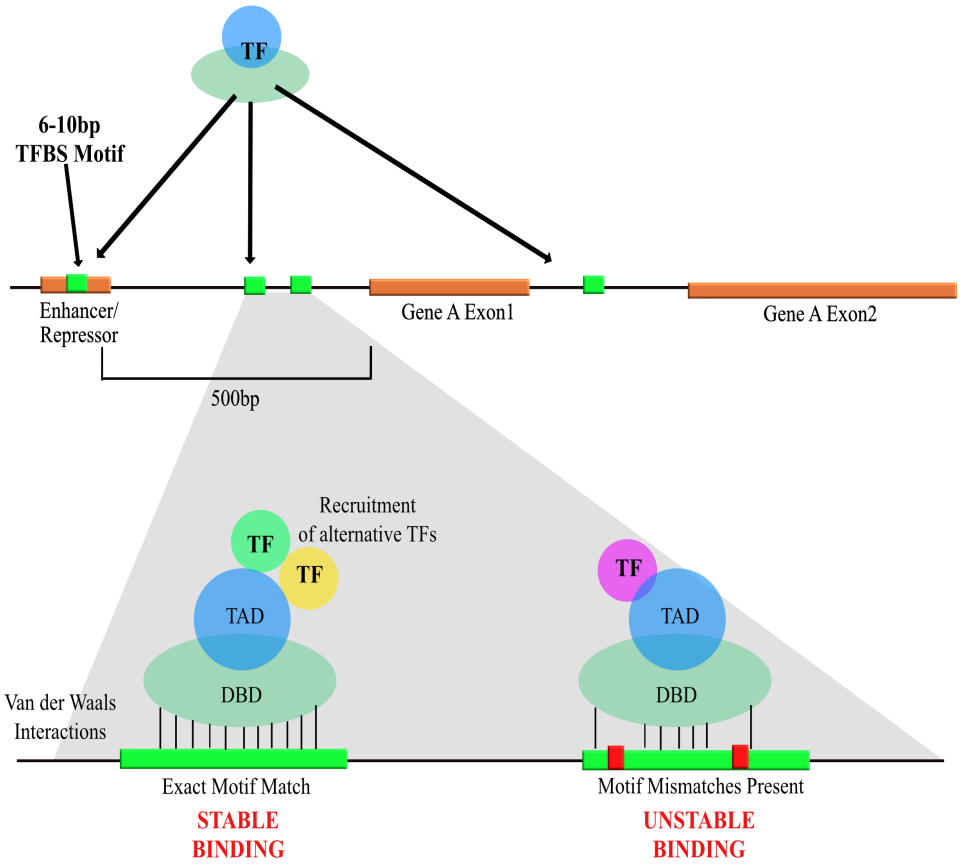


Figure 1.13: *Transcription factor binding in eukaryotes: Highlights TF binding to it's corresponding TFBS (green bar) and that TFBSs can exist in multiple locations and have the ability to facilitate TF binding even if mismatches are present, although stability is compromised. Diagram also depicts the function of both the DBD and TAD of TFs.*

Up and down-regulation of gene expression upon TF binding is determined by a RNAP's ability to access and bind to the gene of interest, this requires coordinated chromosomal compaction/decompaction by histones involving acetylation and deacetylation (Kolovos *et al.*, 2012). This is carried out by histone acetylase (HATs) and deacetylase (HDACs) enzymes (Figure 1.14) and this will be outlined in more detail in Section 1.5.2.3.

TFs can alter gene expression independently or in combination with other regulatory elements, these combinatorial units are known as *cis*-regulatory modules (CRMs) and their construction around promoter sequences are highlighted in Figure 1.15 (Blanchette *et al.*, 2006; Philips and Hoopes, 2008). Enhancer sequences are ~500bp to accommodate for CRM size and this region can potentially contain multiple TFBSs.

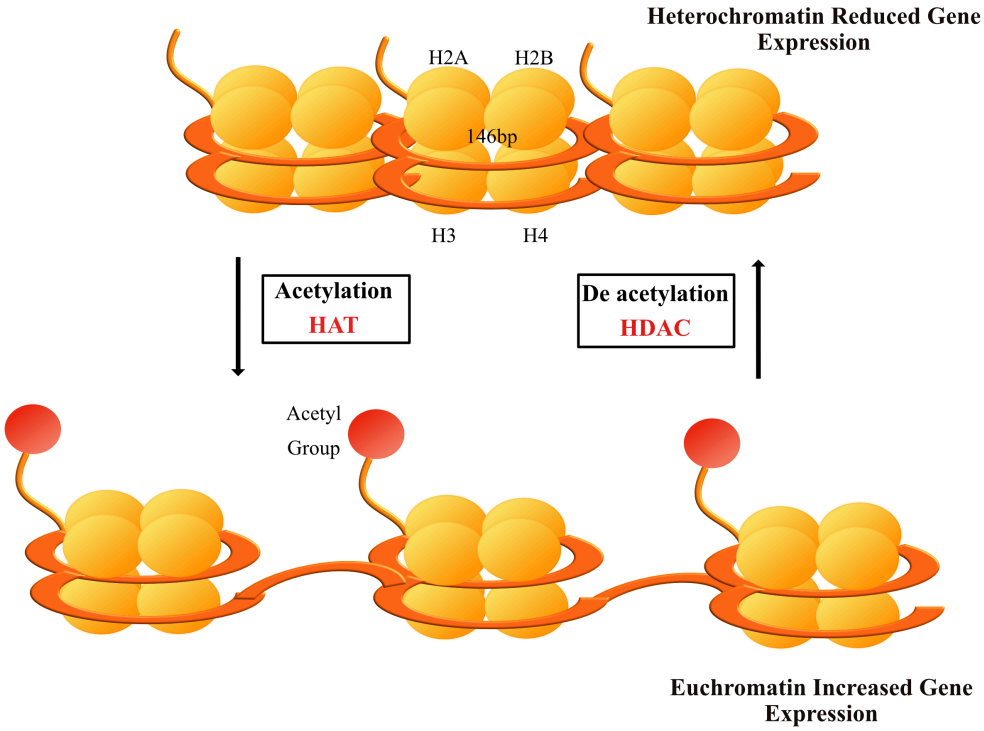


Figure 1.14: *An illustration of the histone modifications required for DNA compaction (heterochromatin) and decompaction (euchromatin). Histone decompaction is carried out by the addition of acetylation of histones by HAT enzymes, promoting gene expression through RNAP accessibility. Histone compaction is carried out by HDACs that remove acetyl groups, inhibiting RNAP entry thus limiting gene expression*

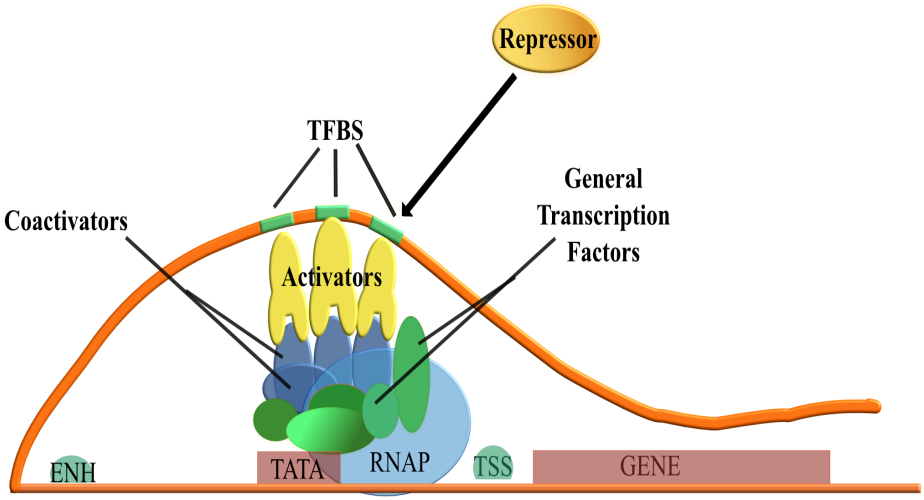


Figure 1.15: *Depiction of CRM construction on eukaryotic promoters that demonstrates the key relationships between general TFs (green), with coactivating (blue) and activator TFs (yellow). This relationship effects transcription efficiency.*

Interestingly, although vertebrate gene expression profiles and transcriptional output is consistent across species, the TFBS sequences across these species have unexpectedly low sequence conservation (Otto *et al.*, 2010). TFBSs across vertebrate species have been investigated and results have shown that across a panel of 51 genes in human and rodent only 32-40% TFBSs are identical across the two species (Dermitzakis and Clark, 2002). An analysis of the CEBPalpha TF across human, macaque, mouse, opossom and chicken revealed that only 50 TFBS were conserved across these species (Schmidt *et al.*, 2010). However, not all TFs follow this pattern, CTCF a genomic imprinting TF has 60-70% TFBS conservation among *metazoa* (Villar, Flicek and Odom, 2014).

The degeneracy of TF function explains some of their rapid turnover rate as the likelihood of such a short sequence being created by random chance is quite high (Tuğrul *et al.*, 2015). Theoretical models of neutral evolution have demonstrated that only a short time-scale is required for the acquisition of enough base pair substitutions to generate a new TFBS (Berg, Willmann and Lässig, 2004), this is in agreement with Kimura's theory of neutral evolution (Kimura, 1989). Furthermore, very few substitutions are required to improve upon the affinity of a mismatched low binding affinity consensus sequence. The dynamic nature of TFBSs sequence and surrounding sequence is a significant challenge to us understanding the impact of TFs on genome evolution however; TFs are not the only *cis*-acting element known to influence trancriptional control.

1.5.2.2) Splicing as a mechanism for transcriptional control

Splicing emerged in the eukaryotic lineage (Keren, Lev-Maor and Ast, 2010) and it is responsible for the removal of intronic sequences located between exons in pre-mRNA and the subsequent ligation of exons to produce a mature mRNA transcript potentially with the potential to produce a viable protein product (Berget *et al*., 1977). Constitutive splicing of a gene occurs when a gene's introns are spliced and exons ligated in the order that they appear in the pre-mRNA transcript (Wickramasinghe *et al.*, 2015). Alternative splicing (AS) occurs when the spliceosome includes/excludes specific exons creating a myriad of transcript isoforms (Wickramasinghe *et al.*, 2015). The evolution of splicing is

responsible for the expansion of the protein repertoire of multicellular eukaryotic organisms (Schmucker *et al.*, 2000) and has been associated with phenotypic differences between species (Xiong *et al.*, 2015).

Spliced genes generally produce a major isoform and minor isoform, the properties of which are detailed in Figure 1.16. Minor isoforms generally have tissue-specific expression profiles and have no greater/less functional importance than ubiquitously expressed, major, isoforms. For example, aberrant AS in the human SMN gene causes spinal muscular atrophy with expression restricted to only neuronal cells (Yoshimoto *et al.*, 2016). Isoforms produced with pre-mature stop codons and those subject to NMD are not necessarily devoid of function but may have a regulatory role (Mitrovich and Anderson, 2000; Cuccurese *et al.*, 2005; Peccarelli and Kebaara, 2014)

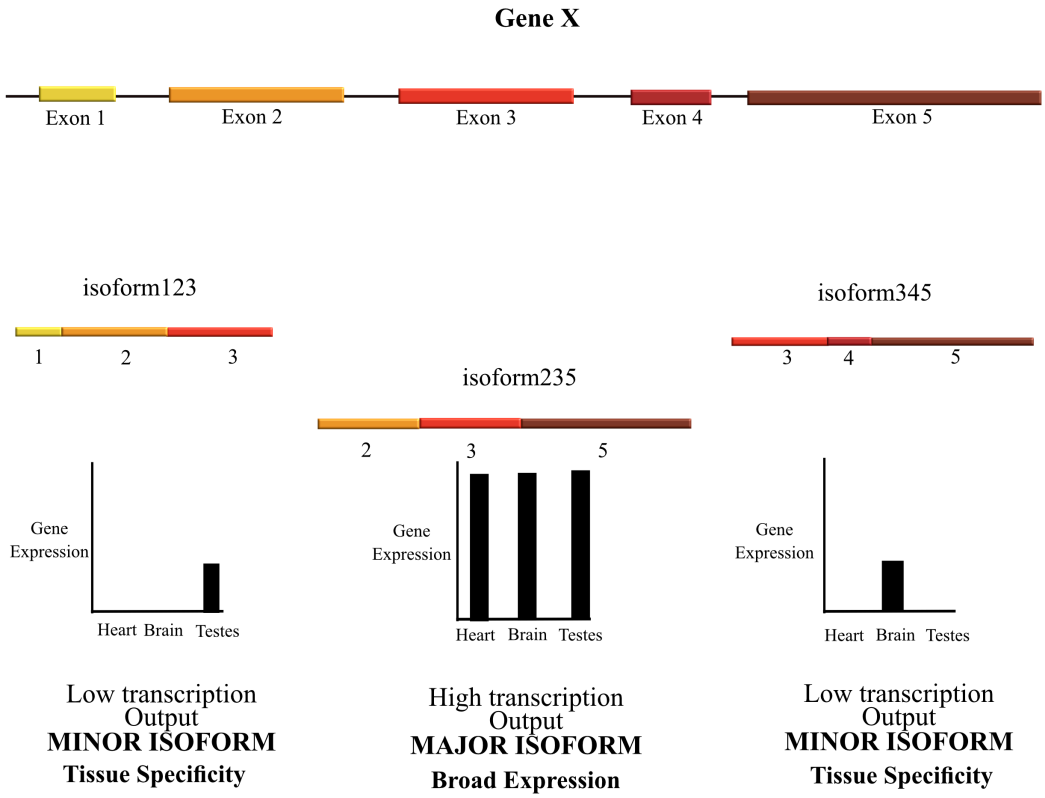


Figure 1.16: *Major and minor splice isoform expression profile characterization: The production of three splice isoforms from Gene X. Isoform123 and Isoform 345 are known as minor isoforms as they have a lower expression profile. These isoforms tend to have a tissue specific expression profile. Isoform 235 has a high, broad expression profile and is known as the major isoform* (Alekseyenko *et al.*, 2007).

The vertebrate spliceosome consists of five ribonucleoproteins that contain small-nuclear RNAs (U1-U6) along with ~150 other accessory proteins. The spliceosome itself is crucial for intron/exon definition requiring four main *cis*-sequences (splice signals). These splice signals are illustrated in Figure 1.17. Once assembled on the pre-mRNA transcript the spliceosome carries out two trans-esterification reactions (Figure 1.17) (Will and Lührmann, 2011). The efficiency of spliceosomal assembly on pre-mRNA sequences is affected by other *cis*-acting elements such as exonic splice enhancers/silencers (ESE/ESS) and intronic splice enhancers/silencers (ISE/ISS) (Fu and Ares, 2014) (Figure 1.17).

Unlike constitutive splicing, AS requires additional regulation in the form of SFs (Busch and Hertel, 2012). SFs bind to specific *cis*-motifs called SFBSs within pre-mRNA transcripts that can either prime repression or enhancement of splicing. Two main SFs control AS among eukaryotes, the serine-rich family, and the hnRNP family of proteins, The sum of either serine-rich or hnRNP proteins on a pre-mRNA determines spliceosome binding, for instance if increased levels of serine-rich SFs are present spliceosome assembly is more efficient and therefore the frequency of alternative isoform transcripts is increased (Busch and Hertel, 2012).

Serine-rich proteins contain extended serine and arginine dipeptides as well as at least one RNA binding domain (Busch and Hertel, 2012) and are known to increase exon inclusion upon binding (Figure 1.17). Conversely, hnRNP proteins repress AS generation through the blockage of spliceosome access by multimeric complexes construction. Both serine-rich and hnRNP families have expanded throughout eukaryotic evolution with human species containing more than ancient eukarya (Busch and Hertel, 2012).

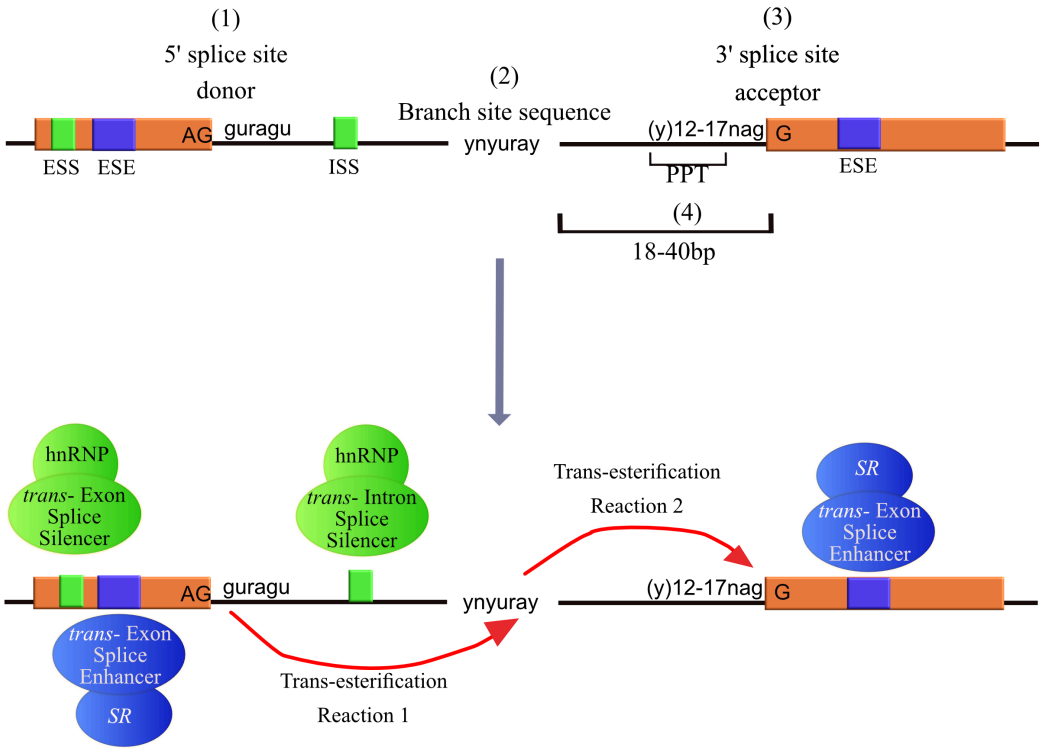


Figure 1.17: *An illustration of the cis-sequences required for accurate, controlled splicing in eukaryotes: Diagram highlights the 4 necessary cis-sequences required for accurate splice patterns ((1) 5' splice site donor, (2) branch site sequence, (3) 3' acceptor sequence and (4) polypyrimidine tract (PPT)) and the two trans-esterification reactions that are required to bring exons together. The graphic also depicts exon splice enhancers (ESE), intronic splice enhancers (ISE), exon splice silencers (ESS) and intron splice silencers (ISS) in controlling the rate of spliceosomal assembly across splice junctions (Busch and Hertel, 2012).*

1.5.2.3) The role of histone modifications in transcriptional control

Due to the expansion of the intergenic and intronic sequences in complex vertebrate genomes it's process of compaction inside the nucleus has in turn increased in sophistication (França *et al.*, 2017). The negatively charged nature of DNA sequence (Watson and Crick, 1953) allows for its coiling around positively charged, highly conserved histone molecules in order to create the nucleosome (McGinty and Tan, 2015) (Figure 1.14) which in turn forms chromatin. Chromatin can be decompacted to create regions of euchromatin allowing transcriptional machinery access or alternatively exist in its default compacted state inaccessible to transcription machinery – heterochromatin (Eissenberg and Elgin, 2014).

In order to switch between states chromatin remodelers are required and five families have been identified (Ho and Crabtree, 2010). All families contain ATPase activity that acts on a sequence 40bp inside of the nucleosome that provides the torque necessary to pump DNA around nucleosomes (Längst and Manelyte, 2015). Each family also contains a nucleosome specificity sequence that restricts specific family members to specific nucleosomes – nucleosome selection (Längst and Manelyte, 2015). These remodelers are added in a stepwise manner in order to localise euchromatic regions only to genes under active transcription (Tyagi *et al.*, 2016).

Although histones are highly conserved sequences, they contain N' terminal tails (Biswas *et al.*, 2011) that can undergo covalent modification by functional groups such as acetylation, methylation, phosphorylation, and ubiquitination to nitrogen group in lysine R groups – "the histone code" (Jenuwein and Allis, 2001). The level of modification alters the rate of transcription in the region and Table 1.4 illustrates each modifications impact on chromatin structure. However, whether each histone modification is monolacetylated/monomethylated or triacetylated/methylated can also contribute to a transcription status alteration. Due to technological advancements the wealth of transcriptomic data becoming publically available is increasing and therefore our understanding of the roles of each *cis*-regulatory element is becoming more clear.

Table 1.4: The 'Histone Code' and its effect on chromatin conformation

Histone	Site	Modification	Transcription Consequence
H1	Ser27	Phosphorylation	Activation
H1	Lys26	Methylation	Repression
H2A	Lys5	Acetylation	Activation
H2B	Lys12	Acetylation	Activation
H2B	Lys15	Acetylation	Activation
H3	Lys9	Acetylation	Activation
H3	Lys14	Acetylation	Activation
H3	Lys23	Acetylation	Activation
H3	Lys18	Acetylation	Activation
H3	Lys23	Acetylation	Activation
H3	Lys27	Acetylation	Activation
H3	Lys4	Methylation	Activation
H3	Lys9	Methylation	Repression
H3	Lys27	Methylation	Repression
H3	Lys36	Methylation	Activation
H3	Ser28	Phosphorylation	Activation
H4	Lys5	Acetylation	Activation
H4	Lys8	Acetylation	Activation
H4	Lys12	Acetylation	Activation
H4	Lys16	Acetylation	Activation
H4	Arg3	Methylation	Activation
H4	Lys59	Methylation	Silencing

Provides an insight into the role of histone modifications in chromatin structure. Column 1 and 2 depict the histone and the position of potential modifications. Column 3 and 4 highlight the modification carried out and the impact it has on local chromatin structure.

1.5.2.4) Technological advancements in gene expression profiling

It is clear that the sophistication of transcriptomic analysis technologies has been a gradual process beginning in 1953 with Francis and Crick's identification of the DNA molecule and it's structure (Watson and Crick, 1953) through to today where a whole human genome can be sequenced routinely in a hospital environment in just 26 hours (Farnaes *et al.*, 2018) at a continuously reducing cost (Wetterstrand, 2016). Instead of dwelling on more classical transcriptomic technologies such as expressed sequence tagging (EST) (Adams *et al.*, 1991), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995) and RT-qPCR (Nolan *et al*., 2006) a brief description on the Sanger sequencing and Maxam-Gilbert methods will be given as a premise of the sequencing technique followed by more recent methods. Platforms are split into three categories first, second and third generation sequencing technologies. The first sequencers were based on the manipulation of fundamental biological research, beginning with Sanger (Sanger *et al*., 1977) whom discovered the ability to radiolabel DNA sequence and separate these sequences by 2 dimensional fractionation methods. On this basis the first sequencing methods were developed by Sanger (Sanger *et al*., 1977) and Maxim Gilbert in 1977 (Maxam and Gilbert, 1977). Sanger's chain terminating technique was based on the initial radiolabelling of nucleotides and their addition to DNA sequences of interest in the presence of DNAP causing cessation of the replication process. Ceased fragments could then be separated by polyacrylamide gel electrophoresis (Figure 1.18) (Sanger *et al*., 1977). Maxam and Gilbert followed this technique but instead of DNAP manipulation depended on chemicals for specific polymerisation cessation (Maxam and Gilbert, 1977). Hydrazine cleaved pyrimidine bonds (C and T), acids cleaved purine bonds (A and G), dimethyl sulphate cleaved G specifically and hydrazine in high Na+ concentrations specifically cleaves C residues (Maxam and Gilbert, 1977). These techniques dominated for over 30 years prior to the advent of 2nd generation sequencing platform (Figure 1.19).

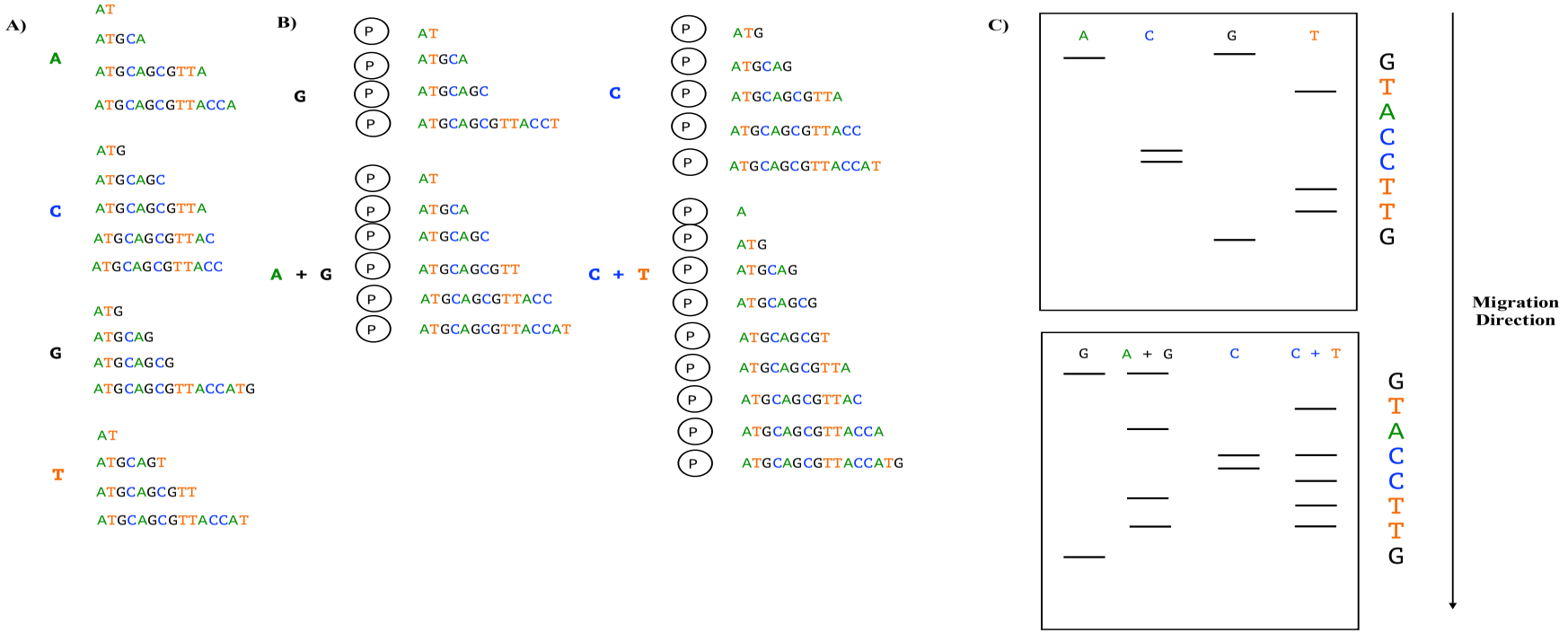


Figure 1.18: *An illustration of DNA sequencing used by the two first generation sequencing methods. A) Illustrates Sanger's chain termination sequencing after incorporation of a radiolabelled dNTP and subsequent separation by a polyacrylamide gel (top(C)). B) Depicts the Maxam-Gibert sequencing mechanism whereby (P) illustrates a P^{32}-tag attached to the 5' end of each strand. Chemicals are added to solution in order to cleave bases at specific sites and again fragments are obtained using the separation mechanism (bottom(C)).*

Second generation sequencing platforms are defined as those reliant on amplified sequence banks for the generation of low cost, small reads that do not require gel separation techniques (Kchouk *et al*., 2017). The magnitude of this advancement is highlighted when compared to the Sanger sequencing of the human genome in 2003 (Consortium, 2001; Venter, 2001) taking almost 15 years and costing ~1 billion dollars (Bentley, 2006) while today 454 sequencing platforms requires only 2 months and cost just 100th of the price. Second generation sequencing platforms include Roche's 454 sequencer (Margulies *et al.*, 2005) and the ion torrent sequencing platform (Boland *et al.*, 2013). The 454 sequencer produces reads of >1000bp in length at ~ 1million reads per minute. The method randomly fragments input DNA and that bind to microbeads covered in DNA specific primers. Emulsion PCR, of each bead using radiolabelled nucleotides is utilised with resultant light production indicative of successful nucleotide binding which can be monitored spectrally (Figure 1.19) (Margulies *et al.*, 2005). The second method, ion torrent sequencing, does not rely on fluorescent tags or light production (Figure 1.19). Instead the DNA sequences are amplified and on nucleotide binding hydrogen ions are released. Hydrogen production changes the pH of the underlying solution and can be monitored sensorally (Boland *et al.*, 2013). Although making significant advancements they are limited in their ability to call INDEL sequences and in dealing with the repetitive units existing in complex genomes (Kchouk *et al*., 2017)

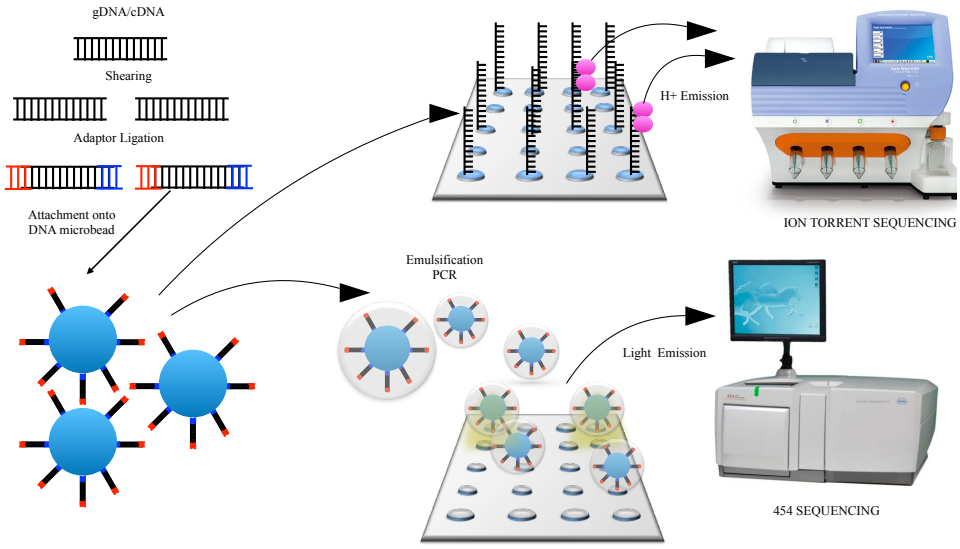


Figure 1.19: *A graphical representation of the 454 sequencing platform and the Ion torrent platform: A depiction of the methodology behind two 2nd generation sequencing technologies – Roche's 454 sequencing platform (bottom) and the Ion Torrent sequencing platform (top).*

3rd generation sequencers were developed to deal with the pitfalls of 2nd generation sequencers. They do not require PCR amplification thus saving on time, labor and expensive laboratory materials (Kchouk, Gibrat and Elloumi, 2017). Their read length of > 60kbps aids dealing with repetitive genomic units. 3rd gen platforms are based on one of two methods:

1) Small molecule real time (SMRT): Single molecule real time sequence production (Roberts *et al*., 2013).
2) Synthetic approach: Utilises pre-existing short read data to construct longer reads (Metzker, 2010).

SMRT approaches are the most commonly used 3rd gen platform, examples include PacBioscience's PacBio sequencer (Rhoads and Au, 2015) and Oxford Nanopore system (Lu, Giordano and Ning, 2016). PacBio is based on a plate of micro-fabricated pores or zeromode wave guides (ZMG) each containing a DNAP. The process itself relies on the binding of fluorescently labeled nuleotides to the ZMG, which generates light in, real-time (Figure 1.20). Sequencing of an entire whole genome can take just ~4-6 hours (Rhoads and Au, 2015).

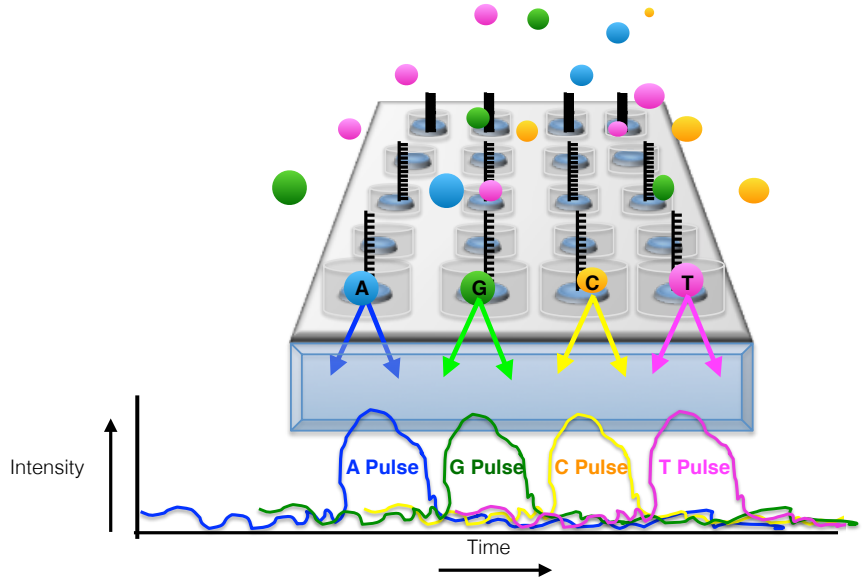


Figure 1.20: *An illustration of Pacific Bioscience's third generation PacBio system: An illustration of the PacBio sequencer. Balls graphically represent fluorescently labelled dNTPs, with blue illustrating 'A' nucleotides, green depicting 'G' nucleotides, orange representing 'C' nucleotides and magenta showing 'T' nucleotides. On binding each nucleotide emits a wavelength specific to its fluorescent tag 's colour, which is measured by the PacBio system.*

The Oxford Nanopore sequence platform is just 4- 6 inches in size. Upon addition of DNA sample, binding occurs to complementary sequences on the device in a hairpin fashion (Lu, Giordano and Ning, 2016). Bound sequences are passed through a protein nanopore with each individual nucleotide emitting a specific torrent that can be monitored in real time. This mechanism is both time and cost efficient as each platform contains 48 flow cells with 3000 nanopores each and yields 500bp per second (Kchouk *et al*., 2017). These technological advancements allow more accurate sequencing of cells transcriptional output and therefore is key to RNA based comparative genomic analyses. Through these technological advancements transcriptomic profiles now exist of organisms spanning the entire phylogenetic tree of life and therefore comparative transcriptome analyses between species will aid our understanding of transcriptional control evolution.

1.5.2.5) Lineage-specific gene expression in vertebrates

Many cases of novel gene expression patterns have been identified in homologous genes across vertebrate species (Sudmant *et al*., 2015). Differences in these profiles are likely to contribute to observed phenotypic changes (Uebbing *et al.*, 2016) and the cause of profile disparities between lineages has been linked to differential regulatory element such as DNA methylation (Laqqan and Hammadeh, 2018), histone modifications and regulatory RNAs (Thiel *et al.*, 2017).

The significant role for regulatory elements in gene expression pattern alterations between species has long been established, for instance a TF binding analysis of five vertebrate livers (human, macaque, mouse, rat and dog) uncovered differential transcriptional patterns with only 35/30,000 binding sites shared across all 5 species (Schmidt *et al.*, 2010). In humans specifically, an elevation in TFBS polymorphism rate correlates to both the loss and gain of gene expression (Tuğrul *et al.*, 2015). For example, an investigation of human and chimpanzee methylation profiles between sperm and neutrophil cells has revealed that methylation caused 12-18% of expression differences found (Hernando-Herraez

et al., 2015) and that 7% of this is due to, H3K4me3 histone modification specifically (Cain *et al.*, 2011).

The challenging nature of identifying regulatory elements that correlate with phenotypic change is now being addressed through technological advances such as microarrays (Narlikar and Ovcharenko, 2009), RNA sequencing (Smith *et al.*, 2012) and CHIP sequencing (Ren and Dynlacht, 2004; Kim and Ren, 2006; Barski *et al.*, 2007). Examples of the application of NGS technologies include the identification of a BMP4 gene regulatory factor and its direct role in beak shape and size in Darwinian finches (Abzhanov *et al.*, 2004), FOXP2 and language development (Becker *et al.*, 2018) and a deletion upstream of AR receptor gene causing the loss of penile spines in humans (McLean *et al.*, 2011).

In summary, pre-existing genes have the potential to gain novel expression patterns through alterations in DNA methylation patterns, histone modifications and mutations in transcription factor binding sites. However, new genes require transcriptional machinery acquisition before their output can be regulated in this manner.

1.5.2.6) The acquisition of expression profiles in new genes

New genes are generated by the mechanisms outlined in Section 1.2 and they have the potential to change the genotype and phenotype of the host organism (Chen *et al*., 2013). However, for this to occur new genes must first acquire the neccesary machinery for RNA and protein production.

The expression profile of a new gens can occur by the acquisition or co-option of pre-existing and highly conserved regulatory elements (Peter and Davidson, 2011), e.g. repressors, proximal enhancers, methyl groups, and/or boundary elements (Kaessmann, 2010). An optic lobe gene (*NEP1*) in *Drosophila*, gained expression through the co-option of a pre-existing enhancer element followed by four point mutations that increased the gene's expression level (Glassford and Rebeiz, 2013). A chromosomal inversion of the tinman gene complex in aphid genomes caused the relocation of the ladybird enhancer element (a gene within

the complex) into a novel location adjacent to a gene thus causing ectopic, elevated expression of this gene (Cande *et al*., 2009).

Through gene-gene interaction networks (GGIs) new genes in human have been predominantly associated with a narrow spatiotemporal expression profile having having a low number of interacting partners (Zhang *et al.*, 2015). In contrast, more ancient genes had an increased number of interacting partners, increasing the probabaility of pleiotrophic functionalities (Zhang *et al.*, 2015). Interestingly, it has been proposed that over time new genes increase their interaction level within the GGI leading to an increased expression breadth and pleiotrophy (Zhang *et al.*, 2015). These findings support the 'out-of-testis' hypothesis which states that the testis is the initial location where new genes are expressed before they functional acquisition and fixation (Kaessmann, 2010).

Exceptions to this limited expression profile of new genes have been demonstrated in analysis of new genes in human and mouse, where 5,000 cases had increased interaction levels (Chen *et al*., 2013) with links to alzheimers, fetal brain development and neuronal connectivity. The origin of these new genes co-occur with the expansion of the neocortex, specifically the lateral prefrontal cortex in primates (Zhang *et al.*, 2015) and were found genomically different to those new genes containing low centrality values due to: 1) an increased number of low complexity regions and 2) containing intrinsically disordered regions (Zhang *et al*., 2005). These genomic factors implicate a protein's flexibility, adaptability and ability to form interacting pairs and could explain their selective advantage in fetal brain development during neocortex expansion (Zhang *et al.*, 2015).

New gene expression profiles have been shown to significantly contribute to transcriptional output between human and chimpanzee where they underpin 2-8% of the expression difference between the species (Blekhman *et al*., 2009). Of course, genes (new or otherwise) are not only regulated at the transcriptomic level but also during translation.

1.6) The Mechanics of Translation

In Section 1.5 the importance of transcription on genome evolution is extensively discussed. Although the RNA output of each cell enhances our understanding of cell and tissue functions a translatomic profile is also required as mRNA output does not equate to protein output in most cases *e.g.* 'the hidden transcriptome' (Dinger *et al.*, 2008).

Translation is the process whereby mRNA transcripts are converted to amino acid polypeptide chains. There are 61 codons responsible for mRNA to amino acid (20 amino acids) conversions, 3 of which are stop codons (UAG, UAA and UGA) and one initiation codon (AUG) (Hinnebusch, 2014; Rozov *et al.*, 2016). Translation ensues in a 5' to 3' direction with the binding of a specific transfer RNA (tRNA) containing a complementary codon sequence (anticodon) to the mRNA (Kringelbach *et al.*, 2007). The tRNA is charged with a corresponding amino acid by an aminoacyl tRNA synthesase enzyme (5' end of ATP bound to COOH+ of amino acid). Specific tRNAs are responsible for the conversion of a specific mRNA codon template into a specific amino acid output and therefore aminoacyl tRNA synthesase enzymes only bind to 'isoaccepting' tRNAs (Kringelbach *et al.*, 2007). As there are 61 codons one may think 61 tRNAs and 61 synthases are required for translation however position 3, "the wobble position", of each codon allows for translation to be carried out by only 32 tRNAs and 20 synthase enzymes in human (Rozov *et al.*, 2016).

A charged tRNA has the ability to bind to an mRNA template only with the aid of the ribosome. The eukaryotic ribosome is composed of a small and large subunit (40S and 60S) as well as >79 regulatory proteins (Wilson and Cate, 2012). The ribosome complex is conserved across the tree of life however as organisms became more complex, expansion segments were required and so the ribosome of human is much larger than that found in bacteria. The ribosomal protein itself contains 2 catalytic sites 1) the P or peptidyl site and 2) the A or aminoacyl site that allow for the loading of a charged tRNA complementary to the bound mRNA (Wilson and Cate, 2012). After successful binding the ribosome's catalytic centre (peptidyl transferase) catalyses a di-peptide bond

between the two tRNAs loaded into the P and A site and the beginning of polypeptide chain production ensues. (Wilson and Cate, 2012).

Translation can be split into three distinct phases 1) initiation 2) elongation and 3) termination. Translational initiation requires 3 major components an mRNA, the small ribosomal subunit (40S), and an initiation or start codon (AUG). During initiation these components come together with the aid of eukaryotic initiation factors (eIFs) 1, 2, and 3 to form a pre-initiation complex (PIC). The PIC contains a bound small ribosomal subunit at the mRNA start codon and a P-site occupied by an initiator tRNA ($tRNA^{fmet}$). Correct placement of the complex on the mRNA transcript is not solely determined by the TSS, but also by an upstream Kozak sequence motif (5-ACCAUGG-3), and by the 5'cap of the mRNA that is attached post-transcriptionally. Once the complex identifies the 5' cap it binds, moves in a 3' direction until an initiation codon is reached – this is known as the 'scanning mechanism' and is aided by eIF3 and eIF4 (helicase activity). Once the initiation codon is reached eIF2 hydrolyses GTP to GDP causing the release of an inorganic phosphate allowing the $tRNA^{fmet}$ into the ribosomal subunit's P-site. eIF1 is then responsible for ensuring that the PIC has been positioned correctly, further stabilising only those with correct placement. After this eIFs dissociate completely allowing access for the larger ribosomal subunit to bind causing eIF2 hydrolysis and initiation arrest (Andreev *et al.*, 2017).

Translation can then enter its second phase, elongation. Elongation occurs after $tRNA^{fmet}$ loading into the P-site, leaving the A site empty and available to bind tRNAs with a complementary sequence to that of the codon of mRNA to which it is attached (Dever and Green, 2012). After this the ribosome's peptidyl transferase enzymes catalyse the formation of a di-peptide link between the initiation methionine amino acid in the P-site and the amino acid in the A-site (Dever and Green, 2012). This catalysis is a GTP dependent process and with the inorganic phosphate release the initiator $tRNA^{fmet}$ leaves the ribosome through the exit or E site giving the tRNA in the A site it's methionine amino acid (Dever and Green, 2012). The tRNA in the A site then shifts to the P site and the mRNA

translocates by 3 residues allowing a new complementary tRNA to enter the A site and a tri-peptide linkage to form between it's charged amino acid and amino acids found in the P-Site. This cycle continues until a termination signal is found (Dever and Green, 2012).

There are 3 codons that signal for translational termination. These are uncharged codons that bind the P site and signal for GTP dependent release factors that cleave the polypeptide allowing it's release from the ribosome through ribosomal dissociation (Dever and Green, 2012). Similarly to transcriptomic data, translatomic-profiling data is also increasing due to the development of increasingly sophisticated next generation profiling technologies.

1.6.1) A brief overview of translatome profiling technologies

Transcriptional analysis of mRNA is not always sufficient in uncovering a cells protein or translational output as mRNA level does not always directly equate to the production of a protein product (Liu *et al*., 2016). With this in mind translational profiling techniques are a requirement for a more accurate insight into a cells translational profile. From Section 1.6 is clear that the importance of translation cannot be underestimated. It's large consumption of the cells energy requirements (Pascal and Boiteau, 2011), ability to undergo global deactivation in response to a stimulus at a rate quicker than a new mRNA creation along with it's tight but dynamic regulation all demonstrate this significance (Mami and Pallet, 2015). The dynamic nature of translation regulation allows each cell to remodel their proteome with both speed and accuracy to adopt different cellular fates in response to differentiation and developmental processes but also makes analysing (Spriggs *et al*., 2010) and obtaining an accurate snapshot of translation quite challenging due to the speed of translation adaption to its environment (López-Maury *et al*., 2008; Gerashchenko *et al*., 2012). With this in mind it is important to select the most appropriate translational profiling tool to address the question of interest. Here, a brief introduction to some of the classic and more recent translation profiling tools will be given.

1.6.1.1) Polysomal Profiling

Polysomal profiling is a classical translational profiling method developed in the 1960s (Chassé *et al.*, 2017). This technique can be used on an individual transcript level by 1) investigating specific proteins bound to mRNA transcripts e.g. ribosome binding proteins (RBPs) bound to untranslated region (UTR) sequences 2) by investigating non-coding mRNAs that can repress translation of mRNA transcripts and lastly 3) by investigating *cis*-acting binding sequences that prevent ribosome binding (Chassé *et al.*, 2017). Polysomal profiling also allows global/holistic profiling of translation though investigating phosphorylation or general modifications of the factors involved in the translational process e.g. investigation of phosphorylation of eIFα as it represses active initiation complexes (DuRose *et al.*, 2009). The mechanism itself can be paired with deep sequencing technologies, RT-qPCR or Northern Blotting technologies to answer a myriad of questions about a cell's translation profile (Chassé *et al.*, 2017).

The technique utilises the basic premise of translation, the requirement of an mRNA transcript to bind with a ribosome in order to undergo active translation. Profiling begins with the freezing of active translation typically carried out by inhibitors of translation such as cyclohexamide (CHX) (Duncan and Mata, 2017), or a 96 well purification processor (Chassé *et al.*, 2017), or even by freezing the cells with liquid nitrogen followed by cryogenic grinding (Rhoads, Dinkova and Jagus, 2007). The cell complement can be loaded into a 4-15% sucrose gradient and separated by ultracentrifugation, separating mRNAs containing monosomes (small ribosomal subunits) that are unlikely to undergo efficient translation from those that contain multiple ribosomes *i.e.* polysomes (Upper fraction) that have an increased likelihood of active translation. After separation the fractions are analysed using a spectrophotometer at A_{524} (Kleene *et al.*, 2010).

Although the technique has advanced our knowledge about the translational profiles of cells and about translation mechanisms in general there are some limitations to the technique. These include the utilization of sucrose gradients as they are fragile and even the most careful, skilled researcher could disrupt the fractionate. Unfortunately, it is impossible to know whether a sample has been

disturbed or not until the end of the analysis thus wasting time, sample and expensive reagents (Edelstein *et al*., 1984). Another limitation includes equipment expenses and if incorrect or damaged equipment is used it can profoundly affect the gradient. The technique is incredibly labour intensive, not many samples can be processed simultaneously, samples are prone to contamination and pairing with deep sequencing techniques requires large sample sizes (Head *et al.*, 2014).

1.6.1.2) Ribosomal Affinity Purification Techniques

Ribosomal affinity purification techniques (TRAPs) allow for the accurate identification of translation profiles for any cell of interest in comparison to more classical techniques. In comparison to more classical techniques this mechanism can distinguish even between tissues that are generally indistinguishable and intermixed *e.g.* the somodendritic tissue containing straitopallidal medium spiny neurons and straionigral cells through ribosomal techniques obtained both cell types translational profile (Heiman *et al.*, 2008). The mechanism itself is based on the fusion of an enhanced green fluorescent protein (GFP) to the 5' end of the largest subunit of the ribosome (10a) and the subsequent attachment of the promoter sequence for the gene of interest. After this the engineered eGFP gene can be placed on a bacterial artificial chromosome (BAC) and ribosomes expressed that preferentially attach the gene of interest promoter binding sequence, causing active translation of the gene of interest (Heiman *et al.*, 2014). The extraction process is carried out by anti-eGFP magnetic beads (contain eGFP antibodies) that are selectively extracted from the cell lysate using magnets (Heiman *et al.*, 2014).

A major advantage of this technique is the ability to tag any genetically pre-determined cell population. It also negates the need for microdissection, cell panning, sorting, fixation or single cell suspensions therefore reducing the time cells are under stressful conditions thus increasing the likelihood of a more accurate snapshot of the unadapted cell's translatome. TRAP has been used successfully in a variety of contexts including the study of the aging process

through neuronal cell translational profiling that identified a potential modulator of Huntington's disease (Heiman *et al.*, 2008).

1.6.1.3) Ribo-tRNA sequencing

Although the determination of a cells translational output through analysis of both specific and global mRNA/ribosome usage gives a great insight into the mechanics of a cells transcriptional and translational profile and subsequent function other aspects of the fundamental process behind translation can be exploited to further our understanding of translation. By investigating the tRNA both inside and outside ribosomes tRNA usage and biases can be analysed and compared across different tissue types, or cells under different environmental conditions (Chen and Tanaka, 2018). The importance of understanding a cells tRNA complement both in the cytosol and ribosome of a given cell is highlighted in examples such as:

1) tRNA and ribo-tRNA profiles between healthy and environmentally stressed cells have shown evidence for differential tRNA usage patterns.
2) Stressful environmental conditions increase the probability of finding either damaged or uncharged tRNAs within the ribosome (Chen and Tanaka, 2018).
3) tRNA and anticodons do not have a direct 1:1 correspondence due to the redundant nature of translation and the "Wobble-position" discovered by Crick in 1966 (Crick, 1966). Differential usage of anticodons has been documented across cells under different environmental conditions and cells types. This level of knowledge can not be determined by mRNA based techniques (Chen and Tanaka, 2018).

To date, tRNA usage is vastly understudied and has a decreased amount of technologies available for its assessment however, both Ribo-tRNA-sequencing and tRNA-sequencing platforms are next generation tools developed to investigate this important component of translation (Kawashima *et al.*, 2009). Both techniques can be paired with and run simultaneously alongside mRNA

sequencing therefore obtaining a more holistic analysis of the translational profile of a cell at the monosomal level.

The technique requires the capture of tRNA inside actively translating ribosomes. Similarly to previously describe mechanisms CHX can be used to freeze translation and the cell complement can then be passed along a sucrose gradient. Here, the 2 mechanisms differ through RNAse treatment of the cell fraction. If no RNase treatment is carried out the upper fraction of the sucrose gradient contains tRNAs of 70-110amino acids in length, this sample is used for tRNA sequencing profiling. However, if the cell fraction is treated with RNAse these cytosolic tRNAs become degraded leaving only those inside an actively translating 80S ribosome. This sample is utilised for Ribo-tRNA sequencing (Chen and Tanaka, 2018).

1.6.1.4) Ribosome Profiling

The aim of ribosomal profiling is to measure ribosomal occupancy and translation on a genome-wide level through the deep sequencing of ribosomally protected mRNA fragments (Ingolia *et al.*, 2012). This technique has three major benefits including ascertaining mRNA abundance, precise delineation of translated regions (almost to base pair precision) as well as providing information on ORFs and frameshifts (Ingolia *et al.*, 2012).

Again these features of translational profiling highlight its power and should be paired with transcriptional based analytical tools. Similarly to previously described translational profiling tools ribosomal profiling/footprinting/ or ribosequencing is a ribosome-centric based approach that exploits the occupancy of ribosomes on an mRNA transcript during active translation. The premise behind the technique exploits the fact that changes in ribosome occupancy level can manifest as changes in translation i.e. an increased ribosomal density suggests more efficient translation.

The technique requires the degradation of all nucleic sequences with the exception of the 21-30 nucleotide mRNA sequences that are ribosomally bound

or protected (footprints). Each footprint is then converted to cDNA using a reverse transcriptase enzyme and primer sequences are added (Gobet and Naef, 2017). The addition of primers aids the ligation and circularisation of the footprint in order to reduce sequence based capturing biases prior to amplification. The technique is then paired with a next generation sequencing platform and reads aligned to a reference genome to obtain the ribosomal profile (Gobet and Naef, 2017) (Figure 1.21)

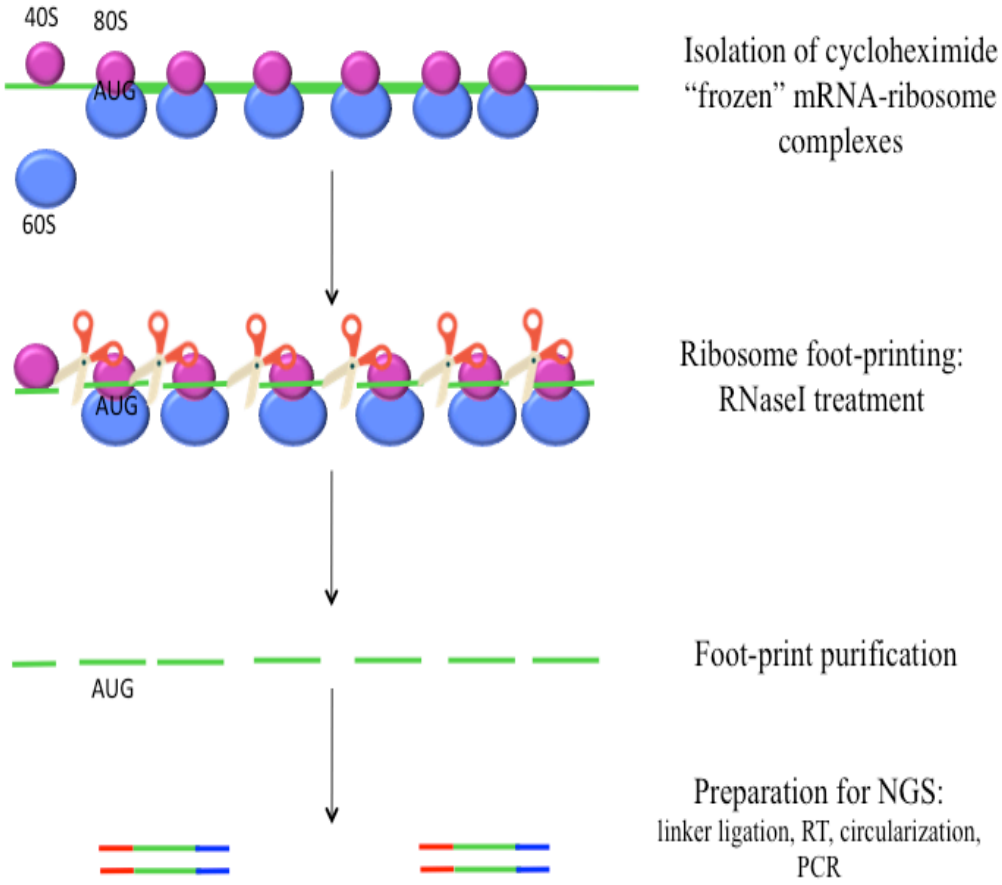


Figure 1.21: *A graphical representation of the ribosomal profiling mechanism: A depiction of the ribosomal profiling technique from freezing active translation of mRNA transcript to the preparation of each read for NGS sequencing* (Ingolia, 2016).

There have been many unanticipated discoveries made by ribosomal profiling including: many mRNA transcripts have more than one initiation codon and that the annotated start site is, in many cases, not even the major start site for active transcription/translation (Jackson and Standart, 2015). Furthermore, transcripts have been identified with alternative N- terminus isoforms *i.e.* with many different terminating codons resulting in longer or truncated mRNA/proteins – 'leaky' translation (Jackson and Standart, 2015). As well as this nineteen examples of protein-coding sequences with ribosomes bound on both strands of mRNA transcripts were uncovered in *S. pombe* (Jackson and Standart, 2015). Lastly, the biotin labelling of ribosomes in close proximity or even in the endoplasmic reticulum or mitochondria has been identified as a powerful tool used for ribosome sub-population separation (Jackson and Standart, 2015).

These four technologies are only the tip of the iceberg and there are many up and coming techniques now becoming extremely powerful in the investigation of both translational profiles and its control. These include: HydroSeq a tRNA profiling tool that can more efficiently deal with the challenging secondary structure of tRNAs (Arimbasseri *et al.*, 2015) and both AlkB-facilitated RNA methylation sequencing (ARM-seq) (Cozen *et al.*, 2015) and DM-tRNA sequencing methods improve profiling by using the *E. coli* dealkalizing enzyme, AlkB (Clark *et al.*, 2016). It is obvious that these technologies have provided great insights into translational protocols however, there limitations must also be considered. The use of inhibitors such as CHX to freeze active translation within cells is a double-edge sword as although active translation is stopped ribosomes accumulate at remaining mRNA initiation start sites potentially resulting in a distorted image of translational output. The impact of CHX on translational profiles has been examined with a gain of 100% ribosome occupancy of vacant initiation sites in under 10 seconds (Duncan and Mata, 2017). Although the tools are now available to accurately acquire high quality sequence data, the statistical tools used to interpret these data are lagging. As with most bioinformatics analyses sample replicates are essential, particularly as next generation sequencing technologies simultaneously analyse thousands of genes. Due to the wealth of data per gene best fit models are not currently possible however single

error models seem an adequate compromise in comparative analyses (Ingolia *et al.*, 2012).

The future of profiling techniques needs to focus on translation events, reduce contaminated samples and the development of additional technologies that accurately decipher between active translation and ribosomal occupancy. Other aspects of translation that require additional understanding include the efficiency of ribosomal release after reaching the stop codon, the length of ribosomal footprints and the 3 nucleotide periodicity of translation.

1.7) Application of network theory to understanding protein evolution

1.7.1) Graph theory and its use in biological data

Graph theory is centred around the mathematical analysis of the properties of graphs based on relationships that exist between objects. The power of graph theory has been demonstrated in a plethora of research fields, including the expansion of the World Wide Web (Sun *et al.*, 2012), social networks (Granovetter and Granovetter, 1983), food webs (Dunne *et al*., 2002), epidemiology (Welch *et al*., 2011) and many more. Mathematical graphs are an incredibly powerful tool largely underutilized in the field of biological sequence data analytics. Although the power of these methods is beginning to grow in niche areas such as protein-protein interaction analysis (Raman, 2010) and metabolic pathway analysis (Jonnalagadda and Srinivasan, 2014) the application of networks to DNA sequence analyses is still lagging behind.

Graphs, or networks, are a set G = (V, E), whereby V denotes vertices and E edges. Edges are drawn between two vertices to represent a specific relationship between the two entities (Huber *et al.*, 2007) (Figure 1.22). Graphs can either be directed or undirected. In a directed graph the edges illustrate the direction of the relationship between two vertices and relationships depicted by edges do not have to be reciprocal, e.g. vertex A may have a relationship with vertex B but B may not have a relationship with A (Pavlopoulos *et al*, 2011). Thus an edge will be drawn with an arrow pointing from A to B, but none from B to A. Undirected edges do not have this requirement and in Figure 1.22 (a) edges are drawn

without relationship direction being considered (Pavlopoulos *et al.*, 2011), (Jachiet *et al.*, 2013).

As the usage and complexity of graphs in biological research is becoming more popular a number of unexpected network properties of biological data have come to light. In general graphs are expected to have a Poisson or Gaussian distribution where all vertices in the graph are close to that of the mean degree. For instance the Erdos-Renyi model of random graph generation is based on a Poisson distribution (Barabasi *et al*, 2004). It is works by selecting a random number of nodes (*N*) and connecting these nodes using a series of random edges. These random edges are placed based on probabilities (*p*) of connections existing between N calculated by $pN(N-1)/2$ (Barabasi *et al*., 2004). However, networks created from biological data tend to have unexpected graph topologies that are scale-free in nature with a power-law based distribution. This means that a low number of vertices have many edge connections or high degree, and these are known as hub nodes (Figure 1.23). This is in direct contrast to the majority of vertices in the graph have a low degree (Zhu *et al*., 2007). Graphs of a scale-free nature have a disproportionate amount of nodes with respect to degree

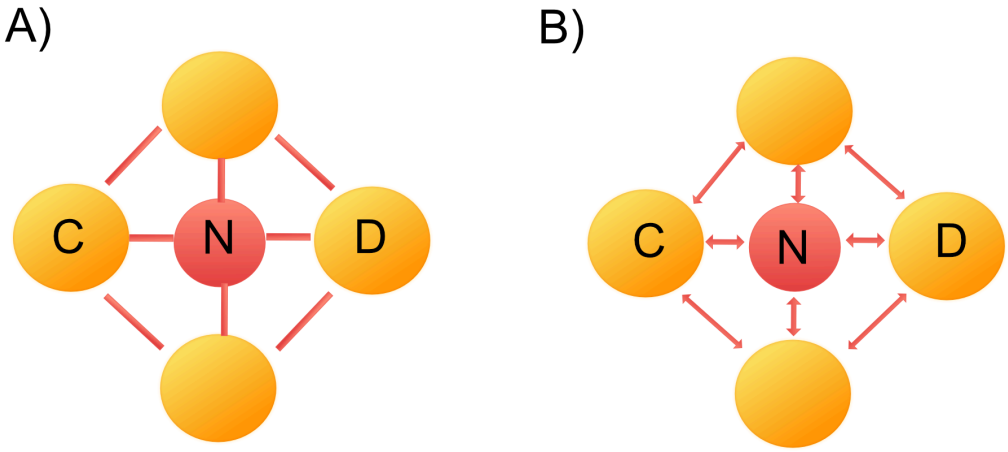


Figure 1.22: *An illustration of undirected and directed graphs:* ***(a)*** *Undirected graph with edges being drawn between all vertices with a relationship despite relationship direction. Therefore, if c shares a relationship with N, but N has no relationship with c an edge will be created.* ***(b)*** *Directed graph with edges only being drawn between reciprocated relationships, whereby only if c shares a relationship with N and N shares a relationship with c will an edge be drawn.*

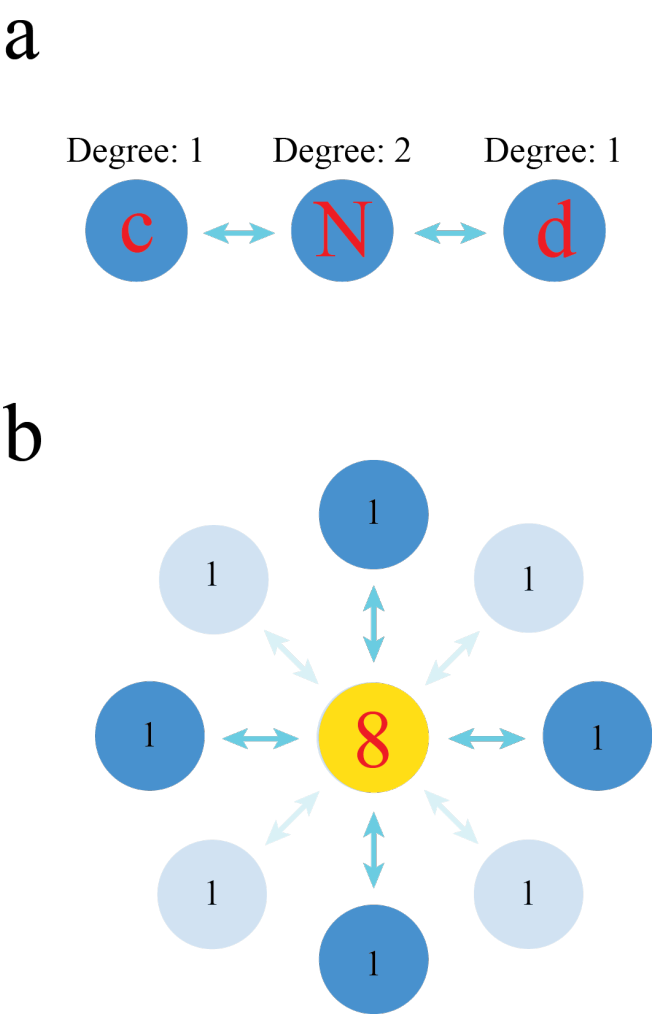


Figure 1.23: *Calculating node degree in networks:* ***(a)*** *Explanation as to how the degree of a node is calculated. Node c and d having a degree of 1 as each only share an edge with one other node, labelled 'N'. N has a degree of 2 as is shares a relationship with both c and d.* ***(b)*** *Nodes with a high degree where connected nodes have a comparatively low degree are known as hub nodes. Here, N is a hub node with a degree of 8, whereas all other nodes in which it shares a connection has a degree of 1. Figure adapted from Webb et al, PhD thesis 2015.*

Distribution pattern was not the only unexpected property of biological graphs, path length also yielded surprising results. The path between any two nodes/vertices, say A and B, in a graph can be found by calculating how many nodes you must pass through to get from A to B, or vice-versa (Figure 1.24). Some calculations that can be made on this principle are (i) the distance of any two vertices in a graph can be obtained by calculating the shortest path between them, (ii) the diameter of a graph can be determined by obtaining the maximum path length of the graph, and (iii) the average distance is calculated from the shortest path between all vertices (Zhu, Gerstein and Snyder, 2007). In biological data the shortest path length was found to be shorter than expected, this fact is known as the 'small world property' of biological graphs (Zhu, Gerstein and Snyder, 2007).

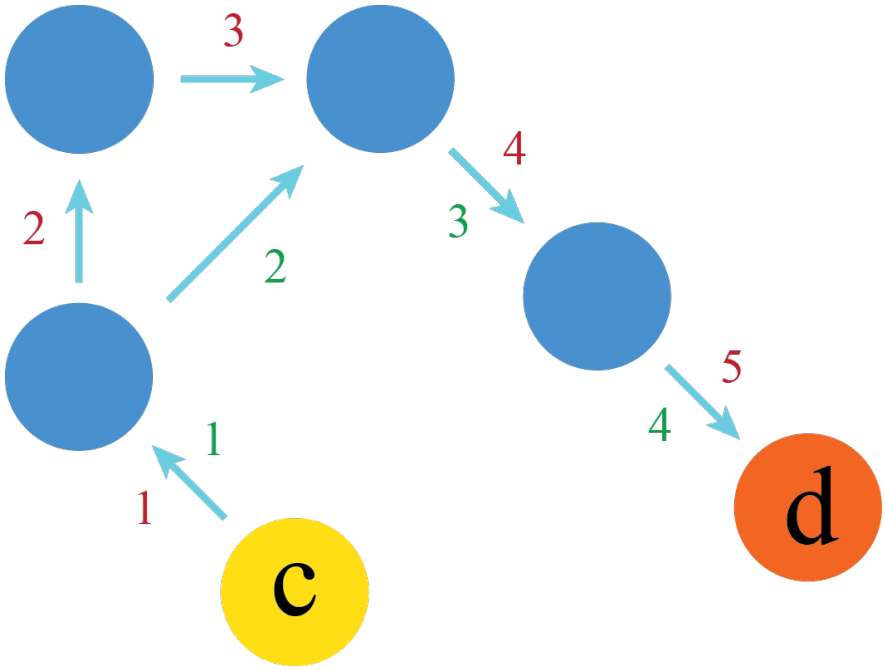


Figure 1.24: *Illustration of path length properties of graphs: Path length illustration depicting both the shortest and longest path between the nodes c and d. The red path represents the shortest path, and the green path represents an existing, but longer path length that would be ignored during this the shortest path length calculation. Figure adapted from Webb et al, PhD thesis 2015.*

Other interesting properties that can be calculated using graph theory include the clustering co-efficient. The clustering coefficient of a given node is the number of edges it shares with it's neighbors in comparison to the number of edges these neighbors have overall (Boccaletti *et al.*, 2006). The clustering coefficient ranges from 0 to 1 where 0 indicates that the neighbors of any given node share no connections whereas a coefficient of 1 indicates they are completely connected.

Assortativity measures the correlation between degree and node connectivity. There are three types of assortatitive patterns: assortative, neutral or disassortative. Assortative mixing patterns create graphs where nodes are likely to connect with other nodes of a similar degree. Contrastingly, disassortative mixing patterns are those where connections between nodes are likely between nodes of a disimilar degree. Neutral mixing patterns exist when a nodes are connected with no bias toward any degree level (Newman, 2003).

In general biological data is modular in nature. Modules are clusters of nodes where each individual node has a low degree on average. In biological graphs these modules of low degree tend to form connections to hub nodes suggesting that hub nodes provide bridges or links between otherwise unconnected modules. These hubs were found to rarely connect with other hub nodes therefore displaying disassortative mixing patterns (Boccaletti *et al.*, 2006) unlike social networks which have been shown to create assortitive mixing patterns (Newman, 2003).

Centrality is particularly helpful in determining the essentiality, or importance, of a node in a graph. There are many different methods for quantifying centrality: 1) degree centrality, 2) closeness centrality, 3) betweeness centrality (Borgatti, 2005). These measures are outlined below:

1. **Degree Centrality**

This measure determines how important a node is by considering the amount of connections or edges it has. More simply, the more connected any given node is the more important it is to the graph overall (Borgatti, 2005).

Equation 2 :

$C_i^{DEG} = \sum_j^N a_{ij}$

$a_{ij} = \{_0^1$ if $i = j$

where i is the given node, j is all other nodes within the network, N is the total number of nodes in the graph, and a_{ij} is the adjacency matrix. The adjacency matrix is needed to interpret whether an edge exists between any given *j*.

2. Closeness Centrality

Closeness centrality is the distance or shortest path between a node and any other node in a graph. The theory behind this is that the more quickly a node can communicate with other nodes the more important the node is to the network (Borgatti, 2005) and is calculated by obtaining the inverse sum of the shortest paths of a given node in a network (*i*) compared to that of all the remaining nodes in the graph (*j*).

Equation 3:

$C_i^{CLO} = N - 1 / \sum_j^{N-1} d_{ij}$

Where N is the total number of nodes in the graph, and d_{ij} is the shortest path in the adjacency matrix.

3. Betweeness Centrality

Betweeness centrality measures the centrality of a given node (*i*) in terms of how many times it lies in the shortest path between two separate nodes in the network, *j* and *k* ($\sigma(j, k)$). The more frequently a node (*i*) acts like a bridge between two nodes, *j* and *k*, the more important it is to overall graph topology ($\sigma(j, k|i)$). Thus,

more important nodes will connect two nodes that would otherwise be unconnected (Freeman, 1977).

Equation 4:

$$C_i^{BET} = \sum_j \sum_k \sigma(j, k|i) / \sigma(j, k)$$

In summary, biological data create hierarchical network structures, where modules cluster together in functionally distinct cliques known as communities. In the future, this inference could be used in predicting gene function in currently unannotated sequences/genomes. To date biological research using graph theory has focussed on protein-protein interaction networks (Makino and Gojobori, 2007; Chen *et al.*, 2014), however with both the wealth and size of sequencing data expanding traditional computational methods are failing to interpret data. Only now is the power of graph theory in tackling sequence similarity questions being fully realised (Qi *et al.*, 2011; Zola, 2014).

1.7.2) Identifying cliques and communities within graphs

Cliques are subgraphs of a graph/network whereby each node within that subgraph are connected to one another. Maximal cliques are those which cannot be expanded any further by neighboring nodes. Calculating all maximal cliques for any graph is considered an non-deterministic polynomial time (NP) problem, i.e. a problem which cannot be solved in an efficient manner but may be verified in polynomial time (Karp, 1971). Although proving difficult there are algorithms available which approximate maximal cliques from graphs (Butenko and Wilhelm, 2006).

Communties are defined as a subset of nodes within a graph that although are highly connected, do not require each each inside to be connected. Therefore they contain many highly connected nodes, but also nodes that contain sparse connections. There are many methods available for deconstructing graphs into its subset of communities but this is beyond the scope of this thesis.

1.7.3) Sequence similarity networks in the identification of fusion genes

For the purpose of comparative genomics, networks are used to identify sequence similarity between DNA and/or protein strings in an approach known as sequence similarity networks (SSNs), where the nodes represent protein-coding genes, transcript sequences or sequencing reads. An edge is drawn between two nodes if they contain sequence similarity, information that can be obtained by a sequence similarity finding software algorithm such as pHMMer (Eddy, 1998), Smith-Waterman (Liu, Maskell and Schmidt, 2009) or BLAST (Altschul *et al.*, 1990) to name but a few. To reduce the amount of edges drawn by random chance a threshold of some sort is utilised, usually either an e-value or percentage identity cut-off. Setting a strict percentage identity threshold, such as 95%, very few mismatches are allowed and therefore only extremely similar sequences will be captured. Both e-value and percentage identity thresholds need careful consideration.

Originally, SSNs were utilized in mining biological data to identify orthologs within poorly annotated genomes as genes which cluster together have an increased probability of them having similar functional role (Figure 1.25). For unannoted genes this aids in the inference of functional predictions (Tatusova *et al.*, 2015). SSNs have been applied for many outcomes including for the detection of fusion genes (Jachiet *et al.*, 2013).

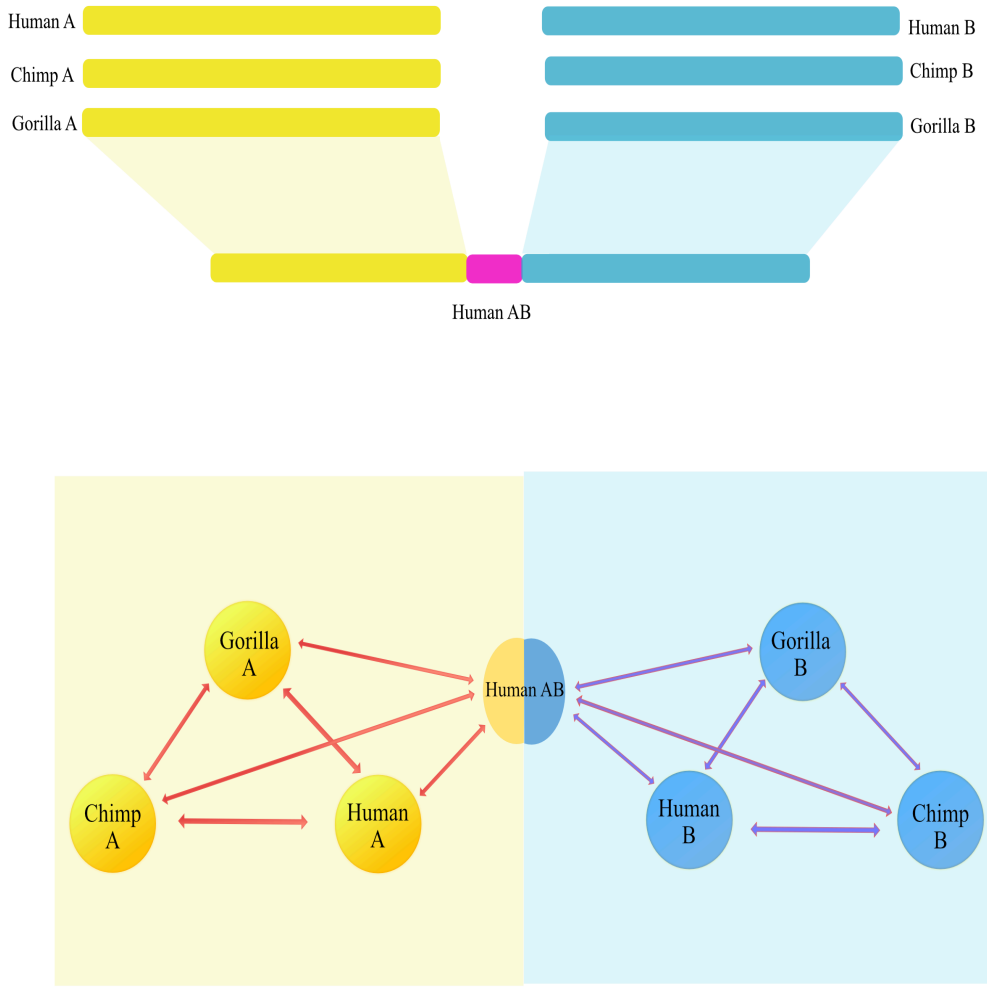


Figure 1.25: *Sequence similarity networks based on global sequence similarity searches explained:* ***Top:*** *Sequence similarity search illustration of a dataset of gene* ***A*** *and* ***B*** *orthologs across gorilla, chimp and human against the human genome. Yellow indicates a sequence similarity hit of gene* ***A*** *orthologs against the human fusion. Blue depicts sequence similarity hit of gene* ***B*** *orthologs within the human fusion gene. Magenta highlights no sequence similarity was found in this region.* ***Bottom:*** *Sequence similarity search results displayed as a sequence similarity network. Here, gene* ***A*** *orthologs are represented by yellow nodes and form a maximal clique. Gene* ***B*** *orthologs have blue nodes and also create a maximal clique. Both maximal cliques share sequence similarity to the human fusion therefore an edge is drawn between both cliques and the human fusion node. However, both cliques share no sequence similarity to each other and therefore share no edges with eachother creating a non-transitive triplet.*

Initially the approaches used to detect fused genes on a genome-wide level failed due to computational issues from the handling of such large datasets as well as an inability to deal with false positives results generated from gene families. It was shown that using network decomposition and clique separation of datasets not only makes the analysis less computationally intense but reduces run time of the analysis (Jachiet *et al.*, 2013). These methods are very useful for the identification of fusion genes from large datasets.

The defining characteristic of a fused gene in a network is that it shares edges with two modules i.e. the parents, and these parents share no similarity to each other. This creates a detectable structure in the network known as a non-transitive triplet where the fused gene is a clique minimal separator. Clique minimal separators are nodes or clusters of nodes that if removed deconstruct the graph. Network decomposition methods can then be used to split a network into designated subgraphs of modules (Palomar and Chiang, 2006) uncovering clique minimal separators within a global network. In summary, clique minimal seperation is used to identify fused genes because if that fused gene were removed from the network then the parent cliques will no longer be connected (Berry *et al*., 2010).

However, there is a potential for high false positive rates using this approach as distant homologs can be interpretted as parents of fused genes. At present, network based packages are unable to completely address this shortcoming and so expertise is required for interpretion and quality control of the results. Another source of error is the possibility of incorrect annotations during sequencing as well as the incorrect joining of short reads causing superficial fused gene calls (Jachiet *et al.*, 2013).

1.7.4) FusedTriplets and MosaicFinder packages for fusion detection

FusedTriplets (python package) and MosaicFinder (C++ package) are two software packages that use different methodologies based on mathematical graphs to predict fusion genes from input genomes (Jachiet *et al.*, 2013).

FusedTriplets **(Version_2.0)** (Jachiet *et al.*, 2013) detects non-transitive triplets within a global sequence similarity network to determine possible fusion gene events. Candidates are re-analysed using a more permissive threshold to investigate whether the triplets remain non-transitive at this level of similarity. More simply, if parents are dissimilar at an e-value threshold of $1e^{-10}$ the same test is carried out at using a threshold of $1e^{-5}$ if they remain dissimilar the fusion passes the re-analysis. After this each parent of a predicted fused gene is aligned and only those with sequence similarity overlap of <20 amino acids are accepted. This slight overlap is allowed because sequence similarity searches, such as BLAST (Altschul *et al.*, 1990), tend to overextend alignments during analyses (Jachiet *et al.*, 2013).

MosaicFinder **(Version_3)** (Jachiet *et al.*, 2013) uses a slightly more complex algorithm to predict gene fusion. Here a global sequence similarity network based on a user defined soft and hard thresholds is generated. The global network is split into subgraphs based on clique minimal separators (CMS) to identify non-transitive patterns (Berry *et al.*, 2010; Jachiet *et al.*, 2013). This allows for the identification of fused genes and their corresponding parent gene families/cliques. As previously mentioned, CMS are used to detect fused genes as each fused gene node itself becomes a separator as its removal disconnects the two parent gene families as no sequence homology is shared (Jachiet *et al.*, 2013). The identification of accurate parent gene families of the each fused gene is calculated using additional decomposition methodologies in the common neighborhood of each CMS (Berry, Pogorelcnik and Simonet, 2010) so that only true parents are retained. At this stage there is an optional crosscheck for similarity between the two identified parent gene families for distant similarity, significantly reducing the possibility of parent gene families being distant homologs. Like fusedTriplets a permissive threshold is used which is optional as the software package yields robust results regardless. After this, an alignment step is carried out to minimize false positives and parent gene family alignments that overlap on the fusion gene by more than 20 amino acids are rejected. If the fused gene family passes all the aforementioned requirements a “fusion point” is

calculated which is the median point between both parent alignments on the fused gene (Jachiet *et al.*, 2013).

Mosaicfinder uses global sequence similarity searches in order to identify non-transitive triplets between cliques. In order to do this a clique minimal separator (CMS) algorithim is used to identify CMSs in the global network that create non-transitive triplets *i.e.* cliques that share no sequence similarity to eachother but share sequence similarity an edge with an alternative, but the same nodes.

Chapter 2: Identification and computational characterisation of RNA-mediated gene fusions across primate genomes

2.1) Introduction

The modular nature of protein-coding sequences along with non-homologous recombination provides the opportunity for new genes to evolve from existing protein coding sequence (Tan *et al.*, 2010). This process can occur at the level of exon/domain i.e. exon/domain shuffling (Section 1.2.4) and also at the level of genes, i.e. gene fusion/fission (Section 1.2.5) (Kaessmann, 2010). Combined we refer to these mechanisms of novel gene genesis as "gene remodeling events". Gene fusion/fission can be further categorised based on the specific process that brought them about: (1) DNA mediated Gene fusion/fission (DMGF) (Bailey *et al.*, 2002; She *et al.*, 2006; Marques-Bonet *et al*., 2009), and (2) RNA mediated-gene fusion/fission (RMGF) (Akiva *et al.*, 2006; Rodríguez-Martín *et al.*, 2017). DNA-mediated gene fusion events have been identified on an gene level across *animalia,* for example the remodeling of *Adh* to derive *jingwei* in *D. melanogaster* 2 million years ago (MYA) (Long and Langley, 1993) and the remodeling of two human genes, *Kua* and *UBE2V1* to derive the *Kua-UEV* fused gene (Thomson *et al.*, 2000). They have also been identified and investigated on a genome-wide level specifically in fungi (Leonard and Richards, 2012), viruses (Jachiet *et al.*, 2014), and human (Y. Wang *et al.*, 2015).

Contrastingly, at present RMGFs are thought a rarity and although their presence has been somewhat investigated in human (Pardigol *et al.*, 1998; Upadhyaya, Lee and Dejong, 1999; Zaphiropoulos, 1999; Thomson *et al.*, 2000; Pradet-Balade *et al.*, 2002; Akiva *et al.*, 2006) only a small number of individual cases of RMGF amongst other vertebrates have been reported in the literature (Veeramachaneni *et al.*, 2004).

An understanding of RMGFs, their impact on genomic architecture and their contribution to phenotypic change across vertebrates is necessary as their significance has been demonstrated to have a substantial impact on the human genome; causing disease (Fears *et al.*, 1996), novel expression patterns (Pardigol *et al.*, 1998), novel functions (Kolfschoten *et al.*, 2003) and novel cellular localisation (Thomson *et al.*, 2000). However, it remains unclear how often this mechanism produces new genes in vertebrates, the impact they have on genome

structure across the vertebrate phylogenetic tree, and what mechanisms could potentially be driving their emergence in these genomes.

The aim of this chapter was to detect RMGFs across vertebrates, assess their mechanism of generation, and characterise the genes to highlight significant differences between RMGFs and non-fused protein-coding genes. In order to do this the frequency of RMGFs across vertebrates was assessed using a sequence similarity network (SSN) based approach (Bapteste *et al.*, 2012). The RMGF identification pipeline was established using a dataset of 6 primates and mouse (out group) with curated RMGFs undergoing further investigation across a panel of 13 high quality vertebrate transcriptomes (Brawand *et al.*, 2011).

Much is known about the mechanisms behind new gene generation across genomes (Section 1.2) but the methodology behind events RMGF specifically remained aloof. In the literature a link between fusion genes and genomic rearrangements had been proposed (Bailey *et al.*, 2002; Stankiewicz *et al.*, 2004; Bailey and Eichler, 2006) therefore their formation through rearrangements, specifically segmental duplication (SD) (Section 1.4) was examined in human RMGFs using the Segmental Duplication Database (Sharp *et al.*, 2005).

RMGF characterisation through codon usage bias, selective pressures and the evolutionary rate established an understanding of RMGF compared to non-fused human protein-coding genes. Both functional and motif analyses provided further characterisation of RMGFs highlighting the potential of these genes to impact the genome at a phenotypic level.

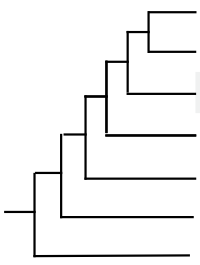
2.2) Materials and Methods

2.2.1) The identification of RMGF across primate and vertebrate genomes

2.2.1.1) Primate dataset acquisition, cleaning and filtering

An assessment of data quality across available primate species data was carried out by comparing both coverage and N50 values and a panel of primates with high quality genomes were selected and downloaded by Ensembl Genome

Browser (Herrero *et al.*, 2016). The mouse genome was also downloaded as a high quality out group (Herrero *et al.*, 2016). Protein-coding DNA genes were downloaded from the Ensembl Genome Browser API (**Version 71**) (Herrero *et al.*, 2016) for the following species (and versions): *Homo sapiens* (GRCh37), *Mus musculus* (GRCm38), *Pan troglodytes* (CHIMP2.1.4), *Gorilla gorilla* (gorGor3.1), *Macaca mulatta* (MMUL_1), *Pongo abelii* (PPYG2) and *Callithrix jacchus* (C_jacchus3.2.1) (Figure 2.1). Quality checks were carried out to ensure adequate sequence quality through assessment of 1) complete codons , and 2) intermittent stop codons across sequences. Any sequences that did not meet these criteria were removed from the dataset. To create an amino acid database the filtered nucleotide dataset was translated with the phase information considered using the lab-designed tool, VESPA (Appendix_B).



Species	Assembly	Coverage	Contig N50	Assembly Size	Gene Number
Human	GRCh37	High	38Mb	3,326,743,047	20,805
Chimpanzee	CHIMP2.1.4	6	50kb	2,995,917,117	18,759
Gorilla	gorGor3.1	2.1 & 35	11kb	2,828,888,833	20,962
Orangutan	PPYG2	6	15kb	3,109,347,532	20,424
Macaque	MMUL_1	6	35kb	3,093,871,206	21,905
Marmoset	culJac3	6	29kb	2,759,289,125	20,978
Mouse	NCBIM37	High	32Mb	2,745,142,291	22,147

Figure 2.1: *A phylogram illustrating selected high quality primate genomes for analysis: Figure highlights the primate and mouse genome assemblies used throughout the RMGF identification process. It also shows the quality of each genome, indicated by coverage and N50 contig size, assembly size and protein coding gene number.*

2.2.1.2) Sequence similarity searches on dataset of primate protein-coding genes

To investigate sequence homology profiles across our dataset sequence similarity searches were required (Section 1.7.3). Each genome within our dataset was assessed individually using a best reciprocal BLASTp approach (Altschul *et al.*, 1990) using an e-value = $1x10^{-5}$ and self-hits were removed. This generated a blast result for each individual genome within our dataset.

2.2.1.3) Sequence similarity network generation to identify RMGFs

Sequence similarity results obtained from Section 2.2.1.2 were used as input for sequence similarity network generation in order to identify gene fusion events. In order to select the most sensitive and robust algorithm for network creation a comparison of available fusion detection software packages was carried out – namely fusedTriplets and MosaicFinder (Section 1.7.4).

MosaicFinder **(Version_3)** detects fusion genes using complex algorithms for global SSN generation based on user defined soft and hard thresholds. Clique minimal separators (CMS) split this SSN into subgraphs of non-transitive triplets (Section 1.7.2). This identifies fused genes and their parent gene families/cliques as each fused gene node itself becomes a separator due to its removal resulting in the disconnection of the two parent gene families that contain no sequence homology. At this stage there is an optional crosscheck for similarity between the two identified parent gene families for distant similarity, significantly reducing false positive predictions caused by parent gene families actually being distant homologs. A slight overlap of 20 amino acids is tolerated to account for any overextensions during BLASTp searches "Fusion points" are then calculated which is the median point between both parent alignments on the fused gene.

FusedTriplets **(Version_2.0**), a second fusion detection software package**,** uses non-transitive triplets within a global SSN to detect gene fusions. Triplet candidates undergo a re-analysis using more permissive thresholds to investigate whether the triplets remain non-transitive. More simply, if parents are dissimilar at an e-value threshold of $1e^{-10}$ the same test is carried out using a threshold of $1e^{-}$

[5] if they remain dissimilar the fusion passes the re-analysis. Like MosaicFinder, pairwise alignments are constructed between the fusion gene against its parents for false positive detection.

Findings were compared and considered and MosaicFinder (Jachiet *et al.*, 2013) was chosen due to its conservative nature and the ability to deal with large datasets in both a time and computationally inexpensive manner. The MosaicFinder algorithm uses a sequence similarity search result to generate a global sequence similarity network followed by a deconstruction step into discrete sub-graphs of clique minimal separators (gene fusions) (Berry, Pogorelcnik and Simonet, 2010). In order to investigate the rates of evolutionary change of RMGFs, MosaicFinder analyses were carried out across each individual genome at four thresholds of sequence percentage identity (PI) (50%, 70%, 80% and 90%) (Jachiet *et al.*, 2013). The iGraph package was used to visually inspect each fusion/parent gene family (Csárdi and Nepusz, 2006).

2.2.1.4) Validation of identified RMGF through pairwise alignment construction

In order to ensure accurate RMGF prediction the protein-coding sequences for RMGFs identified at a 90% PI threshold and their corresponding gene parents were extracted from our database. As shown in Figure 2.2 individual pairwise alignments were constructed using PRANK (Löytynoja and Goldman, 2010) for each RMGF against its corresponding parent gene. For example in Figure 2.2, Fusion AB underwent a pairwise alignment against Parent A and then again against Parent B. Generated alignments were manually assessed using the SeaView software package to ensure 1) RMGF calling was accurate and 2) for distant homology between parent A and parent B (Gouy, Guindon and Gascuel, 2010),

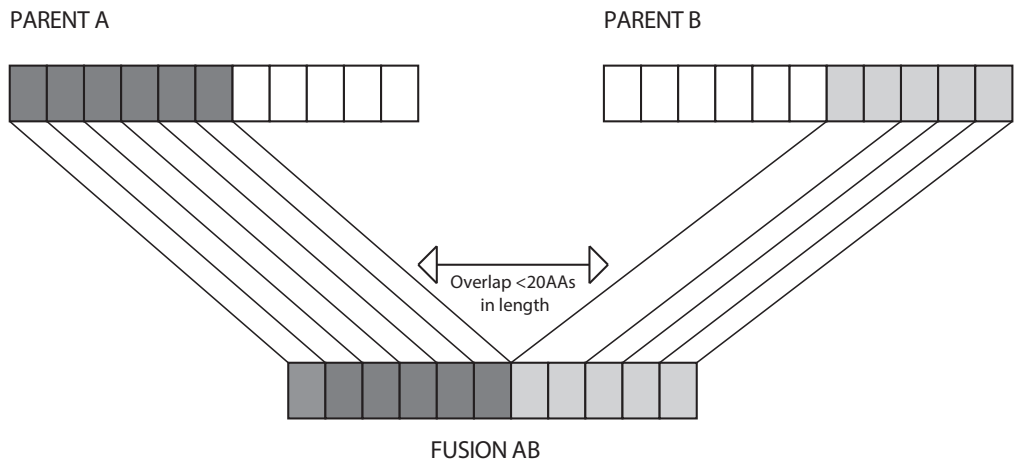


Figure 2.2: *Alignment validation of SSN identified fusion genes: Illustrates a fusion AB's sequence homology to both parent genes, Parent A and Parent B. According to our criteria Parent A and Parent B should not have any sequence homology to one another but a 20 amino acid overlap between the sequences is tolerated to allow for blast over extension.*

2.2.1.5) An investigation of identified RMGF orthologs across vertebrate species

In order to determine the phylogenetic distribution of the RMGFs identified at 90% PI transcriptomic data was obtained for a panel of selected high quality vertebrate species. The RNA datasets used were taken from the NCBI database (O'Leary *et al.*, 2016) for the following species: Bonobo; Cat (Felis_Catus_3.2); Ceolocanth (LatCha1); Chicken (Gallus_gallus4.0); Chimp (PanTro4); Cow (BosTau4); Dog (CanFam3.1); Dolphin (Ttru_1.4); Elephant (Loxafr3.0); fugu (FUGU4.0); Gibbon (Nleu_1.0); Gorilla (Gorgor3.1); Guinea pig (Cavpor3.0); Horse (EquCab2.0); Human (GRCm38.p3); Macaque (Mmul_051212); Marmoset (Callithrix_jacchus3.2); Brown bat (MyoLuc2.0); Mouse (GRCm38.p2); Naked mole rat (hetGla2/hetGla_Female_1.0); Olive baboon (Panu2.0); Opossum (MonDom5); Orangutan (P_pygmaeus2.0.2); Orca (Oorc1.1); Pig (Sscofra10.2); Platypus (Ornithorynchus_anaticus5.01); rat (Rnor.6); Tarsier (Tarsius_syrichta1); Turkey (Turkey2.01); Zebrafish (GRCz10), and Zebrafinch (teaGut3.2.4)). Sequence similarity searches were performed using the RMGFs as queries (Altschul *et al.*, 1990). Results were parsed and alignments generated using MUSCLE (Edgar, 2004). [Note: in this instance MUSCLE (Edgar, 2004) is used rather than PRANK (Löytynoja and Goldman, 2010) as it had adequate sensitivity and increased speed].

2.2.2) To determine if RMGFs coincide with known human segmental duplication break points

To understand the mechanism behind RMGF creation an assessment of our RMGFs identified at a 90% identity threshold for enrichment in known regions of human segmental duplication was carried out. Human chromosomal positions were obtained for RMGFs (if co-ordinates were available) using the Ensembl Genome Browser (**Version 74**) (Herrero *et al.*, 2016). Human SD coordinates were obtained from the Segmental Duplication Database (Khurana *et al.*, 2010). Overlap between human RMGF chromosomal coordinates (at each PI) and the human SD coordinates was assessed (Figure 2.3) using Perl scripts (Appendix_B).

To assess the frequency of occurrence of fused genes and parent genes in regions of segmental duplication simulations were carried out as follows: The coordinates for all human protein coding sequences were downloaded from the Ensembl Genome Browser (**Version 74**) (Herrero *et al.*, 2016). Datasets of fused and parent genes were simulated for each level of percentage identity by random sampling from the entire set of protein coding genes and chromosomal coordinates for human. From this randomly sampled data the number of genes located in a region of SD were recorded. This simulation was carried out on 10,000 replicate sets and p-values were obtained.

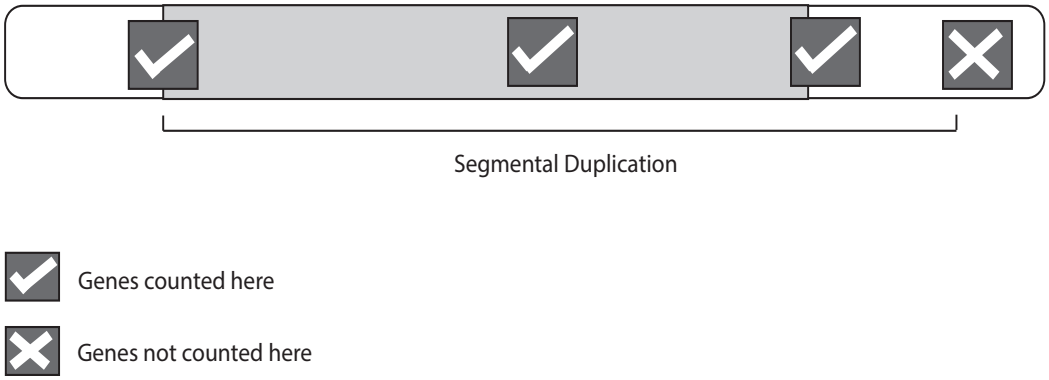


Figure 2.3: *A graphical representation of the methodology behind identifying RMGFs within regions of human segmental duplication: Illustrates theory behind RMGF identification in regions of known human segmental duplication code production. The graphic highlights the process of a Perl script that generates a list of RMGFs located entirely or partially within known regions of SD. The diagram also highlights that RMGFs lying outside of known SDs are ignored.*

2.2.3) RMGF characterisation and comparison to non-fused protein coding genes

2.2.3.1) A functional enrichment analysis across RMGFs and their parents

In order to understand the impact of RMGFs a functional enrichment analysis was carried out using the software package GOrilla (Eden *et al.*, 2009). The Ensembl gene identifiers (Herrero *et al.*, 2016) for each RMGF and their parents from human and mouse across each PI threshold (70%, 80%, and 90%) were used. Although RMGFs genes were included, due to low frequency at high PI thresholds (90PI), power was insufficient for accurate assessment. Using the GOrilla (Eden *et al.*, 2009) software package enrichment was analysed through exact p-value and false discovery rate (FDR) q-value calculation. GO terms were then investigated using the GO term Consortium Database and functions investigated using the Uniprot Database (Huang *et al.*, 2007; UniProt Consortium, 2018).

2.2.3.2) An investigation of human single nucleotide polymorphism and INDELs across human RMGFs

To characterise the rate of polymorphism across fusion genes an analysis of human RMGFs identified at 90 PI were investigated for both single nucleotide polymorphisms (SNP) and insertion/ deletion events (INDEL). Both SNP and INDEL coordinates across Great Apes were obtained (Prado-Martinez *et al.*, 2013) and chromosomal coordinates of RMGFs identified in human (90 PI threshold) were downloaded from the Ensembl Genome Browser **(Version 83)** (Herrero *et al.*, 2016). Co-ordinates were cross-compared to identify polymorphic human RMGFs.

2.2.3.3) An assessment of RMGF motif usage

A motif enrichment analysis was carried out to investigate whether human RMGFs have a bias for specific motifs usage patterns in comparison to a dataset of randomly shuffled human protein coding sequences. Human RMGFs

identified at the 90 PI were investigated for regulatory motif enrichment using the AME function in the MEME software suite (Bailey *et al.*, 2009).

The AME software package assesses input sequences for motifs and compares the output with a user provided control dataset or computationally shuffled sequences generated by the package itself. Sequence data was obtained using Ensembl Genome Browser's Biomart System (**Version 83**) (Herrero *et al.*, 2016). Default settings were used with a threshold of significance of $p < 0.05$ and shuffled input sequences were used as controls. RMGF sequences were analysed against a eukaryote DNA database (Bailey *et al.*, 2009).

2.2.3.4) Codon usage analysis of human RMGF in comparison to non-fused human protein-coding genes

In order to investigate whether human RMGFs codon usage is under selective constraint, the codon usage of identified genes at the 90 PI threshold were assessed (McInerney, 1998). Codon usage was compared between RMGFs and a test dataset of 40,000 human transcripts randomly selected from the human genome through the Ensembl Genome Browser's Biomart System (**Version 83**) (Herrero *et al.*, 2016). The frequency of codon utilization was calculated using the GCUA software package that generates relative synonymous codon usage (RSCU) values across the dataset. This calculation is based on the amount of times a specific codon is used by a gene compared to the amount of times you would expect it to be used in the absence of codon usage bias. If no codon usage bias is present the RSCU value is 1, a codon usage less than expected <1 and a codon usage more than expected has an RSCU of >1. The usage of RSCU values normalizes that result across a dataset. An average of each codon across the datasets was obtained and the average codon usage between RMGFs and protein coding genes were compared.

2.2.3.5) An investigation of RMGF family location across the vertebrate tree

In order to understand the transmission of RMGFs across genomes over both shallow and deep phylogenetic depths an analysis was carried out across the parent's of RMGFs identified at 90% soft threshold. The chromosomal

coordinates for each parent of each human RMGF identified were downloaded from the Ensembl Genome Browser **(Version 83)** (Herrero *et al.*, 2016). The orthologous chromosomal position (syntenic) for each parent's RMGF across a panel of 25 vertebrate species (Human, Chimpanzee, Gorilla, Orangutan, Gibbon, Baboon, Macaque, Marmoset, Tarsier, Rat, Mouse, Guinea pig, Dolphin, Cow, Pig, Horse, Microbat, Dog, Cat, Elephant, Opossum, Platypus, Chicken, Ceolocanth, and Zebrafish) were downloaded from the Ensembl Database (Herrero *et al.*, 2016). The position of each parent was then assessed across all genomes.

2.2.4) Rate heterogeneity and selective pressure heterogeneity analyses across candidate RMGF

2.2.4.1) Selection, sequence acquisition and alignment of candidate RMGFs

In order to investigate the rate of evolution across RMGFs in comparison to non-fused protein coding genes an analysis was carried out of the RMGFs identified at the 90 PI threshold against a dataset of non-fused protein coding genes corresponding to each individual species within the dataset (human, chimpanzee, gorilla, orangutan, macaque and mouse). However, not all RMGFs within this dataset could be assessed due to the sophistication of the analysis. Only the most simplistic cases of RMGF, i.e. those with 2 parents, were considered for branch length analysis and subsequent selective pressure variation. In total 12 RMGF met our criteria as follows: (i) sequence is available for the two parent genes and only 2 parent genes were identified for the RMGF, (ii) the orthologous sequences of each parent sequence are available across six or more vertebrate genomes (this is a prerequisite for CodeML analyses to minimise FDR (Yang, 2007)) and (iii) that the homologous sequences all come from the same set of species to ensure accurate comparability. All sequences were downloaded from the Ensembl Genome Browser Biomart System (**Version 83**) (Herrero *et al.*, 2016). Each RMGF was pairwise aligned to each parent gene individually using PRANK (Löytynoja and Goldman, 2010), and the alignments were visualized and validated using SeaView (Gouy, Guindon and Gascuel, 2010). Each alignment was manually cleaved so that only homologous regions of the RMGF and

corresponding parent gene were retained. Orthologous sequences from selected vertebrate species were then aligned against the initial cleaved pairwise alignment and trimmed in order to retain the corresponding homologous portion (Figure 2.4) 24 multiple sequence alignments were generated, again utilising PRANK (Löytynoja and Goldman, 2010). Cleaving of the alignments was carried out using SeaView (Gouy, Guindon and Gascuel, 2010).

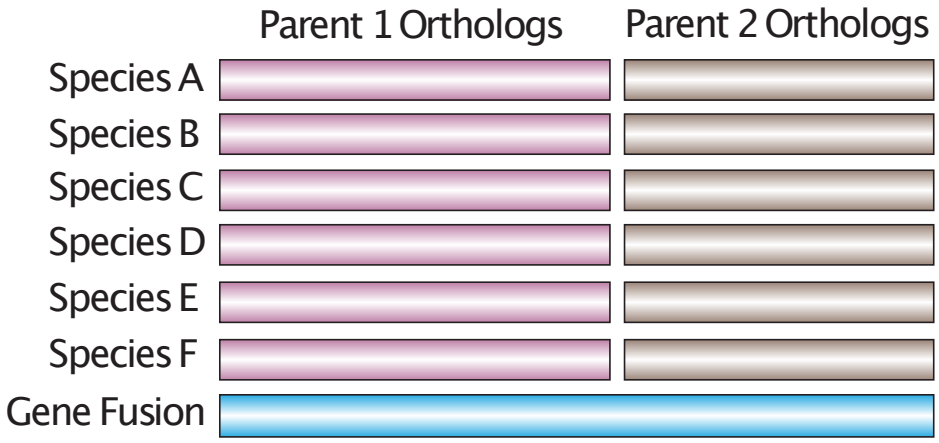


Figure 2.4: *Illustration of candidate RMGF selection process and sequence data acquisition: The image highlights the selection of candidate genes that were eligible for both branch length analyses and phylogenetic reconstruction analyses. The blue gene highlights the sequence of the candidate RMGF and pink bar represents parent gene 1 (pink Species F) and only the region of sequence that is homologous to the RMGF and it's orthologs across 5 other vertebrate species (Species A-E). The pale brown gene shows parent gene 2 (pale brown Species F) highlights the homologous sequence it shares with he RMGF and its orthologous sequence across the same other 5 species (Species A-E) as parent gene 1.*

2.2.4.2) Phylogeny reconstruction of candidate RMGFs

After selection and alignment of the most suitable candidate RMGFs in Section 2.2.4.1 the evolutionary rate of RMGFs could be compared through branch length calculation. The branch length for each selected RMGF was estimated using the heterogeneous phylogenetic modeling approach implemented in P4 (Foster, 2004). We estimated the branch lengths for 24 alignments (12 fused genes each with 2 parents). For each estimate we supplied P4 with an alignment, its associated pre-calculated composition vector and exchange rate matrix (JTT was selected), and a fixed topology (species tree)(Morgan *et al.*, 2013). P4 was run for two million generations with sampling every ten generations. Parameters were assessed during the MCMCMC process and were accepted between 10-80% of the time. Finally, we compared the standard deviation between the checkpoints of the MCMCMC process, where a low standard deviation between checkpoints indicates convergence. To test if the model (composition vector and exchange rate matrix) used on each alignment was appropriate for the data we carried out posterior predictive simulations. The simulations were generated during the MCMCMC process for each alignment. Each simulated dataset was compared to the input data. The real data should look characteristically similar to the simulated data in instances where the model of evolution is adequate for the given data. This simulated data was then compared to the real data using a χ^2 test to determine whether the RMGFs were evolving at a faster rate on average. For each analysis p-values were calculated based on the degrees of freedom for that analysis.

2.2.4.3) Testing the selective pressures acting on RMGFs

To obtain an understanding of the selective pressures acting on new genes of this nature and whether they had the potential to undergo fixation in the genome a selective pressure analyses was carried out using CodeML from the PAML package (Yang, 2007), both site- and lineage-specific codon models of evolution were compared using a maximum likelihood framework. Therefore there is an inherent risk of local minima increasing the frequency of false positive calls. To reduce the chances of reporting results from local minima, analyses were

performed at several starting omega (ω) values across the likelihood plane (ω = 0, 1, 2, and 10). To determine the model of best fit for each alignment a standard Likelihood ratio test (LRT) was carried out under a χ^2 distribution as displayed in Table 2.1 (Yang and Bielawski, 2000; Zhang *et al*., 2005).

Table 2.1: Likelihood ratio test calculations and number of parameters estimated for each model

Comparison	Degrees of Freedom	Δl	χ2 Critical Values
M0 v M3K2	2	X2	≥ 5.99
M3K2 v M3K3		X1	≥ 1.00
M1a v M2a	2	X2	≥ 5.99
M7 v M8	2	X2	≥ 5.99
M8 v M8a	1	X2	≥2.71 ≥ 5.41
M1A v Model A	2	X2	≥ 5.99
Model A v Model A null	1	X2	≥ 3.84

Model comparisons carried out by the CodeML package are depicted here. Each model is compared to a more paramaterised model to ensure the most accurate model is implemented. Each LRT result (Δl) is multiplied by Column 3 above. Critical values for each model are also illustrated for each comparison to be made.

2.2.4.4) An assessment of the accuracy of the Ensembl Genome Browser's transcript annotation pipeline

An analysis was carried out to investigate the validity of the Ensembl annotation process (Zerbino *et al.*, 2018). In order to do this the 9 human-specific RMGFs identified were assessed for orthologs in chimpanzee and gorilla through the Ensembl Biomart System (Zerbino *et al.*, 2018). Results were mapped onto a phylogram for illustrative purposes.

2.3) Results

2.3.1) RMGFs detected in primate and vertebrate genomes

In order to accurately identify RMGFs on a genome–wide level a pipeline was constructed and tested on a dataset of high quality primate genomes and mouse (outgroup species) (Section 2.2.1). A comparison of fusion gene identification tools, namely Fusedtriplets and MosaicFinder was carried out across human protein coding genes, and the frequency of fusion genes were identified and compared (Table 2.2). Mosaicfinder produced a panel RMGFs that was more conservative utilising more sophisticated algorithms and therefore was selected as the detection software of choice. At a PI threshold of 90 MosaicFinder identified 360 fusion genes in comparison to fusedTriplets were 4699 fusions were identified and 4400/4699 (94%) were not identified by the MosaicFinder tool. At a threshold of 80%, 90% of fused genes identified by fusedTriplets were not identified by MosaicFinder, at the 70% threshold 88% were not found, and at 50% 85% were not found.

Table 2.2: MosaicFinder vs FusedTriplets comparative analysis of RMGF in humans at 90, 80, 70 and 50% identity thresholds

Percentage ID Threshold	fusedTriplets (fT)	MosaicFinder (MF)	Found in MF but not fT	Found in fT but not MF
90%	4699	360	7	4400
80%	8125	929	20	7356
70%	11854	1624	40	10523
50%	23275	4738	145	19844

Indicates the frequency of fusion gene identification across human protein coding genes across a range of soft identity thresholds (Column 1). The frequency of fusion genes only found by each individual software package were then investigated and compared.

After MosaicFinder algorithm selection fusion gene detection analysis was carried out (Jachiet *et al.*, 2013). A range of four percentage identity thresholds were tested (50%, 70%, 80% and 90%) and the frequency of fusion genes detected at each threshold in each individual species is highlighted in Figure 2.5 and the specific numbers are represented in Table 2.3**.** From this it is clear that the relationship between the number of RMGFs detected and the PI used is inversely proportionate as the number of fusions detected at 90 PI is less than that identified at a 70 PI threshold across all species examined.

From Table 2.3 it is clear that an increased level of fusion genes were identified across the human genome when compared to the other species in our dataset with 42 fusion events being identified at the most stringent PI threshold of 90 all other species in the dataset identified only a fraction of these events with chimp and gorilla containing only 4% of the amount identified in human, orangutan containing 23%, macaque and marmoset containing only 12% and 2% respectively. A similar profile of fusion gene identification frequency can be seen across all PI thresholds examined. The number of genes identified across each species in the dataset does not increase in a linear fashion across the phylogenetic tree, indicating that fusion genes frequently occur at a species-specific level and their generation rate is also species dependent

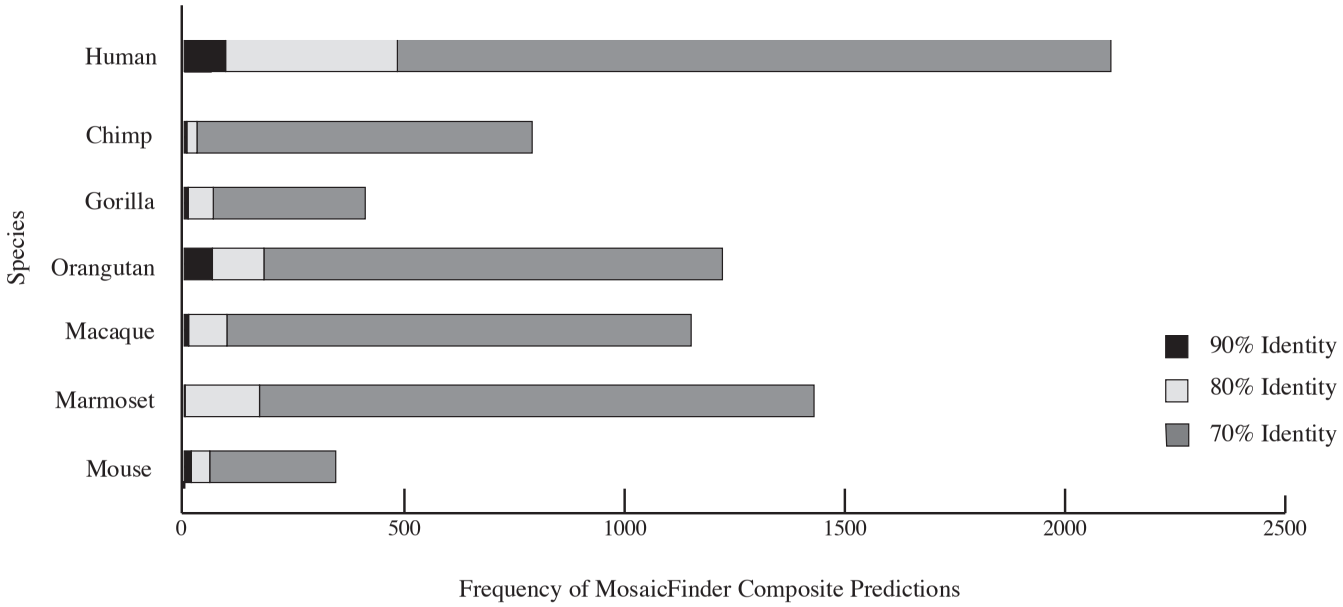


Figure 2.5: *The frequency of RMGF calculated across primates and mouse at 90%, 80%, 70%, and 50% identity thresholds: Displays the frequency of fusion genes in our dataset across a range of soft percentage identity thresholds. Species are depicted on the Y axis and RMGF frequency on the X axis. Black bars represent a percentage threshold of 90% identity, pale grey show 80% identity results, and dark grey depict RMGFs found using a 70% threshold.*

Table 2.3: MosaicFinder RMGF frequency results across our dataset

Species	90%	80%	70%	50%
Human	42	68	98	258
Chimp	2	7	35	152
Gorilla	2	12	24	127
Orangutan	10	16	31	209
Macaque	5	14	70	409
Marmoset	1	18	54	147
Mouse	7	9	14	92

Results table comparing frequency of fusion gene detection across individual species in our dataset at four percentage identity thresholds.

Pairwise validation alignments were carried out as outlined in Section 2.2.1.4 and manually curated. From alignment inspection it became evident that all genes identified using a soft PI threshold of 90% were generated by an RNA-mediated mechanism (See Figure 1.3 image of RMGF compared to DMGF).

Thirty five human RMGFs and 3 mouse RMGFs identified at 90 PI both passed alignment criteria and manual curation and were analysed to determine the frequency and dispersal of these transcripts across a dataset of 13 vertebrate species across a number of tissue types. Vertebrate RNA transcripts were mapped to a panel of constructed RMGF fake reads Section 3.2.1.2, if >=1 fake read mapped successfully to a vertebrate RNA transcript this was taken as evidence for the existence of this transcript within that species, these instances are highlighted as red boxes in Figure 2.6. However, if the fake read did not map this was not sufficient evidence for the non-existence of the RMGF in this species but rather that the RMGF does not exist in the current data using current techniques. Out of the 35 RMGF transcripts examined 29 mapped successfully and in human 9 RMGFs were not found across any other vertebrate analysed. 17/35 RMGF transcripts mapped in chimpanzee, 1/17 were both chimpanzee and human specific and 3 were found across chimpanzee, bonobo and human only. 14/35 genes were identified in bonobo and gorilla. 10 of the RMGFs were found in orangutan and 11 in macaque. Only 5 were found in mouse, 3 of which were not found in any other vertebrate species and finally only 1 RMGF was identified across both opossum and platypus.

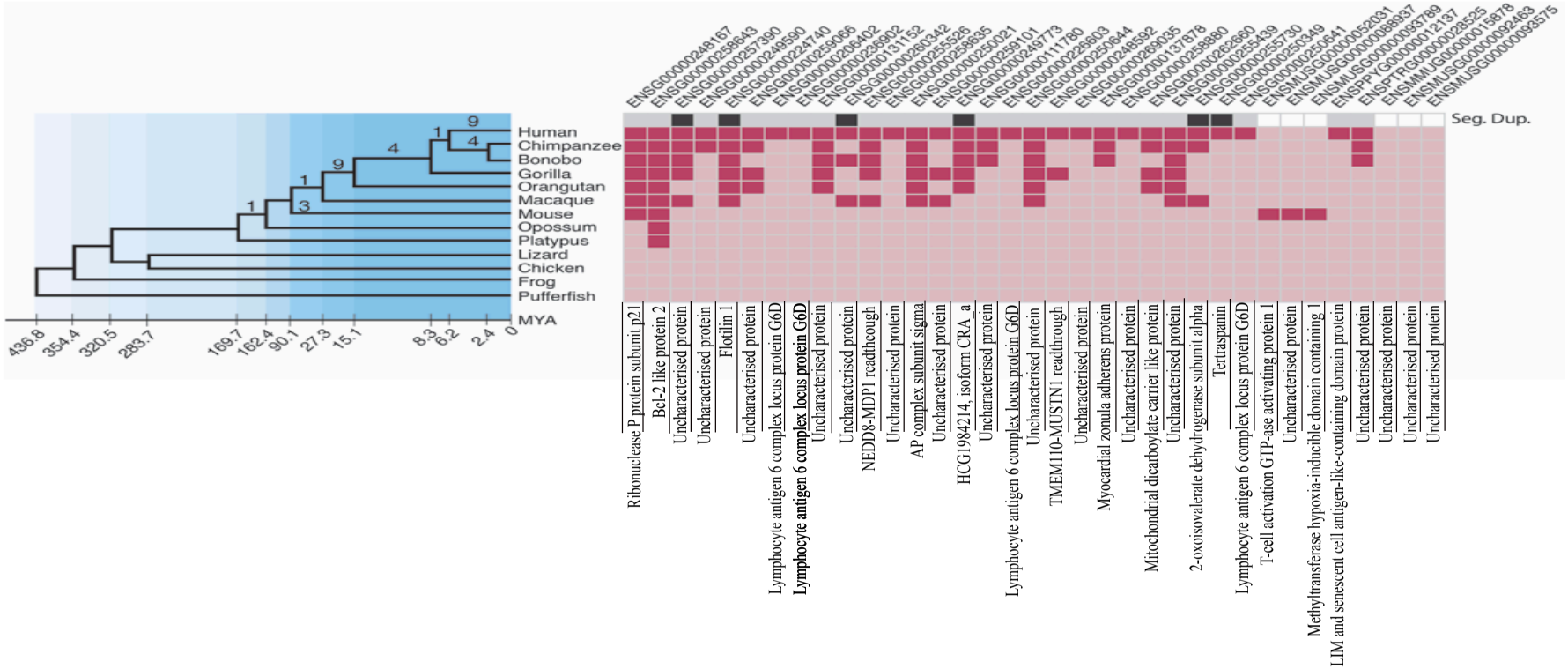


Figure 2.6: *The phylogenetic distribution of RMGFs detected.* ***Left)*** *The species sampled are represented in the phylogeny with their estimated divergence times in million of years ago (MYA). The numbers on the branches represent the number of RMGFs at those nodes.* ***Right)*** *The cells within the matrix correspond to the presence (deep pink) or absence (pale pink) of the gene fusion in that species. The "Seg Dup" row in the matrix shows the RMGFs present at known SD breakpoints from human (dark grey), in pale grey are RMGFs with missing information and in white are the RMGFs not found in human. Uniprot data for each RMGF is indicated below each RMGF column within the matrix.*

2.3.2) To determine if RMGF are overrepresented in regions of primate segmental duplication in comparison to non-fused protein coding genes

As detailed in Section 1.4 there is already a well-established link between new gene generation, such as RMGFs and regions of genomic instability such as SD. Due to the number of human specific-species (9/35 genes) RMGFs identified, the frequency of RMGFs transcripts validated at a soft threshold of 90 PI (29/35) and the availability of high quality human SD sequence co-ordinates (with syntenic regions identified across a number of primate species) human RMGFs were assessed for their frequency within regions of human SD (Section 2.2.2). 8 human RMGFs were identified as being located within regions of human SD and their corresponding segmental co-ordinates are detailed in Table 2.4**,** only 6/8 of these genes had validated transcriptomic data (Figure 2.6) across vertebrate species, these six RMGF transcripts are highlighted in Figure 2.6. In order to assess the significance of this result 10,000 simulations were run on non-fused randomly selected human protein coding gene datasets of the same size as detailed in Section 2.2.2 and a p-value of 0.0282 was obtained

Table 2.4: Human RMGF transcripts located in known regions of human segmental duplication

Ensembl Gene ID	Chromosome Name	Gene Start (bp)	Gene End (bp)	Human Segmental Duplication Co-ordinates
ENSG00000025 5730	19	41350 853	41425 004	HIT(chr19: 41367420: 41589429: 41530045:41336556:41589232:41550041)
ENSG00000025 7390	12	55757 275	55827 546	HIT(chr12:9091898)
ENSG00000025 8465	1	16021 7464	16028 5130	HIT (chr14:77101031:79013537:103840290, chr1:236982858:102251812:154350593, chr3:32232169:110401053,chr2:62373433,chr12:68946695, chr19:24009887, chr8:81471035, chr13:21535374, chrX:44600255:99410808:86958272, chr13:67840884,chr4:81082204, chr15:53177729, chr16:73974469)
ENSG00000026 0342	16	18788 063	18801 519	HIT (chr16:16665156)
ENSG00000026 1740	16	29443 056	29454 651	HIT (chr16:21353861: 21839360: 22496825: 29490604: 30229945: 70239881: 18867664: 21896688: 22480785: 21464446)
ENSG00000024 9773	7	55887 277	55955 239	HIT (chr20:32989148, ch4:22728061, chr7:55705128, chr1:22417917)
ENSG00000025 0349	X	37349 330	38687 674	HIT (chrX:37060162:37363134:37060588: 36986405: 37349792: 37047257: 34476799: 36861477: 36737744, chr:17457999)
ENSG00000025 0588	3	15896 2235	15989 7366	HIT (chr1:231447389, chr2:187812964:187814690, chr6:111750916, chr13:95409580, chr6:118822652)

Human RMGF transcripts identified at a soft threshold of 90% identity and their frequency within regions of known human SD with coordinates obtained from the Human Segmental Duplication Database. Both the chromosomal co-ordinate of the RMGF transcript (Herrero *et al.*, 2016) *and the coordinate of SD are illustrated.*

2.3.4) A computational characterisation of RMGFs

2.3.4.1) A Functional Enrichment Analysis across RMGFs and their Parents

A functional characterisation experiment was carried out across both human and mouse RMGF families (Section 2.2.3.1). Experiments were carried out on both parents and RMGFs at all soft threshold levels in order to uncover any potential bias across RMGF families. RMGF enrichment analysis for 90%, 80% and 70% identity thresholds uncovered no statistically significant functional enrichment across either species.

At a 50 PI threshold RMGF transcripts showed an affinity for calcium binding (Table 2.7). Human RMGF parents identified using a 70 and 80 PI showed a statistical enrichment for binding functionalities specifically for DNA, protein, organic compound and heterocyclic compound binding. An enrichment for enzymatic function was also found (Table 2.5 and Table 2.6). A bias for binding activities was found in mouse RMGF parents identified at 70 PI, specifically for DNA, olfactory receptor, G-couple protein receptor, metal and cationic binding however, no significant enrichment for molecular function was found (Table 2.8). No other significant biological or molecular functional enrichments were identified across both mouse and human datasets.

Table 2.5: Functional enrichment results for human RMGF parents identified using at a soft threshold of 70% identity

GO Terms	Species	Fusion Family Member	PI threshold	Description	P-Values	FDR q-value
GO:0003677	Human	Parent	70%	DNA binding	7.41E-37	2.32E-34
GO:0003676	Human	Parent	70%	Nucleic acid binding	1.30E-31	2.03E-29
GO:1901363	Human	Parent	70%	Heterocyclic compound binding	1.08E-26	1.13E-24
GO:0097159	Human	Parent	70%	Organic cyclic compound binding	1.08E-26	8.45E-25
GO:0043169	Human	Parent	70%	Cation binding	1.53E-24	9.57E-23
GO:0046872	Human	Parent	70%	Metal ion binding	1.53E-24	7.98E-23
GO:0043167	Human	Parent	70%	Ion binding	1.01E-20	4.52E-19
GO:0005488	Human	Parent	70%	Binding	1.27E-12	4.99E-11
GO:0001071	Human	Parent	70%	Nucleic acid binding transcription factor activity	4.53E-08	1.58E-06
GO:0003700	Human	Parent	70%	Sequence-specific DNA binding transcription factor activity	4.53E-08	1.42E-06
GO:0003674	Human	Parent	70%	Molecular_function	8.18E-07	2.33E-05
GO:0005515	Human	Parent	70%	Protein binding	1.47E-04	3.85E-03
GO:0004175	Human	Parent	70%	Endopeptidase activity	3.73E-04	8.98E-03

Gorilla functional enrichment results for human RMGF parents found using at a soft threshold of 70% identity with associated p-value (Column 6) and FDR value (Column 7).

Table 2.6: Functional enrichment results for parent RMGFs of fusions identified using MosaicFinder with a soft threshold of 80% identity

GO Terms	Species	Fusion Family Member	PI threshold	Description	P-Values	FDR q-value
GO:0003677	Human	Parent	80%	DNA binding	1.02E-16	2.65E-14
GO:0003676	Human	Parent	80%	Nucleic acid binding	3.20E-13	4.16E-11
GO:1901363	Human	Parent	80%	Heterocyclic compound binding	2.24E-12	1.94E-10
GO:0097159	Human	Parent	80%	Organic cyclic compound binding	2.24E-12	1.45E-10
GO:0043169	Human	Parent	80%	Cation binding	1.03E-11	5.35E-10
GO:0046872	Human	Parent	80%	Metal ion binding	1.03E-11	4.45E-10
GO:0043167	Human	Parent	80%	Ion binding	5.39E-10	2.00E-08
GO:0005488	Human	Parent	80%	Binding	1.31E-06	4.25E-05
GO:0001071	Human	Parent	80%	Nucleic acid binding transcription factor activity	1.26E-05	3.63E-04
GO:0003700	Human	Parent	80%	Sequence-specific DNA binding transcription factor activity	1.26E-05	3.27E-04
GO:0003674	Human	Parent	80%	Molecular_function	3.76E-05	8.88E-04

Functional enrichment results for GOrilla analyses across parent RMGFs of fusions identified using MosaicFinder with a soft threshold of 80 PI. P-values for each enrichment are located in column 6 and their associated q-value in Column 7.

Table 2.7: Functional enrichment results for human RMGFs using MosaicFinder with 50 PI threshold

GO Terms	Species	Fusion Family Member	PI threshold	Description	P-Values	FDR q-value
GO:0005509	Human	Fusion Gene	50%	Calcium ion binding	4.54E-07	1.15E-04

Gorilla analyses carried out to investigate functional enrichment for human RMGFs identified using a MosaicFinder 50 PI threshold. Each enrichment has an associated p-value (Column 6) and q-value (Column 7).

Table 2.8: Results of mouse RMGF parent's functional enrichment analyses at a 70 PI threshold utilising the MosaicFinder software package

GO Terms	Species	Fusion Family Member	PI threshold	Description	P-Values	FDR q-value
GO:0043169	Mouse	Parent	70%	Cation binding	3.04E-21	2.73E-19
GO:0043167	Mouse	Parent	70%	Ion binding	3.04E-21	1.37E-19
GO:0046872	Mouse	Parent	70%	Metal ion binding	3.04E-21	9.11E-20
GO:0005488	Mouse	Parent	70%	Binding	4.51E-14	1.01E-12
GO:0001071	Mouse	Parent	70%	Nucleic acid binding transcription factor activity	8.67E-06	1.56E-04
GO:0003700	Mouse	Parent	70%	Sequence-specific DNA binding transcription factor activity	8.67E-06	1.30E-04
GO:0004984	Mouse	Parent	70%	Olfactory receptor activity	1.06E-04	1.36E-03
GO:0004930	Mouse	Parent	70%	G-protein coupled receptor activity	1.06E-04	1.19E-03

Results of a GOrilla functional enrichment analysis on mouse RMGF parents using a 70 PI threshold. P-values per enrichment are located in Column 6 and their corresponding FDR in Column 7.

2.3.4.2) An investigation SNPs and INDELs in human RMGFs

To further characterise RMGFs an analyses was carried out to uncover polymorphisms within human RMGFs identified at 90% identity. Here the RMGF dataset was analysed for the presence of known INDEL and SNPs identified across primate species as identified by the Primate Genome Project (Prado-Martinez *et al.*, 2013). Due to chromosomal co-ordinate availability only 29/35 genes were analysed for SNPs and INDELS. 4/29 RMGFs were found to contain SNPs and 2/29 genes contained an INDEL region. In the *Ttl3* gene (ENSG00000214021) a SNP was identified on chromosome 3 at position 9849914 (C-T). In the *Gli1* gene (ENSG00000111087) an INDEL was located on chromosome 12 at position 56151310 having frame-shifting consequences on the gene. The *AC020763.1* (ENSG00000260156) gene on chromosome 16 at position 69357846 contained an INDEL (GC) causing a splice-site creation within the gene. The *Mvd* gene (ENSG00000167508) contained two SNPS both of which were found on chromosome 16. The first SNP was identified at position 87248714 (C-T) and the second was found at 87251937 (A-T) and both SNPs cause non-sensible transcript creation. Finally, an INDEL (TTATTTTTTG) was identified in the *AC138028.3* gene (ENSG00000259813) on chromosome 16 at position 87312797 causing a frame-shift within the gene.

2.3.4.3) A motif enrichment analysis of RMGF

An investigation of motif enrichment across human RMGFs identified at 90 PI was carried out (Section 2.2.3.3). RMGFs at this threshold were significantly enriched for two motifs with a p-value < 0.05, namely EHF and NCAT. EHF is a transcriptional activating motif containing a GGAA sequence that controls gene expression in epithelial cells. The motif drives expression of TNFRSF10B/DR5 by promoter binding (Lim *et al.*, 2006). It has known functions in both proliferation and differentiation control within epithelial cells (Silverman *et al.*, 2002) as well as playing role in kinase signalling cascades (Tugores *et al.*, 2001).

The NCAT1 motif interacts with many immune related signalling cascades by cytokine interactions and has been shown to bind cell surface receptors, causing

aberrant WNT signaling pathways and causing an increased level of migration and invasion of cancer cells (Shimamura *et al.*, 1994).

2.2.4.4) Codon usage bias in RMGFs

An analysis of the multivariate codon usage across human RMGFs identified at 90 PI was carried out and compared to non-fused human protein-coding genes (of the same length) (Section 2.2.3.4).

A comparison of the calculated RSCU values of human RMGFs (Table 2.9) and RSCU values of randomly selected non-fused human protein-coding genes (Table 2.10) was carried out and highlighted columns are indicative of an RSCU difference between the two datasets. Results uncovered 7/64 tested had alternating codon usage patterns with Serine (UCU), Ileum (AUU) and Threonine (ACU), Arginine (AGA), Alanine (GCU) and Glycine (GGA) codons found at a lower than expected when compared to human non fused protein coding genes where they were identified at a greater frequency than expected. In RMGFs Valine (GUC) showed no evidence of codon usage bias whereas the in non-fused protein-coding genes it presented at lower than expected frequencies.

Table 2.9: GCUA Cumulative Codon Usage results for human RMGFs identified 90 PI threshold

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	178	(0.78)	**Ser**	UCU	165	(0.95)
	UUC	276	(1.22)		UCC	273	(1.57)
Leu	UUA	64	(0.26)		UCA	117	(0.67)
	UUG	187	(0.76)		UCG	53	(0.30)
Tyr	UAU	146	(0.83)	**Cys**	UGU	170	(0.85)
	UAC	207	(1.17)		UGC	229	(1.15)
ter	UAA	3	(0.00)	**ter**	UGA	17	(0.00)
ter	UAG	11	(0.00)	**Trp**	UGG	218	(1.00)
Leu	CUU	132	(0.54)	**Pro**	CCU	249	(1.16)
	CUC	322	(1.31)		CCC	307	(1.43)
	CUA	124	(0.51)		CCA	242	(1.13)
	CUG	644	(2.62)		CCG	60	(0.28)
His	CAU	139	(0.84)	**Arg**	CGU	72	(0.48)
	CAC	193	(1.16)		CGC	195	(1.30)
Gln	CAA	142	(0.42)		CGA	96	(0.64)
	CAG	531	(1.58)		CGG	171	(1.14)
Ile	AUU	157	(0.87)	**Thr**	ACU	165	(0.99)
	AUC	324	(1.80)		ACC	269	(1.62)
	AUA	59	(0.33)		ACA	153	(0.92)
Met	AUG	325	(1.00)		ACG	77	(0.46)
Asn	AAU	188	(0.81)	**Ser**	AGU	128	(0.74)
	AAC	277	(1.19)		AGC	307	(1.77)
Lys	AAA	249	(0.72)	**Arg**	AGA	147	(0.98)
	AAG	440	(1.28)		AGG	217	(1.45)
Val	GUU	108	(0.45)	**Ala**	GCU	286	(0.97)
	GUC	238	(1.00)		GCC	519	(1.76)
	GUA	87	(0.37)		GCA	260	(0.88)
	GUG	519	(2.18)		GCG	115	(0.39)
Asp	GAU	259	(0.79)	**Gly**	GGU	119	(0.48)
	GAC	393	(1.21)		GGC	352	(1.42)
Glu	GAA	338	(0.67)		GGA	232	(0.94)
	GAG	670	(1.33)		GGG	286	(1.16)

GCUA Cumulative Codon Usage results for RMGFs identified at a 90 PI threshold with highlighted values indicating a difference in RSCU value of the human RMGFs and the human non-fused protein coding gene dataset.

Table 2.10: Codon usage bias analyses of human non-fused protein coding gene transcripts.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	162481	(0.94)	Ser	UCU	145816	(1.12)
	UUC	183865	(1.06)		UCC	165956	(1.27)
Leu	UUA	75187	(0.48)		UCA	119636	(0.92)
	UUG	122734	(0.78)		UCG	42832	(0.33)
Tyr	UAU	114276	(0.90)	Cys	UGU	98685	(0.93)
	UAC	139272	(1.10)		UGC	113224	(1.07)
Ter	UAA	4651	(0.00)	ter	UGA	8331	(0.00)
Ter	UAG	3719	(0.00)	Trp	UGG	114236	(1.00)
Leu	CUU	125862	(0.80)	Pro	CCU	170423	(1.15)
	CUC	179445	(1.14)		CCC	189078	(1.28)
	CUA	67234	(0.43)		CCA	165311	(1.12)
	CUG	372055	(2.37)		CCG	67119	(0.45)
His	CAU	104408	(0.85)	Arg	CGU	43061	(0.48)
	CAC	141872	(1.15)		CGC	98013	(1.10)
Gln	CAA	120207	(0.53)		CGA	59708	(0.67)
	CAG	331310	(1.47)		CGG	110576	(1.24)
Ile	AUU	153327	(1.11)	Thr	ACU	127713	(1.01)
	AUC	190890	(1.38)		ACC	173897	(1.38)
	AUA	71434	(0.52)		ACA	144837	(1.15)
Met	AUG	203840	(1.00)		ACG	57114	(0.45)
Asn	AAU	165036	(0.96)	Ser	AGU	121036	(0.93)
	AAC	179117	(1.04)		AGC	188202	(1.44)
Lys	AAA	242876	(0.88)	Arg	AGA	113924	(1.28)
	AAG	306171	(1.12)		AGG	110312	(1.24)
Val	GUU	106478	(0.75)	Ala	GCU	175474	(1.06)
	GUC	133438	(0.94)		GCC	265202	(1.60)
	GUA	68630	(0.48)		GCA	154260	(0.93)
	GUG	262258	(1.84)		GCG	69887	(0.42)
Asp	GAU	215892	(0.95)	Gly	GGU	101380	(0.65)
	GAC	240820	(1.05)		GGC	209458	(1.35)
Glu	GAA	295760	(0.87)		GGA	156699	(1.01)
	GAG	387385	(1.13)		GGG	152435	(0.98)

GCUA Cumulative Codon Usage results for 40,000 randomly selected human non-fused protein coding genes.

2.3.4.5) RMGF family location across vertebrate species

35 RMGFs identified at a 90 PI threshold had the necessary data required to carry out a RMGF family location analyses, whereby the location of both the RMGF itself as well as it's two parent genes were examined for synteny across a dataset of 25 high quality vertebrate genomes. Results of the analysis are highlighted in Figure 2.7 and 28 of RMGFs parents were found to lie adjacent to each other in human, 16 in chimpanzee, 15 in gorilla and 1 set of RMGF gorilla parents were found on different contigs. 13 parents were identified as being syntenic in Gibbon but only 3 parents were found as having parents on different contigs.

A confounding issue of this experiment lay with the unavailability of the location of many of the RMGF parent orthologs across vertebrate species due to either no ortholog being currently available for the RMGF parent transcript or because the location of the ortholog has not yet been allocated a position in the species genome.

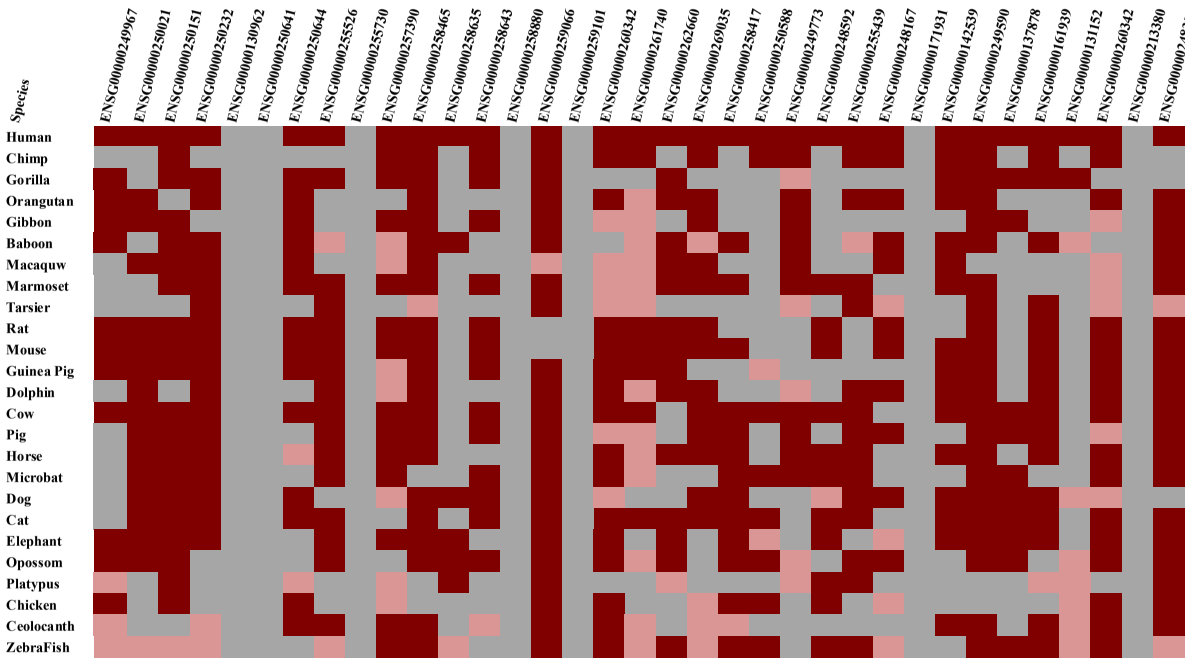


Figure 2.7: *Location of RNMF parent gene orthologs across a selected high quality vertebrate database: A location analysis carried out on human RMGF parent genes and their orthologs across a vertebrate dataset. Red cells indicate that the parent genes are found directly adjacent to one another in that species. Pale pink cells highlight parents are present but exist on different chromosomal contigs in that species. Grey cells represent cases where one or both of the RMGF parents are missing in that species.*

2.3.5) Rate heterogeneity and selective pressure analyses across candidate RMGFs

2.3.5.1) An investigation of evolutionary rates across RMGFs in comparison to non-fused protein coding genes of comparable length

An assessment was carried out across RMGFs in order to decipher whether they are evolving more quickly or slowly than non-fused protein coding genes (See Section 2.2.4.2). Rates of evolution are obtained through branch length calculation with the human rate shown in Table 2.11**,** chimpanzee in Table 2.12**,** mouse in Table 2.13, marmoset in Table 2.14, and orangutan in Table 2.15**,** and. Data indicate no significant difference was found when the evolutionary rate of RMGFs were compared to the rate of non-fused protein coding genes across all species tested. Indicating the new genes remodeled via RNA mediated gene fusion are not evolving at a rate that is significantly different to that of "normal" protein coding genes.

Table 2.11: Candidate human RMGFs branch length analysis results

	Averages per Species	Standard Deviation	P-Values	Significance
Human	0.023083425	0.021005279	0.34546918	No
Marmoset	0.200527			No
Mouse	0.125938	0.023697977	0.33054794	No
Rat	0.110326	0.013739792	0.32542951	No
Chimpanzee	0.0553071	0.042644379	0.80319013	No
Gibbon	0.0214536			No
Gorilla	0.0495997	0.011840927	0.32779482	No
Tarsier	0.4007065	0.14284052	0.36475072	No

Evolutionary rate analysis of human RMGFs compared to non-fused protein coding genes across a dataset of vertebrate species (Column 1) to determine a rate change between the two datasets. Significance is highlighted in Column 5. Empty cells are due to ortholog unavailability in that species.

Table 2.12: Branch length investigations of chimpanzee RMGFs compared to a panel of selected vertebrates

Averages per Species		Standard Deviation	P-Values	Significance
Human	0.023083425	0.021005279	0.34546918	No
Marmoset	0.200527			No
Mouse	0.125938	0.023697977	0.33054794	No
Rat	0.110326	0.013739792	0.32542951	No
Chimpanzee	0.0553071	0.042644379	0.80319013	No
Gibbon	0.0214536			No
Gorilla	0.0495997	0.011840927	0.32779482	No
Tarsier	0.4007065	0.14284052	0.36475072	No

Candidate chimpanzee RMGFs branch length analysis results. Branch length analyses comparing the average evolutionary rate of chimpanzee RMGFs to the average rate of other species analysed. Significance is highlighted in Column 5. Empty cells are due to ortholog unavailability in that species.

Table 2.13: Investigation results of a branch length analysis ran across candidate mouse RMGFs compared to a dataset of vertebrates

	Averages per Species	Standard Deviation	P-Values	Significance
Human	0.030033695	0.030665531	0.35386745	No
Marmoset	0.554099	0.0049629	0.55787322	No
Mouse	0.02521117	0.032980651	0.61947372	No
Chimpanzee	0.026572035	0.034731763	0.36190555	No
Macaque	0.5409064	0.691612092	0.5558049	No
Gibbon	0.0332717	0.025621802	0.34594112	No
Gorilla	0.029674055	0.03129067	0.35489984	No

Candidate mouse RMGFs branch length analysis results comparing selected mouse RMGFs to a panel of ortholog containing vertebrates. Significance is highlighted in Column 5.

Table 2.14: An evolutionary rate analysis of candidate marmoset RMGFs compared to a panel of high quality vertebrate species genomes

	Averages per Species	Standard Deviation	P-Values	Significance
Human	0.016484	0.00252324	0.32120343	No
Marmoset	0.086019925	0.090687787	0.8103296	No
Chimpanzee	0.01623405	0.002621881	0.32131656	No
Macaque	0.05364965	0.055469486	0.36790451	No
Gibbon	0.04490595			No
Gorilla	0.08860595	0.071740296	0.36825628	No

Results of an evolutionary rate analysis through branch length calculation comparing the average evolutionary rate of marmoset RMGFs to the average rate of other species analysed. Vertebrates assessed are found in column 1, missing orthologous information is represented by empty cells and significance is highlighted in Column 5.

Table 2.15: Results of a RMGF branch length analysis in orangutan compared to a panel of vertebrates

	Averages per Species	Standard Deviation	P-Values	Significance
Human	0.01825535	0.01971746	0.34711663	No
Marmoset	0.01102187	0.009555742	0.33552121	No
Mouse	0.024995825	0.023541105	0.34775562	No
Chimpanzee	0.012986185	0.005932081	0.32757014	No
Macaque	0.0111099	0.009677181	0.3356854	No
Orangutan	0.0376787	0.013256774	0.80146247	No

Candidate orangutan RMGFs branch length analysis results comparing the average evolutionary rate of orangutan RMGFs to the average rate of other species analysed (Column 1). Significance is highlighted in Column 5.

2.3.5.2) An investigation of the selective pressures acting on RMGFs

2.3.5.2.1) Lineage-specific selective pressure heterogeneity

Lineage-specific positive selection was detected in the orangutan RMGF parent ENSPPYG00000029627 using modelA (Table 2.15). Four sites were detected as being under positive selection according to the Bayes Empirical Bayes (BEB) estimation. However, 98% of sites were found to be under strict purifying selection. These sites are under positive selection in orangutan alone and are under purifying selection in all other background species tested (p_0=0.98213,p_1=0.00000, p_2=0.01787, p_3=0.00000, ω_0=0.02479, ω_1=1.00000, ω_2=33.33516).

Signatures of positive selection were detected in the chimpanzee lineage across a total of 8 sites when the RMGF ENSPTRG00000020751 transcript was compared against its parent gene ENSPTRG00000032572 (Table 2.16). In this alignment 33% of sites were found to be under strict purifying selection (ω_0=0.10990), 59% were found to be evolving neutrally (ω_1=1.00000) and ~7% were found to be under positive selection (ω_2>>1). Five BEB sites were identified in the region of the RMGF ENSPTRG00000020751 that mapped to its other parent gene, namely ENSPTRG00000032572. In this region of the RMGF homology, 38% of these sites were found to be under purifying selection (ω_0=0.20339), 55% of all sites are evolving neutrally (ω_1=1), and 6% of sites within the alignment are under lineage-specific positive selection in the chimpanzee-fused gene (ω_2=21.83685). This gene has a molecular function in hormone activity and positively selected sites are positioned within an Ilgf domain, which has been shown to have roles in growth, differentiation and reproduction, whilst on a cellular level, also contributes to cell cycle, migration, proliferation and differentiation (Eden *et al.*, 2009; UniProt Consortium, 2018).

A comparison of RMGF ENSG00000266953 with its corresponding parent gene region in ENSG00000105220 revealed three BEB sites indicating positive selection. The gene contains glucose-6-phosphate isomerase enzymatic activity and the positively selected residues are located 5 amino acids away from the

active site of this protein (Table 2.16) (Eden *et al.*, 2009; UniProt Consortium, 2018) In summary, 92% of the sites within the alignment are found to be under purifying selection (ω_0=0.01956), 3% of sites are evolving neutrally (ω_1 =1) and finally ~4% of sequences are found to be under positive selection in the human gene fusion alone and not in any other species (ω_2>>1) (Table 2.16).

Table 2.16: Lineage-specific CodeML positive selection BEB results across candidate RMGF families

	Fusion Family Gene	Lnl	Parameter Estimates	Positive Selection	BEB
Orangutan	ENSPPYG00000029627 Parent Gene Uncharacterised	- 1459.529982	p_0=0.98213 p_1=0.00000 p_2=0.01787 p_3=0.00000 ω_0=0.02479 ω_1=1.00000 ω_2=33.33516	Yes	3 > 0.50 0 > 0.95 1 > 0.99
Chimp	ENSPTRG00000020751 RMGF (Parent 1 alignment) Uncharacterised	-557.052657	p_0=0.33452 p_1=0.59186 p_2=0.02659 p_3=0.04704 ω_0=0.10990 ω_1=1.00000 ω_2=999.0000	Yes	1 > 0.50 2 > 0.95 0 > 0.99
Chimp	ENSPTRG00000020751 RMGF (Parent 2 alignment) Uncharacterised	- 1585.987559	p_0=0.38767 p_1=0.55309 p_2=0.02441 p_3=0.03483 ω_0=0.20339 ω_1=1.00000, ω_2=21.83685	Yes	5 > 0.50 0 > 0.95 0> 0.99
Human	ENSG00000266953 RMGF Uncharacterised	-482.355021	p_0=0.92397 p_1=0.03145 p_2=0.04311 p_3=0.00147 ω_0=0.01956 ω_1 =1.00000 ω_2=999.00000	Yes	0 > 0.50 1 > 0.95 2 > 0.99

Results of a lineage specific selective pressure analysis across a select dataset of human, chimpanzee and orangutan species, BEB site results are contained within column 6 and significant p-values are determined in Column 5.

2.3.5.2.2) Site-specific selective pressure variation of candidate RMGFs

Site-specific selective pressure analyses showed only statistically significant results using discrete models. Simulation studies have shown that results from the discrete models are not as reliable as those found using the more sophisticated model m8 and therefore results are only briefly stated here. The RMGF gene ENSG00000167447 showed signatures of positive selection when aligned to its parent gene ENSG00000265303 (Table 2.17). In total, 23 sites were identified under the m3k2 model (BEB), 41% of sites were found to be under strict purifying selection (ω_0=0.00000), 53% were evolving neutrally (ω_1=0.63648), and 4% of sites contain signatures of positive selection (ω_2=4.40273). Many of these sites are positioned adjacent to one another (aligned codon positions: 12-15, 17-20, 25-29 39, 40, 53, 55, 56, 63, 65-69) however no domain information is available at present for this protein and it has been suggested to be under non-sense mediated decay (UniProt Consortium, 2018). When the appropriate region of the fused gene ENSG00000167447 was aligned to it's second parent gene, the m3k3 model and BEB estimation predicted two sites under positive selection (Table 2.17). These sites are also positioned adjacent to one another (68-69). Signatures of site-specific positive selection were detected in RMGF ENSPTRG00000020751 (ω_1=1.64072) under the m3k2 model, with 31 sites identified as positively selected using Naïve Empirical Bayes (NEB) (13-15, 17, 21, 25, 30, 34-35, 37, 47, 52 ,61, 63, 73, 81, 87, 91-98, 102-103, 108, 109, 112-113) (Table 2.17). This gene is a member of the insulin family of proteins, and has been found to play a role in hormone activity and predicted to regulate key biological processes such as gene expression, angiogenesis and enzymatic activity within cells (UniProt Consortium, 2018). Finally, the ENSPPYG00000012073 RMGF (parent gene to RMGF ENSPPYG00000012137) has evidence of positive selection in 2% of sites (ω_1=3.84880), with 98% of sites under purifying selection according to m3k3 (ω_0=0.08236). Two sites were identified as under site-specific positive selection using NEB (11, 40) (Table 2.17). All 5 sites are located in a Vkc domain within the catalytic subunit of the protein that contains vitamin K epoxide reductase activity (UniProt Consortium, 2018).

Species	Fusion Family Gene	Model Type	Lnl	Parameter Estimates	Positive Selection	NEB
Human	ENSG00000255439 Parent Gene Uncharacterised	m3Discrtk2	-611.316687	p_0=0.92129 p_1=0.07871 ω_0=0.06998 ω_1=1.67921	Yes	3 > 0.50 0 > 0.95 2 > 0.99
Human	ENSG00000167447 Parent Gene 1 SMG8	m3Discrtk2	-541.411807	p_0=0.41861 p_1=0.53361 p_2=0.04778 ω_0=0.00000 ω_1=0.63648 ω_2=4.40273	Yes	15 > 0.50, 6 > 0.95 2 > 0.99
Human	ENSG00000167447 Parent Gene 2 SMG8	m3Discrtk3	-540.206954	p_0=0.41861 p_1=0.53361 p_2=0.04778 ω_0=0.00000 ω_1=0.63648 ω_2=4.40273	Yes	1 > 0.50 1 > 0.95
Chimp	ENSPTRG00000020751 Parent Gene Uncharacterised protein	m3Discrtk2	-568.521978	p_0=0.37699 p_1=0.62301 ω_0=0.10174 ω_1=1.11245	Yes	9 > 0.50 11 > 0.95 7 > 0.99
Chimp	ENSPTRG00000020751 Parent Gene Uncharacterised protein	m3Discrtk2	-1589.12407	p_0=0.68216 p_1=0.31784 ω_0=0.38666 ω_1=1.64072	Yes	29 > 0.50 1 > 0.95 3 > 0.99
Orangutan	ENSPPYG00000012073 Parent Gene Uncharacterised	m3Discrtk2	-728.572752	p_0=0.97237 p_1=0.02763 ω_0=0.08236 ω_1=3.84880	Yes	0 > 0.50 1 > 0.95 1 > 0.99

Table 2.17: *Site-specific CodeML positive selection NEB results across candidate RMGF families: Significant results from a lineage –specific positive selection analysis carried out using the CodeML software package. NEB sites are located in Column 6 and significance is determined by the p-value in Column 4.*

2.3.6) An assessment of the accuracy of the Ensembl Genome Browser's ortholog annotation pipeline

As remodeled genes such as RMGFs are composed of pre-existing sequences, they're prone to misidentification from their parent genes by Genome Browser Orthology databases as the presence of a parent gene is often interpreted as presence of its corresponding RMGF. From SSN networks, PRANK alignments, BLASTs (Section 2.2.1) and follow up transcriptomic validation outlined in Section 2.2.1.5 a total of 9 RMGFs were identified as being human-specific. However, when analysed using the Ensembl Genome Browser's orthology database two of these genes were found to have orthologs across all three species analysed (Figure 2.8) with another found to have an ortholog in gorilla but not chimp. This suggests that the orthology pipeline implemented by the Ensembl Genome Browser has issues deciphering between RMGFs and their corresponding parent genes.

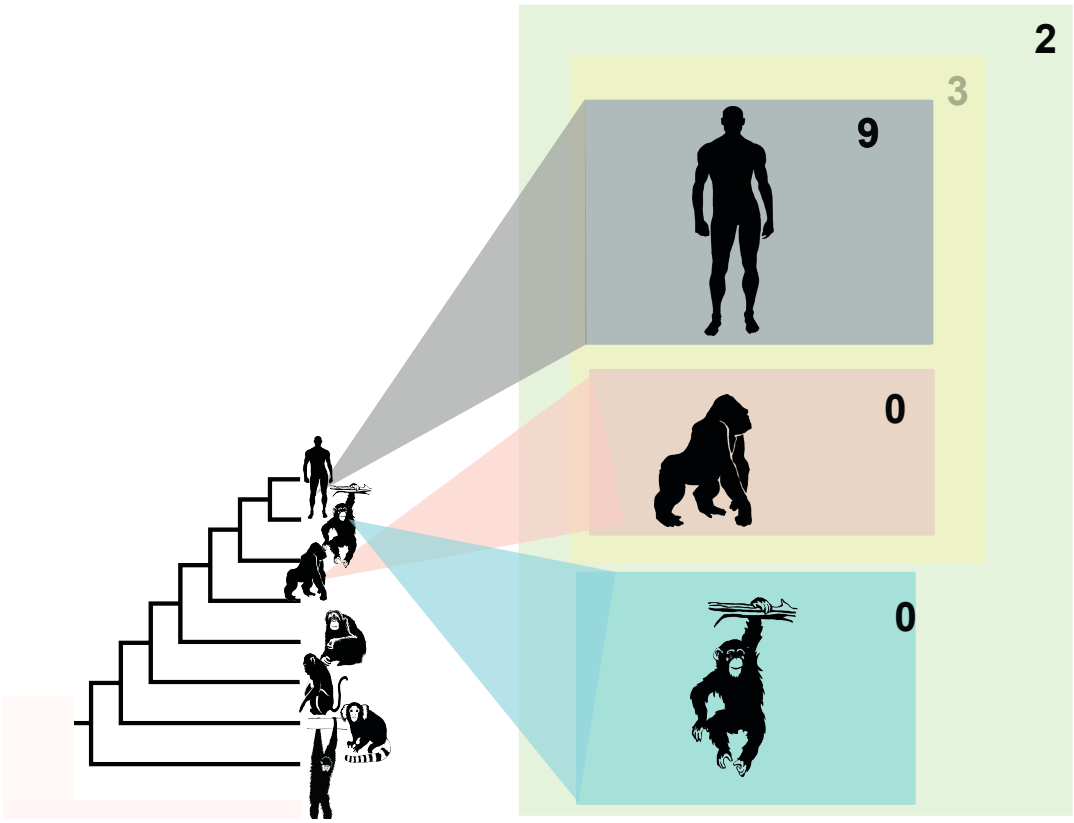


Figure 2.8: *An assessment the Ensembl Genome Browser's orthology pipeline across human, chimpanzee and gorilla species: Results of orthology analysis carried out between identified human-specific RMGFs and the Ensembl Genome Browsers Orthology Database* (Herrero *et al.*, 2016). *The phylogeny on the left hand side of the diagram shows the relationship between human's closest living ancestors. The grey box illustrates the 9 human specific fusion genes identified in Section 2.2.1***,** *the pink box represents how many of these RMGFs were identified and validated in gorilla, and the blue box represents those in chimpanzee. The yellow box represents number of human specific fusions with shared orthologs in gorilla according to the Ensembl Genome Browser* (Herrero *et al.*, 2016). *The green box represents RMGFs orthologs shared across all species according to the Ensembl Ortholog Database* (Herrero *et al.*, 2016).

2.4) Discussion

Prior to the advancement of sequence similarity network (SSN) generating algorithms the identification of fusion genes was challenging due to computational intensity, low sensitivity and robustness levels. MosaicFinder was selected as the fusion gene detection software package of choice due to its conservative nature, lower false positive rate, algorithm sophistication (CMS algorithm), user-friendly fusion family description and recall rates. Earlier in 2018 a new algorithm became available, CompositeFinder, this package is less conservative but carries out analysis at an increased speed. A key objective of our analysis is the detection of RMGFs therefore the usage of a more conservative algorithm such as MosaicFinder is more appropriate given the question being asked of the data. SSN software packages are not limitation free as they can only detect gene fusion events when both fusion parents still exist in the genome being analysed and this dismisses cases where one or both of the parents have become pseudogenised and therefore potentially misses a cohort of interesting fusion genes. Fusion genes were identified across a dataset of six primates and mouse (an out group) at 5 different percentage identity thresholds to determine fusions of more recent origins (highly similar to parent genes – more stringent thresholds) from those that are fast evolving or indeed more ancient (more dissimilar to parent genes – less stringent thresholds). The frequency of fusions detected was found to have an inversely proportionate relationship to the soft percentage identity (PI) used during the analysis or simply, genes with a lower PI had a higher frequency of fusions detected as the frequency of fusions detected at 70% was much greater across all species analysed in comparison to frequencies found using a 90 PI threshold.

A focus was placed on fusion genes detected at 90 PI as all fusion genes identified here were generated via an RNA-mediated fusion mechanism. This sacrificed potentially interesting (more ancient or faster evolving) fused genes detected at lower PI thresholds. The detection pipeline identified 42 human RMGFs at a threshold of 90%, 9 occurred after the divergence of human from the Great Ape species. Three species-specific RMGFs were also identified in mouse. Increased frequencies of RMGFs were identified across human however

this result could be a false representation caused by comparing the high quality human genome to other primate genomes of inferior quality. Eichler *et al* recently released new non-human based chimpanzee and orangutan genome (65x) sequences utilising long read SMRT sequencing platforms obtaining better transcript annotation calling and resolution at over repetitive elements. Using of these genomes more accurate comparisons can potentially be made between human and our Great Ape ancestors (Kronenberg *et al*., 2018).

An analysis of RMGF location revealed an enrichment in regions of SD and this enrichment is supported by individual cases identified in the literature whereby new genes/gene families were created *via* gene fusion events in SD regions. Examples include **1)** the generation of 2 novel fusions (*PMCHL1* and *PMCHL2*) from a SD event on human chromosome 5 (Courseaux and Nahon, 2001) , **2)** the morpheus gene family expansion due to a 15 Mb SD on human chromosome 16 (Johnson *et al.*, 2001) and **3)** a *FAM90A* family expansion after the orang-utan divergence via an SD on human chromosome 8 (Bosch *et al.*, 2007) . 6/29 RMGFs analysed were found polymorphic across humans this could be as a result of their enrichment in regions of genomically unstable regions such as SD.

The location of RMGF's parents across vertebrate orthologs revealed that these genes are more likely to reside adjacent to each other than separated on the same chromosome or indeed on entirely different chromosomes. Most follow this pattern despite no RMGF being identified in that species suggesting that RMGFs are more likely created by a mutation forming a new or deactivating an existing splice site or even deactivating a terminating signal between the two parental genes.

A functional enrichment for binding and enzymatic activities was found across both human and mouse RMGF parent genes, these data were supported by follow-up MEME motif assessments that indicated an EHF and NCAT motif enrichment both of which are associated with binding activities. The enrichment of binding and catalytic function is expected amongst fusion genes particularly when taken in the context of cancer malignancies caused by "oncogenes".

Oncogenes are predominantly created *via* ectopic fusion gene generation. The most common oncogenes found in carcinomas are TF or tyrosine kinase fusion genes containing DNA-binding and catalytic activity respectively (Teixeira, 2006). Also, the *ETS* TF fusion gene family is present in >50% of prostate cancers indicating the fusion genes with binding activities are abundant in the human genome (Rostad *et al.*, 2007; Linn *et al.*, 2016). The analysis was limited to RMGF parent genes due to the low number of RMGFs causing insufficient power. Codon usage bias was identified across 7 codons in RMGFs in comparison to un-fused protein coding genes in the human genome. These codons were found at lower rate than expected, for instance TA nucleotide levels were reduced and therefore are enriched in CG residues at positions 2, 3, and 4 (typical of the human genome (Karlin and Mrázek, 1996)) and are associated with a lower ranges of expression across tissues (Kotlar and Lavner, 2006) . This is in line with the "out-of-testes" hypothesis (Levine *et al.*, 2006; Kaessmann, 2010; Villanueva-Cañas *et al.*, 2017) and many proposals stating that new genes tend to have limited expression profiles (Hou *et al.*, 2012; Lan and Pritchard, 2016; Guschanski, Warnefors and Kaessmann, 2017) Human RMGFs appear to be evolving at a rate consistent with that of human non-fused protein coding genes, this finding was expected as this analysis was run on RMGF parents that still produce their own individual transcripts. However, the identification of lineage-specific positive selection in a single human, chimp and orang-utan RMGF was unexpected as positive selection here also means that the parent gene is under selection pressures. This could be due to adaptive pressures perhaps being placed on these genes to reside in a new cellular compartment.

Chapter 3: Transcriptomic and Translatomic Profiles of RNA-Mediated Gene Fusions

3.1) Introduction

New gene genesis and its transcriptional and translational contribution to phenome changes has been extensively studied across vertebrates, specifically primates (Journal and Society, 1996; Nowick *et al.*, 2009; Kaessmann, 2010). The focus of much literature has been placed on new gene acquisition and usage profiles across primate brain tissues due to an increased brain size (prefrontal cortex and cerebellum specifically) and an enhancement of cognitive abilities found in human in comparison the Great Apes (Barton and Venditti, 2014). The expansion of brain size in human is correlated with an increased size and surface area (lamination) of the cerebellum (Barton and Venditti, 2014). New genes have been linked to co-ordinating this expansion *e.g.* 14mya *NOTCH2* underwent a partial gene duplication creating an inactive duplicate – *NOTCH2NL*, 11my later the gene gained transcriptional capabilities and was subsequently duplicated twice more in human. This gene is known to control the freezing of stem cell division and the triggering of neuron cell differentiation (Suzuki *et al.*, 2018). Secondly, in an analysis of transcription factor (TF) usage between human and chimpanzee brains the *ZNF717* human-specific gene duplication was shown to link two independent modules/clusters in a brain interaction network. The first module contained predominantly TFs and chromatin re-modellers whilst the second contained energy and metabolic gene TFs (Nowick *et al.*, 2009). Both *NOTCH2NL* and *ZNF717* highlight the ability of new gene to contribute to phenotype diversity in human cerebellum in comparison to the Great Apes.

Here, an assessment of how RMGFs transcription and translation profile impact vertebrate phenomes, with a specific focus on primate species, was carried out. Primates were selected as a focus group due to their shallow divergence time, high quality data and obvious phenotype differences (Rogers and Gibbs, 2014). From the literature it is evident that RMGFs can alter their own transcriptional profile but also of the genes within their environment. For instance, RMGF can cause non-sense mediated decay *via* frame-shift mutation occurrence at the fusion breakpoint resulting in pre-mature stop codon formation, which ceases transcription and forces transcript degradation (Neu-Yilik *et al.*, 2011). Also, by

RMGF acquisition of the first 5'exon of an adjacent gene both the transcriptional machinery and profile of expression of that gene can be gained (Akiva *et al.*, 2006). Thirdly, adjacent gene transcription profiles can be obtained through transcriptional leakage whereby the ribosome skips the correct transcription termination signal causing a RMGF read-through event and potential novel expression profile acquisition (Nacu *et al.*, 2011). Examples of the transcriptional impact of gene fusions on the phenome exist in the literature these include; the *MYB-NFIB* fusion identified in cystic carcinomas that results in the loss of 2 miRNA binding sites (miR15a/16 and miR150) causing elevated MYB protein levels leading to tumorigenesis (Persson *et al.*, 2009); the *TMPRSS2-ERG* fusion that causes abnormally high ERG levels causing androgen independence and cancer (Perner, 2005); the *FGFR3-TACC3* fusion resulting in the loss of miR99A binding sites triggering aneuploidy by mitotic interference and consequently causing tumour formation (Parker *et al.*, 2013). These cases highlight the propensity of RMGFs to cause transcriptional changes within genomes resulting in drastic phenotype consequences.

However, RMGFs also contribute phenotypically on the level of translation for instance, the *STON-GTF2A1L* (Upadhyaya, Lee and Dejong, 1999), NM23-LV (Valentijn, Koster and Versteeg, 2006) and *Kua-Uev1* (Thomson *et al.*, 2000) are known examples of fusion genes that produce viable protein products. These protein products can potentially contribute to phenotypic changes within the species they exist. The *EML4-ALK* fusion has been identified in non-small cell lung cancer, the fusion causes the auto-phosphorylation of ALK by a dimerization domain in *EML4*. This dimerization event results in a constitutively active *ALK* causing malignancy (Choi *et al.*, 2008). The BCR-ABL gene (characterises chronic myeloid lymphoma (CML)) illustrates the importance of fusion as a mechanism to create phenotypes. Three versions of the fusion protein exist and determine cancer prognosis; 1) *BCR-ABL* p210: occurs in classic CML resulting from the fusion of a ~5-8kb region in *BCR* known as major breakpoint cluster region with *ABL*, 2) *BCR-ABL* p190: rarely occurs in CML patients but occurs in 2/3 of acute lymphoblastic lymphoma patients (ALL). Fusion forms by the removal of 2 exons from the *BCR* gene prior to its *ABL* fusion creating a

smaller protein, and 3) *BCR-ABL* p230: occurs in benign CML cases (CNL) fuses *BCR* gene that includes 2 exons and thus produces a larger protein (Journal and Society, 1996). This example demonstrates the consequence of fusion genes and their impact on the phenome and therefore an understanding of the likelihood of RMGFs to create such proteins is crucial for our understanding of their contribution to the genome and phenome evolution.

In this chapter we investigate the transcription and translation profiles of RMGFs using a variety of approaches. In the first instance we use high quality RNAseq datasets from the literature (Brawand *et al.*, 2011) and perform an in-depth analyses of the expression of RMGFs and their parents, comparing expression profiles between genes, species and tissues. We then perform qRT-PCR using a limited number of samples from gorilla and chimpanzee obtained from the zoo at Barcelona and using human tissue panels, this is to capture a different origin of the sample and a different time point. Finally we examine the available ribosome profiling datasets to determine if RMGFs have evidence for translation.

3.2) Materials and Methods

3.2.1) An assessment of the RMGF transcriptomic profiles

3.2.1.1) Preparation and quality control of published RNAseq data (Brawand *et al*, 2011)

RNA sequencing data (project number SRP007412) was downloaded from the SRA archive (Leinonen *et al*., 2011) for all seven species in the dataset (*i.e.* Human, Chimpanzee, Macaque, Marmoset, Gorilla, Orangutan, and Mouse), and for six tissues: brain, cerebellum, kidney, heart, liver and testis. Reads were predominantly 76 base pair single-end sequences (paired-end sequences were discarded due to poor quality). SRA files were converted to SAM format using the SRA toolkit (Leinonen *et al*., 2011) and then to FASTQ format using SAMtools (Li *et al.*, 2009). Reads were quality checked using FASTqc (Andrews, 2010). The following characteristics of sequence reads were determined per base: sequence quality; quality/phred scores; sequence content; GC-content and distribution N content; and per sequence for nucleotide

distribution across reads; length distribution; over-represented sequences; kmer and duplication content. Phred scores were low for all reads because the IBIS base caller had been used in the initial study. Reads with phred scores < 20 were removed. The leading 10-13 bases of each sequence read were of poor quality, possibly due to presence of adaptor sequences, and they were trimmed using TrimGalore **(Version 0.4.1)** (Martin, 2011). Human FASTqc phred score results (Figure 3.1) highlight that post trimming the sequence quality of the reads has significantly improved.

Finally, reads were again inspected by FASTqc. Example output for chimpanzee reads (SRR306811) are shown in Figure 3.2 – Figure 3.7 were the quality of reads over the start of each read has been significantly improved, contains no bias in GC% distribution across the dataset, has low/no N base calls, with evidence for slight duplications within the reads and a bias for GC content toward the end of each read.

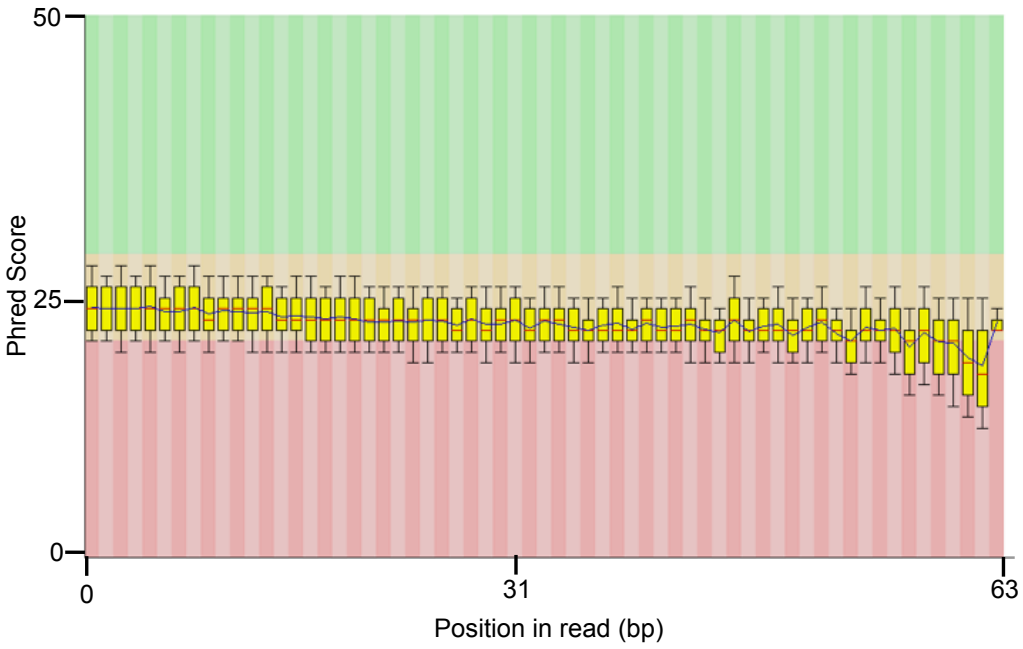


Figure 3.1: *FASTqc quality assessment of Human RNAseq reads after adaptor trimming implementation: Quality of human reads after poor quality read filtration and post trimming of the adaptor sequence (13bp). Y axis depicts the phred (quality) score while the X axis highlights the position in the read in bp.*

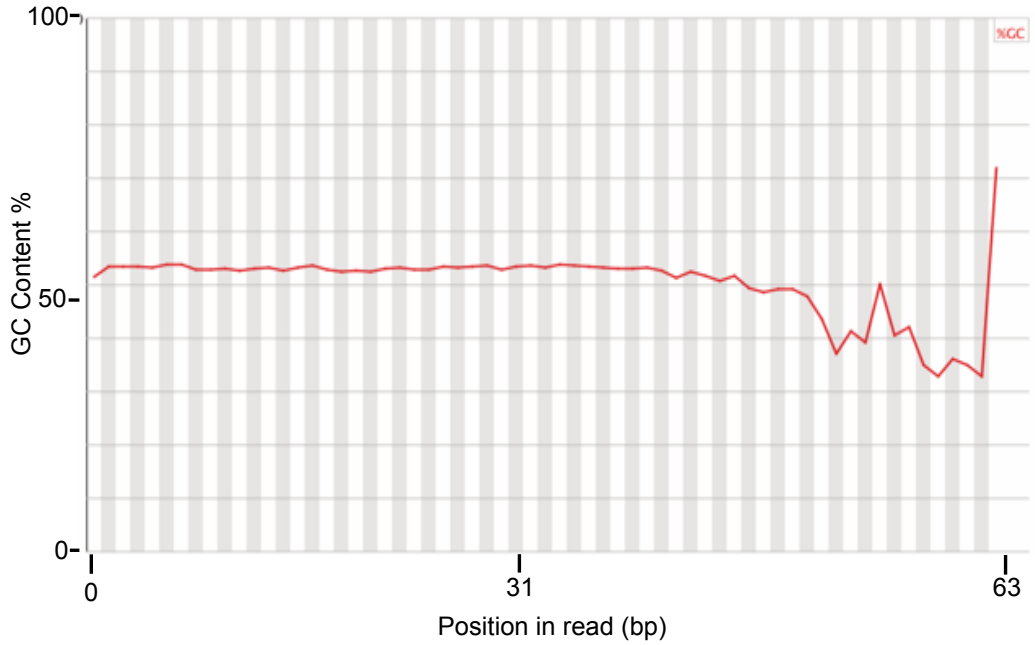


Figure 3.2: *Depiction of per base GC content analysis carried out by the FASTqc software package in chimpanzee RNA sequencing data. A graph illustrating the GC% across reads in the chimpanzee dataset. For high quality reads one would expect a perfectly straight line. A sporadic line is indicative of GC bias and poor quality data. Y axis illustrates the %GC content and the X axis shows the position in the read.*

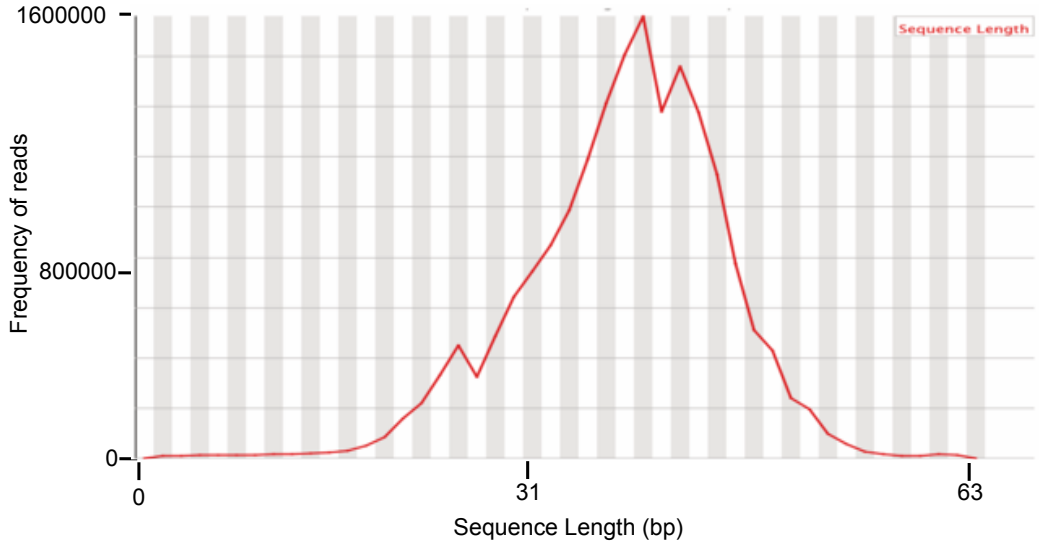


Figure 3.3: *FASTqc analysis of sequence length across chimpanzee RNA sequence reads: An illustration of the read length within the chimpanzee dataset. The distribution found has several peaks suggesting multiple sequence lengths in the data, however the majority are the same length with one much larger peak evident.*

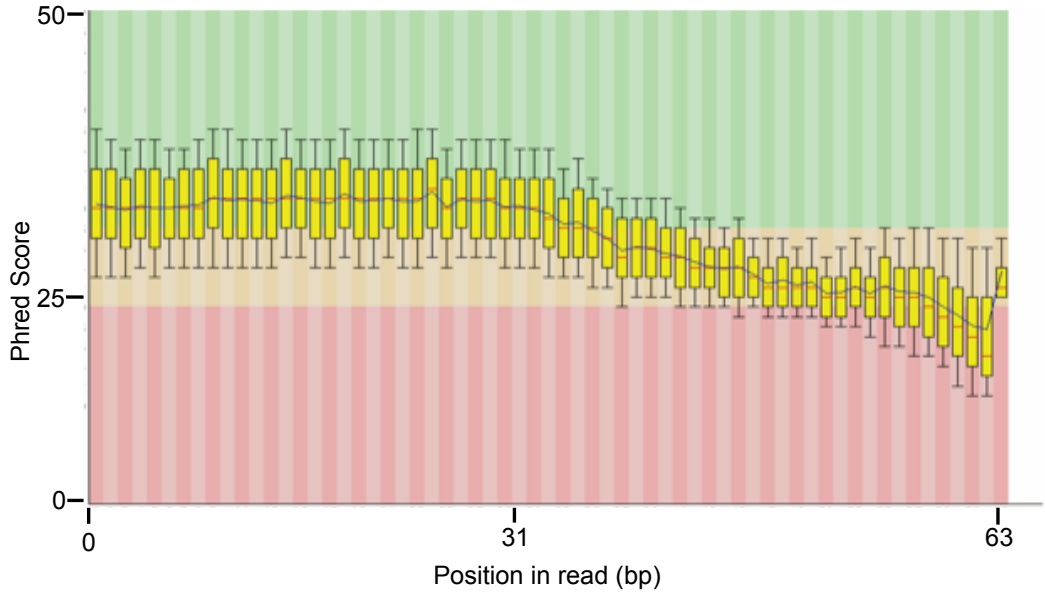


Figure 3.4: *A sequence quality across chimpanzee RNA sequencing reads by FASTqc: Graphical representation of the phred score/sequence quality of the dataset per read (Y-axis), with high quality sequences being placed inside the green background, intermediate quality inside the orange background and poor quality reads inside the read background. The red line is representative of the mean phred score at that position in the read and the blue line indicates the mean quality across all sequences assessed.*

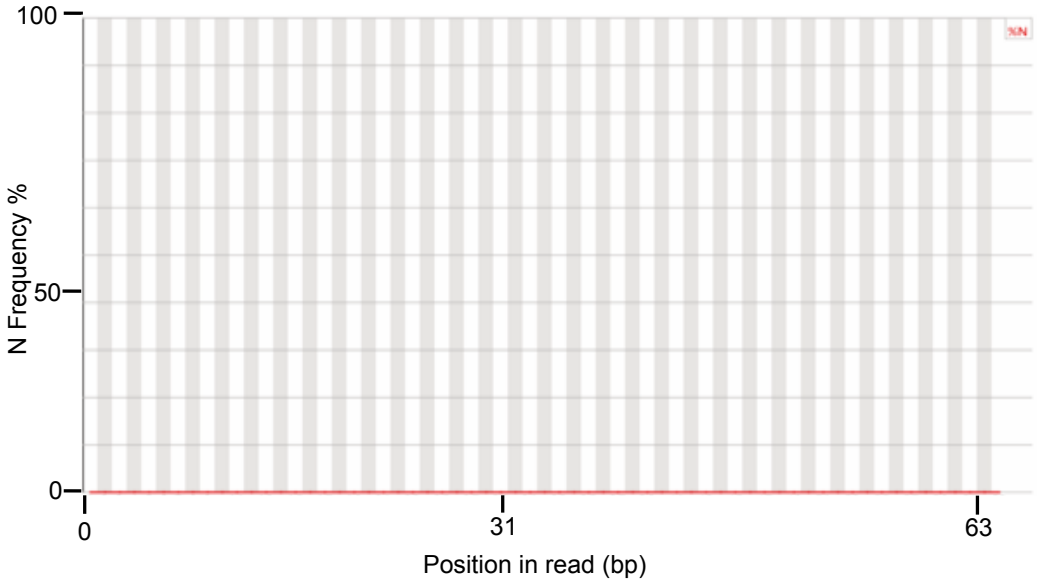


Figure 3.5: *FASTqc analysis of chimpanzee RNA sequencing dataset post-trim for uncalled bases: An assessment of uncalled or 'N' bases (Y axis) across each read (X axis). This graphic illustrates that no reads contain N sequences suggesting that all bases were mapped.*

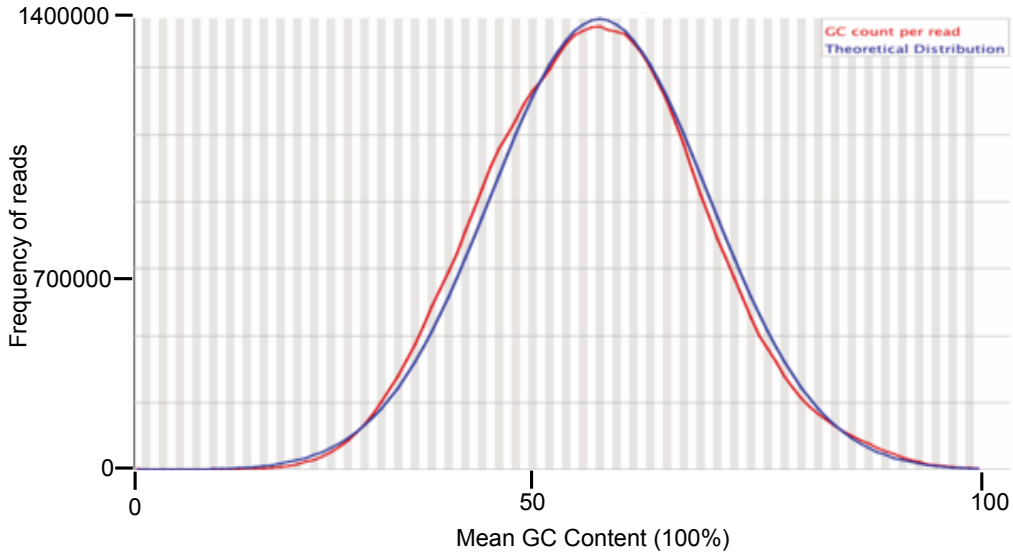


Figure 3.6: *A post adaptor trim quality assessment of GC content per read of the chimpanzee dataset: An analysis of GC content distribution across the entire read. The blue line illustrates a normal distribution created by a randomly generated dataset and the red line represents the distribution of GC in the chimpanzee dataset. As both the red and blue line follows the same distribution this suggests there is no bias in this distribution.*

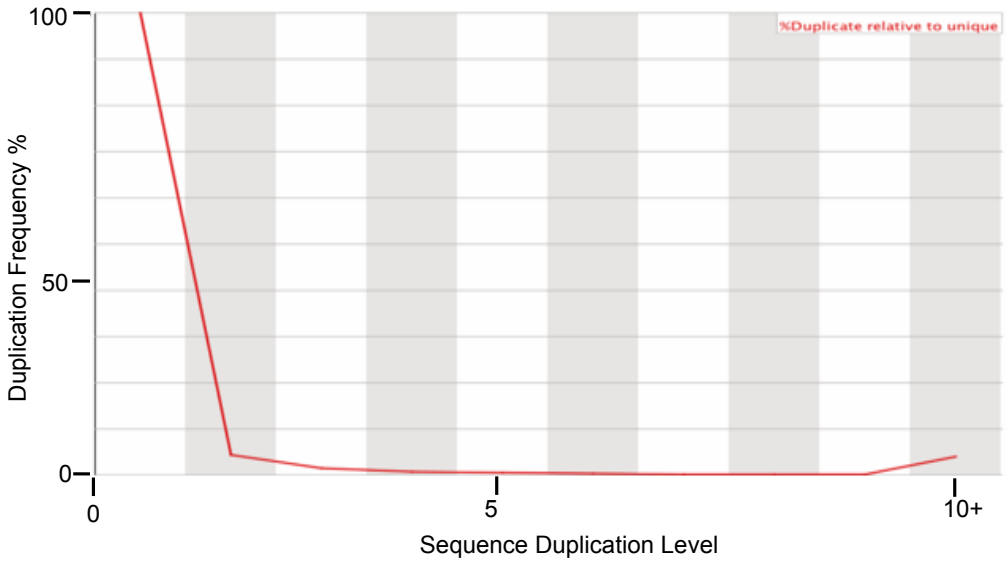


Figure 3.7: *Sequence duplication FASTqc result of the chimpanzee RNA sequence dataset post adaptor trim: An analysis of sequence duplication rate across reads within the chimpanzee data. A rise in the final category of the graph indicates a high duplication level in the dataset. Here a slight rise is shown indicating a slight elevation of duplications across the data.*

3.2.1.2) Mapping of RNA sequence reads to reference genomes

The reference genomes for human, chimpanzee, gorilla, orangutan, macaque, and mouse were downloaded from the Ensembl Genome Browser (**Version 74**) (Herrero *et al.*, 2016) and a selected panel of 36 human RMGFs transcripts were selected from Section 2.2.1 at the 90 PI threshold (Only 36/42 identified at the 90 PI threshold exact sequence was available from the Ensembl Genome Browser (Herrero *et al.*, 2016). RMGFs were also obtained for the 29/35 chimpanzee RMGFs, 18/24 gorilla RMGFs, 25/31 orangutan RMGFs, 34/42 macaque RMGFs, and 14/15 mouse identified at the 70 PI threshold (70 PI threshold was utilised due to the low frequency of RMGFs at 90% across these species). RMGF parents were analysed with 502 parents being assessed in human, 248 in chimpanzee, 305 in gorilla, 431 in orangutan, 569 in macaque and 285 in mouse.

Here, each reference genome was initially mapped to it's corresponding species cleaned transcriptome (prepared in Section 3.2.1.1) using STAR (Dobin *et al.*, 2013). This prevented humanisation of results. After reference mapping fake reads were then mapped to the transcriptome of each species in order to identify RMGF transcript production.

STAR was selected over TopHat (Trapnell *et al*., 2009) as it is 50 times faster, has a higher mapping frequency and precision of mapping (Engström *et al.*, 2013). STAR is a standalone C++ software package that carries out mapping in two steps; step one uses seed searching followed by a subsequent stitching and clustering step. Seed searching involves Maximal Mappable Prefix (MMP) whereby each read is mapped base-by-base until an alignment can no longer be formed. The remaining unaligned reads are then searched against the remaining genome for a hit. By mapping reads only to the unaligned portion of the genome mapping time is significantly reduced. The clustering stage takes all reads mapped during seed searching and stitches them together these become known as anchor seeds. A second round of mapping then ensues and if the second round of mapped reads are within a specified window size from the first round they are stitched together for transcript production.

In the case of RMGFs, only reads that span the junction/breakpoint of both parents were mapped and read counts after each stage of the analyses are highlighted in Figure 3.8 and illustrate the success of the mapping protocol across all species examined. Reads that mapped successfully were then counted on a species-by-species basis. For each species, the genome annotation file (".gtf") was downloaded from the Ensembl Genome Browser (Herrero *et al.*, 2016). HTseq Count software package (**Version 0.5.3**) (Anders, Pyl and Huber, 2015) was used to identify the reads that mapped to annotated transcripts and to count the number of reads mapped per transcript. The union overlap resolution method (Ander *et al*., 2015) was used to ensure the accurate counting of RMGFs and their parents as separate entities. Transcripts containing >1 mapped read spanning the fusion breakpoint were considered expressed, however analyses were also carried out for reads >3, and at >5, mapped reads as reads with >=5 read successfully mapped were more supported than those with >=1 read mapped.

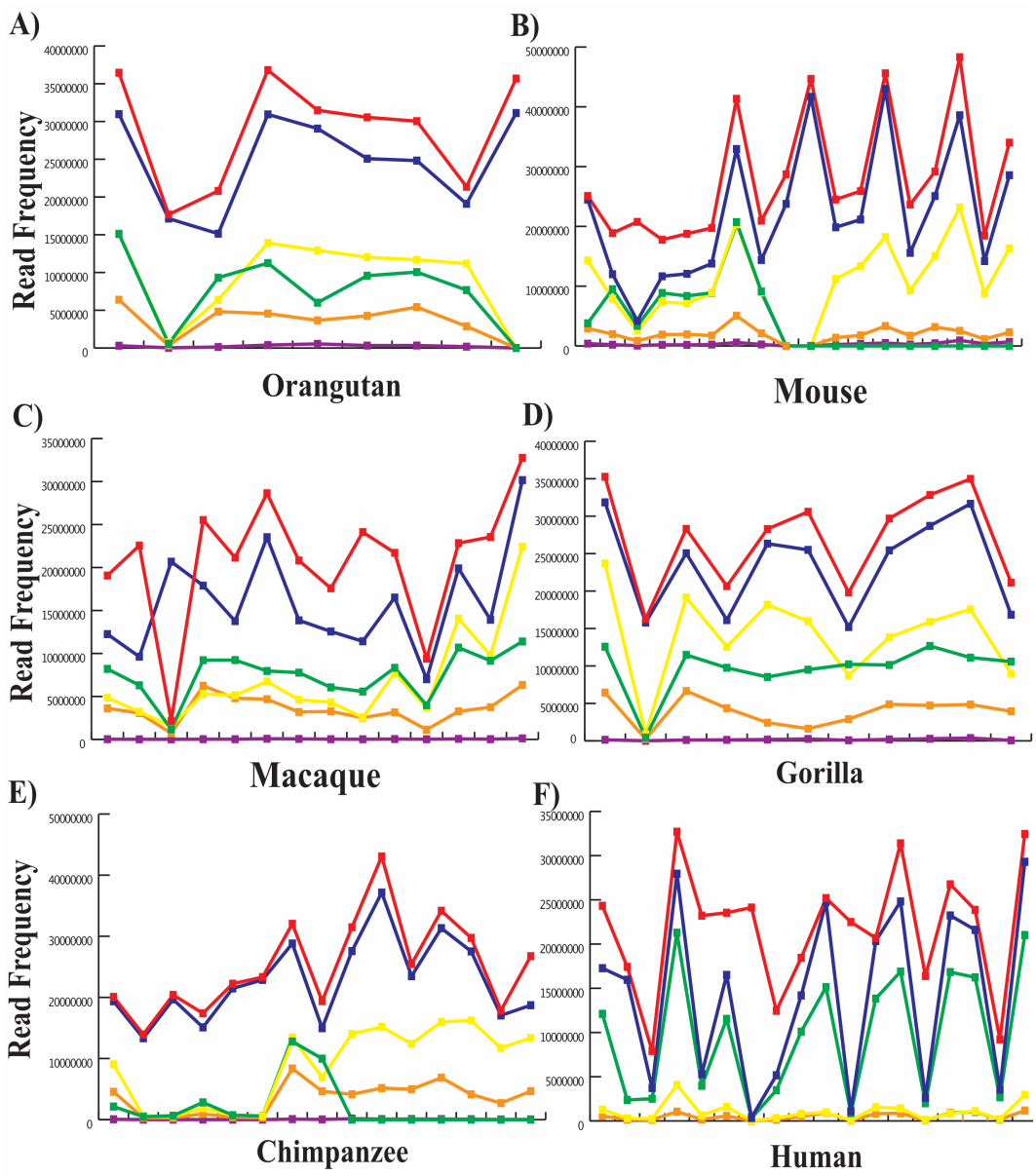


Figure 3.8: *Read counts taken after each step of the mapping protocol across each species transcriptome: Each data point depicts an individual .sra file partition as determined by the SRA archive. Red lines represent raw reads, blue lines highlight read counts after adaptors had been trimmed, yellow lines correspond to reads successfully mapped to the entire genome and green lines show those reads mapping to genes. Orange lines illustrate frequency of reads with no feature i.e. did not map to a gene and purple represent ambiguous reads i.e. have mapped equally to multiple regions. A) Orangutan genome, B) Mouse genome, C): macaque genome, D) gorilla genome, E) chimpanzee genome and F) human genome.*

3.2.1.3) Differential Gene Expression Analysis

Differential gene expression analysis was carried out on the dataset of RMGFs prepared in Section 3.2.1 using the EdgeR package in R (Robinson *et al*., 2010) and the results from the HTseq count analysis (Section 3.2.1.2) (Anders *et al*., 2015). The differential expression analysis was performed on human RMGFs, that had an annotated 1:1 ortholog in all other species being assessed (Human, Chimpanzee, Gorilla, Orangutan, Macaque, Mouse) across a panel of tissue samples (brain, cerebellum, heart, liver, kidney, liver, testes) for each mapped species (Human, Chimpanzee, Gorilla, Orangutan, Macaque, Mouse) logFC, logCPM, and p-value were calculated to assess for significance followed by an FDR to adjust for multiple testing. Genes with significant FDR results were plotted.

A subsequent analysis was then carried out to compare RMGF expression levels across the same tissue panel but between species within the database. With RMGFs with 1:1 orthologs across all other species being examined for differential expression. Again, logFC, logCPM, and p-values were calculated and corrected for multiple testing through FDR.

3.2.2) Wet-bench validation of RMGFs transcription profiles obtained through computational analysis

3.2.2.1) Quantitative RT-PCR to assess transcription of human RMGFs at 90 PI

Total human RNA was purchased from Life Technologies® and RNA was extracted from the following tissues: liver (AM7960), brain (AM7962), placenta (AM7950), lung (AM7968) and testes (AM7972). Five µg was digested with DNAseI (Sigma AMP-D1) for 15 minutes at room temperature (RT). cDNA was synthesized from the DNAse-free RNA using the Tetro cDNA synthesis kit (Bioline BIO-65042) as per manufacturers instructions. Quantitative real-time PCR was carried out on the cDNA using ABI fast SYBR-green qPCR kit (4385616) and on the 7900 HT ABI thermal-cycler. Each reaction contained

20ng/µl cDNA amplified with 0.2 µM of each primer, and was undertaken in triplicate. Primer sequences and their targets can be found in Appendix_C: Table 1 and the ACTB gene was used as an internal reference. Expression was assessed in two ways: (1) The primer pair displayed a single reproducible dissociation curve in at least one tissue analyzed, and (2) The delta threshold cycle (CT) value for a given primer pair compared with ACTB > 0.1, which we determined was our detection limit of a true positive.

3.2.2.2) RT-PCR analysis of two human RMGFs and their orthologs in gorilla and chimpanzee

Follow up RT-PCR analysis of two RMGFs from donated heart and frontal lobe tissue samples of gorilla and chimpanzee was carried out. DNA panels were limited as due to conservation laws, primate samples were difficult to obtain (samples were donated from Zoo De Barcelona). The two human RMGFs used were ENSG00000250021 and ENSG00000249773, and their orthologs in chimpanzee were ENSPTRG00000007442 and ENSPTRG00000019203 respectively and in gorilla ENSGGOG00000007765 and ENSGGOG00000022240 respectively. According to the RNA sequencing metadata analysis in Section 3.2.1 both of the selected human RMGFs had ubiquitous expression in human, chimp and gorilla. Unique primers were designed spanning the fusion breakpoints of each RMGF (Table 3.1). The RT-PCR experimental protocol used for zoo samples such as these is outlined in Table 3.2. In summary, prior to cDNA amplification a reverse transcription step was carried out across the tissues (human, chimpanzee and gorilla heart and frontal lobe/total brain tissue) being examined to convert RNA to a cDNA sample was required. Here a high capacity cDNA reverse transcription (part number: 4368814, 200rxn) kit was used: 2 µl of 10X RT Buffer, 0.8 µl 25X dNTPs, 2 µl of 10X Random primers; multiscribe, 1µl of reverse transcriptase, and 10 µl of sample RNA. Next RT-PCR was carried using the prepared cDNA sample for our 2 human RMGFs (ENSG00000250021 and ENSG00000249773 and their orthologs across chimpanzee and gorilla), the analysis was performed using 2.5µl of both forward and reverse primers (sequences in Table 3.1) along with a customised PCR kit: 2.5 µl of 10x buffer and 2µl of extracted DNA, 2µl

dNTP mix, 0.75µl of MgCl2, 0.25 µl of BioTaq polymerase, and the 7900 HT ABI thermocycler was run for 40 cycles. A 100bp DNA ladder (1-1517bp) from New England BioLabs; catalog number N3231l was selected based on our gene lengths and 4µl this of ladder was used per gel. Samples were ran on a prepared 1% agarose gel (50ml TBE 1X + 0.5g of agarose) and 2µl of SYBR safe was used. This was ran at 90V for 40mins.

Table 3.1: RT-PCR primer design for across-species expression analysis of 2 human RMGFs and their orthologs in gorilla and chimpanzee

Primer Name	Sequence	Gene Length (bp)	PCR Product Size (bp)
ENSG00000250021_1L	GATGACAACATCTGTAACTTC	2439	101
ENSG00000250021_1R	CAAAATACAAAGTAGAGGGTAG	2439	101
ENSG00000250021_2L	AGAGACTTTCCATCTAGTCC	2439	101
ENSG00000250021_2R	CAAAATACAAAGTAGAGGGTAG	2439	101
ENSG00000250021_3L	GATGACAACATCTGTAACTTC	2439	101
ENSG00000250021_3R	ATCCACACAAAATACAAAGTAG	2439	101
ENSG00000249773_1L	TGTACCCTGCCCAAAAGAAC	1732	100
ENSG00000249773_1R	CGCGATTACCTCTGGCTTAC	1732	100
ENSG00000249773_2L	ATGATGGCTCACAGATGGTG	1732	100
ENSG00000249773_2R	CCAGCATCACGTCTCGATAG	1732	100
ENSG00000249773_3L	ACGAGCAAAGCATGTGAAAC	1732	100
ENSG00000249773_3R	CAACGGACTCTCCAGGTAGG	1732	100
ENSGGOG00000007765_1L	GATGACAACATCTGTAACTTC	5667	101
ENSGGOG00000007765_1R	CAAAATACAAAGTAGAGGGTAG	5667	101
ENSGGOG00000007765_2L	AGAGACTTTCCATCTAGTCC	5667	101
ENSGGOG00000007765_2R	CAAAATACAAAGTAGAGGGTAG	5667	101
ENSGGOG00000007765_3L	GATGACAACATCTGTAACTTC	5667	101
ENSGGOG00000007765_3R	ATCCACACAAAATACAAAGTAG	5667	101
ENSPTRG00000019203_1L	ACGAGCAAAGCATGTGAAAC	583	115
ENSPTRG00000019203_1R	CAACGGACTCTCCAGGTAGG	583	115
ENSPTRG00000019203_2L	ACCAAAGCCACGTAATGTCC	583	115
ENSPTRG00000019203_2R	CAGAACAAGCCTGGTCACTC	583	115
ENSPTRG00000019203_3L	ACCAAAGCCACGTAATGTCC	583	115
ENSPTRG00000019203_3R	AACAAGCCTGGTCACTCTCAC	583	115
ENSGGOG00000022240_1L	GGCCGAGATTGTTTTCAAAG	393	100
ENSGGOG00000022240_1R	GGTTTCCGAACTCAATGGAC	393	100

ENSGGOG00000022240_2L	GCGAACAAAGCATGTGAAAC	393	100
ENSGGOG00000022240_2R	GGTTTCCGAACTCAATGGAC	393	100
ENSGGOG00000022240_3L	AACTGGCCGAGATTGTTTTC	393	100
ENSGGOG00000022240_3R	GGTTTCCGAACTCAATGGAC	393	100
ENSPTRG00000007442_1L	GATGACAACATCTGTAACTTC	6255	101
ENSPTRG00000007442_1R	CAAAATACAAAGTAGAGGGTAG	6255	101
ENSPTRG00000007442_2L	AGAGACTTTCCATCTAGTCC	6255	101
ENSPTRG00000007442_2R	CAAAATACAAAGTAGAGGGTAG	6255	101
ENSPTRG00000007442_3L	GATGACAACATCTGTAACTTC	6255	101
ENSPTRG00000007442_3R	ATCCACACAAAATACAAAGTAG	6255	101

Depicts unique primer sequences for each human RMGF and their orthologs in gorilla and chimpanzee for RT-PCR analysis. Three unique, break point spanning primers were designed per experiment in order to run triplicate experiments. Gene lengths are indicated for optimal polymerase selection.

Table 3.2: Experimental design of RT-PCR cross-species human RMGF RNA-sequencing validation

RMGF analysed	Sample	uL RNA	H20	Pos
ENSPTRG00000019203	Chimp Heart	1.4	8.6	4
ENSPTRG00000019203	Chimp_frontal_cortex	3.05	6.95	3
ENSGGOG00000022240	Gorilla Heart	24	0	6
ENSGGOG00000022240	Gorilla_frontal_cortex	6	4	5
ENSG00000249773	Human_Heart	1	9	2
ENSG00000249773	Human_Brain	1	9	1
ENSGGOG00000007765	Gorilla_Heart	24	0	6
ENSGGOG00000007765	Gorilla_frontal_cortex	6	4	5
ENSG00000250021	Human_Heart	1	9	2
ENSG00000250021	Human_Brain	1	9	1
ENSPTRG00000007442	Chimp Heart	1.4	8.6	4
ENSPTRG00000007442	Chimp_frontal_cortex	3.05	6.95	3
	Control A	0	10	
	Control B	0	10	

RT-PCR experimental design of human RMGFs (ENSG00000249773 and ENSG00000250021) and their orthologs in chimp and gorilla. Analysis was carried out in heart and with total brain or frontal cortex tissues. Table also includes sample preparation concentration calculations for each experimental setup.

3.2.3) An investigation of RMGFs with annotated alternative transcripts

An analysis was carried out across RMGFs to discover whether they are prone to the creation of alternative transcripts. Human RMGFs identified at 90 PI were assessed. Using the Biomart function on the Ensembl Genome Browser (**Version 74**) (Herrero *et al.*, 2016) we identified any annotated alternative transcripts present for each RMGF.

3.2.4) Uncovering translation profiles of detected RMGFs through ribosomal profiling

Four human ribosomal profiling datasets were selected from the GWIPS Web Browser including a skeletal muscle dataset, a glioma cell dataset and 2 fibroblast cell datasets (Michel *et al.*, 2014). Datasets were selected, as they were the most recently published datasets available on the GWIPs Browser. SRA files GSE45833 (Loayza-Puch *et al.*, 2013), GSE51424 (Gonzalez *et al.*, 2014), GSE48933 (Rooijers *et al.*, 2013) and GSE56148 were downloaded from the NCBI database (O'Leary *et al.*, 2016) and. FASTq file conversions were carried out using fastq-dump package from the SRAtoolkit (Leinonen *et al*., 2011). Adaptors were removed and reads were trimmed using the Fastx-toolkit's (http://hannonlab.cshl.edu/fastx_toolkit/index.html) fastx_trimmer function and cutadapt (Martin, 2011) , and reads of >25 nucleotides were retained. Data quality was assessed using the FASTQC package after each cleaning step (Andrews, 2010)(Figure 3.9-Figure 3.15). rRNA depletion of each dataset was carried out using BowTie2 against a reference human rRNA dataset (Langmead and Salzberg, 2012). Read counts were obtained at each step for quality control purposes (Figure 3.16). Reads of 48bp (16 amino acids) were constructed that spanned the fusion breakpoint centrally (i.e. 8 amino acids on each side of the breakpoint). Reads were mapped to each cleaned ribosomal profiling dataset using the Bowtie2 function to allow for split read mapping. Reads that successfully mapped to a ribosome profiling dataset where further mapped using BowTie to the latest human RefSeq genome **(hg19)** available on the UCSC Genome Browser (Karolchik *et al*., 2011; Langmead and Salzberg, 2012). In this way the exact chromosomal coordinates of each positive read hit were obtained.

Positive hits were viewed using the IGV Web Browser (IGV) (Integrative Genomic Viewer), 2013).

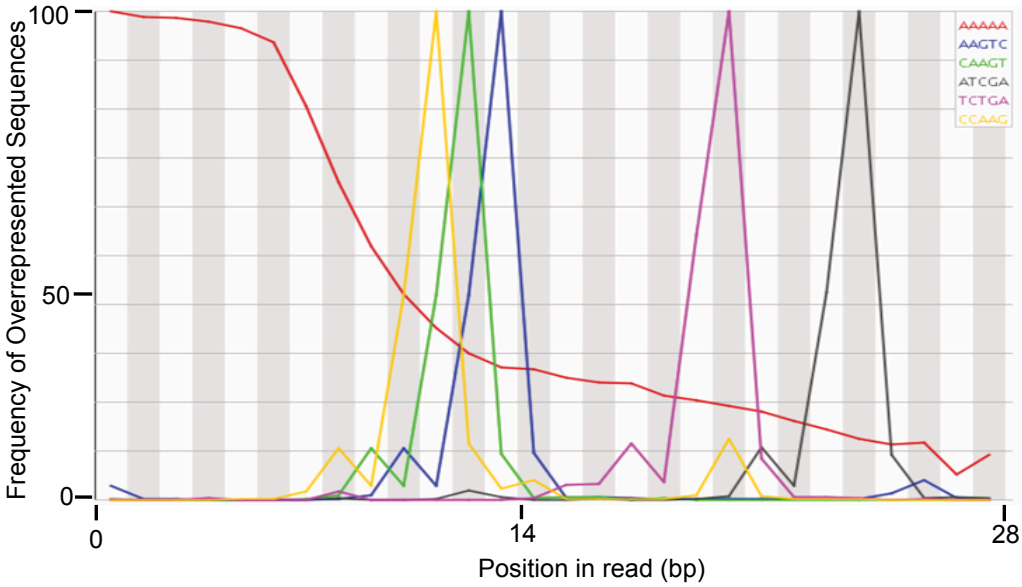


Figure 3.9: *FASTqc relative enrichment quality check output of the Rooijer Dataset after adaptor trimming: Figure highlights the frequency of overrepresented exactly duplicated sequences across reads examined. Each line colour corresponds to a specific overrepresented 5-mer highlighted in the top right corner of the graph. Poor quality sequences will have reduced kmer content due to sequencing errors across the reads. The red line indicated a potential bias in reads for the AAAAA kmer.*

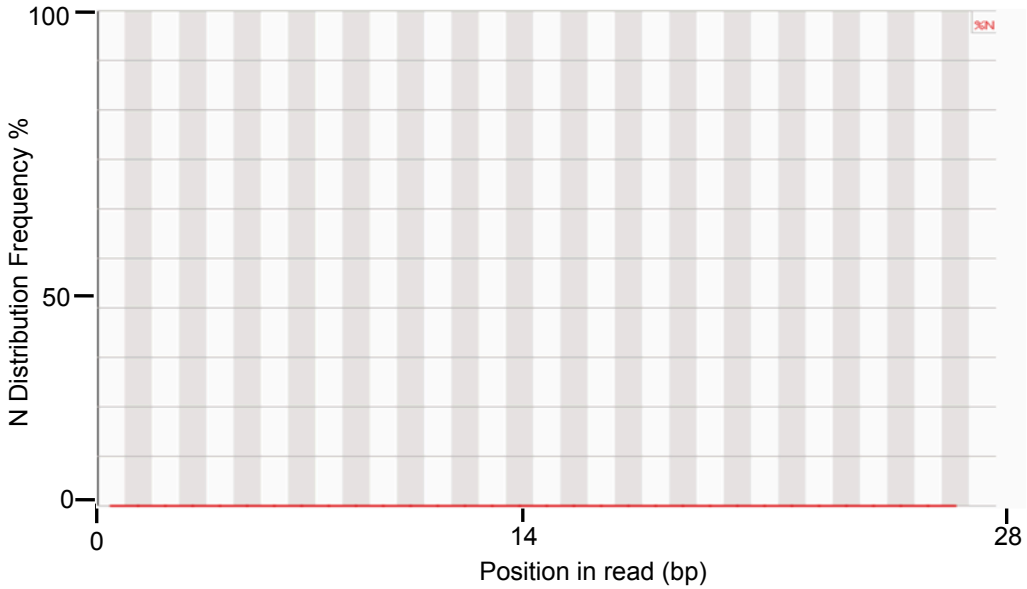


Figure 3.10: *An inspection of N content within the Rooijer Dataset after adaptor trimming by the FASTqc package: The Y axis highlights N count frequency and the X axis determines the position in base pairs of each read. Graph highlights that no N or uncalled bases were identified along the reads. This is indicative of a good quality sequence.*

Figure 3.11: *GC content quality check of the Rooijer Dataset after adaptor trimming: The results of FASTqc quality checker depicting the frequency of GC content (Y axis) across each base of a read (X axis). For good quality sequences this line is expected to be as straight as possible. Data show that there is bias amongst sequences. This confirms the overrepresentation found in Figure 3.9 of the AAAAA kmer.*

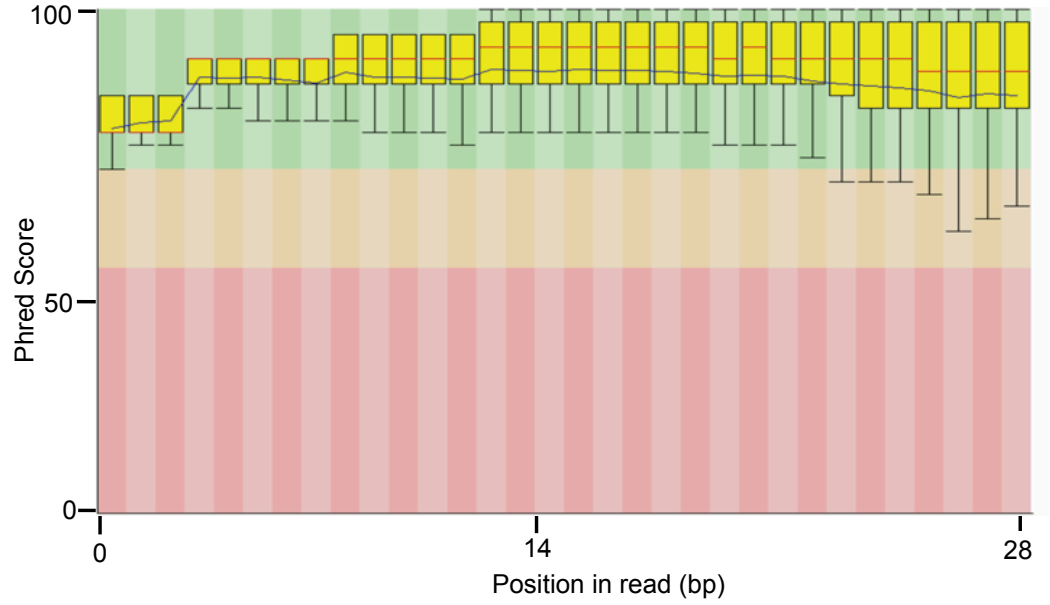


Figure 3.12: *An overall quality score assessment across bases in reads from the Rooijer dataset: This image represents the overall quality scores (Y axis) of each base across reads (X axis) with the blue line represents the overall mean quality score and the read line indicates the mean quality for that base. The green background is indicative of high quality data, the orange background represents intermediate quality and red represents poor quality data.*

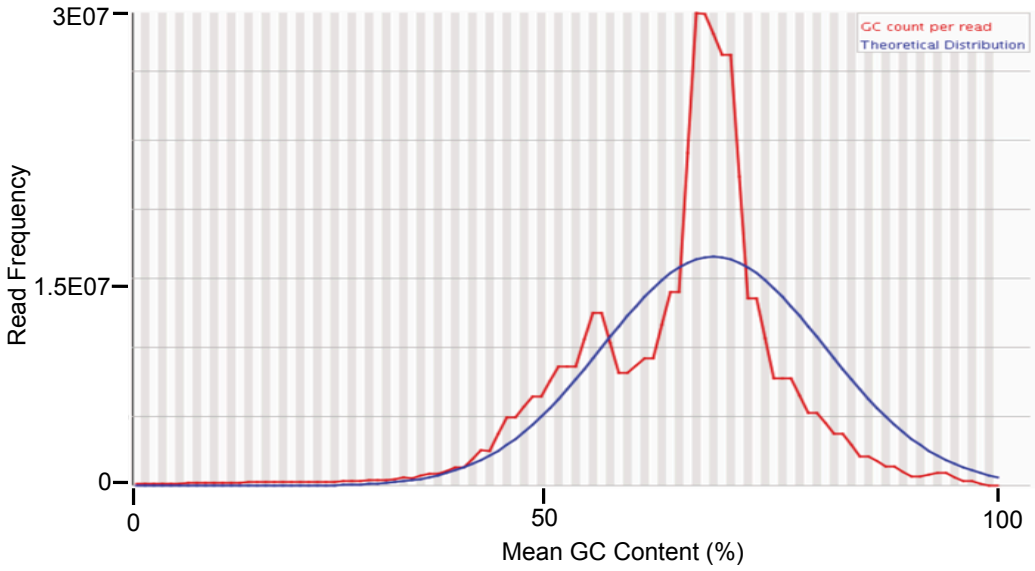


Figure 3.13: *FASTqc analysis result of GC distribution over all reads in the Rooijer dataset: Graphical representation of the distribution of GC (Y axis) across all reads in the dataset compared to the overall mean GC content (X axis). The blue line represents a normal randomly generated library with peak indicating overall GC distribution of the genome. The red line corresponds to the GC content of the actual genome being examined. The red line does not follow the normal distribution of the randomly generated blue line, thus suggesting some bias in the dataset.*

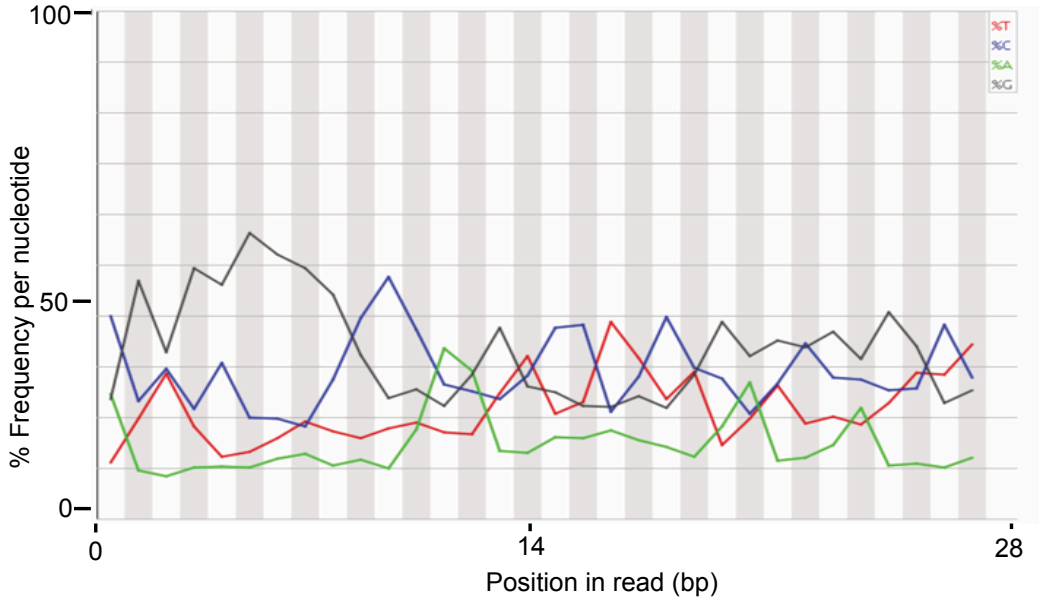


Figure 3.14: *Graphical display of the percentage of each nucleotide across each base in a read generated after adaptor trimming in the Rooijer dataset: Each coloured line depicts the percentage of a nucleotide; red representing 'T' nucleotides, blue depicting 'C' nucleotides, green highlighting 'A' nucleotides and black showing 'G' nucleotides. In a good quality dataset parallel lines would be expected, as each base should be represented equally across the read. However, here the lines are very sporadic, particularly across the start of the read, highlighting potential bias.*

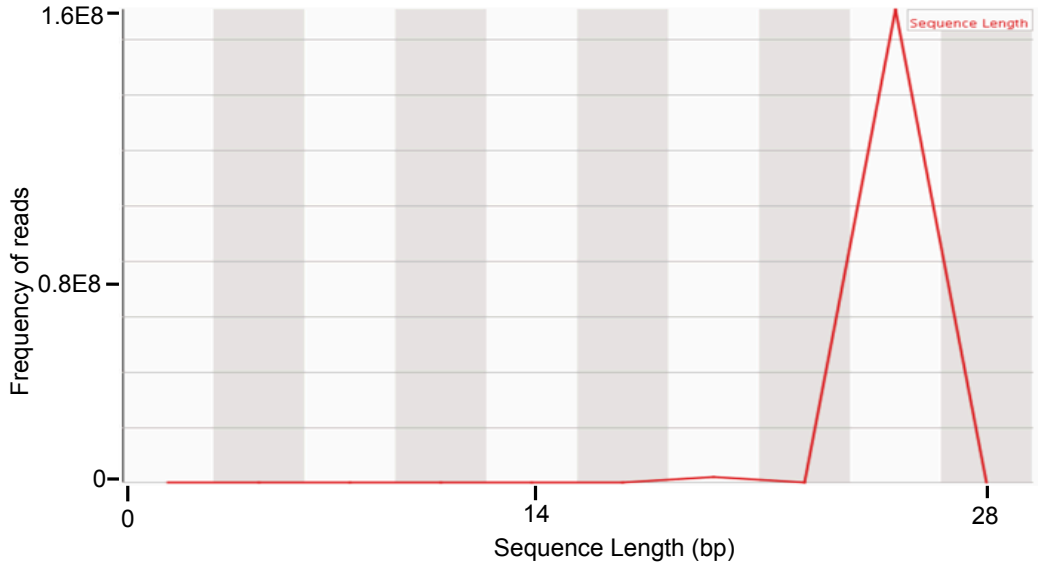


Figure 3.15: *An analysis of the read length distribution within the Rooijer dataset: An illustration of the distribution of read lengths (Y axis) across each position in the read (X axis). An individual peak is indicative of a single read length, this is expected after trimming.*

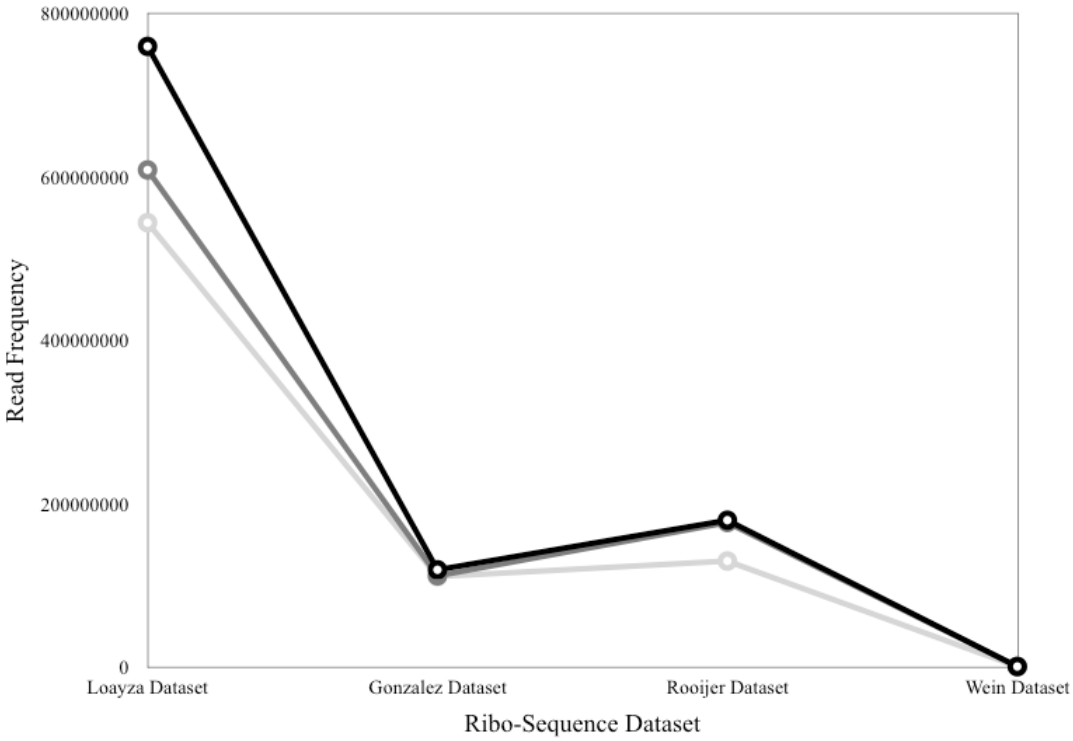


Figure 3.16: *An investigation of read frequency across all four ribosomal profiling datasets analysed: Each of the 4 datasets analysed are given on the (X-axis) – named as per first author of their corresponding publications. The Y-axis illustrates the number of reads. The black line indicates number of raw reads within each dataset, the dark grey line indicates number of reads post adaptor removal and trimming. Light grey line illustrates the number of reads in each dataset post rRNA depletion.*

3.3) Results

3.3.1) Transcription profiles of RMGFs from analysis of RNA sequencing data

To determine the expression profiles of the RMGFs and their parents, a metadata analysis of high quality transcription data was carried out (Brawand *et al.*, 2011). The dataset consisted Illumina RNA sequencing data from 6 species (human, chimpanzee, gorilla, orangutan, macaque and mouse) and 5 tissues (brain, cerebellum, lung, liver, placenta, and testes). RMGFs across all species were selected, obtained, filtered and cleaned as per Section 3.2.1.1 and mapped to it's corresponding reference genome to prevent humanisation of results. The number of reads mapping successfully across the RMGF panel for each species across all tissues were counted and split into 3 categories; **1)** >=1 read successfully mapped, **2)** >=3 read successfully mapped and **3)** >=5 reads successfully mapped. All read category findings are shown in Table 3.3 however only the >=1 read mapping category will be discussed in detail as >1 read is sufficient evidence for transcript existence.

Results of the RMGF analysis are shown in Figure 3.17 in human it is clear that across all read mapping categories that both testes and cerebellum have the greatest frequency of RMGFs expressed. For instance out of the 36 human RMGFs analysed 9 (25%) were found with signatures of expression in testes and 8 (22.2%) in the cerebellum. Interestingly, across the 29 chimp RMFGs examined 96% of RMGFs (28/29) were expressed with heart tissues showing the second most abundant number of RMGF genes (70.6%) showing expression. In gorilla 18 RMGFs were tested for expression and 94% of these genes showed both testes and liver expression. Across the 34 RMGFs examined in macaque 29 (85%) were found to have testes expression and 74% showed evidence for cerebellum expression. Across the 14 mouse RMGFs 32% of these were expressed in the testes. Finally, 25 RMGFs were analysed across 5 orangutan tissues (no testes tissue available) and 19/25 (76%) of these genes were found with expression signatures in the cerebellum. In summary, across all species analysed the frequency of RMGF expression was found at elevated levels in the testes which supports the tissue-specific expression of new gene hypotheses such

as the “out-of-testes” hypothesis. Interestingly, in human not only are RMGF expressed more frequently in the testes but also in the cerebellum.

A graphical representation of an investigation carried out on RMGF parent expression using the same panel of 6 species across the same tissue panel is highlighted in Figure 3.18. Here, out of 502 human parent RMGFs analysed the highest frequency of genes were found expressed in both the testes and cerebellum (11%). This is the same trend found in across human’s corresponding RMGFs. In chimpanzee a total of 248 parent RMGFs were assessed and the highest number (2.4%) of parent RMGFs were found to map in testes, again this follows the trend of their associated RMGFs. In gorilla 305 parent RMGFs were assessed with highest numbers found in testes (2.2%) followed secondly by brain and cerebellum with 2.1% of parent RMGFs found to be expressed here. An analysis of 431 orangutan parent RMGFs (no testes sample available for testing) uncovered that 9.2% of the parent RMGFs were expressed in cerebellum correlating with the result from it’s corresponding RMGFs. A total of 569 macaque parent RMGFs were analysed and 1.4% were found with testes expression a signature also found in their RMGF counterparts. Finally, 285 mouse parent genes were assessed and highest numbers were uncovered in the heart (5.2%) and kidney (4.9%). Interestingly this pattern was not found in their associated RMGFs whom showed heightened numbers of RMGFs expressed in the testes. To summarise, across all species examined parent RMGFs (Figure 3.18) follow the same expression profile as their corresponding RMGF (Figure 3.17) apart from mouse whose RMGFs are found to be expressed at elevated levels in testes but the corresponding parent genes show heightened numbers of expression in heart and kidney tissues. Again, human parent RMGFs were shown to have heightened numbers of expression not only in testes but also cerebellum.

The frequency of RMGF and parent genes that have >=5 reads mapping to them are shown in Table 3.3. For all the species tested in this read mapping category (no testes sample available in orangutan) the majority of RMGFs are expressed in the testes. This pattern remains to be true at the lesser stringent mapping categories, >=3 and >=1.

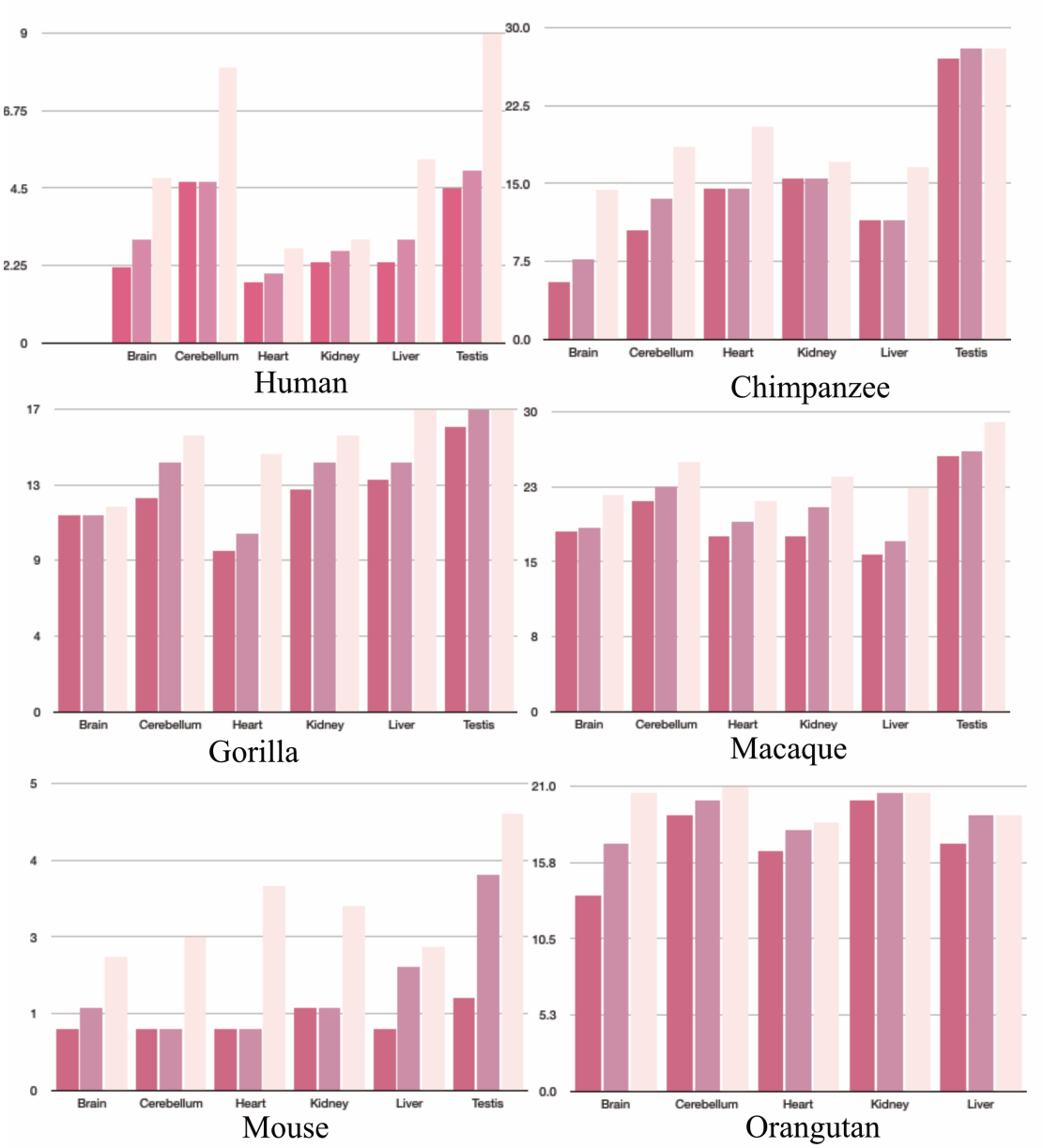


Figure 3.17: *RNAsequencing metadata analysis of RMGFs across a dataset of 6 primate species and mouse: Image depicts number of RMGFs expressed in each tissue (X axis) across each species (human, chimpanzee, gorilla, macaque, mouse and orangutan) examined. The Y Axis represents the frequency of genes transcribed per tissue (if more than one sample per tissue was analysed a tissue average was calculated). Light pink represents transcription level of RMGFs with >1 read mapped per gene. Medium pink bars illustrate >3 reads were mapped to each RMGF and dark pink bars show instances were >5 reads mapped to each RMGF.*

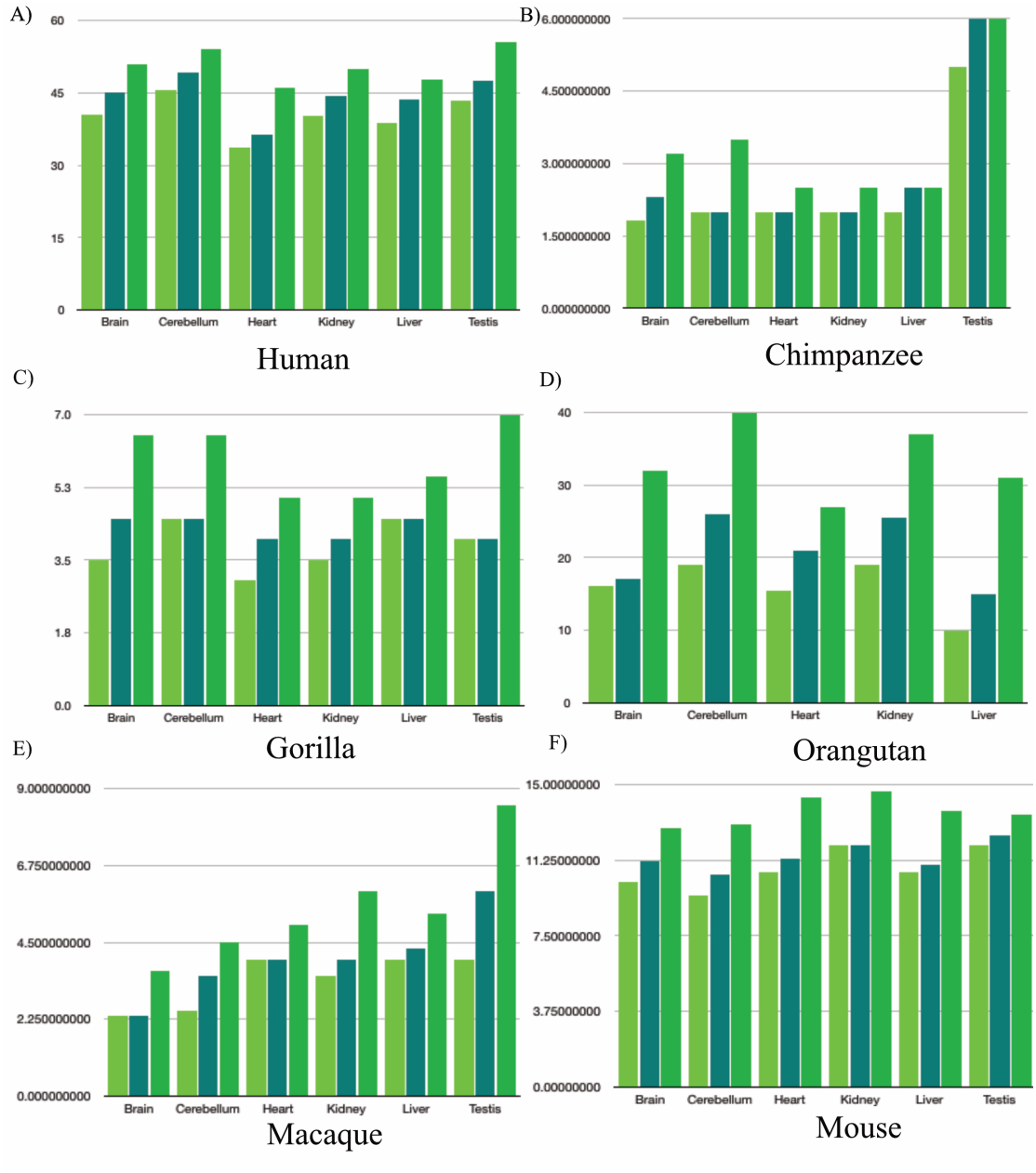


Figure 3.18: *Expression profiles for RMGF parents from RNAseq metadata analysis of 6 primate species and mouse across a panel of 6 tissues: The six tissues examined are displayed on the X-axis. The Y-axis corresponds to the frequency of RMGF expression in that tissue (if more than one tissue sample was used an average was extrapolated). Right medium green bars represent RMGF frequencies were at least one read successfully mapped, middle dark green bars indicate >=3 reads mapped and left light green bars illustrate RMGFs were >=5 genes successfully mapped.*

Table 3.3: The frequency of RMGFs expressed across 3 read mapping categories in a dataset of 5 primates and mouse

	Human			Chimp			Gorilla			Macaque			Mouse			Orangutan		
	Reads >5	Reads >3	Reads >1	Reads >5	Reads >3	Reads >1	Reads >5	Reads >3	Reads >1	Reads >5	Reads >3	Reads >1	Reads >5	Reads >3	Reads >1	Reads >5	Reads >3	Reads >1
Brain	2.2	3	4.8	5.5	7.66	14.3	11	11	11.5	18	18.3	22	1	1.33	2.16	13.5	17	21
Cerebellum	4.7	4.7	8	10.5	13.5	18.5	12	14	15.5	21	22.5	25	1	1	2.5	19	20	21
Heart	1.8	2	2.8	14.5	14.5	20.5	9	10	14.5	17.5	19	21	1	1	3.33	16.5	18	19
Kidney	2.3	2.7	3	15.5	15.5	17	12.5	14	15.5	17.5	20.5	24	1.33	1.33	3	20	21	21
Liver	2.3	3	5.3	11.5	11.5	16.5	13	14	17	15.7	17	22	1	2	2.33	17	19	19
Testis	4.5	5	9	27	28	28	16	17	17	25.5	26	29	1.5	3.5	4.5	n/a	n/a	n/a

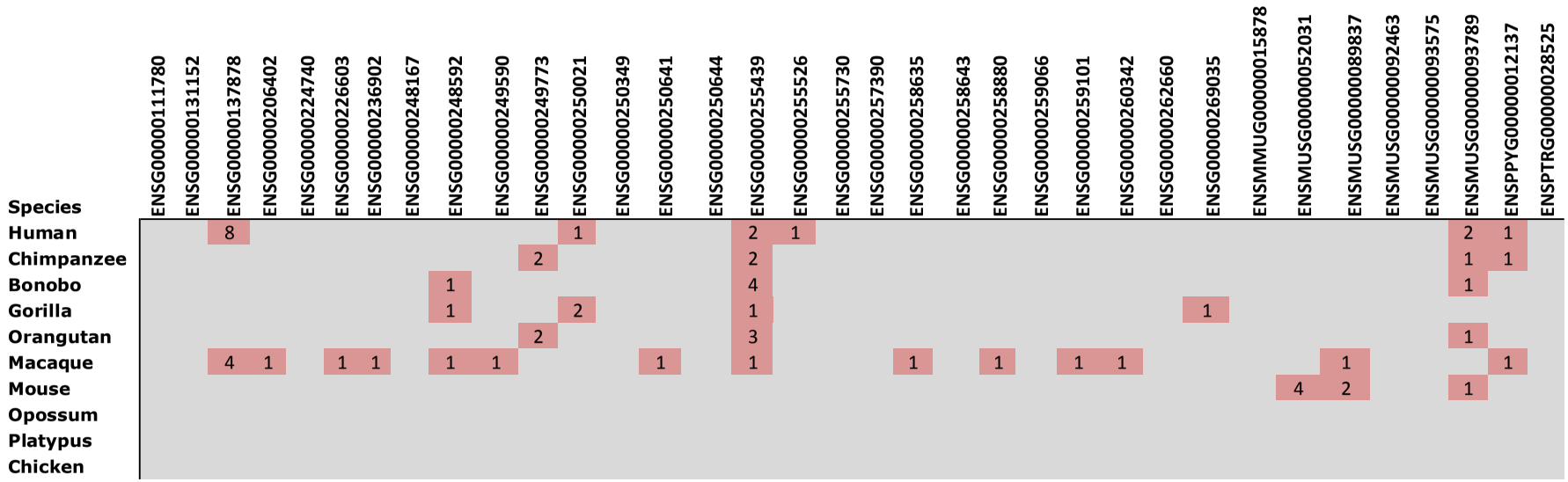
RNA sequencing results for RMGFs across 5 primates (chimpanzee, gorilla, macaque, mouse and orangutan) and mouse. over a panel of tissues including brain, cerebellum, heart, kidney, liver and testis. The analysis illustrates the number of genes containing successfully mapped genes at three read mapping thresholds; >=5, >=3 and >=1 reads mapped.

3.3.2) Alternative transcript frequency for human RMGFs and human ortholog containing primates

Human RMGFs identified at a 90% identity threshold and primates RMGFs with human 1:1 orthologs were assessed for alternative transcript frequency (Section 3.2.3). In total 19/35 (54%) RMGFs tested have a known annotated alternative transcript in at least one species analysed (Table 3.4). Surprisingly only 6 of the 35 (17%) RMGFs tested had an alternative transcript in human. However, 14/35 alternative transcripts were found in macaque. This could be as a result of mis-annotations in the lower quality macaque genome.

From the data ENSG00000255439 appears to have annotated alternative transcripts across primates coinciding with the RMGF's ubiquitious expression profile in primates (Section 3.2.1)**.** However, species contain differing numbers of transcript isoforms suggesting that this RMGF plays a flexible role in primate genomes. ENSMUSG000000093789 also appears to have alternative transcripts in mouse. No alternative transcripts were identified in opossum, platypus or chicken according to the Ensembl Genome Browser, this is probably due to the lower number of alternative transcript analyses that have been carried out on these genomes and consequently resulting in a lower alternative transcript annotation rate. To assess the length of RMGFs in comparison to non-fused human protein coding genes a simulation was carried out and compared using non-parametric Mann Whitney tests. It was found that fusion genes were statisitically significantly longer than non-fused genes with a p-value of 0.05606

Table 3.4: An analysis of alternative transcripts across RMGFs identified at 90 PI and their orthologs across a panel of high quality vertebrate species



Results of an analysis of alternative transcript frequency across a panel of 35 RMGF identified at a 90 PI. Pink cells indicate an annotated has been associated with the gene according to the Ensembl Genome Browser (Herrero *et al.*, 2016). *The number of currently identified isoforms is provided within the cell each pink cell. Grey cells indicate no alternative transcripts have been annotated in the database at present.*

3.3.3) qRT-PCR analysis of 27 RMGFs using human as a representative of the Great Ape Clade

As only a small amount of primate DNA was available due to conservation limitations we took human as a representative of the Great Apes and carried out qRT-PCR experiments across 27 candidate human RMGF orthologs. Candidate RMGFs are those whom unique primers can be generated, the requirements of which are detailed in Section 3.2.1.2. RMGFs were tested across five tissue panels: brain, testes, liver, lung and placenta and the results of the qRT-PCR experiment are shown in Figure 3.19. 10/27 RMGFs analysed were found to have ubiquitous expression across all tissues examined, 3/27 have expression profiles in 3 tissues examined, 2/27 in 3 tissues, 6/27 in 2 tissues, 4/27 in 1 tissue and 2 RMGFs were found to have no expression across any human tissue analysed. 20/27 (74%) genes have testes expression, 13/27 (48%) in brain, 19/27 (705%) in liver, and 16/27 (59%) in both lung and placenta. In summary, the greatest number of RMGFs analysed had expression verified through qRT-PCR analysis in the testes, again supporting the signature of expression found in the RNA sequencing analysis carried out in Section 3.2.1.2. However, although these genes are more frequently expressed in the testes, a lower number are also found expressed across the other tissues examined.

Figure 3.19: qRT-PCR results for 27 candidate human RMGFs identified at 90 PI across a panel of five human tissues

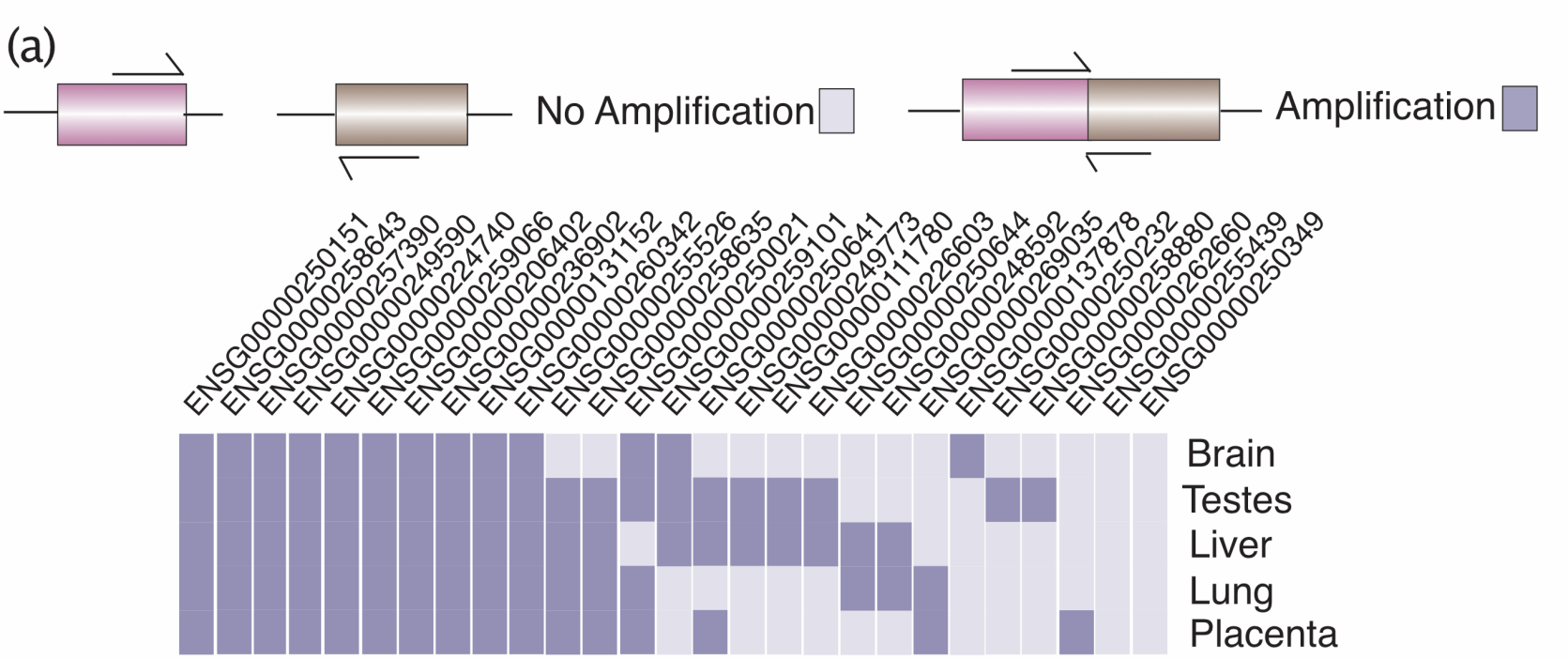


Figure 3.19: *The Ensembl gene ID's for all 27 human RMFG orthologs are given as columns and the 5 human tissue panels analysed are given as rows. Depiction of a qRT-PCR analysis with evidence for gene expression signatures within a tissue represented by blue cells, grey cells indicate no gene expression signature was detected in the tissue examined.*

3.3.4) Differential expression analyses of primate and mouse RMGFs using the EdgeR package

The panel of RMGFs tested in Section 3.3.3 were assessed for patterns of differential expression, this dataset included 36 human, 29 chimpanzee, 18 gorilla, 25 orangutan, 34 macaque and 14 mouse. RMGFs These 156 RMGFs were investigated for differential expression (DE) profiles across a panel of 5 tissues (brain, cerebellum, heart, kidney and testes) using the EdgeR package (Robinson *et al*., 2010).

In human, 3/36 (8.3%) RMGFs showed signatures of differential expression (Figure 3.20), the ENSG00000137878 shows DE in brain compared to cerebellum, heart, liver and testis. DE was also found when cerebellum was compared to heart, liver, testis and brain and heart when compared to all other tissues. For the ENSG00000250588 gene, DE was found when brain was compared to all other tissues and in the ENSG00000185304 gene brain was found to be under DE when compared to testis and cerebellum and the cerebellum was found differentially expressed when compared to heart. In summary, across all 3 genes examined all 3 genes showed differential expression when brain was compared to both testes and cerebellum tissues.

In chimpanzee significant levels of DE was found in 4/29 (13.79%) RMGFs. Interestingly in both the ENSPTRG0000012066 and ENSPTRG0000008624 gene the testis was found differentially expressed when compared to all other tissues and across all four genes (ENSPTRG0000012066, ENSPTRG0000008624, ENSPTRG00000020751, and ENSPTRG00000028525), showed no DE between brain and cerebellum tissues, unlike the profile found in human RMGFs where DE was found in these tissues across all three RMGFs.

In gorilla 6/18 (33%) RMGFs showed signatures of DE across tissues. DE was identified in 2/6 RMGFs when brain was compared to both testes and liver. A further 3/6 RMGFs also showed DE between cerebellum and heart tissues. All other significant DE events are indicated in Figure 3.20. DE was identified across 8/25 (32%) RMGFs in orangutan with 5/8 genes showing evidence of DE

3.3.4) Differential expression analyses of primate and mouse RMGFs using the EdgeR package

The panel of RMGFs tested in Section 3.3.3 were assessed for patterns of differential expression, this dataset included 36 human, 29 chimpanzee, 18 gorilla, 25 orangutan, 34 macaque and 14 mouse. RMGFs These 156 RMGFs were investigated for differential expression (DE) profiles across a panel of 5 tissues (brain, cerebellum, heart, kidney and testes) using the EdgeR package (Robinson *et al*., 2010).

In human, 3/36 (8.3%) RMGFs showed signatures of differential expression (Figure 3.20), the ENSG00000137878 shows DE in brain compared to cerebellum, heart, liver and testis. DE was also found when cerebellum was compared to heart, liver, testis and brain and heart when compared to all other tissues. For the ENSG00000250588 gene, DE was found when brain was compared to all other tissues and in the ENSG00000185304 gene brain was found to be under DE when compared to testis and cerebellum and the cerebellum was found differentially expressed when compared to heart. In summary, across all 3 genes examined all 3 genes showed differential expression when brain was compared to both testes and cerebellum tissues.

In chimpanzee significant levels of DE was found in 4/29 (13.79%) RMGFs. Interestingly in both the ENSPTRG0000012066 and ENSPTRG0000008624 gene the testis was found differentially expressed when compared to all other tissues and across all four genes (ENSPTRG0000012066, ENSPTRG0000008624, ENSPTRG00000020751, and ENSPTRG00000028525), showed no DE between brain and cerebellum tissues, unlike the profile found in human RMGFs where DE was found in these tissues across all three RMGFs.

In gorilla 6/18 (33%) RMGFs showed signatures of DE across tissues. DE was identified in 2/6 RMGFs when brain was compared to both testes and liver. A further 3/6 RMGFs also showed DE between cerebellum and heart tissues. All other significant DE events are indicated in Figure 3.20. DE was identified across 8/25 (32%) RMGFs in orangutan with 5/8 genes showing evidence of DE

between brain and heart tissues and 3/8 indicating DE between brain and liver (Figure 3.20). There was no evidence for any DE between brain and cerebellum across any RMGF examined in orangutan, therefore in contrast to human RMGFs all Great Ape RMGFs show no DE between brain and cerebellum tissues. Out of the 34 RMGFs investigated in macaque 5 were found with evidence of DE across tissues examined (14.7%). Evidence for DE between testes and all other tissues examined was identified for 2/5 RMGFs and 4/5 genes showed DE between brain and testes and 3/5 show DE between brain and liver and 2 uncovered DE between brain and both kidney and liver. Only 1 gene showed DE between brain and cerebellum. DE across 3/5 tissues was found in cerebellum when compared to both kidney and liver tissues. All other DE found in the macaque RMGF panel is represented in Figure 3.20.

Finally, only 1/14 RMGFs (ENSMUSG00000052031) examined in mouse showed evidence of DE. Here, testes were differentially expressed when compared with all other tissues examined and DE was identified between brain and heart tissues (Fig 3.20).

A second differential expression analysis was carried out in order to compare human RMGF brain expression to their orthologous genes in chimpanzee, gorilla, orangutan, macaque and mouse. Here, only human RMGFs with 1:1 orthologs across the species in the dataset could be used (13 chimp orthologs, 7 gorilla orthologs, 6 orangutan orthologs, 16 macaque orthologs and 21 mouse orthologs) (Table 3.5). A comparison of human RMGFs and their chimp orthologs revealed DE in 8/13 cases (61.5%). When compared to gorilla 3/7 instances of DE were identified and 4/6 cases were identified in orangutan. Across the 16 orthologs assessed in macaque 14 showed signatures of expression differentiation and 16/21 in mouse (76%). Both ENSG00000213380 and ENSG00000258465 show evidence of DE across 4 species being examined. The ENSG00000213380 gene has been associated with anterograde transport which is the movement from the cell body toward the synapse and it's erroneous expression has been associated with neurodevelopment disorders, psychomotor retardation, and dysmorphic features to name but a few (Zolov and Lupashin,

2005; UniProt Consortium, 2018) The ENSG00000258465 remains uncharacterised however it does contain WD40 repeats, which have a strong association with nucleotide binding and also act as intermediaries in transduction pathways and cell communication (UniProt Consortium, 2018).

ENSG000000250021, ENSG000000257390 and ENSG00000249773 show brain DE across 3 species. The ENSG000000250021 gene remains uncharacterised but has been associated with type-2-diabetes in many GWAS studies (Consortium, 2014). The ENSG000000257390 gene resides in the nucleus and contains a j-domain, these domains are known to have chaperone activities that are essential for the folding of proteins, translocation of polypeptides across membranes, response to stress and degradation targeting (Byron *et al.*, 2012). Lastly the ENSG00000249773 gene has been characterised and a protein has been predicted. The genes protein product has been identified as a component of the ribosome and contains a KRAB domain found in zinc finger proteins (ZFPs) that are known to cause transcriptional repression (Hendrickson *et al.*, 2010).

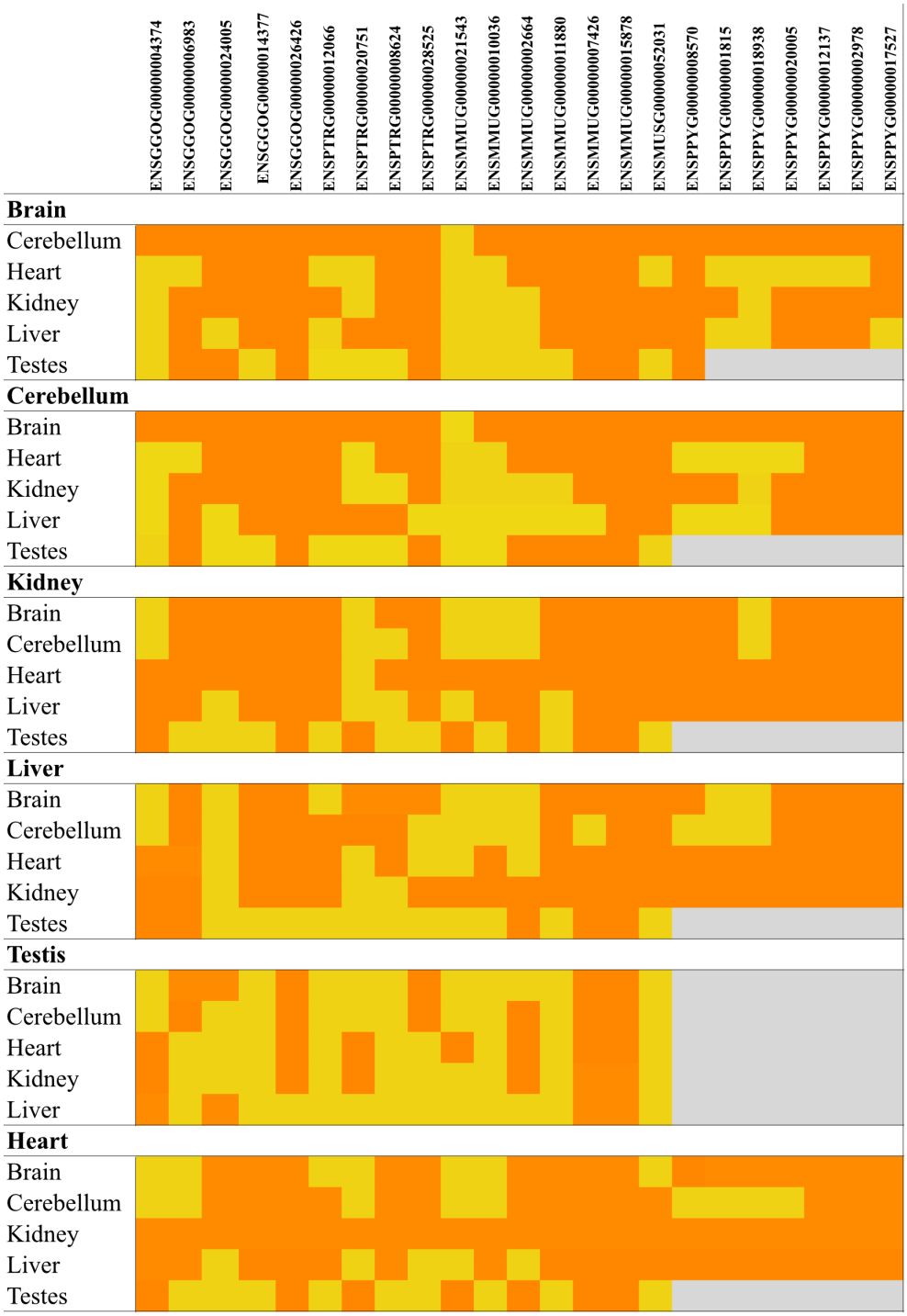


Figure 3.20: *Differential expression results of RMGFs across primate species and mouse: RMGF gene ID's from Ensembl are provided in the columns and tissues compared against one another are provided as rows. Yellow cells indicate no differential expression was identified between those particular pair of tissues in that species whilst orange cells refer to statistically significant differential expression between the tissues.*

Table 3.5: Differential expression analysis of brain tissues between human RMGFs and their primate and mouse orthologs

Human RMGF Ensembl ID	Orthologous Species	FDR
ENSG00000213380	Chimpanzee	3.83×10^{-25}
ENSG00000258465	Chimpanzee	3.97×10^{-15}
ENSG00000250021	Chimpanzee	3.33×10^{-11}
ENSG00000255526	Chimpanzee	2.17×10^{-9}
ENSG00000262660	Chimpanzee	6.96×10^{-9}
ENSG00000250232	Chimpanzee	1.91×10^{-8}
ENSG00000131152	Chimpanzee	7.47×10^{-8}
ENSG00000257390	Chimpanzee	0.013643351
ENSG00000258465	Gorilla	3.39×10^{-10}
ENSG00000213380	Gorilla	0.00017843
ENSG00000250588	Gorilla	0.0002297765
ENSG0000025846	Orangutan	1.98×10^{-8}
ENSG00000259066	Orangutan	0.001262971
ENSG00000213380	Orangutan	0.02514
ENSG00000249773	Orangutan	0.02514
ENSG00000258465	Macaque	6.22×10^{-38}
ENSG00000259066	Macaque	4.62×10^{-23}
ENSG00000255730	Macaque	1.33×10^{-22}
ENSG00000250151	Macaque	5.03×10^{-13}
ENSG00000261740	Macaque	1.81×10^{-7}
ENSG00000257390	Macaque	1.12×10^{-6}
ENSG00000250021	Macaque	6.55×10^{-6}
ENSG00000255439	Macaque	0.000150658
ENSG00000137878	Macaque	0.000841705
ENSG00000249773	Macaque	0.005636885
ENSG00000250644	Macaque	0.019311731
ENSG00000248592	Macaque	0.019311731
ENSG00000131152	Macaque	0.035910986
ENSG00000171931	Macaque	0.039070723
ENSG00000248167	Mouse	7.09×10^{-69}
ENSG00000258465	Mouse	4.11×10^{-64}
ENSG00000259066	Mouse	6.40×10^{-64}
ENSG00000249590	Mouse	1.18×10^{-54}
ENSG00000250232	Mouse	1.12×10^{-47}
ENSG00000255526	Mouse	6.08×10^{-41}
ENSG00000258643	Mouse	5.53×10^{-39}

ENSG00000257390	Mouse	$5.06x10^{-36}$
ENSG00000249773	Mouse	$7.39\ x10^{-31}$
ENSG00000250151	Mouse	$1.70\ x10^{-24}$
ENSG00000255730	Mouse	$2.15\ x10^{-24}$
ENSG00000250021	Mouse	$1.59\ x10^{-23}$
ENSG00000250644	Mouse	$3.58\ x10^{-20}$
ENSG00000248592	Mouse	$3.84\ x10^{-18}$
ENSG00000213380	Mouse	$3.95\ x10^{-10}$
ENSG00000137878	Mouse	0.0366474

Depicts results of a differential expression analysis of brain tissues between human RMGFs and their 1:1 orthologs across chimpanzee, gorilla, orangutan, macaque and mouse species. RMGFs with evidence of DE are represented here.

3.3.5) RT-PCR analysis of gorilla and chimpanzee tissue samples from the Barcelona Zoo

RNA samples were donated from the Barcelona Zoo for, chimpanzee and gorilla species across frontal cortex and heart tissues. Human DNA was donated from Dr. Tomas Marques-Bonet's Comparative Genomics Lab, PRBB Institute, Barcelona for total brain and heart tissues. RT-PCRs were carried out in the Comparative Genomics Lab in the PRBB, Barcelona under the supervision of Prof. Tomas Marques-Bonet. Table 3.6 illustrates the expected expression level as per RNA sequencing result (Section 3.3.1) and Figure 3.21 displays the RT-PCR polyacrylamide gel results.

Out of the 15 RMGFs RT-PCR analyses (Figure 3.21) 13/15 correspond to the expected result obtained from the RNA sequencing experiment. This excludes lane 8 where high expression levels were expected (Figure 3.21) according to the RNA sequencing result but no band is present after gel separation. Erroneous double band formation in lane 9 could be as a result of primer dimer formation. However the band that was created had an intense signal that was expected according to the RNA sequencing result (Table 3.6).

Table 3.6: Expected expression level per RT-PCR experiment as determined by RNA sequencing metadata analysis (Section 3.3.1)

Lane	Experiment	Species	Tissue	Expected Expression
1	ENSPTRG00000019203	Chimp	Frontal Cortex	Low
2	ENSPTRG00000019203	Chimp	Heart	High
3	ENSGGOG00000022240	Gorilla	Frontal Cortex	Low
4	ENSGGOG00000022240	Gorilla	Heart	High
5	ENSG00000249773	Human	Brain	Low/None
6	ENSG00000249773	Human	Heart	None
8	ENSPTRG00000007442	Chimp	Frontal Cortex	High
9	ENSPTRG00000007442	Chimp	Heart	High
10	ENSGGOG00000007765	Gorilla	Frontal Cortex	High
11	ENSGGOG00000007765	Gorilla	Heart	High
12	ENSG00000250021	Human	Brain	None
13	ENSG00000250021	Human	Heart	None
14	ENSPTRG00000034246	Chimp	Frontal Cortex	Low
15	ENSPTRG00000034246	Chimp	Heart	Low
16	- Control			None
17	- Control			None

Illustrates the expression levels obtained from RNA sequencing metadata analysis across each experiment run, these findings are the expected results of the RT-PCR validations. The expected result for each RT-PCR run is highlighted in column 5. Negative controls are cells containing H_20 in replacement of RNA.

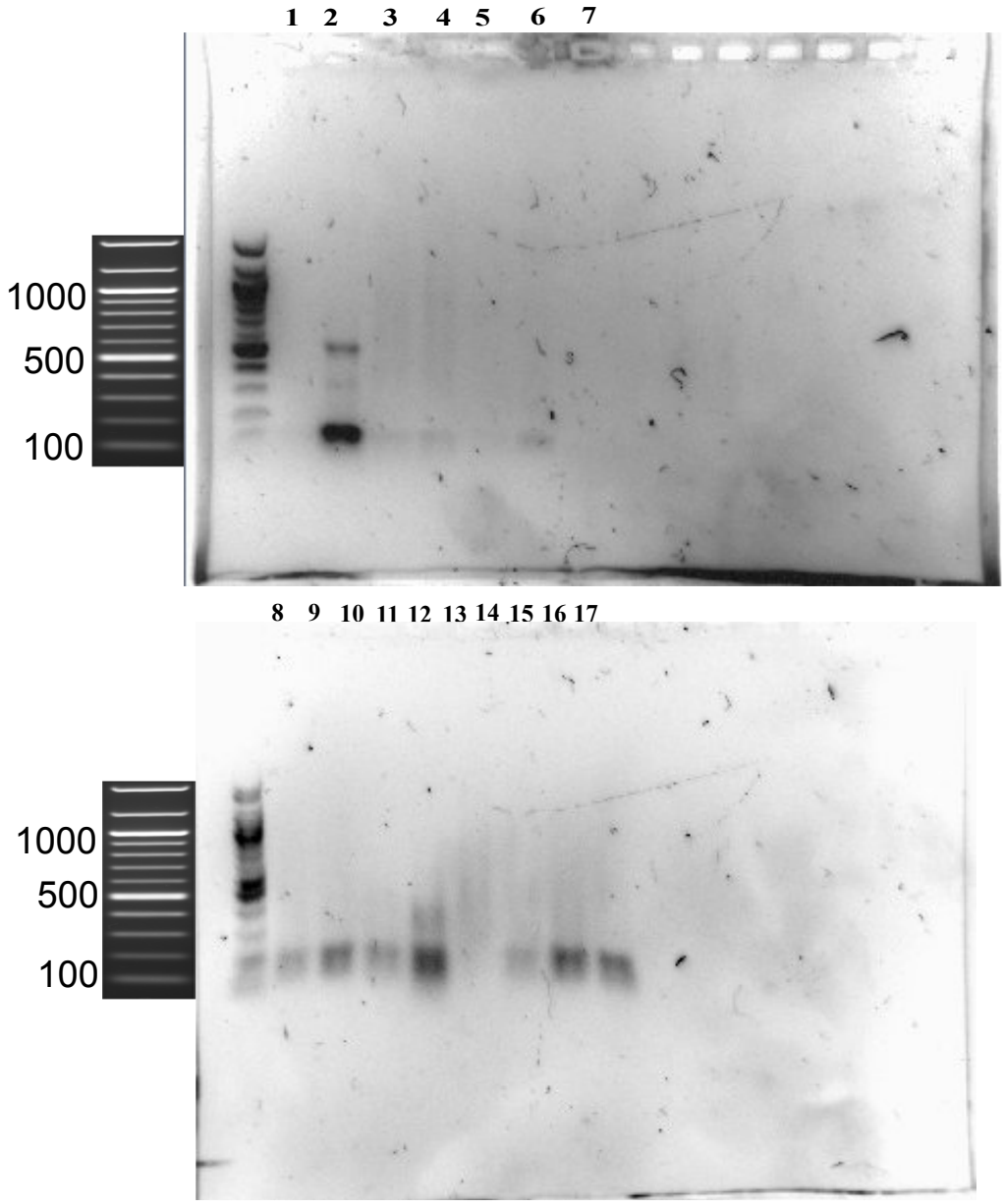


***Figure 3.21**: RT-PCR polyacrylamide gel results of 2 human RMGFs and their corresponding orthologs in chimpanzee and gorilla: RT-PCR experiment to validate transcription in 2 human RMGFs (ENSG00000249773 and ENSG00000250021) and their orthologs in chimpanzee and gorilla. Lane 1-17 contain RNA transcripts as detailed in Table 3.6. Lanes alternate between human, gorilla and chimpanzee and between tissues sampled.*

3.3.6) Analysis of ribosome profiles from limited datasets reveals signature of translation for small number of RMGFs

The translatomic profiles of 35 human RMGFs identified at 90 PI were assessed using four ribosome profiling datasets two from human fibroblasts and one from both skeletal muscle and glioma cells. Unique reads spanning each RMGF were constructed (Section 3.2.1). Footprints were identified, with no mismatches for 3 human RMGFs as they successfully mapped to the ribosome profiling fibroblast dataset using BowTie2 (Langmead and Salzberg, 2012) (Figure 1.22). These include ENST00000529564, ENST00000446072 and ENST00000567078. Subsequent mapping to the human genome **(hg19)** (O'Leary *et al.*, 2016) showed that these 3 fused genes were unique and had not yet been annotated and entered into the latest RefSeq gene database (O'Leary *et al.*, 2016).

The ENST00000529564 RMGF transcript and its corresponding protein have only been computationally predicted by Uniprot (UniProt Consortium, 2018) and not yet been experimentally proven. However, from our ribosome profiling results it is clear that a protein has the potential to be generated from this RMGF transcript due to clear ribosome binding. Analysis of GO functional data suggest a serine-type endopeptidase activity and vitamin-K-epoxide reductase activity for this RMGF (Ashburner *et al.*, 2000). This transcript has been predicted to play a role in blood coagulation, organic cyclic compound and organonitrogen responses (Ashburner *et al.*, 2000). Due to its enzymatic activity it is unsurprising that this protein is suspected to be intra-cellularly membrane bound. Generation of this novel gene transcript occurred due to an RNA-mediated fusion event between the *PRSS53-201* and *VKORC1-206* genes. However, qRT-PCR and RNA sequencing across selected panels of human tissues uncovered no signatures of expression for this gene.

The ENST00000446072 gene transcript has been derived from Ensembl's automatic analysis pipeline (Aken *et al.*, 2017). However, expression has been detected in human testes and liver tissues in RNAseq datasets, and our qRT-PCR analysis revealed expression across all tissues sampled (brain, heart, liver, kidney, testes and cerebellum)**.** These results provide evidence that this transcript

is broadly expressed as indicated by qRT-PCR analyses, and ribosome profiling data suggests it also produces a protein product in human.

Similarly, the ENST00000567078 transcript has only been predicted from Ensembl's annotation pipeline (Zerbino *et al.*, 2018). The gene is generated from a fusion event between the genes *ARL6IP1* (enhanced expression in brain) and RPS15A (enhanced expression in the lung and ovary)**.** The 2 adjacent gene transcripts fuse over intron 2 and aberrant splicing ligates ENSE00003490828 and ENSE00003544843 exons to produce the RNA-mediated fusion transcript**.** qRT-PCR analyses support the expression of this gene ubiquitiously across all tissues tested**.** The ENST00000567078 fusion itself has been predicted to be part of the ribosome and unsurprisingly has an associated biological GO term for translation (Ashburner *et al.*, 2000).

3.4) Discussion

The aim of this chapter was to assess RMGFs both transcriptionally and translationally in order to enhance our understanding of their role in the evolution of vertebrates, specifically primate genomes, and their potential contribution to the evident phenotypic disparities across these species.

RMGFs were transcriptionally assessed for 1) the identification of RMGF transcripts and validating our identification protocol and 2) the comparison of RMGFs and their parents to other new genes brought about by alternative mechanisms (Section 1.2). RNA sequencing data was obtained for 6 species (Human, Chimpanzee, Gorilla, Orangutan, Macaque and Mouse) across a panel of 5 tissues (brain, cerebellum, heart, kidney and testis). Transcriptional profiling of a panel of RMGFs identified in Section 3.2.1.1 was carried out using 75bp Illumina Sequence Data (Brawand *et al.*, 2011), RT-PCR and qRT-PCR technologies.

RMGF analysis for frequency of genes expressed/present across all species was carried out at 3 read mapping thresholds; >=1 read mapped, >=3 reads mapped and >=5 reads mapped. However due to the limited number of proposed lowly expressed RMGFs >=1 category was chosen as providing sufficient evidence for

both the identification of the transcripts thus validating Section 2.2.1 pipeline but also that the RMGF is transcriptionally active in that tissue type. Those RMGFs with >=5 reads mapping suggests a greater support for the existence of these RMGFs in comparison to those with only 1 read mapping. Across all species examined elevated frequencies of RMGFs were found expressed in the testis, 25% of human RMGFs, 96% of chimpanzee RMGFs, 94% of gorilla RMGFs, 85% of macaque RMGFs and 3.2% of mouse RMGFs. Orangutan did not contain a testes RNA sample. Here, across all species examined elevated numbers of RMGFs were also expressed in the cerebellum (22.5%).

An identical analysis of RMGFs and their corresponding RMGF parents was carried out and the trends of expression fell in line with those of their corresponding RMGF apart from mouse where no testes elevation was identified but rather in heart and kidney tissues. These results must not be taken without all caveats considered. Although RMGF transcription can be validated *via* sequencing mechanisms the lack of transcription cannot, but rather only that the RMGF is not undergoing active transcription at that particular spatio-temporal time-point within the tissues being analysed. However these RMGFs could still potentially be transcribed in another tissue or in the same tissue at a different time point. Another caveat of the analysis is relying on external sequencing datasets that were produced ~8 years ago utilising the Illumina Genome II Analyzer (Brawand *et al.*, 2011). Since, a new Illumina short read (150bp) sequence by synthesis sequencer has become available. Other long read technologies have also become available from Pacific BioSciences and Oxford Nanopore (Rhoads and Au, 2015; Jain *et al.*, 2016). However due to the suspected low level of expression across RMGFs shorter read technologies are more appropriate (Conesa *et al.*, 2016). Future work could include the re-sequencing of the dataset using more up to date sequencing platforms and reanalysing RMGFs expression profiles. The importance of the most appropriate platform required to answer the question being asked of your data can not be underestimated as it is essential that high quality raw data is utilised for the production of accurate, high quality sequences (Conesa *et al.*, 2016). With this in mind our findings suggest that RMGFs have elevated numbers in testes which

falls in line with the proposed 'out-of-testis' hypothesis (Kaessmann, 2010); Nyberg and Carthew, 2017) whereby new genes have been found to test their expression in the testes with those found to enhance spermatogenesis, show an advantage for sexual conflict, or even demonstrate an ability to avoid germline pathogenesis undergoing positive selection (Nyberg and Carthew, 2017). These selective pressures can lead to the new gene becoming fixed in the genome where over time their range of expression can be broadened (Nyberg and Carthew, 2017). Our findings across RMGF parents are also supported by the literature as it has been shown that mammal specific genes tend to be expressed in the testes illustrating the strong sexual selection on the developmental organs of mammals.

To further characterise RMGFs transcriptional profile a differential expression (DE) analysis was carried out across all species. Here, two questions were being asked of the data 1) are RMGFs expressed differently in any particular tissue type and 2) are they expressed differently in the same tissue type across species. RMGFs were identified as having signatures of positive selection with only 3/36 human RMGFs showing DE, 4/29 in chimp, 33% in gorilla, 32% in orangutan, 5/34 in macaque and only a single RMGF in mouse. All three human genes showed DE between brain and cerebellum tissues. This signature was not found across any other species examined. This finding coincides with the divergence of *hominidae* and *homininae* species and the expansion of the cerebellum in human (Barton and Venditti, 2014).

A follow up experiment was then carried out to determine whether RMGFs were being expressed differently in brain tissue across species and results suggest they show signatures of differential expression across all species examined. This is interesting due to our finding that brain/cerebellum tissues were found also differentially expressed across human specifically. Therefore, it is unsurprising that this difference in expression levels is found between species particularly as primate and mouse brain size differs dramatically. DE analysis do come with certain limitations particularly when comparing expression levels of lowly expression levels of lowly expressed genes such as RMGFs through RNA

sequencing platforms (Łabaj and Kreil, 2016), and due to the complexity of the statistics required for detection with. In 2017 a new DE software package became available, Slueth, that utilises the most sophisticated statistics for DE detection currently available (Pimentel *et al.*, 2017). RNA sequencing platforms have been shown inferior to traditional microarray analysis in their sensitivity for lowly expressed genes and therefore have elevated false positive rates in comparison to traditional microarray analyses, however they have a more broad range of transcript detection and this must also be taken into consideration as. In future work these genes could be re-examined using high density human exon junction array (HJAY) technologies to determine the best platform to investigate RMGF DE as it has been shown that microarrays lack the ability to find novelties within genomes (Liu *et al.*, 2011; Su *et al.*, 2014).

An investigation of 35 RMGFs (human RMGFs with known sequences as well as primate RMGFs with human orthologs) alternative splicing profiles were carried out across a panel of vertebrates and results are shown in (Table 3.4). This analysis highlighted the probability of finding an RMGF transcript associated with genes with high numbers of AS isoforms, or indeed low levels of AS isoforms. Our analysis found 19/35 RMGFs examined had an associated AS isoform in at least one of the vertebrates analysed. Only 6/35 RMGFs had AS isoforms identified in human while 14/35 RMGFs has AS isoforms.

These results do not fall in line with what is currently present in the literature. It has been shown that 94% of human genes have an associated AS isoform (Pan *et al.*, 2008) and that vertebrates in general have elevated AS isoforms when compared to invertebrates (Kim *et al*., 2007). This suggests that high levels of AS is associated with genome complexity. From this one would think RMGF genes should have higher levels of AS in human in comparison to all other species examined, but this is not found in our dataset. This could suggest RMGFs in primates and human do not form AS isoforms at present due to their 'young' evolutionary age. It could also be due to the limitations of data quality available, and the ability to make comparisons between species that have such differences in coverage values (Chen *et al*., 2012). This fact is one that has been ignored for

the most part as all three databases; ASAPII (Kim *et al.*, 2007), AspAlt (Bhasi *et al.*, 2009), and ECgene (Lee *et al.*, 2007) of AS isoforms across vertebrates do not account for genome quality. For instance, the AspAlt database utilises Ensembl AS annotations (Table 3.7) (Zerbino *et al.*, 2018).

Table 3.7: A summary of Ensembl's alternative isoform information present in the AltAsp database

Ensembl Species	# of mRNA isoforms	# of genes with AS event
Human	3.27	41%
Chimp	2.8	32%
Macaque	2.79	31%
Mouse	2.89	30%
Rat	2.66	55%
Opossom	3.01	31%
Platypus	2.18	36%
Dog	2.38	21%
Cow	2.33	18%
Chicken	2.4	22%
Zebrafish	2.66	25%

A summary of AS isoforms annotated by the Ensembl Genome Browser (Zerbino *et al.*, 2018) *and utilised by the AltAsp database* (Bhasi *et al.*, 2009).

The AltAsp database is a publically available dataset whereby researchers can assess AS isoforms across a panel of 46 eukaryotic species (Bhasi *et al.*, 2009). However, no caveat is given to highlight that the lack of isoform evidence in the database does not mean that the isoform does not exist but rather it may be un-annotated within the genome being analysed. A complete assessment and functional annotation on a genome-wide scale is currently unfeasible and at present is carried out on a case-by-case basis. Therefore an accurate representation of AS isoform across the phylogenetic tree remains a huge challenge and a very slow, laborious process with genomes of interest accumulating a greater profile than non-model organisms this is unhelpful from a comparative genomics viewpoint. Although challenging the importance of AS analyses of this nature cannot be disregarded as isoforms have been demonstrated to play a significant role in vertebrate evolution (Barbosa-Morais *et al.*, 2012). However, it is not only genes that are actively translated that are of genomic importance as untranslated AS isoforms have been correlated with life-history traits such as longevity suggesting a role in genomic evolution e.g. POLB species-specific frequency across primate genomes (Skandalis *et al.*, 2010).

As validation of results obtained computationally by RNA sequencing follow-up RT-PCR and RT-qPCR were carried out. RT-qPCR was run on a panel of 29 human RMGFs across placenta, testis, lung, brain and liver tissues. Here, 25/27 genes analysed had an expression signature in at least one tissue and the highest number of RMGFs expressed was found in the testes (20/27 RMGFs) followed closely by the liver (19/27). These results support findings for RMGFs frequency of expressing found using RNA sequencing data, again supporting the 'out-of-testis' hypothesis (Nyberg and Carthew, 2017). However, low numbers of RMGFs are expressed ubiquitously across tissues. This could be as a result of these genes being slightly older and broadening of their expression profile has occurred due to mutational accumulation (Kaessmann, 2010).

RT-PCR analyses were then carried out on two human RMGFs (ENSG00000249773 and ENSG0000025062) with expected high expression levels across chimp and gorilla orthologs. Results supported expression of all

analyses apart from in frontal cortex and heart tissue of the chimp orthologs ENSPTRG000007442 where in frontal tissue expression was expected but no band appeared on the gel and in heart tissue two bands were formed suggesting primer dimer occurrence. However, due to the difficulty in obtaining Great Ape RNA samples the analysis could not be re-run to decipher the exact nature of the abnormality. Future work could use sequencing of fusion breakpoints to validate existence if required samples become available.

Experimental validations do come with a set of caveats that require consideration. Only RMGFs with unique, breakpoint spanning primers could be utilised for analysis therefore further reducing our RMGF sample size. Again, lack of expression does not mean lack of existence rather that is not transcribed in the dataset examined. The premise of experimental validation in itself also requires consideration. As assessing a small number of 'selected' genes to validate proof of concept appears to be the norm but may be a waste of valuable resources. Genome-scale analysis such as these require a randomised sampling of genes for validation as well as sufficient numbers of these genes to measure any kind of sensitivity, false positive and false discovery rates (Hughes, 2009). With this in mind a correlation has been found between RNA sequencing and qPCR (Everaert *et al.*, 2017).

Finally, in order to test for the possibility of RMGF translation four ribosomal datasets (skeletal, glioma and two fibroblast cells) were assessed. Here, 3/27 genes analysed had >=1 read mapping suggesting that ENST0000529564, enst00000446072, and ENST00000567078 are all undergoing active translation in fibroblast cell lines.

However, this finding is likely an underestimate caused by the limited data available for analyses as the low numbers of available datasets are due to the technology being relatively new in comparison to other profiling mechanisms. This does prove that RMGFs can undergo active translation and produce viable protein products. Although relatively new ribosomal sequencing technology has been frequently used in the literature (Gonzalez *et al.*, 2014; Duncan and Mata,

2017) and its power is unquestionable. However, similarly to RNA sequencing analysis care must be taken when presenting negative results. A further caveat of profiling sequencing technologies is the use of translation inhibitors to freeze translation with samples. This can cause the sample to alter its translation profile in response to its new stressful environment and may not give an accurate representation of the sample's translation output (Duncan and Mata, 2017).

Chapter 4: Computational prediction of the regulation of expression in RMGFs.

4.1) Introduction

Vertebrates have developed sophisticated mechanisms of gene expression regulation to deal with both their increased complex nature in comparison to prokaryotes, and the decoupling of the transcription and translation process after eukaryote and prokaryote divergence (Lynch and Conery, 2003). The regulation of the transcription process does not occur at transcription initiation but rather beforehand through epigenetic mechanisms (Jaenisch and Bird, 2003).

Epigenetics means 'around genetics', alluding to the fact that any epigenetic alternations made to genetic information are not permanent fixture but rather are both added and removed depending on the cell's requirements, for example a shift in the cells spatio-temporal context, development stage or even in response to either an internal or external stimulus (Turner, 2009). These epigenetic mechanisms require both activating and repressing regulatory elements that can bind to *cis*-regulatory sequences surrounding a given gene in order to either enhance it's transcription or indeed impede it. These regulatory elements include 1) chromatin remodellers, 2) histone modifications, 3) transcription factors, and 4) splice factors. This panel of elements help control the gene expression profile across vertebrate genomes.

The first barrier to eukaryotic gene expression is the default heterochromatic state of DNA that yields genes inaccessible to transcriptional machinery such as RNA polymerase. For gene expression initiation chromatin remodellers cause a conformational transition from a tightly coiled heterochromatic state into a more loosely bound 'beads-on-a-string' euchromatic conformation optimised for transcription machinery accessibility (Li, Carey and Workman, 2007).

With the release of the 111 epigenomes in 2015 from the NIH RoadMap Epigenomics Consortium (Roadmap Epigenomics Consortium *et al.*, 2015) it became possible to assess genomes for regions of both heterochromatin alluding to transcriptional inactivity and euchromatin suggesting gene expression. Not only this but it became possible to predict chromatin conformation on a gene level and infer transcriptional activity level.

Chromatin conformation alone however does not confirm gene expression but rather just increases the likelihood of it occurring whilst in certain conformations therefore additional regulatory elements are required such as histone modifications. As outlined in Section 1.5.2.3. DNA is wrapped around two sets of 4 histones with each histone containing a protruding tail that can be modified and these modifications either cause transcriptional enhancement or repression. There are four main histone modifications; phosphorylation, ubiquitination, methylation and acetylation all of which effect transcription rate (Dong and Weng, 2013). For example, in an examination of four upregulated genes found across suicide victims, the histone h3k4me3 marker was found across promoter regions highlighting the modifications role in transcriptional activation (Fiori, Gross and Turecki, 2012). Moreover, an investigation of memory lymphocytes in mouse revealed a positive correlation between transcriptional activation and h3k4me3 and a negative correlation with h3k27me3 (Araki *et al.*, 2009). The current release of the Roadmap Epigenomics Database **(Version 9)** (Roadmap Epigenomics Consortium *et al.*, 2015) contains 1821 histone datasets examining a selection of activating modifications; h3k36me3, h3k4me1, h3k4me3 and h3k9ac and a repressive marker; h3k27me3 across a plethora of human tissues.

Transcription factors (TFs) are another cohort of regulatory elements essential for transcriptional control. TFs are DNA binding proteins that bind 6-8bp *cis*-sequence motifs positioned at gene promoters either activating or inhibiting gene expression through RNA polymerase recruitment toward or impedance from the TSS of a given gene (Pan *et al.*, 2010). For instance the PU.1 TF has been found to repress gene expression in macrophages and B lymphocytes (Borras *et al.*, 1995) whilst the Sp1 and Sp3 TFs enhance gene expression in human endometrium (Krikun *et al.*, 2000). Some TFs can even change from an activating to a repressive role depending on binding context, for example the Pit1 TF can function as both a transcription activator or repressor depending on the motif it binds (Latchman, 2001). These regulatory elements can significantly alter a genes expression profile and therefore accurate control of TFs is essential for cell survival with aberrant TF expression commonly resulting in

malignancies, for example both p53 and c-myc TFs are the most common genes found altered within tumours (Patricia A.J. Muller1, 2014; Tansey, 2014).

Another level of transcriptional control is implemented by splice factors (SFs). Splice factors are DNA binding proteins that bind specific *cis*-sequences within genes in order to remove/include introns prior to mature mRNA creation (Long and Caceres, 2009; Geuens *et al*., 2016). The sequential removal of intronic sequences from between adjacent exons is known as constitutive splicing and this produces 'major isoforms', but when an exon is skipped, an alternative 3'/ 5' splice site used or an intron retained this is known as alternative splicing and produces 'minor isoforms' (Lees *et al*., 2015). 95% of human transcripts are now thought to have minor isoforms (Wang *et al.*, 2008). Splicing is a powerful tool and has been essential in expanding the protein repertoire of complex eukaryotes (Nilsen and Graveley, 2010) with the removal or insertion of novel sequences into proteins potentially affecting protein structure (Birzele, Csaba and Zimmer, 2008), functional activity (Neverov *et al.*, 2005) and viability (Birzele *et al*., 2008). However, its importance is also demonstrated in its ability to control transcription of genes this is highlighted through mechanisms such non-sense mediated decay of mRNAs. Non-sense mediated decay can be caused by the retention of a large intron sequennce (Ge and Porse, 2014) followed by a frameshift mutation due to alternative splice site usage. This can result in early transcriptional termination (Withers *et al*., 2012).

As previously mentioned the sophistication of vertebrate genomes requires a highly efficient, tightly controlled method of regulating gene expression across tissues and all four regulatory elements described provide vertebrate genomes with this level of intricacy. Fortunately these elements have been extensively studied, initially on a case-by-case basis but with the advancement in sequencing technologies these elements are now being analysed on genome-wide levels, with many datasets becoming publically available such as the RoadMap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015) and predictive software packages to analyse these datasets being developed e.g. SFmap (Paz *et al.*, 2010) for SF binding site predictions and the JASPAR dataset (Khan *et al.*,

2018) for transcription factor predictions. However, although datasets and software packages are abundant caution must be taken particularly whilst using publically available datasets. Only high quality, comparable datasets should be used in analyses and quality checks should be ran on all data prior to analyses. The use of predictive software packages like SFmap and JASPAR also come with limitations as these packages limit their database only to include experimentally validated elements, therefore results could provide a misleading incomplete picture of how expression is being controlled. Another confounding feature of predictive software packages is high false positives rates caused by algorithm insensitivity or even the user using the tool as a 'black-box' and not customising parameters for their own specific analyses.

Once these limitations are considered, both the datasets and predictive tools provide insightful information about regulatory elements and their potential impact on transcription. To date, not much is known about the regulation of new gene expression apart from their prevalence for testes-specific expression (Kaessmann, 2010). This chapter aimed to provide an insight into the transcriptional regulation of new genes through the analyses of RMGFs. More specifically the following questions were addressed:

1) Do RMGFs have a tendency to be positioned in either euchromatin or heterochromatin?
2) Do RMGFs have a SF usage pattern different to that of non-fused protein coding genes? Do certain SF binding sites show a bias toward co-existing on the same gene as another SF binding site? If so is the same bias found across non-fused protein coding genes?
3) Are RMGFs enriched for either activating or repressive histone modifications? And is this bias tissue dependent?
4) Are RMGFs controlled by specific TFs? Or TFs that restrict expression to certain human tissues? Do the TFs found across RMGFs co-occur with histone modifications?

4.2) Materials and Methods

4.2.1) An investigation of activating and repressing signal profiles across RMGFs

NIH Roadmap Epigenomics core15-state model mnemonic BED files were downloaded across lung, liver, heart, brain (frontal lobe), spleen, small intestine, placenta (amnion) and embryonic cell lines from the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015). These data contain chromosomal start and finish positions as well as the mnemonic/abbreviation associated with the corresponding chromosomal region. Mnemonics and their descriptions are provided in Table 4.1.

The chromosomal coordinates of 14 human RMGFs were downloaded using the Ensembl Genome Browser (Herrero *et al.*, 2016). These genes were selected based on human RMGFs identified at 90 PI and their presence in the NIH Roadmap Epigenomics core15-state model database. An in-house Perl script (**Code_Box 1**) extracted only those mnemonics within the RMGFs chromosomal co-ordinates. Activator mnemonics were grouped together to create activator panels and repressor mnemonics were grouped together to create repressor panels. Activator and repressor panels were then compared on a gene-by-gene basis firstly by the calculation both the number of activators and the number of repressors across the 14 RMGFs in all 8 human tissues examined and secondly by the calculation of the average number of both activators and repressors present across all RMGFs.

Table 4.1: Characteristics of mnemonics from the Roadmap Epigenomic Consortium (Roadmap Epigenomics Consortium *et al.*, 2015)

State Number	Mnemonic	Description	Transcriptional Impact
1	TssA	Active TS	Activator
2	TssAFlnk	Flanking Active TSS	Activator
3	TxFlnk	Transcr. at gene 5' and 3'	Activator
4	Tx	Strong transcription	Activator
5	TxWk	Weak transcription	Activator
6	EnhG	Genic enhancers	Activator
7	Enh	Enhancers	Activator
8	ZNF/Rpts	ZNF genes & repeats	Activator
9	Het	Heterochromatin	Repressor
10	TssBiv	Bivalent/Poised TSS	Activator
11	BivFlnk	Flanking Bivalent TSS/Enh	Activator
12	EnhBiv	Bivalent Enhancer	Activator
13	ReprPC	Repressed PolyComb	Repressor
14	ReprPCWk	Weak Repressed PolyComb	Repressor
15	Quies	Quiescent/Low	Repressor

The state number assigned to each mnemonic within the core15-state model mnemonic database, along with a brief description of the mnemonic (column 3) and its activator or repressor status (column 4) are shown.

Code_Box 1: Perl script to extract activator and repressor information for RMGFs from core15-state model mnemonic BED files

```
1  open(IN, "input_file");
2  while(<IN>){
3    chomp;
4  #   print "$_\n";
5    if ($_ =~ /chr_number/){
6      @chr = split/\s/,$_;
7      for $e (0..$#chr){}
8  #     print "$chr[1]\n";
9      if ($chr[1] >= start_chr_coord && $chr[2] <=
10 end_chr_coord){
11       print "$chr[0]\t$chr[1]\t$chr[2]\t$chr[3]\n";
12     }
13   }
14 }
15
   close(IN);
```

Code_Box_1: *Line 1 inputs a list of RMGF chromosomal coordinates and compares them to the chromosomal coordinates of known mnemonics in BED files obtained from the Roadmap Epigenomic consortium database* (Roadmap Epigenomics Consortium *et al.*, 2015). *Line 1 and 2 read in a specified .bed file and assess it for content. Once the file is not empty line 3 and 4 remove new lines and a check is carried out. Line 5 and 6 identify lines that contain a user specified chromosome number, if this is present it splits the line based on negative space and inserts it into an array. Line 7 and 8 read in the array and line 9 checks if each element of the array contains a number between 2 user defined chromosomal coordinates, if it does line 10 prints the line.*

4.2.2) An analysis of splice factors across RMGFs

4.2.2.1) A comparison of splice factor binding sites (SFBSs) across RMGFs and human non-fused protein coding genes

Chromosomal coordinate data was acquired from the Ensembl Genome Browser (**Version_73)** (Aken *et al.*, 2017) for each of the 37 RMGFs and across all non-fused human protein-coding sequences. The non-fused protein coding dataset was used to generate random samplings of 100 datasets of 37 genes in size. The number and position of SFBSs identified in each of the randomly sampled

datasets were then compared to those found in the RMGFs. A python script ("randomsimulation.pl", Appendix_B) was designed to perform the sampling.

SFmap **(Version 1.8)** identifies putative SFBSs for known functionally annotated splice factors across datasets (Table 4.2) (Paz *et al.*, 2010). This software was selected as it bases all binding site predictions on both the genomic environment of the flanking sequence as well as the evolutionary conservation of the SF *cis*-regulatory sequences – unlike other available packages that base predictions solely on the overabundance of the SFBS in regulatory regions (Fairbrother *et al.*, 2004) or those based completely on experimental binding data (Cartegni *et al.*, 2003).

Due to the short sequence length and degenerative nature of SFBSs a two pronged approached is implemented by the SFmap algorithm that calculates a conservation of score (COS) and weighted rank (WR) for each predicted SFBS. COS (WR) carries out a sequence similarity search to identify SFBSs, this is followed by an assessment of the sequence's environment to analyse it's ability to cluster and a WR is given (Paz *et al.*, 2010). Each SFBS is examined for evolutionary conservation by alignment with the mouse genome and additionally assigned a COS score. Each predicted SFBS is assigned a COS (WR) score ranging from 60-90. SFmap was ran across all motifs (23 annotated motifs available in total (Table 4.2) at medium stringency levels, as per the software package recommendation for analyses of this nature (Paz *et al.*, 2010).

Table 4.2: The panel of SFBSs represented in the SFmap software package (Paz *et al.*, 2010)

SFmap Annotated SFBS	Binding Motif	Reference
FOX1	-	(Zhang *et al.*, 2008)
SF2ASF	ugrwgv	(Tintaru *et al.*, 2007)
YB1	-	(Lasham *et al.*, 2003)
hnRNPF	gggug	(Caputi and Zahler, 2001)
hnRNPF	gugkau	(Caputi and Zahler, 2001)
hnRNPF	gukgykg	(Caputi and Zahler, 2001)
hnRNPH1	gargag	**Buckanovich, Posner, & Darnell, 1993)**
MBNL		(Timchenko *et al.*, 1996)
NOVA1	-	(Buckanovich, Posner and Darnell, 1993)
PTB	cucucu	(Gil *et al.*, 1991)
PTB	ucuu	(Gil *et al.*, 1991)
QK1	acuaay	(Feng and Bankston, 2010)
hn**RNPA1**	guaguagu	(Biamonti *et al.*, 1989)
hn**RNPA2B1**	aggwuhgr	(Kozu, Henrich and Schäfer, 1995)
SC35	gryymcyr	(Richardson *et al.*, 2011)
SC35	ugcygyy	(Richardson *et al.*, 2011)
SF2ASF	ugrwgvh	(Tintaru *et al.*, 2007)
SRp20	cuckucy	(Huang and Steitz, 2001)
SRp20x	wcwwc	(Huang and Steitz, 2001)
SRp55	yrckm	(Tran and Roesser, 2003)
Tra2alpha	gaagaggaag	(Tacke *et al.*, 1998)
Tra2beta	aguguu	(Tacke *et al.*, 1998)
Tra2beta	ghvvganr	(Tacke *et al.*, 1998)

Highlights the 23 annotated human SFs (Column 1) and their specific binding site motif (Column 2) used by the SFmap software package for the assessment of RMGFs. SFmap motifs were determined based on the literature (column 3), genome-wide computational prediction and subsequent experimentally validation (Akerman *et al.*, 2009).

In order to accurately compare across RMGFs a gene length adjustment was required to account for differences in gene length distribution. The Ensembl Genome Browser (Aken *et al.*, 2017) was used to acquire gene length in kilobases (kB) for all genes within both RMGF and the 100 simulated datasets. This calculation allowed for a per kB adjustment of SFBS frequency across genes per kB so that SFBS frequencies could be compared more accurately between datasets.

To investigate and potentially identify bias in SFBS usage between human RMGFs and non-fused genes a comparison of the frequency of each of the 23 SFBS identified across all human RMGFS was made to the frequency of SFBS found in the human non-fused protein coding gene simulated datasets. Data distribution was assessed **(Code_Box 2)** and was not normally distributed therefore standard parametric comparisons could not account for the stochastic nature of the distributions and thus non-parametric Mann Whitney U tests were used to compare distributions across datasets. P-values were adjusted for multiple testing using the R p.adjust function. Co-occurence of each SFBS on each RMGF was investigated on the same datasets in order to compare the co-occurence of SFBSs within RMGFs and the relationship between RMGFs in non-fused protein coding genes. This was carried out using in-house scripts (Appendix_B). Again, as the same datasets were used as per frequency distribution, significance testing required non-parametric statistics such as the Mann Whitney U test followed by and p-values extrapolation and multiple testing adjustments by the R function p.adjust (Benjamini-Hochberg correction) (Appendix_B).

Code_Box 2: Python script to generate human non-fused protein coding SFBS frequency distribution graphs prior to statistical significance testing

```
1  #!/usr/bin/python
2  #usage python3 annGraph.py infile outfile title
3  import sys
4  infile=sys.argv[1]
5  outfile=sys.argv[2]
6  rmgf_av=sys.argv[3]
7  title=sys.argv[4]
8  import matplotlib.pyplot as plt
9  import pandas as pd
10 db=pd.read_csv(infile)
11 db.set_index(db['gene'])
12 plt.figure()
13 db.hist(column=['non-fused SFBS'], grid=False, bins=20)
14 plt.axvline(rmgf_av, color='red', linestyle='dashed',
15 linewidth=3)
16 plt.title(title)
17 plt.savefig(outfile)
```

Code_Box 2: *Python script run across human un-fused protein coding genes in order to obtain distribution pattern. If a normal distribution was found standard parametric t-tests can be carried out, however if distributions were non normal a logarithmic conversion or Mann Whitney U was needed to accurately infer statistical significance. Line 4,5, 6, and 7 read are user specified inputs, input file name, output file name and the average of a particular SFBS found across all RMGFs analysed and graph title. Line 8 and 9 imports the specific software packages need to generate distributions, namely matplotlib and pandas. Line 10 and 11 reads in a specified .csv file and send all the elements of the column 'gene' into an index. Line 12 and 13 takes this index and plots a histogram of the data with blue bars indicating human non-fused distribution and the red dashed line illustrating the RMGF average. Line 16 and 17 save this plot to an output file.*

4.2.2.2) An assessment of splice factor transcriptional profiles across the ENCODE database (Harrow *et al.*, 2012)

From Section 4.2.2.1 the frequency of each SFBS was obtained, in order to understand what impact this may have on RMGF transcription, the transcription profiles for each SFBSs corresponding splice factor (SF) were obtained from the Expression Atlas' ENCODE dataset for the following human tissues: testes, spleen, intestine, colon, pancreas, ovary, lung, liver, kidney, heart, brain, adrenal gland and adipose tissue (Consortium *et al.*, 2012). The abundance of each SFBS across RMGFs calculated in Section 4.2.2.1 as well as the SF transcriptional profiles were mapped onto a weighted bipartite network using the Cytoscape software package (Shannon *et al.*, 2003). Edges were drawn between SFBS nodes and RMGF nodes to represent SFBS presence within that RMGF and the width of the bar is based on the abundance of the binding site within that gene, for example in Figure 4.5 the NOVA1 binding contains a large number binding sites across the RMGF - ENST00000540732 as indicated by a thick edge joining the two individual node. Contrastingly, PTBucuu contains a low number of binding sites across this gene highlighted by a thin edge connecting the two nodes. The SFBSs were categorised according to the tissue in which their corresponding SF indicated the highest level of expression according to the Expression Atlas Database (Papatheodorou *et al.*, 2018).

4.2.2.3) An assessment of SFBS frequency across the fusion breakpoint of RMGFs

SFBSs present across the fusion breakpoint have a greater likelihood of playing a role in controlling the gene expression profile of the RMGF specifically and not their parents. In order to examine SFBS abundance around the fusion breakpoint of RMGFs pairwise alignments of RMGFs and their parent genes were constructed using the PRANK alignment software package (Löytynoja and Goldman, 2010) and precise fusion breakpoints were identified from the alignment. An investigation of SFBSs spanning the exons surrounding each RMGF's fusion breakpoint was carried out and the expression of each SFBS identified corresponding SF gene expression profile ascertained (Papatheodorou *et al.*, 2018). Gene expression for each RMGF's parent gene was downloaded

from the ENCODE database (Harrow *et al.*, 2012) and compared with each SF gene predicted to control the RMGF's transcriptional profile. Through the investigation of 1) SFBS usage across fusion breakpoints, 2) their corresponding SFs expression profile and 3) the expression profile of each RMGFs parent gene an understanding of the impact of SFs may have on the transcriptional control of RMGFs could be determined. Results will help determine whether RMGF expression profiles are likely to be consistent with that of their corresponding parents or whether a novel profile is more likely.

4.2.3) An analysis of histones across RMGFs

4.2.3.1) Data acquisition and histone abundance calculation across RMGFs

Five specific histone markers were selected based on tissue availability and the effect they have on gene control, i.e. transcriptional activation (h3k36me3, h3k9ac, h3k4me1 and h3k4me3) and repression (h3k27me3). Histone marker binding sites were downloaded from the Roadmap Epigenomic Database (**Release 9)** through the GEO repository (Roadmap Epigenomics Consortium *et al.*, 2015)**.** The number of histone markers predicted for each RMGF and each randomly sampled non-fused dataset were determined. For histone h3k4me3, h3k4me1, h3k9ac, h3k36me3, h3k27me3 tissues examined are shown in represented in Table 4.3 – 4.7.

Table 4.3: Tissue samples utilised during the analysis of the h3k4me3 histone modification

Tissue Sample	Reference ID
Embryonic stem cell	GSM409308
Embryonic stem cell	GSM410808
Embryonic stem cell	GSM433170
Fetal lung	GSM469970
Liver	GSM537697
Liver	GSM537697
Brain	GSM669624
Brain	GSM670016
Kidney	GSM773005
Pancreas	GSM910581
Lung	GSM915336

Illustrates each tissue sample used in the h3k4me3 histone marker analysis and its corresponding project reference number in the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015).

Table 4.4: Tissue samples used during a h3k4me1 histone modification analysis of RMGFs

Tissue Sample	Reference ID
Embryonic stem cell	GSM409307
Embryonic stem cell	GSM433177
Embryonic stem cell	GSM466739
Ovary	GSM1013148
Placenta	GSM1127129
Liver	GSM537706
Brain Frontal Lobe	GSM670015
Kidney	GSM670025
Kidney	GSM773001
Pancreas	GSM910576
Heart	GSM910575

Tissue samples used and their corresponding reference numbers from the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015) *in a h3k4me1 histone modification analysis of RMGFs.*

Table 4.5: Tissue samples used during an analysis of the histone modification h3k9ac across RMGFs

Tissue Sample	Reference ID
Embryonic stem cell	GSM410807
Embryonic stem cell	GSM433171
Embryonic stem cell	GSM434785
Liver	GSM537705
Brain Frontal Lobe	GSM670021
Kidney	GSM772811

An assessment of h3k9ac histone modifications utilising the tissue samples from column 1 from the data repository reference number in column 2 across RMGFs and their corresponding reference numbers from the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015) .

Table 4.6: Tissue panel obtained from the Roadmap Epigenomics Database for a h3k36me3 histone modification analysis

Tissue Sample	Reference ID
Embryonic stem cell	GSM409312
Embryonic stem cell	GSM428296
Embryonic stem cell	GSM433176
Embryonic stem cell	GSM450268
Embryonic stem cell	GSM466737
Lung	GSM956014
Ovary	GSM1013143
Brain frontal lobe	GSM669982
Pancreas	GSM910570

Depiction of the tissue panel obtained from the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015) *to assess the usage of h3k36me3 histone modification across RMGFs.*

Table 4.7: Tissue panel obtained from for a h3k27me3 histone modification analysis of RMGFs obtained from the Roadmap Epigenomics Database

Tissue Sample	Reference ID
Embryonic stem cell	GSM428295
Embryonic stem cell	GSM433167
Embryonic stem cell	GSM434776
Embryonic stem cell	GSM466734
Liver	GSM537698
Placenta	GSM1127139

An illustration of the tissues examined during an assessment of h3k27me3 histone modifications across RMGFs and their corresponding project reference numbers from the Roadmap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015).

4.2.3.2) Assessing the relationship between histone markers and splice factor usage in RMGFs

The number of histone marker binding site and SFBSs were taken from Section 4.2.3.1and Section 4.2.2.1 respectively. Linear regression analyses were carried out using the RStudio package (R Development Core Team, 2011).

4.3.4) An assessment of annotated and experimentally validated transcription factor binding sites from the JASPER database across RMGFs

Co-ordinates of 37 RMGFs identified at 90 PI were downloaded from the Ensembl Genome Browser **(Version_90)** (Aken *et al.*, 2017)**.** There were 18 deemed suitable as input for JASPER analysis **(Release 7)** (Wasserman and Sandelin, 2004). Suitability was based on chromosomal co-ordinate availability of the RMGF within the JASPER reference genome. Default settings were used and only TFBS on the correct strand (sense or anti-sense) were considered. Expression profiles of each TFBS's corresponding TF gene were assessed using the ExpressionAtlas ENCODE (Harrow *et al.*, 2012) database in order to identify TFs with tissue specific or ubiquitous profiles of expression.

4.3) Results

4.3.1) Computational characterisation of activator and repressor signals in human RMGFs across 8 tissues

Gene regulation is dependent on RNAP access to a TSS, this is chromatin dependent- either in a euchromatic or heterochromatic state as described in Section 4.1. Chromatin state is controlled by many *cis*-regulatory features, these features can either act to loosen the chromatin into a euchromatic state to allow RNAP entry or act as repressors tightening the chromatin through heterochromatin formation. Through the use of both activation and repression signalling data across 127 human epigenomes (Roadmap Epigenomics Consortium *et al.*, 2015) an insight was gained into the combinatorial interactions between different chromatin (both activating and repressing) marks in their spatial context. This provided the adequate data to investigate these

signals and consequently gain insights into the chromatin context of RMGFs furthering our understanding of their transcriptional regulation.

A full profile of activation and repressor signals for 14 RMGFs across 8 human tissues (lung, liver, small intestine, spleen, brain (frontal lobe), placenta, ovary and embryonic cells) was obtained and results are summarised in Figure 4.1. In lung tissues most RMGFs contain a higher number of activation signals apart from the ENST00000465127 transcript. This gene not only has higher repressor signal but overall has higher activator and repressor signals in lung than all other RMGFs (Figure 4.1 (h)). In brain tissue activating signals appear more common which may be indicative of transcriptional activation in human brain frontal lobe tissues. However, again the ENST00000465127 transcript appears to be highly repressed (Figure 4.1 (b)). In human liver tissue panels activating signals are more prominent across all RMGFs (Figure 4.1 (a)) and this holds true in placenta amnion cell lines also (Figure 4.1 (e)). In the human small intestine (Figure 4.1 (f)), embryonic stem cells (Figure 4.1 (c)) and ovary (Figure 4.1 (d)) cell line analyses across RMGFs activating signals are more prominent with the only transcript showing signals of repression again being ENST00000465127 transcript.

4.3.2) Computational characterisation of splice factor binding site usage across RMGFs

Using the SFmap, 21 gold standard SFs and 35 of their corresponding SFBSs were examined across 37 RMGFs using a range of COS (WR) score thresholds – 70, 80 and 90 scores**.** As COS (WR) scores increase from 70 to 90 the number of SFBS identified decreases across a gene. Here, 1861 SFBSs spanning 34/35 SFBS were predicted at the 70 score threshold, 1400 SFBSs spanning 28/35 SFBS at 80 score and 628 spanning only 19/35 SFBS were predicted at the 90 threshold. This indicates that binding sites identified at 90 scores are more stringent than those at 70 (Figure 4.2). A COS (WR) score of 90 or above was selected for all further analyses to reduce mis-matched motif calls due to its stringent two step calculation; 1) motif clustering analysis within regulatory regions and 2) an evolutionary conservation analysis through human and mouse ortholog alignment.

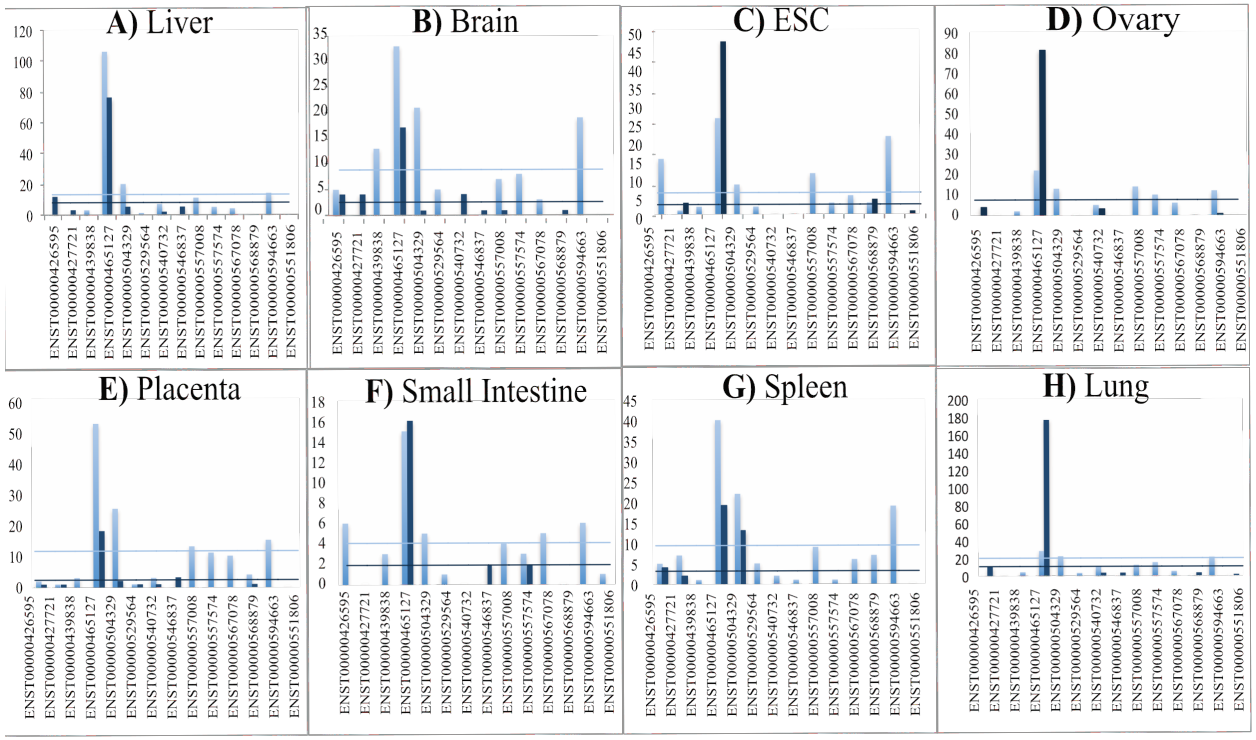


Figure 4.1: *An assessment of human RMGF activator and repressor signal profiles across 8 human tissues:* *Results of an RMGF activator and repressor abundance analysis carried out across 8 human tissues. Frequency of activation is shown by pale blue bars with the pale blue trend line highlighting the average frequency of activators across all 14 RMGFs examined. Dark blue bars indicate the frequency of repression signals in each RMGF, with the dark blue line depicting the average repressor frequency across all RMGFs under assessment.*

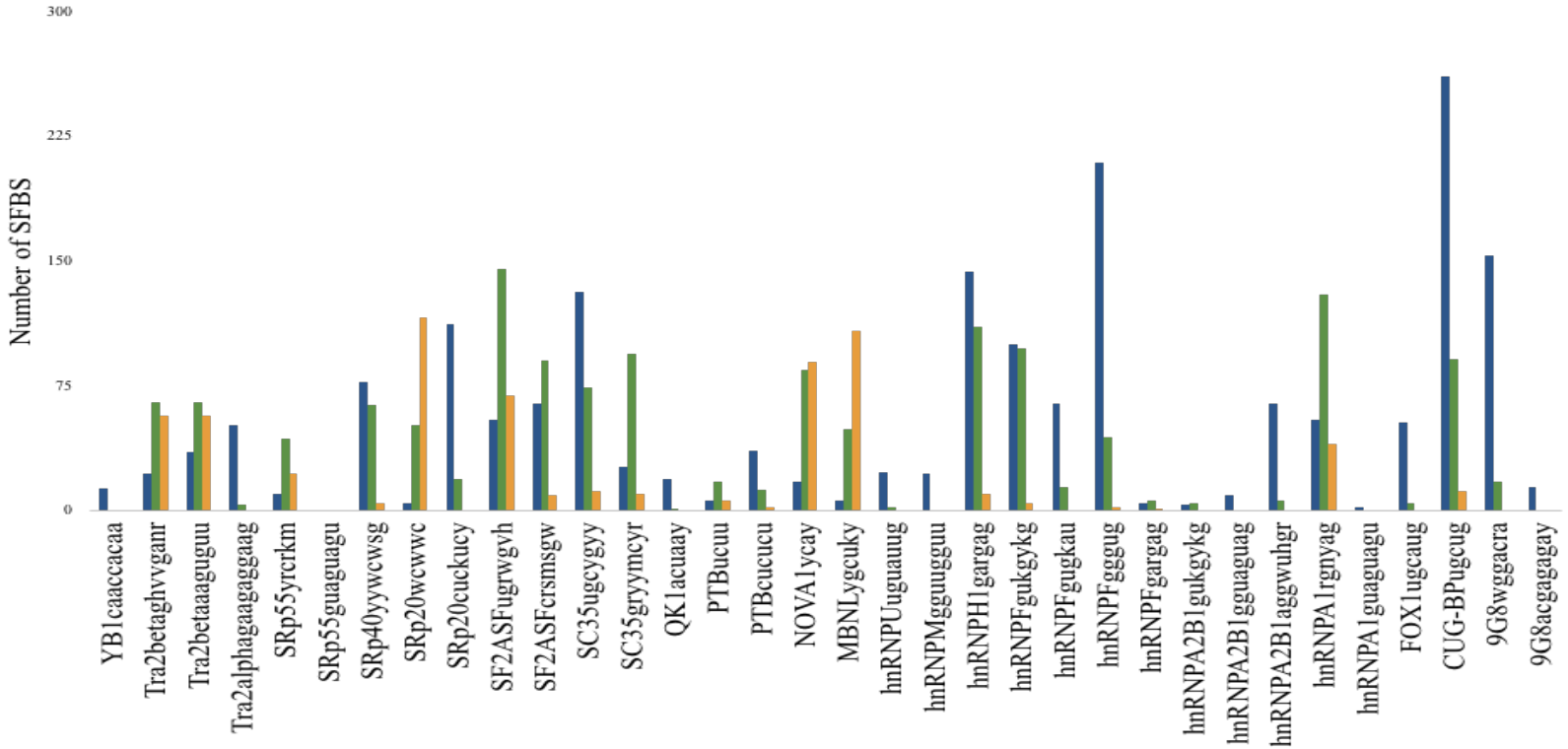


Figure 4.2: *A comparison of SFmap COS (WR) scores calculated during an assessment of human RMGFs for experimentally validated SFBSs: Known gold standard SFBSs are illustrated on the X axis and the number of each SFBS present across RMGFs on the Y axis. This image examines the number of SFBS predicted across each COS (WR) score category; 90 score (yellow bar), 80 score (green bar), 70 score (blue bar).*

The presence or absence analysis of the 35 SFBS (specific to 21 SFs) across the fusion breakpoint of 20/37 human RMGFs was carried out (only 20 RMGFs exact fusion breakpoint co-ordinates were recognised by the SFmap software package). 19/21 SF genes were found to contain binding sites spanning the breakpoint of these 18/20 RMGFs analysed. In order to investigate the potential transcriptional profiles of these genes based on the SFBS analysis the expression pattern each SFBS's corresponding SF was also obtained (Papatheodorou *et al.*, 2018).

Results uncovered 85% of RMGFs had SF2ASF binding sites, which are predominantly, expressed spleen, the adrenal gland and brain indicating a high number of RMGFs may be transcribed within these human tissues (Figure 4.3) (Papatheodorou *et al.*, 2018). 80% had SRp20 binding sites and SRp20 shows highest expression levels in human spleen, lung and testes tissues (Figure 4.3). NOVA1 binding sites were located across 70% of RMGFs examined and this SF has elevated profiles of expression in brain, adrenal gland and testes tissues (Figure 4.3), 60% had Tra2Beta binding sites and Tra2Beta has been found with high expression across spleen, the adrenal gland and testes tissues (Figure 4.3). 55% of RMGFs were found with SRp55 binding sites spanning their fusion breakpoint and this SF has been found with high signatures of expression in the spleen, pancreas and adrenal gland (Figure 4.3). MBNL binding sites were identified for 50% of RMGFs and MBNL had high expression profiles identified in human spleen, lung and adipose tissue (Figure 4.3). 40% had hnRNPA1 binding sites and this SF has elevated levels of expression in ovary, spleen and lung (Figure 4.3), 25% of RMGFs contained SC35 binding sites but no expression information was found for the SF, however 25% of RMGFs also contained CUG-BP binding sites which have elevated expression in the spleen, testes and adrenal gland (Figure 4.3). 20% of RMGFs had hnRNPF binding sites however expression data for this SF remains unavailable. 15% had PTB binding sites and this SF was found highly expressed in testes, spleen and the colon tissues (Figure 4.3). hnRNPH1 binding sites were also found across 15% of RMGFs examined and show elevated levels of expression in spleen, small intestine and adipose (Figure 4.3). 10% contained SRp40 binding sites and have

an elevated expression signature in spleen, adipose tissue and the sigmoid colon (Figure 4.3), and 5% had both hnRNPA2B1 and hnRNPM binding sites with hnRNPA2B1 showing high expression in spleen, the adrenal gland and testis and the hnRNPM SF in testis, spleen and brain (Figure 4.3).

Not only was the presence or absence pattern of each SFBS obtained across RMGFs but also the sum of each binding site spanning each RMGF's breakpoint was analysed and compared across RMGFs (Figure 4.4). Interestingly, NOVA1 binding sites were present across 70% of RMGFs examined and 93% of these RMGFs were found to have NOVA binding sites at increased levels in comparison to other SFBSs examined this is indicated by orange cells in Figure 4.4. In RMGFs where SRp20 binding sites were present, 88% contained this binding site at increased levels (Figure 4.4). This demonstrates that not only are these binding sites present across RMGFs more frequently than all other SFBS, but also demonstrates that when they are present they are found in much higher numbers thus increasing their probability of controlling the RMGF expression profile.

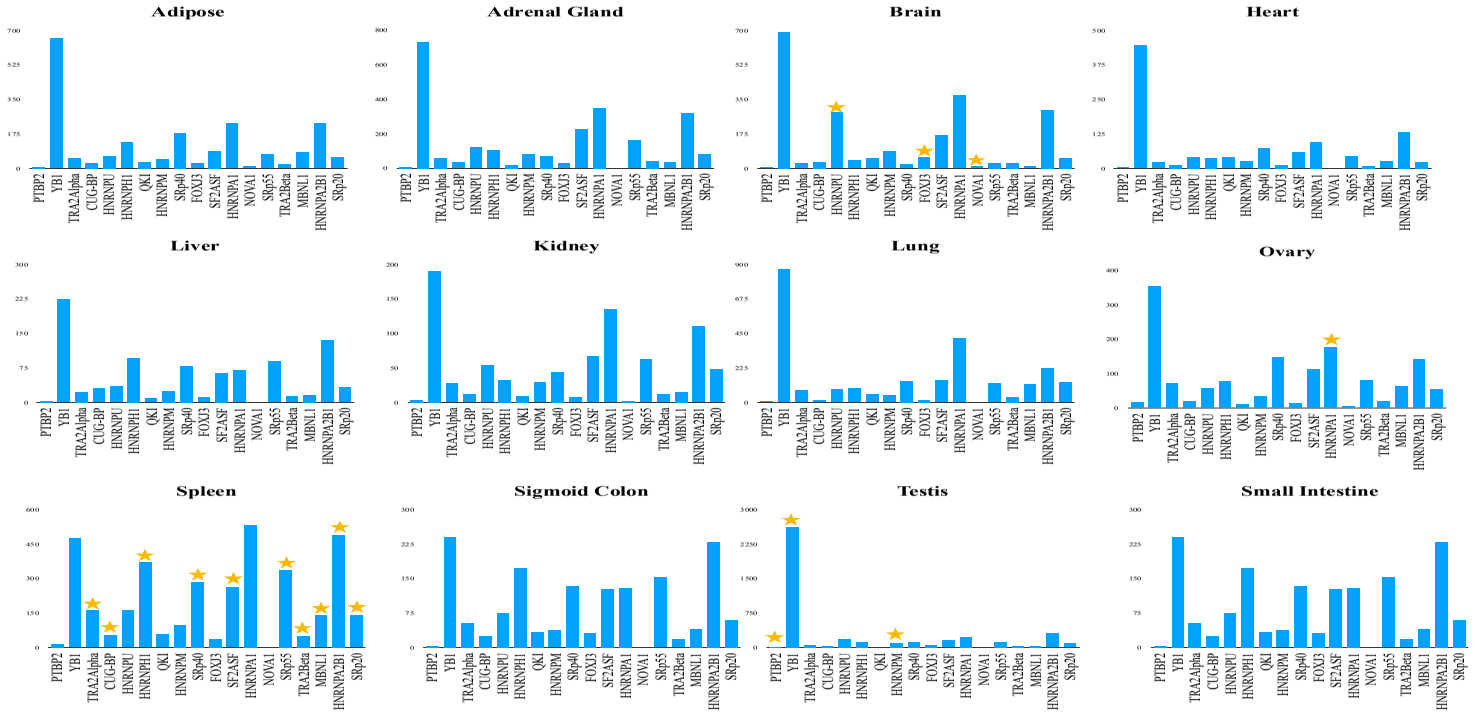


Figure 4.3: *Expression profiles of Splice factors associated with SFBS motif's across a panel of 12 human tissues obtained from the ENCODE database: A transcriptomic analyses of the 21 splice factor genes corresponding to each 35 SFBS analysed across 12 human tissues. Human tissue data was available for 18 SF genes from ExpressionAtlas ENCODE database* (Harrow *et al.*, 2012)*. Yellow stars represent that SF shows it's highest level of expression in that tissue when compared to all other tissues analysed.*

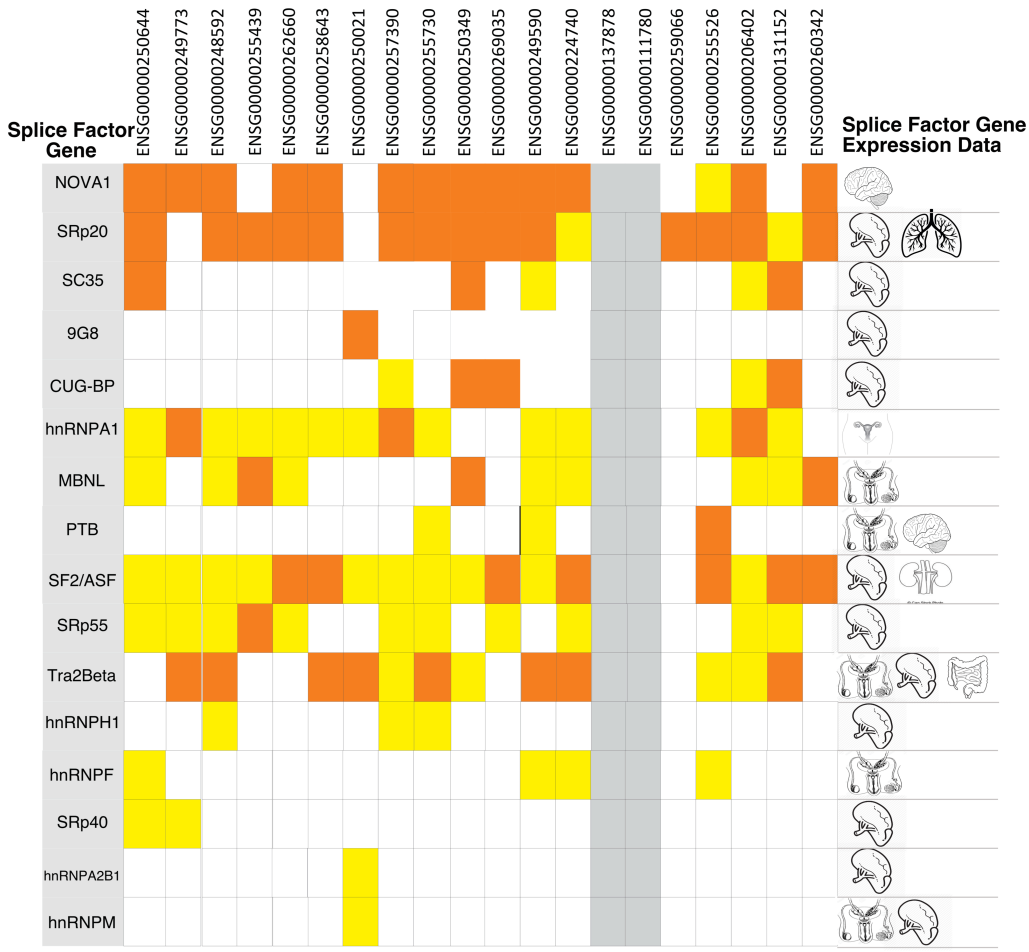


Figure 4.4: *Presence/absence of 16 SFBSs across 20 RMGFs and the tissue in which their corresponding splice factor is expressed at the highest level: SFBS presence/absence analysis carried out across RMGFs with orange or yellow cells indicating SFBS presence, however if white this indicates absence. Orange cells represent high numbers of SFBSs are present and yellow cells a low abundance of SFBS present, and grey cells represent unknowns. On the left hand side of the diagram the most abundantly expressed tissue for that corresponding splice factor according to the Expression Atlas Database is depicted* (Papatheodorou *et al.*, 2018).

An undirected weighted bipartite network was constructed to understand the relationship between SFBS usage across human RMGFs in order to uncover any potential usage bias (Figure 4.5). The figure highlights 1) the degree distribution across SFBS through the number of edges drawn between each SFBS node and the RMGF node in which they are found present and 2) the number of each SFBS present across each individual RMGF's fusion breakpoint, this is highlighted by the width of the edge between the RMGF and the SFBS and may be indicative of how strongly it contributes to the expression profile of the RMGF. The SFBS with the highest degree is SRp20 with 100% of RMGFs containing at least one binding site (motif: wcwwc). This SF also contained the greatest number of binding sites across RMGFs with each of the 37 RMGFs examined containing on average 42 SRp20 SFBSs across it's fusion breakpoint. Interestingly, the SRp20 cuckucy motif is only present in 2 RMGFs. The SFBS with the second largest degree distribution was MBNL; with 31/37 (84%) RMGFs containing the BS and on average each of the RMGFs contains 13 binding sites per gene. The hnRNPA1 RMGF also contains BSs across 31 of the RMGFs exampled and on average each RMGF contains 18 hnRNPA1 binding sites. Both SRp55 and Tra2beta (motif: ghvvgaur) contain BSs in 29 RMGFs (78%) and both contain an average of 11 and 17 BSs in each RMGF examined respectively. Interestingly, the Tra2Beta gaagaggagg and gaagaa motif are found only in 1 and 5 genes respectively.

CUG-BP has binding sites in 18 RMGFs (49%) and contains 8 BS on average across each RMGF. The SC35 SF has 2 binding motifs (motif: gryymcyr and motif: ugcgyy) and these contain BSs in 9 and 17 RMGFs respectively. For the ugcgyy motif 12/17 BS are present on the same RMGFs as the gryymcyr motif. On average the SC35 motifs contain 2 and 6 BSs across each RMGF respectively. The hnRNPF gene has 3 BS motifs (motif: gukgykg, motif: gggug, and gugkau) and they contain BSs in 9 (24%), 2 (5%) and 0 RMGFs respectively. The gukgykg motif on average has 3 BS per RMGF and the gukgykg has 0.5 BS present on average. The PTB SF has 2 motifs (motif: cucucu and motif:ucuu) and they are present in 2 (5%) and 6 (16%) of RMGFs

respectively with on average 0.2 and 1 BSs being present on each RMGF and both of the motifs are never found on the same RMGF.

SRp40 had a degree of 6 edges/RMGFs (16%) (2 BS per RMGF on average), hnRNPH1 had 13 RMGFs present (35%) (8 BS per RMGF on average) and Tra2alpha had 1 SFBS (0.01 BS per RMGF on average).

All other SFBS under examination had a degree of zero and therefore no edges were drawn between their nodes and the RMGF nodes in Figure 4.5, these include FOX1, hnRNPM, hnRNPU, hnRNPA2B1, QK1, and YB1 SFs. Interestingly, specific binding sites of other SFs are devoid across RMGFs such as the aguguu motif for Tra2beta the gugkau motif for hnRNPF

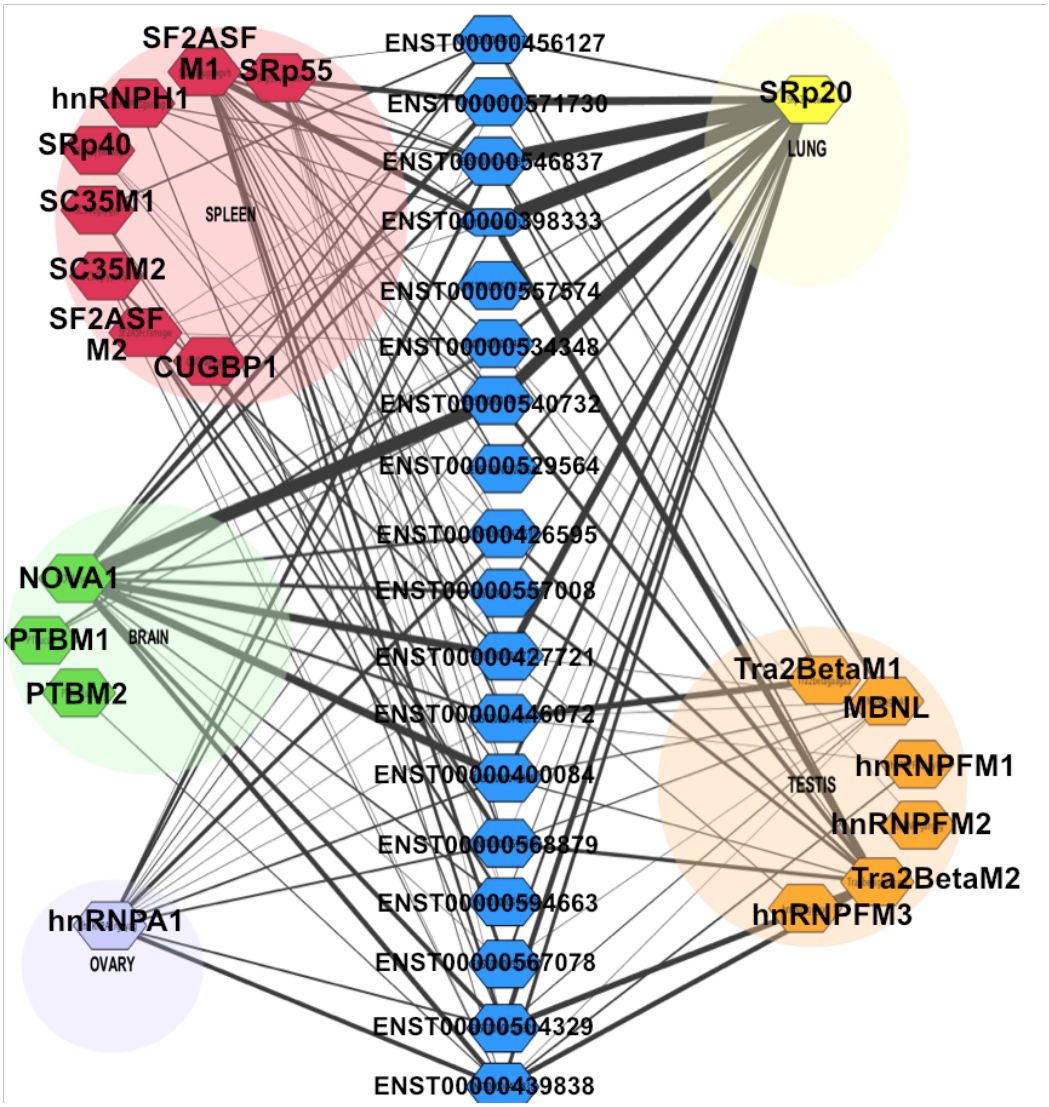
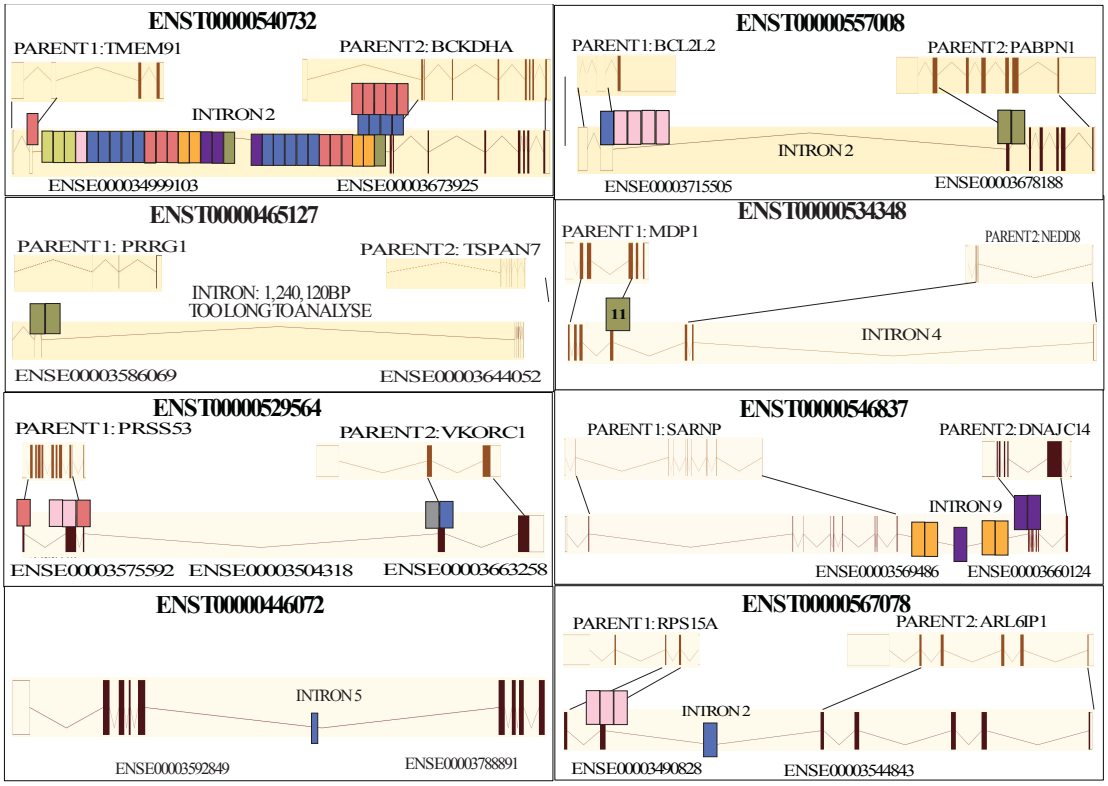
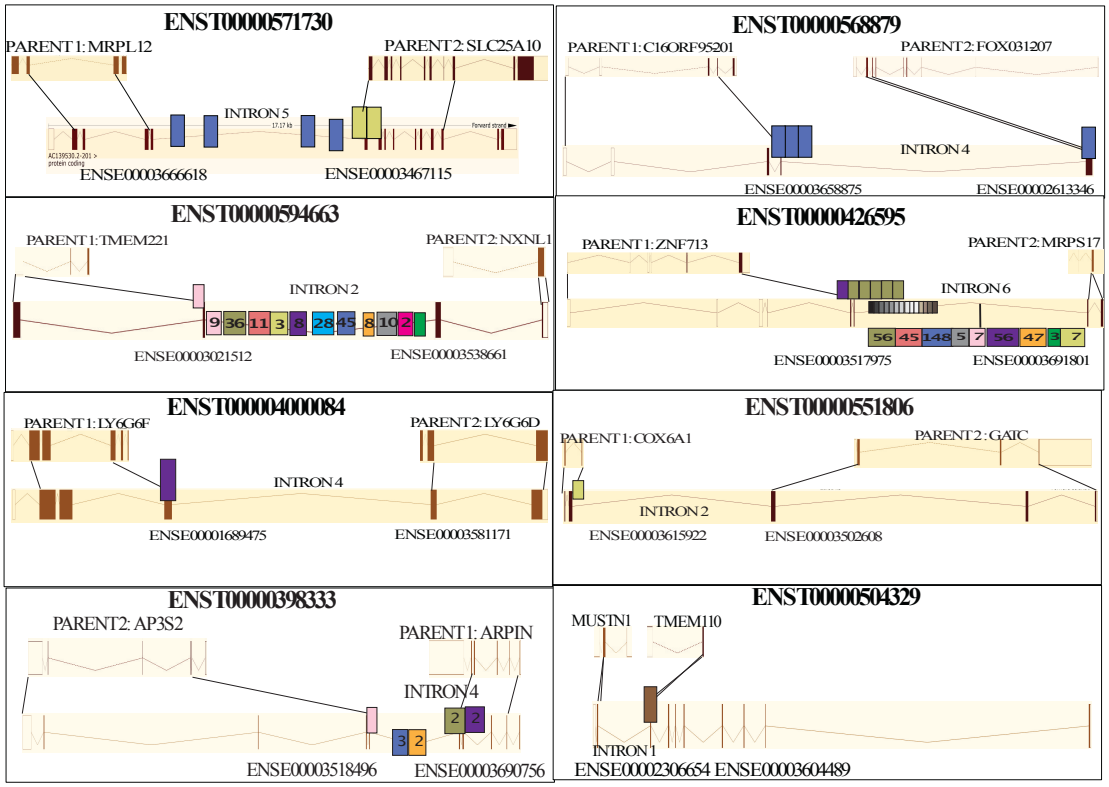


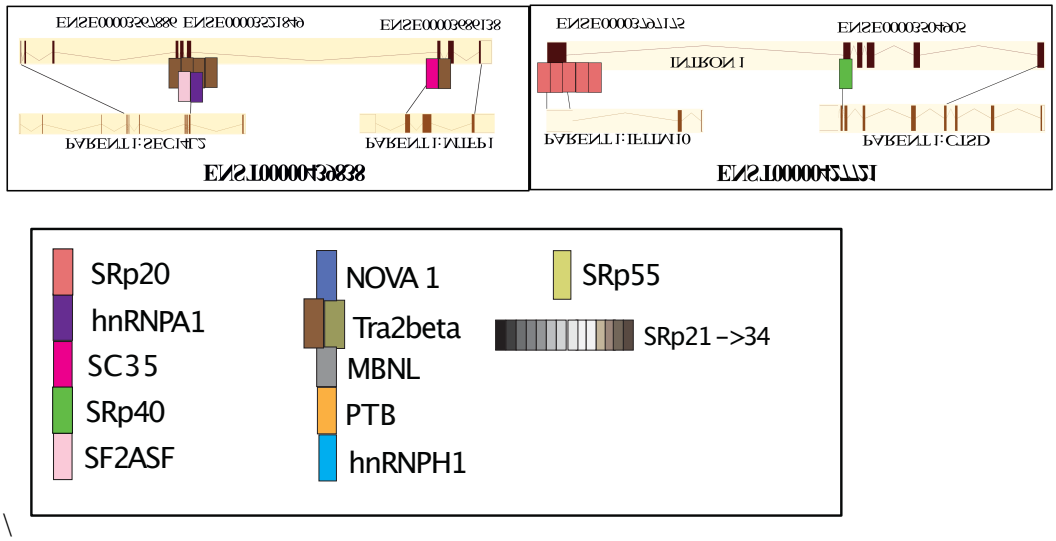
Figure 4.5: *A weighted bipartite network constructed to analyse the relationship between the number of SFBSs and tissue expression profiles associated with splice factors across 18 RMGFs: Bipartite network constructed to highlight the relationship between RMGFs and the SFBSs. If an SFBS is found to be present an edge is drawn between the RMGF node and the SFBS node. A lack of an edge between a SFBS node and a RMGF node suggests no SFBS present. Each group of SFBSs were grouped into categories based on their corresponding SF genes expression profile. For instance SFBSs located in the green circle have a corresponding SF that has its highest level of expression found in brain tissues, orange depicts highest expression in testes, lilac for ovary, red for spleen and lung. Each edge is weighted as an indication of SFBS frequency levels, the thicker the edge the more SFBS is present across the gene.*

4.3.3) An investigation into the impact of SF presence on RMGF parent transcription profiles

The presence of SFBSs across RMGF fusion breakpoints may play a critical role in regulating tissue specific gene expression profiles across RMGFs. SFBSs located here could potentially facilitate expression differences between RMGFs and their parent genes. Overall, instances were identified whereby a SFBS profile fell in line with at least one of the parent gene's expression profile, therefore it is expected that the RMGF co-opted parent gene pre-existing regulatory machinery for example ENST0000056078 (Figure 4.6), ENST00000426595 (Figure 4.6) ENST00000427721 (Figure 4.6), and ENST00000540732 (Figure 4.6). However, cases were also identified where the SFBS profile identified across RMGFs was different to that of the RMGF parents expression profile which indicates that novel splice sites have contributed to a novel expression pattern for the RMGF in comparison to it's parent gene for instance ENST00000557008 (Figure 4.6), ENST00000546837 (Figure 4.6), and ENST00000504329 (Figure 4.6). Analysis of the RMGF transcript ENST000000398333 revealed the presence of NOVA1, PTB, Tra2Beta and hnRNPA1 SFBSs (Figure 4.6). According to the Expression Atlas Database (Papatheodorou *et al.*, 2018) both PTB and NOVA1 SFs are predominantly expressed in brain tissue, Tra2Beta in testis and hnRNPA1 in spleen, ovary and testes. The driving of these SFs in these tissues suggests that ENST000000398333 may also have elevated expression levels across these tissues. The ENST000000427721 contains the SRp20 and SRp40 SFBS that have clear preferential expression in lung and spleen tissues respectively (Papatheodorou *et al.*, 2018). This falls in line with that of its parent gene, CTSD's expression profile (Figure 4.7) that also has a predominant expression signature in the lung. The panel of SFBS found spanning the ENST000000439838 transcript (SF2ASF, hnRNPA1, SC35, and Tra2Beta) have corresponding SFs showing elevated expression in spleen and testes (Papatheodorou *et al.*, 2018) in contrast to the RMGFs parent genes SEC14L2 that shows elevated liver and small intestine expression and MTFP1 that has a very low ubiquitious expression profile. This indicated that the ENSGT000000439838 transcript potentially contains an expression profile different to that of its parent genes Figure 4.7(a).


ENST00000540732
PARENT 1: TMEM91
PARENT 2: BCKDHA
INTRON 2
ENSE00003499103
ENSE00003673925
ENST00000557008
PARENT 1: BCL2L2
PARENT 2: PABPN1
INTRON 2
ENSE00003715505
ENSE00003678188
ENST00000465127
PARENT 1: PRRG1
PARENT 2: TSPAN7
INTRON: 1,240,120BP
TOO LONG TO ANALYSE
ENSE00003586069
ENSE00003644052
ENST00000534348
PARENT 1: MDP1
PARENT 2: NEDD8
11
INTRON 4
ENST00000529564
PARENT 1: PRSS53
PARENT 2: VKORC1
ENSE00003575592
ENSE00003504318
ENSE00003663258
ENST00000546837
PARENT 1: SARNP
PARENT 2: DNAJC14
INTRON 9
ENSE00003569486
ENSE00003660124
ENST00000446072
INTRON 5
ENSE00003592849
ENSE00003788891
ENST00000567078
PARENT 1: RPS15A
PARENT 2: ARL6IP1
INTRON 2
ENSE00003490828
ENSE00003544843


ENST00000571730
PARENT 1: MRPL12
PARENT 2: SLC25A10
INTRON 5
ENSE00003666618
ENSE00003467115
ENST00000568879
PARENT 1: C16ORF95201
PARENT 2: FOX031207
INTRON 4
ENSE00003658875
ENSE00002613346
ENST00000594663
PARENT 1: TMEM221
PARENT 2: NXNL1
INTRON 2
9 36 11 3 8 28 45 8 10 2
ENSE00003021512
ENSE00003538661
ENST00000426595
PARENT 1: ZNF713
PARENT 2: MRPS17
INTRON 6
56 45 148 5 7 56 47 3 7
ENSE00003517975
ENSE00003691801
ENST00000400084
PARENT 1: LY6G6F
PARENT 2: LY6G6D
INTRON 4
ENSE00001689475
ENSE00003581171
ENST00000551806
PARENT 1: COX6A1
PARENT 2: GAIC
INTRON 2
ENSE00003615922
ENSE00003502608
ENST00000398333
PARENT 2: AP3S2
PARENT 1: ARPIN
INTRON 4
2 2
3 2
ENSE00003518496
ENSE00003690756
ENST00000504329
MUSTN1
TMEM110
INTRON 1
ENSE00002306654 ENSE00003604489



\

Figure 4.6: *RMGF breakpoint analysis of SFBS results and location within their corresponding parents genes: Panel of RMGFs assessed for SFBS across fusion breakpoint and their region of homology with their corresponding parent genes with the final panel providing a legend in which to identify each SF.*

The ENST00000446072 RMGF transcript contains only NOVA1 SFBSs across its fusion breakpoint suggesting brain specific signatures of expression however no parent information was available for comparison here (Figure 4.6). For RMGF ENST00000465127, Tra2Beta BS were found suggesting preferential testes expression (Figure 4.6) whilst its parent genes TSPAN7 contains a brain specific expression profile and PRRG1 is ubiquitously expressed (Figure 4.7(a)), suggesting a broadening of expression in the RMGF in comparison to its parent gene**.** The RMGF ENST00000504329 again contains Tra2Beta BS (Figure 4.6) but it's parents have both expression profiles specifically in the ovaries (Arpin) and ubiquitiously (Asps2) this finding supports the "out of testes hypothesis" of new genes in comparison to more ancient parent genes (Figure 4.7 (a). The RMGF ENST00000529564 contains 4 SFBSs all with different tissue specificities (Figure 4.6), some with narrow expression profiles whilst others have a more broad expression profile - its parent gene VKORC1 supports this broader panel of expression potentially highlighting the co-option of the transcriptional profiles of parent genes by its corresponding fused gene (Figure 4.7(b)).

Assessments of the ENST00000534348 (Figure 4.6) transcript revealed testes specific SFBSs (Tra2Beta) however it's parent gene NEDD8 contains only lung and adrenal signatures of expression again potentially supporting the 'out of testes' hypothesis of new genes (Figure 4.7 (b)). Interestingly, the ENST00000546837 parent gene DNAJC14 shows testes specific expression but the RMGF itself contains only SFBSs specific for brain and ovary suggesting that the RMGF adapted and gained it's own mechanism of transcriptional recruitment (Figure 4.7 (b)). Contrastingly, both the ENST000000540732 RMGF transcript and its parent**s** support a ubiquitous expression profile (Figure 4.7 (b)). Both of the ENST00000557008 transcript's parents appear to share a broad expression profile (Figure 4.7 (b)) however it's SFBS profile of SFASF, NOVA1 and Tra2beta support spleen, brain and testes expression perhaps again illustrating the potential of new genes to ascertain novel transcription patterns (Figure 4.6)**.** Both SF2ASF (spleen expression) and NOVA1 (brain expression) are found to span the ENST00000567078 RMGF (Figure 4.6), these expression

patterns are supported by both parents, RPS15AA and ARL61P1 (Figure 4.7 (b)). .

It is clear from comparing RMGF SFBS profiles Figure 4.6 against their corresponding parent genes expression profiles (Figure 4.7(a) and Figure 4.7(b)) that SF choice at RMGF fusion breakpoints may impact the expression profile of the RMGF transcript causing expression changes between it and it's parent genes. The most predominant SFBS across the RMGF transcripts examined is the NOVA1 SFBS. This is an interesting result due to its known drive toward a cerebellum and brain-specific transcriptional profile (Jensen *et al.*, 2000). The frequency of this marker across RMGFs therefore suggests that RMGFs in human are targeted towards brain-specific tissues and that remodelled genes of this fashion could contribute to the complexity across more recently diverged species.

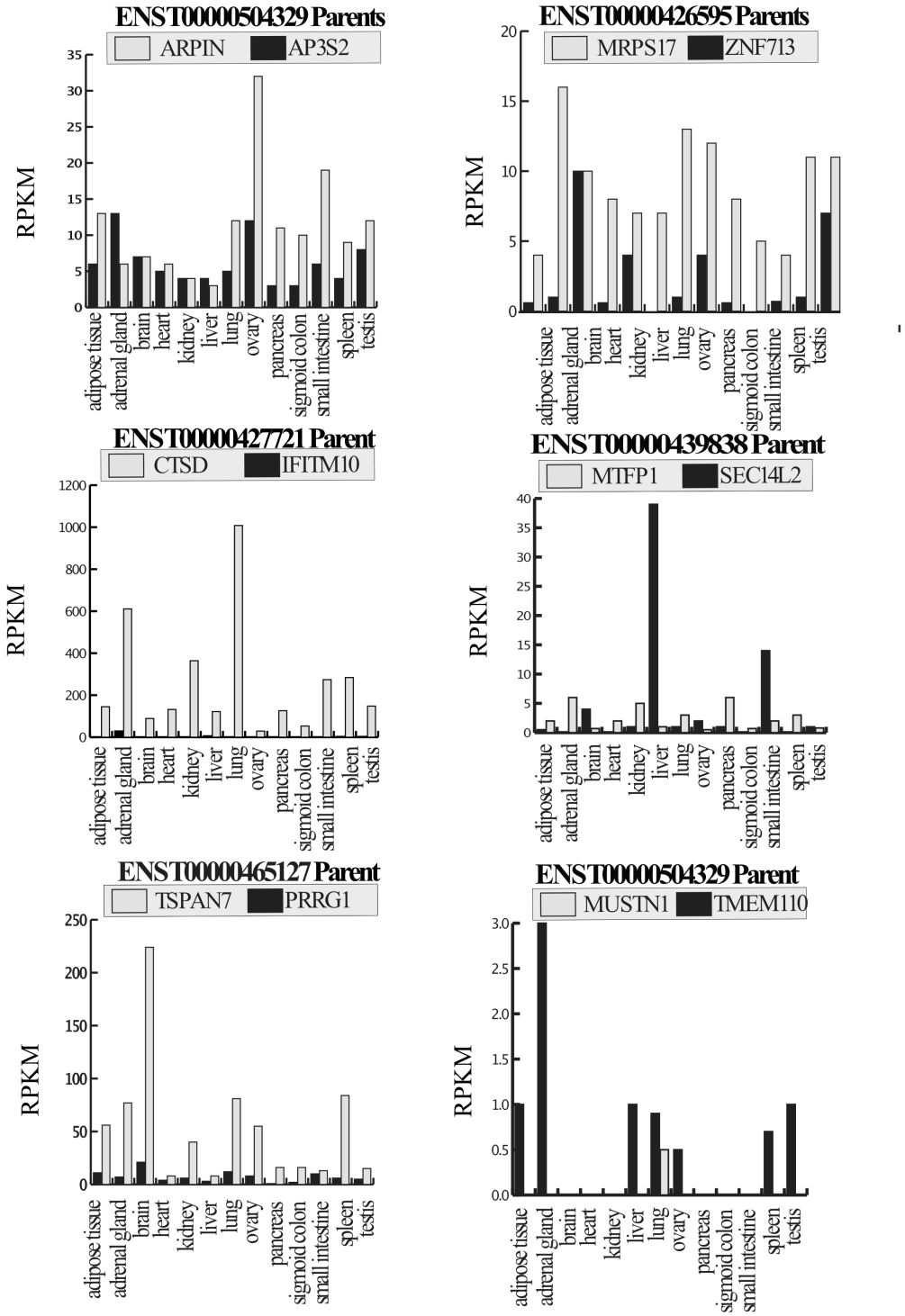


Figure 4.7(a): *Illustration of the gene expression profile for each RMGF parent: An illustration of the gene expression profiles for each RMGF's parents. Expression data obtained by the ExpressionAltas Database* (Papatheodorou *et al.*, 2018).

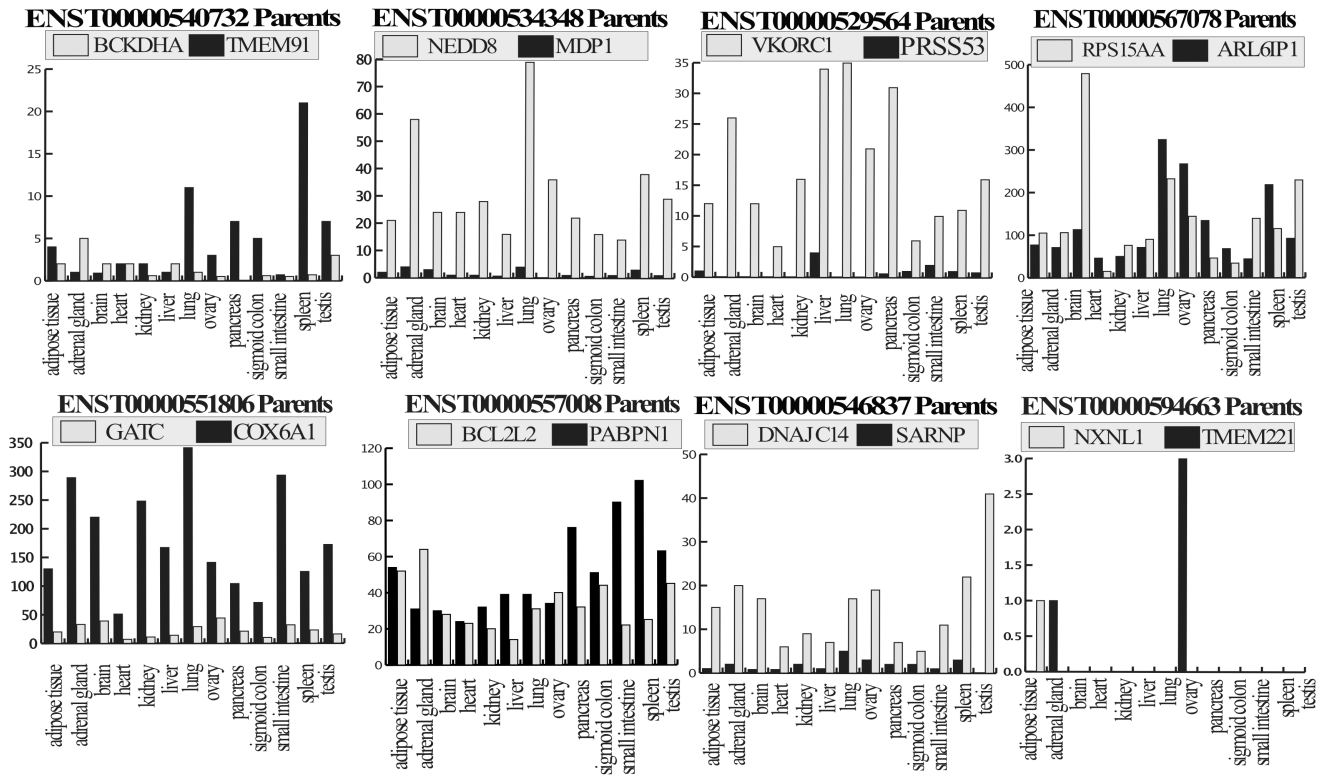


Figure 4.7 (b): *Illustration of the gene expression profile for each RMGF parent: An illustration of the gene expression profiles for each RMGF's parents. Expression data obtained by the Gene Expression Altas Database* (Papatheodorou *et al.*, 2018)**.**

A comparison of the occurrence of binding sites for the 31 SFBSs for 21 SF genes between non-fused human protein coding genes and 37 RMGFs was carried out to determine if SFBS usage differs across these two categories of genes. The number of each SFBS in each RMGF per kB was obtained and compared with the number of SFBSs for non-fused randomly sampled protein-coding datasets (per kB) (see Section 4.2.2.1). To identify the most appropriate statistical test the distribution of each SFBS across human protein coding genes in general needed to be determined. The distributions obtained for the non-fused datasets did not follow a normal distribution across any of the SFBSs analysed (Figure 4.8) with a stochastic distribution being identified. Due to the non-normal distribution statistical comparisons could not be made using standard parametric tests such as a t-test thus a non-parametric Mann Whitney U test was required for accurate comparison**.** Mann Whitney U tests were used to compare frequencies of SFBSs per kilobase between RMGFs and a 100 randomly sampled datasets of 37 human non-fused protein genes and no significant difference was uncovered between the two datasets (Table 4.8) ($p = 0.9408$). However, the analysis did uncover that 39% of SFBSs were not present at all across the RMGFs and 17/31 SFBSs were at an increased rate in RMGFs in comparison to non-fused human protein coding genes (Table 4.8).

A second analysis was carried out in order to determine if the pattern of SFBS co-occurrence in RMGFs is significantly different than that expected in human non-fused protein coding genes. Here, 31 SFBSs found across 37 RMGF were analysed to investigate the likelihood of each SFBS to co-occur on the same RMGF as another SFBS, and subsequently whether the same probability of SFBS co-occurance exists across non-fused protein-coding genes. Significant p-values were obtained for 20 SFBSs (65%) and p-values are shown in Table 4.9. In summary, a bias was not identified across human RMGFs for any specific SFBS however co-occurence results suggest that although their presence is unbiased SFBS co-occurence on RMGFs with other SFBS is strongly biased with 65% of all SFBSs examined pairing with other SFBSs differently in RMGFs than in non –fused human protein coding genes. More simply, SFBSs in RMGFs pair with different SFBSs than non-fused genes.

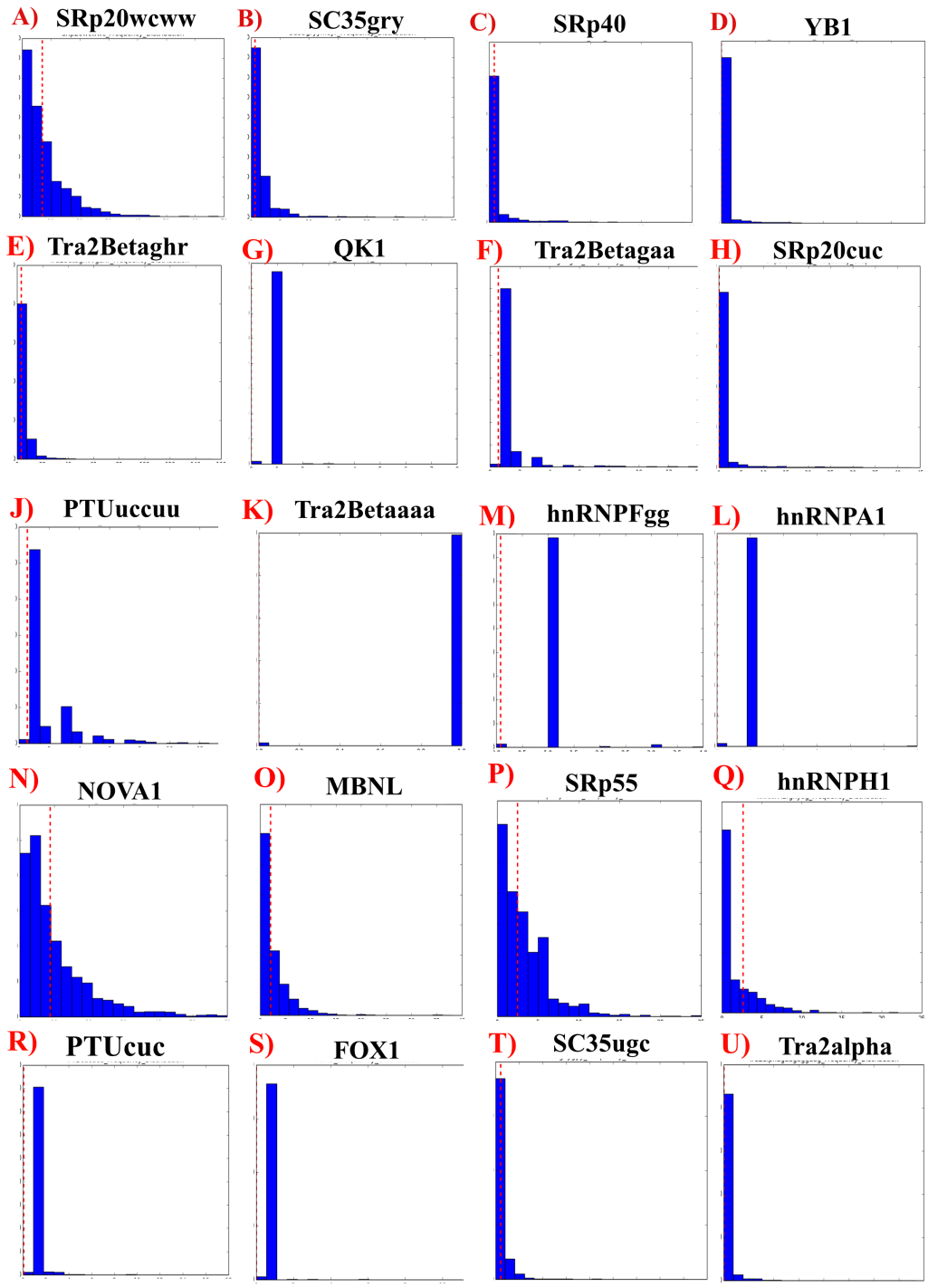


Figure 4.8: *An investigation of the distribution of SFBSs across human non-fused protein coding genes in comparison to human RMGF genes: Distributions of the number of SFBSs found in 100 randomly simulated datasets containing 37 randomly sampled non-fused human protein coding genes per dataset (blue) compared to the number found in 37 human RMGFs identified at 90 PI (red line).*

Table 4.8: The number of SFBSs identified in RMGFs as compared to human non-fused protein coding genes

SFBS	Motif	Average SFBS per Simulation per Kb	Average SFBS per Fusion per KB
9G8	-	0.083243257	0
CUGBP	-	0	0
FOX1	-	0.017567561	0
hnRNPA1	rgnyag	0.994054045	2.67568
hnRNPA2 B1	gguaguag	0.013783771	0
hnRNPF	gggug	0.059459442	0.0810811
hnRNPF	gugkau	0.029999992	0
hnRNPF	gukgykg	0.29054047	0.432432
hnRNPH1	gargag	0.588648714	1.35135
hnRNPM	gguugguu	0.012432422	0
hnRNPU	gauug	0.04702702	0
MBNL	-	0.966486602	2.08108
NOVA1	-	4.254864342	8.83784
PTB	cucucu	0.06027025	0.0540541
PTB	ucuu	0.486756796	0.567568
QK1	-	0.008378374	0
hnRNPA1	guaguagu	0.010270261	0
hnRNPA2 B1	aggwuhgr	0.027297283	0
SC35	gryymcyr	0.413243233	0.702703
SC35	ugcygyy	0.429459349	1.08108
SF2ASF	crsmsgw	0.346486367	0.675676
SF2ASF	ugrwgvh	2.153513372	4.18919
SRp20	cuckucy	0.149189154	0.0810811
SRp20	wcwwc	3.44513495	6.94595
SRp40	yywcwsg	0.196756654	0.513514
SRp55	yrckm	1.279729843	2.51351
Tra2alpha	gaagaggaag	0.045135107	0.162162
Tra2beta	aguguu	0.03486482	0
Tra2beta	gaagaa	0.27270264	0.567568
Tra2beta	ghvvganr	1.629999764	3.40541
YB1	caaccacaa	0.035945961	0
	W = 444.5	**p-value**	**0.9408**

Results of SFBS frequency (per Kb) comparisons between RMGFs and simulated data using non-parametric Mann-Whitney U tests. The R package actually utilises the Wilcoxin test and subsequent W score located in column 2 to determine significance. This is equivalent to the U-statistic obtained by the Mann-Whitney test.

Table 4.9: A comparison of SFBS co-occurrence between human RMGFs and simulated human non-fused protein coding genes

Co-occurance results	W Statistic	P-value	Adjusted P-value
hnRNPFgggug	W = 0	0.00000011	3.428571e-07
NOVA1	W = 140	0.2614	2.751579e-01
hnRNPFgugkau	W = 0	0.00000002	1.000000e-07
hnRNPFgukgykg	W = 0	0.00000002	1.000000e-07
hnRNPH1gargag	W = 0	0.00000012	3.428571e-07
PTBcucucu	W = 378	0.00007263	1.210500e-04
PTBucuu	W = 56	0.0001865	2.869231e-04
QK1acuaa	W = 10.5	0.00000001	1.000000e-07
SRp20wcwwc	W = 159	0.1223	1.358889e-01
SF2ASFugrwgvh	W = 96	0.004385	5.846667e-03
SRp20cuckucy	W = 26	0.00000123	3.075000e-06
SRp40yywcwsg	W = 116.5	0.007965	9.956250e-03
SRp55yrckm	W = 157	0.493	4.930000e-01
SC35gryymcyr	W = 25	0.001654	2.362857e-03
Tra2alphagaagaggaag	W = 37	0.00001825	3.318182e-05
Tra2betagaagaa	W = 21	0.00000012	3.428571e-07
Tra2betaghvvganr	W = 81.5	0.02771	3.260000e-02
YB1caaccacaa	W = 0	0.00000001	1.000000e-07
FOX1	W = 0	0.00001027	2.054000e-05
MBNL	W = 0	0.00000169	3.755556e-06

A comparison of the co-occurrence of 21 SFBSs between human RMGFs and human non-fused protein codon genes using parametric Mann Whitney U test. The R package utilises the Wilcoxin rank test to generate a W statistic (column2), this W statistic is then used to generate p-values (Column 3) that were corrected for multiple testing using the Benjamini Hochberg statistic (Column 4).

4.3.4) An exploration of histone binding sites across RMGFs

4.3.4.1) An investigation of histone marker frequencies in RMGFs across a panel of human tissues

To further characterise RMGFs and understand their mechanism of transcriptional regulation an investigation of 4 transcriptional activation histone markers (h3k4me3, h3k36me3, h3k4me1, and h3k9ac) and the h3k27me3 transcriptional repression marker was carried out across a panel of 18 RMGFs. Only 18 genes could be analysed due to the limited data available for analysis in the RoadMap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015).

The analysis of h3k27me3 marker frequency across RMGFs showed an elevation of this marker in RMGFs particularly within embryonic tissues (4776 markers present) (Figure 4.9). Embryonic tissues contain a four fold increase in BS frequency when compared to liver tissues (1154 markers), a 7 fold increase when compared with brain (675 markers) and an 8 fold increase in comparison to the number of BS present in heart tissues (582) (Figure 4.9). The h3k27me3 marker is associated with heterochromatin formation and transcriptional repression across tissues however, there is evidence in the literature that supports an association of the h3k27me3 marker with transcriptional activation under certain circumstances (Akkers *et al.*, 2009).

H3k36me3 markers too are present at an increased number in embryonic tissues (5435 markers) when compared to brain a 4 fold decrease in markers was observed (1403 markers), with ovary (1033 markers) a 5 fold decrease was shown and when compared to lung a 6 fold decrease in marker frequency was found (915 markers) (Figure 4.9). This marker is associated with euchromatin with high levels of this marker associated with transcriptional activation (Dong and Weng, 2013).

h3k4me1 histone markers again are found at the highest level in kidney tissues (18171 markers) however levels are also found across embryonic tissues at a 7

fold decreased rate (2436 markers), in brain at a 13 fold decreased frequency (1388markers) and in liver and ovaries at a 17 fold decelerated frequency level (1048 and 1035 markers respectively) (Figure 4.9).

Another transcription activating histone marker (Dong and Weng, 2013), h3k4me3 has elevated levels of binding sites specifically in the embryonic tissues (2858 markers) but also found at high levels across all other tissue examined including brain (1250), lung (1016 markers), kidney 1103 (markers), liver (860 markers) and pancreas (749 markers). The h3k4me3 marker is a signal for transcription activation and its presence at such high levels across all tissues examined suggests this histone plays a key role in the activation of RMGFs ubiquitously (Dong and Weng, 2013) (Figure 4.9). The h3k9ac marker is present at elevated levels in embryonic tissues (5403 markers) but at lower levels in brain (1296 markers), kidney (1124 markers) and liver (755 markers).

Across embryonic and brain tissues the h3k36me3 activational marker is the most abundant and across liver, placenta, pancreas, kidney and heart tissues the h3k4me1 marker. In ovary tissues both the h3k36me3 and h3k4me1 markers are present at equal levels and in lung the h3k4me3 marker occurs most frequently.

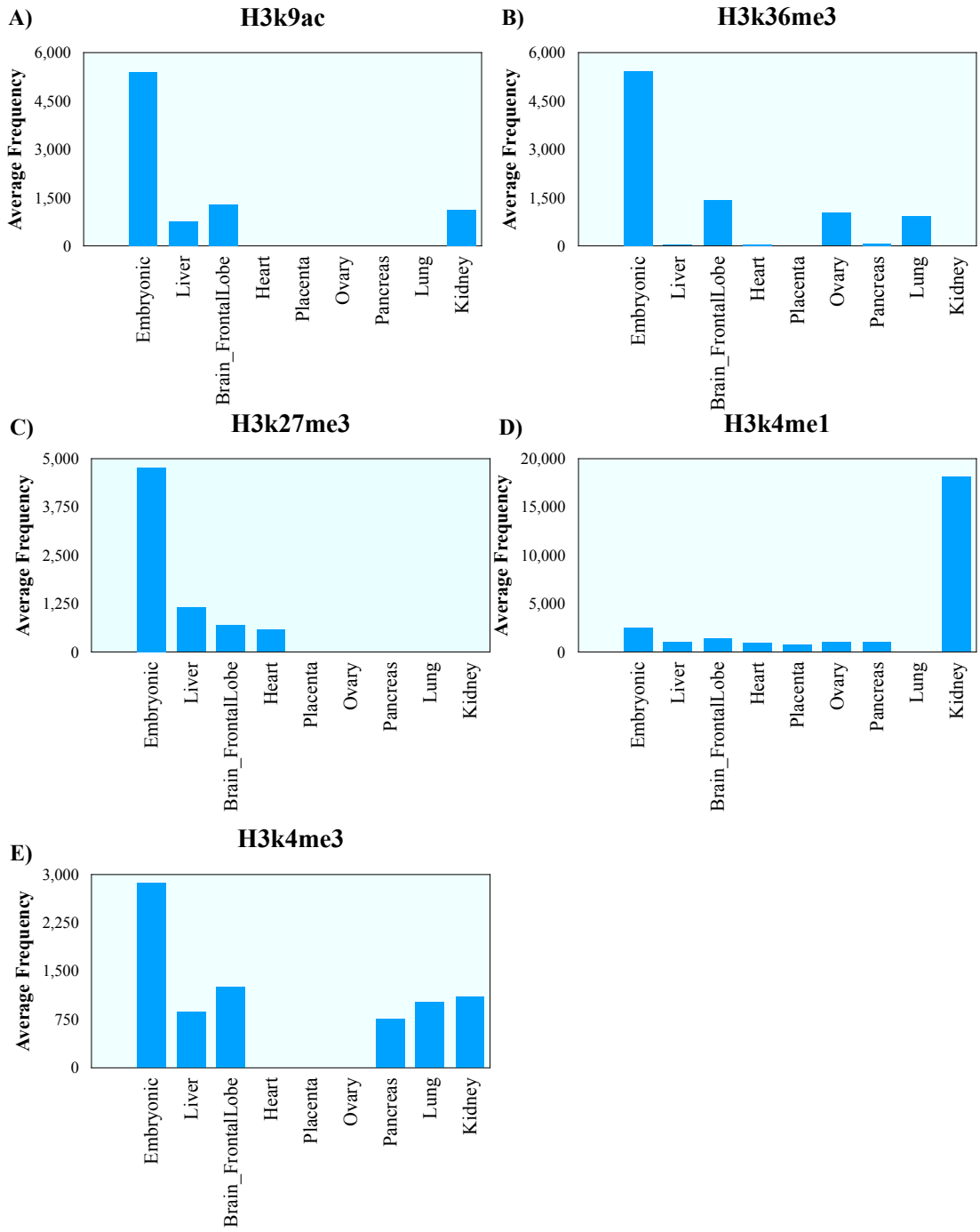


Figure 4.9: *An illustration of the average number of histone binding sites in RMGFs across a panel of 9 human tissues: Histone frequency analysis results depicting the average frequency (Y axis) of each histone binding site tested across identified human RMGFs at a 90% PI threshold in a panel of human tissues (X axis).*

The number of activating histone binding sites predicted across individual RMGF transcripts varies between 53,081 markers in the ENST00000465127 RMGF on average across the 5 markers examined and 647 in the ENST00000534348 RMGF transcript on average (Figure 4.10) this indicates an 82 fold difference of histone marker abundance across RMGFs. Within the ENST00000465127 RMGF two activating histone markers were found at an increased level when compared to the average abundance of total histone markers, the h3k4me3 (1.2 fold increase) and h3k4me1 (1.2 fold increase) marker (Figure 4.10). The second highest activating histone marker abundance was found in the ENST0000038056 RMGF transcript containing 12689 histone markers on average and again both h3k4me3 (1 fold increase) and h3k4me1 (1.2 fold increase) markers were found at elevated levels when compared to the average abundance. The third RMGF with the highest activating histone marker presence was the ENST0000039833 RMGF transcript with 11526 markers present on average and 4 markers were present at levels above this average, namely h3k4me1 (1 fold increase), h3k4me3 (1 fold increase), and h3k36me3 (1.4 fold increase) (Figure 4.10). The average of the activating histone marker presence was compared to the h3k27me3 repressive histone marker and although it is clear that ENST00000465127 is highly abundant in markers for all 5 histone markers examined, it contains a greater abundance of transcription activating histone markers and therefore is likely to be expressed. Similarly, the ENST000000380569 gene contains a low level (2 fold decrease) of the h3k27me3 repressive marker and contains an increased level of all activating histone markers, h3k3me3, h3k4me1, h3k4me1 and h3k9ac. This RMGF transcript is more likely to be located in transcriptionally active, euchromatic regions of the genome across the human tissue panels investigated. The overall trend is that RMGF transcripts are more likely to be in regions of transcriptional activity across human tissues with 12/18 (66%) of RMGFs showing elevated levels of histone marker that promote transcriptional activation (Figure 4.10). Only 5/18 (27%) genes examined are more abundant in histone markers that repress transcriptional activity and 1/18 RMGF examined had insufficient data to carry out the analysis.

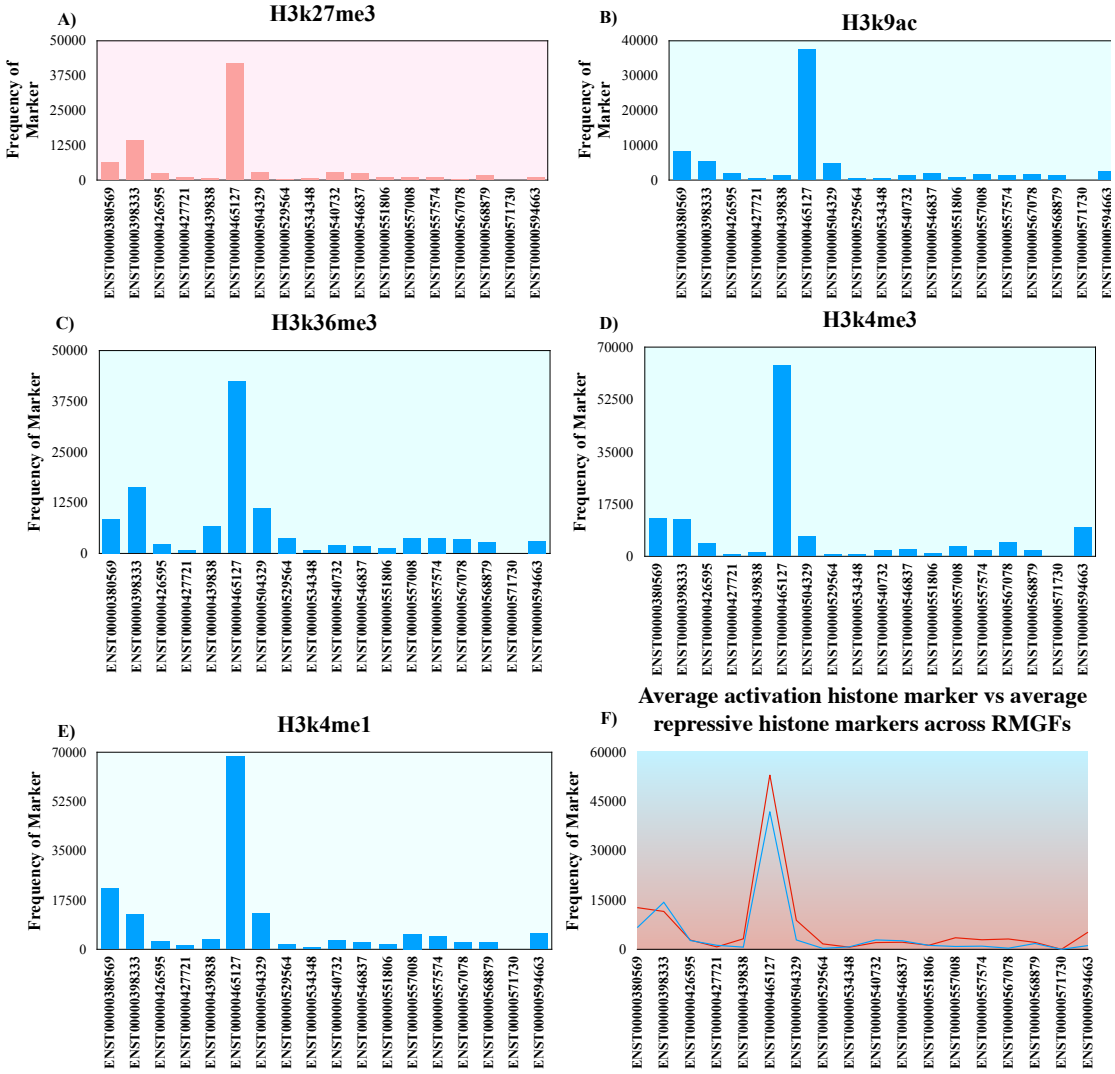


Figure 4.10: *Histone marker frequency across RMGF transcripts across a panel of human tissues: Graphs with a blue background colour depict histone marker analysis of transcription activating histone markers whereas a red background highlights repressive histone markers. (a) Illustrates h3k27me3 histone marker frequencies across RMGFS (b)* depicts h3k9ac activational markers, (c) shows h3k36me3 histone marker abundance, (d) represents h3k4me3 marker numbers, (e) shows h3k4me1 marker abundance and *(f) illustrates a comparison between an average of the activational markers identified across RMGFs (red line) in comparison to the repressive histone marker (blue line).*

4.3.4.2) Linear regression analyses identify correlation between histone marker and specific splice binding site usage in RMGFs.

In order to understand whether particular splice sites co-occur within particular transcriptional states (active or inactive) a linear regression analysis was carried out for 18 RMGFs to identify whether a dataset of 31 SFBS correlate with any of the 5 histone markers. The results of the linear regression analyses are represented per tissue; frontal lobe, liver, kidney**,** pancreas and heart (Appendix_C: Table 2) and a summary of significant results is represented in Figure 4.11.

Correlations are seen between 4 SFBSs and all specific histone marker binding sites for the 18 RMGFs examined. Elevated levels of binding sites for the h3k27me3 marker were found in heart ($p = 0.04566$, $p = 0.06959$, $p = 0.005358$), brain ($p = 0.03403$, $p = 0.0488$, $p = 0.003598$), and liver ($p = 0.03761$, $p = 0.02842$, $p = 0.003675$), corresponding to an increase in CUG-BP, MBNL, and SC35 SFBSs respectively.

The h3k4me3 marker in brain ($p = 0.04715$, $p = 0.03749$, $p = 0.004641$), liver ($p = 0.05181$, $p = 0.03004$, $p = 0.005362$) and pancreas ($p = 0.08767$, $p = 0.07207$, $p = 0.01001$) tissues, as well as the h3k9ac marker in brain ($p = 0.05285$, $p = 0.03457$, $p = 0.004641$), embryonic stem cells ($p = 0.05689$, $p = 0.04719$, $p = 0.005994$), and liver ($p = 0.04733$, $p = 0.02629$, $p = 0.004688$) tissues were both found to be correlated with increased frequencies of CUG-BP, MBNL, and SC35 SFBSs respectively.

Increased h3k4me1 markers were found to correlate with an increased number of binding sites for CUG-BP, MBNL, and SC35 in pancreas ($p = 0.0982$, $p = 0.04426$, $p = 0.01172$), CUG-BP, MBNL, SC35, and SF2ASF in liver ($p = 0.08716$, $p = 0.02316$, $p = 0.01082$, $p = 0.01082$), MBNL, SC35, and SF2ASF in heart ($p = 0.08961$, $p = 0.04387$, $p = 0.04387$) and CUG-BP, MBNL, and SC35 binding sites in embryonic stem cells ($p = 0.0464$, $p = 0.05123$, $p = 0.005216$) tissues. If h3k36me3 is elevated in brain and heart CUG-BP, MBNL and SC35 also increase, however in pancreas only MBNL and SC35 increase and in

embryonic stem cells only SC35 and SF2ASF will increase. From Figure 4.11 it is clear that only 4/19 SFBS show elevations significantly correlated with the 5 histone markers tested in RMGFs namely, CUG-BP, MBNL, and SC35 binding sites in RMGFs.

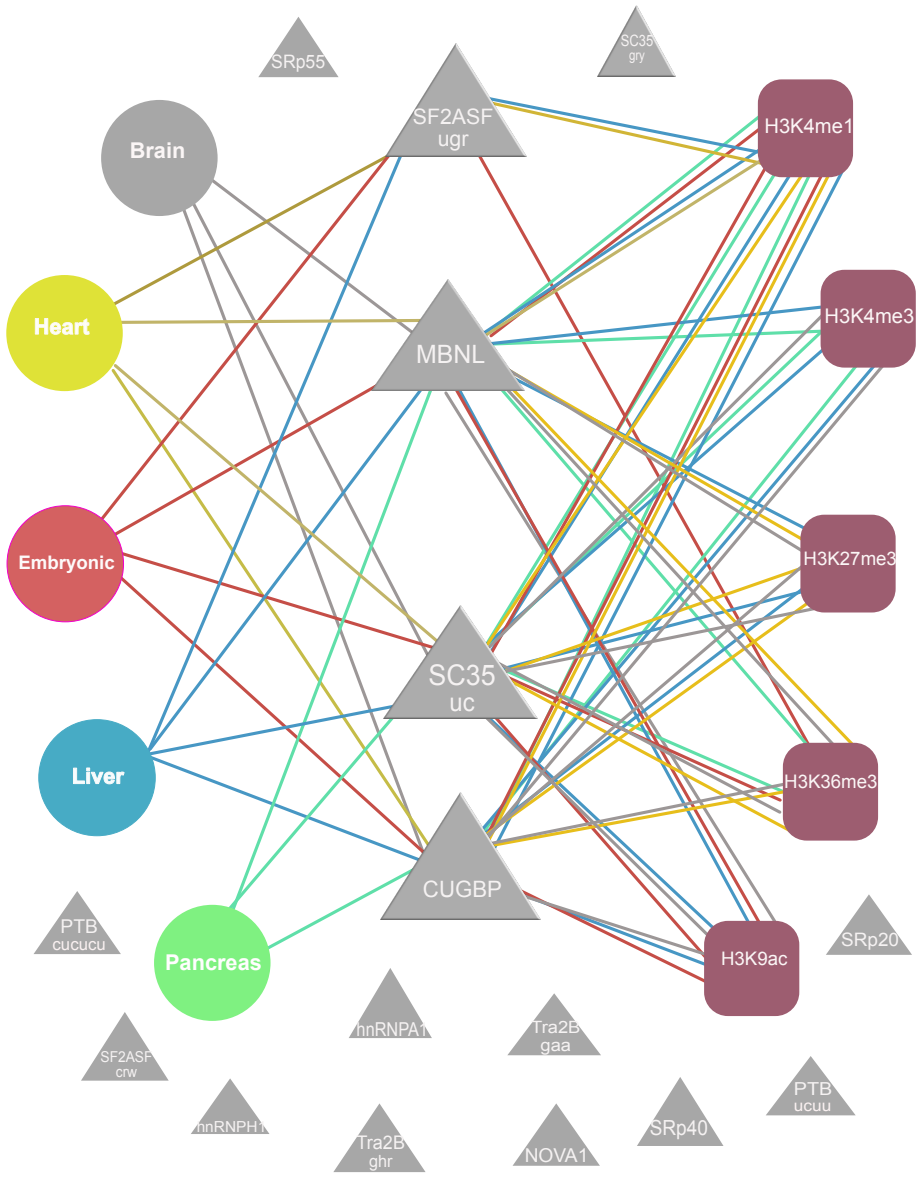


Figure 4.11: *A network constructed based on statistically significant correlations between histone marker and splice site usage in a panel of human tissues across RMGFs: An undirected network graphically displaying statistically significant correlations between splice factor and histone marker binding site frequency across a panel of human tissues. The 5 human tissues examined (heart, liver, pancreas, embryonic and brain) are represented by circles, the 19 SFBSs are represented by triangles based on their presence across RMGF fusion breakpoints, and rounded rectangles depict the 5 histone markers assessed. A line or edge is drawn between nodes if a statistically significant p-value result was obtained from the linear regression.*

The Human Protein Atlas's HPA, GTEX and Fantom5 datasets were used to determine the RNA expression profiles of the 4 splice factors (CUG-BP, MBNL, SF2ASF and SC35) across 36 human tissues (Uhlén *et al.*, 2015) (Table 4.10). From these data it is clear that these splice factors are more commonly expressed in brain tissues in comparison to any other tissue examined, particularly for the splice factor binding sites SF2ASF and CUG-BP. However, an investigation of MBNL2 has subsequently shown a dominant brain-specific expression pattern (Konieczny, Stepniak-Konieczna and Sobczak, 2014).

Table 4.10: An RNA expression profile analysis of SFBSs correlated with histone markers using Human Protein Atlas Data (Uhlén *et al.*, 2015)

	Human Tissue	HPA per 36 tissues	GTEx per 36 tissues	FANTOM5 per 36 tissues
SF2ASF	Heart	31	32	36
	Kidney	20	25	29
	Pancreas	23	30	36
	Liver	30	34	28
	Cerebellum	18	10	2
SC35	Heart	33	32	29
	Kidney	31	24	19
	Pancreas	36	29	27
	Liver	31	29	28
	Cerebellum	25	16	20
MBNL	Heart	28	25	24
	Kidney	29	25	31
	Pancreas	36	35	36
	Liver	33	30	33
	Cerebellum	28	33	31
CUG-BP	Heart	33	36	30
	Kidney	20	31	21
	Pancreas	29	33	36
	Liver	22	25	16
	Cerebellum	10	1	1

RNA expression information from the Human Protein Atlas Database (Uhlén *et al.*, 2015) *for the 4 splice factors (SF2ASF, SC35, MBNL, and CUG-BP) that had statistically significant correlations with 5 histone marker binding sites analysed in RMGFs. Analyses were carried out over 3 datasets (HPA, GTEx and FANTOM5) and a panel of 36 human tissues. Numbers indicate out of a total of 36 tissues analysed how highly the splice factor's expression is usually ranked, with 36 being the lowest (red cells) and 1 being the highest (orange cells).*

4.3.5) An assessment of TFBS usage and frequency comparison across RMGFs

Transcription factors binding sites (TFBSs) are another *cis*-regulatory element that can profoundly affect the spatio-temporal transcriptional output of a given cell by either enhancing or repressing transcript production through transcription factor recruitment. To understand *cis*-regulation at a greater level an analysis of TFBSs was carried out across RMGFs. Firstly, a presence/absence investigation was carried out in a panel of 15 gold standard experimentally validated and curated TFBSs available from the JASPAR database (Khan *et al.*, 2018) in RMGFs (Figure 4.11) (Khan *et al.*, 2018). Results were obtained for 15/37 RMGFs analysed and they uncovered a slight bias towards the EWSR1-FLI1, SP1, and Zfx TFBS with 73% of RMGFs containing a EWSR1-FLI1 binding site and 66% containing both a SP1, and Zfx binding sites

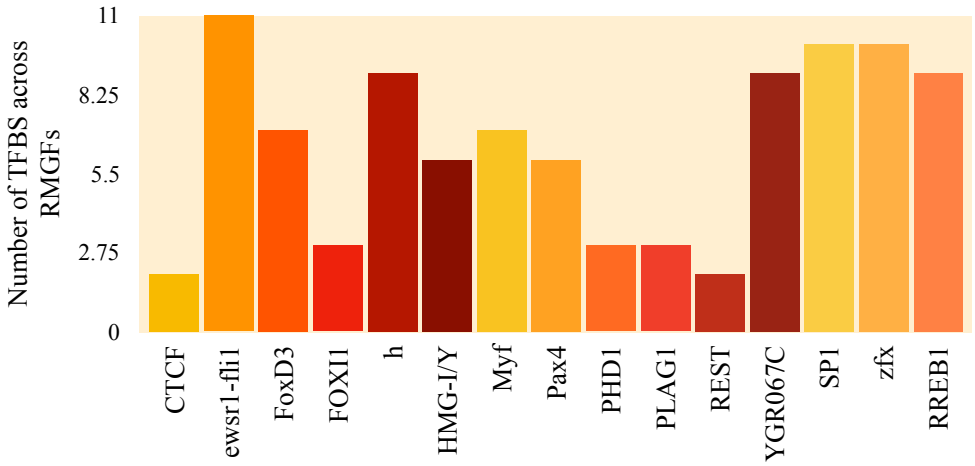


Figure 4.11: *An investigation of 15 gold standard TFBSs and their usage across human RMGFs: A TFBS analysis run using TFBS provided by the JASPER database* (Khan *et al.*, 2018) *on RMGFs identified at a 90 PI threshold and using default settings. X axis illustrates the panel of TFBSs under investigation while the Y axis shows the number of RMGFs that contain each TFBS.*

A gene expression analysis of each 9/15 TFBSs corresponding TF gene was carried out (CTCF, FOXD3, PAX4, PLAG1, FOXI1, REST, Zfx, SP1 an d MYF6) in order to investigate if each TF showed a tissue-specific or more ubiquitous expression profile (Papatheodorou *et al.*, 2018). The CTCF TF indicated a ubiquitous expression profile with highest expression found across brain, ovary and testes tissues. The FoxD3 TF had a brain, testis and colon specific expression pattern (Papatheodorou *et al.*, 2018). PLAG1, REST, SP1 and Zfx all had a ubiquitous profile with highest levels again found in testes and spleen (Papatheodorou *et al.*, 2018). FOXI1 had a kidney-specific expression profile, MYF6 a testes-specific pattern and Pax4 a testis-specific pattern. Interestingly, 7/15 genes contained the FOXD3 TF with highest expression in brain tissues and 6/15 had Pax4 BS indicative of testes-specific expression profiles (Papatheodorou *et al.*, 2018).

4.4) Discussion

The results obtained from Chapter 3 suggest that our identified dataset of human RMGFs have both transcriptional and translational outputs however, it is not yet known how these outputs are being controlled and whether they are being controlled differently to that of human non-fused protein coding genes. In order to enhance our understanding of RMGFs transcriptional activation and repression a characterisation of their chromatin structure and their usage of three *cis*-regulatory elements was carried out; 1) Splice factor binding sites, 2) histone markers and 3) transcription factor binding sites.

Cis-regulatory sequence elements require an uncoiled chromatin structure for accessibility to recruit their corresponding trans-acting factors and thus are highly dependent on signals that promote either euchromatin or heterochromatin formation. Prior to RMGF *cis*-regulatory characterisations an investigation of both activation and repressive chromatin markers was carried out using epigenomic datasets spanning 8 human tissues across RMGFs (Roadmap Epigenomics Consortium *et al.*, 2015) (Figure 4.1). On average activation markers are increased in comparison to repressive markers suggesting a euchromatic structure across RMGFs allowing RNA polymerase entry, TSS

binding and transcription initiation this supports results obtained in Chapter 3. Interestingly, across all 8 tissues examined the ENST00000465127 transcript has an elevated level of both activating and repressing signals in comparison to all other RMGFs in the dataset. Across ovary, embryonic stem cell, small intestine and lung tissues this transcript had increased repressive markers suggesting a lack of transcriptional outputs for this transcript across these tissues. ENST00000465127 awaits further annotation however the literature suggests evidence for a viable transcript, transmembrane localisation and calcium ion binding domains suggesting a role for the transcript in cell-cell communication and cystolic calcium homeostasis (UniProt Consortium, 2018). It is possible by the increase in transcriptional repressive markers in lung, ovary, embryonic stem cells and small intestine that specific signalling cascades are inactive. This increased level of transcriptional repressing markers was an exception and generally RMGFs were found with comparatively higher levels of activators present, suggesting transcriptional activation.

1) A characterisation of splice factor usage across identified RMGFs

Splice factors have the ability to control transcriptional activity by the addition or indeed the removal of sequences resulting in either transcription enhancement or repression (Kornblihtt *et al.*, 2004). The SFmap software package (Paz *et al.*, 2010) was to identify splice factor binding sites (SFBS) across RMGFs and those identified with a >=90 COS(WR) were summed for each SFBS across each RMGF. A conservation of score (weighted rank) (COS (WR)) score of 90 was utilised as the specified threshold value due to its high stringency as less motif mismatches as well as less false positives are found at this threshold. These calculations were then compared to a randomly sampled dataset of simulated non-fused human protein coding genes of equivalent length.

At this threshold the top 3 SFBS present across the RMGFs tested were SRp20, MBNL, and NOVA1. The SRp20 SF has been thoroughly examined in the literature and has been shown to play a vital role in spliceosomal assembly prior to transcription initiation and alternative splicing (Corbo, Orrù and Salvatore, 2013). GO term analyses of the SF also indicate a role in DNA replication, DNA

elongation, and telomere maintenance (Ashburner *et al.*, 2000). Due to the necessity of these functions for cell survival the abundance of SRp20 binding sites is perhaps expected. NOVA1 has been highly associated with brain specific expression across the literature (Ule *et al.*, 2005) and GO term analysis revealed a role for the splice factor in neuron cell-cell adhesion, vocalisation behaviour, alpha-amino-3 hydroxxy-5-methyl-iso propionate selective glutamate receptor and has also been associated with neuronal synaptic plasticity (Ashburner *et al.*, 2000). The abundance of NOVA1 SFs in RMGFs suggests that these genes are involved in brain-specific activities. This is expected across new genes in human tissues as the divergence of humans and Great Apes coincided with an expansion of brain size in humans as well as increased neural connectivity in both the cerebral cortex and cerebellum tissues.

The calculation of SFBS abundance levels highlights SFs role in the control of transcription. It is also important however to understand each SFs gene expression profile as the RMGFs that they contain motifs have an increased likelihood of also being expressed here. Figure 4.3 depicts the expression profile of all 21 SF genes analysed across a panel of 12 human tissues obtained from the Expression Atlas's ENCODE database (Papatheodorou *et al.*, 2018). Analysis revealed SRp20 is predominantly expressed across spleen, lung and testis while NOVA1's expression signature is confined to brain, adrenal gland and testes tissues while MBNL shows signatures of expression across lung and adipose tissue samples. This may indicate potential bias in RMGF expression within these tissues but certainly supports results obtained in Chapter 3 suggesting a broad range of expression for RMGFs

A follow up analysis was carried out specifically across RMGF fusion breakpoint as SFBS identified in this region have a greater probability of effecting RMGF transcription than across the entire gene, results are depicted in Figure 4.4 and show a definite bias for both high levels of NOVA1 and SRp20, supporting trends found across the entire gene. A bipartite network was constructed to illustrate the relationship between SFBSs and their abundance across RMGF breakpoints (Figure 4.5). This network highlights the high degree distribution of

the SRp20 wcwwc motif amongst RMGFs with 42 binding sites on average present across each RMGF. Interestingly, the alternative SRp20 motif, cuckucy is present only in 2 RMGFs examined suggesting motif selection bias across splice factors in RMGFs.

The SC35 SF gene was also assessed for 2 motifs (gryymyer and ugcgyy) and again a motif bias is present with grymyer motifs present in just 9 RMGFs whereas the ugcgyy motif is found across a total of 17 RMGFs. Another point of interest are those SFs that contain no SFBS at all across RMGFs these include; FOX1, hnRNPM, hnRNPU, hnRNPA2B1, QK1, and YB1. Interestingly, hnRNPM, hnRNPU,and hnRNPA2B1 and QK1 (formerly known as hnRNPK) are all members of the heterogeneous nuclear ribonucleoparticle family that are responsible for splicing repression through either antagonising splice site recognition or interfering with the binding of proteins to enhancer elements, namely SR proteins (Busch and Hertel, 2012). This supports our findings that although hnRNP SFBSs are not present across RMGFs, SR proteins such as SRp20 are frequent in number. hnRNPM specifically has been shown to act antagonistically against the NOVA1 SF and as NOVA1 is found at increased levels across RMGFs it is unsurprising that hnRNPM is therefore low in frequency (Park *et al.*, 2011).

This result falls in line with results obtained from Section 4.3.1 where activation markers were found at higher levels across the assessed human tissue panel in comparison to repressive markers. The lack of FOX1 however was perhaps somewhat unexpected as it plays a key role in splicing recognition particularly in brain and muscle tissue (Zhang *et al.*, 2008). The YB1 SF has been identified in the literature as a spliceosomal associated protein therefore its lack of binding site is also unexpected (Wei *et al.*, 2012). Further investigations of SF promiscuity across the human genome will uncover whether the lack of these SFBSs is normal however at present the data required for this analysis is unavailable.

From the literature SFs have been shown to control gene expression and tissue specificity (Grosso *et al.*, 2008). Like most other types of new genes, RMGFs have the potential to either keep their parent genes expression profile or adapt and gain a novel expression profile. In order to investigate this, all SFBSs identified across fusion breakpoints were taken and their corresponding SF genes expression profile was obtained and compared to that of the RMGFs fusion parent. If the SF gene and the RMGF parent gene had similar expression profiles it was expected that the RMGF co-opted the expression profile of one of their two parents. However if the expression profile was different it was thought that the RMGF could have potentially ascertained a novel transcriptional profile. When the abundance of RMGF SFBSs across breakpoints (Figure 4.6) was examined alongside their corresponding parents expression profile (Figure 4.7(a) and Figure 4.7(b)) both of these cases were identified with no bias towards either. Finally, SFBS usage was compared to that of randomly simulated non-fused protein coding gene datasets. Here, through non-parametric Mann Whitney U tests (Table 4.8) no significant difference was identified between RMGFs and human non-fused protein coding genes. This is expected as SFBSs present in RMGFs also exist in their corresponding parent gene and therefore should be consistent unless the parent gene itself is under positive selection gaining an unexpected SF usage that is, by proxy, ascertained by the RMGF transcript.

A second analysis was carried to investigate co-occurrence patterns of SFBSs with RMGFs and these results were then compared to the co-occurrence profiles found across non-fused human protein coding genes (Table 4.9). Here 65% of SFBSs analysed were found to co-occur differently between RMGFs and human non-fused datasets. Therefore SFBS present on the same gene in RMGFs will co-occur in a significantly different pattern in non-fused protein coding genes. This is an interesting result as it suggests that in RMGFs SFBSs tend to be present alongside different SFBSs than would be expected. Unfortunately, although a lot is known about individual SFs and their binding sites little is known about their interactions with other SFs. Future work may identify and understand SF pairings of protein coding genes and indeed the repulsion or aversion of some SFs to co-exist on the same gene.

Although investigating SF binding sites at present is informative it does have certain caveats. The 127 epigenomes available from the RoadMap Epigenomic project are a fantastic resource in which to investigate activation and repression markers, however to date there is only human data available making comparative genomic analyses impossible (Roadmap Epigenomics Consortium *et al.*, 2015). In order to preform the necessary SFBS characterisation experiments different software packages were utilised each of which rejected the analysis of certain RMGFs based on their packages own specific criteria for analysis, thus reducing the amount of genes and power of the overall analysis as well as making solid comparisons across tests challenging. SFBS predictions were carried out using the SFmap software package (Paz *et al.*, 2010), however other software packages exist such as the RBPmap package released in 2014 (Paz *et al.*, 2014). This package also utilises the COS (WR) calculation but has two additions 1) conservation based PSSM filtering and 2) a background model for different genomic regions e.g. splice sites, 3' and 5' UTR sequences, non-coding RNA and intragenic sequences. The background model has enhances accuracy due to its ability to account for the likelihood of each binding site occurring in RNA regulatory regions. This package assessed a much more broad range of SFBSs most of which have yet to be experimentally validated which is a requirement for the SFmap database. Future work may be to compare results across these two packages.

2) Histone marker characterisation experiments across RMGFs

A second set of *cis*-regulatory elements that play a role in the transcriptional regulation of protein coding genes are histone markers. Histone modifications such as acetylation and methylation can have an impact on whether the transcriptional profile of the gene is enhanced or repressed (Dong and Weng, 2013). Here both histone activator markers (h3k9ac, h3k36me3, h3k4me1, and h3k4me3) and repressive markers (h3k27me3) were examined across RMGFs in a panel of 8 human tissues obtained from the RoadMap Epigenomics dataset (Roadmap Epigenomics Consortium *et al.*, 2015). Figure 4.8 highlights that all histone markers analysed are present more frequently across embryonic datasets

in comparison to all other tissues, apart from the H3k4me1 marker where kidney is found to contain the highest number of markers. These findings fall in line with the lower frequency of the default chromatin structure, heterochromatin, across embryonic stem cells and thus may require additional markers to maintain a euchromatin conformation. The elevation of the repressive marker h3k27me3 could also potentially be explained by acting as a repressor of particular developmental gene sets during particular stages of differentiation (Tee and Reinberg, 2014). The average of the 4 activating histone markers was calculated and compared to the single repressive marker examined (Figure 4.9). Overall it was found that across RMGFs there are more activating histone markers than repressive markers, and as demonstrated in Figure 4.9 the RMGF transcript ENT00000456127 contains an overabundance of markers when compared to all other RMGFs. Interestingly, this is the same transcript highlighted in Section 4.3.1 as being over abundant in both activating and repressing markers,

In order to understand whether *cis*-regulatory factors co-operate during transcriptional control a linear regression analysis was carried out to uncover potential existing correlations between SFBS usage and histone markers present across RMGFs. One would expect that the binding sites for activating SFs would co-occur with histone markers that contribute to euchromatin formation i.e. h3k9ac, h3k4me1, h3k4me3, and h3k36me3 and conversely that deactivating SFs would correlate repressive heterochromatin forming histone modifications – h3k27me3. In summary, results suggest that both repressive and activating markers correlate with 3 SFs specifically, namely; SC35, MBNL, and CUG-BP across RMGFs across a broad range of tissues. When assessed for their transcriptional profile across 36 human tissues (Table 4.10) these SFs showed a bias toward expression in the cerebellum, particularly the SC35 SF.

MBNL is highly conserved across the tree of life with fish containing a single copy whereas three are found across mammals (MBNL1, MBNL2, MBNL3) (Konieczny *et al*., 2014). Each copy has its own specific expression profile e.g. MBNL2 is specifically associated with brain tissues (Charizanis *et al.*, 2012). They control key developmental transitions e.g. the transition between foetal and

adult splice forms and control both constitutive and alternative splicing (Konieczny *et al*., 2014). Depending on its binding location it can cause either exon inclusion by outcompeting intron splice silencers (ISSs) or removal by blocking intron splice enhancers (ISEs) (Konieczny, Stepniak-Konieczna and Sobczak, 2014). Ectopic MBNL expression is found in individuals suffering with neuromuscular degeneration as the SF is incorrectly localised at synapses and neuromuscular junctions (Lee *et al.*, 2013). With the SFs ability to adjust transcriptional profiles so dynamically it is perhaps unsurprising that it has been found correlated with both activating and repressing histone markers.

Interestingly, across 100s of targets the MBNL SF is was found to have an antagonistic relationship with CUG-BP (also correlated with histone markers) with MBNL binding causing RNA stabilisation whereas CUG-BP binding causes RNA repression (Wang *et al.*, 2015). The CUG-BP SF itself is conserved across complex eukaryotes including *Xenopus* (*EDEN-BP*), *C. elegans* (*etr1*) and *Drosophila* (*BRUNO*) (Dasgupta and Ladd, 2012). The SFs functional role is location dependent with nuclear localisation resulting in the SF exerting AS control e.g. human TroponinT (*TNNT*) and the insulin gene (*INSR*) (Warf and Berglund, 2007) while a cytoplasmic location results in the SF effecting translation efficiency.by recruitment of deadenylases resulting in transcript degradation (Beisang *et al.*, 2012).

The SC35 SF has been identified across metazoan and is known for its role in spliceosomal assembly (Fu and Maniatis, 1992), RNA splicing commitment (Fu, 1993) and plays a role in thymus, pituitary and T cell development (Xiao *et al.*, 2007). Its ectopic expression results in inappropriate cell development and impacts mammalian organogenesis (Xiao *et al.*, 2007). As analyses revealed a correlation between SFBS and histone markers this demonstrates that *cis*-regulatory elements may influence each other across RMGFs

Although useful, the usage of RoadMap Epigenomics resources has limitations, for instance inconsistent tissue availability across histone markers therefore markers cannot be compared over a consistent set of tissue samples (Roadmap

Epigenomics Consortium *et al.*, 2015). The datasets available are all from different sources, using different specimens, and different equipment that introduces additional unwanted variability between tissue samples. The database also only provides information on one repressive marker making any comparison between activating and repressing histone markers a challenge. An additional challenge of histone marker analyses is their multifunctional role depending on the spatio-temporal nature of a given cell or cell type. Future work could focus on the acquisition of a more complete profile of histone modifications across RMGFs when more histone marker datasets and tissue samples become available.

3) Transcription factor usage across RMGF characterisation

A third cis-regulatory element, transcription factors and their binding sites were investigated for their influence on RMGF transcriptional control. Using the JASPAR software package (Khan *et al.*, 2018) 15 gold standard experimentally validated TFs were assessed for abundance levels across RMGFs (Figure 4.11), three of which were found to be enriched namely; ESWR-FLI1, SP1, and Zfx. The most abundant TF across RMGFs was ESWR-FLI1 a TF generated *via* gene fusion of ESWR (chromosome 22) and the TF gene *FLI1* (chromosome 11) (May *et al.*, 1993). The TF is present across 83% of Ewing sarcoma and when ectopically expressed has been identified in neuro-ectodermal tumours patients and is a more potent transcriptional activator than its parent *FLI1* (May *et al.*, 1993). The SF1 enhancing TF is found to influence chromatin remodelling, cell growth and apoptosis and when incorrectly expressed results in lung and glial tumorgenesis. (Doghman *et al.*, 2013). Zfx binding sites are also abundant across RMGFs and when ectopically expressed are found in glioblastoma cells (Fang *et al.*, 2014). Interestingly, all three TFs when incorrectly expressed cause brain malignancies suggesting they play a role in the control of these genes within brain tissues.

Again, this analysis had its own set of caveats. As the JASPAR software package was utilised the use of only RMGFs with exact chromosomal co-ordinates in the reference human genome utilised by the software package could be assessed.

This reduced numbers of genes suitable for assessment as well as overall power. Also, JASPAR only analyses TFs that have been experimentally validated thus limiting comparisons that could be made and perhaps future work could compare results against software packages that do not require experimental validation for inclusion into their dataset. By comparing in this way it will become evident if by selecting validated TFs we are obtaining a holistic viewpoint of TF transcriptional control or just obtaining a misleading snapshot into their potential role in regulation.

Chapter 5: An application of sequence similarity networks to understanding domain shuffling in vertebrates

5.1) Introduction

Chapters 2-4 addressed various aspects of how pre-existing genetic content contributed to the generation of new genes through RNA-mediated gene fusion and how these new genes are regulated, expressed and translated. The application of sequence similarity networks (SSNs) allowed us to identify fused genes as non-transitive triplets in the networks produced. As sequences can share similarity at the level of domains rather than whole genes, we have applied SSNs to domain level data across 30 vertebrates so we can characterize domain shuffling across the RMGFs in comparison to all other genes. Domain shuffling has been shown to play a major role in vertebrate evolution by increasing diversity and complexity (Kawashima *et al.*, 2009). Domains are typically small units of genetic information with discrete structural folding and a functionally annotated purpose (Kawashima *et al.*, 2009). As domains are therefore far shorter than whole protein coding genes, the homology searching in this chapter was done using the hidden markov model (HMMs) based algorithm, pHMMer (Finn, Clements and Eddy, 2011).

The questions addressed in this chapter are as follows:

1) Are the properties of RMGFs domain usage networks typical of biological networks?
2) Is domain usage in RMGFs different to that of human non-fused genes?
3) How does domain co-occurrence in RMGFs compare to non-fused protein-coding genes in human non-fused genes?
4) Are there promiscuous/reclusive domains and if so what are they and are they present at similar levels in RMGFs and non-fused gene sets?

5.2) Materials and Methods

The approach taken in this chapter is outlined here: pHMMer compared a dataset of predefined functionally annotated domains from the pFam database (Finn *et al.*, 2014) to a dataset of: (1) non-fused protein coding genes across vertebrate species, (2) human RMGFs, and (3) human non-fused protein coding genes of the same length as the identified human RMGF dataset. The results obtained were then used to characterize domain properties across vertebrate protein-

coding genes and this was used as a benchmark to compare to both human non-fused protein coding genes in general and then specifically to RMGFs. In order to compare domains in a genome-wide manner SSNs were generated and the transmission of domains across RMGFs was assessed and compared to both human non-fused protein coding genes as well as vertebrate protein-coding genes. The HMM profiles provided by pHMMer (Finn *et al*., 2011) were then used to construct a global SSN of domain usage across both datasets. This global network then underwent two deconstructions. Firstly, the global network was decomposed into a series of undirected bipartite networks where one set of nodes (genes) was compared to a second set of nodes (domains). An edge was drawn between the two nodes if they shared sequence similarity. Simply, bipartite network construction was used to identify domains within genes across vertebrates. A further deconstruction was carried out based on these bipartite networks whereby all gene nodes were removed leaving only domains, these unipartite graphs were used to identify domains that coexist on the same gene. These decomposition events allowed for the characterization of domain frequency and co-occurrence across vertebrate genomes in general to provide a benchmark against which to compare human RMGF and non-fused human protein-coding gene datasets.

5.2.1) Creation of bipartite and unipartite networks from Pfam-A domains

The profile hidden markov models (HMMs) of 16,712 PFam-A domain families were retrieved from Pfam v31.0 (Finn et al. 2016). Using the HMMscan function of HMMER v3.1b1(Finn et al. 2011) and an e^{-20} we identified sequence homology between Pfam-A domains across a dataset of 30 vertebrate species (Table 5.1).

Table 5.1: Species names and database versions of the 30 vertebrate genomes used for domain usage assessment

Species Name	Database Version
Anolis_carolinensis	AnoCar2
Bos_taurus	UMD3.1
Callithrix_jacchus	C_jacchus3.2
Canis_familiaris	CanFam3.1
Carlito_syrichta	tarSyr1
Cavia_porcellus	cavPor3
Danio_rerio	GRCz10
Dasypus_novemcinctus	Dasnov3
Equus_caballus	EquCab2
Felis_catus	Felis_catus_6.2
Gallus_gallus	Gallus_gallus-5
Gorilla_gorilla	gorGor3.1
Homo_sapiens	GRCh38
Latimeria_chalumnae	LatCha1
Loxodonta_africana	loxAfr3
Macaca_mulatta	Mmul_8
Meleagris_gallopavo	UMD2
Monodelphis_domestica	BROADO5
Mus_musculus	GRCm38
Myotis_lucifugus	Myoluc2
Nomascus_leucogenys	Nleu1
Ornithorhynchus_anatinus	OANA5
Pan_troglodytes	CHIMP2.1
Papio_anubis	PapAnu2
Pongo_abelii	PPYG2
Rattus_norvegicus	Rnor_6
Sus_scrofa	Sscrofa10.2
Taeniopygia_guttata	taeGut3.2
Takifugu_rubripes	FUGU4
Xenopus_tropicalis	JGI_4.2

The 30 vertebrate species and their corresponding database versions obtained from the Ensembl Genome Browser (Version_90)(Herrero *et al.*, 2016) *for both unipartite and bipartite network construction to investigate domain usage across vertebrates.*

The vertebrate dataset was cleaned and filtered using python script - PfamFilter.py (Appendix_B) to create a high quality dataset of vertebrate with each domain only being represented once across the dataset. The use of appropriate filters was an important step as smaller domains in PFam can be nested inside much larger domains. Therefore the PFam data has to be processed according to the following filtering steps:

(1) If multiple nested sub-domains exist within a larger domain, only the largest domain with the lowest e-value was retained.

(2) If 80% or more of a domain/s overlapped with another domain on the same protein only the single domain with the lowest e-value was kept.

These filters are necessary to minimise false positives caused by multiple members of the same pFam family (those with similar sequence motifs) aligning to the same position on a gene.

Using the filtered pFam data and the 30 vertebrate genomes a global undirected bipartite network was constructed to illustrate relationships between vertebrate protein-coding genes (node set 1) and pFam domains (node set 2). Edges were drawn if sequence similarity between node set 1 and node set 2 was above the specified e-value threshold – e^{-20} . This global network was deconstructed so that each individual connected component produced a bipartite network (Appendix_A). From this, a unipartite projection was made by removing vertebrate gene nodes and retaining the edges between pFam domain nodes that co-occur on the same gene (Appendix_A).

5.2.2) Domain co-occurrence network centrality

Three measures of centrality were calculated for our dataset of vertebrate protein-coding genes:

1) degree centrality
2) closeness centrality
3) betweeness centrality

Due to the undirected nature of our networks across our dataset, centrality calculations were based on the edge frequency of each node. Using a lab-generated program namely "networkStats.py" (Appendix_B), both closeness and

betweenness centrality were normalised as follows. Closeness centrality was normalised by accounting for the remaining number of nodes in the network, where "n" in the number of nodes in the network (n-1), and betweenness centrality by the maximum number of pairs of nodes across the network excluding the node of interest ((2/(n-1)(n-2))).

5.2.3) Comparison of degree distribution of the vertebrate gene network to a typical scale-free network

In order to compare vertebrate network structure to structures typically found across biological networks we compared and visualized degree distribution of the vertebrate protein-coding co-occurrence networks, against (1) a randomly generated network, and (2) a network generated with a typical scale-free property. The vertebrate protein-coding gene dataset had identical node frequencies when compared to both the randomly generated and scale-free networks. The analysis was carried out using the lab-designed code "Degree_Graph_generator.py" (Appendix_B) that specifically used the networkX functions gnp_random_network and scale_free_network for random network generation (Hagberg *et al*., 2008).

5.2.4) An investigation of domain usage across RMGFs and comparison to simulated dataset of human non-fused protein-coding genes

An analysis of domain usage across human RMGFs was carried out and compared with that of human non-fused protein coding genes. The protein coding sequences of 27 human RMGFs were obtained from the Ensembl Genome Browser (**Version_92**)(Herrero *et al.*, 2016). A dataset of human non-fused protein coding genes were prepared using the latest human genome (**GRCh38**) (Herrero *et al.*, 2016). To compare the domain usage of RMGFs to that of the human non-fused proteincoding gene dataset, 100 randomly sampled datasets (without replacement) of 27 genes (with lengths +/-10% of RMGFs) were extracted from the human dataset using "Dataset_randomgrabber.py" (**Code_Box 3).** Both human RMGFs and the 100 simulation dataset underwent pHMMer analysis (Version 31.0) (Finn, Clements and Eddy, 2011) using an e^{-20} threshold. Results of both pHMMer analyses were further filtered using lab

generated software "Pfam_filter.py" (Appendix_B) for the construction of bipartite and unipartite graphs (Section 5.2.1). Results of the RMGF analysis were compared with those of the human non-fused protein-coding database.

Code_Box 3: "Dataset_randomgrabber.py" code used to generate 100 randomly sampled human non-fused protein coding genes

```
1  import os
2  import sys
3  import random
4
5  from Bio import SeqIO
6
7  res_root = 'res/2018-05-26/'
8  out_root = f'{res_root}/out'
9  fasta_path = f'{res_root}/Homo_sapiens.GRCh38.pep.all.fa'
10 lengths_path = f'{res_root}/proteinlengths.txt'
11
12 with open(lengths_path) as lengths_fh:
13    lengths_raw = lengths_fh.read()
14    lengths = list(map(int, lengths_raw.strip().split('\n')))
15
16 records = SeqIO.to_dict(SeqIO.parse(fasta_path, 'fasta'))
17
18 for i in range(1, 101):
19    out_records = []
20    for length in lengths:
21       min_length = round(length*0.9)
22       max_length = round(length*1.1)
23
24       ids_of_apt_length = []
25       for record in records.values():
26          if min_length <= len(record.seq) <= max_length:
27             ids_of_apt_length.append(record.id)
28       chosen_id = random.choice(ids_of_apt_length)
29       if chosen_id in SeqIO.to_dict(out_records).keys():
30          print('Duplicate! Resampling...')
31          chosen_id = random.choice(ids_of_apt_length)
32       if chosen_id in SeqIO.to_dict(out_records).keys():
33          print('Fail')
34       chosen_record = records[chosen_id]
35       chosen_record.description = str(length)
36       out_records.append(chosen_record)
37    SeqIO.write(out_records, f'{out_root}/{i}.fasta', 'fasta')
```

Code_Box 3: *Code to create 100 datasets of human non-fused protein coding genes of the same length (+/- 10%) as the 27 RMGFs analysed. Line 1-4 reads in packages and sets up the standard input/output interface, line 6-9 gives the path to 2 user defined files 1) fasta file of non-fused human protein coding genes and 2) list of lengths. Line 15-18 reads in lengths file, takes each number and creates a min/max length so that a 10% range of length is accepted. Lines 20-30 reads the fasta file and examines it for genes that contains a length of between the min/max length value, The code then checks for duplicates andoutputs the fasta formatted genes into a file and this is carried out 100times.*

5.2.5) A functional analysis of domains identified across RMGFs and human simulated data

Functional annotation data was provided by pFam2go (Finn *et al.*, 2014) from the pFam database (**Release_31**) (Finn *et al.*, 2014) that associates pFam domains with a functional annotation, if known. The GO hierarchical terms for biological processes are shown in (Figure 5.1) and for functional annotations (Ashburner *et al.*, 2000) are summarised in (Figure 5.2).

This hierarchy of annotations allows each domain to be given several GO terms depending on the specificity needed. Lists of domains were acquired for both datasets from the unipartite projection networks created in Section 5.2.4 and their corresponding functional annotation extrapolated from the pFamtogo tool. Functional annotations were compared across both datasets in order to determine if any functional trend differences were evident between the two datasets.

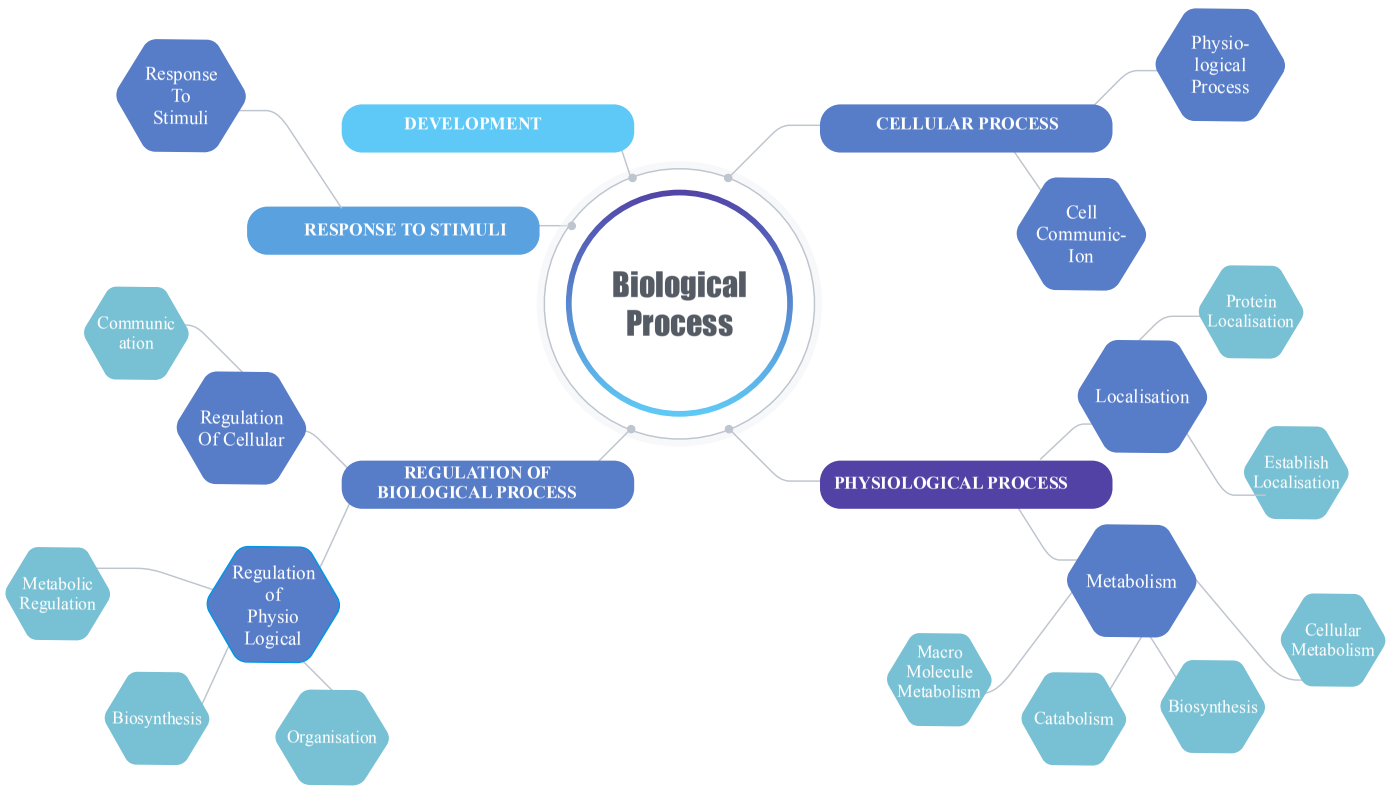


Figure 5.1: *Biological Process GO term hierarchical structure: A depiction of the GO term categorisation of Biological Processes with more centric terms being much broader than those terms located at the periphery that are more specific* (Ashburner *et al.*, 2000).

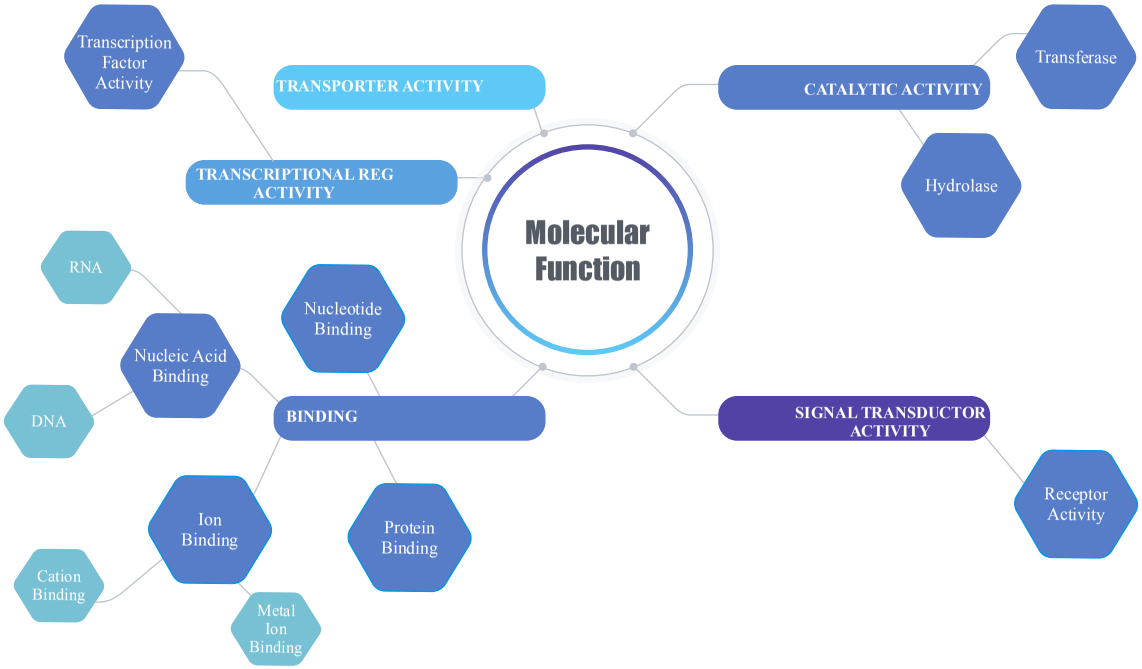


Figure 5.2: *GO term molecular function hierarchy: A depiction of GO term categorisation of Molecular Function with more centric terms being much broader than those terms located at the periphery* (Ashburner *et al.*, 2000).

5.3) Results

5.3.1) Bipartite network of vertebrate protein coding genes and pFam domains

A total of 3,702 networks were generated containing 2 node types: genes and domains. This represents 3,702 connected components (Finn *et al.*, 2014). Interestingly, the analysis uncovered a single graph (graph_0000) that contained 48,823 edges equating to 28% of the edges present over the entire global graph, this connected component contained hits from 427 pFam domains (8% of the entire dataset). The statistics of the largest 5 connected component graphs are shown in Table 5.2**.**

The largest component (Graph_0) contained 427 nodes, representative of 8% of the nodes within the global vertebrate network before connected component deconstruction. This component contains 48,823 edges that connect these nodes and these edges represent 28% of the nodes that exist in the global vertebrate graph. Interestingly, The second largest component found across the network (graph_3) only containing 0.6% of the domains within the pFam database and only containing 1162 edges, or only 0.9% of the total edges generated from the global network. Interestingly, 84% of all connected components are singletons, containing only one domain per component. This suggests that the vast majority of pFam domains that are found in the vertebrate protein-coding dataset are not promiscuous in nature, preferring isolation on specific genes over the presence on multiple genes.

Table 5.2: Statistics for the 5 largest connected components in the bipartite network of vertebrate protein coding genes and pFam domains

Graph	pFam _Count	pFam_Percent	Edge_Count	Edge_Percent
0	427	8%	48823	28%
3	33	0.60%	1662	0.90%
31	16	0.30%	550	0.30%
40	14	0.20%	476	0.20%
49	13	0.20%	515	0.20%

Representation of the 5 largest connected components after deconstruction of the vertebrate protein coding gene – pFam domain analysis. Column 1 is the number assigned to the connected component prior to deconstruction. Column 2 counts the amount of pFam domains within each component, and column 3 represents the number of pFam_Counts as a percentage of the total domain count of the global vertebrate network and a percentage is generated. Column 4 highlights the number of edges in each connected component and column 5 represents the number of Edge_Count as a percentage of the Edge_Count of the total vertebrate network.

5.3.2) Unipartite network creation in vertebrate protein coding gene dataset

To uncover the relationship between domains that co-exist on the same gene unipartite graphs were generated based on the graphs made in Section 5.2.1. These undirected unipartite graphs were constructed to determine domains that have a higher probability of co-occurring on the same gene, or contrastingly have a very low chance of being present on the same gene. In total, 585 projections were created form the 3702 bipartite graphs generated (Section 5.2.1). A statistical profile was carried out across the three largest unipartite graphs and results are displayed in Table 5.3.

As depicted in Table 5.3 the largest unipartite graph (Graph_2269) consists of 427 pFam (20% of entire pFam dataset) domains and 31% of the edge counts of the entire global network. Again when the second largest graph (Graph_2305) containing only 1% and 2% of domains and edges existing here respectively was compared against the largest unipartite network it clear to the drastically decreased connectivity. Of the 582 unipartite projections found 62% of these contained 2 domains (1 edge). This highlights the nature of these domains having slight preference against the formation of more modular proteins that contain multi-domain proteins.

Table 5.3: Top 3 largest unipartite networks based on bipartite human RMGFs and pFam domain database networks

Graph	Pfam_Count	Pfam_Percent	Edge_Count	Edge_Percent
2269	427	20%	709	31%
2305	33	1%	46	2%
835	16	0.70%	22	0.90%

Results of unipartite network creation based on vertebrate protein coding and pFam domain bipartite networks. Displayed are the top 3 largest projections. With column 1 representing the assigned number given to the unipartite connected component by the algorithm. Column 2 represents the number of pFam domains within the each connected component and column 3 illustrates a percentage of how many nodes pFam_Count represents in the global bipartite network from which the unipartite graph was constructed. The Edge_Count column represents the amount of edges connecting each domain and Edge_Percent uses this to calculate the number of edges in each unipartite graph in comparison to the bipartite graph it was generated from.

5.3.3) Centrality of RMGF unipartite networks

Unipartite centrality was investigated through degree, closeness and betweenness centrality calculations across unipartite projections of the bipartite networks generated from the global vertebrate and pFam domain network. The degree distribution of the unipartite graphs in comparison to both random graphs and scale free graphs (Figure 5.3) shows the scale free nature of this network. This finding is in line with other biological graphs such as protein-protein interaction networks or gene expression networks that exhibit scale free properties (Albert, 2005). A scale free topology suggests that most nodes contain a low degree (don't share many connections with other nodes) and that very few high degree or "hub " nodes exist in the network. This result indicates that most domains restrict interactions with other domains and very few domains preferentially interconnect with a plethora of other domains. However as the largest co-occurrence graph contains 20% of pFam families suggests that some domains do preferentially coexist alongside other domains on the same gene contributing the modular structure of genes.

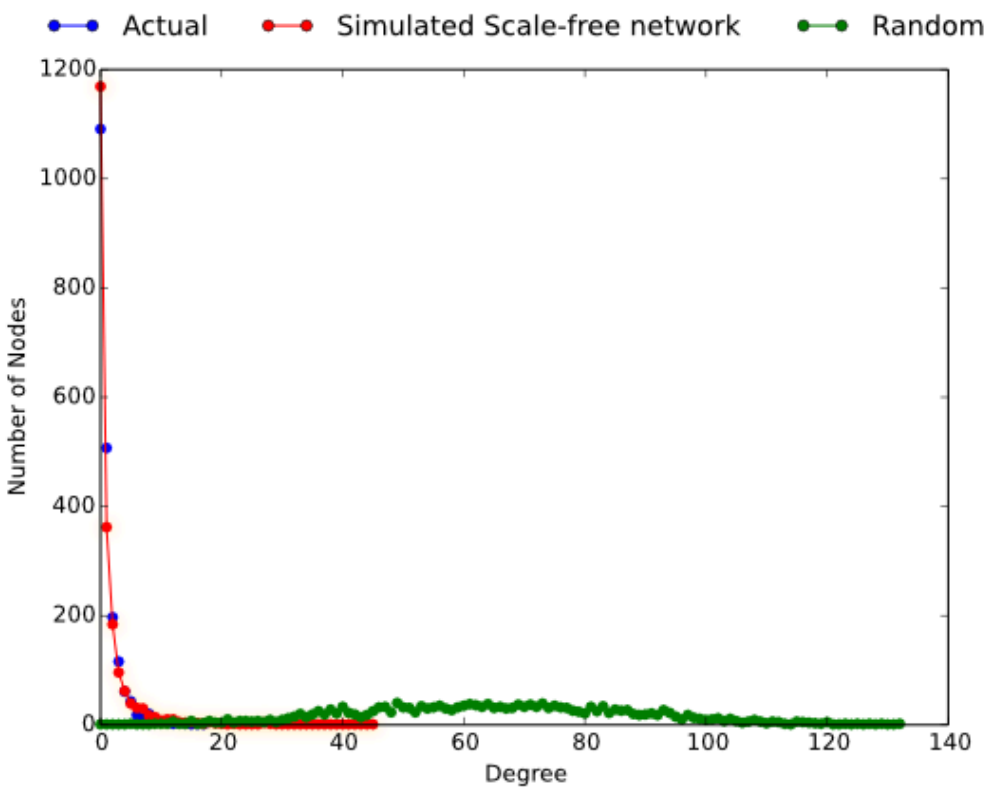


Figure 5.3: *A comparison of graph topology between vertebrate protein coding gene networks and randomly generated networks: Illustration of vertebrate protein coding gene unipartite network degree distributions (blue line) in comparison to randomly generated graphs of the same degree distribution (green line) as well as a graph with a simulated scale free topology (red line).*

5.3.4) An analysis of domains in RMGFs in comparison to non-fused human protein coding genes

An investigation of domain usage patterns was carried out across the 27 RT-qPCR validated human RMGF genes identified in Section 3.3.3 and compared to non-fused protein coding genes of similar length. Similarly to that carried out in the vertebrate analysis a global bipartite network was generated in order to identify domains across the RMGFs, after bipartite decomposition into it's connected components (Section 5.2.1) 48% of RMGFs had no domains present, 37% of the genes had only 1 domain present and 11% of the genes had 2 domains present on the same gene. This trend suggests that human RMGFs in general contain domains of low promiscuity. The corresponding unipartite network projections supported these data with only 3 genes showing evidence of co-occurrence, these domains include; PRY and Rpr2 (ENST00000513556), UPF0552 and Clat_adaptor_S (ENSG00000250021), and STIMATE and Mustang (ENSG00000248592). A pFam2go analysis was used to highlight specific functional enrichment across 1) domains (Table 5.4) and 2) the genes containing domains (Table 5.5). The lack of information available across identified domains suggests that these domains are new annotations. The PANTHER analysis of genes containing these domains (Table 5.5) highlights the predominance of enzymatic, membrane receptors, transport and apoptotic functions.

Table 5.4: pFam2go results of domains identified in domain usage analysis of human RMGFs

Domain	Pfam2go Description	Co-occurence
Jiv90	No information available	No
Jirairya	No information available	No
Asp	Protein Transport	No
TMEM251	No information available	No
PRY	No information available	Yes
Rpr2	No information available	Yes
UPF0552	No information available	Yes
Clat_Adaptor_S	No information available	Yes
Tetraspanin	Integral component of membrane	No
STIMATE	No information available	Yes
Mustang	Chrondrocyte differentiation, tissue regeneration, chondrocyte proliferation	Yes
KRAB	Nucleic acid binding, regulation of transcription	No
Bcl-2	Regulation of the apoptotic process	No
Acid-PPase	Phophotase activity	No

Human RMGF domain Pfam2go information. The left column contains the domains identified across the RMGF gene panel and the right hand side signifies the Pfam2go results if available and the third column indicating whether it is present on a gene alongside another domain i.e. co-existing domains.

Table 5.5: PANTHER investigation across RMGFs with identified pFam domains

RMGF_Gene_ID	PANTHER Analysis
ENSG00000257390	Molecular Function: Protein binding Cell Component: Cytoplasmic ribonucleoprotein, nuclear speck Panther Subfamily: DNAj Homolog member 14
ENSG00000269035	Cell Component: Integral membrane component Panther Subfamily: Transmembrane protein 221
ENSG00000250644	Molecular Function: aspartic-type endonuclease activity, proteolysis, protein catabolic process. Biological Process: Response to biotic function autophagy Cell Component: Integral membrane component Panther Subfamily: Cathepsin D
ENSG00000248167	No information available
ENSG00000255730	Molecular Function: 3metyl-2oxobutanoate dehydrogenase, alpha keto-acid dehydrogenease Biological Process: Branch chain a.a, catabolic process, oxidation reduction Cell Component: mitochondrial alpha- ketoglutarate dehydrogenase Panther Subfamily: 2-oxoisovalerate dehydrogenase subunit α
ENSG00000250021	Biological Process: Negative regulation of actin nucleation, vesicle mediated transport, protein transport Cell Component: Membrane coat, intracellular membrane bound organelle Panther Subfamily: Arpin related
ENSG00000250349	Molecular Function: Calcium ion binding Biological Process: Cell surface receptor signaling Cell Component: Integral component of plasma membrane, extracellular region. Panther Subfamily: Tetraspannin
ENSG00000249590	Panther Subfamily: SEC-14 like protein 2
ENSG00000248592	Biological Process: Chrondrocyte differentiation/proliferation, tissue regeneration Cell Component: Nucleus, integral component of membrane Panther Subfamily: Store operated Ca entry regulator
ENSG00000258643	Molecular Function: Identical protein binding, protein homo/hetero dimerization, distorted domain binding Biological Process: Sertoli cell proliferation, spermatogenesis, negative regulator of apoptosis, extrinsic signaling of apoptosis without ligand Cell Component: Bcl2 protein complex, integral comp of membrane, cytosol, mitochondrial outer membrane Panther Subfamily: Signalling molecule
ENSG00000255526	Molecular Function: phosphotase Biological Process: dephosphorylation

A GO term functional analysis of human 27 RMGFs. Left column is the Ensembl Gene Identifier (Herrero *et al.*, 2016) *and the right column outlines GO under specific categories; Molecular function, biological process and cell component. Gene identifiers in red are those that contain 2 domains* (Ashburner *et al.*, 2000).

In order to contextualize our findings we compared them to that of 100 randomly sampled human non-fused protein coding genes of the same size**.** Here, 15 (55%) bipartite networks across 100 random samples were generated in comparison to the 13 (48%) in the RMGF dataset. Of these bipartite networks 44% on average contained 1 domain compared to RMGF dataset's 37% and 11% of the human bipartite networks contain > 2 identical to the RMGF database. Therefore RMGFs co-occurrence pattern falls in line with what would be expected i.e. result identical to randomly simulated non-fused dataset but they have a lower than expected rate of genes containing just 1 domain and genes that contain no domains. From these analyses it is clear that non-fused genes tend to have a single domain but both RMGFs and non-fused genes share an equal likelihood of having more than 1 domain.

5.3.5) A GO term functional assessment of domains identified within RMGFs in comparison to human non-fused protein coding genes

This comparison was carried out across 3 broad GO term categories namely; Biological Function, Cellular Components and Molecular Functions (Ashburner *et al.*, 2000). Results indicate across both datasets that catalytic activity, physiological processes and binding domains are important (Figure 5.4). However, the RMGF dataset preferentially contains "developmental" and "response to stress" domains whereas the human simulated dataset appears to stick to "house-keeping" cell maintenance domains involved in protein localization, TF regulator activities and regulation of biological processes (Figure 5.4).

Within the cellular components category RMGFs contained domains that suggest localization in the integral membrane and nucleus, supporting associations with response, catalysis, and catalytic activity shown in Table 5.6. Cellular component GO term analyses results found in Table 5.6 highlight that domains identified across randomly sampled genes contain a much broader range of cellular components than found in the RMGF dataset alone. However, results did show some crossover with the RMGF dataset showing both membrane and nucleus localization preference.

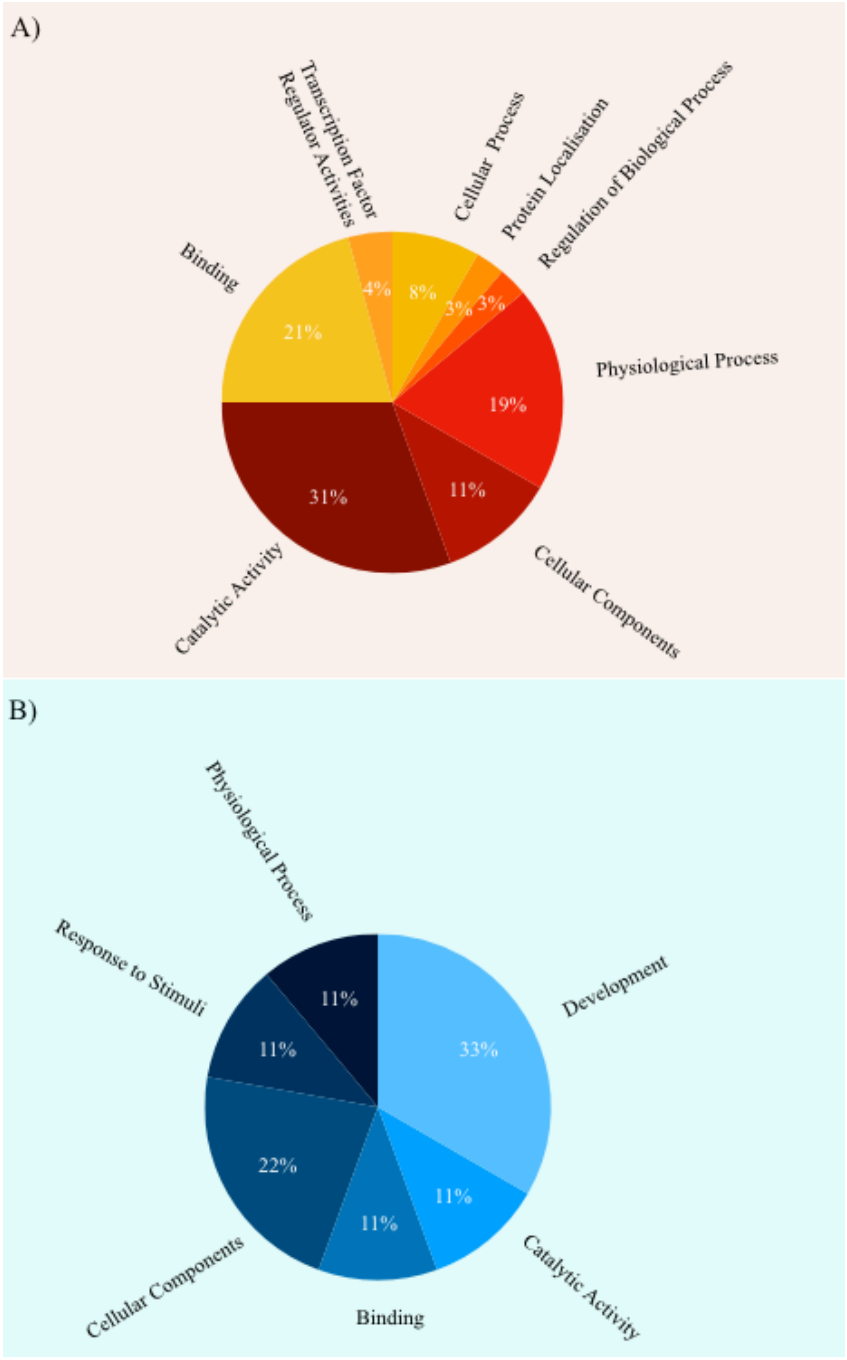


Figure 5.4: *GO term analysis of identified domains across RMGFs and non-fused human protein coding genes: Pie charts displaying proportion of GO terms corresponding to the domains identified across both RMGF and randomly sampled genes. (A) This chart depicts a GO categorization of RMGF domains identified as a percentage of the total domains identified, and (B) represents the same information but in the human RMGF dataset.*

Table 5.6: Results of cellular component GO term analysis across the human simulated datasets

Cellular Component
Mitochondrial outer membrane
COPII vesicle coat
Signal recognition particle
MHC class II protein complex
Membrane
Integral component of membrane
Nucleus

Cellular components identified across non-fused human protein coding.

5.4) Discussion

In this chapter we have used graph theory for the global characterization of 1) domain abundance and 2) domain combinations or co-occurrence that underpin the evolution of modular proteins, with a specific focus across RNA-mediated fusion genes identified in primates.

From the literature it has been suggested that the number of domains identified per species across the phylogenetic tree is correlated with an increase in both genomic and organismal complexity with the identification of a broader spectrum of domains in eukaryotes in comparison to prokaryotes (Gerstein, 1998) and moreover with animals having a higher domain frequency than unicellular eukaryotes (Basu *et al* ., 2009). With this in mind it would be expected that human would contain a much greater range of domains than that of more simple organisms such as Fugu however (illustrating a positive correlation between domain number and complexity), the most recent pFam Database release **(Version31.0)** does not support these findings (Table 5.7) (Finn *et al.*, 2014).

Table 5.7: A statistical analysis of the current pFam Database (Finn *et al.*, 2014) across vertebrates

Species	Number of domains	Sequence Coverage
Human	98,236	70%
Mouse	83,785	79%
Orangutan	45,305	86%
Macaque	68,795	85%
Marmoset	85,290	88%
Elephant	50,750	92%
Platypus	34,382	79%
Rat	61,228	91%
Pig	45,391	86%
Fugu	122,502	95%
ZebraFinch	33,904	95%
Cow	49,496	93%
Chicken	35,958	89%
Turkey	32,729	91%

A statistical assessment of the latest pFam Database (Finn *et al.*, 2014) *across a panel of vertebrate genomes with column 2 illustrating the number of domains that have been identified across each species and column 3 highlighting the percentage of the genome that has been assessed to date.*

The findings in Table 5.7 illustrate that human, an example of a highly complex genome, has 98,236 identified domains whereas Fugu and ZebraFinch, two genomes of lower complexity, contain 33,904 and 122,502 domains respectively (Finn *et al.*, 2014). When taken together no correlation between complexity and domain frequency is evident but rather support a more species-specific domain abundance level and further investigation is required to decipher any evident correlations at the genus or family level.

The protein repertoire has expanded and is drastically different across taxa and this concept is well supported across the literature *(Chothia et al., 2003; Buljan et al*., 2010) Domain architecture could potentially contribute to this novel protein generation as these mobile domain elements are readily transmitted across genomes with the potential to locate into new environments providing the necessary machinery to drive novel protein evolution a process known as domain shuffling (Kawashima *et al.*, 2009). Although, domains are versatile and mobile across genomes there is a constraint on the number of domains co-occurring within proteins that alludes to selective pressures restricting their transmission rate across genomes i.e. constraining domain promiscuity (Basu *et al*., 2009). Despite this evidence for constraint most proteins contain at least two domains (Chothia *et al.*, 2003) but rarely co-exist in a protein with a multitude of other domains, for instance out of the 98,236 domains identified across the human genome only 97 domains have shown signatures of promiscuity (Basu *et al.*, 2008). The top 20 highly promiscuous domains are highlighted in Table 5.8.

Table 5.8: An analysis of the 20 largest identified domain families and promiscuous domains

	Largest Domain Families	**Most Highly Promiscuous Domains**
1	WD40	RING
2	ABC_trans	AAA
3	Zf-C2H2	UCH
4	PKinase	PH
5	Mfs-1	PHD
6	Response_reg	SET
7	Ank_2	ANK
8	BPD_transop_1	UBQ
9	HATPase_C	C2
10	LRK_8	BROMO
11	RRM_1	Biotin_lipoyl
12	Helicase_C	MySc
13	PPR_2	S_TKc
14	Mito_carr	DEXDc
15	Fn3	DNAj
16	AMP_binding	BRCT
17	I-set	CHROMO
18	Adh_short	UBa
19	PPR	Cyt_b5
20	HisKA	GTP_ETFU

Column 2 lists the top 20 largest domain families currently identified in the pFam Database (Finn *et al.*, 2014) *while column 3 shows the top 20 promiscuous domains within the human genome* (Basu *et al.*, 2008).

Although the promiscuous domains in Table 5.8 are highly versatile across eukaryotes in general domain promiscuity is species-specific (Basu *et al.*, 2008). From these findings it is clear that domain usage across vertebrates requires additional investigation before their role in genome and protein architecture is fully understood. It is an important consideration that domain usage across individual genes may not be heterogeneous as new gene domain usage could potentially be different than would be expected in comparison to protein coding genes. Therefore, a further investigation of domain usage in the protein coding genomes of vertebrates is required before new genes can be assessed for an alternative domain usage pattern.

Here, an investigation of domain usage across RMGFs was carried out in order to determine domain usage patterns of a cohort of new genes against 1) randomly sampled non-fused human protein coding genes and 2) protein coding genes across a panel of 30 high quality vertebrate genomes.

For benchmarking purposes a panel of 30 vertebrate genomes (Table 5.1) domain usage pattern was assessed through sequence similarity networks. pHMMer profiles were used to generate a global bipartite vertebrate domain networks whereby an edge was drawn between a domain node and a gene node if and only if they shared sequence similarity, utilising networks in this way is a powerful way to uncover relationships in datasets. The global network was decomposed generating 3702 sub-graphs each containing a single connected component. As would be expected form the literature these graphs fall in line with most biological datasets and demonstrated a scale-free network topology (Figure 5.3) (Albert, 2005). Evidence for the scale free nature of the connected component graphs is highlighted in Table 5.2 were the largest connected component (Graph_0) contains 427 nodes which is 8% of the total domain utilised in the initial global vertebrate network therefore this projection contains 2.5% of the total domains within the pFam database (Finn *et al.*, 2014). However, the second largest network projection (Graph_2305) with 33 nodes and 46 edge connections contains only 0.19% of domains in the total pFam Database - a significant decrease. It was found that 62% of the total unipartite projections constructed contained a single edge, this implies that 62% of the domains identified in vertebrate genomes occur on a gene with another domain and this finding is

supported by the current literature (Basu *et al.*, 2008). It can be inferred that 48% of the domains identified in vertebrates have a preference to have >1 domain present on a single gene, again this is supported by the literature (Basu *et al.*, 2008). For instance, the frequencies of the Protein Databank's (PDBs) domains were compared across *S. cerevisiae*, a single celled eukaryote, and *E. coli*, a prokaryote. It was found that whilst 56% of genes in *S. cerevisiae* had >2 domains present only 38% of *E. coli* had this domain co-occurrence frequency. When compared for genes containing >5 domains *S. cerevisiae* had 12.6% with this domain usage frequency but *E. coli* only had 2.8%. This highlights an expansion of domains across eukaryotes in comparison to prokaryotes and a preference for eukaryotes to have proteins of a multi-domain nature (Gerstein, 1998). Although the bipartite projections in Table 5.2 and the unipartite projections in Table 5.3 results alluded to a scale-free topology for these relationships a follow up centrality analysis cemented this finding as an assessment of degree, closeness and betweenness of the global vertebrate network fell in line with both the trend shown in a randomly generated graph and a simulated graph of scale-free topology (Figure 5.3), again an expected result based on our current understanding of the scale-free nature of biological data (Albert, 2005).

This scale free topology of domain usage across vertebrates suggests that most domains are present on genes with a low number of other domains whilst very few are present on a gene with multiple other domains e.g. an analyses carried out on domain super-families uncovered that few super-families are highly versatile and most have very low versatility (Dasu *et al*., 2015). In summary, the vertebrate genome domains are usually found on a gene with another domain present and a little less than half of these domains actually have >2 domains present. The vertebrate networks scale free topology supported that very few domains are promiscuous and most are potentially under selective constraint to inhibit their transmission thus most reside on very few genes and have much lower network degree centrality therefore in general the vertebrate network supports current analyses from the literature and is a good benchmarking tool for further comparisons.

A panel of 27 RT-qPCR validated human RMGFs were investigated in the same manner as for the vertebrate dataset. Here, 48% of the RMGFs had no domains

present, 37% contained just 1 domain whilst 11% contained >1 domain. These findings suggest that most RMGFs contain domains of a non-promiscuous nature, which could be expected of their evolutionary history by gene fusion. Three co-occurring domains were identified across the RMGF panel assessed 1) PRY and Rpr2 domains, 2) UPF0552 and Clat_adaptor_S and 3) STIMATE and Mustang (Table 5.4). Rpr2 is a known ribosomal P subunit that is used during RNA processing particularly in the mitochondria, but also plays a role in the processing of 5.8rRNA (Cook *et al.*, 2018). Unfortunately information is limited on the PRY domain other than its association with another domain SPRY (Cook *et al.*, 2018). The Clat_Adaptor_S domain is commonly found as a chaperone protein during the vesicular transport of molecules within cells, whist no information is yet known about its co-occurring domain UPF0552 (Cook *et al.*, 2018). Lastly the STIMATE domain is a known transmembrane protein within the endoplasmic reticulum and its co-existing domain Mustang has a role in tissue regeneration and chrondrocyte differentiation and proliferation (Cook *et al.*, 2018). A PANTHER analysis of the genes these co-occurring domains highlight a preference of enzymatic, membrane and transportation functionalities suggesting a clear role for both inter and intra cellular communication (Figure 5.4). The interpretation of any biological result is challenging when interpreted in isolation therefore all results were compared to a simulated dataset of random non-fused human protein coding genes for contextualisation purposes. Overall, an increased number of bipartite networks were generated with 15 created in the non-fused dataset in comparison to just 13 in the RMGF dataset. Not only this but a higher percentage of these graphs contained domains that had a preference to co-exist on a gene with just 1 domain (non-fused: 44%, RMGFs 37%) but in both non-fused and RMGFs 11% of networks displayed a preference for co-existence on a gene with >1 domain. In summary, RMGFs do not have a domain usage pattern different to that of human non-fused protein coding genes, which is expected as these fused genes were created as a result of an RNA-mediated event, and therefore their un-fused parents (potentially present in the un-fused human dataset) are still present in the human genome.

A major limitation of this analysis was the lack of power due to the very low numbers of RMGFs that could be investigated. Future analysis could potentially

reduce the stringency threshold implemented during the RMGF identification process (Section 2.2.1.1) allowing additional potential gene fusions to be assessed, however by lowering this stringency a focus would not be placed solely on RMGF events but rather gene fusion/fission events in general. The filtration process (Section 5.2.1) was another confounding feature of the analysis, as leaving only the largest domain within the dataset ignores all nested domain events and therefore a layer of complex information is ignored by the analysis. This is particularly important as it has been shown in the literature that tandemly duplicated domains expanded in metazoan species and that these within these repeated regions a domain does not work in isolation but rather only becomes functionally active when co-existing with other domains within the repeated region (Björklund, Ekman and Elofsson, 2006).

Future work could construct protocols to deal with the complexity of domain architecture amongst genomes and thus provide a more accurate picture of domain usage. Although much effort has been placed on the identification of domains across species there is still a significant amount of sequence that has not yet been covered (Table 5.7) and so many domains await discovery. Our analysis illustrates the trend of domain usage of both new genes, with a focus on RMGF events, and protein coding genes in general given the current information and using the most recent protocols (Mi *et al.*, 2009).

Chapter 6: Discussion and Conclusions

With the release of the Great Ape Genome Project (Prado-Martinez *et al.*, 2013) and the development of novel network based algorithms (Jachiet *et al.*, 2013) we had the unique opportunity to establish how remodelling, specifically through RNA-mediated gene fusion (RMGF) events has played a role in the evolution of Great Apes (Kaessmann, 2010; Tung *et al*., 2010).

Sequence similarity networks (SSNs) are a powerful way to detect gene fusions (Alvarez-Ponce *et al.*, 2013) and these methods have seen considerable development (Jachiet *et al.*, 2013; Pathmanathan *et al.*, 2018) At the time of performing the analysis in Chapter 2 the most recent method available was MosaicFinder (Jachiet *et al.*, 2013). There are limitations to consider and both the data and the methods have undergone new iterations and releases since the analyses for this thesis were performed. For instance, MosaicFinder (Jachiet *et al.*, 2013) has since been replaced by CompositeFinder (Pathmanathan *et al.*, 2018) which is better able to deal with larger datasets with higher precision, recall and accuracy. However, a strength of Mosaicfinder is that it is more suited to identifying highly conserved remodelled events whereas CompositeFinder's unique selling point is in it's ability to account for fast evolving fusion genes across very large datasets. Therefore the application of the MosaicFinder algorithm is more appropriate for the investigation of RMGFs in a dataset of this size and at this phylogenetic depth. Regardless of which algorithm is applied to the sequence data there is a requirement for the parent genes to be present within the genome, and scenarios where one or more parent genes have been lost are not detectable. Therefore it is likely that the number of gene fusions we report is an underestimate of the total number of gene fusions that occur and are retained in the genome.

The data used in the analyses of expression levels of RMGFs was obtained from an analyses carried out in 2010 by an Illumina Genome Analyzer II producing single-end reads of 75bp in length (Brawand *et al.*, 2011). This technology has been supplanted by the NovaSeq 6000 sequencing Illumina platform that produces short end reads and can produce 20 billion reads per run essentially providing more opportunities to uncover lowly transcribed genes. In addition, the

genomic sequence data used throughout this thesis have also been improved since the analyses were carried out, e.g. a new chimpanzee and orangutan genome were released with 65x coverage produced by advanced real time (SMRT) long read sequencing technology (Kronenberg, ZN *et al*., 2018).

On comparison of the number of RMGFs in human to other species analysed human contains a larger number of RMGFs. Additionally, we determined that these fusions were enriched in regions of segmental duplication. Through GO term analyses we uncovered functional enrichment for both binding and catalytic activity across the remodelled genes (Eden *et al.*, 2009). Of course GO term analysis is a crude way of assessing function as more 'popular' proteins are more intensively analysed yielding intricate annotations whilst less 'popular' proteins remain poorly annotated or not annotated at all (Gaudet and Dessimoz, 2017). Due to the stringent nature of my analysis low numbers of RMGFs were identified and thus enrichment analyses did not have statistical power, thus all enrichment analyses were based off fusion parents. GO terms are based aggregated data and Simpson's paradox suggests that this aggregation may lead to different results if compared to the data in an un-aggregated fashion therefore caution needs to be taken when considering results obtained (Gaudet and Dessimoz, 2017). Also, 99% of GO annotations are computationally predicted and not experimentally validated - a factor that also requires consideration when interpreting results (Alterovitz *et al.*, 2007). Conversely, in the absence of resources, capacity and licenses/permits to do specific functional analyses GO terms can provide useful insights into putative function.

Human segmental duplications are known to be enriched for genic content in comparison to segmental duplications in other Great Apes (Khurana *et al.*, 2010). Our discovery that human RMGFs are enriched in known breakpoints for segmental duplication is perhaps unsurprising given that segmentally duplicated regions are known to be highly volatile (Marques-Bonet *et al*., 2009) and contain mobile sequences that provide the necessary instability to increase the likelihood of gene fusion (Marques-Bonet *et al*., 2009). DNA-mediated gene fusions have been associated with important roles in both inter/intra cellular communication

and signal transduction (Latysheva *et al.*, 2016) and the enrichment for catalytic and interaction based functions in RMGFs fits with this observation suggesting that these functions are common in both forms of gene fusion. However, our sample size for RMGF is small, these functional categories are broad and there are a large number of genes in these GO categories.

A substantial body of research has been carried out on the impact of remodelled genes on phenotype as reviewed in (Chen *et al*., 2013). Briefly, RMGFs have the potential to contribute to phenotype on two levels: (1) by the production of stable protein products with either an identical function to one of its parents or with a novel function that is advantageous for the survival of the organism (Latysheva *et al.*, 2016), or (2) on a transcriptional level by the production non-coding regulatory RNA transcripts (Polychronopoulos *et al.*, 2017). Non-coding RNA transcript outputs can potentially affect both the transcriptional and translational output of a given cell through mechanisms such as chromatin remodelling (Magistri *et al.*, 2012), non-sense mediated decay (Smith and Baker, 2015) and can even effect translation efficiency across a cell (Bazin *et al.*, 2017). Non-coding RNAs have been extensively studied and can impact a cells transcription and translational output. Non-coding RNA transcripts have been shown to be present across genomes at unexpectedly high frequencies, e.g., 1/3 of all alternatively spliced isoforms produce non-coding transcripts with regulatory functions (Wang *et al.*, 2008). It is clear that deciphering a new genes transcriptional and translational profile is essential to understanding their contribution to the evolution of both the genome and phenome.

Both computational and follow-up wet bench analyses highlighted an increase in RMGF expression in the testes across all species examined. This supports the proposals from the current literature stating that new genes are initially expressed exclusively in a testes specific manner and over time this expression can potentially become more broad (Kaessmann, 2010). In human only, the cerebellum tissue contained a high number of expressed RMGFs - a finding that was not shared across other species in the dataset. RMGFs in human brain had a differential expression pattern when compared to all other species in the dataset

and when RMGFs were compared between tissue samples, human was the only species containing a brain-cerebellum differential expression profile. This differential expression is not limited to RMGFs however as ~1400 genes in the cerebellum are differentially expressed in the human cerebellum in comparison to all other brain tissues (Khaitovich *et al*., 2004).

The exploitation of translational profiles for each RMGFs using a recently established translatomic sequencing technology – Ribo-sequencing (Jackson and Standart, 2015) – was an exciting and important aspect of this work. In contrast to transcriptomic technologies and sequence datasets, translatomics is lagging behind. Most techniques for translatomics still rely on labour intensive fractionation followed by Western Blotting (Burnette, 1981). However, with the advent of ribo-sequencing technologies this bottleneck in data availability is set to improve and now translatomic profiles can be obtained computationally on a genome-wide level (Chassé *et al.*, 2017). Out of 19 human RMGFs assessed across 4 publically available datasets (2 fibroblast, 1 skeletal muscle and 1 glioma dataset) (Michel *et al.*, 2014) three genes were identified as having protein products. However, tissue sample availability and resultant datasets for ribo-sequencing are currently limited to a small set of tissues. This makes any conclusion around the production of translated protein products from RMGFs beyond the current data.

In general, all RNA sequencing experiments (be they RNA-seq or Ribo-seq) are subject to spatial and temporal variation issues. Therefore it is challenging to accurately classify an RNA molecule as definitively expressed or translated. A negative gene expression result does not necessarily infer lack of transcription, but rather only the lack of transcription in that tissue at that specific time in that particular cellular context. Microarrays have been proposed instead of RNA sequencing technology to investigate differential expression across lowly expressed genes for better sensitivity and more robust results (Liu *et al.*, 2011). This could be an area for future investigation.

Human tissue panels are readily available for qRT-PCR style analyses and all the RMGFs present in human were assessed using 5 human tissue panels. It is entirely possible that the expression profile for these genes would be different in other Great Ape species but RT-PCR style experimental approaches to determine expression were limited by the supply of Great Ape RNA samples, therefore only 2 of the RMGFs could be examined across the Great Apes.

Despite these limitations it is clear that RMGFs have the ability to produce mRNA transcripts, potentially rRNA transcripts and viable protein products. With this knowledge and the availability of 127 human epigenomes (Roadmap Epigenomics Consortium *et al.*, 2015) we characterised their mechanisms of transcriptional control. Our results show that human RMGFs are predominantly located in loose, coiled euchromatic regions and have a preference for activating histone modifications. These findings offer further support for the production of transcriptional output for the RMGFs.

The SFmap package (Paz *et al.*, 2010) used uncovered a bias for both SRp20 (splice factor expression: spleen, lung and testes (Papatheodorou *et al.*, 2018)) and NOVA1 (splice factor expression: brain, adrenal and testes (Papatheodorou *et al.*, 2018)) splice factors. SRp20 has been associated with spliceosomal assembly, DNA replication and elongation and NOVA1 is a known brain-specific splice factor responsible for neural cell-cell adhesion and aids neural plasiticity (Eden *et al.*, 2009). Although SRp20 contains 2 binding motifs only the wcwwc motif was enriched across RMGFs suggesting a potential motif bias. This is an interesting avenue for future investigation as bias in splice factors binding motifs usage remains under investigated to date. We also established that splice factor co-occurrence patterns are significantly different between RMGFs and non-fused human genes and this requires further investigation.

We observed an absence of binding sites in RMGFs for a total of 6/21 splice factors. Four of these splice factors are members of the hnRNP splice factor family known to promote transcriptional silencing (Martinez-Contreras *et al.*, 2007; Busch and Hertel, 2012; Geuens, Bouhy and Timmerman, 2016). This

result would suggest that there is a lack of repression amongst RMGFs, a finding that correlates well with the broad transcriptional profiles of RMGFs established in Chapter 3, and their presence in euchromatic regions.

SFmap (Paz *et al.*, 2010) is a powerful tool in the assessment of splice factor binding sites however since our dataset was investigated additional parameters have been incorporated (Paz *et al.*, 2014). The incorporation of additional PSSM filtering and background models that account for the heterogeneous nature of sequences (splice sites, 3' and 5' UTR sequences, non-coding RNA and intragenic sequences specifically) reduce false discovery rates (Paz *et al.*, 2014). Future work could re-analyse this data using this updated algorithm.

The RoadMap Epigenomics Database (Roadmap Epigenomics Consortium *et al.*, 2015) only provides information on human epigenetic markers and therefore comparative genomic analyses were not possible but could be an avenue for future work. The histone modification content is confined to just 5 histone modifications, four activating and only a single repressive marker - making comparisons between activating and repressing modifications somewhat biased. Furthermore, investigating a selected panel of 5 modifications only provides a snapshot into histone modification usage across these epigenomes. On top of data limitations, histone modifications can be functionally dynamic depending on: (1) their position, *e.g.* a histone modification might be a promoter of transcription in one binding region but a repressor if bound elsewhere (Bernstein *et al.*, 2005), and (2) cross-talk between histones. Therefore, analysing histone modifications individually, and not in context of their genic environment may not provide the most biologically realistic results (Bernstein *et al.*, 2005).

The incorporation of extra-genic domains into proteins (either RMGFs or other) by non-homologous recombination has the potential to generate novel function (e.g. jingwei (Long and Langley, 1993)). Domain shuffling can be thought of as a finer-scale form of the gene remodelling we assessed in Chapter 2, and therefore can be identified in sequence data using similar network based tools. However, the length of the sequences involved requires the application of HMM

based sequence similarity algorithms. The pFam Database (Finn *et al.*, 2014) provides a 16,712 set of domain data which we used as input for our sequence similarity network. The homologous regions were detected using the profile based sequence similarity search algorithm (pHMMer) (Finn, Clements and Eddy, 2011)). The domain abundance and co-occurrence rate was established and found through centrality calculations to have a scale free topology typical of biological graphs according to recent literature (Albert, 2005). We found that RMGFs have a lower than expected number of genes containing only a single domain. This finding is expected due to the fusion process, i.e., the fusion gene will most likely inherit at least one domain from each parent. One might have expected that a remodelled gene could therefore generate unique domain patterns. However, this was not the case. We did however find evidence to support a significant difference in domain co-occurrence in the RMGF set when compared to our simulated dataset of non-fused human protein-coding genes. None of the domains identified across RMGFs were members of the top 20 promiscuous domains identified across human (Table 5.8) (Basu *et al.*, 2008).

The 90% sequence identity threshold set during initial network analysis identified a small number of high quality RMGFs. This threshold was selected as lower thresholds; the 80% identity threshold for instance, identified DNA-mediated gene fusion events as well as an increased potential false positive rate (higher number of mismatches between the identified fusion genes and their corresponding parent genes). The level of sequence divergence (mismatches) can result in distant homologs being identified as fusion parents (Jachiet *et al.*, 2013). In future investigations the percentage identity threshold could be dropped, and DNA-mediated fusion genes could also be examined. For our analysis involving domains, nested domains and tandemly repeated domains could not be considered due to their repetitive nature and the algorithm being used being based on sequence homology. Although much of the literature assumes that 90% of domains are un-nested (Aroul-Selvam, Hubbard and Sasidharan, 2004), domains can exist with two or even three additional insertions and this contributes to protein structure and function (Aroul-Selvam *et al*., 2004) and complicates the process of identifying domain shuffling events. For example,

Thermoplasma acidophilum contains an archael chaperonin domain that has been is inserted into an intermediate domain that is subsequently placed into an ATPase domain (Ditzel *et al.*, 1998). Tandemly repeated domains are also of fundamental importance to protein structure and function as some domains only function if adjacent to another (Björklund *et al*., 2006) and thus analysing domains individually and not in their genic context may be misleading or biologically unrealistic. For instance, the RNA methyltransferase gene contains a domain that is tandemly duplicated, and one domain is responsible for the correct folding of the other copy of the domain (Sunita *et al.*, 2007). A final confounding factor of this analysis is the use of the pFam Database. Although the pFam Database is a vital resource when assessing domains across species it is important to consider that not all species have had domains identified across a comparable amount of their sequence (Finn *et al.*, 2014).

Additional work is required to build on the results from this thesis. Sequencing the breakpoint of the 9 human specific RMGFs, functional annotation of the 3 RMGF protein products identified by Ribosequencing, reanalysis using the latest chimpanzee and orangutan genomes and perhaps reanalysing the data in the context of all gene fusion events (both DNA and RNA-mediated). This will help us further our understanding of RNA-mediated gene fusions and their impact on both primate genome and phenome evolution.

Conclusion

In this thesis we explored the impact of newly remodeled genes, specifically RNA mediated gene fusion events (RMGFs) on the evolution vertebrate species, with a specific focus on primates. Through our stringent network based analyses 69 RMGFs were identified across a dataset of 5 Great Apes, 1 non-Great Ape primate and mouse. The highest numbers of RMGFs were identified in human (42) and 9 were human-specific. Through both the transcriptomic and translatomic profiling of these genes we determined that RMGFs can be fixed in genomes and have the ability to be both transcribed and translated. Like other new genes RMGFs show a bias toward testes-specific expression in support of

the ‘out-of testis hypothesis’. This expression profile is supported by their *cis*-regulatory element usage as the identified RMGFs were more commonly associated with; activating histone modification profiles, regions of euchromatin and levels of splice factors with testes and brain biased expression profiles. However, RMGF domain usage did not siginificantly differ from non-fused protein coding genes across primates or indeed vertebrates in general. Finally, we uncovered that regions of genomic instability such as segmental duplication provide a mechanism by which these genes can be both created and transmitted across genomes with potential phenotypic consequences.

Chapter 7: Bibliography

Abzhanov, A. *et al.* (2004) 'Bmp4 and morphological variation of beaks in Darwin's finches', *Science*, 305(5689), pp. 1462–1465. doi: 10.1126/science.1098095.

Adams, M. D. *et al.* (1991) 'Complementary DNA sequencing: expressed sequence tags and human genome project', *Science*, 252(5013), pp. 1651–1656. doi: 10.1126/science.2047873.

Aken, B. L. *et al.* (2017) 'Ensembl 2017', *Nucleic Acids Research*, 45(D1), pp. D635–D642. doi: 10.1093/nar/gkw1104.

Akerman, M. *et al.* (2009) 'A computational approach for genome-wide mapping of splicing factor binding sites', *Genome Biology*, 10(3). doi: 10.1186/gb-2009-10-3-r30.

Akiva, P. *et al.* (2006) 'Transcription-mediated gene fusion in the human genome', *Genome Research*, 16(1), pp. 30–36. doi: 10.1101/gr.4137606.

Akkers, R. C. *et al.* (2009) 'A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in Xenopus Embryos', *Developmental Cell*, 17(3), pp. 425–434. doi: 10.1016/j.devcel.2009.08.005.

Albert, R. (2005) 'Scale-free networks in cell biology.', *Journal of cell science*, 118(Pt 21), pp. 4947–4957. doi: 10.1242/jcs.02714.

Alberts, B. *et al.* (2002) *Molecular Biology of the Cell, 4th edition*, *Garland Science*. doi: 10.3389/fimmu.2015.00171.

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(5), pp. 661–670. doi: 10.1261/rna.325107.

Alföldi, J. and Lindblad-Toh, K. (2013) 'Comparative genomics as a tool to understand evolution and disease', *Genome Research*, pp. 1063–1068. doi: 10.1101/gr.157503.113.

Alterovitz, G. *et al.* (2007) 'GO PaD: The Gene Ontology partition database', *Nucleic Acids Research*, 35(SUPPL. 1). doi: 10.1093/nar/gkl799.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Alvarez-Ponce, D. *et al.* (2013) 'Gene similarity networks provide tools for understanding eukaryote origins and evolution', *Proceedings of the National Academy of Sciences*, 110(17), pp. E1594–E1603. doi:

10.1073/pnas.1211371110.

Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.

Andolfatto, P. (2001) 'Adaptive hitchhiking effects on genome variability', *Current Opinion in Genetics and Development*, pp. 635–641. doi: 10.1016/S0959-437X(00)00246-X.

Andreev, D. E. *et al.* (2017) 'Insights into the mechanisms of eukaryotic translation gained with ribosome profiling', *Nucleic Acids Research*, pp. 513–526. doi: 10.1093/nar/gkw1190.

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data.*, *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. doi: citeulike-article-id:11583827.

Anisimova, M., Nielsen, R. and Yang, Z. (2003) 'Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites', *Genetics*, 164(3), pp. 1229–1236. doi: 10.1093/bioinformatics/btn086.

Araki, Y. *et al.* (2009) 'Genome-wide Analysis of Histone Methylation Reveals Chromatin State-Based Regulation of Gene Transcription and Function of Memory CD8+ T Cells', *Immunity*, 30(6), pp. 912–925. doi: 10.1016/j.immuni.2009.05.006.

Aravind, L. *et al.* (2005) 'The many faces of the helix-turn-helix domain: Transcription regulation and beyond', *FEMS Microbiology Reviews*, pp. 231–262. doi: 10.1016/j.femsre.2004.12.008.

Arimbasseri, A. G. *et al.* (2015) 'RNA Polymerase III Output Is Functionally Linked to tRNA Dimethyl-G26 Modification', *PLoS Genetics*, 11(12). doi: 10.1371/journal.pgen.1005671.

Aroul-Selvam, R., Hubbard, T. and Sasidharan, R. (2004) 'Domain insertions in protein structures', *Journal of Molecular Biology*, 338(4), pp. 633–641. doi: 10.1016/j.jmb.2004.03.039.

Ashburner, M. *et al.* (2000) 'Gene ontology: Tool for the unification of biology', *Nature Genetics*, pp. 25–29. doi: 10.1038/75556.

Babushok, D. V. *et al.* (2007) 'A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids', *Genome Research*, 17(8), pp. 1129–1138.

doi: 10.1101/gr.6252107.

Baele, G. and Lemey, P. (2013) 'Bayesian evolutionary model testing in the phylogenomics era: Matching model complexity with computational efficiency', *Bioinformatics*, 29(16), pp. 1970–1979. doi: 10.1093/bioinformatics/btt340.

Bailey, J. A. *et al.* (2002) 'Recent segmental duplications in the human genome', *Science*, 297(5583), pp. 1003–1007. doi: 10.1126/science.1072047.

Bailey, J. A. *et al.* (2004) 'Analysis of segmental duplications and genome assembly in the mouse', *Genome Research*, 14(5), pp. 789–801. doi: 10.1101/gr.2238404.

Bailey, J. A. and Eichler, E. E. (2006) 'Primate segmental duplications: Crucibles of evolution, diversity and disease', *Nature Reviews Genetics*, pp. 552–564. doi: 10.1038/nrg1895.

Bailey, T. L. *et al.* (2009) 'MEME Suite: Tools for motif discovery and searching', *Nucleic Acids Research*, 37(SUPPL. 2). doi: 10.1093/nar/gkp335.

Baldwin, M. W. *et al.* (2014) 'Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor', *Science*, 345(6199), pp. 929–933. doi: 10.1126/science.1255097.

Bapteste, E. *et al.* (2012) 'Evolutionary analyses of non-genealogical bonds produced by introgressive descent', *Proceedings of the National Academy of Sciences*, 109(45), pp. 18266–18272. doi: 10.1073/pnas.1206541109.

Barabasi, A.-L., Oltvai, Z. N. Z. N. and Barabási, A.-L. (2004) 'Network biology: understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113. doi: 10.1038/nrg1272.

Barbosa-Morais, N. L. *et al.* (2012) 'The evolutionary landscape of alternative splicing in vertebrate species', *Science*, 338(6114), pp. 1587–1593. doi: 10.1126/science.1230612.

Barski, A. *et al.* (2007) 'High-Resolution Profiling of Histone Methylations in the Human Genome', *Cell*, 129(4), pp. 823–837. doi: 10.1016/j.cell.2007.05.009.

Bartek, J., Hamerlik, P. and Lukas, J. (2010) 'On the origin of prostate fusion oncogenes', *Nature Genetics*, pp. 647–648. doi: 10.1038/ng0810-647.

Barton, N. H. and Otto, S. P. (2005) 'Evolution of recombination due to random drift', *Genetics*, 169(4), pp. 2353–2370. doi: 10.1534/genetics.104.032821.

Barton, R. A. and Venditti, C. (2014) 'Rapid evolution of the cerebellum in

humans and other great apes', *Current Biology*, 24(20), pp. 2440–2444. doi: 10.1016/j.cub.2014.08.056.

Basu, M. K. *et al.* (2008) 'Evolution of protein domain promiscuity in eukaryotes', *Genome Research*, 18(3), pp. 449–461. doi: 10.1101/gr.6943508.

Basu, M. K., Poliakov, E. and Rogozin, I. B. (2009) 'Domain mobility in proteins: Functional and evolutionary implications', *Briefings in Bioinformatics*, 10(3), pp. 205–216. doi: 10.1093/bib/bbn057.

Bazin, J. *et al.* (2017) 'Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation', *Proceedings of the National Academy of Sciences*, p. 201708433. doi: 10.1073/pnas.1708433114.

Becker, M. *et al.* (2018) 'Mapping of Human FOXP2 Enhancers Reveals Complex Regulation', *Frontiers in Molecular Neuroscience*, 11. doi: 10.3389/fnmol.2018.00047.

Beisang, D. *et al.* (2012) 'Regulation of CUG-Binding Protein 1 (CUGBP1) binding to target transcripts upon T cell activation', *Journal of Biological Chemistry*, 287(2), pp. 950–960. doi: 10.1074/jbc.M111.291658.

Bentley, D. R. (2006) 'Whole-genome re-sequencing', *Current Opinion in Genetics and Development*, pp. 545–552. doi: 10.1016/j.gde.2006.10.009.

Berg, J., Willmann, S. and Lässig, M. (2004) 'Adaptive evolution of transcription factor binding sites', *BMC Evolutionary Biology*, 4. doi: 10.1186/1471-2148-4-42.

Berget, S. M., Moore, C. and Sharp, P. A. (1977) 'Spliced segments at the 5′ terminus of adenovirus 2 late mRNA', *Proceedings of the National Academy of Sciences*, 74(8), pp. 3171–3175. doi: 10.1073/pnas.74.8.3171.

Bergthorsson, U., Andersson, D. I. and Roth, J. R. (2007) 'Ohno's dilemma: Evolution of new genes under continuous selection', *Proceedings of the National Academy of Sciences*, 104(43), pp. 17004–17009. doi: 10.1073/pnas.0707158104.

Bernstein, B. E. *et al.* (2005) 'Genomic maps and comparative analysis of histone modifications in human and mouse', *Cell*, 120(2), pp. 169–181. doi: 10.1016/j.cell.2005.01.001.

Berry, A., Pogorelcnik, R. and Simonet, G. (2010) 'An introduction to clique minimal separator decomposition', *Algorithms*, pp. 197–215. doi:

10.3390/a3020197.

Berthonneau, E. and Mirande, M. (2000) ‘A gene fusion event in the evolution of aminoacyl-tRNA synthetases’, *FEBS Letters*, 470(3), pp. 300–304. doi: 10.1016/S0014-5793(00)01343-0.

Bertoli, C., Skotheim, J. M. and De Bruin, R. A. M. (2013) ‘Control of cell cycle transcription during G1 and S phases’, *Nature Reviews Molecular Cell Biology*, pp. 518–528. doi: 10.1038/nrm3629.

Bhasi, A. *et al.* (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(1), pp. 48–54. doi: 10.1016/j.ygeno.2009.02.006.

Biamonti, G. *et al.* (1989) ‘Isolation of an active gene encoding human hnRNP protein A1. Evidence for alternative splicing’, *Journal of Molecular Biology*, 207(3), pp. 491–503. doi: 10.1016/0022-2836(89)90459-2.

Birzele, F., Csaba, G. and Zimmer, R. (2008) ‘Alternative splicing and protein structure evolution’, *Nucleic Acids Research*, 36(2), pp. 550–558. doi: 10.1093/nar/gkm1054.

Biswas, M. *et al.* (2011) ‘Role of histone tails in structural stability of the nucleosome’, *PLoS Computational Biology*, 7(12). doi: 10.1371/journal.pcbi.1002279.

Björklund, Å. K., Ekman, D. and Elofsson, A. (2006) ‘Expansion of protein domain repeats’, *PLoS Computational Biology*, 2(8), pp. 0959–0970. doi: 10.1371/journal.pcbi.0020114.

Blackburne, B. P. and Whelan, S. (2012) ‘Measuring the distance between multiple sequence alignments’, *Bioinformatics*, 28(4), pp. 495–502. doi: 10.1093/bioinformatics/btr701.

Blanchette, M. *et al.* (2006) ‘Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression’, *Genome Research*, 16(5), pp. 656–668. doi: 10.1101/gr.4866006.

Blau, J. *et al.* (1996) ‘Three functional classes of transcriptional activation domain.’, *Molecular and cellular biology*, 16(5), pp. 2044–2055. doi: 10.1128/MCB.16.5.2044.

Blekhman, R., Oshlack, A. and Gilad, Y. (2009) ‘Segmental duplications

contribute to gene expression differences between humans and chimpanzees', *Genetics*, 182(2), pp. 627–630. doi: 10.1534/genetics.108.099960.

Boccaletti, S. *et al.* (2006) 'Complex networks: Structure and dynamics', *Physics Reports*, pp. 175–308. doi: 10.1016/j.physrep.2005.10.009.

Boeke, J. D. and Chapman, K. B. (1991) 'Retrotransposition mechanisms', *Current Opinion in Cell Biology*, 3(3), pp. 502–507. doi: 10.1016/0955-0674(91)90079-E.

Boland, J. F. *et al.* (2013) 'The new sequencer on the block: Comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing', *Human Genetics*, 132(10), pp. 1153–1163. doi: 10.1007/s00439-013-1321-4.

Borgatti, S. P. (2005) 'Centrality and network flow', *Social Networks*, 27(1), pp. 55–71. doi: 10.1016/j.socnet.2004.11.008.

Borras, F. E. *et al.* (1995) 'Repression of I-A beta gene expression by the transcription factor PU.1', *The Journal of biological chemistry*, 270(41), pp. 24385–24391. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7592651.

Bosch, N. *et al.* (2007) 'Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events', *Human Molecular Genetics*, 16(21), pp. 2572–2582. doi: 10.1093/hmg/ddm209.

Boyd, J. L. *et al.* (2015) 'Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex', *Current Biology*, 25(6), pp. 772–779. doi: 10.1016/j.cub.2015.01.041.

Brawand, D. *et al.* (2011) 'The evolution of gene expression levels in mammalian organs', *Nature*, 478(7369), pp. 343–348. doi: 10.1038/nature10532.

Buckanovich, R. J., Posner, J. B. and Darnell, R. B. (1993) 'Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system', *Neuron*, 11(4), pp. 657–672. doi: 10.1016/0896-6273(93)90077-5.

Buermans, H. P. J. and den Dunnen, J. T. (2014) 'Next generation sequencing technology: Advances and applications', *Biochimica et Biophysica Acta*, 1842(10), pp. 1932–1941. doi: 10.1016/j.bbadis.2014.06.015.

Buljan, M., Frankish, A. and Bateman, A. (2010) 'Quantifying the mechanisms

of domain gain in animal proteins', *Genome Biology*, 11(7). doi: 10.1186/gb-2010-11-7-r74.

Burnette, W. N. (1981) '"Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A', *Analytical Biochemistry*, 112(2), pp. 195–203. doi: 10.1016/0003-2697(81)90281-5.

Busch, A. and Hertel, K. J. (2012) 'Evolution of SR protein and hnRNP splicing regulatory factors', *Wiley Interdisciplinary Reviews: RNA*, pp. 1–12. doi: 10.1002/wrna.100.

Byron, A. *et al.* (2012) 'Proteomic analysis of alpha4beta1 integrin adhesion complexes reveals alpha-subunit-dependent protein recruitment', *Proteomics*, 12(13), pp. 2107–2114. doi: 10.1002/pmic.201100487.

Cáceres, E. F. and Hurst, L. D. (2013) 'The evolution, impact and properties of exonic splice enhancers', *Genome Biology*, 14(12). doi: 10.1186/gb-2013-14-12-r143.

Cai, J. *et al.* (2008) 'De novo origination of a new protein-coding gene in Saccharomyces cerevisiae', *Genetics*, 179(1), pp. 487–496. doi: 10.1534/genetics.107.084491.

Cain, C. E. *et al.* (2011) 'Gene expression differences among primates are associated with changes in a histone epigenetic modification', *Genetics*, 187(4), pp. 1225–1234. doi: 10.1534/genetics.110.126177.

Callinan, a, Batzer, M. a. and Callinan, P. a (2006) 'Retrotransposable Elements and Human Disease', *Genome Dynamics*, 1, pp. 104–115. doi: 10.1159/000092503.

Cande, J. D., Chopra, V. S. and Levine, M. (2009) 'Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum', *Development*, 136(18), pp. 3153–3160. doi: 10.1242/dev.038034.

Caputi, M. and Zahler, A. M. (2001) 'Determination of the RNA Binding Specificity of the Heterogeneous Nuclear Ribonucleoprotein (hnRNP) H/H???/F/2H9 Family', *Journal of Biological Chemistry*, 276(47), pp. 43850–43859. doi: 10.1074/jbc.M102861200.

Cardoso-Moreira, M. *et al.* (2016) 'Evidence for the fixation of gene duplications

by positive selection in Drosophila', *Genome Research*, 26(6), pp. 787–798. doi: 10.1101/gr.199323.115.

Cartegni, L. *et al.* (2003) 'ESEfinder: A web resource to identify exonic splicing enhancers', *Nucleic Acids Research*, 31(13), pp. 3568–3571. doi: 10.1093/nar/gkg616.

Charizanis, K. *et al.* (2012) 'Muscleblind-like 2-Mediated Alternative Splicing in the Developing Brain and Dysregulation in Myotonic Dystrophy', *Neuron*, 75(3), pp. 437–450. doi: 10.1016/j.neuron.2012.05.029.

Chassé, H. *et al.* (2017) 'Analysis of translation using polysome profiling', *Nucleic acids research*, 45(3), p. e15. doi: 10.1093/nar/gkw907.

Chen, C. W. and Tanaka, M. (2018) 'Genome-wide Translation Profiling by Ribosome-Bound tRNA Capture', *Cell Reports*, 23(2), pp. 608–621. doi: 10.1016/j.celrep.2018.03.035.

Chen, C. Y. *et al.* (2014) 'Dissecting the human protein-protein interaction network via phylogenetic decomposition', *Scientific Reports*, 4. doi: 10.1038/srep07153.

Chen, L., Tovar-Corona, J. M. and Urrutia, A. O. (2012) 'Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome', *International Journal of Evolutionary Biology*, 2012, pp. 1–10. doi: 10.1155/2012/596274.

Chen, S., Krinsky, B. H. and Long, M. (2013) 'New genes as drivers of phenotypic evolution', *Nature Reviews Genetics*, pp. 645–660. doi: 10.1038/nrg3521.

Cheng, Z. *et al.* (2005) 'A genome-wide comparison of recent chimpanzee and human segmental duplications', *Nature*, 437(7055), pp. 88–93. doi: 10.1038/nature04000.

Choi, Y. L. *et al.* (2008) 'Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer.', *Cancer research*, 68(13), pp. 4971–6. doi: 10.1158/0008-5472.CAN-07-6158.

Chothia, C. *et al.* (2003) 'Evolution of the protein repertoire', *Science*, pp. 1701–1703. doi: 10.1126/science.1085371.

Clark, W. C. *et al.* (2016) 'tRNA base methylation identification and quantification via high-throughput sequencing', *RNA*, 22(11), pp. 1771–1784.

doi: 10.1261/rna.056531.116.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. doi: 10.1186/s13059-016-0881-8.

Consortium, Dia. G. R. A. M. (DIAGRAM) (2014) 'Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility.', *Nature Genetics*, 46(3), pp. 234–244. doi: 10.1038/ng.2897.

Consortium, E. P. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Consortium, I. H. G. S. (2001) 'Initial sequencing and analysis of the human genome.', *Nature*, 409, pp. 860–921. doi: http://dx.doi.org/10.1038/35057062.

Consortium, M. G. S. (2002) 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, 420(December), pp. 520–562. doi: 10.1038/nature01262.

Cook, C. E. *et al.* (2018) 'The European Bioinformatics Institute in 2017: Data coordination and integration', *Nucleic Acids Research*, 46(D1), pp. D21–D29. doi: 10.1093/nar/gkx1154.

Corbo, C., Orrù, S. and Salvatore, F. (2013) 'SRp20: An overview of its role in human diseases.', *Biochemical & Biophysical Research Communications*, 436(1), pp. 1–5. Available at: http://10.0.3.248/j.bbrc.2013.05.027%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=88987518&site=ehost-live.

Cordaux, R. and Batzer, M. A. (2009) 'The impact of retrotransposons on human genome evolution', *Nature Reviews Genetics*, pp. 691–703. doi: 10.1038/nrg2640.

Courseaux, A. and Nahon, J. L. (2001) 'Birth of two chimeric genes in the Hominidae lineage', *Science*, 291(5507), pp. 1293–1298. doi: 10.1126/science.1057284.

Cozen, A. E. *et al.* (2015) 'ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments', *Nature Methods*, 12(9), pp. 879–884. doi: 10.1038/nmeth.3508.

Crichton, J. H. *et al.* (2014) 'Defending the genome from the enemy within: Mechanisms of retrotransposon suppression in the mouse germline', *Cellular and Molecular Life Sciences*, pp. 1581–1605. doi: 10.1007/s00018-013-1468-0.

Crick, F. H. (1966) 'The genetic code--yesterday, today, and tomorrow.', *Cold Spring Harbor Symposia on Quantitative Biology*, 31, pp. 1–9. doi: 10.1101/SQB.1966.031.01.007.

Csárdi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal Complex Systems*, 1695, pp. 1–9. doi: 10.3724/SP.J.1087.2009.02191.

Cuccurese, M. *et al.* (2005) 'Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression', *Nucleic Acids Research*, 33(18), pp. 5965–5977. doi: 10.1093/nar/gki905.

Cuscó, I. *et al.* (2008) 'Copy number variation at the 7q11.23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion', *Genome Research*, 18(5), pp. 683–694. doi: 10.1101/gr.073197.107.

Darwin, C. (1859) *On the origin of species by means of natural selection*, *Darwin*. doi: 10.1016/S0262-4079(09)60380-8.

Darwin, C. (1968) 'On the origin of species by means of natural selection. 1859', *Murray*. Available at: http://www.robmacdougall.org/1805/h1805-18-origin-of-species.pdf%5Cnpapers3://publication/uuid/3190FE49-79AB-444B-832F-57F0CE03DA98.

Das, S., Dawson, N. L. and Orengo, C. A. (2015) 'Diversity in protein domain superfamilies', *Current Opinion in Genetics and Development*, pp. 40–49. doi: 10.1016/j.gde.2015.09.005.

Dasgupta, T. and Ladd, A. N. (2012) 'The importance of CELF control: Molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins', *Wiley Interdisciplinary Reviews: RNA*, pp. 104–121. doi: 10.1002/wrna.107.

Dehal, P. and Boore, J. L. (2005) 'Two rounds of whole genome duplication in the ancestral vertebrate', *PLoS Biology*, 3(10). doi: 10.1371/journal.pbio.0030314.

Deininger, P. (2011) 'Alu elements: know the SINEs', *Genome Biol*, 12(12), p. 236. doi: gb-2011-12-12-236 [pii]\r10.1186/gb-2011-12-12-236.

Deng, C. *et al.* (2010) 'Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict', *Proceedings of the National Academy of Sciences*, 107(50), pp. 21593–21598. doi: 10.1073/pnas.1007883107.

Dermitzakis, E. T. and Clark, A. G. (2002) 'Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover', *Molecular Biology and Evolution*, 19(7), pp. 1114–1121. doi: 10.1093/oxfordjournals.molbev.a004169.

Dever, T. E. and Green, R. (2012) 'The elongation, termination, and recycling phases of translation in eukaryotes', *Cold Spring Harbor Perspectives in Biology*, 4(7), pp. 1–16. doi: 10.1101/cshperspect.a013706.

Dinger, M. E. *et al.* (2008) 'Differentiating protein-coding and noncoding RNA: Challenges and ambiguities', *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1000176.

Ditzel, L. *et al.* (1998) 'Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT', *Cell*, pp. 125–138. doi: 10.1016/S0092-8674(00)81152-6.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dobzhansky, T. (1973) 'Nothing i n Biology Makes Sense Except i n the Light of Evolution', *The American Biology Teacher*. Available at: http://papodeprimata.com.br/wp-content/uploads/2016/02/Dobzhansky.pdf.

Doghman, M. *et al.* (2013) 'Integrative analysis of SF-1 transcription factor dosage impact on genome-wide binding and gene expression regulation', *Nucleic Acids Research*, 41(19), pp. 8896–8907. doi: 10.1093/nar/gkt658.

Dong, X. and Weng, Z. (2013) 'The correlation between histone modifications and gene expression', *Epigenomics*, pp. 113–116. doi: 10.2217/epi.13.13.

Doolittle, R. F. (1995) 'The Multiplicity of Domains in Proteins', *Annual Review of Biochemistry*, 64(1), pp. 287–314. doi: 10.1146/annurev.bi.64.070195.001443.

Duncan, C. D. S. and Mata, J. (2017) 'Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe /631/337 /631/337/574 /38/91 /38/39 article', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-10650-1.

Dunne, J. A., Williams, R. J. and Martinez, N. D. (2002) 'Food-web structure and network theory: The role of connectance and size', *Pnas*, 99(20), pp. 12917–12922. doi: 10.1073/pnas.192407699.

DuRose, J. B. *et al.* (2009) 'Phosphorylation of eukaryotic translation initiation

factor 2alpha coordinates rRNA transcription and translation inhibition during endoplasmic reticulum stress.', *Molecular and cellular biology*, 29(15), pp. 4295–307. doi: 10.1128/MCB.00260-09.

Durrens, P., Nikolski, M. and Sherman, D. (2008) 'Fusion and fission of genes define a metric between fungal genomes', *PLoS Computational Biology*, 4(10). doi: 10.1371/journal.pcbi.1000200.

Eddy, S. (1998) 'Profile hidden Markov models.', *Bioinformatics*, 14(9), pp. 755–763. doi: btb114 [pii].

Edelstein, C., Pfaffinger, D. and Scanu, a M. (1984) 'Advantages and limitations of density gradient ultracentrifugation in the fractionation of human serum lipoproteins: role of salts and sucrose.', *Journal of lipid research*, 25, pp. 630–637.

Eden, E. *et al.* (2009) 'GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists', *BMC Bioinformatics*, 10. doi: 10.1186/1471-2105-10-48.

Edgar, R. C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Edgren, H. *et al.* (2011) 'Identification of fusion genes in breast cancer by paired-end RNA-sequencing', *Genome Biology*, 12(1). doi: 10.1186/gb-2011-12-1-r6.

Eicher, E. M. *et al.* (1976) 'Evolution of mammalian carbonic anhydrase loci by tandem duplication: Close linkage of Car-1 and Car-2 to the centromere region of chromosome 3 of the mouse', *Biochemical Genetics*, 14(7–8), pp. 651–660. doi: 10.1007/BF00485843.

Eissenberg, J. C. and Elgin, S. C. (2014) 'Heterochromatin and Euchromatin', in *eLS*. doi: 10.1002/9780470015902.a0001164.pub3.

Enard, W. (2015) 'Human evolution: Enhancing the brain', *Current Biology*, 25(10), pp. R421-RR423. doi: 10.1016/j.cub.2015.03.031.

Engström, P. G. *et al.* (2013) 'Systematic evaluation of spliced alignment programs for RNA-seq data', *Nature Methods*, 10(12), pp. 1185–1191. doi: 10.1038/nmeth.2722.

Everaert, C. *et al.* (2017) 'Benchmarking of RNA-sequencing analysis

workflows using whole-transcriptome RT-qPCR expression data', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-01617-3.

Fairbrother, W. G. *et al.* (2004) 'RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons', *Nucleic Acids Research*, 32(WEB SERVER ISS.). doi: 10.1093/nar/gkh393.

Fang, X. *et al.* (2014) 'The zinc finger transcription factor ZFX Is required for maintaining the tumorigenic potential of glioblastoma stem cells', *Stem Cells*, 32(8), pp. 2033–2047. doi: 10.1002/stem.1730.

Farnaes, L. *et al.* (2018) 'Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization', *npj Genomic Medicine*, 3(1). doi: 10.1038/s41525-018-0049-4.

Farrel, A., Murphy, J. and Guo, J. (2016) 'Structure-based prediction of transcription factor binding specificity using an integrative energy function', *Bioinformatics* , 32(12), pp. i306–i313. doi: 10.1093/bioinformatics/btw264.

Fay, J. C. and Wu, C. I. (2003) 'Sequence divergence, functional constraint, and selection in protein evolution', *Annual Review of Genomics and Human Genetics*, 4(1), pp. 213–235. doi: 10.1146/annurev.genom.4.020303.162528.

Fears, S. *et al.* (1996) 'Intergenic splicing of MDS1 and EVI1 occurs in normal tissues as well as in myeloid leukemia and produces a new member of the PR domain family.', *Proceedings of the National Academy of Sciences of the United States of America*, 93(4), pp. 1642–1647. doi: 10.1073/pnas.93.4.1642.

Feng, Y. and Bankston, A. (2010) 'The STAR family member: QKI and cell signaling', *Advances in Experimental Medicine and Biology*, pp. 25–36. doi: 10.1007/978-1-4419-7005-3_2.

Finn, R. D. *et al.* (2000) 'Escherichia coli RNA polymerase core and holoenzyme structures', *EMBO Journal*, 19(24), pp. 6833–6844. doi: 10.1093/emboj/19.24.6833.

Finn, R. D. *et al.* (2014) 'Pfam: The protein families database', *Nucleic Acids Research*. doi: 10.1093/nar/gkt1223.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) 'HMMER web server: Interactive sequence similarity searching', *Nucleic Acids Research*, 39(SUPPL. 2). doi: 10.1093/nar/gkr367.

Fiori, L. M., Gross, J. A. and Turecki, G. (2012) 'Effects of histone

modifications on increased expression of polyamine biosynthetic genes in suicide', *International Journal of Neuropsychopharmacology*, 15(8), pp. 1161–1166. doi: 10.1017/S1461145711001520.

Foster, P. G. (2004) 'Modeling compositional heterogeneity', *Systematic Biology*, 53(3), pp. 485–495. doi: 10.1080/10635150490445779.

França, G. S. *et al.* (2017) 'Unveiling the impact of the genomic architecture on the evolution of vertebrate microRNAs', *Frontiers in Genetics*. doi: 10.3389/fgene.2017.00034.

Franchini, L. F. and Pollard, K. S. (2015) 'Genomic approaches to studying human-specific developmental traits', *Development*, 142(18), pp. 3100–3112. doi: 10.1242/dev.120048.

Freeman, L. C. (1977) 'A Set of Measures of Centrality Based on Betweenness', *Sociometry*, 40(1), p. 35. doi: 10.2307/3033543.

French, C. A. *et al.* (2008) 'BRD-NUT oncoproteins: A family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells', *Oncogene*, 27(15), pp. 2237–2242. doi: 10.1038/sj.onc.1210852.

Fu, X. D. (1993) 'Specific commitment of different pre-mRNAs to splicing by single SR proteins', *Nature*, 365(6441), pp. 82–85. doi: 10.1038/365082a0.

Fu, X. D. and Ares, M. (2014) 'Context-dependent control of alternative splicing by RNA-binding proteins', *Nature Reviews Genetics*, pp. 689–701. doi: 10.1038/nrg3778.

Fu, X. D. and Maniatis, T. (1992) 'The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site.', *Proceedings of the National Academy of Sciences of the United States of America*, 89(5), pp. 1725–9. doi: 10.1073/pnas.89.5.1725.

Galtier, N. and Duret, L. (2007) 'Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution', *Trends in Genetics*, 23(6), pp. 273–277. doi: 10.1016/j.tig.2007.03.011.

Gaspari, M., Larsson, N. G. and Gustafsson, C. M. (2004) 'The transcription machinery in mammalian mitochondria', in *Biochimica et Biophysica Acta - Bioenergetics*, pp. 148–152. doi: 10.1016/j.bbabio.2004.10.003.

Gaudet, P. and Dessimoz, C. (2017) 'Gene ontology: Pitfalls, biases, and remedies', in *Methods in Molecular Biology*, pp. 189–205. doi: 10.1007/978-1-4939-3743-1_14.

Gavrilets, S. and Vose, A. (2005) 'Dynamic patterns of adaptive radiation', *Proceedings of the National Academy of Sciences*, 102(50), pp. 18040–18045. doi: 10.1073/pnas.0506330102.

Ge, Y. and Porse, B. T. (2014) 'The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression', *BioEssays*, 36(3), pp. 236–243. doi: 10.1002/bies.201300156.

Gerashchenko, M. V, Lobanov, A. V and Gladyshev, V. N. (2012) 'Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. TL - 109', *Proceedings of the National Academy of Sciences of the United States of America*, 109 VN-(43), pp. 17394–17399. doi: 10.1073/pnas.1120799109.

Gerstein, M. (1998) 'How representative are the known structures of the proteins in a complete genome? A comprehensive structural census', *Folding and Design*, 3(6), pp. 497–512. doi: 10.1016/S1359-0278(98)00066-2.

Geuens, T., Bouhy, D. and Timmerman, V. (2016) 'The hnRNP family: insights into their role in health and disease', *Human Genetics*, pp. 851–867. doi: 10.1007/s00439-016-1683-5.

Giannuzzi, G. *et al.* (2013) 'Evolutionary dynamism of the primate LRRC37 gene family', *Genome Research*, 23(1), pp. 46–59. doi: 10.1101/gr.138842.112.

Gil, A. *et al.* (1991) 'Characterization of cDNAs encoding the polypyrimidine tract-binding protein', *Genes and Development*, 5(7), pp. 1224–1236. doi: 10.1101/gad.5.7.1224.

Girton, J. R. and Johansen, K. M. (2008) 'Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila.', in *Advances in genetics*, pp. 1–43. doi: 10.1016/S0065-2660(07)00001-6.

Glassford, W. J. and Rebeiz, M. (2013) 'Assessing constraints on the path of regulatory sequence evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632), pp. 20130026–20130026. doi: 10.1098/rstb.2013.0026.

Glémin, S. *et al.* (2015) 'Quantification of GC-biased gene conversion in the

human genome', *Genome Research*, 25(8), pp. 1215–1228. doi: 10.1101/gr.185488.114.

Gobet, C. and Naef, F. (2017) 'Ribosome profiling and dynamic regulation of translation in mammals', *Current Opinion in Genetics and Development*, pp. 120–127. doi: 10.1016/j.gde.2017.03.005.

Gonzalez, C. *et al.* (2014) 'Ribosome Profiling Reveals a Cell-Type-Specific Translational Landscape in Brain Tumors', *Journal of Neuroscience*, 34(33), pp. 10924–10936. doi: 10.1523/JNEUROSCI.0084-14.2014.

Gould, S. J. and Lewontin, R. C. (1979) 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme', *Proceedings of the Royal Society B: Biological Sciences*, 205(1161), pp. 581–598. doi: 10.1098/rspb.1979.0086.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building', *Molecular Biology and Evolution*, 27(2), pp. 221–224. doi: 10.1093/molbev/msp259.

Graham, T. and Boissinot, S. (2006) 'The genomic distribution of L1 elements: The role of insertion bias and natural selection', *Journal of Biomedicine and Biotechnology*. doi: 10.1155/JBB/2006/75327.

Granovetter, M. and Granovetter, M. (1983) 'The Strength of Weak Ties: A Network Theory Revisited', *Sociological Theory*, 1, pp. 201–233. doi: 10.2307/202051.

Grant, S. G. N. (2016) 'The molecular evolution of the vertebrate behavioural repertoire', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1685), p. 20150051. doi: 10.1098/rstb.2015.0051.

De Grassi, A., Lanave, C. and Saccone, C. (2008) 'Genome duplication and gene-family evolution: The case of three OXPHOS gene families', *Gene*, pp. 1–6. doi: 10.1016/j.gene.2008.05.011.

Griffiths, A., Miller, J. and Suzuki, D. (2000) *Introduction to Genetic Analysis. 7th edition.*, *Transcription: an overview of gene regulation in eukaryotes*.

Grosso, A. R. *et al.* (2008) 'Tissue-specific splicing factor gene expression signatures', *Nucleic Acids Research*, 36(15), pp. 4823–4832. doi: 10.1093/nar/gkn463.

Gunnelius, L. *et al.* (2014) 'The omega subunit of the RNA polymerase core directs transcription efficiency in cyanobacteria', *Nucleic Acids Research*, 42(7), pp. 4606–4614. doi: 10.1093/nar/gku084.

Guschanski, K., Warnefors, M. and Kaessmann, H. (2017) 'The evolution of duplicate gene expression in mammalian organs', *Genome Research*, 27(9), pp. 1461–1474. doi: 10.1101/gr.215566.116.

Hagberg, A. a, Schult, D. a and Swart, P. J. (2008) 'Exploring network structure, dynamics, and function using NetworkX', *Network*, 836(SciPy), pp. 11–15.

Hall, B. G. (2013) 'Building phylogenetic trees from molecular data with MEGA', *Molecular Biology and Evolution*, 30(5), pp. 1229–1235. doi: 10.1093/molbev/mst012.

Han, J. H. *et al.* (2007) 'The folding and evolution of multidomain proteins', *Nature Reviews Molecular Cell Biology*, pp. 319–330. doi: 10.1038/nrm2144.

Han, M. V. *et al.* (2009) 'Adaptive evolution of young gene duplicates in mammals', *Genome Research*, 19(5), pp. 859–867. doi: 10.1101/gr.085951.108.

Harrow, J. *et al.* (2012) 'GENCODE: The reference human genome annotation for the ENCODE project', *Genome Research*, 22(9), pp. 1760–1774. doi: 10.1101/gr.135350.111.

Head, S. R. *et al.* (2014) 'Library construction for next-generation sequencing: Overviews and challenges', *BioTechniques*, 56(2), pp. 61–77. doi: 10.2144/000114133.

Heiman, M. *et al.* (2008) 'A Translational Profiling Approach for the Molecular Characterization of CNS Cell Types', *Cell*, 135(4), pp. 738–748. doi: 10.1016/j.cell.2008.10.028.

Heiman, M. *et al.* (2014) 'Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP)', *Nature Protocols*, 9(6), pp. 1282–1291. doi: 10.1038/nprot.2014.085.

Hendrickson, S. L. *et al.* (2010) 'Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression', *PLoS ONE*, 5(9), pp. 1–8. doi: 10.1371/journal.pone.0012862.

Hernando-Herraez, I. *et al.* (2015) 'DNA Methylation: Insights into Human Evolution', *PLoS Genetics*. doi: 10.1371/journal.pgen.1005661.

Herrero, J. *et al.* (2016) 'Ensembl comparative genomics resources', *Database*,

2016. doi: 10.1093/database/bav096.

Hickman, M. A. and Rusche, L. N. (2010) 'Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication', *Proceedings of the National Academy of Sciences*, 107(45), pp. 19384–19389. doi: 10.1073/pnas.1006436107.

Hinnebusch, A. G. (2014) 'The Scanning Mechanism of Eukaryotic Translation Initiation', *Annual Review of Biochemistry*, 83(1), pp. 779–812. doi: 10.1146/annurev-biochem-060713-035802.

Ho-Huu, J. *et al.* (2012) 'Contrasted patterns of selective pressure in three recent paralogous gene pairs in the Medicago genus (L.)', *BMC Evolutionary Biology*, 12(1). doi: 10.1186/1471-2148-12-195.

Ho, L. and Crabtree, G. R. (2010) 'Chromatin remodelling during development', *Nature*, pp. 474–484. doi: 10.1038/nature08911.

Hou, Z. C. *et al.* (2012) 'Elephant transcriptome provides insights into the evolution of eutherian placentation', *Genome Biology and Evolution*, 4(5), pp. 713–725. doi: 10.1093/gbe/evs045.

Hrdy, I. *et al.* (2004) 'Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I', *Nature*, 432(7017), pp. 618–622. doi: 10.1038/nature03149.

Huang, D. W. *et al.* (2007) 'The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists', *Genome Biology*, 8(9). doi: 10.1186/gb-2007-8-9-r183.

Huang, Y. and Steitz, J. A. (2001) 'Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA', *Molecular Cell*, 7(4), pp. 899–905. doi: 10.1016/S1097-2765(01)00233-7.

Huber, W. *et al.* (2007) 'Graphs in molecular biology', *BMC Bioinformatics*. doi: 10.1186/1471-2105-8-S6-S8.

Huelsenbeck, J. P. and Rannala, B. (2004) 'Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models', *Systematic Biology*, 53(6), pp. 904–913. doi: 10.1080/10635150490522629.

Hughes, T. R. (2009) '"Validation" in genome-scale research', *Journal of Biology*. doi: 10.1186/jbiol104.

Hurst, L. D. and Pál, C. (2001) 'Evidence for purifying selection acting on silent sites in BRCA1', *Trends in Genetics*, pp. 62–65. doi: 10.1016/S0168-9525(00)02173-9.

IGV (Integrative Genomic Viewer) (2013) 'Integrative Genomics Viewer', *Broad Institute*, 29(1), pp. 24–26. doi: 10.1038/nbt0111-24.

Imashimizu, M. *et al.* (2014) 'Transcription elongation. Heterogeneous tracking of RNA polymerase and its biological implications', *Transcription*, p. e28285. doi: 10.4161/trns.28285.

Ingolia, N. T. *et al.* (2012) 'The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments', *Nature Protocols*, 7(8), pp. 1534–1550. doi: 10.1038/nprot.2012.086.

Ingolia, N. T. (2016) 'Ribosome Footprint Profiling of Translation throughout the Genome', *Cell*, 165(1), pp. 22–33. doi: 10.1016/j.cell.2016.02.066.

Inukai, S., Kock, K. H. and Bulyk, M. L. (2017) 'Transcription factor–DNA binding: beyond binding site motifs', *Current Opinion in Genetics and Development*, pp. 110–119. doi: 10.1016/j.gde.2017.02.007.

Ishibashi, T. *et al.* (2014) 'Transcription factors IIS and IIF enhance transcription efficiency by differentially modifying RNA polymerase pausing dynamics', *Proceedings of the National Academy of Sciences*, 111(9), pp. 3419–3424. doi: 10.1073/pnas.1401611111.

Jachiet, P. A. *et al.* (2013) 'MosaicFinder: Identification of fused gene families in sequence similarity networks', *Bioinformatics*, 29(7), pp. 837–844. doi: 10.1093/bioinformatics/btt049.

Jachiet, P. A. *et al.* (2014) 'Extensive gene remodeling in the viral world: New evidence for nongradual evolution in the mobilome network', *Genome Biology and Evolution*, 6(9), pp. 2195–2205. doi: 10.1093/gbe/evu168.

Jackson, R. and Standart, N. (2015) 'The awesome power of ribosome profiling', *RNA*, 21(4), pp. 652–654. doi: 10.1261/rna.049908.115.

Jacob, F. and Monod, J. (1961) 'On the Regulation of Gene Activity', *Cold Spring Harbor Symposia on Quantitative Biology*, 26(0), pp. 193–211. doi: 10.1101/SQB.1961.026.01.024.

Jaenisch, R. and Bird, A. (2003) 'Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals', *Nature Genetics*, pp.

245–254. doi: 10.1038/ng1089.

Jain, M. *et al.* (2016) 'The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community', *Genome Biology*, 17(1), p. 239. doi: 10.1186/s13059-016-1103-0.

Jayaswal, P. K. *et al.* (2017) 'A tree of life based on ninety-eight expressed genes conserved across diverse eukaryotic species', *PLoS ONE*, 12(9). doi: 10.1371/journal.pone.0184276.

Jensen, K. B. *et al.* (2000) 'Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability', *Neuron*, 25(2), pp. 359–371. doi: 10.1016/S0896-6273(00)80900-9.

Jenuwein, T. and Allis, C. D. (2001) 'Translating the histone code', *Science*, pp. 1074–1080. doi: 10.1126/science.1063127.

Jiang, S. Y. and Ramachandran, S. (2016) 'Expansion Mechanisms and Evolutionary History on Genes Encoding DNA Glycosylases and Their Involvement in Stress and Hormone Signaling', *Genome biology and evolution*, 8(4), pp. 1165–1184. doi: 10.1093/gbe/evw067.

Johnson, M. E. *et al.* (2001) 'Positive selection of a gene family during the emergence of humans and African apes', *Nature*, 413(6855), pp. 514–519. doi: 10.1038/35097067.

Jones, F. C. *et al.* (2012) 'The genomic basis of adaptive evolution in threespine sticklebacks', *Nature*, 484(7392), pp. 55–61. doi: 10.1038/nature10944.

Jonnalagadda, S. and Srinivasan, R. (2014) 'An efficient graph theory based method to identify every minimal reaction set in a metabolic network', *BMC Systems Biology*, 8(1). doi: 10.1186/1752-0509-8-28.

Journal, T. and Society, T. A. (1996) 'The Diversity of BCR-ABL Fusion Proteins and Their Relationship to Leukemia Phenotype', *Blood*, 88(7), pp. 1697–1702. doi: 10.1007/s007830050018.The.

Kaessmann, H. (2010) 'Origins, evolution, and phenotypic impact of new genes', *Genome Research*, pp. 1313–1326. doi: 10.1101/gr.101386.109.

Kapranov, P., Willingham, A. T. and Gingeras, T. R. (2007) 'Genome-wide transcription and the implications for genomic organization', *Nature Reviews Genetics*, pp. 413–423. doi: 10.1038/nrg2083.

Karlin, S. and Mrázek, J. (1996) 'What drives codon choices in human genes?',

Journal of Molecular Biology, 262(4), pp. 459–472. doi: 10.1006/jmbi.1996.0528.

Karolchik, D., Hinrichs, A. S. and Kent, W. J. (2011) 'The UCSC genome browser', *Current Protocols in Human Genetics*, (SUPPL. 71). doi: 10.1002/0471142905.hg1806s71.

Karp, R. M. (1971) 'REDUCIBILITY AMONG COMBINATORIAL PROBLEMS t Richard M. Karp University of California at Berkeley', *Science*, pp. 85–103. doi: 10.1007/978-1-4684-2001-2_9.

Katzman, S. *et al.* (2011) 'Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots', *Genome Biology and Evolution*, 3(1), pp. 614–626. doi: 10.1093/gbe/evr058.

Kawasaki, K., Buchanan, A. V. and Weiss, K. M. (2007) 'Gene duplication and the evolution of vertebrate skeletal mineralization', *Cells Tissues Organs*, 186(1), pp. 7–24. doi: 10.1159/000102678.

Kawashima, T. *et al.* (2009) 'Domain shuffling and the evolution of vertebrates', *Genome Research*, 19(8), pp. 1393–1403. doi: 10.1101/gr.087072.108.

Kchouk, M., Gibrat, J. F. and Elloumi, M. (2017) 'Generations of Sequencing Technologies: From First to Next Generation', *Biology and Medicine*, 09(03). doi: 10.4172/0974-8369.1000395.

Kellis, M. *et al.* (2003) 'Sequencing and comparison of yeast species to identify genes and regulatory elements', *Nature*, 423(6937), pp. 241–254. doi: 10.1038/nature01644.

Keren, H., Lev-Maor, G. and Ast, G. (2010) 'Alternative splicing and evolution: Diversification, exon definition and function', *Nature Reviews Genetics*, pp. 345–355. doi: 10.1038/nrg2776.

Khaitovich Muetzel, B.,She, X.,Lachmann, M., Hellmann, I.,Dietzsch, J.,Steigele, S.,Do, H-H.,Weiss, G., P. and Enard Heissig, F.,Arendt, T.,Nieselt-Struwe, K., Eichler, E.E.,and Paabo, S., W. (2004) 'Regional Patterns of Gene Expression in Human\nand Chimpanzee Brains', *Genome Research*, 14, pp. 1462–1473.

Khan, A. *et al.* (2018) 'JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework', *Nucleic Acids Research*, 46(D1), pp. D260–D266. doi: 10.1093/nar/gkx1126.

Khurana, E. *et al.* (2010) 'Segmental duplications in the human genome reveal details of pseudogene formation', *Nucleic Acids Research*, 38(20), pp. 6997–7007. doi: 10.1093/nar/gkq587.

Kim, E., Magen, A. and Ast, G. (2007) 'Different levels of alternative splicing among eukaryotes', *Nucleic Acids Research*, 35(1), pp. 125–131. doi: 10.1093/nar/gkl924.

Kim, N. *et al.* (2007) 'The ASAP II database: Analysis and comparative genomics of alternative splicing in 15 animal species', *Nucleic Acids Research*, 35(SUPPL. 1). doi: 10.1093/nar/gkl884.

Kim, T. H. and Ren, B. (2006) 'Genome-Wide Analysis of Protein-DNA Interactions', *Annual Review of Genomics and Human Genetics*, 7(1), pp. 81–102. doi: 10.1146/annurev.genom.7.080505.115634.

Kimura, M. (1989) 'The neutral theory of molecular evolution and the world view of neutralists', *Genome*, 31, pp. 24–31. doi: 10.1139/g89-009.

Kleene, K. C. *et al.* (2010) 'Quantitative analysis of mRNA translation in mammalian spermatogenic cells with sucrose and Nycodenz gradients', *Reproductive Biology and Endocrinology*, 8. doi: 10.1186/1477-7827-8-155.

Knowles, D. G. and McLysaght, A. (2009) 'Recent de novo origin of human protein-coding genes', *Genome Research*, 19(10), pp. 1752–1759. doi: 10.1101/gr.095026.109.

Ko, J. Y., Oh, S. and Yoo, K. H. (2017) 'Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development', *Mol. Cells*, 40(3), pp. 169–177. doi: 10.14348/molcells.2017.0033.

Kolfschoten, G. M. *et al.* (2003) 'TWE-PRIL; a fusion protein of TWEAK and APRIL', in *Biochemical Pharmacology*, pp. 1427–1432. doi: 10.1016/S0006-2952(03)00493-3.

Kolkman, J. A. and Stemmer, W. P. C. (2001) 'Directed evolution of proteins by exon shuffling', *Nature Biotechnology*, pp. 423–428. doi: 10.1038/88084.

Kolovos, P. *et al.* (2012) 'Enhancers and silencers: An integrated and simple model for their function', *Epigenetics and Chromatin*. doi: 10.1186/1756-8935-5-1.

Konieczny, P., Stepniak-Konieczna, E. and Sobczak, K. (2014) 'MBNL proteins and their target RNAs, interaction and splicing regulation', *Nucleic Acids*

Research, pp. 10873–10887. doi: 10.1093/nar/gku767.

Koonin, E. V. (2009) 'Evolution of genome architecture', *International Journal of Biochemistry and Cell Biology*, pp. 298–306. doi: 10.1016/j.biocel.2008.09.015.

Koonin, E. V., Wolf, Y. I. and Karev, G. P. (2002) 'The structure of the protein universe and genome evolution', *Nature*, pp. 218–223. doi: 10.1038/nature01256.

Korkhin, Y. *et al.* (2009) 'Evolution of complex RNA polymerases: The complete archaeal RNA polymerase structure', *PLoS Biology*, 7(5). doi: 10.1371/journal.pbio.1000102.

Kornblihtt, A. R. *et al.* (2004) 'Multiple links between transcription and splicing', *RNA*, pp. 1489–1498. doi: 10.1261/rna.7100104.

Kotlar, D. and Lavner, Y. (2006) 'The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids', *BMC Genomics*, 7. doi: 10.1186/1471-2164-7-67.

Kozu, T., Henrich, B. and Schäfer, K. P. (1995) 'Structure and expression of the gene (HNRPA2B1) encoding the human hnRNP protein A2/B1', *Genomics*, 25(2), pp. 365–371. doi: 10.1016/0888-7543(95)80035-K.

Krikun, G. *et al.* (2000) 'Regulation of tissue factor gene expression in human endometrium by transcription factors Sp1 and Sp3.', *Molecular endocrinology (Baltimore, Md.)*, 14(3), pp. 393–400. doi: 10.1210/mend.14.3.0430.

Kringelbach, M. L. *et al.* (2007) 'Translational principles of deep brain stimulation', *Nature Reviews Neuroscience*, pp. 623–635. doi: 10.1038/nrn2196.

Krishnamurthy, S. and Hampsey, M. (2009) 'Eukaryotic transcription initiation', *Current Biology*. doi: 10.1016/j.cub.2008.11.052.

Kronenberg, ZN., Fiddes, IT., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, OS., Eichler, E. (2018) 'High-resolution comparative analysis of great ape genomes', *Science*, 360(6393).

Kuhlman, T. C. *et al.* (1999) 'The General Transcription Factors IIA, IIB, IIF, and IIE Are Required for RNA Polymerase II Transcription from the Human U1 Small Nuclear RNA Promoter', *Molecular and Cellular Biology*, 19(3), pp. 2130–2141. doi: 10.1128/MCB.19.3.2130.

Łabaj, P. P. and Kreil, D. P. (2016) 'Sensitivity, specificity, and reproducibility

of RNA-Seq differential expression calls', *Biology Direct*, 11(1). doi: 10.1186/s13062-016-0169-7.

Lamichhaney, S. *et al.* (2015) 'Evolution of Darwin's finches and their beaks revealed by genome sequencing', *Nature*, 518(7539), pp. 371–375. doi: 10.1038/nature14181.

Lan, X. and Pritchard, J. K. (2016) 'Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals', *Science*, 352(6288), pp. 1009–1013. doi: 10.1126/science.aad8411.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Längst, G. and Manelyte, L. (2015) 'Chromatin remodelers: From function to dysfunction', *Genes*, pp. 299–324. doi: 10.3390/genes6020299.

Laqqan, M. and Hammadeh, M. E. (2018) 'Alterations in DNA methylation patterns and gene expression in spermatozoa of subfertile males', *Andrologia*, 50(3). doi: 10.1111/and.12934.

Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Lartillot, N. *et al.* (2013) 'Phylobayes mpi: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment', *Systematic Biology*, 62(4), pp. 611–615. doi: 10.1093/sysbio/syt022.

Lasham, A. *et al.* (2003) 'The Y-box-binding protein, YB1, is a potential negative regulator of the p53 tumor suppressor', *Journal of Biological Chemistry*, 278(37), pp. 35516–35523. doi: 10.1074/jbc.M303920200.

Latchman, D. S. (2001) 'Transcription factors: Bound to activate or repress', *Trends in Biochemical Sciences*, pp. 211–213. doi: 10.1016/S0968-0004(01)01812-6.

Latysheva, N. S. *et al.* (2016) 'Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer', *Molecular Cell*, 63(4), pp. 579–592. doi: 10.1016/j.molcel.2016.07.008.

Lee, K. Y. *et al.* (2013) 'Compound loss of muscleblind-like function in myotonic dystrophy', *EMBO Molecular Medicine*, 5(12), pp. 1887–1900. doi: 10.1002/emmm.201303275.

Lee, Y. *et al.* (2007) 'ECgene: An alternative splicing database update', *Nucleic*

Acids Research, 35(SUPPL. 1). doi: 10.1093/nar/gkl992.

Lees, J. G., Ranea, J. A. and Orengo, C. A. (2015) 'Identifying and characterising key alternative splicing events in Drosophila development', *BMC Genomics*, 16(1). doi: 10.1186/s12864-015-1674-2.

Leinonen, R., Sugawara, H. and Shumway, M. (2011) 'The sequence read archive', *Nucleic Acids Research*, 39(SUPPL. 1). doi: 10.1093/nar/gkq1019.

Leonard, G. and Richards, T. A. (2012) 'Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life', *Proceedings of the National Academy of Sciences*, 109(52), pp. 21402–21407. doi: 10.1073/pnas.1210909110.

Levine, M. T. *et al.* (2006) 'Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression', *Proceedings of the National Academy of Sciences*, 103(26), pp. 9935–9939. doi: 10.1073/pnas.0509809103.

Li, B., Carey, M. and Workman, J. L. (2007) 'The Role of Chromatin during Transcription', *Cell*, pp. 707–719. doi: 10.1016/j.cell.2007.01.015.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, W. H., Tanimura, M. and Sharp, P. M. (1987) 'An evaluation of the molecular clock hypothesis using mammalian DNA sequences.', *Journal of molecular evolution*, 25(4), pp. 330–342. doi: 10.1007/BF02603118.

Lim, J. H. *et al.* (2006) 'ESE-3 transcription factor is involved in the expression of death receptor (DR)-5 through putative Ets sites', *Biochemical and Biophysical Research Communications*, 350(3), pp. 736–741. doi: 10.1016/j.bbrc.2006.09.102.

Linn, D. E. *et al.* (2016) 'Deletion of interstitial genes between TMPRSS2 and ERG promotes prostate cancer progression', *Cancer Research*, 76(7), pp. 1869–1881. doi: 10.1158/0008-5472.CAN-15-1911.

Liu, S. *et al.* (2011) 'A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species', *Nucleic Acids Research*, 39(2), pp. 578–588. doi: 10.1093/nar/gkq817.

Liu, Y., Beyer, A. and Aebersold, R. (2016) 'On the Dependency of Cellular Protein Levels on mRNA Abundance', *Cell*, pp. 535–550. doi:

10.1016/j.cell.2016.03.014.

Liu, Y., Maskell, D. L. and Schmidt, B. (2009) 'CUDASW++: Optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units', *BMC Research Notes*, 2. doi: 10.1186/1756-0500-2-73.

Livak, K. J. and Schmittgen, T. D. (2001) 'Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.', *Methods (San Diego, Calif.)*, 25(4), pp. 402–8. doi: 10.1006/meth.2001.1262.

Loayza-Puch, F. *et al.* (2013) 'p53 induces transcriptional and translational programs to suppress cell proliferation and growth', *Genome Biology*, 14(4). doi: 10.1186/gb-2013-14-4-r32.

Long, J. C. and Caceres, J. F. (2009) 'The SR protein family of splicing factors: master regulators of gene expression', *Biochemical Journal*, 417(1), pp. 15–27. doi: 10.1042/BJ20081501.

Long, M. (2000) 'A new function evolved from gene fusion', *Genome Research*, pp. 1655–1657. doi: 10.1101/gr.165700.

Long, M. *et al.* (2003) 'The origin of new genes: Glimpses from the young and old', *Nature Reviews Genetics*, pp. 865–875. doi: 10.1038/nrg1204.

Long, M. and Langley, C. (1993) 'Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila', *Science*, 260(5104), pp. 91–95. doi: 10.1126/science.7682012.

López-Maury, L., Marguerat, S. and Bähler, J. (2008) 'Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation', *Nature Reviews Genetics*, pp. 583–593. doi: 10.1038/nrg2398.

Lorberbaum, D. S. and Barolo, S. (2013) 'Gene regulation: When analog beats digital', *Current Biology*, 23(23). doi: 10.1016/j.cub.2013.10.004.

Lorente-Galdos, B. *et al.* (2013) 'Accelerated exon evolution within primate segmental duplications', *Genome Biology*, 14(1). doi: 10.1186/gb-2013-14-1-r9.

Loughran, N. B. *et al.* (2012) 'Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase', *Molecular Biology and Evolution*, 29(8), pp. 2039–2046. doi: 10.1093/molbev/mss073.

Löytynoja, A. and Goldman, N. (2010) 'WebPRANK: A phylogeny-aware multiple sequence aligner with interactive alignment browser', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-579.

Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', *Genomics, Proteomics and Bioinformatics*, pp. 265–279. doi: 10.1016/j.gpb.2016.05.004.

Lupski, J. R. (1998) 'Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits', *Trends in Genetics*, pp. 417–422. doi: 10.1016/S0168-9525(98)01555-8.

Luse, D. S. (2013) 'Promoter clearance by RNA polymerase II', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, pp. 63–68. doi: 10.1016/j.bbagrm.2012.08.010.

Lynch, M. (2007) 'The frailty of adaptive hypotheses for the origins of organismal complexity', *Proceedings of the National Academy of Sciences*, 104(Supplement 1), pp. 8597–8604. doi: 10.1073/pnas.0702207104.

Lynch, M. and Conery, J. S. (2003) 'The Origins of Genome Complexity', *Science*, 302(5649), pp. 1401–1404. doi: 10.1126/science.1089370.

Macadangdang, B. R. *et al.* (2014) 'Evolution of histone 2A for chromatin compaction in eukaryotes', *eLife*, 2014(3). doi: 10.7554/eLife.02792.

Mack, K. L., Campbell, P. and Nachman, M. W. (2016) 'Gene regulation and speciation in house mice', *Genome Research*, 26(4), pp. 451–461. doi: 10.1101/gr.195743.115.

MacKintosh, C. and Ferrier, D. E. K. (2017) 'Recent advances in understanding the roles of whole genome duplications in evolution', *F1000Research*, 6, p. 1623. doi: 10.12688/f1000research.11792.1.

Magistri, M. *et al.* (2012) 'Regulation of chromatin structure by long noncoding RNAs: Focus on natural antisense transcripts', *Trends in Genetics*, pp. 389–396. doi: 10.1016/j.tig.2012.03.013.

Mahdi, L. K. *et al.* (2013) 'A Transcription Factor Contributes to Pathogenesis and Virulence in Streptococcus pneumoniae', *PLoS ONE*, 8(8). doi: 10.1371/journal.pone.0070862.

Makino, T. and Gojobori, T. (2007) 'Evolution of protein-protein interaction network', *Genome Dynamics*, pp. 13–29. doi: 10.1159/000107601.

Makino, T. and McLysaght, A. (2008) 'Interacting gene clusters and the evolution of the vertebrate immune system', *Molecular Biology and Evolution*, 25(9), pp. 1855–1862. doi: 10.1093/molbev/msn137.

Mami, I. and Pallet, N. (2015) ‘Transfer RNA fragmentation and protein translation dynamics in the course of kidney injury.’, *RNA biology*, p. 0. doi: 10.1080/15476286.2015.1107704.

Marais, G., Mouchiroud, D. and Duret, L. (2001) ‘Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes’, *Proceedings of the National Academy of Sciences*, 98(10), pp. 5688–5692. doi: 10.1073/pnas.091427698.

Margulies, E. H. *et al.* (2007) ‘Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome’, *Genome Research*, 17(6), pp. 760–774. doi: 10.1101/gr.6034307.

Margulies, M. *et al.* (2005) ‘Genome sequencing in microfabricated high-density picolitre reactors’, *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

Marques-Bonet, T., Girirajan, S. and Eichler, E. E. (2009) ‘The origins and impact of primate segmental duplications’, *Trends in Genetics*, pp. 443–454. doi: 10.1016/j.tig.2009.08.002.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Martinez-Contreras, R. *et al.* (2007) ‘hnRNP Proteins and Splicing Control’, in *Advances in experimental medicine and biology*, pp. 123–147. doi: 10.1007/978-0-387-77374-2.

Maxam, a M. and Gilbert, W. (1977) ‘A new method for sequencing DNA.’, *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp. 560–4. doi: 10.1073/pnas.74.2.560.

May, W. A. *et al.* (1993) ‘The Ewing’s sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1.’, *Molecular and cellular biology*, 13(12), pp. 7393–8. doi: 10.1128/MCB.13.12.7393.

Maynard Smith, J. and Haigh, J. (2008) ‘The hitch-hiking effect of a favourable gene’, *Genetics Research*, 89(5–6), pp. 391–403. doi: 10.1017/S0016672308009579.

McCarthy, A. D. and Hardie, D. G. (1984) ‘Fatty acid synthase - an example of protein evolution by gene fusion’, *Trends in Biochemical Sciences*, 9(2), pp. 60–63. doi: 10.1016/0968-0004(84)90184-1.

McGinty, R. K. and Tan, S. (2015) 'Nucleosome structure and function', *Chemical Reviews*, pp. 2255–2273. doi: 10.1021/cr500373h.

McInerney, J. O. (1998) 'GCUA: General codon usage analysis', *Bioinformatics*, 14(4), pp. 372–373. doi: 10.1093/bioinformatics/14.4.372.

McLean, C. Y. *et al.* (2011) 'Human-specific loss of regulatory DNA and the evolution of human-specific traits', *Nature*, 471(7337), pp. 216–219. doi: 10.1038/nature09774.

Metzker, M. L. (2010) 'Sequencing technologies the next generation', *Nature Reviews Genetics*, pp. 31–46. doi: 10.1038/nrg2626.

Mi, H. *et al.* (2009) 'PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium', *Nucleic Acids Research*, 38(SUPPL.1). doi: 10.1093/nar/gkp1019.

Michel, A. M. *et al.* (2014) 'GWIPS-viz: Development of a ribo-seq genome browser', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1035.

Mita, P. and Boeke, J. D. (2016) 'How retrotransposons shape genome regulation', *Current Opinion in Genetics and Development*, pp. 90–100. doi: 10.1016/j.gde.2016.01.001.

Mitrovich, Q. M. and Anderson, P. (2000) 'Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans', *Genes and Development*, 14(17), pp. 2173–2184. doi: 10.1101/gad.819900.

Moreira, D. and Philippe, H. (2000) 'Molecular phylogeny: pitfalls and progress.', *International microbiology : the official journal of the Spanish Society for Microbiology*, 3(1), pp. 9–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10963328.

Morgan, C. C. *et al.* (2013) 'Heterogeneous models place the root of the placental mammal phylogeny', *Molecular Biology and Evolution*, 30(9), pp. 2145–2156. doi: 10.1093/molbev/mst117.

Morgante, M. *et al.* (2005) 'Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize', *Nature Genetics*, 37(9), pp. 997–1002. doi: 10.1038/ng1615.

Muller, J. *et al.* (2010) 'AQUA: Automated quality improvement for multiple sequence alignments', *Bioinformatics*, 26(2), pp. 263–265. doi: 10.1093/bioinformatics/btp651.

Munoz-Lopez, M. and Garcia-Perez, J. (2010) 'DNA Transposons: Nature and Applications in Genomics', *Current Genomics*, 11(2), pp. 115–128. doi: 10.2174/138920210790886871.

Nacu, S. *et al.* (2011) 'Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples', *BMC Medical Genomics*, 4. doi: 10.1186/1755-8794-4-11.

Narlikar, L. and Ovcharenko, I. (2009) 'Identifying regulatory elements in eukaryotic genomes', *Briefings in Functional Genomics and Proteomics*, pp. 215–230. doi: 10.1093/bfgp/elp014.

Neu-Yilik, G. *et al.* (2011) 'Mechanism of escape from nonsense-mediated mRNA decay of human β-globin transcripts with nonsense mutations in the first exon', *RNA*, 17(5), pp. 843–854. doi: 10.1261/rna.2401811.

Neverov, A. D. *et al.* (2005) 'Alternative splicing and protein function.', *BMC bioinformatics*, 6(1), p. 266. doi: 10.1186/1471-2105-6-266.

Newman, M. E. J. (2003) 'Mixing patterns in networks', *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(2), p. 13. doi: 10.1103/PhysRevE.67.026126.

Nilsen, T. W. and Graveley, B. R. (2010) 'Expansion of the eukaryotic proteome by alternative splicing', *Nature*, pp. 457–463. doi: 10.1038/nature08909.

Nolan, T., Hands, R. E. and Bustin, S. A. (2006) 'Quantification of mRNA using real-time RT-PCR', *Nature Protocols*, 1(3), pp. 1559–1582. doi: 10.1038/nprot.2006.236.

Nowick, K. *et al.* (2009) 'Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain.', *Proceedings of the National Academy of Sciences of the United States of America*, 106(52), pp. 22358–22363. doi: 10.1073/pnas.0911376106.

Nyberg, K. G. and Carthew, R. W. (2017) 'Out of the testis: biological impacts of new genes', *Genes & Development*, 31(18), pp. 1825–1826. doi: 10.1101/gad.307496.117.

O'Bleness, M. *et al.* (2012) 'Evolution of genetic and genomic features unique to the human lineage', *Nature Reviews Genetics*, pp. 853–866. doi: 10.1038/nrg3336.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI:

Current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733–D745. doi: 10.1093/nar/gkv1189.

O'Toole, Á. N., Hurst, L. D. and McLysaght, A. (2018) 'Faster Evolving Primate Genes Are More Likely to Duplicate', *Molecular Biology and Evolution*, 35(1), pp. 107–118. doi: 10.1093/molbev/msx270.

Ohno, S. (1970) 'Evolution by Gene Duplication', *(1970)*. doi: 10.1007/978-3-642-86659-3.

Ohta, T. (1992) 'The Nearly Neutral Theory Of Molecular Evolution', *Annual Review of Ecology and Systematics*, 23(1992), pp. 263–286. doi: 10.2307/2097289.

Orphanides, G., Lagrange, T. and Reinberg, D. (1996) 'The general transcription factors of RNA polymerase II.', *Genes & development*, 10(21), pp. 2657–83. doi: 10.1101/gad.10.21.2657.

Otto, W. *et al.* (2010) 'Measuring Transcription Factor-Binding Site Turnover: A Maximum Likelihood Approach Using Phylogenies', *Genome Biology and Evolution*, 1(0), pp. 85–98. doi: 10.1093/gbe/evp010.

Pal, M. and Luse, D. S. (2003) 'The initiation-elongation transition: lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23.', *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp. 5700–5. doi: 10.1073/pnas.1037057100.

Palomar, D. P. and Chiang, M. (2006) 'A tutorial on decomposition methods for network utility maximization', *IEEE Journal on Selected Areas in Communications*, 24(8), pp. 1439–1451. doi: 10.1109/JSAC.2006.879350.

Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature Genetics*, 40(12), pp. 1413–1415. doi: 10.1038/ng.259.

Pan, Y. *et al.* (2010) 'Mechanisms of transcription factor selectivity', *Trends in Genetics*, pp. 75–83. doi: 10.1016/j.tig.2009.12.003.

Panov, K. I., Friedrich, J. K. and Zomerdijk, J. C. (2001) 'A step subsequent to preinitiation complex assembly at the ribosomal RNA gene promoter is rate limiting for human RNA polymerase I-dependent transcription.', *Molecular and cellular biology*, 21(8), pp. 2641–9. doi: 10.1128/MCB.21.8.2641-2649.2001.

Papatheodorou, I. *et al.* (2018) 'Expression Atlas: Gene and protein expression

across multiple studies and organisms', *Nucleic Acids Research*, 46(D1), pp. D246–D251. doi: 10.1093/nar/gkx1158.

Pardigol, A. *et al.* (1998) 'HCC-2, a human chemokine: gene structure, expression pattern, and biological activity', *Proc Natl Acad Sci U S A*, 95(11), pp. 6308–6313.

Park, E. *et al.* (2011) 'Regulatory roles of hnRNP M and Nova-1 in the alternative splicing of the dopamine D2 receptor pre-mRNA.', *Journal of Biological Chemistry*, 286(28):25(28), pp. 25301–8. doi: 10.1074/jbc.M110.206540.

Parker, B. C. *et al.* (2013) 'The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma', *Journal of Clinical Investigation*, 123(2), pp. 855–865. doi: 10.1172/JCI67144.

Parmley, J. L., Chamary, J. V. and Hurst, L. D. (2006) 'Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers', *Molecular Biology and Evolution*, 23(2), pp. 301–309. doi: 10.1093/molbev/msj035.

Pascal, R. and Boiteau, L. (2011) 'Energy flows, metabolism and translation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1580), pp. 2949–2958. doi: 10.1098/rstb.2011.0135.

Pasek, S., Risler, J. L. and Brézellec, P. (2006) 'Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins', *Bioinformatics*, 22(12), pp. 1418–1423. doi: 10.1093/bioinformatics/btl135.

Pathmanathan, J. S. *et al.* (2018) 'CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection', *Molecular Biology and Evolution*, 35(1), pp. 252–255. doi: 10.1093/molbev/msx283.

Patricia A.J. Muller1, ∗ and Karen H. Vousden2 (2014) 'Mutant p53 in Cancer: New Functions and Therapeutic Opportunities', *Cancer Cell*. doi: 10.1016/j.ccr.2014.01.021.

Patthy, L. (1999) 'Genome evolution and the evolution of exon-shuffling - A review', *Gene*, 238(1), pp. 103–114. doi: 10.1016/S0378-1119(99)00228-0.

Patthy, L. (2003) 'Modular assembly of genes and the evolution of new functions', *Genetica*, pp. 217–231. doi: 10.1023/A:1024182432483.

Pavlopoulos, G. A. *et al.* (2011) 'Using graph theory to analyze biological

networks', *BioData Mining*. doi: 10.1186/1756-0381-4-10.

Paz, I. *et al.* (2010) 'SFmap: A web server for motif analysis and prediction of splicing factor binding sites', *Nucleic Acids Research*, 38(SUPPL. 2). doi: 10.1093/nar/gkq444.

Paz, I. *et al.* (2014) 'RBPmap: A web server for mapping binding sites of RNA-binding proteins', *Nucleic Acids Research*, 42(W1). doi: 10.1093/nar/gku406.

Peccarelli, M. and Kebaara, B. W. (2014) 'Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway', *Eukaryotic Cell*, 13(9), pp. 1126–1135. doi: 10.1128/EC.00090-14.

Pereira, V. (2004) 'Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome.', *Genome Biology*, 5(10), p. R79. doi: 10.1186/gb-2004-5-10-r79.

Perner, S. (2005) 'Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer Transcription Factor Genes in Prostate Cancer', *science*, 310(November 2015), pp. 644–648. doi: 10.1126/science.1117679.

Persson, M. *et al.* (2009) 'Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck', *Proceedings of the National Academy of Sciences*, 106(44), pp. 18740–18744. doi: 10.1073/pnas.0909114106.

Peter, I. S. and Davidson, E. H. (2011) 'Evolution of gene regulatory networks controlling body plan development', *Cell*, pp. 970–985. doi: 10.1016/j.cell.2011.02.017.

Phatnani, H. P. and Greenleaf, A. L. (2006) 'Phosphorylation and functions of the RNA polymerase II CTD', *Genes and Development*, pp. 2922–2936. doi: 10.1101/gad.1477006.

Philips, T. and Hoopes, L. (2008) 'Transcription Factors and Transcriptional Control in Eukaryotic Cells', *Nature Education*, 1(1), p. 119. doi: 10.1016/0092.

Pimentel, H. *et al.* (2017) 'Differential analysis of RNA-seq incorporating quantification uncertainty', *Nature Methods*, 14(7), pp. 687–690. doi: 10.1038/nmeth.4324.

Polychronopoulos, D. *et al.* (2017) 'Conserved non-coding elements: Developmental gene regulation meets genome organization', *Nucleic Acids Research*, pp. 12611–12624. doi: 10.1093/nar/gkx1074.

Porrua, O., Boudvillain, M. and Libri, D. (2016) 'Transcription Termination: Variations on Common Themes', *Trends in Genetics*, pp. 508–522. doi: 10.1016/j.tig.2016.05.007.

Posada, D. and Crandall, K. A. (2002) 'The effect of recombination on the accuracy of phylogeny estimation', *Journal of Molecular Evolution*, 54(3), pp. 396–402. doi: 10.1007/s00239-001-0034-9.

Potaman, V. N. and Sinden, R. R. (2005) 'CHAPTER 1 DNA: Alternative Conformations and Biology', *DNA Conformation and Transcription*, pp. 1–16. Available at: https://www.landesbioscience.com/curie/chapter/2078/.

Pradet-Balade, B. *et al.* (2002) 'An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK-APRIL fusion protein', *EMBO Journal*, 21(21), pp. 5711–5720. doi: 10.1093/emboj/cdf565.

Prado-Martinez, J. *et al.* (2013) 'Great ape genetic diversity and population history', *Nature*, 499(7459), pp. 471–475. doi: 10.1038/nature12228.

Qi, X. *et al.* (2011) 'A novel model for DNA sequence similarity analysis based on graph theory', *Evolutionary Bioinformatics*, 2011(7), pp. 149–158. doi: 10.4137/EBO.S7364.

R Development Core Team, R. (2011) *R: A Language and Environment for Statistical Computing*, *R Foundation for Statistical Computing*. doi: 10.1007/978-3-540-74686-7.

Rajalingam, R., Parham, P. and Abi-Rached, L. (2004) 'Domain Shuffling Has Been the Main Mechanism Forming New Hominoid Killer Cell Ig-Like Receptors', *The Journal of Immunology*, 172(1), pp. 356–369. doi: 10.4049/jimmunol.172.1.356.

Raman, K. (2010) 'Construction and analysis of protein-protein interaction networks', *Automated Experimentation*. doi: 10.1186/1759-4499-2-2.

Ratnakumar, A. *et al.* (2010) 'Detecting positive selection within genomes: the problem of biased gene conversion', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552), pp. 2571–2580. doi: 10.1098/rstb.2010.0007.

Redelings, B. (2014) 'Erasing errors due to alignment ambiguity when estimating positive selection', *Molecular Biology and Evolution*, 31(8), pp. 1979–1993. doi: 10.1093/molbev/msu174.

Ren, B. and Dynlacht, B. D. (2004) 'Use of Chromatin Immunoprecipitation Assays in Genome-Wide Location Analysis of Mammalian Transcription Factors', *Methods in Enzymology*, 376, pp. 304–315. doi: 10.1016/S0076-6879(03)76020-0.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Rhoads, R. E., Dinkova, T. D. and Jagus, R. (2007) *Translation Initiation: Extract Systems and Molecular Genetics*, *Methods in enzymology*. doi: 10.1016/S0076-6879(07)29013-5.

Ricarte-Filho, J. C. *et al.* (2009) 'Mutational profile of advanced primary and metastatic radioactive iodine-refractory thyroid cancers reveals distinct pathogenetic roles for BRAF, PIK3CA, and AKT1', *Cancer Research*, 69(11), pp. 4885–4893. doi: 10.1158/0008-5472.CAN-09-0727.

Richard, P. and Manley, J. L. (2009) 'Transcription termination by nuclear RNA polymerases', *Genes & development*, pp. 1247–1269. doi: 10.1101/gad.1792809.

Richards, S. *et al.* (2005) 'Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution', *Genome Research*, 15(1), pp. 1–18. doi: 10.1101/gr.3059305.

Richardson, D. N. *et al.* (2011) 'Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: Insights into sub-family classification and extent of alternative splicing', *PLoS ONE*, 6(9). doi: 10.1371/journal.pone.0024542.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–329. doi: 10.1038/nature14248.

Roberts, R. J., Carneiro, M. O. and Schatz, M. C. (2013) 'The advantages of SMRT sequencing', *Genome Biology*, 14(6). doi: 10.1186/gb-2013-14-6-405.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.', *Bioinformatics (Oxford, England)*, 26(1), pp. 139–40. doi: 10.1093/bioinformatics/btp616.

Rodríguez-Martín, B. *et al.* (2017) 'ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data', *BMC genomics*, 18(1),

p. 7. doi: 10.1186/s12864-016-3404-9.

Rogers, J. and Gibbs, R. A. (2014) 'Comparative primate genomics: Emerging patterns of genome content and dynamics', *Nature Reviews Genetics*, pp. 347–359. doi: 10.1038/nrg3707.

Romiguier, J. *et al.* (2010) 'Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes', *Genome Research*, 20(8), pp. 1001–1009. doi: 10.1101/gr.104372.109.

Rooijers, K. *et al.* (2013) 'Ribosome profiling reveals features of normal and disease-associated mitochondrial translation', *Nature Communications*, 4. doi: 10.1038/ncomms3886.

Rostad, K. *et al.* (2007) 'ERG upregulation and related ETS transcription factors in prostate cancer.', *International journal of oncology*, 30(1), pp. 19–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17143509.

Rozov, A. *et al.* (2016) 'Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code', *Nature Communications*, 7. doi: 10.1038/ncomms10457.

Ruiz I Altaba, A. (2011) 'Hedgehog signaling and the Gli code in stem cells, cancer, and metastases', in *Science Signaling*. doi: 10.1126/scisignal.2002540.

Ruzycki, P. A. *et al.* (2015) 'Graded gene expression changes determine phenotype severity in mouse models of CRX-associated retinopathies', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0732-z.

Sabeti, P. C. *et al.* (2006) 'Positive natural selection in the human lineage', *Science*, pp. 1614–1620. doi: 10.1126/science.1124309.

Sabir, N. *et al.* (2012) 'Prognostically Significant Fusion Oncogenes in Pakistani Patients with Adult Acute Lymphoblastic Leukemia and their Association with Disease Biology and Outcome', *Asian Pacific Journal of Cancer Prevention*, 13(7), pp. 3349–3355. doi: 10.7314/APJCP.2012.13.7.3349.

Samonte, R. V. and Eichler, E. E. (2002) 'Segmental duplications and the evolution of the primate genome', *Nature Reviews Genetics*, pp. 65–72. doi: 10.1038/nrg705.

San Mauro, D. *et al.* (2006) 'A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome', *Molecular Biology and Evolution*, 23(1), pp. 227–234. doi: 10.1093/molbev/msj025.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Sassaman, D. M. *et al.* (1997) 'Many human L1 elements are capable of retrotransposition', *Nature Genetics*, 16(1), pp. 37–43. doi: 10.1038/ng0597-37.

Schlötterer, C. (2015) 'Genes from scratch - the evolutionary fate of de novo genes', *Trends in Genetics*, pp. 215–219. doi: 10.1016/j.tig.2015.02.007.

Schmidt, D. *et al.* (2010) 'Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.', *Science (New York, N.Y.)*, 328(5981), pp. 1036–40. doi: 10.1126/science.1186176.

Schmucker, D. *et al.* (2000) 'Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity', *Cell*, 101(6), pp. 671–684. doi: 10.1016/S0092-8674(00)80878-8.

Schneider, A. *et al.* (2010) 'Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment', *Genome Biology and Evolution*, 1(0), pp. 114–118. doi: 10.1093/gbe/evp012.

Schubert, F. R., Nieselt-Struwe, K. and Gruss, P. (1993) 'The Antennapedia-type homeobox genes have evolved from three precursors separated early in metazoan evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 90(1), pp. 143–7. doi: 10.1073/pnas.90.1.143.

Shannon, P. *et al.* (2003) 'Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks', *Genome Research*, (Karp 2001), pp. 2498–2504. doi: 10.1101/gr.1239303.

Sharp, A. J. *et al.* (2005) 'Segmental duplications and copy-number variation in the human genome.', *American journal of human genetics*, 77(1), pp. 78–88. doi: 10.1086/431652.

She, X. *et al.* (2006) 'A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications', *Genome Research*, 16(5), pp. 576–583. doi: 10.1101/gr.4949406.

Shimamura, K. *et al.* (1994) 'Wnt-1-dependent regulation of local E-cadherin and alpha N-catenin expression in the embryonic mouse brain.', *Development*

(Cambridge, England), 120, pp. 2225–2234.

Silverman, E. S. *et al.* (2002) 'Constitutive and cytokine-induced expression of the ETS transcription factor ESE-3 in the lung', *American Journal of Respiratory Cell and Molecular Biology*, 27(6), pp. 697–704. doi: 10.1165/rcmb.2002-0011OC.

Simmons, M. P., Ochoterena, H. and Freudenstein, J. V. (2002) 'Amino acid vs. nucleotide characters: Challenging preconceived notions', *Molecular Phylogenetics and Evolution*, 24(1), pp. 78–90. doi: 10.1016/S1055-7903(02)00202-6.

Simonsen, K. L., Churchill, G. A. and Aquadro, C. F. (1995) 'Properties of statistical tests of neutrality for DNA polymorphism data', *Genetics*, 141(1), pp. 413–429. doi: drosophila-melanogaster; population-genetics; x-chromosome; region; recombination; hitchhiking; mutations; models.

Skandalis, A. *et al.* (2010) 'The adaptive significance of unproductive alternative splicing in primates', *RNA*, 16(10), pp. 2014–2022. doi: 10.1261/rna.2127910.

Smith, J. E. and Baker, K. E. (2015) 'Nonsense-mediated RNA decay-a switch and dial for regulating gene expression.', *BioEssays : news and reviews in molecular, cellular and developmental biology*, 37(6), pp. 612–23. doi: 10.1002/bies.201500007.

Smith, R. P. *et al.* (2012) 'Pharmacogene regulatory elements: From discovery to applications', *Genome Medicine*. doi: 10.1186/gm344.

Snel, B., Bork, P. and Huynen, M. (2000) 'Genome evolution. Gene fusion versus gene fission', *Trends Genet*, 16(1), pp. 9–11. doi: S0168-9525(99)01924-1 [pii].

Sowdhamini, R., Rufino, S. D. and Blundell, T. L. (1996) 'A database of globular protein structural domains: Clustering of representative family members into similar folds', *Folding and Design*, 1(3), pp. 209–220. doi: 10.1016/S1359-0278(96)00032-6.

Spencer, C. A. and Groudine, M. (1990) 'Transcription elongation and eukaryotic gene regulation', *Oncogene*, 5(6), pp. 777–785.

Spitz, F. and Furlong, E. E. M. (2012) 'Transcription factors: From enhancer binding to developmental control', *Nature Reviews Genetics*, pp. 613–626. doi: 10.1038/nrg3207.

Spriggs, K. A., Bushell, M. and Willis, A. E. (2010) 'Translational Regulation of Gene Expression during Conditions of Cell Stress', *Molecular Cell*, pp. 228–237. doi: 10.1016/j.molcel.2010.09.028.

Stankiewicz, P. *et al.* (2004) 'Serial segmental duplications during primate evolution result in complex human genome architecture.', *Genome research*, 14(11), pp. 2209–2220. doi: 10.1101/gr.2746604.

Su, Z. *et al.* (2014) 'Focus on RNA sequencing quality control (SEQC)', *Nature Biotechnology*, 32(9), pp. vii–vii.

Sudmant, P. H., Alexis, M. S. and Burge, C. B. (2015) 'Meta-analysis of RNA-seq expression data across species, tissues and studies', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0853-4.

Sullivan, L. L., Chew, K. and Sullivan, B. A. (2017) 'α satellite DNA variation and function of the human centromere', *Nucleus*, pp. 331–339. doi: 10.1080/19491034.2017.1308989.

Sun, Y. *et al.* (2012) 'Using complex network theory in the Internet engineering', in *ICCSE 2012 - Proceedings of 2012 7th International Conference on Computer Science and Education*. doi: 10.1109/ICCSE.2012.6295099.

Sunita, S. *et al.* (2007) 'Functional specialization of domains tandemly duplicated within 16S rRNA methyltransferase RsmC', *Nucleic Acids Research*, 35(13), pp. 4264–4274. doi: 10.1093/nar/gkm411.

Suzuki, I. K. *et al.* (2018) 'Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation', *Cell*, 173(6), p. 1370–1384.e16. doi: 10.1016/j.cell.2018.03.067.

Tacke, R. *et al.* (1998) 'Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing', *Cell*, pp. 139–148. doi: 10.1016/S0092-8674(00)81153-8.

Tan, I. K. L. *et al.* (2010) 'A recombination hotspot leads to sequence variability within a novel gene (AK005651) and contributes to type 1 diabetes susceptibility', *Genome Research*, 20(12), pp. 1629–1638. doi: 10.1101/gr.101881.109.

Tansey, W. P. (2014) 'Mammalian MYC Proteins and Cancer', *New Journal of Science*, 2014, pp. 1–27. doi: 10.1155/2014/757534.

Tatusova, T. *et al.* (2015) 'Erratum: RefSeq microbial genomes database: New representation and annotation strategy (Nucleic Acids Res. (2014) 42 (D553-

D559))', *Nucleic Acids Research*, p. 3872. doi: 10.1093/nar/gkv278.

Tee, W.-W. and Reinberg, D. (2014) 'Chromatin features and the epigenetic regulation of pluripotency states in ESCs', *Development*, 141(12), pp. 2376–2390. doi: 10.1242/dev.096982.

Teixeira, M. R. (2006) 'Recurrent fusion oncogenes in carcinomas.', *Critical reviews in oncogenesis*, 12, pp. 257–271. doi: 10.1615/CritRevOncog.v12.i3-4.40.

Tennessen, J. A. (2008) 'Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence', *Bioinformatics*, 24(12), pp. 1421–1425. doi: 10.1093/bioinformatics/btn205.

Thiel, C. S. *et al.* (2017) 'Dynamic gene expression response to altered gravity in human T cells', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-05580-x.

Thomas, C. A. (1971) 'The Genetic Organization of Chromosomes', *Annual Review of Genetics*, 5(1), pp. 237–256. doi: 10.1146/annurev.ge.05.120171.001321.

Thompson, J. D. *et al.* (2001) 'Towards a reliable objective function for multiple sequence alignments.', *Journal of molecular biology*, 314(4), pp. 937–51. doi: 10.1006/jmbi.2001.5187.

Thomson, T. M. *et al.* (2000) 'Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene.', *Genome research*, 10(11), pp. 1743–56. doi: 10.1101/gr.GR-1405R.

Timchenko, L. T. *et al.* (1996) 'Identification of a (CUG)(n) triplet repeat RNA-binding protein and its expression in myotonic dystrophy', *Nucleic Acids Research*, 24(22), pp. 4407–4414. doi: 10.1093/nar/24.22.4407.

Tintaru, A. M. *et al.* (2007) 'Structural and functional analysis of RNA and TAP binding to SF2/ASF', *EMBO Reports*, 8(8), pp. 756–762. doi: 10.1038/sj.embor.7401031.

Tran, Q. and Roesser, J. R. (2003) 'SRp55 is a regulator of calcitonin/CGRP alternative RNA splicing', *Biochemistry*, 42(4), pp. 951–957. doi: 10.1021/bi026753a.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat: Discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105–1111. doi: 10.1093/bioinformatics/btp120.

Tsochatzidou, M. *et al.* (2017) 'Genome urbanization: Clusters of topologically co-regulated genes delineate functional compartments in the genome of Saccharomyces cerevisiae', *Nucleic Acids Research*, 45(10), pp. 5818–5828. doi: 10.1093/nar/gkx198.

Tugores, A. *et al.* (2001) 'The Epithelium-specific ETS Protein EHF/ESE-3 is a Context-dependent Transcriptional Repressor Downstream of MAPK Signaling Cascades', *Journal of Biological Chemistry*, 276(23), pp. 20397–20406. doi: 10.1074/jbc.M010930200.

Tuğrul, M. *et al.* (2015) 'Dynamics of Transcription Factor Binding Site Evolution', *PLoS Genetics*, 11(11). doi: 10.1371/journal.pgen.1005639.

Tung, J., Alberts, S. C. and Wray, G. A. (2010) 'Evolutionary genetics in wild primates: Combining genetic approaches with field studies of natural populations', *Trends in Genetics*, pp. 353–362. doi: 10.1016/j.tig.2010.05.005.

Turner, B. M. (2009) 'Epigenetic responses to environmental change and their evolutionary implications.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1534), pp. 3403–3418. doi: 10.1098/rstb.2009.0125.

Tyagi, M. *et al.* (2016) 'Chromatin remodelers: We are the drivers!!', *Nucleus*, pp. 388–404. doi: 10.1080/19491034.2016.1211217.

Uebbing, S. *et al.* (2016) 'Divergence in gene expression within and between two closely related flycatcher species', *Molecular Ecology*, 25(9), pp. 2015–2028. doi: 10.1111/mec.13596.

Uhlén, M. *et al.* (2015) 'Proteomics. Tissue-based map of the human proteome.', *Science (New York, N.Y.)*, 347(6220), p. 1260419. doi: 10.1126/science.1260419.

Ule, J. *et al.* (2005) 'Nova regulates brain-specific splicing to shape the synapse', *Nature Genetics*, 37(8), pp. 844–852. doi: 10.1038/ng1610.

UniProt Consortium, T. (2018) 'UniProt: the universal protein knowledgebase', *Nucleic Acids Research*, 46(5), pp. 2699–2699. doi: 10.1093/nar/gky092.

Upadhyaya, A. B., Lee, S. H. and Dejong, J. (1999) 'Identification of a general transcription factor TFIIAalpha/beta homolog selectively expressed in testis', *Journal of Biological Chemistry*, 274(25), pp. 18040–18048. doi: 10.1074/jbc.274.25.18040.

Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003) 'Comparative genomics:

Genome-wide analysis in metazoan eukaryotes', *Nature Reviews Genetics*, pp. 251–262. doi: 10.1038/nrg1043.

Uzman, A. *et al.* (2000) 'Molecular Cell Biology (4th edition) New York, NY, 2000, ISBN 0-7167-3136-3', *Biochemistry and Molecular Biology Education*, 29, p. Section 1.2The Molecules of Life. doi: 10.1016/S1470-8175(01)00023-6.

Valentijn, L. J., Koster, J. and Versteeg, R. (2006) 'Read-through transcript from NM23-H1 into the neighboring NM23-H2 gene encodes a novel protein, NM23-LV', *Genomics*, 87(4), pp. 483–489. doi: 10.1016/j.ygeno.2005.11.004.

Vaquerizas, J. M. *et al.* (2009) 'A census of human transcription factors: Function, expression and evolution', *Nature Reviews Genetics*, pp. 252–263. doi: 10.1038/nrg2538.

Varki, A. and Altheide, T. K. (2005) 'Comparing the human and chimpanzee genomes: Searching for needles in a haystack', *Genome Research*, pp. 1746–1758. doi: 10.1101/gr.3737405.

Veeramachaneni, V. *et al.* (2004) 'Mammalian overlapping genes: The comparative perspective', *Genome Research*, 14(2), pp. 280–286. doi: 10.1101/gr.1590904.

Velculescu, V. E. *et al.* (1995) 'Serial Analysis of Gene Expression', *Science*, 270(5235), pp. 484–487. doi: 10.1126/science.270.5235.484.

Venter, J. C. (2001) 'The Sequence of the Human Genome', *Science*, 291(5507), pp. 1304–1351. doi: 10.1126/science.1058040.

Ventura, M. *et al.* (2012) 'The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2', *Genome Research*, 22(6), pp. 1036–1049. doi: 10.1101/gr.136556.111.

Verrijzer, C. P., Kal, A. J. and Van der Vliet, P. C. (1990) 'The DNA binding domain (POU domain) of transcription factor oct-1 suffices for stimulation of DNA replication.', *The EMBO journal*, 9(6), pp. 1883–8. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC551894/pdf/emboj00233-0202.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/2347308%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC551894.

Villanueva-Cañas, J. L. *et al.* (2017) 'New genes and functional innovation in mammals', *Genome Biology and Evolution*, 9(7), pp. 1886–1900. doi: 10.1093/gbe/evx136.

Villar, D., Flicek, P. and Odom, D. T. (2014) 'Evolution of transcription factor binding in metazoans-mechanisms and functional implications', *Nature Reviews Genetics*, pp. 221–233. doi: 10.1038/nrg3481.

Vogel, C. and Morea, V. (2006) 'Duplication, divergence and formation of novel protein topologies', *BioEssays*, pp. 973–978. doi: 10.1002/bies.20474.

Voldoire, E. *et al.* (2017) 'Expansion by whole genome duplication and evolution of the sox gene family in teleost fish', *PLoS ONE*, 12(7). doi: 10.1371/journal.pone.0180936.

Voliotis, M. *et al.* (2008) 'Fluctuations, pauses, and backtracking in DNA transcription', *Biophysical Journal*, 94(2), pp. 334–348. doi: 10.1529/biophysj.107.105767.

Wang, E. T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, 456(7221), pp. 470–476. doi: 10.1038/nature07509.

Wang, E. T. *et al.* (2015) 'Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins', *Genome Research*, 25(6), pp. 858–871. doi: 10.1101/gr.184390.114.

Wang, W., Yu, H. and Long, M. (2004) 'Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species', *Nature Genetics*, 36(5), pp. 523–527. doi: 10.1038/ng1338.

Wang, Y. *et al.* (1998) 'Otoconin-90, the mammalian otoconial matrix protein, contains two domains of homology to secretory phospholipase A2.', *Proceedings of the National Academy of Sciences of the United States of America*, 95(26), pp. 15345–15350. doi: 10.1073/pnas.95.26.15345.

Wang, Y. *et al.* (2015) 'FusionCancer: A database of cancer fusion genes derived from RNA-seq data', *Diagnostic Pathology*, 10(1). doi: 10.1186/s13000-015-0310-4.

Warf, M. B. and Berglund, J. A. (2007) 'MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T', *RNA*, 13(12), pp. 2238–2251. doi: 10.1261/rna.610607.

Wasserman, W. W. and Sandelin, A. (2004) 'Applied bioinformatics for the identification of regulatory elements', *Nature Reviews Genetics*, pp. 276–287. doi: 10.1038/nrg1315.

Watson, J. D. and Crick, F. H. C. (1953) 'Molecular structure of nucleic acids',

Nature, pp. 737–738. doi: 10.1097/BLO.0b013e3181468780.

Webb, A. E. *et al.* (2015) 'Adaptive evolution as a predictor of species-specific innate immune response', *Molecular Biology and Evolution*, 32(7), pp. 1717–1729. doi: 10.1093/molbev/msv051.

Wei, L. *et al.* (2013) 'New insights into nested long terminal repeat retrotransposons in brassica species', *Molecular Plant*, 6(2), pp. 470–482. doi: 10.1093/mp/sss081.

Wei, W. J. *et al.* (2012) 'YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts', *Nucleic Acids Research*, 40(17), pp. 8622–8636. doi: 10.1093/nar/gks579.

Weichenhan, D. *et al.* (1998) 'Evolution by fusion and amplification: the murine Sp100-rs gene cluster', *Cytogenet Cell Genet*, 80(1–4), pp. 226–231.

Welch, D., Bansal, S. and Hunter, D. R. (2011) 'Statistical inference to advance network models in epidemiology', *Epidemics*, 3(1), pp. 38–45. doi: 10.1016/j.epidem.2011.01.002.

Werner, F. and Grohmann, D. (2011) 'Evolution of multisubunit RNA polymerases in the three domains of life', *Nature Reviews Microbiology*, pp. 85–98. doi: 10.1038/nrmicro2507.

Wetterstrand, K. A. (2016) 'DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program', *www.genome.gov/sequencingcostsdata*, (2), p. Accessed [4 September 2016].

Wickramasinghe, V. O. *et al.* (2015) 'Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0749-3.

Wilkinson, M. *et al.* (2004) 'Some desiderata for liberal supertrees', in *Phylogenetic supertrees: combining information to reveal the Tree of Life*, p. 564. doi: 10.1007/978-1-4020-2330-9_11.

Will, C. L. and Lührmann, R. (2011) 'Spliceosome structure and function', *Cold Spring Harbor Perspectives in Biology*, 3(7), pp. 1–2. doi: 10.1101/cshperspect.a003707.

Wilson, D. N. and Cate, J. H. D. (2012) 'The structure and function of the eukaryotic ribosome', *Cold Spring Harbor Perspectives in Biology*, 4(5), p. 5. doi: 10.1101/cshperspect.a011536.

Withers, J. B., Ashvetiya, T. and Beemon, K. L. (2012) 'Exclusion of Exon 2 Is a Common mRNA Splice Variant of Primate Telomerase Reverse Transcriptases', *PLoS ONE*, 7(10). doi: 10.1371/journal.pone.0048016.

Xiao, R. *et al.* (2007) 'Splicing Regulator SC35 Is Essential for Genomic Stability and Cell Proliferation during Mammalian Organogenesis', *Molecular and Cellular Biology*, 27(15), pp. 5393–5402. doi: 10.1128/MCB.00288-07.

Xiong, H. Y. *et al.* (2015) 'The human splicing code reveals new insights into the genetic determinants of disease', *Science*, 347(6218). doi: 10.1126/science.1254806.

Yamaguchi, Y., Shibata, H. and Handa, H. (2013) 'Transcription elongation factors DSIF and NELF: Promoter-proximal pausing and beyond', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, pp. 98–104. doi: 10.1016/j.bbagrm.2012.11.007.

Yang, Z. (2007) 'PAML 4: Phylogenetic analysis by maximum likelihood', *Molecular Biology and Evolution*, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.

Yang, Z. and Bielawski, J. R. (2000) 'Statistical methods for detecting molecular adaptation', *Trends in Ecology and Evolution*, pp. 496–503. doi: 10.1016/S0169-5347(00)01994-7.

Yi, S. V. (2013) 'Morris Goodman's hominoid rate slowdown: The importance of being neutral', *Molecular Phylogenetics and Evolution*, pp. 569–574. doi: 10.1016/j.ympev.2012.07.031.

Yi, X. *et al.* (2010) 'Sequencing of 50 human exomes reveals adaptation to high altitude', *Science*, 329(5987), pp. 75–78. doi: 10.1126/science.1190371.

Yoshihito Niimura (2012) 'Olfactory Receptor Multigene Family in Vertebrates: From the Viewpoint of Evolutionary Genomics', *Current Genomics*, 13(2), pp. 103–114. doi: 10.2174/138920212799860706.

Yoshimoto, S. *et al.* (2016) 'Alternative splicing of a cryptic exon embedded in intron 6 of SMN1 and SMN2', *Hum Genome Var*, 3, p. 16040. doi: 10.1038/hgv.2016.40.

Zaphiropoulos, P. G. (1999) 'RNA molecules containing exons originating from different members of the cytochrome P450 2C gene subfamily (CYP2C) in human epidermis and liver', *Nucleic Acids Research*, 27(13), pp. 2585–2590.

doi: 10.1093/nar/27.13.2585.

Zerbino, D. R. *et al.* (2018) 'Ensembl 2018', *Nucleic Acids Research*, 46(D1), pp. D754–D761. doi: 10.1093/nar/gkx1098.

Zhang, C. *et al.* (2008) 'Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2', *Genes and Development*, 22(18), pp. 2550–2563. doi: 10.1101/gad.1703108.

Zhang, J. (2018) 'Neutral Theory and Phenotypic Evolution', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msy065.

Zhang, J., Nielsen, R. and Yang, Z. (2005) 'Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level', *Molecular Biology and Evolution*, 22(12), pp. 2472–2479. doi: 10.1093/molbev/msi237.

Zhang, L. *et al.* (2005) 'Patterns of segmental duplication in the human genome', *Molecular Biology and Evolution*, 22(1), pp. 135–141. doi: 10.1093/molbev/msh262.

Zhang, W. *et al.* (2015) 'New genes drive the evolution of gene interaction networks in the human and mouse genomes', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0772-4.

Zhang, X. *et al.* (2007) 'Role of RNA polymerase IV in plant small RNA metabolism.', *Proceedings of the National Academy of Sciences of the United States of America*, 104(11), pp. 4536–41. doi: 10.1073/pnas.0611456104.

Zhu, X., Gerstein, M. and Snyder, M. (2007) 'Getting connected: Analysis and principles of biological networks', *Genes and Development*, pp. 1010–1024. doi: 10.1101/gad.1528707.

Zola, J. (2014) 'Constructing similarity graphs from large-scale biological sequence collections', in *Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS*, pp. 500–507. doi: 10.1109/IPDPSW.2014.63.

Zolov, S. N. and Lupashin, V. V. (2005) 'Cog3p depletion blocks vesicle-mediated Golgi retrograde trafficking in HeLa cells', *Journal of Cell Biology*, 168(5), pp. 747–759. doi: 10.1083/jcb.200412003.