Alternative routes to optimal expression levels:
evolutionary evidence for competitive endogenous
RNAs and dosage compensation by gene duplication

by

Cian Glenfield

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College
The University of Dublin

September, 2018

Date awarded: 2019
Declaration of Authorship

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University’s open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

___________________________  ______________________
Signature  Date
Summary

Dosage constraints play an important role in shaping genomic content, and gene duplication provides new material for genomic innovation and expression evolution. In this thesis we explore two potential mechanisms of gene expression optimisation arising from different fates of gene duplicates: a widespread role for pseudogenes as competitive endogenous RNAs (ceRNAs), and tandem duplication of high maximally expressed X-linked genes as a means of escaping dosage constraints resulting from upper limits on transcriptional rates from a single locus.

It has become increasingly clear that long non-coding RNAs (lncRNAs) perform various regulatory functions in different cellular processes. The ceRNA hypothesis is an attractively simple model to explain the biological role of many putatively functionless lncRNAs. However, the role of ceRNAs in normal biological processes and at physiological levels is disputed. By comparison of parent genes and pseudogenes we show, both for a specific example and genome-wide, that the pseudo-3'UTRs (untranslated regions) of expressed pseudogenes are frequently retained and are under selective constraint in mammalian genomes. In Chapter 3 we find that the pseudo-3'UTR of BRAF1, a previously described oncogenic ceRNA, has reduced substitutions relative to its pseudo-coding region (pseudo-CDS), and we show sequence constraint on MREs shared between the parent gene, BRAF, and the pseudogene. Investigation of RNA-seq data reveals expression of BRAF1 in normal somatic tissues in human and in other primates, consistent with biological ceRNA functionality of this pseudogene in non-pathogenic cellular contexts.
Furthermore, in Chapter 4 we identify human pseudogene orthologs across 20 mammalian genomes. We find that on a genome-wide scale pseudo-3'UTRs of mammalian pseudogenes are under stronger selective constraint than their pseudo-CDS counterparts, and these pseudogenes are more frequently retained and expressed when a 3'UTR is present. Our results suggest that many human pseudogenes, often considered non-functional, may have an evolutionarily constrained role, consistent with the ceRNA hypothesis.

Evolution of mammalian sex chromosomes is an interesting case of how dosage constraints have shaped gene content on the X. The unusual biology of sex chromosomes necessitates resolution of dosage imbalances through several different mechanisms, including X chromosome inactivation and upregulation of some X-linked genes. Here we explore an additional model for dosage resolution, namely that the most highly expressed genes on the X chromosome are more likely to have duplicated (i.e. are more duplicable) after X chromosome formation, thus resolving constraints placed on them by transcriptional haplidy. In Chapter 5 we examine the relationship between X-linked gene duplication status and expression level. We find that maximal expression levels of X-linked genes that have duplicated in tandem after sex chromosome formation (young X-linked duplicates) are lower than singleton levels, and also have lower levels than X-linked genes that duplicated before sex chromosome formation. In addition, young X-linked duplicates have lower maximal expression levels than young autosomal duplicates. After separating duplicates into maximal expression level categories we find no enrichment of young X-linked duplicates in the most highly expressed category compared to young autosomal duplicates. Our observations here do not currently support the hypothesis that tandem duplication on the X is a means of escaping upper limits on transcriptional rates from single loci.
Acknowledgements

First and foremost, this thesis would not have been possible without the support and encouragement of my family over these past four years. Thank you for constantly believing in me through all those times when I didn’t believe in myself. I will always be grateful.

Thank you to my friends in the Genetics department for chats over coffee, the occasional monthly social, and our (almost) annual ski trips. I’ve had a blast here, and I’ll be sad to leave. Special mention to our D&D group for giving me something great to look forward to each week, and thanks in particular to Eoin for putting so much effort into keeping us happy.

Nicola, Gina, and Aoife, thank you for being constant companions for the last three years or so, and for not getting sick of living with me (yet). I’m probably emigrating to the other side of the world soon so you won’t have to put up with me much longer either way.

I am thankful to all the past and present members of the McLysaght lab for their help, ideas, and friendship during my time here. Without frequent coffee, pastry, and pasta trips I don’t think I would have remained sane.

Finally, thank you Aoife for your support and guidance throughout my time in the lab. It has been an unforgettable experience.
## Contents

1 Introduction .......................................................... 1
   1.1 Pseudogenes ..................................................... 2
      1.1.1 Origins and types of pseudogenes. .................... 3
      1.1.2 Pseudogenes throughout evolution. .................... 8
      1.1.3 Functional pseudogenes. ................................ 12
   1.2 The competitive endogenous RNA hypothesis. ............... 16
      1.2.1 Examples of specific ceRNA activity. ................. 19
      1.2.2 Cellular ceRNA networks. ................................. 26
   1.3 MicroRNA biogenesis and evolution. .......................... 28
      1.3.1 Biogenesis and mechanism of action. .................... 29
      1.3.2 MiRNA and target site conservation. .................... 33
   1.4 Sex chromosome evolution. ..................................... 35
      1.4.1 Therian sex chromosome dosage compensation. ......... 37
      1.4.2 Dosage compensation in other taxa. .................... 39
   1.5 Aim ..................................................................... 41

2 Materials and Methods .................................................. 42
   2.1 Protein-coding gene and pseudogene annotations. .......... 42
   2.2 Identifying pseudogene orthologs. ............................... 45
   2.3 Gene expression data. ............................................. 46
   2.4 Gene trees and paralogy. ......................................... 48
   2.5 Statistical analyses. ................................................. 50
   2.6 Source code and data availability. .............................. 50
3 Sequence constraint of the BRAFP1 pseudogene indicates conservation of a ceRNA function.

3.1 Introduction ........................................... 51
3.2 Materials and Methods .................................. 55
   3.2.1 Phylogenetic inference of BRAFP1 and Braf-rs1 evolutionary history. .................................. 55
   3.2.2 Region-specific substitution analysis. ................. 57
   3.2.3 Prediction of MREs and analysis of validated BRAF and BRAFP1 miRNA interactions. ...................... 58
   3.2.4 BRAF and BRAFP1 expression data. .................. 59
   3.2.5 MiRNA expression analysis. ........................ 60
3.3 Results .................................................... 60
   3.3.1 BRAFP1 and Braf-rs1 evolved independently. ......... 60
   3.3.2 The 3’UTR of BRAFP1 is under greater evolutionary sequence constraint. ............................ 64
   3.3.3 Greater density of shared MREs in the 3’UTR of BRAF and BRAFP1. .................................. 67
   3.3.4 MREs associated with BRAF and BRAFP1 ceRNA activity are conserved. ............................... 69
   3.3.5 Expression of BRAFP1 in human somatic and developmental tissues. .................................... 74
   3.3.6 MiRNA expression is not predictive of BRAFP1 or BRAF expression. .................................. 79
3.4 Discussion ................................................. 81

4 Human pseudogenes exhibit patterns of retention, expression and 3’UTR constraint consistent with ceRNA function. 86

4.1 Introduction ............................................. 86
4.2 Materials and Methods ........................................ 89
  4.2.1 Identifying accurate human pseudogene coordinates. . . . . . 89
  4.2.2 Identifying human pseudogene orthologs. .................... 90
  4.2.3 Determination of pseudogene retention and expression. . . 91
  4.2.4 Analysis of pseudogene CDS and 3’UTR substitution differences. ........................................ 91
  4.2.5 Analysis of cancer genes. .................................... 93
  4.2.6 Positive selection analysis. .................................. 93
4.3 Results .................................................................... 95
  4.3.1 Human pseudogene orthologs across 20 mammalian species. 95
  4.3.2 Pseudogene genome-wide expression patterns. ............... 98
  4.3.3 Expressed pseudogenes with 3’UTRs have retained orthologs across mammalian species. ................. 100
  4.3.4 Pseudogene 3’UTRs are under stronger evolutionary constraint. .................................................. 104
  4.3.5 Sequence constraint on expression-correlated pseudogenes. 109
  4.3.6 Tumour suppressor genes are not enriched for pseudogene paralogs. ........................................... 112
  4.3.7 Recent positive selection at pseudogene loci. ................. 114
4.4 Discussion ............................................................ 119

5 Testing gene duplication as a mechanism to restore dosage of X-linked genes. ........................................ 125
  5.1 Introduction ......................................................... 125
  5.2 Materials and Methods ............................................ 130
    5.2.1 Gene duplication status and stratification. ................. 130
    5.2.2 Expression data. ................................................ 131
  5.3 Results .............................................................. 131
5.3.1 Tandem duplications on the X chromosome and autosomes. 131
5.3.2 More frequent tandem duplications on the X relative to
autosomes after sex chromosome formation. . . . . . . . . . 132
5.3.3 Young X-linked duplicates have lower expression levels than
other duplicate categories. . . . . . . . . . . . . . . . . . . . 135
5.3.4 Highest expressed genes are not enriched for young X-linked
duplicates. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137
5.3.5 Net expression of X-linked and autosomal duplicate pairs. 139
5.3.6 Constrained X-linked gene duplication after sex chromo-
some formation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 141
5.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 143

6 Conclusions 147
6.1 Future directions. . . . . . . . . . . . . . . . . . . . . . . . . . 151

Bibliography 156
List of Tables

3.1 Confirmed and predicted miRNAs targeting \textit{BRAF} and \textit{BRAFP1}. 70

4.1 Numbers of retained and 3'UTR-containing pseudogenes expressed in human. 97

4.2 Numbers of pseudogenes with tumour suppressive or oncogenic parents. 113

4.3 Positive selection scores across 22 pseudogene loci. 116

4.4 Top 22 potentially positively selected pseudogenes. 117

4.5 Gene names and descriptions of parent genes from top 22 positively selected pseudogenes. 118

5.1 X chromosome and autosomal tandem duplicate and singleton genes. 132
List of Figures

1.1 Mechanisms of pseudogene formation. .................. 4
1.2 Pathways of small RNA silencing. ..................... 15
1.3 The competitive endogenous RNA hypothesis. ............ 17
1.4 The PTEN ceRNA network. ................................ 27
1.5 The miRNA and siRNA biogenesis pathway. ............... 30
1.6 Progression of vertebrate sex chromosome evolution. .... 36
2.1 PsiDR pseudogene annotation workflow. ................. 43
2.2 Timing of therian sex chromosome formation. ............ 49
3.1 BRAFP1 overexpression can cause B-cell lymphoma. ....... 54
3.2 Pairwise and multiple sequence alignments of human and mouse
  BRAF homologs. .............................................. 61
3.3 Reconstruction of BRAFP1 originating from the BRAF locus. 63
3.4 BRAF and BRAFP1 region-specific substitution rates. ...... 66
3.5 MRE densities across BRAF and BRAFP1 regions. .......... 68
3.6 Predicted MRE locations and sequence conservation in BRAF and
  BRAFP1. ...................................................... 71
3.7 Alignments of additional predicted MREs across BRAF and BRAFP1. 73
3.8 Expression of BRAF and BRAFP1 in primates and mouse. .... 76
3.9 Heatmap of BRAF and BRAFP1 expression in human brain tis-
  sues during embryogenesis. ................................ 78
3.10 Expression of BRAF- and BRAFP1-targeting miRNAs in human. 80
4.1 Human pseudogene orthologs across 20 mammalian species. ... 96
4.2 Human pseudogene and parent gene expression patterns. 99
4.3 Pseudogene expression, biotype and 3’UTR retention. 102
4.4 Pseudogene age correlates with median number of substitutions. 105
4.5 Constraint of 3’UTRs across pseudogene orthologs. 106
4.6 Parent genes exhibit greater constraint in their CDS regions. 108
4.7 Expression correlation of PTEN and PTENP1. 110
4.8 Region substitution comparison of pseudogene and parent gene
   expression correlations. 111
4.9 Human cancer genes with pseudogene paralogs. 114

5.1 Post sex chromosome formation duplication event distributions. 133
5.2 Expression distributions of tandem duplicate and singleton genes. 136
5.3 Tandem duplicate genes by expression categories. 138
5.4 Expression levels by gene duplicability. 139
5.5 X-linked and autosomal dosage-constrained genes. 142
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>ceRNA</td>
<td>competitive endogenous RNA</td>
</tr>
<tr>
<td>circRNA</td>
<td>circular RNA</td>
</tr>
<tr>
<td>dsRNA</td>
<td>double-stranded RNA</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>MRE</td>
<td>miRNA recognition element</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>mya</td>
<td>million years ago</td>
</tr>
<tr>
<td>piRISC</td>
<td>piRNA-Induced Silencing Complex</td>
</tr>
<tr>
<td>piRNA</td>
<td>PIWI-interacting RNA</td>
</tr>
<tr>
<td>psiDR</td>
<td>GENCODE Pseudogene Decoration Resource</td>
</tr>
<tr>
<td>PTC</td>
<td>premature termination codon</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-Induced Silencing Complex</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>siRNA</td>
<td>short-interfering RNA</td>
</tr>
<tr>
<td>TPM</td>
<td>transcripts per million reads</td>
</tr>
<tr>
<td>TSG</td>
<td>tumour suppressor gene</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Evolution of new species and phenotypes often necessitates change in regulation of gene expression. With the exception of some well known cases, such as the FOXP2 gene involved in the onset of human speech and language (Enard et al., 2002), there are surprisingly few functional changes within protein-coding genes between species that are evolutionarily closely related, yet great morphological differences between those species exist. Since 1975 it has been argued that evolution of expression regulation contributes significantly to phenotypic variation (King and Wilson, 1975), and it is clear that dosage constraints have shaped gene expression throughout the genome (Ohno, 1967, 1970; Hurst et al., 2015; Rice and McLysaght, 2017a). Much research has focused on identifying both novel and evolutionarily conserved regulatory elements - and the mechanisms governing them - and this remains an important challenge in elucidating how current trends of gene expression regulation have evolved (Brawand et al., 2011; Preuss, 2012; Franchini and Pollard, 2017).

Gene duplication provides material for expression evolution, and can result in non-coding - yet functional - fates for duplicated paralogs (Stephens, 1951; Nei, 1969; Ohno, 1970; Balakirev and Ayala, 2003; Conrad and Antonarakis, 2007; Conant and Wolfe, 2008). Deleterious or neutral gene duplications that arise can often consequently have one duplicate removed from the repertoire of coding genes by pseudogenisation - the process by which a gene becomes non-coding - leading
to the accumulation of non-coding paralogs in the genome (Jacq et al., 1977; Vanin, 1985; Kent et al., 2003). These genes can also skip pseudogenisation and be removed by gene deletion. Alternatively, both duplicates can be maintained as protein-coding genes that share expression between them, which may represent a mechanism by which gene dosage balance is restored during sex chromosome evolution (Ohno, 1970; Force et al., 1999; Stoltzfus, 1999; Ohno, 1967; Hurst et al., 2015).

1.1 Pseudogenes

The term “pseudogene” was coined in 1977 by Jacq et al. in their paper which examined the DNA structure of tandem repeats encompassing a 5S ribosomal RNA gene in *Xenopus laevis*. They found that each copy of the repeat contained the 5S RNA gene in addition to an almost identical untranscribed pseudogene copy. No RNA product could be purified from the pseudogene, which they subsequently reason to be a non-functional relic of evolution by gene duplication. Pseudogenes are now defined as genomic elements with homology to functional protein-coding or RNA genes, but which are no longer capable of coding for a functional product due to either frame shifts, premature termination codon (PTC) mutations, or abrogation of their regulatory elements (Vanin, 1985; Mighell et al., 2000; Harrison et al., 2001; Balakirev and Ayala, 2003; Zheng et al., 2007).

Protein-coding paralogs of pseudogenes are usually referred to as their parent genes. It is known that, in general, as pseudogenes become older they gradually acquire more mutations that distinguish them from their parents, until eventually little to no sequence similarity can be found, making their identification more difficult (Zhang, 2003). However, since their initial discovery, many thousands of pseudogenes have been identified and annotated using combinations of experimen-
tal and computational approaches including the PseudoPipe (Zhang et al., 2006), PseudoFinder (Lu et al., 2006), and RetroFinder (Zheng and Gerstein, 2006) algorithms, as well as HAVANA manual and automated annotation (Zhang and Gerstein, 2004; Solovyev et al., 2006; Karro et al., 2007; Pei et al., 2012; Navarro and Galante, 2015). Additionally, in the past two decades an increasing number of studies using both molecular and computational experiments have unearthed pseudogenes with functional roles as long non-coding RNAs (lncRNAs) (Balakirev and Ayala, 2003; Poliseno et al., 2010; Milligan and Lipovich, 2015; Hon et al., 2017), suggesting their assumed status as mere non-functional evolutionary relics may be, to some extent, incorrect.

1.1.1 Origins and types of pseudogenes.

Pseudogenes are broadly characterised into three main classes - or biotypes - based on their mode of formation: processed pseudogenes; duplicated pseudogenes; and unitary pseudogenes (Figure 1.1). The relative numbers of each of these types of pseudogenes often varies between species (Karro et al., 2007). Other types of pseudogenes include immunoglobulin (IG) and T-cell receptor (TR) pseudogenes (Pei et al., 2012).

Processed pseudogenes.

In primate lineages processed pseudogenes - also known as retrotransposed pseudogenes or retrocopies - are continually being generated and are more abundant than any other type of pseudogene (Zhang and Gerstein, 2004; Navarro and Galante, 2015). These pseudogenes are formed by the process of retrotransposition, which involves reverse transcription of messenger RNA (mRNA) originating from protein-coding genes (Figure 1.1A). This process is facilitated by two proteins encoded by long interspersed nuclear element 1 (LINE1) transposable ele-
Figure 1.1 | Mechanisms of pseudogene formation.

(A) Processed pseudogene formation involves retrotransposition of mRNA by reverse transcription and reintegration into the genome at an alternative locus. Occasionally these pseudogenes can co-opt neighbouring regulatory elements (green arrow).

(B) Duplicated pseudogenes arise as a consequence of direct duplication of protein-coding gene genomic DNA and subsequent loss of coding potential of one paralog. This can occur as a result of small-scale duplication or whole-genome duplication.

(C) A protein-coding gene which has lost coding potential due to disabling mutations without prior duplication is termed a unitary pseudogene. Pseudogenes can retain the regulatory elements (orange arrow) and expression patterns of their parents despite coding disablements.

*This figure was adapted from Poliseno, 2012.*
ments: an endonuclease and reverse transcriptase enzyme; and an RNA-binding protein (Mathias et al., 1991; Feng et al., 1996; Hohjoh and Singer, 1997). In the cytoplasm, single-stranded complementary DNA (cDNA) retrocopies are produced from coding mRNA strands by reverse transcriptase, which then replicates this DNA strand to form a double-stranded DNA retrocopy (Esnault et al., 2000). This retrocopy is thereupon reintegrated into the genome in a different location to its parent gene. Several taxa, including marmoset, have a large number of species-specific retrocopies (Navarro and Galante, 2015). Retrocopies with no coding potential but with homology to functional protein-coding genes are thus called processed pseudogenes. In human and mouse, 10,844 and 9,422 processed pseudogenes are annotated in their current Ensembl genome builds, respectively (Aken et al., 2017).

As a consequence of how they are created, processed pseudogenes typically possess only the exonic coding sequence (CDS) regions and untranslated regions (UTRs) from their parent genes, in addition to a poly-A tail - which is frequently present - derived from the parental mRNA transcript. As introns are usually spliced out before retrotransposition occurs they are not present in most processed pseudogene DNA sequences, and regulatory regions from their parent genes are also not replicated (Vanin, 1985). However, partially spliced transcripts can also be retrotransposed, resulting in a subset of processed pseudogenes possessing one or more introns from their parents. This can cause difficulty in distinguishing processed and duplicated pseudogenes using computational approaches that rely on presence or absence of intronic DNA as an indicator of molecular biotype (Pei et al., 2012). Additionally, processed pseudogene 5’ ends are frequently truncated compared to their parent genes (Pavlček et al., 2002). Given their mode of formation and presumed lack of regulatory features at their new loci, processed pseudogenes are often considered “dead-on-arrival” when they are created (Vanin,
Notably, in spite of lacking their parental regulatory features, many retrocopies in primates are nevertheless expressed, often with tissue-specific expression patterns (Navarro and Galante, 2015). Some of these retrocopies, termed retrogenes, retain open reading frames and protein-coding functionality, while others are transcribed processed pseudogenes with non-coding regulatory functions (Balakirev and Ayala, 2003; Kaessmann et al., 2009; Milligan and Lipovich, 2015). Retrogenes and processed pseudogenes may gain expression potential by recruiting regulatory features of neighbouring genes and elements (Kaessmann et al., 2009). Alternatively, expression could be achieved through de novo promoter emergence from random sequences, which has recently been shown can occur over relatively short evolutionary time-scales in prokaryotes (Yona et al., 2018). Noteworthy examples of expressed, functional, non-coding pseudogenes include PTENP1, KRASP1, and BRAFP1 (Poliseno et al., 2010; Karreth et al., 2015). These will be discussed in detail in later sections and chapters.

**Duplicated pseudogenes.**

In contrast to processed pseudogenes, duplicated (or unprocessed) pseudogenes are less abundant in the human genome, with 3,390 annotated in the current genome build (Aken et al., 2017). These pseudogenes originate from DNA-based gene duplications resulting from processes such as small-scale duplication by non-allelic homologous recombination, or whole-genome duplication (Olmo, 1970; Dehal and Boore, 2005; Bailey and Eichler, 2006; Kim et al., 2008; Khurana et al., 2010). Gene duplication in this manner often creates functional redundancy, typically with no fitness benefit to retaining both copies as functional proteins (Lynch and Conery, 2000). Thus, while duplicated pseudogenes have characteristics resembling those of coding genes, including the exon, intron, and regula-
tory structures derived from their parents, they have lost coding potential owing to a gain of disabling mutations in the absence of purifying selection (Figure 1.1B). These disablements can result in abrogation of expression or produce a non-functional protein structure. Duplicated pseudogenes often possess the same upstream regulatory elements as their parents, and it has been shown that young DNA-based duplicates have expression patterns that closely mirror their parent genes (Guschanski et al., 2017).

**Unitary pseudogenes.**

Genes that were formerly protein-coding which have acquired abrogating mutations - compromising their coding capability - without prior duplication are termed unitary pseudogenes (Figure 1.1C). They are distinct from processed and duplicated pseudogenes in that there is no parent gene and thus no functional redundancy when a unitary pseudogene is created. For this reason, these pseudogenes are much rarer than either processed or duplicated pseudogenes, with only 103 unitary pseudogenes currently identified in human (Aken et al., 2017). GULOP is a pseudogene in primates and other taxa and is a well-studied example of a unitary pseudogene. It retains functionality (as GULO) in many other mammalian species (Nishikimi et al., 1992). In those mammals, GULO (L-gulono-γ-lactone oxidase) is involved in the biosynthesis of ascorbic acid, also known as vitamin C, which is not synthesised naturally in humans as a result of this pseudogenisation (Drouin et al., 2011). Interestingly, while GULO is largely disabled in bats, it has been shown that this unitary pseudogene has been “reactivated” by specific mutations in a subset of species and is capable of producing a functional GULO protein (Birney, 1976; Cui et al., 2011). This suggests that some unitary pseudogenes, and possibly other types of pseudogenes, with minor disablements could become protein-coding again through disablement-reversing
mutations. However, frequent pseudogenisation and reactivation of a gene such as \textit{GULO} indicates lack of selective pressure to maintain its functional status, at least under recent environmental pressures.

\subsection*{1.1.2 Pseudogenes throughout evolution.}

Processed, duplicated, and unitary pseudogenes are continually being created throughout evolution via various processes (Pei et al., 2012). For instance, unitary pseudogenes have been shown to arise at a relatively uniform rate (Zhang et al., 2010b), and since divergence with mouse \(~75\) million years ago (mya), at least 76 unitary pseudogenes with functional orthologs in other mammalian taxa have arisen in the human lineage. It was also shown that 11 of these are polymorphic pseudogenes in human populations, meaning both coding and non-coding alleles exist (Zhang et al., 2010b). These pseudogenes are non-coding in other primates, at least in the individuals from these species which have been sequenced, suggesting unitary pseudogenes could be resurrected in some human populations or individuals, similar to what is observed for \textit{GULO} in bat species (Cui et al., 2011).

Whole-genome duplication events occurred at the base of the vertebrate lineage (Ohno, 1970; Dehal and Boore, 2005), and although all genes were initially duplicated, only around 30\% of protein-coding genes in human have paralogs resulting from these events (Nakatani et al., 2007; Makino and McLysaght, 2010), suggesting many duplicate copies were lost, potentially by pseudogenisation.

Pseudogenes have been well studied in primate genomes, which were all shown to possess many lineage-specific processed pseudogenes (Navarro and Galante, 2015). Specifically, an average of \(~7,500\) retrocopies were shown to be present across Catarrhine species (Old World monkeys and apes, including human), comprising both retrogenes and processed pseudogenes (Navarro and Galante, 2015).
This study identified 7,831 retrocopies in the human lineage, with 91% of these also annotated in the GENCODE Pseudogene Decoration Resource (psiDR; Pei et al., 2012), suggesting the majority of retrocopies identified are processed pseudogenes rather than retrogenes. Additionally, transcription of these retrocopies was investigated in five Catarrhine species (Navarro and Galante, 2015). 1,304 retrocopies were shown to be transcribed in human, 1,500 in chimpanzee, 1,461 in gorilla, 846 in orangutan, and 1,324 in rhesus, though a distinction was not made between retrogenes and processed pseudogenes. These retrocopies were more likely to be expressed in testis and nervous tissues, and were less frequently found in more highly specialised tissues. Furthermore, in Platyrrhini (New World monkeys), an average of ~10,000 such elements have been found, suggesting a greater number of retrotransposition events within this lineage, which may be a result of higher LINE1 activity (Navarro and Galante, 2015). It has been shown that a burst of retrotransposition by LINE1 elements in ancestral primates resulted in the generation of many processed pseudogenes, in addition to Alu repeat elements (Ohshima et al., 2003; Liu et al., 2009).

Most pseudogenes gradually acquire greater numbers of substitutions with age. It follows then that with time the relationship to the parent gene may become unrecognisable. This loss of similarity would result in difficulties determining an accurate number of pseudogenes in any given genome. It is therefore possible that many duplicated pseudogenes arising from the ancestral vertebrate whole-genome duplications may no longer be recognisable, at which point they might no longer be considered pseudogenes. Additionally, if they have diverged to such an extent it is unlikely that they retain any kind of function, as pseudogene functionality is frequently associated with sequence similarity to its parent gene (Tam et al., 2008; Polisen et al., 2010; Watanabe et al., 2015). Pseudogenes with conserved biological functionality are expected to be under sequence constraint,
and as a consequence may exhibit reduced substitutions relative to pseudogenes of a comparable age.

Some pseudogenes can undergo ectopic gene conversion with their parent genes (Balakirev and Ayala, 2003). It has been shown in chicken that gene conversion between donor pseudogene sequences and IG genes, in the germline and during antibody diversification in somatic cells, can give rise to increased IG gene diversity (Benatar and Ratcliffe, 1993). In addition, IG pseudogenes appear to be highly conserved (Rothenfluh et al., 1995), and together these observations suggest that IG pseudogenes may contribute significantly to antibody diversity in somatic cells (McCormack et al., 1993; Balakirev and Ayala, 2003).

Conservation of pseudogene sequences has been estimated previously (Svensson et al., 2006). As a group it appears that pseudogenes exhibit little to no sequence constraint, though a subset of pseudogenes are likely to be under greater constraint (Pei et al., 2012). It has been suggested that duplicated pseudogenes are more often conserved than processed pseudogenes, though this is due to the fact that many processed pseudogenes are largely primate-specific, and these conservation estimates rely on the presence of orthologous pseudogenes in mouse. Around ~1,000 human pseudogenes are considered conserved by psiDR, based on sequence divergence estimates between human, chimp, and mouse pseudogene orthologs (Pei et al., 2012). This set is enriched for transcribed pseudogenes, hinting at biological functionality. Pseudogene preservation rates were also investigated across 16 organisms using whole-genome multiple sequence alignments to identify pseudogene orthologs, with lizard (*Anolis carolinensis*) being the most distantly related species from human examined. It was shown that human orthologs of duplicated pseudogenes gradually decrease in the extent of their preservation with increasing evolutionary distance (Pei et al., 2012). By contrast, there is an abrupt decrease in the preservation rate of processed pseudogenes between macaque and
mouse, corresponding to the burst of retrotranspositional activity in ancestral pri-
mates which gave rise to many of the processed pseudogenes observed in modern 
primate genomes (Balasubramanian et al., 2009).

Sequence identity between human pseudogene regions (pseudo-CDS and pseudo-
3’UTR) and their parent gene counterparts was also examined by pairwise se-
quence alignment (Pei et al., 2012). It was shown that some pseudogenes exhib-
ited high sequence identity to their parents in their 3’UTRs and low sequence 
identity in their CDS regions, and some pseudogenes showed the opposite trend. 
Additionally, the overall distributions of sequence identity across pseudogenes for 
the CDS regions and 3’UTRs differ, and it was proposed that these regions may 
be under different evolutionary sequence constraints (Pei et al., 2012).

A small set (~20) of long-lived pseudogenes, originating before the human-
mouse divergence, with potentially functional properties have been identified, in-
cluding processed pseudogenes originating from the ATX7NL3 and ATX1 parent 
genes (Svensson et al., 2006). The majority of pseudogenes in this set exhibited 
lower sequence divergence between human and mouse than would be expected 
for sequences under no constraint, and have levels of conservation that would 
typically be found for protein-coding genes. Of particular importance to note is 
that the downstream region (1000 nucleotides long) of the ATX1 pseudogene -
corresponding to the pseudo-3’UTR - showed unexpectedly high sequence con-
straint (Svensson et al., 2006), indicative of biological functionality of this region. 
These studies suggest pseudogenes which have been retained over long evolu-
tionary time-scales may represent an important set of elements to investigate 
for non-coding functionality, though more recently evolved pseudogenes should 
not be discounted. For instance, the PTENP1 and BRAFP1 pseudogenes are 
lineage-specific - found only in Catarrhine primates, and thus are <$\sim$40 mya (Ku-
mar et al., 2017) - yet they exhibit functionality in the form of ceRNA activity
(see Chapter 3 for an evolutionary analysis of BRAFP1).

### 1.1.3 Functional pseudogenes.

While often considered evolutionary vestiges of gene duplication and retrotransposition, experimental studies have allowed several functional pseudogenes to be characterised. Various roles have been attributed to these non-coding - and coding - RNAs, including as sources of short-interfering RNAs (siRNAs; Muro et al., 2010; Chan et al., 2013) and PIWI-interacting RNAs (piRNAs; Siomi et al., 2011). Interestingly, the Xist RNA gene, which has an important role in X chromosome silencing in eutherians, appears to have evolved from a former pseudogene (Duret et al., 2006), suggesting pseudogenes can be co-opted for novel essential functions. Additionally, pseudogenes can participate in microRNA (miRNA)-mediated post-transcriptional regulation in the form of competitive endogenous RNA (ceRNA) activity (Seitz, 2009; Poliseno et al., 2010; Karreth et al., 2015). Many pseudogenes linked to ceRNA activity have been discovered in recent years, and these will be discussed in later sections.

### Protein-coding “pseudo-pseudogenes”.

*In silico* analyses have been used to detect pseudogenes based on the presence of sequence substitutions that introduce PTCs which should result in a non-functional protein product. Interestingly, other lines of evidence have shown that some of these elements do in fact produce a functional protein, giving rise to the term “pseudo-pseudogene” (Prieto-Godino et al., 2016). For instance, *in silico* evidence suggested that the chimeric jingwei gene in Drosophila species represented a processed pseudogene derived from mRNA transcribed from the alcohol dehydrogenase gene (Jeffs and Ashburner, 1991). However, subsequent *in vivo* analysis demonstrated that this gene has acquired exons and introns from
an upstream unrelated gene, and a functional alcohol dehydrogenase protein can be produced from the locus with expression patterns differing from the original parent gene (Long and Langley, 1993; Wang et al., 2000).

Furthermore, in *Drosophila sechellia* the gene encoding ionotropic glutamate receptor Ir75a possesses an apparent PTC, initially marking this element as an olfactory receptor pseudogene. However, through the mechanism of translational readthrough, which allows translation of mRNA downstream of a stop codon, a functional receptor protein is produced in neuronal cells (Prieto-Godino et al., 2016). This PTC does not appear to markedly affect the overall activity of the receptor, and thus was proposed to be selectively neutral. Three additional olfactory receptor pseudo-pseudogenes that code for functional proteins were also identified in different Drosophila species (Prieto-Godino et al., 2016), and in human 8 new coding genes that were previously annotated as pseudogenes were discovered based on mass spectrometry of the proteome (Kim et al., 2014; Wright et al., 2016). These studies suggest that many other elements that have been annotated as pseudogenes may code for functional proteins by translational readthrough or by other mechanisms (Betrán et al., 2002; Jungreis et al., 2011).

**Non-coding pseudogenes.**

Transcribed pseudogenes with non-coding functional properties are also observable in several genomes, including mouse, and RNA sequencing (RNA-seq) data suggest at least 5-10% of currently annotated pseudogenes show some evidence of expression in human (Pei et al., 2012). These pseudogenes can be sources of small RNA molecules that are capable of regulating parental gene expression levels (Figure 1.2). Antisense expressed pseudogenes have been shown to undergo processing to produce piRNA molecules which have target sites within protein-coding mRNA and IncRNA sequences (Zhou et al., 1992). Such activ-
ity has been shown in mammalian germline cells, including in human mature sperm cells (Pantano et al., 2015) and mouse late spermatocytes (Watanabe et al., 2015), where it was shown that pseudogene-derived piRNAs can regulate their cognate parent genes. In general, piRNAs interact with PIWI proteins to form the piRNA-Induced Silencing Complex (piRISC), which has been shown to mediate epigenetic and post-transcriptional silencing of specific elements in the germline, including retrotransposons, to maintain germline genome integrity (Figure 1.2; Malone and Hannon, 2009; Siomi et al., 2011; Iwasaki et al., 2015). This mechanism of retrotransposon silencing appears to be of crucial importance for mammalian embryonic development and spermatogenesis (Aravin et al., 2008; Wang and Reinke, 2008). Furthermore, pseudogenes can function as natural antisense regulators of their parent genes, such as transcripts of the nitric oxide synthase (NOS) pseudogene which regulate synthesis of NOS protein in snails (Korneev et al., 1999).

Pseudogene transcripts can also give rise to siRNAs, which are utilised in the RNA interference (RNAi) post-transcriptional regulatory pathway. Initially, long endogenous double-stranded RNA transcripts are formed from hairpin or complementary sequences, which are subsequently processed into shorter siRNAs by DICER (Figure 1.2). The siRNA is then unwound to become single-stranded and is incorporated into the RNA-Induced Silencing Complex (RISC), which targets specific transcript sequences for degradation (Filipowicz et al., 2005). In mouse oocytes, it has been shown that siRNA precursor transcripts can be derived from dsRNAs produced by hybridisation of processed protein-coding mRNAs and their homologous pseudogene transcripts (Tam et al., 2008; Watanabe et al., 2008). From these studies it is unclear whether the process of siRNA production alone is sufficient for mRNA repression, due to hybridisation of the mRNA, or if siRNAs directly target sequences for degradation.
One example of pseudogene-derived siRNA activity is the transcribed pseudogene \( \Psi PPM1K \), which has been shown to generate endogenous siRNAs capable of targeting not only its parent gene but also the \( NEK8 \) gene, and potentially others (Chan et al., 2013). Expression of \( \Psi PPM1K \)-derived siRNAs, as well as the parent gene, were significantly downregulated in hepatocellular carcinomas, whereas \( NEK8 \) exhibited higher expression levels in these tumours. Experiments which induced overexpression of \( \Psi PPM1K \) demonstrated that \( PPM1K \) and \( NEK8 \) had reduced expression, provided the siRNA region of \( \Psi PPM1K \) was
not mutated (Chan et al., 2013). This study highlights the potential of transcribed pseudogenes-derived siRNAs as tumour suppressive agents acting independently of their parent genes.

Pseudogenes with inverted repeats can also self-hybridise to form dsRNAs, and can potentially target their parents for degradation by associating with RISC (Tam et al., 2008). These pseudogenes have been shown to possess siRNA functionality in developing mouse oocytes. Their siRNA activity was verified by knocking out DICER - an essential component of the miRNA and siRNA machinery (Figure 1.2) - which resulted in significant upregulation of the would-be target genes. DICER in these cells is essential for normal development, and deletion of DICER results in non-functional gametes with abnormal chromosomal segregation, suggesting pseudogene-derived siRNAs may contribute to this phenotype (Murchison et al., 2007). That some pseudogenes are shown to be sources of siRNAs and piRNAs indicates that the RNAi and/or RISC-based mechanisms are pathways by which a potentially large number of pseudogenes can participate in post-transcriptional expression regulation of protein-coding mRNAs. However, the extent of pseudogene-derived siRNA and piRNA activity remains to be elucidated.

1.2 The competitive endogenous RNA hypothesis.

The term “competitive endogenous RNA” was first coined by Salmena et al. (2011) after their discovery of how the PTENP1 and KRASP1 pseudogenes are capable of regulating the expression of their parent genes in trans by competing for a common pool of miRNA molecules (Poliseno et al., 2010; Figure 1.3A). However, target mimicry - a similar mechanism - had been described in plants
Figure 1.3 | The competitive endogenous RNA hypothesis.
(A) CeRNAs compete in trans for a common pool of miRNA molecules. CeRNAs can be coding or non-coding transcripts. (B) A simplified model of ceRNA activity is illustrated. Expression of a lncRNA, for example an expressed pseudogene, can alter the amount of protein-coding mRNA available for translation by sequestering miRNAs that otherwise would have targeted the coding mRNA. MiRNA concentration can be limiting and they bind to miRNA-recognition elements (MREs) that can be present in different transcripts. Increasing expression of one transcript can thus result in increased expression of the other.

*Part B of this figure was adapted from Salmena et al, 2011.*
previously (Franco-Zorrilla et al., 2007), and the concept of miRNA competition and sequestration was first proposed by Seitz in 2009. In the past decade the ceRNA hypothesis has become an attractive, albeit controversial, model to explain the function of many putatively functionless lncRNAs, including many expressed pseudogenes (Thomson and Dinger, 2016; An et al., 2017). Under this generalised model, there exist transcripts in the cell, in a similar manner to the \(PTENP1\) and \(KRASP1\) pseudogenes (Poliseno et al., 2010), whose role is to titrate out miRNAs such that the expression level of another target sequence is altered (Seitz, 2009; Salmena et al., 2011). Mechanistically, upregulation of a ceRNA transcript will usually result in a greater proportion of cellular miRNAs binding to that transcript. This effect frees up protein-coding mRNA for translation, which otherwise would have been repressed by those same miRNAs, and allows for greater amounts of protein to be produced (Figure 1.3B). Conversely, ceRNA downregulation can result in more miRNAs available to target the coding mRNA partner and reduce the amount of mRNA available for translation. In this fashion, aberrant expression of ceRNAs can have knock-on effects on mRNA availability and protein abundance, with potentially pathogenic consequences (Salmena et al., 2011; Karreth et al., 2015; Thomson and Dinger, 2016).

Several expressed pseudogenes and lncRNAs with experimentally characterised ceRNA functions have been discovered since the ceRNA hypothesis was first proposed. The majority of pseudogene ceRNAs were discovered in cancer contexts (Poliseno et al., 2010; Poliseno, 2012; Karreth et al., 2015; Poliseno and Pandolfi, 2015), though some pseudogenes and lncRNAs have exhibited ceRNA functionality in normal cellular processes (Cesana et al., 2011; Wang et al., 2013b; Zhu et al., 2017). Multiple studies have attempted to experimentally elucidate and/or computationally predict networks of competing transcripts using multiple variables, including common miRNA recognition elements (MREs), and co-expression
or expression correlation of interacting partners (Sumazin et al., 2011; Gennarino et al., 2012; Ala et al., 2013; Yip et al., 2014; Xia et al., 2014a), from both normal and cancer transcriptomes. While there are some examples of ceRNA activity affecting normal cellular processes, in the majority of cases their effect under physiological conditions remains unclear. Additionally, the evolutionary patterns and sequence constraint of these elements has not been investigated in the context of the ceRNA hypothesis.

1.2.1 Examples of specific ceRNA activity.

Long non-coding RNAs

LncRNAs are usually - somewhat arbitrarily - defined as RNA transcripts longer than 200 nucleotides in length which are not translated into a protein product (Ma et al., 2013). Multiple functional roles have been ascribed to lncRNAs, including as ceRNAs. Initial studies of lncRNA evolution suggested that many lncRNAs in the human genome have conserved sequence elements which are under purifying selection across mammals (Ponjavic et al., 2007; Guttman et al., 2009; Haerty and Ponting, 2013), though investigation of transcription revealed that lncRNA expression is not conserved between species (Kutter et al., 2012; Necsulea et al., 2014; Ulitsky, 2016). Given that sequence constraint is generally indicative of biological functionality (Doolittle, 2013; Graur et al., 2013), it is possible that apparently conserved lncRNAs are transcribed in tissues or at developmental stages not interrogated (Hon et al., 2017). However, further investigation of lncRNA sequence constraint and expression in a greater number of species and tissues should reveal the extent of conserved functionality of these elements.

While a general consensus on lncRNA conservation is lacking there have been a plethora of lncRNA elements annotated in human using computational prediction
(Quek et al., 2015; Hon et al., 2017), and there is good evidence in support of many specific lncRNAs with post-transcriptional roles in various cellular processes. Of particular interest are the lncRNAs *linc-MD1* and *lnc-mg*, which have both demonstrated ceRNA activity in myoblast cells (Cesana et al., 2011; Zhu et al., 2017). Specifically, in mouse and human, *linc-MD1* is a muscle-specific lncRNA which regulates the timing of muscle differentiation by competing for miR-133 and miR-135. This competition in turn regulates the *MAML1* and *MEF2C* transcription factors, which are also targeted by these miRNAs. These transcription factors regulate muscle-specific gene expression, and *linc-MD1* downregulation or overexpression experiments have the opposing effects of decreasing or increasing, respectively, biomarkers of muscle differentiation (MYOG and MHC proteins), and both effects are linked to abnormal muscle differentiation (Cesana et al., 2011). Additionally, human myoblasts from patients with Duchenne muscular dystrophy exhibit reduced levels of *linc-MD1*, indicative of a role in human disease.

Further lncRNAs are involved in myogenesis and myogenic differentiation, including *H19* which acts as a molecular sponge for the *let-7* family of miRNAs (Kallen et al., 2013; Neguembor et al., 2014). This miRNA family has been shown to have important roles in regulating developmental processes in *C. elegans* and Drosophila (Reinhart et al., 2000; Caygill and Johnston, 2008), with inferred roles in vertebrate development also (Kloosterman and Plasterk, 2006). Furthermore, the lncRNA *lnc-mg* in mouse also exhibits muscle-specific expression enrichment and has a role in promoting myogenesis and muscle stem cell differentiation by competing for miR-125b to regulate expression of insulin-like growth factor 2 (*Igf-2*) (Zhu et al., 2017). It was shown that overexpression of *lnc-mg* resulted in acceleration of muscle stem cell differentiation, increased levels of *Myod* and *Myog* expression, and enhanced muscle mass in vivo. By contrast,
knockout experiments resulted in differentiation inhibition, and muscular atrophy and weakness (Zhu et al., 2017). In human, another lncRNA, Linc-ROR, has exhibited important ceRNA activity in regulation of embryonic stem cell self-renewal (Wang et al., 2013b). Collectively these studies suggest important, potentially conserved, functionally specialised roles for non-coding ceRNAs in tissue-specific developmental processes.

In addition to ceRNA activity, there have been several different functions ascribed to a number of well-studied lncRNAs. For instance, HOTAIR was the first discovered lncRNA which is capable of influencing transcription in trans on a separate chromosome to which it is located (Rinn et al., 2007). This lncRNA interacts with Polycomb Repressive Complex 2 (PRC2) and facilitates PRC2-mediated histone H3 lysine-27 trimethylation (H3K27me3) of the HOXD locus, resulting in silencing of this gene (Tsai et al., 2010). Dysregulation of HOTAIR has been implicated in the formation of several cancer types, including metastatic breast cancer, gastric cancer, and esophageal squamous cell carcinoma, by either acting as a ceRNA causing HER2 overexpression or reprogramming genomic chromatin states to promote metastasis (Gupta et al., 2010; Chen et al., 2013; Liu et al., 2014).

FIRRE is another example which highlights the varied roles of lncRNAs, where it has been shown to modulate interchromosomal nuclear architecture by interfacing with nuclear matrix factor hnRNPU (Hacisuleyman et al., 2014). Additionally, this X-linked lncRNA interacts with the CTCF transcriptional repressor and has been shown to be capable of anchoring the inactive X chromosome to the nucleolus in mouse embryonic stem cells to maintain X chromosome inactivation mediated by H3K27me3, with evidence for roles in adipogenesis also (Sun et al., 2013; Bergmann et al., 2015; Yang et al., 2015). Furthermore, the ANRIL lncRNA is has been shown to facilitate transcriptional repression of INK4A and
INK4B by interacting with CBX7 in PRC1 and SUZ12 in PRC2, respectively, and polymorphisms which affect the expression of ANRIL have been implicated in a number of diseases including diabetes and various cancers (McPherson et al., 2007; Cunnington et al., 2010; Uno et al., 2010; Yap et al., 2010; Kotake et al., 2011). The diverse roles and disease associations of these lncRNAs highlight the importance of these genomic elements in transcriptional, translational, and epigenetic regulation, and many more lncRNAs with such functions are likely to be elucidated.

Circular RNAs

Circular RNAs (circRNAs) are yet another type of lncRNA. They are often formed from alternatively spliced transcripts of protein-coding genes after their 5' and 3' ends have self-joined at the splice donor and acceptor sites (Capel et al., 1993; Memczak et al., 2013). While some circRNAs have been shown to code for protein (Pamudurti et al., 2017), the majority that have currently been identified appear to be non-coding (Jeck et al., 2013). They are mostly found in the cell cytoplasm at comparatively high levels relative to their linear counterparts from the same gene, and they are generally more stable with longer half-lives (Hansen et al., 2013; Enuka et al., 2016).

Studies have shown that circRNAs make effective miRNA sponges and can participate in ceRNA regulation (Memczak et al., 2013). In particular, an antisense transcript of the CDR1 gene in human has been shown to form an endogenous circRNA, termed either CiRS-7 (circular RNA sponge of miR-7) or CDR1as (CDR1 antisense). CiRS-7 is expressed in human neuronal tissues and possesses over 60 binding sites for miR-7, making it a highly effective sponge for these miRNAs (Hansen et al., 2013; Memczak et al., 2013). Ectopic expression of CiRS-7 in zebrafish resulted in impaired midbrain development as a result
of miR-7 sequestration. Subsequent studies have shown that this circRNA also regulates insulin transcription and secretion in islet cells, again functioning as an endogenous sponge of miR-7 (Xu et al., 2015). Additionally, the first discovered circRNA, derived from the Sry gene in mouse, has also demonstrable ceRNA activity, acting as an effective sponge of miR-138 (Hansen et al., 2013). These studies highlight the important functional roles of non-coding circRNAs as ceRNAs, and show that they likely represent more effective competitors due to their increased abundance and stability (Wilusz and Sharp, 2013).

**Pseudogenes**

Expressed pseudogenes, a subset of IncRNAs, have garnered particular attention as post-transcriptional regulators through ceRNA action. The majority of currently known pseudogene ceRNAs compete with and regulate the expression levels of their parent genes (Thomson and Dinger, 2016), as these two elements are likely to have many MREs in common owing to their shared homology, resulting in more effective ceRNA competition (Smillie et al., 2018). Among the most well-studied ceRNAs is the PTENP1 pseudogene, which has been shown to regulate the expression levels of its parent, the PTEN tumour suppressor gene, in a miRNA-dependent manner (Poliseno et al., 2010; Johnsson et al., 2013; Yu et al., 2014). This pseudogene shares high sequence similarity with its parent, but is a processed retrocopy and possesses a substitution which prevents translation. Multiple different miRNAs, including miR-19a and miR-20b in DU145 prostate cancer cells, bind to MREs common to both PTENP1 and PTEN, resulting in repression of both transcripts. Of particular interest is the fact that expression specifically of the 3’UTR of PTENP1 (i.e., the region of PTENP1 homologous to the 3’UTR of PTEN) was sufficient for ceRNA functionality of the pseudogene, upregulating the levels of PTEN by ~50% in vitro and leading to cellular
growth inhibition. As expected, inhibition of PTENP1 resulted in reduced PTEN mRNA and protein levels. The tumour suppressive role of the PTENP1 3’UTR also extends beyond PTEN, as it was shown that this pseudogene was a more powerful tumour suppressive agent than its parent, and was capable of exerting its effects independently by also competing for miRNAs with additional genes such as p21, a cyclin-dependent kinase inhibitor (Poliseno et al., 2010). PTENP1 is expressed in both normal human tissues and prostate tumour cells, and is significantly correlated with expression of PTEN in both cases. Notably, the PTENP1 locus exhibits copy number losses in a subset of sporadic colon cancers and additional cancer genomes, suggesting PTENP1 may be selectively lost in human cancers (Poliseno et al., 2010; Liu et al., 2017a).

The role of the KRASP1 pseudogene as a regulator of the KRAS proto-oncogene was also investigated (Poliseno et al., 2010). Upregulation of KRASP1 resulted in increased expression of its parent and lead to an increase in cellular growth rates. Amplification of the KRASP1 locus in several human cancers suggests an oncogenic role for this pseudogene. Similarly to KRASP1 and in contrast to PTENP1, the BRAFP1 processed pseudogene (termed Braf-rs1 in mouse) exhibits oncogenic properties in human cancers via ceRNA activity, again regulating its parent gene BRAF (Karreth et al., 2015). This pseudogene contributes to the development of B-cell lymphoma in vitro in human and mouse cell lines and in vivo in mouse, and these effects are again miRNA-dependent. In mouse cells, both the CDS region and 3’UTR of Braf-rs1 exhibit ceRNA activity and oncogenic properties, suggesting MREs important for competition are present in both regions (Karreth et al., 2015). BRAFP1 is discussed in more detail in Chapter 3 of this thesis, where we focus on examining the evolutionary history and signatures of sequence constraint of this pseudogene.

Various other pseudogenes also have described ceRNA functions, including
**Introduction**

*OCT4-P4, TUSC2P, CYP4Z2P, and GBAP1* (Wang et al., 2013a; Rutnam et al., 2014; Zheng et al., 2015; Straniero et al., 2017), each competing with their parents via common MREs in their 3’UTRs. A particular pseudogene of interest is the parent-less *Pbcas4* unitary pseudogene in mouse, which is orthologous to the breast carcinoma amplified sequence 4 (*BCAS4*) gene in human and other mammalian species. *Pbcas4* differs from the trend of parent-pseudogene ceRNA competition, as it was shown that downregulation of *Pbcas4* in the mouse N2A neuroblastoma cell line resulted in significant differential regulation of 96 genes, none paralogs (Marques et al., 2012). Interestingly, many human orthologs of these genes showed positive expression correlation with *BCAS4*. Five of these human and mouse orthologs, in addition to *BCAS4/Pbcas4*, were downregulated upon transfection of miR-185 in both species, which targets these transcripts (Marques et al., 2012).

The fact that human *BCAS4* shares 3’UTR-located MREs with the *Pbcas4* transcript in mouse, and *BCAS4* exhibits expression correlation with orthologs affected by *Pbcas4* silencing, suggests this protein-coding gene in human has a dual role as a ceRNA, which has been retained in *Pbcas4* after having lost coding capability. Several other protein-coding genes also exhibit this dual non-coding function via their 3’UTRs (Jeyapalan et al., 2011; Rutnam and Yang, 2012; Fang et al., 2013; Yang et al., 2014; Li et al., 2015; Gao et al., 2016). These observations of *PTENP1* and *Pbcas4* highlight the potential for expressed pseudogenes to post-transcriptionally regulate genes other than their parents.

In Chapter 4 of this thesis we explore the evolutionary trends and sequence constraint of expressed pseudogenes across mammalian species, with a focus on investigating their potential as widespread post-transcriptional regulators via 3’UTR ceRNA activity.
1.2.2 Cellular ceRNA networks.

Many studies have demonstrated specific interactions between ceRNA pairs, such as between pseudogenes and their parent genes. However, it has also been suggested that ceRNAs can form extensive, global cellular networks of competing RNA transcripts, where the perturbation of one transcript in the network could have widespread knock-on effects that alter expression of a number of different transcripts (Karreth et al., 2011; Sumazin et al., 2011; Tay et al., 2011; Ala et al., 2013; Bosia et al., 2013; Bosson et al., 2014; Jens and Rajewsky, 2014; Tay et al., 2014; Chiu et al., 2017; Smillie et al., 2018). This is made possible due to the existence of many mammalian miRNAs which can potentially target hundreds of transcripts (Bartel, 2009). However, the existence of such networks and the relevance of cellular ceRNA activity in general in non-pathogenic circumstances is disputed. In particular, it is not clear that the ratio of ceRNA to miRNA expression at physiological levels is amenable to competition, because it has been suggested that the number of MREs present in the transcriptome for a given miRNA far outweigh the effective number of miRNA molecules (Denzler et al., 2014, 2016). In this case it was suggested that potential ceRNAs would require upregulation to unrealistic cellular concentrations for any meaningful upregulation of a rival ceRNA to occur.

By contrast, there are experiments utilising single-cell assays that demonstrate extensive ceRNA interaction (Bosia et al., 2017), as well as the multiple examples of functional ceRNAs under normal physiological conditions, including linc-MD1, lnc-mg, CiRS-7, PTEN, and GBAP1, as discussed previously. Indeed, several localised networks of ceRNA interactivity have also been established, and one of the most intensely studied ceRNA networks involves PTEN. In addition to competing with its pseudogene, PTEN appears to exist within a network of ceR-
Figure 1.4 | The PTEN ceRNA network.
A simplified model of the ceRNA network involved in PTEN regulation is illustrated. PTEN transcription and protein translation is indicated in red. Reciprocal or mono-directional ceRNA interactions are given by double-edged or single-edged black arrows, respectively. Green circles highlight indirect interactions, whereas transcripts in the centre circle represent direct interactions. Green motifs within transcripts indicate common or partially common MRE profiles.

This figure was adapted from Poliseno & Pandolfi, 2015.

NAs and interacts in trans with multiple different lncRNAs and protein-coding mRNAs (Figure 1.4), including CNOT6L, VAPA, VCAN, and ZEB2 (Karreth et al., 2011; Sumazin et al., 2011; Tay et al., 2011; Poliseno and Pandolfi, 2015). Silencing of 13 predicted ceRNAs of PTEN using siRNAs resulted in increased tumour cell growth due to decreased PTEN expression, and this effect was shown to be dependent on the 3’UTR of PTEN (Sumazin et al., 2011). Furthermore, studies have shown that vast ceRNA networks of ~7,000 transcripts may exist and have the ability to regulate oncogenic pathways in glioblastoma (Sumazin et al., 2011), as well as in prostate and breast adenocarcinomas (Chiu et al., 2017).
While some ceRNA networks have been elucidated experimentally, there have been several databases and web resources developed which use computational algorithms to predict ceRNA networks based on common MREs and expression profiles. These include StarBase (Li et al., 2014a), InCeDB (Das et al., 2014), ceRDB (Sarver and Subramanian, 2012), miRcode (Jeggari et al., 2012), DIANA-LncBase (Paraskevopoulou et al., 2013), and LncACTdb (Yip et al., 2014). However, many of these resources are based solely on sequence data and they infer miRNA binding interactions using tools such as TargetScan (Lewis et al., 2005) and miRanda (Enright et al., 2003), and there are often vastly different results between them (Thomson et al., 2011). Thus, it appears none yet have the requisite combination of datasets or the correct parameters to accurately and consistently infer reliable networks of ceRNA activity (Thomson and Dinger, 2016). Incorporating comparative genomics data to determine evolutionarily conserved ceRNA elements could result in more accurate ceRNA prediction. The concept of pseudogene 3'UTR constraint as an indicator of biological ceRNA functionality is explored in Chapter 4.

1.3 MicroRNA biogenesis and evolution.

It has become increasingly clear that post-transcriptional regulation of gene expression is of great importance for proper cellular function and homeostasis (Franks et al., 2017; Corbett, 2018). Among the different regulatory mechanisms is the miRNA-mediated silencing system, which is central to the ceRNA hypothesis. MiRNA molecules are ~20-22 nucleotides long and represent an abundant class of small RNAs in eukaryotic transcriptomes that target both coding and non-coding transcripts for degradation and/or translational repression. Target downregulation is achieved by complementary base pairing between MREs in the
target sequence and the miRNA guide sequence which is associated with a protein complex responsible for mediating repression (Bartel, 2004). MiRNAs are now known to be important regulators of gene expression, thought to affect over 60% of protein-coding transcripts (Lewis et al., 2005; Friedman et al., 2009). It is thus not surprising that studies have shown that miRNA dysfunction is associated with disease (Jiang et al., 2009; De Pontual et al., 2011), and that they have been investigated as possible therapeutic agents (Trang et al., 2008; Li et al., 2009). Additionally, miRNAs are the mediators of competition between ceRNAs with common MRE profiles, and dysregulation of miRNA targets can be pathogenic in trans (Poliseno et al., 2010; Salmena et al., 2011; Karreth et al., 2015).

### 1.3.1 Biogenesis and mechanism of action.

MiRNA molecules are transcribed from nuclear DNA, with approximately 40% of miRNA genes found within the introns of coding genes and usually co-regulated with them (Rodriguez et al., 2004). In animals, the primary pathway of miRNA biogenesis (Figure 1.5) involves the initial transcription of primary miRNAs (pri-miRNAs), which can be hundreds of nucleotides long and are capped, polyadenylated, and spliced (Bartel, 2004; Cai et al., 2004; He and Hannon, 2004). Pri-miRNAs are subsequently processed in the nucleus by the DROSHA enzyme into precursor miRNAs (pre-miRNAs), which form into RNA stem-loop, or hairpin, structures often 70 nucleotides in length. Alternatively, similar pre-miRNAs can also be formed directly from the intronic processing of protein-coding RNA transcripts (Figure 1.2), bypassing the DROSHA-mediated cleavage step (Ruby et al., 2007). These miRNA genes are termed mirtrons, with > 1000 mirtrons currently identified computationally in human and mouse (Wen et al., 2015). Some experimentally validated mirtrons exhibit patterns of sequence constraint and evolution consistent with biological regulatory function (Ruby et al., 2007).
Figure 1.5 | The miRNA and siRNA biogenesis pathway.
The biogenesis pathway of miRNAs and siRNAs is illustrated. Pri-miRNAs are processed into pre-miRNAs by DROSHA and exported into the cytoplasm. There, they are further processed into an miRNA:miRNA duplex by DICER, unwound by a helicase enzyme, and one strand is incorporated into the RISC assembly. These miRNAs are used to guide RISC to complimentary transcripts for translational repression and/or degradation.
This figure was adapted from He & Hannon, 2004.
After these initial processing steps pre-miRNAs, from both canonical miRNA genes and miRtrons, are then exported to the cytosol and further processed by DICER, an RNase enzyme, resulting in a 22 base pair (bp) imperfectly bound (i.e., partially mismatched) double-stranded miRNA:miRNA duplex structure (Bartel, 2004). After unwinding of the duplex by a helicase enzyme, one strand usually becomes the mature miRNA and is incorporated into the RISC assembly, while the other strand is degraded (Pratt and MacRae, 2009). The miRNA:RISC assembly targets transcripts for degradation or translational repression using the miRNA sequence as a guide (Figure 1.5). Occasionally, both strands of a miRNA:miRNA duplex can become mature miRNAs that target separate RNA transcripts (Okamura et al., 2008).

RISC refers to a number of different protein complexes that, via guide RNA molecules, have the ability to target specific elements and co-ordinate their silencing (Pratt and MacRae, 2009). These complexes each include a member of the Argonaut (AGO) family; proteins which incorporate the guide RNAs, such as mature miRNA molecules, and which are essential for proper silencing by positioning them in the correct conformation, thus facilitating guide RNA binding to target transcripts (Figure 1.5). Specifically, AGO2 is responsible for miRNA and siRNA-mediated silencing in mammals (Meister et al., 2004; Filipowicz, 2005). Additional members of the RISC assembly include DICER2 (DCR2) and the transactivating response RNA-binding protein (TRBP). TRBP has been shown to recruit DCR2 to the RISC assembly, allowing for DCR2-mediated processing of miRNAs (Chendrimada et al., 2005). Additionally, TRBP and DCR2 associate with AGO2 to allow transfer of the mature siRNAs or miRNAs to AGO2, and this overall complex is subsequently utilised for target specification and silencing (Filipowicz, 2005).

Target silencing via the RISC assembly can take the form of RNA transcript
degradation or prevention of translation, as well as through the formation of heterochromatin (Pratt and MacRae, 2009). The seed sequence of a miRNA is classified as nucleotides 2-8 from the 5' end of the miRNA and this 7 nt region has been shown to be the most important for target recognition (Filipowicz et al., 2008). Transcript degradation can be catalysed by AGO2 when perfect sequence complementarity is present between the miRNA seed region and the MRE within the target transcript, usually within their 3'UTRs, and results in target cleavage (Valencia-Sanchez et al., 2006; Bartel, 2009). However, in metazoans it is more often the case that there are mismatches between the miRNA seed and its target MRE, i.e. partial complementarity. In these instances silencing is usually achieved through translational inhibition via mRNA destabilisation, direct translational repression, or a combination of both (Lim et al., 2005). Upon miRNA binding, accelerated mRNA deacylation (i.e., poly(A) tail removal) occurs via recruiting of the CAF1-CCR4-NOT1 deadenylase complex (Wu et al., 2006; Eulalio et al., 2009). Subsequently, the target mRNA is decapped and undergoes exonucleolytic-mediated degradation (Behm-Ansmant et al., 2006; Eulalio et al., 2007).

Through these silencing mechanisms protein-coding mRNAs and lncRNAs, including expressed pseudogenes and circRNAs, can be repressed by many different miRNA molecules (Lewis et al., 2005). Additionally, it has been shown that individual miRNAs can have hundreds of diverse targets due to abundant target sites among 3'UTRs, coupled with the fact that miRNA seed sequences and MREs are generally only 7-8 nucleotides long (Bartel, 2009). Given this short target size and the abundance of different miRNA families (Kozomara and Griffiths-Jones, 2014; Hammond, 2015), it has been suggested that new MREs are readily able to evolve and influence post-transcriptional regulation, indicating a potential route for novel ceRNAs to acquire functionality (Stark et al., 2005;
Bartel, 2018). Additionally, it has become increasingly important to accurately predict miRNA targets and binding interactions. There are now a diverse set of algorithms dedicated to predicting these interactions, including miRanda and TargetScan (Enright et al., 2003; Lewis et al., 2005), which potential interactions are often scored based on the extent of sequence complementarity between the miRNA seed and its target. However, one of the key factors in accurate prediction is the presence of evolutionary conservation within the MREs of target transcripts (Bartel, 2009). Searching for conserved base pairing between the miRNA seed region and its target MRE(s) appears to greatly reduce the number of false positives that would normally be present when sequence conservation is not considered.

1.3.2 MiRNA and target site conservation.

MiRNA-mediated repression is thought to represent an ancient, conserved, and widespread method of post-transcriptional regulation (Axtell and Bartel, 2005; Chen and Rajewsky, 2007), though the precise mechanisms are different in animals and plants and appear to have evolved independently (Axtell et al., 2011). Computational prediction has yielded estimates of the number of miRNAs in the human genome ranging from 2,588 (Kozomara and Griffiths-Jones, 2014; Hammond, 2015) down to ~600 (Fromm et al., 2015). Given this large number of post-transcriptional regulators with short recognition sequences it is not surprising that many genes are subject to miRNA-mediated regulation. MiRNAs and their targets are often conserved in plants (Axtell and Bartel, 2005), and separate miRNA families and targets are also conserved in animals (Friedman et al., 2009; Bartel, 2018), with studies showing that 90 miRNA families in human are well conserved and are present in species as evolutionarily distant as fish. Among the total complement of human gene 3’UTRs there are an average of over 300 MREs
for each of these conserved miRNAs (Friedman et al., 2009), indicating that they may have abundant conserved targets in the transcriptome.

Novel miRNA genes can evolve from random DNA sequences or by duplication of existing miRNA genes (Wheeler et al., 2009), but are thought to do so relatively slowly despite the apparent plethora of miRNA families in modern genomes, including human. For instance, it has been shown that the average rate of miRNA gene gain per million years is 1.2 to 3.3 genes in *Arabidopsis thaliana* (Fahlgren et al., 2010), though this number likely differs in other taxa. Higher rates of miRNA gene gain at different time periods have been associated with morphological innovation (Heimberg et al., 2008; Wheeler et al., 2009), suggesting evolution of novel miRNA genes may have an important role in genomic evolution.

In general, most conserved MREs have been located within the 3’UTRs of coding genes, though this may be due to difficulties in disentangling signals of MRE conservation from codon constraint in CDS regions. Indeed, some have argued that conserved miRNA sites are present in CDS regions in equal abundance to 3’UTRs (Schnall-Levin et al., 2010). However, the presence of conserved target sites in coding gene 3’UTRs can provide an avenue through which investigation of conserved ceRNA functionality can take place. Functional miRNAs and their target MREs are thought to be under purifying selection (Nozawa et al., 2010), suggesting that ceRNAs may also exhibit patterns of evolutionary constraint. In particular, non-coding RNAs, such as expressed pseudogenes, which have important ceRNA functions would be expected to exhibit sequence constraint of MREs mediating said function, and mirrored constraint of MREs in their ceRNA partners should also be observed. Examining lncRNAs for MRE constraint could suggest a ceRNA function of many lncRNAs with currently unknown biological relevance. Chapter 3 of this thesis demonstrates conservation of specific MREs
in a known expressed pseudogene ceRNA, *BRAFP1*, which are mirrored in its parent gene also.

1.4 Sex chromosome evolution.

An interesting phenomenon in genome evolution is the origins of sex chromosomes, which has occurred multiple times independently across different taxa including many animal species and some plant species (Bachtrog et al., 2014). Sex chromosomes commonly have vastly different sizes, specialised gene content with sex-specific functions and expression, and exhibit reduced rates of recombination (Úbeda et al., 2014; Abbott et al., 2017). The currently accepted model of sex chromosome evolution suggests that they initially evolve from an ancestral pair of homologous autosomes (Figure 1.6) that contain one or more sex-determining genes (Charlesworth et al., 2005; Bachtrog et al., 2011; Abbott et al., 2017), which can occur either in genomes with no pre-existing pair of sex chromosomes or in a system that already has sex chromosomes. In cases with pre-existing sex chromosomes the new sex determining factors would usually require fitness benefits over the old system in order to become fixed (Bachtrog et al., 2014).

Suppression of recombination between the nascent sex chromosomes is thought to occur next, and can be a result of large inversion events (Lahn and Page, 1999; Charlesworth et al., 2005; Lemaitre et al., 2010) or gradual degradation and divergence (Mensah et al., 2014). In regions of newly suppressed recombination the sex-specific chromosome begins to degenerate due to the accumulation of mutations (Figure 1.6; Charlesworth, 1978; Abbott et al., 2017). Due to the attrition of genes on the non-recombining sex chromosome dosage imbalances arise between the sex chromosomes and autosomes in the heterogametic sex. Various mechanisms have been discovered which resolve this problem through
Introduction

Figure 1.6 | Progression of vertebrate sex chromosome evolution from ancestral autosomes.
Sex chromosomes begin to evolve from a pair of autosomes after one allele acquires the male determining factor (MDF) or female determining factor (FDF), resulting in an XY or WZ system, respectively. Some time later recombination is suppressed, either through large inversion events or gradual degradation, leading to divergence between the proto-sex chromosomes. Dashes between chromosomes represent regions that still recombine. The current genomic structures of the sex chromosomes in different species are illustrated.

This figure was adapted from Graves, 2016.

dosage compensation (Ohno, 1967; Charlesworth, 1978), and these mechanisms are variable between vertebrate taxa with independent sex chromosome evolution (Mank, 2013; Graves, 2016).
1.4.1 Therian sex chromosome dosage compensation.

In human and other therian species, males possess one copy of the X chromosome as well as one copy of the Y chromosome, the latter containing the male-determining gene \textit{SRY} (Figure 1.6). This gene is homologous to the \textit{SOX3} gene on the X chromosome, and is a single autosomal locus in some other vertebrates with different sex-determining factors (Foster and Graves, 1994). However, while evolution of the \textit{SRY} gene was specific to therians, \textit{SOX3} and other genes have in fact given rise to sex-determining genes multiple times independently in different lineages, suggesting some genes may be predisposed to function as sex-determining factors (Graves and Peichel, 2010; O’Meally et al., 2012; Graves, 2016).

In contrast to males, females have two copies of the X chromosome, and the majority of the genes on one copy are silenced giving an effective X:autosome expression ratio of 1:2, as in males. This silencing mechanism is termed X chromosome inactivation (XCI), or lyonisation (Lyon, 1961), and is mediated epigenetically by the spreading and coating of the chromosome by the X-inactive specific transcript (\textit{XIST}) lncRNA (Ng et al., 2007; Hoki et al., 2009), which is transcribed solely from its locus on the inactive X (Brown et al., 1991). \textit{XIST} coating prevents transcription of most genes by DNA methylation or post-translational histone modifications (Mohandas et al., 1981; Peterson and Laniel, 2004). The mechanism of XCI involves whole-chromosome silencing, and this appears to be an exceptional case specific to therian species rather than a common mechanism of dosage compensation across vertebrates (Graves, 2016). The timing of XCI onset for different parts of the X chromosome is associated with the timing of evolutionary stratum formation, suggesting XCI evolved in tandem with sex chromosome divergence (McLysaght, 2008). In contrast to placental mammals where
the X chromosome that becomes silenced is randomly chosen, XCI also inactivates one X chromosome in marsupials but appears to always target the paternal chromosome, and XCI in these animals is more tissue-specific and less complete compared to placentals (Cooper et al., 1993). Additionally, the molecular mechanism that mediates silencing differs between these taxa (Kaslow and Migeon, 1987; Mahadevaiah et al., 2009; Rens et al., 2010). However, XCI does not resolve differences in gene expression between the sex chromosomes and autosomes in therian mammals after sex chromosome formation, it extends this problem to females and results in haploid-like, similar expression levels of X-linked genes in both sexes (Julien et al., 2012; Sado and Sakaguchi, 2013).

Further dosage compensation mechanisms are also present in therians, since it was likely there existed expression imbalances between genes on the sex chromosomes and those on the autosomes (Ohno, 1967). These genes were at a 1:1 ratio of expression levels before the sex chromosomes evolved, but after these events this ratio would have been closer to 2:1 autosomal to X-linked expression without dosage compensation, which would likely result in deleterious effects for dosage-sensitive genes (Pessia et al., 2012). It was proposed that a global, chromosome-wide upregulation of X-linked genes occurred to resolve these differences, creating an imbalance in females subsequently resolved by XCI (Ohno, 1967), but evidence suggests otherwise (Xiong et al., 2010; Casci, 2011) and expression level upregulation may be restricted to certain dosage-sensitive genes (Deng et al., 2011; Pessia et al., 2012). Additionally, studies have suggested that downregulation of autosomal interacting partners or retention of homologous Y-linked genes could aid dosage compensation (Julien et al., 2012; Bellott et al., 2014), though it appears likely that resolving expression disparities between X-linked and autosomal genes is mostly on a gene-by-gene basis.

Expression of X-linked genes are constrained by maximal expression levels
due to their haploid-like status (Hurst et al., 2015). The traffic jam hypothesis suggests there is a maximum expression level possible from a single genomic locus before the rate of transcription can physically increase no more (Vicoso and Charlesworth, 2009). These observations suggest that the X chromosome is an unfavourable environment for genes that require high maximal expression for proper function. In Chapter 5 we explore the potential of tandem duplication of high maximally expressed genes on the X chromosome as a mechanism for resolving dosage constraints imposed by maximal expression levels.

1.4.2 Dosage compensation in other taxa.

It has been shown that sex chromosomes among vertebrate lineages have arisen from different pairs of ancestral autosomes, with human genes on the X chromosome mapped to autosomal regions in birds (Nanda et al., 1999), reptiles (Ezaz et al., 2009), and monotremes (Veyrunes et al., 2008; Graves, 2016). Sex chromosomes can have different fates in different species, with some species having completely lost the sex-specific chromosome (Figure 1.6), e.g. crickets and dragonflies (Bachtrog et al., 2014; Abbott et al., 2017). Alternatively, some genomes can undergo frequent turnover of sex chromosomes, resulting in consistently minimal sequence divergence between them (Perrin, 2009). This turnover appears to be common in fish species, either through evolution of a novel sex-determining factor or relocation of the factor(s) from the sex chromosomes to an autosome (Kitano and Peichel, 2012).

While XCI mostly removes any difference in gene expression levels between males and females after sex chromosome formation in therians, other lineages have developed different mechanisms which often do not equalise gene dosage between sexes (Mank, 2013; Graves, 2016). For instance, monotremes possess 5 genetically distinct pairs of sex chromosomes, 10 X chromosomes in females and 5 each of the
X and Y chromosomes in males (Grützner et al., 2004). In general there appears to be a two-fold increase in X-linked gene expression in females compared to males (Julien et al., 2012), though some gene-specific dosage compensation has been observed (Deakin et al., 2008).

Birds and snakes have adopted the ZW system of sex-determination, which differs from the mammalian XY system in that the ovum, rather than the sperm, determines the sex of the offspring (Shetty et al., 1999; Graves, 2016). Thus, in these species (Figure 1.6) males are the homogametic sex (ZZ) whereas females are heterogametic (ZW). The Z chromosome is well conserved among bird species and each species carries an almost identical complement of Z-linked genes (Zhou et al., 2014), though as with monotremes dosage of Z-linked genes between males and females do not appear to be similar, with Z-linked genes exhibiting expression from both Z alleles (Baverstock et al., 1982). Specifically, Z-linked genes in males appear to be expressed at levels 30-40% higher than their female counterparts for most genes (Ellegren et al., 2007; Itoh et al., 2007), though some gene-specific and regional dosage compensation does occur (Melamed and Arnold, 2007; Mank and Ellegren, 2009; McQueen and Clinton, 2009). This pattern is similar in snakes, which do not share homologous sex chromosomes or a sex-determining factor with birds, though they also do not possess global dosage compensatory mechanisms (Vicoso et al., 2013). Additionally, sex chromosomes in fish appear to be quite diverse, and some dosage compensation has been observed, albeit in a regionalised and gene-by-gene manner rather than whole-chromosome based (Chen et al., 2014; Shao et al., 2014; Schultheiß et al., 2015; Graves, 2016). Overall, it appears that mechanisms which create similar dosage levels between males and females in non-therian species are not chromosome-wide, at least in the lineages investigated so far, and in some taxa this appears to be sufficient and compatible with normal functionality.
1.5 Aim

Between sex chromosome dosage compensation and post-transcriptional regulation via trans-acting ceRNA activity it is clear that fine-tuning of gene expression levels for some genes is beneficial for maintaining and/or optimising proper cellular function and homeostasis. In this thesis comparative genomics data is used to explore evolutionary trends exhibited by pseudogenes with potential ceRNA roles. Chapter 3 provides a detailed evolutionary case-study of a previously described pseudogene ceRNA, *BRAFP1*, and the patterns of constraint indicative of such elements. In Chapter 4, informed by observations from the study of *BRAFP1*, I expand the analysis genome-wide and examine expression and evolutionary trends of mammalian pseudogenes, including whether they exhibit signatures of selection consistent with ceRNA roles. Lastly, Chapter 5 investigates the hypothesis that highly expressed genes on the X chromosome have duplicated in tandem to escape constraints imposed by transcriptional traffic jams owing to haploidy after sex chromosome formation.
Chapter 2

Materials and Methods

This chapter provides an introduction to data sources, gene annotations and methods used throughout this thesis.

2.1 Protein-coding gene and pseudogene annotations.

Protein-coding gene IDs and annotations, including gene names, coordinates, descriptions, homolog IDs, transcript IDs, and sequences, were obtained from the Ensembl genome browser (version 93), unless otherwise specified (Aken et al., 2017). The most recent genome build for each species was used where applicable (e.g. GRCh38 for human).

To explore human pseudogene evolutionary trends we made use of pseudogene annotations from the GENCODE Pseudogene Decoration Resource (psiDR; http://pseudogene.org/psidr/) (Pei et al., 2012). Pseudogenes in this resource were identified and annotated using models created by the Human and Vertebrate Analysis and Annotation (HAVANA) manual annotation team (Figure 2.1). A pseudogene was manually annotated based on whether it possesses at least one of the following characteristics, provided transcriptional, functional, or published evidence contradicts it’s potential status as a pseudogene: (1) presence of a PTC relative to the parent CDS; (2) a frame-shift mutation introduced in a
Figure 2.1 | PsiDR pseudogene annotation workflow.
The integrated GENCODE pseudogene annotation method is shown, including manual annotation by HAVANA and automated annotation by the PseudoPipe and RetroFinder algorithms. The 2-way consensus set from the automated pipelines is intersected with the manually annotated pseudogenes. The surveyed set includes pseudogene loci that were identified by all three methods (level 1 pseudogenes), as well as those identified by manual curation only (level 2 pseudogenes). Functional genomics data from ENCODE and the 1000 Genomes project are combined with this surveyed set to create the Pseudogene Decoration Resource (psiDR).

This figure was taken from Pei et al, 2012.
functional domain; (3) truncation of the 5' or 3' end of the CDS relative to the
parent CDS; and (4) a deletion in the CDS relative to the parent CDS (Pei et al.,
2012). Additionally, processed pseudogenes are annotated as such if they do not
possess locus-specific transcriptional evidence, despite a lack of apparent disabling
mutations. The automated pipelines PseudoPide and RetroFinder, capable of
computationally predicting potential pseudogene loci, were used to inform these
models (Zhang et al., 2006; Baertsch et al., 2008). Each pseudogene is assigned
a particular biotype, most commonly either processed or duplicated, based on
their evolutionary origins and mode of formation. Potential pseudogene loci were
checked for protein-coding potential to ensure coding genes were not incorrectly
annotated as pseudogenes, as coding genes that arose through retrotransposition
(i.e., retrogenes) could be mistaken for processed pseudogenes due to structural
differences from their original parent gene paralogs. In total, psiDR provides
a surveyed set of 11,216 manually annotated human pseudogenes (Figure
2.1), though this resource does not appear to be actively maintained or routinely up-
dated. The estimated total number of pseudogenes in the human genome varies
between \( \sim 12,700 \) and \( \sim 19,700 \) (Torrents et al., 2003; Zhang et al., 2004; Pei
et al., 2012). Additionally, there are currently 14,692 pseudogenes are annotated
in Ensembl GRCh38.p12 (Aken et al., 2017).

The functional, protein-coding paralogous of pseudogenes are often called their
parent genes. PsiDR is an incredibly valuable resource as parent genes for the
majority of human pseudogenes have been identified (Pei et al., 2012). These
annotations are crucially important for studying the evolutionary histories of
potentially functional pseudogenes. In psiDR a functional protein-coding paralog
with the greatest sequence similarity to a given pseudogene is considered its parent
gene. Parent genes for 9,368 pseudogenes were identified using this method, with
1,848 pseudogenes remaining ambiguous with regards to parent identity. 3,391
parent genes were identified, with 2,071 genes having a single pseudogene, and 1,320 parent genes that have multiple pseudogenes.

All pseudogene annotations used in this thesis, including their transcript IDs, associated parent gene and transcript IDs, and biotypes, were obtained from the psiDR surveyed set. Pseudogene transcript IDs were used to extract gene IDs and genomic coordinates from Ensembl. Unless otherwise specified we limited our pseudogene set to those which have a previously identified parent gene (n = 9,368). Additionally, we filtered for pseudogenes which have either the processed or duplicated biotype, GTEx expression data, current Ensembl annotations, and for which orthologous relationships could be resolved. Thus, we examine a total set of 8,704 human pseudogenes and their associated parent genes.

2.2 Identifying pseudogene orthologs.

Pseudogenes are typically only annotated in the human genome. We employed several steps to identify pseudogene orthologs across mammalian species. Facilitated by psiDR annotations we extracted pseudogene and parent gene coordinates from Ensembl, for cases where both IDs could be unambiguously resolved (n = 8,704). We then made use of a 20-way whole genome multiple sequence alignment of 20 mammalian species from the UCSC genome browser (Kent et al., 2002) to extract orthologous pseudogene sequences using the mafsInRegion tool. We used parent gene annotations to ensure pseudogene, and not parent gene, sequences were correctly identified as orthologs. The mammalian species included in this alignment are *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys*, *Papio anubis*, *Macaca mulatta*, *Macaca fascicularis*, *Chlorocebus sabaeus*, *Rhinopithecus roxellana*, *Nasalis larvatus*, *Cal lithrix jacchus*, *Saimiri boliviensis*, *Carlito syrichta*, *Otolemur garnetti*, *Microce-
bus murinus, Tupaia belangeri, Mus musculus, and Canis lupus. We manually inspected 100 pseudogenes using BLASTN (version 2.7.1) with default parameters to determine the accuracy of our ortholog identification, and 91 pseudogenes had correctly inferred branches of origin. The remaining 9 pseudogenes had orthologs in more distant branches that were not detected, indicating this method may not be optimised for some pseudogenes. However, we are confident that most of the pseudogenes we analysed have correctly inferred ortholog sequences. Future higher quality genome builds and more robust whole-genome multiple sequence alignments would aid in identifying pseudogene orthologs with greater accuracy. Those pseudogenes with missing orthologs from more distant branches do not affect our subsequent substitution analysis, as the human parent gene is used as the outgroup in each case (see Chapter 4).

2.3 Gene expression data.

The number of publicly available RNA sequencing (RNA-seq) samples and tissues interrogated has expanded massively in the past few years with the development of new, large scale RNA-seq technologies and mapping algorithms. These datasets allow for analyses of gene expression for many genes across a broad range of tissues on a scale previously not possible. In this thesis we frequently use expression data derived from the GTEx Consortium (version 7: https://www.gtexportal.org/home/) to examine expression of both pseudogene and parent gene transcripts (GTEx Consortium, 2015). The Illumina TruSeq library construction protocol was used by GTEx to perform RNA-seq on the samples. Sequenced reads were aligned to the human reference genome based on GENCODE version 19 annotations using STAR v2.4.2a universal RNA-seq aligner (Dobin et al., 2013). This alignment software maps reads to the genome
using an algorithm based on the Maximal Mappable Prefix (MMP) concept, where given a read sequence, read location and a reference genome sequence, the MMP (seed) is taken as the longest read substring, either in the forward or reverse direction, that matches to exactly one or more substrings of the genome sequence. Alignments of entire read sequences are subsequently generated by stitching together all the seeds that were aligned to the genome. A local linear transcription model is assumed when seeds are stitched together, and seeds from the mates of paired-end RNA-seq reads are clustered and stitched concurrently, with each paired-end read represented as a single sequence. STAR generates chimeric alignments if an alignment within one genomic window does not cover the entire read sequence. A local alignment scoring scheme guides the stitching, with user-defined scores (penalties) for matches, mismatches, insertions, deletions and splice junction gaps, which therefore allows for an assessment of the alignment qualities. All alignments with scores within a certain user-defined range below the highest score are reported for multimapping reads (Dobin et al., 2013), though in GTEx reads are uniquely mapped and multi-mapping reads are discarded. This may result in the appearance of lower expression for younger duplicate genes which are more similar in sequence.

Transcripts per million (TPM) values and read counts for each gene were evaluated with RNA-SeQC v1.1.8 using the “-strictMode” option (Li et al., 2010; Deluca et al., 2012). Filters were then applied so that only uniquely mapped reads, reads that were aligned in proper pairs, reads with an alignment distance of \( \leq 6 \) (i.e., no alignments with more than six non-reference bases), and reads that were fully contained within exon boundaries were used. Any reads that overlap intronic sequence were not counted. TPM refers to a measurement of the proportion of transcripts in the pool of RNA for each sample, and is often considered the best metric for quantifying RNA-Seq data and comparing between-
sample expression (Conesa et al., 2016). TPM is given as the rate, or number of read counts per base for each transcript, which is dependent on the total number of fragments sequenced. This is adjusted for by dividing by the sum of all rates to give the proportion of transcripts in the sample. Additionally, this is multiplied by 1,000,000 to enable easier comprehension. TPM calculation is summarised by the following equation, where $X_i$ is the read count and $l_i$ is the effective length:

$$TPM_i = \frac{X_i}{l_i} \times \frac{1}{\sum \frac{X_j}{l_j}} \times 10^6$$

Tissue samples were taken from recently deceased donors based on the following inclusion criteria: $\geq 21$ and $\leq 70$ years of age, BMI $> 18.5$ and $< 35$, less than 24 hours between time of death and tissue collection, no whole blood transfusion within 48 hours before death, no history of metastatic cancer, no chemotherapy or radiation therapy within the 2 years prior to death, and generally unselected for presence or absence of diseases or disorders (GTEx Consortium, 2015). 635 individuals were sampled in total, resulting in RNA-seq data from 11,688 samples across 53 somatic adult tissues. Further information about the protocols used is also available online (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2).

2.4 Gene trees and paralogy.

Human protein-coding gene paralogous relationships, including timing of duplication, used for the analyses in Chapter 5 were obtained from Ensembl Compara annotations (Vilella et al., 2009; Herrero et al., 2016). These relationships were inferred previously from the Gene Orthology/Paralogy prediction method pipeline, which involves reconciling gene trees with species trees to distinguish between duplication and speciation events (i.e., between paralogs and orthologs). Here,
we take the last common ancestor of paralogous genes as their estimated timing of duplication. For instance, it is inferred for two paralogs with last common ancestor Theria that their duplication event occurred after the divergence of therians and monotremes, but before the divergence of eutherians and metatherians (Figure 2.2). These timings were subsequently used to determine whether gene duplications occurred before or after sex chromosome formation. In short, human paralog pairs with duplication events originating on branches after the therian divergence with monotremes are considered young (post-X) duplicates, and those originating before this divergence are considered old (pre-X) duplicates (Figure 2.2)
2.5 Statistical analyses.

Various paired statistical tests are used throughout this thesis. Paired statistical tests are used to compare two related, or dependent, samples to determine whether or not they are derived from the same distribution. Here, paired t-tests (for example for the *BRAF* and *BRAFP1* substitution analyses and the shared MRE analyses) were performed by pairing values from the CDS region and 3’UTR of each pseudogene. These were considered dependent or related values since these regions are located next to one another in the genome and are unlikely to be completely independent, with the null hypothesis being that no difference exists between them. Similarly, for each Wilcoxon signed-rank test between the CDS regions and 3’UTRs across pseudogenes in the genome we pair the median CDS and 3’UTR value of each pseudogene or parent gene and then compare the overall CDS and 3’UTR distributions. All statistical tests are Bonferroni-corrected where appropriate. Statistical analyses were performed in R (version 3.3.1; R Development Core Team, 2016) or in the Python scipy stats package (version 1.0.0; Oliphant, 2007).

2.6 Source code and data availability.

The source code and summary data files associated with the work in this thesis will be made available online at [https://github.com/glenfiec/Thesis](https://github.com/glenfiec/Thesis).
Chapter 3

Sequence constraint of the *BRAFP1* pseudogene indicates conservation of a ceRNA function.

*Much of the research presented in this chapter has been published in Molecular Biology and Evolution (Glenfield and McLysaght, 2018)*

3.1 Introduction

The ceRNA hypothesis proposes that RNAs, expressed concurrently and with a similar complement of MREs, are capable of indirectly regulating one another by competing for a shared, limited pool of miRNA molecules (Seitz, 2009; Poliseno et al., 2010; Salmena et al., 2011; Thomson and Dinger, 2016). CeRNAs can be coding and non-coding transcripts, such as lncRNAs, circRNAs, or expressed pseudogenes (Poliseno et al., 2010; Karreth et al., 2011, 2015; Thomson and Dinger, 2016; Zheng et al., 2016b; Zhu et al., 2017). These non-coding transcripts are capable of sequestering miRNAs that otherwise would have targeted protein-coding mRNAs, and thus reduce the amount of mRNA undergoing miRNA-mediated degradation and/or repression of translation initiation. This method of post-transcriptional regulation has been suggested as a unifying theory
explaining the function of many heretofore uncharacterised expressed non-coding elements in the genome (Thomson and Dinger, 2016), and allows for the possibility of small- or large-scale networks of competing RNA transcripts (Tay et al., 2011). However, the ceRNA hypothesis remains controversial (Thomson and Dinger, 2016; Smillie et al., 2018).

There is good evidence in support of ceRNA activity for specific cases, most often for ceRNAs acting in disease states such as cancer. Several non-coding RNAs, mainly pseudogenes, have been shown to be oncogenic or tumour suppressive in trans via ceRNA activity when up-regulated or down-regulated in various distinct cancer types. In particular, processed pseudogenes of the tumour suppressor gene PTEN and oncogene BRAF, PTENP1 and BRAFP1 respectively, alter the mRNA levels of their parent genes and thus have potential tumour suppressive and oncogenic properties (Poliseno et al., 2010; Johnsson et al., 2013; Yu et al., 2014; Karreth et al., 2015). Copy number loss and copy number gain of PTENP1 and BRAFP1, respectively, have also been associated with their respective tumour suppressive and oncogenic potential (Poliseno et al., 2010; Karreth et al., 2015). Additional pseudogenes, such as KRASP1, TUSC2P, OCT4P4, CYP4Z2P, GBAP1, and Pbcas4 (a unitary pseudogene in mouse), have also been shown to exert a ceRNA effect, most often by regulating their parental gene’s expression (Poliseno et al., 2010; Marques et al., 2012; Wang et al., 2013a; Rutnam et al., 2014; Zheng et al., 2015, 2016a; Straniero et al., 2017). Dysregulation of expression of these pseudogenes by transcriptional or copy number alteration can thus result in pathogenic consequences.

The relevance of cellular ceRNA activity in general in non-pathogenic circumstances is disputed, as it is not clear that the ratio of ceRNA to miRNA expression at physiological levels is amenable to competition (Denzler et al., 2014, 2016). However, there are several documented examples of ceRNA activity oc-
currying during normal cellular processes. One such ceRNA is the lncRNA *lnc-mg* in mice, which regulates protein concentration of Insulin-like growth factor 2 to promote myogenesis and myogenic differentiation by competing for miRNA-125b (Zhu et al., 2017). Knockout of this lncRNA resulted in increased muscular atrophy, whereas overexpression had the opposite effect of increasing muscular hypertrophy, effected by miRNA competition. Additionally, during stem cell self-renewal in human, the lncRNA *linc-ROR* has been shown to regulate *OCT4*, *NANOG*, and *SOX2* by functioning as a miRNA sponge and competing with the transcripts of these genes to regulate stem cell maintenance and differentiation (Wang et al., 2013b). These studies hint at a more widespread role for ceRNAs with potentially evolutionarily conserved functionality.

An evolutionary analysis of *PTEN* pseudogenes has found multiple evolutionarily independent origins of these pseudogenes across the mammalian lineage (Tang et al., 2016). Intriguingly, *Heterocephalus glaber* (naked mole rat), a species famed for its longevity and resistance to developing cancer, possesses 17 *PTEN* pseudogene copies, with each copy sharing a common MRE profile in their 3’UTRs (Buffenstein and Jarvis, 2002; Buffenstein, 2008; Seluanov et al., 2009; Tang et al., 2016). The *PTENP1* pseudogene has demonstrated tumour suppressive properties in human and mouse (Poliseno et al., 2010; Poliseno and Pandolfi, 2015), suggesting that the many copies of this pseudogene in *H. glaber* may in part contribute to it’s cancer resistance. Pseudogenes may not always be the defunct genomic relics of gene duplication they are usually thought to be.

In contrast to *PTENP1*, over-expression of the oncogenic *BRAFP1* pseudogene in human and the *Braf-rs1* pseudogene in mouse contributes to the formation of B-cell lymphoma by acting as a ceRNA for its parent gene, *BRAF* (Karreth et al., 2015), thus increasing BRAF protein levels (Figure 3.1). BRAF functions in the MAPK signaling pathway to promote cellular growth and proliferation,
Figure 3.1 | **BRAFP1** overexpression can cause B-cell lymphoma.
Upregulation of **BRAFP1** can result in B-cell lymphoma, and potentially other cancers, by sequestering miRNAs that otherwise would have repressed its parent gene, **BRAF**. Increasing BRAF protein levels enhances MAPK signaling, driving cellular growth and proliferation. Additional mechanisms and/or targets of **BRAFP1**-mediated ceRNA activity may also contribute to lymphoma pathogenesis.

*This figure was adapted from Karreth et al., 2015.*

and upregulation of BRAF activity is associated with many different cancers in human (Davies et al., 2002). **BRAFP1** is also expressed in human melanoma, prostate cancer, and lung cancer cell lines, and copy number gains and transcriptional amplification of the **BRAFP1** locus are present in several additional cancer types found in the Cancer Genome Atlas (Karreth et al., 2015). However, **BRAFP1** is not expressed in primary human B-cells, and whether this pseudogene performs a ceRNA function in normal tissues or developmental processes
Evolutionary constraint is an unbiased arbiter of biological functionality; if a genomic element is under natural selection then it follows that that element contributes to the fitness of the organism (Doolittle, 2013; Graur et al., 2013). Tests for these constraints on potential ceRNA candidates should illuminate further whether these elements have biological functions at normal physiological levels.

Here we perform an evolutionary analysis of $BRAFP1$ in human and $Braf-rs1$ in mouse. We confirm that, despite exhibiting similar oncogenic properties in both species, these pseudogenes have evolved independently post-speciation. We find greater sequence constraint acting on the 3’UTR of $BRAFP1$ in Catarrhine primates, and we show conservation of specific MREs both in the CDS and the 3’UTR of $BRAF$ and $BRAFP1$. Furthermore, we find evidence for expression of $BRAFP1$ in multiple primate species and in many human tissues, including relatively high expression during certain stages of embryonic brain development. These results suggest a physiologically relevant role for $BRAFP1$, and our finding of 3’UTR sequence constraint provides an evolutionary metric that may be of use in identifying additional potentially important pseudogene ceRNAs.

### 3.2 Materials and Methods

#### 3.2.1 Phylogenetic inference of $BRAFP1$ and $Braf-rs1$ evolutionary history.

Human $BRAF$ (ENSG00000157764) and mouse $Braf$ (ENSMUSG00000002413) cDNA along with human $BRAFP1$ (ENSG00000224775) genomic sequences were retrieved from Ensembl (GRCh38.p7 and GRCm38.p5) (Aken et al., 2017). As $Braf-rs1$ is unannotated in mouse this sequence was identified in the mouse
genome using BLASTN (version 2.5.0) with default parameters with Braf as a query (Altschul et al., 1990). Pairwise comparisons between BRAF, Braf, BRAFP1, and Braf-rs1 were conducted using Emboss Needle pairwise sequence alignment (Rice et al., 2000). Since these pseudogenes are not annotated in other species, BRAFP1 and Braf-rs1 sequences were used to query reference genomes in additional primate and rodent species using BLASTN (version 2.5.0) with default parameters to identify orthologous sequences. Potential orthologs were verified by manual inspection for conserved gene order flanking the hits. The resulting sequences were obtained from GenBank (Benson et al., 2013).

To reconstruct the evolutionary origins of these pseudogenes, BRAF and Braf cDNA sequences and BRAFP1 and Braf-rs1 genomic sequences were aligned using MUSCLE multiple sequence alignment (Edgar, 2004). Rat, marmoset and bushbaby BRAF sequences were used as outgroups for mouse and Catarrhine primates, respectively, as Braf-rs1 or BRAFP1 orthologs were not found in these species. In addition, an alignment was also generated for Braf-rs1 and Braf across multiple mouse strain genomes, sourced from the Mouse Genomes Project (Keane et al., 2011). The resulting phylogenies were inferred using the maximum likelihood method based on the Tamura 3-parameter (T92) model (Tamura, 1992). This was deemed to be the best model for these data based on the Bayesian Information Criterion score calculated for each model using the Find Best Model function in MEGA7 (Kumar et al., 2016). Phylogenies constructed using the Tamura-Nei (TN93) model are also consistent with those from T92 (Tamura and Nei, 1993). The trees are drawn to scale with branch lengths measured in number of substitutions per site. For each alignment all positions with less than 60% site coverage were eliminated (i.e. positions in the alignments where there were gaps in more than 40% of the sequences). In total, there were 2,792 nucleotide positions in the final dataset for the BRAFP1 phylogeny and 3,329 positions for
the Braf-rs1 phylogeny. To test each phylogeny the bootstrap method with 2000 replicates was used. For BRAF and BRAFP1 the consensus sequence at each position was calculated using Jalview and the proportion of sequences matching the consensus was also determined (Waterhouse et al., 2009). The sliding window analysis was performed by averaging the proportions in 10 base pair windows and the resulting jitter plot (Figure 3.4A) was generated in R using ggplot2 (Wilkinson, 2011; R Development Core Team, 2016). These evolutionary analyses were conducted in MEGA7 and phylogenetic tree figures were produced in FigTree v1.4.3 (Kumar et al., 2016; Rambaut, 2016).

3.2.2 Region-specific substitution analysis.

3’UTRs and CDS regions of BRAFP1 and it’s orthologs in Catarrhini were identified using BLASTN (version 2.5.0) with default parameters, as they have not been specifically annotated previously. These regions in BRAFP1 and BRAF, including marmoset BRAF, were separated and aligned independently using MUSCLE multiple sequence alignment. The resulting phylogenies were inferred as above for the full BRAFP1 and BRAF sequences. Region-specific average number of substitutions per site for each species were obtained by calculating the distance, or branch length, of each sequence to the most recent common ancestor of BRAF and BRAFP1. Paired t-tests and Mann-Whitney U tests with Bonferroni multiple testing correction were used to determine differences in numbers of substitutions between the different sequence regions.
3.2.3 Prediction of MREs and analysis of validated \textit{BRAF} and \textit{BRAFP1} miRNA interactions.

Table 3.1 lists miRNAs experimentally validated or predicted to bind to \textit{BRAF} and/or \textit{BRAFP1} in human, identified from various sources. Mature sequences for these miRNAs were obtained from \url{http://www.mirbase.org/} and locations of MREs for these miRNAs in \textit{BRAF} and \textit{BRAFP1} across Catarrhine species were predicted using miRanda target prediction software (Enright et al., 2003; Kozomara and Griffiths-Jones, 2014). The multiple sequence alignments with consensus sequences, shown in Figure 3.6B-G and Figure 3.7A-D, were visualised using JalView (Waterhouse et al., 2009). The phastCons constraint track shown in Figure 3.6A is from a 30-way mammalian whole genome alignment and was adapted from the UCSC genome browser (GRCh38, \url{http://genome.ucsc.edu}) (Kent et al., 2002).

In addition, a total set of 2588 human mature miRNA sequences were obtained from miRBase 21 (Kozomara and Griffiths-Jones, 2014). Also, 23,000 genomic sequences of 3,000 bp each from randomised genomic locations, 1,000 sequences from each human chromosome, were generated. MREs for each of these sequences were predicted using miRanda with the total set of human miRNAs to estimate the genomic density of spurious MREs per 100 bp (Figure 3.5). Predicted MRE density was also calculated in a similar fashion for each \textit{BRAF} and \textit{BRAFP1} sequence across Catarrhine species (marmoset \textit{BRAF} was also included), with further separation into CDS regions and 3’UTRs. The distribution of shared MRE density was also calculated. MREs are considered “shared” if a given miRNA is predicted to bind to each sequence in a given group at least once.
3.2.4 *BRAF* and *BRAFP1* expression data.

RNA-seq data for human, gorilla, macaque and mouse tissues were retrieved from multiple studies with publicly available data in the NCBI Gene Expression Omnibus (GEO) database (GSE30352, GSE50781, and GSE57096) (Brawand et al., 2011; Hammoud et al., 2014; Wunderlich et al., 2014). From these datasets *BRAF* and *BRAFP1* FPKM values for each sample were evaluated using the Tuxedo suite of tools with default parameters (Tophat, Bowtie, Cufflinks, Cummerbund) (Trapnell et al., 2012). As the pseudogenes are not annotated in non-human species, *BRAFP1* locations in the gorilla and macaque genomes and *Braf-rs1* in the mouse genome were manually annotated in their respective species’ GTF files to facilitate mapping of *BRAFP1* and *Braf-rs1* RNA reads. Additional data for human *BRAF* and *BRAFP1* for 11,688 samples comprising 53 tissues were retrieved from the GTEx Consortium database (GTEx Consortium, 2015). In this resource a gene is counted as expressed if TPM is $\geq 0.1$ and has a read count of $\geq 6$ in more than 20% of samples in a given tissue. Dot plots were generated in R using ggplot2 and the Spearman correlation analysis was performed using the Python scipy stats package (Oliphant, 2007; Wilkinson, 2011; R Development Core Team, 2016).

RNA-seq data for *BRAF* and *BRAFP1* expression shown in Figure 3.9 was obtained from Gerrelli et al., 2015. In this particular resource a gene is considered expressed at TPM $\geq 0.5$, therefore we only show tissues where *BRAFP1* was expressed $\geq 0.5$ TPM in at least one developmental stage, though values below this level are also shown for stages at which *BRAFP1* drops below 0.5 TPM.
3.2.5 MiRNA expression analysis.

MiRNA expression data was obtained from https://ccb-web.cs.uni-saarland.de/tissueatlas/ for miRNAs predicted to mediate the ceRNA interaction between \textit{BRAF} and \textit{BRAFP1} from Table 3.1 (Ludwig et al., 2016). Raw miRNA expression levels from this resource were used in these analyses. For single tissues interrogated in this resource corresponding to multiple tissues in the GTEx resource we calculated the average \textit{BRAF} and \textit{BRAFP1} expression levels of those tissues. Additionally, total miRNA expression for each tissue was calculated by summing the raw expression values of \textit{BRAF}- and \textit{BRAFP1}-targeting miRNAs for that tissue. Figures were generated in R using ggplot2 (Wilkinson, 2011; R Development Core Team, 2016).

3.3 Results

3.3.1 \textit{BRAFP1} and \textit{Braf-rs1} evolved independently.

Pseudogenes of the oncogene \textit{BRAF} are present in human (\textit{BRAFP1}) and in mouse (\textit{Braf-rs1}), and these processed pseudogenes are capable of functioning as oncogenic ceRNAs in both of these species. It was suggested that these two pseudogenes do not share a common origin due to the high sequence similarity between the pseudogene and the parent gene in each species in addition to their different genomic locations (Karreth et al., 2015). We confirm that the locations of \textit{BRAFP1} and \textit{Braf-rs1} are not syntenic, and that the pseudogenes share greater sequence similarity to their \textit{BRAF} parent genes (88.5\% and 85.9\% in human and mouse, respectively) than to each other (57.8\%; Figure 3.2A). We also confirm that, while the CDS regions of \textit{BRAFP1} and \textit{Braf-rs1} share homology, their 3′UTRs are not homologous, suggesting they have arisen via retrotransposition.
Figure 3.2 | Pairwise and multiple sequence alignments of human and mouse BRAF homologs.

(A) Overview of pairwise sequence alignments using Needle (Rice et al., 2000) between BRAF and BRAFP1 in human and Braf and Braf-rs1 in mouse. Linkers between sequences indicate sequence similarity comparisons. Grey shading outlines homologous sequences. (B) Maximum likelihood phylogeny showing the evolutionary origins of BRAFP1 and Braf-rs1 in primates and rodents. The phylogeny was generated from the MUSCLE (Edgar, 2004) multiple sequence alignment of BRAF cDNA transcripts and BRAFP1/Braf-rs1 genomic DNA from each species. (C) Maximum likelihood phylogeny of Braf and Braf-rs1 showing Braf-rs1 is species, but not strain, specific. The phylogeny was generated as above for (B). The human BRAF sequence was used as the outgroup. Sequences from wild-derived mouse strains are highlighted in red (Keane et al., 2011). Bootstrap values (2000 replicates) for each phylogeny are shown beside each branch where confidence is < 100%. The scales are given in number of substitutions per site.
of different \textit{BRAF} mRNA splice variants. For convenience we refer to the CDS or 3'UTR of the pseudogenes even though we understand these to be non-coding. The pseudo-CDS region and pseudo-3'UTR are identified by sequence alignment with the parent gene, in this case \textit{BRAF} in human and \textit{Braf} in mouse. No orthologs of \textit{BRAFP1} or \textit{Braf-rs1} (or additional \textit{BRAF}-derived pseudogenes) could be detected in mouse or in human, respectively, suggesting independent pseudogene formation or reciprocal loss post-speciation.

To determine the timing of the origins of \textit{BRAFP1} and \textit{Braf-rs1} we performed a multiple sequence alignment and maximum likelihood phylogenetic analysis of these pseudogenes and their \textit{BRAF/Braf} parent genes (Figure 3.2B,C, see Materials and Methods for details). Orthologs of \textit{BRAFP1} were discovered in each species of the Catarrhine lineage, which comprises the apes and Old World monkeys, but no other species, supporting a single origin at the base of \textit{Catarrhini} (28 - 43 mya) (Kumar et al., 2017). \textit{Braf-rs1} could not be found in any other species tested, including rat, suggesting a species-specific pseudogene origination event in mouse (< 23 mya). The phylogenetic analysis of \textit{Braf-rs1} and \textit{Braf} in multiple wild- and laboratory-derived mouse strains (Keane et al., 2011) shows that \textit{Braf-rs1} is not specific to the canonical mouse reference genome (C57BL), but is present in every strain tested. This suggests that \textit{Braf-rs1} originated shortly after divergence of mouse and rat lineages. The fact that \textit{Braf-rs1} clusters with mouse \textit{Braf} separate from rat \textit{Braf} further indicates \textit{Braf-rs1} is not orthologous with \textit{BRAFP1} in primates.

From our alignments of \textit{BRAF} and \textit{BRAFP1} we note a region proximal to the 5' end of \textit{BRAFP1} that does not have homology with the CDS of \textit{BRAF}, initially suggesting an insertion event post pseudogene formation (Figure 3.2A). Using sequence similarity searches and reconstructing the pseudogene “exon” structure we find that this region - which is not homologous to \textit{BRAF} cDNA - actually orig-
**BRAFP1** constraint indicates conservation of ceRNA function

Figure 3.3 | Reconstruction of **BRAFP1** originating from the **BRAF** locus.  
(A) Illustration of homology between **BRAFP1** genomic DNA and the exons and one intron of **BRAF**. Genomic co-ordinates of both elements are indicated (genome build GRCh38).  
(B) 6-way reading frame of **BRAF** intronic DNA with homology to **BRAFP1**. Stop codons are denoted by red boxes with stars, with amino acids depicted in other colours. Strand direction and genomic co-ordinates of the intronic sequence are indicated also. The illustration was adapted from Ensembl genome browser (Aken et al., 2017).

...ated from the intron between the first and second exons of **BRAF**, and occurs in the pseudogene in the same order as the exonic DNA (Figure 3.3A). Additionally, examining the homologous intronic region in both Catarrhine species and in the marmoset outgroup species we see that this region has multiple stop codons present in every reading frame (Figure 3.3B), and a substitution rate analysis reveals the region is evolving at a similar rate to surrounding intronic DNA.  

These observations suggest that this intronic fragment is unlikely to have formerly been exonic DNA that was lost post pseudogene formation, indicating that **BRAFP1** may have originated from a non-canonical transcript of **BRAF**. It has previously been shown that long introns are more prone to alternative splicing, and since the **BRAF** intron from which the pseudo-intronic portion of **BRAFP1** is derived from is the longest of the gene, it is possible **BRAFP1** is derived from...
one of these alternative splicing events, possibly produced in error (Kandul and Noor, 2009). Pseudogene annotation efforts suffer from difficulties involved in recognising partially spliced processed pseudogenes such as *BRAFP1* (Pei et al., 2012). In fact, *BRAFP1* is currently annotated as unprocessed - or duplicated - though manual inspection clearly identifies this pseudogene as having arisen through retrotransposition.

These results suggest that the putative ceRNAs *BRAFP1* in primates and *Braf-rs1* in mouse have evolved independently, despite sharing a common parent gene. This finding is especially interesting as these pseudogenes exhibit oncogenic properties in both human and mouse.

### 3.3.2 The 3’UTR of *BRAFP1* is under greater evolutionary sequence constraint.

That *BRAFP1* and *Braf-rs1* both have ceRNA functionality is demonstrated by the observation that an increase in their expression can be oncogenic in both species (Karreth et al., 2015). Specifically, ectopic expression of *Braf-rs1* in mouse cell lines and *BRAFP1* in human cell lines resulted in increased levels of *Braf* and *BRAF*, respectively, elevating ERK phosphorylation and rates of cellular proliferation. These effects were shown to be both *BRAF* and *DICER* dependent, as cell lines with knockouts of these genes did not exhibit a similar phenotype, suggesting a miRNA-mediated mechanism. Additionally, co-expression of *BRAFP1* with a *BRAF* 3’UTR-luciferase reporter in human cells revealed an increase in *BRAF* expression, again in a *DICER* dependent manner, further validating a ceRNA effect. Furthermore, luciferase experiments showed that *Braf-rs1* and *BRAFP1* are able to decoy several different miRNAs which were predicted to target both the pseudogene and its parent gene transcript, particularly at lower miRNA concentrations. Mutation of these miRNA binding sites in the pseudogene transcripts
BRAFP1 constraint indicates conservation of circRNA function

reduced their overall circRNA effectiveness and inhibited their ability to alter the expression of their parents.

However, the question remains as to whether BRAFP1 and Braf-rs1 have a function, circRNA or otherwise, in normal cellular processes. To address this we assessed the level of sequence constraint acting on BRAFP1 in Catarrhine primates. Sequence constraint is an unbiased indicator of biological functionality (Doolittle, 2013; Graur et al., 2013). Pseudogenes (and other IncRNAs) functioning as circRNAs should be expected to show greater sequence constraint acting on their MREs, which for protein-coding genes are most commonly found in the 3’UTR, relative to most of the remainder of their sequence.

Our sliding window analysis of the overall multiple sequence alignment for BRAF and BRAFP1 indicate that some sequence regions from both the CDS and 3’UTR are absent from one or more species (Figure 3.4A). In particular, the 5’ region in the CDS of BRAFP1 appears to be deleted in orangutan. To distinguish between a deletion and poor sequencing depth/genome assembly we BLASTed (BLASTN version 2.7.1, default parameters) human BRAFP1 against a newer, more complete orangutan genome build (Susie.PAby2; https://www.ncbi.nlm.nih.gov/assembly/GCF_002880775.1). We found this region absent in this genome also, suggesting that a species-specific deletion is the most likely scenario. We also note the small region of low similarity proximal to the 5’ end of the CDS region, corresponding to the intronic sequence present in BRAFP1 but not in the transcript of BRAF (also shown in Figure 3.3). We performed separate phylogenetic analyses on the CDS and 3’UTR of BRAF and BRAFP1 orthologs (Figure 3.4C,D, see Materials and Methods for details). From these phylogenies we identified the average number of substitutions per site for each CDS and 3’UTR in each species by calculating the branch lengths of each sequence back to the most recent common ancestor. Rates of mutation and evolution can vary
Figure 3.4 | BRAF and BRAFP1 region-specific substitution rates.  
(A) Sliding window analysis, illustrated as a jitter plot, of primate BRAF and BRAFP1 multiple sequence alignment. The percentage of bases in each window that match the consensus sequence is plotted against position on the transcript. The CDS region and 3’UTR are plotted in blue and orange, respectively.  
(B) Substitutions per site for the CDS/pseudo-CDS and 3’UTR/pseudo-3’UTR are plotted for BRAF and BRAFP1 from each Catarrhine species. Bonferroni-corrected P-values were calculated by Paired t-test for within transcript comparisons (indicated by filled circles • in (B)), and Mann-Whitney U test for between transcripts comparison (indicated by open circles ○).  
(C,D) Independent maximum likelihood phylogenies for the 3’UTR/pseudo-3’UTR (C) and CDS/pseudo-CDS (D), used to calculate the substitutions per site used in (B). Scale is in substitutions per site and branch lengths are indicated.
BRAFP1 constraint indicates conservation of ceRNA function

between different parts of the mammalian genome (Wolfe et al., 1989). Testing for differences in constraint between the 3'UTR and CDS regions controls for this genome-wide variance given these regions are at the same locus.

Comparing the number of substitutions between regions we find that the BRAFP1 3'UTRs have a significantly lower number of substitutions relative to the CDS (Bonferroni-corrected $P = 0.002$, Paired t-test; Figure 3.4B). The pseudogene CDS has a higher number of substitutions than the pseudo-3'UTR and the parent gene CDS, as would be expected for a sequence under no or weaker constraint. Conversely, we find no significant difference between the number of substitutions in the BRAF 3'UTRs and the BRAFP1 3'UTRs (Bonferroni-corrected $P = 0.68$, Mann-Whitney U test), suggesting these sequences are under similar levels of constraint.

These observations show that the 3'UTR of BRAFP1 is under stronger evolutionary constraint than its CDS, which ought to share the same mutation rate. This indicates that the 3'UTR of BRAFP1 has functional properties - which is in keeping with the suggested role of this pseudogene as a ceRNA in normal cellular processes - and raises the question as to whether additional pseudogenes in the genome exhibit similar patterns of constraint. While BRAFP1 nucleotide conservation indicates sequence constraint it may also be useful to compare the RNA secondary structure and RNA accessibility of BRAFP1 transcripts to further evaluate the extent of the constraint acting on this element.

3.3.3 Greater density of shared MREs in the 3'UTR of BRAF and BRAFP1.

If BRAFP1 functions as a ceRNA competing with BRAF for a shared pool of miRNAs, then it follows that we expect to see conservation of MREs corresponding to those miRNAs. To determine whether MRE sequence constraint is present
**Figure 3.5 | MRE densities across **BRAF** and **BRAFP1 regions.**

(A) MRE density per 100bp of **BRAF** and **BRAFP1** by region for all human miRNAs is plotted with the MRE density of random sequences for the X-chromosome and autosomes. Bonferroni-corrected P-values for comparisons between the CDS and 3’UTR of **BRAF** and **BRAFP1** individually are by Paired t-test (●) and for between **BRAF** and **BRAFP1**, and random autosomal and X-chromosome sequence, by Mann-Whitney U test (○). (B) Shared MRE density per 100bp is plotted for separate regions of **BRAF** and **BRAFP1**. Each parent and pseudogene region has a single value for the number of shared MREs per 100bp. MREs are considered shared if the corresponding miRNA is predicted to bind at least once to each **BRAF** and **BRAFP1** sequence, including a related outgroup **BRAF** sequence (marmoset). Bonferroni-corrected P-values for comparisons between the CDS/pseudo-CDS and 3’UTR/pseudo-3’UTR of **BRAF** and **BRAFP1** individually are by Paired t-test (●) and for between **BRAF** and **BRAFP1** by Mann-Whitney U test (○).

We examined the predicted overall MRE profile for **BRAF** and **BRAFP1**, as well as specific MREs for previously validated miRNAs that mediate competition between these transcripts (Karreth et al., 2015). We predicted MREs for the total list of human miRNAs (2,588) across each **BRAF** and **BRAFP1** sequence using miRNA target prediction (Enright et al., 2003; Kozomara and Griffiths-Jones, 2014).

We find that the density of MREs predicted for these sequences are not signif-
BRAFP1 constraint indicates conservation of ceRNA function

icantly different from a set of random genomic sequences (Figure 3.5A). However, these random sequences were derived from genomic DNA and may not accurately reflect the spurious distribution of MREs within transcript sequences. An analysis examining the distribution of spurious hits across the transcriptome, or using shuffled nucleotide sequences of BRAF and BRAFP1, may prove more useful in this estimation. In addition, the MRE densities of the CDS and 3’UTR of BRAFP1 do not differ significantly when considering all MREs, though the CDS of BRAF does have a significantly greater density than its 3’UTR ($P = 1$ and $P = 9.2 \times 10^{-4}$, respectively; Paired t-tests, Bonferroni-corrected). However, comparing the number of shared MREs between the separate regions of these elements - i.e. the number of MREs that are predicted at least once in the CDS or 3’UTR for each BRAF and BRAFP1 ortholog - we find that the 3’UTRs of BRAFP1 and BRAF have a greater number of shared MREs relative to their CDS regions ($P = 0.002$ and $P = 3 \times 10^{-7}$, respectively; Paired t-tests, Bonferroni-corrected; Figure 3.5B). These data suggest that there are a greater number of conserved MREs present in the 3’UTR of these elements.

3.3.4 MREs associated with BRAF and BRAFP1 ceRNA activity are conserved.

We next wished to determine whether specific miRNAs have conserved binding sites within both BRAF and BRAFP1. Focusing on miRNAs previously predicted or validated as regulators of BRAF ($n = 17$, Table 3.1) we found 10 also have MREs in BRAFP1 (Figure 3.6A). Intriguingly, we found that predicted MRE locations match up well with phastCons estimates of local constraint acting on BRAFP1. This is not necessarily surprising as such constraint is often used to initially predict conserved MREs. However, it is interesting to note that MRE locations that we predicted to be present only in BRAFP1 and not in BRAF
**Table 3.1** | Confirmed and predicted miRNAs targeting *BRAF* and *BRAFP1*.

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Location</th>
<th><em>BRAF</em></th>
<th><em>BRAFP1</em></th>
<th>miRanda Score</th>
<th>mirSVR Score</th>
<th>PhastCons Score</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-9</td>
<td>3'UTR</td>
<td>8</td>
<td>8</td>
<td>152</td>
<td>-1.186</td>
<td>0.5751</td>
<td>[1,2,3,4]</td>
</tr>
<tr>
<td>miR-590(1)</td>
<td>3'UTR</td>
<td>8</td>
<td>8</td>
<td>146-147</td>
<td>-0.946</td>
<td>0.5303</td>
<td>[1,2,3]</td>
</tr>
<tr>
<td>miR-410</td>
<td>3'UTR</td>
<td>8</td>
<td>8</td>
<td>120^d</td>
<td>-0.208</td>
<td>0.5341</td>
<td>[1,2]</td>
</tr>
<tr>
<td>miR-182</td>
<td>CDS</td>
<td>8</td>
<td>8</td>
<td>149-154</td>
<td>-</td>
<td>-</td>
<td>[1,3]</td>
</tr>
<tr>
<td>miR-539-5p</td>
<td>3'UTR</td>
<td>7</td>
<td>7</td>
<td>152</td>
<td>-0.125</td>
<td>0.566</td>
<td>[1,2,3]</td>
</tr>
<tr>
<td>miR-30a-5p</td>
<td>CDS</td>
<td>8</td>
<td>6</td>
<td>145-156</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-138(1)</td>
<td>CDS</td>
<td>8</td>
<td>6</td>
<td>150-154</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-539-3p(1)</td>
<td>CDS</td>
<td>8</td>
<td>4</td>
<td>169-173</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-30a-3p(1)</td>
<td>CDS/3'UTR</td>
<td>8</td>
<td>5</td>
<td>144-152</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-138(2)</td>
<td>CDS</td>
<td>8</td>
<td>2</td>
<td>147</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-186</td>
<td>3'UTR</td>
<td>2</td>
<td>6</td>
<td>143</td>
<td>-0.208</td>
<td>0.5014</td>
<td>[1,2]</td>
</tr>
<tr>
<td>miR-579(1)</td>
<td>CDS</td>
<td>8</td>
<td>0</td>
<td>151-155</td>
<td>-</td>
<td>-</td>
<td>[2,5]</td>
</tr>
<tr>
<td>miR-876-3p</td>
<td>CDS</td>
<td>8</td>
<td>0</td>
<td>140-144</td>
<td>-</td>
<td>-</td>
<td>[2,4]</td>
</tr>
<tr>
<td>miR-876-5p(1)</td>
<td>CDS</td>
<td>0</td>
<td>8</td>
<td>142</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-30a-3p(2)</td>
<td>CDS</td>
<td>0</td>
<td>8</td>
<td>140-151</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-590(2)</td>
<td>3'UTR</td>
<td>0</td>
<td>8</td>
<td>143</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-579(2)</td>
<td>CDS</td>
<td>0</td>
<td>7</td>
<td>146</td>
<td>-</td>
<td>-</td>
<td>[2,5]</td>
</tr>
<tr>
<td>miR-579(3)</td>
<td>3'UTR</td>
<td>6</td>
<td>0</td>
<td>145</td>
<td>-</td>
<td>-</td>
<td>[2,5]</td>
</tr>
<tr>
<td>miR-496</td>
<td>3'UTR</td>
<td>6</td>
<td>0</td>
<td>148</td>
<td>-1.232</td>
<td>0.5747</td>
<td>[1,2]</td>
</tr>
<tr>
<td>miR-122</td>
<td>CDS</td>
<td>0</td>
<td>6</td>
<td>152</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-539-3p(2)</td>
<td>CDS</td>
<td>0</td>
<td>5</td>
<td>140</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-579(4)</td>
<td>CDS</td>
<td>0</td>
<td>4</td>
<td>144</td>
<td>-</td>
<td>-</td>
<td>[2,5]</td>
</tr>
<tr>
<td>miR-876-5p(2)</td>
<td>CDS</td>
<td>0</td>
<td>3</td>
<td>140-143</td>
<td>-</td>
<td>-</td>
<td>[2,3]</td>
</tr>
<tr>
<td>miR-137</td>
<td>3'UTR</td>
<td>0</td>
<td>0</td>
<td>^e</td>
<td>-1.215</td>
<td>0.5884</td>
<td>[1]</td>
</tr>
<tr>
<td>miR-340</td>
<td>3'UTR</td>
<td>0</td>
<td>0</td>
<td>^e</td>
<td>-0.201</td>
<td>0.5598</td>
<td>[1]</td>
</tr>
</tbody>
</table>

^a^ mirSVR scores for predicted binding sites were taken from microRNA.org (Betel et al., 2010).

^b^ PhastCons score predicts level of constraint acting on local sequence regions.

^c^ Sources of miRNA binding site prediction and/or confirmed binding interactions.

^d^ Predicted binding site is below miRanda threshold score of 140.

^e^ Predicted by mirSVR but binding site unable to be located using miRanda.

[4] Lewis et al. (2005)
Figure 3.6 | Predicted MRE locations and sequence conservation in BRAF and BRAFP1.
(Continued on the next page.)
Figure 3.6 | Predicted MRE locations and sequence conservation in BRAF and BRAFP1.

(A) Locations of MREs for miRNAs confirmed and/or predicted to regulate BRAF and BRAFP1 are shown with respect to the gene and pseudogene structures. MiRanda target prediction (Enright et al., 2003) was used to locate the putative MREs. Red bars spanning both elements indicate MREs present in both, whereas red bars on one element only indicate element specific MREs. The full list of miRNAs and number of species that each MRE is predicted to be present in can be found in Table 3.1. Lengths of each region, corresponding to the human sequence, are indicated. The phastCons score along the length of BRAFP1 was extracted from the UCSC genome browser (GRCh38) and is from a 30-way mammalian MultiZ alignment, which includes 27 primate genomes (Felsenstein and Churchill, 1996; Kent et al., 2002). (B-G) Predicted conserved MREs across BRAF and BRAFP1 CDS (B-D) and 3’UTR (E-G) in Catarrhini primates. Red shading indicates the seed region, and blue consensus bars for each base denote the proportion of sequences that match the consensus sequence. The position along sequences relative to human BRAFP1 is indicated. The miRNA mature sequences are also shown. Bases in bold red font indicate substitutions predicted to be detrimental to miRNA binding. The long deletion in the pseudo-CDS of orangutan is supported by BLAST searches of two independent genome builds and inspection of intact contigs. Alignments were visualised using JalView (Waterhouse et al., 2009).

also appear to be under local constraint, though some of these do have MREs in BRAF located elsewhere.

Previous experiments have shown that mutation of the MREs for miR-30a-5p (CDS), miR-182 (CDS), and miR-590 (3’UTR) resulted in reduced ceRNA competition between BRAF and BRAFP1 in cancer cell lines (Karreth et al., 2015). We find that these MREs have highly conserved sequence across the pseudogene and parent gene sequences (Figure 3.6C-E). In addition, we find that MREs for miRNAs that have been predicted to be regulators of BRAF by TargetScan (Lewis et al., 2005) and mirSRV (Betel et al., 2010) are conserved in both the parent gene and pseudogene (Figure 3.6B,F,G). Specifically, MREs for miR-9 and miR-539-5p are particularly well conserved, with 100% sequence identity across Catarrhini for miR-9 and a single substitution in BRAFP1 in chimp for miR-539-5p. Both of these MREs are located in the 3’UTR of these transcripts.
**Figure 3.7 | Alignments of additional predicted MREs across *BRAF* and *BRAFP1*.**

(A-D) MREs predicted to bind to both *BRAF* and *BRAFP1* in at least one species are shown for the CDS (A,B) and 3'UTR (C,D). Bases in bold red font indicate substitutions predicted to be detrimental to miRNA binding. The red shaded regions indicate the MRE seed regions. The consensus sequence along with the proportions of bases matching this sequence (blue bars) are shown beneath each alignment. Alignments were visualised using JalView (Waterhouse et al., 2009).
Examples of MREs that are predicted regulators but are not as well conserved are also shown (Figure 3.7A-D).

The greater constraint of the 3’UTR of this pseudogene may potentially reflect a greater density of conserved sites in the 3’UTR, leading to a greater proportion of its sequence exhibiting constraint. However, the spacing of these MREs indicates that those in the 3’UTR may be more important than those in the CDS. Conserved binding sites are more likely to be closer together, separated by 10-130 nucleotides, with the optimal distance deemed experimentally to be 8-40 nucleotides (Grimson et al., 2007; Sætrom et al., 2007). Proximal MREs in this manner have been shown to cooperate to enhance ceRNA activity, and fewer transcripts are required to effectively compete when this cooperation occurs (Denzler et al., 2016). The binding sites we find in the 3’UTR of *BRAF* and *BRAFP1* are within ~40 nucleotides of each other, suggesting these sites may be more important for effective competition between these transcripts (Figure 3.6E-G). Also, conserved sites appear more densely packed in the 3’UTR relative to the CDS region (Figure 3.6A).

These data overall suggest that more conserved MREs are likely to occur within the 3’UTR of *BRAF* and *BRAFP1*, and certain MREs in particular are well conserved between these elements.

### 3.3.5 Expression of *BRAFP1* in human somatic and developmental tissues.

In order for ceRNA post-transcriptional regulation to occur, both the pseudogene and its parent gene must be expressed together. Previous work examining RNA-seq data has indicated that *BRAFP1* is not expressed in normal, non-malignant B-cells (Karreth et al., 2015). Given that overexpression of *BRAFP1* in B-cells has been shown to be oncogenic it is not surprising that it is not expressed in
this tissue under normal circumstances. This does not, however, preclude the possibility of expression of this pseudogene in other tissues or developmental stages.

We first examined expression levels of *BRAF* and *BRAFP1* in RNA-seq data (see Materials and Methods) from multiple tissues in three Catarrhine species, human, gorilla, and macaque (Figure 3.8A-C), as well as *Braf* and *Braf-rs1* in mouse (Figure 3.8D). *BRAFP1* has low or no expression in most samples (< 0.1 fragments per kilobase per million mapped reads (FPKM)), though we find relatively high expression in macaque cerebellum and brain tissues. Additionally, at least one brain tissue sample in each species exhibits expression of *BRAFP1*. By contrast, *BRAF* has a greater breadth of expression levels and for each species is most highly expressed in testis.

To validate these observations in a greater number of tissue samples we utilised additional human expression data from GTEx (GTEx Consortium, 2015) which includes 11,688 samples across 53 tissues. We found that *BRAF* is expressed (i.e., TPM ≥ 0.1 for a given sample) in all samples, whereas *BRAFP1* is expressed in only 1,701 samples (~15% of the total number of samples, including samples from 49/53 tissues; Figure 3.8E). While we observe that *BRAFP1* is consistently expressed at much lower levels than its parent gene, it was previously shown that *BRAFP1* transcripts appear to be more rapidly degraded than *BRAF* transcripts (Karreth et al., 2015). Additionally, these levels are consistent with previous observations showing that *BRAFP1* is an effective ceRNA at comparably low levels, with *BRAFP1* levels being ~15- to ~30-fold less abundant than *BRAF* (Karreth et al., 2015).

*BRAF* and *BRAFP1* expression levels are weakly correlated for those samples in which *BRAFP1* is expressed (\( \rho = 0.14, P = 6.8 \times 10^{-9} \), Spearman rank-order correlation; Figure 3.8E). Though the biological significance of this observation...
BRAFP1 constraint indicates conservation of ceRNA function.

Figure 3.8 | Expression of *BRAF* and *BRAFP1* in primates and mouse. (Continued on the next page.)
in the context of the ceRNA hypothesis is unclear as expression levels do not necessarily need to be correlated in order for a ceRNA effect to take place.

To conservatively determine in which human tissues *BRAFP1* is expressed we follow the GTEx criteria, namely ≥ 0.1 TPM and ≥ 6 reads aligning unambiguously in at least 20% of samples for a given tissue. We found that this pseudogene is expressed in 16 of the 53 tissues based on these criteria (Figure 3.8F). As mentioned, *BRAF* levels are high in testis tissue across several species, and we can see that *BRAFP1* is also expressed in this tissue in human. Notably, none of these somatic brain tissues examined had *BRAFP1* expression levels exceeding this threshold, despite having observed expression in the other RNA-seq brain samples. Furthermore, we find almost no samples of whole blood with *BRAFP1* expression present, consistent with previous observations that it is not expressed in primary human B-cells (Karreth et al., 2015). Examining additional expression data catalogued in the EBI Expression Atlas (Gerrelli et al., 2015) we found that *BRAFP1* and *BRAF* are relatively highly expressed in human developmental brain tissues at certain stages of development (Figure 3.9). In particular, *BRAFP1* is more highly expressed in these tissues (> 5 TPM in choroid plexus at 10 weeks post-conception) than in any samples examined previously, suggesting
**Figure 3.9 | Heatmap of BRAF and BRAFP1 expression in human brain tissues during embryogenesis.**

RNA-seq TPM data for different brain tissues and different stages of human embryogenesis were retrieved from the Human Brain Development Resource for both BRAF and BRAFP1 (Gerrelli et al., 2015). In this resource, elements are considered expressed with a TPM $\geq 0.5$, therefore we only include here tissues and stages which have BRAFP1 expression $\geq 0.5$ in at least one tissue/stage. However, we also display TPM levels between 0.1 and 0.5 for these samples, as BRAFP1 is typically expressed in this range (Figure 3.8A-E). Blank squares indicate stages at which no data was available for that tissue.
this pseudogene may have an important role in embryonic brain development.

Overall, we observe that BRAFP1 is expressed in many human tissues, as well as showing potential expression in additional primate species. These data, coupled with our observation of greater constraint on the 3'UTR in BRAFP1 and presence of conserved MREs, suggests BRAFP1 has a conserved ceRNA role at normal physiological levels in certain tissues and/or at certain stages of development. Additional expression data in non-human Catarrhine species would further illuminate whether BRAFP1 is expressed throughout this clade.

3.3.6 MiRNA expression is not predictive of BRAFP1 or BRAF expression.

One possibility that may account for our observation of BRAFP1 expression in some adult somatic tissues, but not in others, is that increased BRAF levels in these tissues is advantageous. Alternatively, expression levels of BRAF-targeting miRNAs may be higher in these tissues, and BRAFP1 expression may be a compensatory mechanism for this high miRNA expression, restoring some optimal level of BRAF expression. We examined the relationship between BRAF and BRAFP1 expression levels and the raw expression levels of miRNAs predicted to bind to both transcripts, in tissues where expression data for both the transcripts and miRNAs is available. Comparing individual miRNA expression (n = 15) with BRAFP1 expression across 25 tissues we find no significant correlation between expression levels for any miRNA ($P > 0.05$ in all cases, Spearman rank-order correlation tests, Bonferroni-corrected; Figure 3.10A). Similarly, we also find no correlation between BRAF expression and individual miRNA expression levels ($P > 0.05$ in all cases, Spearman rank-order correlation tests, Bonferroni-corrected).

While individual miRNA expression levels do not correlate with BRAF or BRAFP1 expression, the possibility remains that overall miRNA expression of
BRAFP1 constraint indicates conservation of ceRNA function

Figure 3.10 | Expression of BRAF- and BRAFP1-targeting miRNAs in human.
(Continued on the next page.)
Figure 3.10 | Expression of BRAF- and BRAFP1-targeting miRNAs in human.

(A) Dot plots comparing the raw expression levels of individual miRNAs and expression levels of BRAFP1. (B) Raincloud plot of miRNA expression levels in tissues with BRAFP1 expression compared with those in tissues without BRAFP1 expression. P-value was calculated by Mann-Whitney U test. (C) Dot plot of total raw miRNA expression per tissue compared with BRAF expression. \( p \) (rho) and P-value were calculated by Spearman rank-order correlation test. A linear regression is also shown.

those targeting BRAF is higher in tissues in which BRAFP1 is expressed. We pooled miRNA expression levels and compared their expression in tissues with BRAFP1 expression and tissues without BRAFP1 expression. We find no significant difference between these two categories \( (P = 0.91, \) Mann-Whitney U test; Figure 3.10B), suggesting miRNA expression is not higher in BRAFP1-expressing tissues. To briefly examine whether total miRNA expression is predictive of target gene expression level we summed BRAF-targeting miRNA expression levels for each tissue and compared this total expression level with that of BRAF. We find no significant correlation between these measures \( (P = 0.27, \) Spearman rank-order correlation test; Figure 3.10C).

These results suggest that miRNA expression levels are not predictive of expression levels of parent genes or their pseudogenes, at least in the case of BRAF and BRAFP1. Instead, localised sub-cellular concentrations of these transcripts and the miRNAs targeting them may represent a more meaningful predictor.

3.4 Discussion

There has been substantial interest in the ceRNA hypothesis in recent years, with much of the research in the area revolving around how dysregulation of ceRNA expression can affect cancer pathogenicity and progression (Poliseno et al., 2010; Karreth et al., 2011, 2015; Poliseno and Pandolfi, 2015). While some studies
have demonstrated the effects of certain ceRNAs in normal tissues and cells, e.g. *lnc-mg* and *linc-ROR* (Wang et al., 2013b; Zhu et al., 2017), examples of ceRNA function at normal physiological levels is lacking. In particular, some studies dispute that this function could be widespread in normal tissues, claiming that any *bona fide* ceRNAs found would be exceptional cases only (Chang et al., 2004; Denzler et al., 2014, 2016). However, these studies focused on a small number of very highly expressed miRNAs, which are perhaps least amenable to ceRNA activity as the miRNA abundance is unlikely to be limiting.

Here, we examined the *BRAFP1* and *Braf-rs1* pseudogenes with reported oncogenic ceRNA properties in human and mouse, respectively, to determine whether they display the hallmarks of selection - indicative of non-pathogenic cellular roles - that would be expected of a conserved ceRNA (Karreth et al., 2015). Intriguingly, we confirmed that these pseudogenes evolved independently, and for *BRAFP1* we showed that the 3’UTR is under greater evolutionary sequence constraint, explainable as retention of the MREs in this region responsible for mediating ceRNA competition between its parent gene *BRAF* and/or other ceRNAs. The CDS of *BRAFP1* is under reduced constraint relative to the 3’UTR, though we found conserved MREs present across *BRAF* and *BRAFP1* in both the CDS region and the 3’UTR. In mouse, the CDS and 3’UTR of *Braf-rs1* were previously shown to function as ceRNAs independently of one another, suggesting that important MREs may exist in both regions of this transcript (Karreth et al., 2015). This observation was not seen in human, however, although it remains possible considering the distribution of conserved MREs across both regions.

Furthermore, some MREs appear under constraint in *BRAFP1* only, though a subset of these MREs are present in different locations in *BRAF*. This is interesting as it suggests that a particular MRE that contributes to ceRNA competition may not necessarily need to be conserved at one particular location between two
competing transcripts, as long as the same MRE is present somewhere along their sequence. However, we do see MREs that are well conserved between both *BRAF* and *BRAFP1*, indicating that both conserved sites and non-conserved sites could contribute to ceRNA functionality. The possibility that either *BRAF* or *BRAFP1* have other ceRNA partners cannot be discounted, as this could contribute to non-reciprocal MRE constraint in either the parent or the pseudogene.

While both regions can likely exert ceRNA effects, the importance of the 3’UTR specifically to ceRNA crosstalk has recently been demonstrated. Protein-coding transcripts that are predicted to act in trans as ceRNAs for tumour suppressor genes are enriched for 3’UTR shortening through alternative polyadenylation in breast cancer tumours (Xia et al., 2014b; Park et al., 2018). 3’UTR shortening in this manner can significantly alter ceRNA activity and result in reduced protein levels and activity of these tumour suppressor genes. This is explained by a loss of MREs in their ceRNA partners, leading to increased quantities of miRNAs targeting the tumour suppressors instead. Repression of these genes by abolition of ceRNA crosstalk can thus lead to the formation and growth of cancers, highlighting the importance of ceRNA effects and, in particular, of the 3’UTRs of these mRNAs for mediating these effects.

By contrast, *BRAF* is an oncogene rather than a tumour suppressor gene, and upregulation of *BRAFP1* is also oncogeneic (Karreth et al., 2015). However, in certain situations greater amounts of *BRAF* protein are likely required, such as during periods of increased cellular growth and differentiation. *BRAF* is a MEK kinase and its normal function as part of the RAS/MAPK pathway is to activate MEK/ERK and thus promote cellular proliferation, movement, and differentiation. *BRAF* has been shown to be necessary for extra-embryonic placental development, by activating ERK and promoting vascular development (Galabova-Kovacs et al., 2006). *BRAFP1* has high expression in several brain tissues during
various stages of embryonic development, as compared to adult somatic cells, which could facilitate increased \textit{BRAF} levels and is consistent with this pseudogene having an important role during development. Interestingly, \textit{BRAF} and \textit{BRAFP1} have a well conserved MRE in their 3’UTRs for miR-9, which has been shown to have important roles during brain development in vertebrates including neuronal migration and differentiation in the telencephalon (Shibata et al., 2011; Radhakrishnan and Alwin Prem Anand, 2016).

These expression data and our observations of increased 3’UTR sequence constraint suggests \textit{BRAFP1} does have a ceRNA role at normal physiological levels in certain tissues and/or at certain stages of development. However, there is still a large expression disparity between \textit{BRAFP1} and \textit{BRAF} in these tissues, possibly explained by its short half life due to a rapid turnover of \textit{BRAFP1} transcripts relative to \textit{BRAF} (Karreth et al., 2015). This indicates that a ceRNA may not necessarily need to reach high expression levels itself to compete and regulate its competitor(s), even if they are highly expressed. Instead, it may rely more on local subcellular concentrations of the various participants, including the miRNAs. We found that total raw miRNA expression was not predictive of \textit{BRAF} or \textit{BRAFP1} levels. However, these analyses were performed on miRNA and gene/pseudogene expression data derived from difference resources. Testing for miRNA/target expression correlation would benefit from sequencing miRNAs and ceRNAs derived from the same cells, as expression levels are likely more dynamic than those represented by RNA-seq, which essentially provides a snapshot of expression levels at a given time point.

Since pseudogenes have many binding sites for different miRNAs in common with their parents this likely makes them much more effective competitors for single targets (Smillie et al., 2018). Other lncRNAs may be more specialised to sequester much of one specific miRNA, such as the circRNA \textit{CiRS-7} which con-
tains over 60 binding sites for miR-7 (Hansen et al., 2013). These ceRNAs can potentially affect the expression levels of a large number of target genes. Our results are suggestive of a ceRNA role for \textit{BRAFP1} under normal physiological conditions, however our inferences are based on comparative sequence analyses and large-scale RNA-seq data, and are therefore limited. Molecular experiments involving knockdown of \textit{BRAFP1} or mutation of MREs in various different primary cell and tissue types, perhaps informed by tissues interrogated here, should further illuminate the functional relevance of this pseudogene and the extent to which its expression affects its parent gene.

In summary, our results suggest that \textit{BRAFP1} has a conserved ceRNA function in several different tissue types, with a potentially important role during embryonic development. Given that \textit{BRAFP1} has been shown to be overexpressed in several cancer types this pseudogene may also represent a potential target for future drug therapies. As \textit{BRAFP1} is not as widely expressed, targeting this element instead may exhibit reduced side effects compared to a drug which targets its constitutively expressed parent gene (Welsh and Corrie, 2015). Furthermore, we showed that 3’UTR constraint is an evolutionary metric by which additional pseudogenes with biological ceRNA functions may be identified.
Chapter 4

Human pseudogenes exhibit patterns of retention, expression and 3’UTR constraint consistent with ceRNA function.

Much of the research presented in this chapter has been published in Molecular Biology and Evolution (Glenfield and McLysaght, 2018)

4.1 Introduction

Large-scale RNA sequencing efforts have uncovered pervasive transcription throughout the human genome which has resulted in computational identification of thousands of different lncRNAs (Carninci et al., 2005; Cabili et al., 2011; Derrien et al., 2012; Djebali et al., 2012; Iyer et al., 2015; Hon et al., 2017). Accordingly, there are currently 14,720 lncRNAs annotated in the Ensembl GRCh38.p12 genome build (Aken et al., 2017). It is apparent from recent studies that specific expressed pseudogenes and other lncRNAs have functional properties, including as ceRNAs (Poliseno et al., 2010; Schmidt et al., 2011; Karreth et al., 2015; Quek et al., 2015; Thomson and Dinger, 2016; Franchini and Pollard, 2017; Liu et al., 2018).
Human pseudogene evolution consistent with ceRNA function

Some of these genomic elements are well characterised, and often exhibit pathogenic consequences when aberrantly expressed. However, the functional relevance of the majority of annotated lncRNAs remains uncertain (Balakirev and Ayala, 2003; Palazzo and Lee, 2015).

The vast majority of pseudogenes are assumed to be functionless relics of genome evolution, having arisen mainly through retrotransposition or genomic duplication. However, a large number of pseudogenes in the human genome are expressed - at least 5-10% - and some exhibit signatures of evolutionary sequence constraint (Pei et al., 2012). Experimental evidence has shown that a subset of these expressed pseudogenes have functional regulatory properties, capable of acting as ceRNAs, siRNAs, or piRNAs (Poliseno et al., 2010; Muro et al., 2010; Karreth et al., 2015; Siomi et al., 2011; Chan et al., 2013; Thomson and Dinger, 2016). The ceRNA hypothesis proposes that different RNA transcripts within the cell are capable of competing for a shared, limited pool of miRNA molecules (Seitz, 2009; Poliseno et al., 2010; Salmena et al., 2011). In this context, transcripts are able to regulate one another, such as pseudogene or lncRNA transcripts regulating the mRNA levels of protein coding genes. Most currently known pseudogene ceRNAs were shown to regulate expression levels of their parent genes, and were discovered in pathogenic contexts (Poliseno et al., 2010; Karreth et al., 2015; Marques et al., 2012; Wang et al., 2013a; Rutnam et al., 2014; Zheng et al., 2015, 2016a; Straniero et al., 2017). While some experimental studies have demonstrated the potential of dysregulated pseudogene and lncRNA ceRNAs to promote carcinogenesis, such as the BRAFP1 and PTENP1 pseudogenes, it is as of yet unclear whether these elements represent a driving force in cancer formation or whether they are mere passengers of a larger effect.

The overall relevance of cellular ceRNA activity in non-pathogenic circumstances and at physiological levels is largely unknown (Thomson and Dinger,
There has been some suggestion that ceRNAs may comprise a vast network of trans-interacting RNA transcripts, where perturbation of one transcript in the network could have widespread knock-on effects that alter expression of a number of different transcripts (Karreth et al., 2011; Tay et al., 2011; Ala et al., 2013; Bosia et al., 2013; Bosson et al., 2014; Jens and Rajewsky, 2014; Chiu et al., 2017). For example, the CiRS-7 circular RNA has over 60 binding sites for miR-7 and functions as an efficient miRNA sponge (Hansen et al., 2013; Memczak et al., 2013; Xu et al., 2015). CeRNAs with such potent effects may well have far-reaching consequences to a ceRNA network if dysregulated. However, the existence of such networks and the activity of ceRNAs in general is disputed (Tay et al., 2011; Denzler et al., 2014, 2016).

That many pseudogenes with uncertain functions are expressed creates the possibility for a widespread function as ceRNAs, and evolutionary analyses of these pseudogenes should aid in uncovering such a function. Evolutionary constraint is an unbiased arbiter of biological functionality; if a genomic element displays hallmarks of natural selection then it follows that that element contributes to the fitness of the organism (Doolittle, 2013; Graur et al., 2013). Tests for these constraints on potential ceRNA candidates should illuminate further whether these elements have biological functions at normal physiological levels.

In Chapter 3 we demonstrated how the 3'UTR of the BRAFP1 pseudogene is under evolutionary sequence constraint consistent with a ceRNA function. Here, we extend those analyses by providing genome-wide evolutionary evidence suggestive of a ceRNA function for many mammalian pseudogenes. We reason that pseudogenes in particular make excellent ceRNA candidates, provided they are expressed. When formed they automatically have almost identical MRE profiles as their parent genes, and therefore should make effective competitors for the same miRNAs, making these elements an ideal test set to examine conserva-
tion of ceRNA function. We identify previously unannotated human pseudogene orthologs across 20 mammalian genomes. We then show that pseudogenes possessing an identifiable 3’UTR are more likely to be retained across these genomes and are more frequently expressed than pseudogenes without a 3’UTR. Additionally, genome-wide we see that pseudogene 3’UTR sequences are constrained, and, in particular, 3’UTRs of expressed pseudogenes exhibit the strongest constraint. Furthermore, we examine evidence for pseudogenes potentially under recent positive selection in human populations. Our results are indicative of a widespread biological ceRNA function for expressed pseudogenes resulting in sequence conservation of pseudogene 3’UTRs.

4.2 Materials and Methods

4.2.1 Identifying accurate human pseudogene coordinates.

To facilitate evolutionary study we restricted our analyses to pseudogenes with previously identified parent genes, annotated in psiDR by sequence similarity searches (Pei et al., 2012). Gene IDs and genomic coordinates for each pseudogene were obtained from Ensembl (version 90) using the transcript IDs provided by psiDR. Parent gene cDNA sequences were also obtained using parent gene transcript IDs from psiDR. To ensure accurate and full pseudogene coordinates, as many current pseudogene annotations do not include the pseudo-3’UTR, parent gene cDNA sequences were queried against the human reference genome (GRCh38) using BLASTN (version 2.7.1) with output formatted using the following parameter: -outfmt "6 qacc sacc evalue qstart qend sstart send". Only the top 100 hits were selected for further analysis. If the BLAST hit for a given parent gene overlapped with their respective pseudogene, any extension on either flank, with up to a 100 bp gap allowed, was added to the existing pseudogene
coordinates. 4,084 pseudogene annotations were extended using this method, and these new coordinates were used alongside existing coordinates for pseudogenes where no extended overlap was detected. Pseudogene genomic sequences using these updated coordinates were obtained from Ensembl (Aken et al., 2017).

4.2.2 Identifying human pseudogene orthologs.

As human pseudogene orthologs are not currently annotated in the genomes of non-human species we wished to identify these orthologs in mammalian species. A high-quality 20-way mammalian whole genome alignment (GRCh38) was downloaded from the UCSC genome browser (Kent et al., 2002) in MAF format (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/multiz20way/). Using the updated pseudogene coordinates, and limiting our analyses to pseudogenes with parent genes identified, we used the mafsInRegion tool provided by UCSC (hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/mafsInRegion) to extract syntenic orthologous pseudogene sequences across these 20 mammalian species. As this alignment is in MAF format each pseudogene ortholog was extracted in a series of alignment blocks, which were subsequently concatenated to generate full-length pseudogene ortholog sequences. We also do this for the parent gene sequences, to ensure homologous parent genes are not incorrectly identified as pseudogene orthologs. Furthermore, to ensure accurate identification, only orthologs with a length of 50% or more of the human pseudogene sequence were considered. We identify mammalian orthologs (or lack thereof) for 8,704 human pseudogenes across these 20 species.
4.2.3 Determination of pseudogene retention and expression.

The age of each pseudogene was determined by identifying the most distantly related species for which a pseudogene ortholog can be found. The branch of origin of the pseudogene is inferred to be the branch leading to the last common ancestor of all species where the pseudogene is present (Dollo parsimony). A pseudogene is considered “retained” if it is present in every species descending from the branch of origin. This retention analysis was only performed for pseudogenes with presence/absence information in at least 5 species (i.e. pseudogenes originating on the 4th branching point from human or older).

Human pseudogene expression data were obtained from the GTEx Consortium median TPM per tissue data file (version 7, detailed in Chapter 2). A pseudogene is considered expressed if median TPM ≥ 0.1 in its most highly expressed tissue (GTEx Consortium, 2015). Expression correlations between pseudogene-parent gene pairs were calculated by Spearman rank correlation, Bonferroni-corrected for multiple testing. Only pairs where the pseudogene and parent gene are both expressed in at least one tissue were considered (n = 4,779). All samples from each tissue where the pair is expressed were pooled for correlation testing.

Pseudogene biotypes, mostly either processed or duplicated (unprocessed), were determined based on psiDR annotations (Pei et al., 2012).

4.2.4 Analysis of pseudogene CDS and 3’UTR substitution differences.

To identify the locations of pseudo-3’UTRs for each pseudogene, 3’UTR sequences from associated parent gene transcripts were queried against pseudogene sequences using BLASTN and the resulting 3’UTR start coordinate at the 3’ end
was identified. 3'UTRs for 2,331 pseudogenes were detected using this method (3'UTR present), with the remaining pseudogenes considered as having no identifiable 3'UTR (3'UTR absent). For each processed pseudogene with a 3'UTR the CDS and 3'UTR sequences were separated according to the 3'UTR start coordinate identified. For duplicated human pseudogenes and their orthologs the sequence regions corresponding to the CDS of their parent genes were identified by querying the human parent gene CDS against their respective pseudogenes. The resulting hits were joined to form a pseudo-CDS sequence for each duplicated pseudogene and its orthologs.

In order to ensure accurate alignment of these orthologous CDS regions and 3'UTRs the pseudogene and parent gene sequences were filtered according to the following criteria: must be ≥ 200 bp and ≤ 50,000 bp in each region in each species, and must be present in at least 5 species including human. Additionally, using Ensembl gene annotations (GRCh38) we excluded pseudogenes with potentially functional overlapping loci on the opposite strand to ensure these loci did not confound the results (Aken et al., 2017). After filtering there were 1,629 pseudogenes that each had at least 6 (including human parent gene outgroup) orthologous sequences for both the CDS regions and the 3'UTRs. These sequences were subsequently aligned using MUSCLE. We tested these alignments for the best fitting maximum likelihood phylogenetic model for all sites, ranked based on the Bayesian Information Criterion score, using the Find Best Model function in MEGA Compute-Core (Kumar et al., 2012, 2016). The Tamura 3-parameter model (T92) was most frequently the best model based on these sequence data, and we therefore used this model in our subsequent analyses. We also tested other models such as the Tamura-Nei model (TN93), T2 + Gamma distribution, and the Kimura 2-parameter model (K2) (Kimura, 1980; Tamura, 1992; Tamura and Nei, 1993). Our results were consistent across each of these models. 3,258
phylogenies were generated in total. The Dendropy package for python was used to analyse the resulting phylogenies and calculate the distance, in number of substitutions, from each pseudogene ortholog sequence to the most recent common ancestor of the pseudogene (Sukumaran and Holder, 2010). Each pseudogene thus has a single median value for its CDS and 3'UTR, corresponding to the median branch length to the most recent common ancestor of the pseudogene orthologs. These values were then used in the subsequent analyses.

As a positive control, a similar analysis was performed for the CDS regions and 3'UTRs of parent genes (n = 274) using the T92 model. Parent genes with one-to-one orthologs in at least 5 of the species since pseudogene formation and with ≥ 200 bp in each region were used for the analysis, with the corresponding human pseudogene regions used as the outgroups.

### 4.2.5 Analysis of cancer genes.

Lists of cancer genes, including tumour suppressor genes (n = 247), oncogenes (n = 243), and genes annotated as both types (n = 72), were retrieved from the Cancer Gene Census (https://cancer.sanger.ac.uk/census; Futreal et al., 2004) and from Vogelstein et al., 2013. Almost all genes from the latter list were already present in the Cancer Gene Census; thus, only 6 genes from Vogelstein et al. were added to our final set. Human protein-coding gene IDs (n = 22,644) were retrieved from Ensembl (Aken et al., 2017). Pseudogene-parent gene associations were taken from psiDR, as before (Pei et al., 2012).

### 4.2.6 Positive selection analysis.

We obtained pre-computed positive selection statistical measures from the dbP-SHP database of recent positive selection in human populations (http://jjwanglab.org/dbpshp), based on SNP data from the 1000 Genomes Project Phase 1 (http:
Human pseudogene evolution consistent with ceRNA function

(www.internationalgenome.org/data) (The 1000 Genomes Project Consortium, 2010; Li et al., 2014b; Auton et al., 2015). Genotype data from 1,092 individuals from the following 14 populations were included in the dataset: Americans of African Ancestry in SW USA (ASW); Utah Residents with Northern and Western European Ancestry (CEU); Han Chinese in Beijing, China (CHB); Southern Han Chinese (CHS); Colombians from Medellin, Colombia (CLM); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian Population in Spain (IBS); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Mexican Ancestry from Los Angeles USA (MXL); Puerto Ricans from Puerto Rico (PUR); Toscani in Italia (TSI); and Yoruba in Ibadan, Nigeria (YRI). Pseudogene coordinates were converted to GRCh37 to match the database using UCSC LiftOver (Kent et al., 2002). Four different selection measures were chosen to explore pseudogene positive selection: Tajima’s D, derived allele frequency difference (ΔDAF), cross-population composite likelihood ratio (XP-CLR), and fixation index (FST). Pseudogenes were selected based on whether they contain SNPs with any of these measures exceeding their threshold scores as recommended by Li et al., 2014b: Tajima’s D < 0, ΔDAF > 0.2, XP-CLR > 5, and FST > 0.05. Only SNPs located within the bounds of our updated pseudogene coordinates were considered for pseudogene positive selection. It is possible that while these positively selected SNPs do occur within pseudogenes they may be reflecting selection acting on neighbouring elements rather than on the pseudogene itself. For instance, Tajima’s D was calculated using a fixed bin width of 10 kb, and XP-CLR was calculated using 0.1 centimorgan sliding windows across each chromosome, suggesting these individual measures are relatively non-specific (Li et al., 2014b). Additionally, these measures may reflect positive selection acting at different time periods in recent human history. Tajima’s D indicates positive selection occurring <250,000 years ago, ΔDAF <80,000 years ago, and XP-CLR and FST <50,000
to 75,000 years ago (Sabeti et al., 2006).

4.3 Results

4.3.1 Human pseudogene orthologs across 20 mammalian species.

Given that there are several well-characterised pseudogenes with ceRNA functions in human, either under pathogenic or non-pathogenic conditions, it is of interest to examine whether this is a widespread mechanism for post-transcriptional regulation of parent protein-coding gene levels (Poliseno et al., 2010; Karreth et al., 2015; Marques et al., 2012; Wang et al., 2013a; Rutnam et al., 2014; Zheng et al., 2015, 2016a; Straniero et al., 2017). In the previous chapter we showed that a signature of selection on the 3'UTR of the \textit{BRAFP1} pseudogene can be detected by comparing the number of substitutions between its CDS region and 3'UTR. We also infer that such a signature of selection is likely attributable to a ceRNA function. Here we wish to determine if this trend is also seen for pseudogenes on a genome-wide scale. While pseudogenes are annotated, at least in part, in the human genome, the same is not the case for other mammalian genomes, save for mouse. As we wish to explore signatures of selection at pseudogene loci genome-wide, identifying orthologs of human pseudogenes in other mammalian species is necessary for inter-species sequence comparison. Additionally, it is important to ensure the accuracy of pseudogene coordinates in the human genome, such that coordinates include the 3'UTR, for example. Again, for convenience we refer to the CDS or 3'UTR of the pseudogenes even though we understand these to be non-coding.

We analysed pseudogene-parent gene pairs from psiDR (Pei et al., 2012). We
Human pseudogene orthologs across 20 mammalian species.

A time-calibrated phylogenetic tree of 20 mammalian species is shown with the number of pseudogenes originating at each branch, calculated by identifying the most distantly related species in which a pseudogene ortholog can be found. The numbers in brackets indicate the number of retained pseudogenes originating at that branch, i.e. the number of pseudogenes which have an ortholog present in each descendant from that point. Time calibrations were obtained from TimeTree (Kumar et al., 2017).

limited our analyses to pseudogenes with an identified parent gene, in order to have a reliable outgroup sequence. Firstly, we updated genomic coordinates for 4,084 human pseudogenes using sequence similarity to their respective parent gene cDNA sequences. Next, we identified orthologs of 8,704 human pseudogenes across 20 mammalian species by extracting genomic sequences syntenic to human pseudogene annotations from a 20-way high quality whole-genome multiple sequence alignment. The age of each pseudogene was estimated by ascertaining the most distantly related species for which an ortholog can be discovered (Figure 4.1).
Table 4.1 | Numbers of retained and 3’UTR-containing pseudogenes expressed in human.

<table>
<thead>
<tr>
<th></th>
<th>Total set</th>
<th>Retained</th>
<th>Not retained</th>
<th>3’UTR Present</th>
<th>3’UTR Absent</th>
<th>Retained + 3’UTR</th>
<th>Retained - 3’UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressed (≥0.1 TPM)</td>
<td>4944</td>
<td>928</td>
<td>3824</td>
<td>1564</td>
<td>3380</td>
<td>326</td>
<td>602</td>
</tr>
<tr>
<td>Not expressed (&lt;0.1 TPM)</td>
<td>3760</td>
<td>746</td>
<td>2899</td>
<td>767</td>
<td>2993</td>
<td>193</td>
<td>539</td>
</tr>
<tr>
<td>Total #</td>
<td>8704</td>
<td>1674</td>
<td>6723</td>
<td>2331</td>
<td>6373</td>
<td>519</td>
<td>1141</td>
</tr>
<tr>
<td>All Median TPMa</td>
<td>0.1545</td>
<td>0.1311</td>
<td>0.1566</td>
<td>0.2660</td>
<td>0.1291</td>
<td>0.1878</td>
<td>0.1122</td>
</tr>
<tr>
<td>All P-valueb</td>
<td>-</td>
<td>0.015</td>
<td></td>
<td>4.86x10^{-35}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressed only Median</td>
<td>0.6070</td>
<td>0.4732</td>
<td>0.6332</td>
<td>0.6647</td>
<td>0.5919</td>
<td>0.5289</td>
<td>0.4581</td>
</tr>
<tr>
<td>TPMa</td>
<td>-</td>
<td>4.03x10^{-7}</td>
<td></td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressed only P-valueb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processed</td>
<td>7364</td>
<td>1605</td>
<td>5603</td>
<td>2101</td>
<td>5263</td>
<td>508</td>
<td>1097</td>
</tr>
<tr>
<td>Duplicated</td>
<td>1195</td>
<td>61</td>
<td>984</td>
<td>226</td>
<td>969</td>
<td>11</td>
<td>50</td>
</tr>
<tr>
<td>Total #</td>
<td>8559</td>
<td>1666</td>
<td>6587</td>
<td>2327</td>
<td>6232</td>
<td>519</td>
<td>1147</td>
</tr>
</tbody>
</table>

a  Median expression in the most highly expressed tissue per pseudogene.

b  Bonferroni-corrected P-values for all (including non-expressed) and expressed only pseudogenes calculated by Mann-Whitney U test.

We also examined patterns of pseudogene retention across these genomes, and we define a pseudogene as “retained” if it is present in each descendant of the branch where we infer origin. In total, we consider 1,674 (~19.2%) pseudogenes to be retained (Figure 4.1, Table 4.1). Notably, 93 pseudogenes (~9.3%) of the 993 that originated at or before the furthest branching point, the divergence with canines, are retained and present across all 20 species. Additionally, we find that processed pseudogenes are more likely to be retained than duplicated pseudogenes (Bonferroni-corrected P = 1.34 × 10^{-34}, \chi^2-test, Table 4.1). Our pseudogene ortholog identification allows us to further examine their evolutionary trends.
4.3.2 Pseudogene genome-wide expression patterns.

A pseudogene by necessity should be expressed in order to function as a ceRNA. Utilising GTEx RNA-seq data we examined pseudogene expression patterns in more human tissues than previously examined by others (Pei et al., 2012). To determine an adequate expression threshold we plotted the distribution of median TPM values for each pseudogene in the tissue where they are most highly expressed (Figure 4.2A). We selected a threshold value of 0.1 TPM, and we consider a pseudogene expressed if it has a median TPM ≥ 0.1 in at least one tissue. This expression cut-off minimises false positives and false negatives, and GTEx uses a similar cut-off to determine expression. However, our threshold is more conservative as we require a pseudogene to be expressed at ≥ 0.1 TPM in at least 50% of samples in a given tissue (i.e. median ≥ 0.1 TPM). By comparison, GTEx only requires ≥ 0.1 TPM in at least 20% of samples. In total, of our 8,704 pseudogenes with identified orthologs, we define 4,944 (~57%) pseudogenes as expressed in at least one human tissue (Table 4.1), a much higher proportion than previously reported (Pei et al., 2012). However, it remains likely that some of these pseudogenes represent false positives despite this more conservative approach.

We also wished to examine the relationship between pseudogene and parent gene expression. Pseudogenes which ubiquitously function as ceRNAs for their parents should be expressed in all of the same tissues. By contrast, tissue-specific pseudogenes affecting expression of their parents in certain tissues only, exemplified by BRAF and BRAFP1, will be expressed in fewer tissues. We plotted the ratio of the number of tissues with pseudogene expression compared to the number of tissues with parent gene expression (Figure 4.2B). Additionally, we examined the overall distribution of numbers of tissues with pseudogene, parent gene, and non-parent gene expression (Figure 4.2C). These analyses suggest that
Figure 4.2 | Human pseudogene and parent gene expression patterns.  

(A) The density distribution of median expression values between 0 and 10 TPM is shown - with a sub-panel highlighting the region between 0 and 0.5 TPM - for pseudogenes in their most highly expressed tissue. The BRAFP1 maximal median expression level is given by the red bar, for comparison. The blue bar indicates our expression level cut-off of 0.1 TPM. (B) The distribution of the ratio of the number of tissues with pseudogene expression compared to the number of tissues with their respective parent gene expression is plotted for processed and duplicated pseudogenes. To highlight important trends the plot was limited to ratios between 0 and 3; thus 111 pseudogene-parent gene pairs with ratios $> 3$ were excluded from the plot. (C) Numbers of tissues with pseudogene, parent gene and non-parent gene expression are illustrated in the raincloud plot. Sample means are given by the black dots with medians indicated by the box and whisker plots. $P$-value was calculated by Mann-Whitney U test.
the majority of pseudogenes are expressed in many fewer tissues relative to their parent genes and appear to be more tissue specific, consistent with previous reports (Pei et al., 2012; Navarro and Galante, 2015). However, some pseudogenes are expressed in either the same number of tissues as their parents (n = 322, ratio of 1 means the same number of tissues) or in more tissues than their parents (n = 261, ratio > 1 means the pseudogene is expressed in more tissues; Figure 4.2B).

In addition, we find that processed pseudogenes on average have a significantly lower mean ratio than duplicated pseudogenes (Processed = 0.33, Duplicated = 0.87; \( P = 1.35 \times 10^{-8} \), Mann-Whitney U test), suggesting that duplicated pseudogenes are more often expressed in the same number of - or more - tissues as their parents. This is expected, as duplicated pseudogene formation usually involves the copying of the regulatory elements from the parent gene also. Furthermore, we find that protein-coding genes that have pseudogenes (i.e., parent genes) are expressed in more tissues than protein-coding genes with no associated pseudogenes (i.e., non-parent genes) (\( P = 5.93 \times 10^{-60} \), Mann-Whitney U test; Figure 4.2C). This result is consistent with previous reports suggesting many human pseudogenes, particularly processed pseudogenes, originate from housekeeping genes, which are usually more broadly expressed (Zhang et al., 2002; Balasubramanian et al., 2009; Liu et al., 2009; Pei et al., 2012). Our results indicate that many more human pseudogenes than previously estimated may be expressed in at least one tissue type, and are generally more tissue-specific than their parents.

4.3.3 Expressed pseudogenes with 3’UTRs have retained orthologs across mammalian species.

Current pseudogene annotations frequently do not differentiate between sequence regions (Pei et al., 2012; Aken et al., 2017). We identified the pseudo-3’UTR and pseudo-CDS regions by querying the 3’UTRs of the parent genes against their
respective pseudogenes using BLASTN (version 2.7.1, see Materials and Methods for details) Each parent 3’UTR used here was derived from the parent transcript identified by PsiDR as most likely to have given rise to their respective pseudogene(s). Alternative parent gene 3’UTRs were not evaluated as most of them are likely not homologous to their pseudogene 3’UTRs. For duplicated pseudogenes, which also contain pseudo-introns, we queried the parent gene CDS regions against these pseudogenes to delimit their pseudo-CDS regions. We identify 3’UTRs in 2,331 pseudogenes using this method (Table 4.1). We find that identification of 3’UTRs was not biased towards the youngest pseudogenes, which might be expected if these elements are not under constraint. However, frequency of 3’UTR presence, or detection, does vary by branch of origin of its associated pseudogene (Bonferroni-corrected $P = 2.81 \times 10^{-56}$, $\chi^2$-test; Figure 4.3A). Branches 5, 6 and 7 (see Figure 4.1) are enriched for the presence of 3’UTRs, whereas the oldest branches, 11 and 12, are depleted. This may suggest that 3’UTRs of older pseudogenes are more difficult to identify or have been lost.

We also examined the relationship between pseudogene branch of origin and their biotype status and expression in human tissues. We find that, while some branches are enriched for pseudogene expression and some depleted (again using $\geq 0.1$ TPM to indicate expression), there is no discernible pattern beyond the oldest branches being depleted for expressed pseudogenes (Bonferroni-corrected $P = 3.51 \times 10^{-10}$, $\chi^2$-test; Figure 4.3B). Furthermore, while processed pseudogenes ($n = 7,364$) are in general more abundant than duplicated pseudogenes ($n = 1,195$), branches 7, 8, 9, and 10 are significantly enriched for the formation of processed pseudogenes, and duplicated pseudogenes are enriched on every other branch, including the oldest two branches (Bonferroni-corrected $P = 7.93 \times 10^{-277}$, $\chi^2$-test; Figure 4.3C). This is consistent with previous reports suggesting a burst of retrotransposition occurred at the base of several primate lineages (Ohshima
Figure 4.3 | Pseudogene expression, biotype and 3’UTR retention.
(Continued on the next page.)
Figure 4.3 | Pseudogene expression, biotype and 3’UTR retention.

(A) Barplot showing the proportion of pseudogenes with an identifiable 3’UTR compared to those where a 3’UTR is not present (i.e. could not be identified) separated by their branch of origin (age). (B) Barplot showing the proportion of expressed pseudogenes (median TPM ≥ 0.1 in at least one tissue) compared to not expressed pseudogenes by their branch of origin. (C) Barplot showing the proportion of processed vs. duplicated pseudogenes by their branch of origin. (D) Barplot showing the proportion of expressed pseudogenes with an identifiable 3’UTR compared to the proportion where a 3’UTR is not present. (E) Barplot showing the proportion of fully retained pseudogenes (minimum of 5 species) with a 3’UTR compared to those without a 3’UTR. Pseudogenes are considered retained if they are present in every descendant since their branch of origin. (F) Barplot showing the proportion of expressed pseudogenes that are retained and also have a 3’UTR present, compared to those that are retained but do not have a 3’UTR. All P-values were calculated by χ²-test and are Bonferroni-corrected. Red and blue stars indicate enrichment or depletion as determined by deviation of standard residuals.

et al., 2003; Zhang and Gerstein, 2004; Pei et al., 2012; Navarro and Galante, 2015). Contrary to expectation, duplicated pseudogenes are not more likely to be expressed than processed pseudogenes (P = 0.45, χ²-test). However, among expressed pseudogenes, duplicated pseudogenes are more highly expressed than processed pseudogenes (P = 8.97 × 10⁻¹⁹, Mann-Whitney U test). Given that branches 11 and 12 are depleted for presence of pseudogene 3’UTRs but enriched for duplicated pseudogenes, the possibility arises that duplicated pseudogenes are either less likely to retain their 3’UTRs, or it is more difficult to detect them. As expected, we find that processed pseudogenes are more likely to have an identifiable 3’UTR than duplicated pseudogenes (Bonferroni-corrected P = 1.06 × 10⁻¹¹, χ²-test; Table 4.1).

We next pooled pseudogenes regardless of their branch of origin in order to examine their overall trends. We find that pseudogenes possessing a 3’UTR are more highly expressed when considering the total set of pseudogenes (Bonferroni-corrected P = 4.86 × 10⁻³⁵, Mann-Whitney U test; Table 4.1). This trend is not mirrored for the expressed pseudogene only set (Bonferroni-corrected P =
0.18, Mann-Whitney U test). Importantly, however, we find that pseudogenes possessing a 3’UTR are more likely to be expressed in the first place (Bonferroni-corrected $P = 4.89 \times 10^{-28}$, $\chi^2$-test; Figure 4.3D). Additionally, they are more likely to be retained than pseudogenes without a 3’UTR (Bonferroni-corrected $P = 3.30 \times 10^{-5}$, $\chi^2$-test; Figure 4.3E). While in general retained pseudogenes are not more likely to be expressed than those not retained (Bonferroni-corrected $P = 1$, $\chi^2$-test; Table 4.1), pseudogenes that are both retained and have a 3’UTR present are more likely to be expressed than those that are retained but do not have a 3’UTR (Bonferroni-corrected $P = 4.80 \times 10^{-4}$, $\chi^2$-test; Figure 4.3F). Overall these results suggest a significant role for expressed pseudogenes that have retained their 3’UTRs over evolutionary history, consistent with a conserved ceRNA function.

### 4.3.4 Pseudogene 3’UTRs are under stronger evolutionary constraint.

To further elucidate a functional role for pseudogene 3’UTRs we next examined genome-wide sequence constraint on these regions. As mentioned previously, evolutionary constraint is an unbiased arbiter of biological functionality, and tests for these constraints on potential ceRNA candidates should illuminate whether these elements have biological functions at normal physiological levels (Doolittle, 2013; Graur et al., 2013). To examine sequence constraint in a similar manner to *BRAF* and *BRAFp1* we limited our pseudogene set to only those that had CDS regions and 3’UTRs that both pass our filtering criteria (n = 1,629, Figure 4.5A). We also ensured none of these pseudogenes were overlapping functional elements on the opposite strand. Potential signals of selection at such loci that are not associated with these pseudogenes may otherwise confound our analyses. To determine the median number of substitutions for each region of these pseu-
Figure 4.4 | Pseudogene age correlates with median number of substitutions.

Median number of substitutions in the CDS region (A) and in the 3'UTR (B) of each pseudogene by their branch of origin is shown. Plot is jittered for clarity. $\rho$ (rho) and Bonferroni-corrected $P$-values were calculated by Spearman rank-order correlation. Blue lines indicate linear regressions.

For human pseudogenes we performed multiple sequence alignment with MUSCLE (Edgar, 2004) and performed maximum likelihood phylogeny generation for each pseudogene region using the T92 model (Tamura, 1992; see Materials and Methods). We then calculated the median number of substitutions for each phylogeny using the parent genes as out-groups.

We find significant positive correlation between pseudogene age and the number of substitutions in both the pseudogene CDS regions and their 3'UTRs ($\rho = 0.35$, $P = 1.43 \times 10^{-48}$ & $\rho = 0.36$, $P = 2.15 \times 10^{-53}$, respectively, Spearman rank-order correlation tests; Figure 4.4A,B). These results suggest our method of identifying human pseudogene orthologs and their branches of origin is accurate, as the median number of substitutions should increase over time after pseudogene formation and subsequent divergence from their parent genes.

To determine whether a difference in sequence constraint between regions exists for these pseudogenes we compared the median number of substitutions of their CDS regions and 3'UTRs. Examining the total set of pseudogenes we
Figure 4.5 | Constraint of 3'UTRs across pseudogene orthologs. (Continued on the next page)
find that the 3’UTRs have significantly fewer substitutions than their CDS regions (Bonferroni-corrected \( P = 1.89 \times 10^{-24} \), Wilcoxon signed-rank test; Figure 4.5B). These results mirror our findings of increased 3’UTR sequence constraint compared to the CDS region in \textit{BRAFP1}. We also find that this is true for both young \((n = 806, \text{ originated on branch } 7 \text{ or younger, at the base of the simian lineage})\) and old \((n = 823)\) pseudogenes (Bonferroni-corrected \( P = 1.47 \times 10^{-10} \) & \( P = 2.66 \times 10^{-14} \), respectively, Wilcoxon signed-rank tests; Figure 4.5C), suggesting that this trend is not biased by pseudogene age.

By contrast, a similar analysis of parent genes \((n = 274)\) shows that their CDS regions have significantly fewer substitutions relative to their 3’UTRs, which is expected of protein-coding genes (Bonferroni-corrected \( P = 2.7 \times 10^{-41} \), Wilcoxon signed-rank test; Figure 4.6). Notably, when we restrict the pseudogene analysis to the duplicated \((n = 91)\) pseudogenes (as distinct from retropseudogenes), we observe that, unlike the general trend for pseudogenes, the CDS regions have
fewer substitutions ($P = 0.0028$, Wilcoxon signed-rank test; Figure 4.5B). Additionally, the CDS regions of duplicated pseudogenes have fewer substitutions than the corresponding regions of processed pseudogenes (Bonferroni-corrected $P = 4.01 \times 10^{-5}$, Mann-Whitney U test). Young DNA-based duplicates have expression patterns that closely mirror the parent gene (Guschanski et al., 2017).

Even though the genes we analysed are annotated as pseudogenes in human, those that originated by DNA-based duplication may have retained functionality (and thus evolutionary constraint) in some non-human species or for an initial period post-duplication. Such a scenario would consequently reduce the overall median
number of substitutions for their CDS regions.

Furthermore, we find that the median number of substitutions in pseudogene 3’UTRs is significantly lower than their CDS regions for both expressed and non-expressed pseudogenes (Bonferroni-corrected $P = 1.42 \times 10^{-11}$ & $P = 2.94 \times 10^{-14}$, respectively, Wilcoxon signed-rank tests; Figure 4.5D). We also find that the CDS regions and 3’UTRs of expressed pseudogenes have significantly fewer substitutions than their non-expressed counterparts (Bonferroni-corrected $P = 8.47 \times 10^{-8}$ & $P = 5.52 \times 10^{-5}$, respectively, Mann-Whitney U tests). Overall, the 3’UTRs of expressed pseudogenes have the lowest median number of substitutions per site among each of these groups.

In summary, we find that the 3’UTRs of pseudogenes in general are under greater sequence constraint than their CDS regions, though this appears to be true mainly for processed pseudogenes. Expressed pseudogenes are also under greater sequence constraint than not expressed, with the 3’UTRs of expressed pseudogenes exhibiting the greatest amount of constraint. These data are consistent with a ceRNA function for expressed pseudogenes.

4.3.5 Sequence constraint on expression-correlated pseudogenes.

Recent studies have shown how known pseudogene ceRNAs can show significant expression correlation with their parent genes, exemplified by the PTENP1 pseudogene (Poliseno et al., 2010) ($\rho = 0.36$, $P$-value $= 0$, Spearman rank-order correlation test; Figure 4.7). However, it remains unclear whether expression correlation is a prerequisite for ceRNA functionality and interactivity between pseudogene and parent. Using GTEx RNA-seq data we examined expression correlation of pseudogene-parent gene pairs in tissues where both are expressed by performing Spearman rank-order correlation tests on each pair. We find that
Figure 4.7 | Expression correlation of $PTEN$ and $PTENP1$.
The $PTENP1$ pseudogene and its parent gene $PTEN$ are plotted. Expression data in TPM was derived from GTEx RNA-seq data (GTEx Consortium, 2015). A linear regression line is given in blue. Number of samples included is indicated and the $P$-value and $\rho$ (rho) were calculated by Spearman rank-order correlation test.

duplicated pseudogenes were more likely to show correlated expression with their parents than processed pseudogenes (Bonferroni-corrected $P = 1.10 \times 10^{-12}$, $\chi^2$-test). This is expected given duplicated pseudogenes usually also possess copies of the regulatory elements associated with their parent genes, and are more likely to be found in close proximity to their parents (Guschanski et al., 2017). Additionally, pseudogenes with parent-correlated expression were not more likely to possess 3’UTRs (Bonferroni-corrected $P = 0.3$, $\chi^2$-test) and were less likely to be retained (Bonferroni-corrected $P = 0.0003$, $\chi^2$-test).

When we examined expressed pseudogenes only (n = 1,061) we find that those with correlated and non-correlated expression patterns have significantly
Figure 4.8 | Region substitution comparison of pseudogene and parent gene expression correlations.

(A) Median number of substitutions for expressed pseudogenes that have correlation or no correlation with expression of their parent genes. Expression correlation was determined by Spearmann rank-order correlation test. For each pseudogene only tissues with both pseudogene and parent gene expression were analysed. Correlation P-values were Bonferroni-corrected for multiple testing. (B) Median number of substitutions for pseudogenes with positive or negative expression correlation with their parent genes. Bonferroni-corrected P-values for comparisons between different regions are calculated by Wilcoxon signed-rank test (●), and between same regions by Mann-Whitney U test (○).

fewer substitutions in their 3'UTRs (Bonferroni-corrected $P = 4.89 \times 10^{-5}$ & $P = 3.08 \times 10^{-8}$, respectively, Wilcoxon signed-rank tests; Figure 4.8A). The CDS regions and 3'UTRs have fewer substitutions in expression-correlated pseudogenes than in their non-correlated counterparts (Bonferroni-corrected $P = 5.65 \times 10^{-7}$ & $P = 9.68 \times 10^{-5}$, respectively, Mann-Whitney U tests). Separating the correlated pseudogenes into positively or negatively correlated we find that both of these groups showed fewer 3'UTR substitutions (Bonferroni-corrected $P = 0.014$ & $P = 0.0012$, respectively, Wilcoxon signed-rank tests; Figure 4.8B) and the CDS regions of positively correlated pseudogenes have fewer substitutions than that of negatively correlated pseudogenes (Bonferroni-corrected $P =$
However, we find that the 3’UTRs were not significantly different between positively and negatively correlated pseudogenes (Bonferroni-corrected $P = 1$, Mann-Whitney U test), though the median number of substitutions for pseudogenes that have positively correlated expression with their parents is lower than all other groups tested.

Overall, these results suggest that pseudogenes that have correlated expression with their parent genes are under greater sequence constraint. Furthermore, positively correlated pseudogenes appear to also be under greater constraint than negatively correlated pseudogenes, though not significantly for the 3’UTRs.

**4.3.6 Tumour suppressor genes are not enriched for pseudogene paralogs.**

It has recently been shown how transcripts acting as ceRNAs for tumour suppressor genes are selectively targeted for 3’UTR shortening via alternative polyadenylation in breast cancer tumours (Park et al., 2018). Shortening of these transcripts restricts their ability to function as ceRNAs by removing their MREs and reduces tumour suppressor gene activity in trans. These observations indicate that many tumour suppressor genes may have ceRNAs which participate in post-transcriptional regulation of their activity. However, 3’UTR shortening of pseudogene transcripts was not examined. Given these findings we hypothesise that tumour suppressor genes, as opposed to oncogenes, more frequently have associated pseudogene paralogs that function as ceRNAs.

We identified a list of protein-coding genes that are tumour suppressive (TSGs, $n = 247$), oncogenic ($n = 243$), or both ($n = 72$), from the manually curated Cancer Gene Census (Futreal et al., 2004) and from a list compiled by Vogelstein et al., 2013. We find that cancer genes (including TSGs, oncogenes, and both) are enriched for having at least 1 pseudogene compared to non-cancer genes.
Table 4.2 | Numbers of pseudogenes with tumour suppressive or oncogenic parents.

<table>
<thead>
<tr>
<th></th>
<th>Total set</th>
<th>Retained</th>
<th>Not retained</th>
<th>3'UTR Present</th>
<th>3'UTR Absent</th>
<th>Expressed</th>
<th>Not Expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSG pseudogenes</td>
<td>174</td>
<td>30</td>
<td>144</td>
<td>45</td>
<td>129</td>
<td>100</td>
<td>74</td>
</tr>
<tr>
<td>Oncogene pseudogenes</td>
<td>164</td>
<td>39</td>
<td>125</td>
<td>67</td>
<td>97</td>
<td>91</td>
<td>73</td>
</tr>
<tr>
<td>Both pseudogenes</td>
<td>47</td>
<td>7</td>
<td>40</td>
<td>25</td>
<td>22</td>
<td>29</td>
<td>18</td>
</tr>
<tr>
<td>Total #</td>
<td>562</td>
<td>76</td>
<td>309</td>
<td>137</td>
<td>248</td>
<td>220</td>
<td>165</td>
</tr>
</tbody>
</table>

a Tumour suppressor gene
b Genes annotated as both tumour suppressive and oncogenic

(Bonferroni-corrected $P = 1.09 \times 10^{-5}$, $\chi^2$-test; Figure 4.9A). However, we find that, among cancer genes, neither TSGs nor oncogenes were enriched for having pseudogenes ($P = 0.74$, $\chi^2$-test; Figure 4.9B). Additionally, TSGs do not have more pseudogenes on average than oncogenes ($P = 0.81$, Mann-Whitney U test). Despite this result it is possible that pseudogenes with TSG parents (TSG pseudogenes) are more likely to function as ceRNAs by being more frequently expressed, retained, or by possessing a 3’UTR, as compared to pseudogenes with oncogenic parents (oncogene pseudogenes). Examining the properties of pseudogenes with cancer gene parents we find no association between cancer gene type and pseudogene expression or retention (Bonferroni-corrected $P = 0.79$ & $P = 0.18$, respectively, $\chi^2$-tests; Table 4.2). Among expressed pseudogenes, TSG pseudogenes are not expressed at higher levels than oncogene pseudogenes ($P = 0.1$, Mann-Whitney U test).

Contrary to our expectations, we find that TSG pseudogenes are less likely to possess a 3’UTR than oncogene pseudogenes (Bonferroni-corrected $P = 0.0022$, $\chi^2$-test; Table 4.2, Figure 4.9C). Interestingly, pseudogenes with parents annotated as both tumour suppressive and oncogenic are most strongly enriched for possessing a 3’UTR. While cancer genes are enriched for pseudogenes as a whole our results do not support the hypothesis that TSGs are more likely to have
4.3.7 Recent positive selection at pseudogene loci.

We have addressed whether pseudogenes exhibit sequence constraint over long evolutionary time-scales by examining differences in numbers of substitutions between sequence regions. However, we may also observe recent positive selection acting on pseudogenes with ceRNA potential in certain lineages. Several pseudogene ceRNAs than oncogenes. Given their enrichment for 3’UTRs, further investigation of pseudogenes with parents that are both oncogenic and tumour suppressive may reveal whether pseudogenes contribute to these dichotomous effects.
Human population genomic datasets now exist, including The 1000 Genomes Project (The International HapMap Consortium, 2007) and HapMap3 (Auton et al., 2015), which can be interrogated to examine recent selection acting on human pseudogenes. We tested pseudogene loci for signatures of recent positive selection - within human populations - using pre-computed statistical measures from dbPSHP on SNP data from the 1000 Genomes Project (Li et al., 2014b; Auton et al., 2015). We extracted data on Tajima’s D, fixation index ($F_{ST}$), derived allele frequency difference ($\Delta$DAF), and the cross-population composite likelihood ratio (XP-CLR) from this study.

We find 4,741 (54.5%) of 8,693 pseudogene loci contain at least one SNP in at least one population that passes one or more score threshold (Li et al., 2014b). More conservatively, 503 pseudogenes pass positive selection thresholds in at least three of the tests, and 22 pseudogenes pass all four, the latter being the most confident inferences (Tables 4.3 & 4.5). We find that 11 (50%) of the top 22 potentially positively selected pseudogenes possess an identifiable 3’UTR, compared to only ~26.8% of the total set of pseudogenes ($P = 0.026$, Fisher’s exact test; Table 4.4). Additionally, 6 (~27.3%) of these top pseudogenes are retained across species, compared with ~20.2% for the total set ($P = 0.42$, Fisher’s exact test). These observations highlight how analysis of positive selection using large-scale genomic datasets could identify particular pseudogenes with potentially novel ceRNA functions that warrant further study, such as the top 22 pseudogenes illustrated here.
Table 4.3 | Positive selection scores across 22 pseudogene loci.

<table>
<thead>
<tr>
<th>Pseudogene ID</th>
<th>SNP ID</th>
<th>Population</th>
<th>Tajima’s D</th>
<th>ΔDAF</th>
<th>XP-CLR</th>
<th>F_ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000182965</td>
<td>rs6977118</td>
<td>CHB</td>
<td>-0.5202</td>
<td>0.6061</td>
<td>55.5315</td>
<td>0.0842</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-1.291</td>
<td>0.6261</td>
<td>64.4372</td>
<td>0.0915</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JPT</td>
<td>-1.6576</td>
<td>0.6733</td>
<td>59.7076</td>
<td>0.0868</td>
</tr>
<tr>
<td>ENSG00000178631</td>
<td>rs2165139</td>
<td>CHB</td>
<td>-0.9312</td>
<td>0.6093</td>
<td>9.6806</td>
<td>0.0923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-0.914</td>
<td>0.6133</td>
<td>6.7663</td>
<td>0.0968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JPT</td>
<td>-1.321</td>
<td>0.6336</td>
<td>6.7028</td>
<td>0.0871</td>
</tr>
<tr>
<td>ENSG00000237417</td>
<td>rs10786077</td>
<td>CHB</td>
<td>-0.7394</td>
<td>0.569</td>
<td>42.1715</td>
<td>0.0952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-1.092</td>
<td>0.6068</td>
<td>47.0514</td>
<td>0.1074</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JPT</td>
<td>-0.7699</td>
<td>0.2204</td>
<td>26.0036</td>
<td>0.0908</td>
</tr>
<tr>
<td>ENSG00000249531</td>
<td>rs939207</td>
<td>CHB</td>
<td>-0.6319</td>
<td>0.2344</td>
<td>39.2295</td>
<td>0.1063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-0.314</td>
<td>0.2814</td>
<td>12.4747</td>
<td>0.0737</td>
</tr>
<tr>
<td>ENSG0000024520</td>
<td>rs16865471</td>
<td>CHB</td>
<td>-1.6192</td>
<td>0.2696</td>
<td>17.7816</td>
<td>0.0679</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-0.314</td>
<td>0.2916</td>
<td>7.6983</td>
<td>0.0756</td>
</tr>
<tr>
<td>ENSG00000237842</td>
<td>rs3795741</td>
<td>CHB</td>
<td>-1.6192</td>
<td>0.2692</td>
<td>8.4771</td>
<td>0.0668</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHS</td>
<td>-0.314</td>
<td>0.2916</td>
<td>7.6983</td>
<td>0.0756</td>
</tr>
<tr>
<td>ENSG00000233333</td>
<td>rs1625234</td>
<td>CHB</td>
<td>-0.1337</td>
<td>0.532</td>
<td>45.1222</td>
<td>0.0739</td>
</tr>
<tr>
<td>ENSG00000230201</td>
<td>rs4609912</td>
<td>CHS</td>
<td>-0.5913</td>
<td>0.408</td>
<td>22.6532</td>
<td>0.1424</td>
</tr>
<tr>
<td>ENSG00000237672</td>
<td>rs4769937</td>
<td>MXL</td>
<td>-1.0232</td>
<td>0.2179</td>
<td>17.6207</td>
<td>0.0986</td>
</tr>
<tr>
<td>ENSG00000237951</td>
<td>rs1025543</td>
<td>JPT</td>
<td>-0.2617</td>
<td>0.6271</td>
<td>17.3324</td>
<td>0.094</td>
</tr>
<tr>
<td>ENSG00000251478</td>
<td>rs1318743</td>
<td>JPT</td>
<td>-1.4969</td>
<td>0.3429</td>
<td>14.5028</td>
<td>0.2065</td>
</tr>
<tr>
<td>ENSG00000234568</td>
<td>rs13119500</td>
<td>YRI</td>
<td>-0.0564</td>
<td>0.2116</td>
<td>12.5887</td>
<td>0.0741</td>
</tr>
<tr>
<td>ENSG00000218713</td>
<td>rs4487576</td>
<td>MXL</td>
<td>-0.125</td>
<td>0.2784</td>
<td>11.9151</td>
<td>0.1992</td>
</tr>
<tr>
<td>ENSG00000236296</td>
<td>rs4835206</td>
<td>JPT</td>
<td>-0.0468</td>
<td>0.3422</td>
<td>9.3285</td>
<td>0.0666</td>
</tr>
<tr>
<td>ENSG00000229001</td>
<td>rs10823277</td>
<td>CHB</td>
<td>-0.5354</td>
<td>0.363</td>
<td>8.8884</td>
<td>0.0728</td>
</tr>
<tr>
<td>ENSG00000248162</td>
<td>rs1881523</td>
<td>LWK</td>
<td>-0.3485</td>
<td>0.237</td>
<td>8.0523</td>
<td>0.0722</td>
</tr>
<tr>
<td>ENSG00000216518</td>
<td>rs9398187</td>
<td>CHB</td>
<td>-1.0563</td>
<td>0.2157</td>
<td>7.5387</td>
<td>0.0606</td>
</tr>
<tr>
<td>ENSG00000236253</td>
<td>rs1766642</td>
<td>LWK</td>
<td>-0.286</td>
<td>0.3197</td>
<td>7.3286</td>
<td>0.2283</td>
</tr>
<tr>
<td>ENSG00000226549</td>
<td>rs9889729</td>
<td>YRI</td>
<td>-0.9508</td>
<td>0.2582</td>
<td>7.0046</td>
<td>0.2474</td>
</tr>
<tr>
<td>ENSG00000251557</td>
<td>rs7935232</td>
<td>FIN</td>
<td>-0.3026</td>
<td>0.2052</td>
<td>5.4861</td>
<td>0.0505</td>
</tr>
<tr>
<td>ENSG00000234670</td>
<td>rs12230513</td>
<td>CHS</td>
<td>-0.6744</td>
<td>0.2705</td>
<td>5.2774</td>
<td>0.0829</td>
</tr>
<tr>
<td>ENSG00000237164</td>
<td>rs9513048</td>
<td>CEU</td>
<td>-0.9542</td>
<td>0.2036</td>
<td>5.1345</td>
<td>0.0711</td>
</tr>
</tbody>
</table>
Table 4.4 | Top 22 potentially positively selected pseudogenes.

<table>
<thead>
<tr>
<th>Pseudogene ID</th>
<th>Biotype</th>
<th>3'UTR</th>
<th>Retained</th>
<th>Expressed</th>
<th>Parent Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000182965</td>
<td>processed</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>NPM1</td>
</tr>
<tr>
<td>ENSG00000178631</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>ACTG1</td>
</tr>
<tr>
<td>ENSG00000237417</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>XRCC6</td>
</tr>
<tr>
<td>ENSG00000249531</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>UGT2B4</td>
</tr>
<tr>
<td>ENSG00000224520</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>KRT8</td>
</tr>
<tr>
<td>ENSG00000237842</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>SMU1</td>
</tr>
<tr>
<td>ENSG00000233333</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>RNFT1</td>
</tr>
<tr>
<td>ENSG00000230201</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>ATP6V0C</td>
</tr>
<tr>
<td>ENSG00000237672</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>KRR1</td>
</tr>
<tr>
<td>ENSG00000237951</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>PPIL1</td>
</tr>
<tr>
<td>ENSG00000251478</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>TCP1</td>
</tr>
<tr>
<td>ENSG00000234568</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>BIN2</td>
</tr>
<tr>
<td>ENSG00000218713</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>RCC1</td>
</tr>
<tr>
<td>ENSG00000236296</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>GUSB</td>
</tr>
<tr>
<td>ENSG00000229001</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>ACTB</td>
</tr>
<tr>
<td>ENSG00000248162</td>
<td>processed</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>CHORDC1</td>
</tr>
<tr>
<td>ENSG00000216518</td>
<td>processed</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>EEF1G</td>
</tr>
<tr>
<td>ENSG00000236253</td>
<td>processed</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>SLC25A3</td>
</tr>
<tr>
<td>ENSG00000226549</td>
<td>processed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>SCD</td>
</tr>
<tr>
<td>ENSG00000251557</td>
<td>processed</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>HNRNPK</td>
</tr>
<tr>
<td>ENSG00000234670</td>
<td>unprocessed</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>OR6C2</td>
</tr>
<tr>
<td>ENSG00000237164</td>
<td>unprocessed</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>SLC25A15</td>
</tr>
</tbody>
</table>
Table 4.5 | Gene names and descriptions of parent genes from top 22 positively selected pseudogenes.

<table>
<thead>
<tr>
<th>Parent Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPM1</td>
<td>nucleophosmin 1</td>
</tr>
<tr>
<td>ACTG1</td>
<td>actin gamma 1</td>
</tr>
<tr>
<td>XRCC6</td>
<td>X-ray repair cross complementing 6</td>
</tr>
<tr>
<td>UGT2B4</td>
<td>UDP glucuronosyltransferase family 2 member B4</td>
</tr>
<tr>
<td>KRT8</td>
<td>keratin 8</td>
</tr>
<tr>
<td>SMU1</td>
<td>SMU1, DNA replication regulator and spliceosomal factor</td>
</tr>
<tr>
<td>RNFT1</td>
<td>ring finger protein, transmembrane 1</td>
</tr>
<tr>
<td>ATP6V0C</td>
<td>ATPase H+ transporting V0 subunit c</td>
</tr>
<tr>
<td>KRR1</td>
<td>KRR1, small subunit processome component homolog</td>
</tr>
<tr>
<td>PPIL1</td>
<td>peptidylprolyl isomerase like 1</td>
</tr>
<tr>
<td>TCP1</td>
<td>t-complex 1</td>
</tr>
<tr>
<td>BIN2</td>
<td>bridging integrator 2</td>
</tr>
<tr>
<td>RCC1</td>
<td>regulator of chromosome condensation 1</td>
</tr>
<tr>
<td>GUSB</td>
<td>glucuronidase beta</td>
</tr>
<tr>
<td>ACTB</td>
<td>actin beta</td>
</tr>
<tr>
<td>CHORDC1</td>
<td>cysteine and histidine rich domain containing 1</td>
</tr>
<tr>
<td>EEF1G</td>
<td>eukaryotic translation elongation factor 1 gamma</td>
</tr>
<tr>
<td>SLC25A3</td>
<td>solute carrier family 25 member 3</td>
</tr>
<tr>
<td>SCD</td>
<td>stearoyl-CoA desaturase</td>
</tr>
<tr>
<td>HNRNPK</td>
<td>heterogeneous nuclear ribonucleoprotein K</td>
</tr>
<tr>
<td>OR6C2</td>
<td>olfactory receptor family 6 subfamily C member 2</td>
</tr>
<tr>
<td>SLC25A15</td>
<td>solute carrier family 25 member 15</td>
</tr>
</tbody>
</table>
4.4 Discussion

The emergence of the ceRNA hypothesis within the past decade has led to the identification of many lncRNAs, including expressed pseudogenes, that are capable of indirectly regulating protein-coding mRNAs (Poliseno et al., 2010; Marques et al., 2012; Wang et al., 2013a,b; Rutnam et al., 2014; Karreth et al., 2015; Zheng et al., 2015; Thomson and Dinger, 2016; Zheng et al., 2016a; Straniero et al., 2017; Zhu et al., 2017). Aberrant expression of these ceRNAs has been shown to contribute to the development of different diseases, with perhaps the most notable cases being ceRNAs involved in the pathogenesis of various forms of cancer, including the well known PTENP1 and BRAFP1 pseudogenes (Poliseno et al., 2010; Karreth et al., 2011, 2015; Poliseno and Pandolfi, 2015). By contrast, the linc-mg and linc-ROR lncRNAs are examples of ceRNAs with proven functions in normal cellular processes. However, despite the discovery of these important miRNA sponges, little is known about the evolution of ceRNAs as a whole and whether or not they are maintained over evolutionary time-scales.

While many assume pseudogenes to be non-functional relics of gene duplication and retrotransposition, we examined these elements as potential sources of ceRNAs using evolutionary metrics. Modeling of ceRNA activity suggests that the greater the number of miRNAs mediating competition between two transcripts, the more likely it is that expression changes experienced by one transcript will be physiologically relevant to the other (Chiu et al., 2018). Given these observations it becomes clear that pseudogenes in particular make excellent ceRNA candidates, as when formed they automatically have identical MRE profiles as their parent genes. Therefore, provided they are expressed these pseudogenes should make effective competitors for the same miRNAs. In fact, we find that a larger proportion of human pseudogenes than previously estimated (~56.8%)
appear to be expressed in at least one tissue, allowing for the possibility of ceRNA functionality.

Evolutionary constraint is an unbiased indicator of biological functionality (Doolittle, 2013; Graur et al., 2013). In order to examine pseudogene constraint over evolutionary time-scales we identified orthologs for 8,704 human pseudogenes across 20 mammalian genomes. While most of these pseudogenes do not have orthologs dating to the base of these lineages, 993 (~11.4%) are as old or older than the divergence between Canis lupis and the rest of the 19 mammals tested. 1,674 (~19.2%) pseudogenes are retained in each descendant since their formation. Intriguingly, 93 (~9.3% of the 993 originating before the Canis lupis divergence) of these are present/retained in all 20 species, suggesting that these pseudogenes may have important functions that have led to their maintenance for at least 96 million years. Additionally, 2,331 (~26.8%) pseudogenes have identifiable 3'UTRs, and these are enriched for expressed and retained pseudogenes, a finding consistent with a ceRNA role for pseudogenes.

Previous studies have claimed that any bona fide ceRNA activity is likely to be restricted to rare cases only, and dispute that this functionality could be widespread in normal tissues at physiological levels (Chang et al., 2004; Denzler et al., 2014, 2016). In general, 3'UTRs of protein-coding genes are under sequence constraint in mammals, though at a much lower level than their CDS regions (Davydov et al., 2010). However, as seen for BRAFP1 in Chapter 3, increased sequence constraint of the 3'UTR relative to the CDS of a pseudogene can be used as a signature of constraint for identifying other potential pseudogene ceRNAs. Using sequence alignment and phylogeny generation we showed that for parent protein-coding genes the CDS region is under much greater constraint. Conversely, we see the opposite trend for pseudogenes, where our observations of increased constraint on the 3'UTR of BRAFP1 is mirrored genome-wide. Se-
sequence similarity between pseudogene and parent gene CDS regions and 3’UTRs was examined previously, though not across multiple species (Pei et al., 2012). Greater constraint of pseudogene 3’UTRs in this manner is suggestive of conservation of miRNA binding sites. Considering many currently validated ceRNA interactions occur between pseudogenes and their respective parent genes our findings indicate a widespread ceRNA role for pseudogenes (Poliseno et al., 2010; Wang et al., 2013a; Rutnam et al., 2014; Karreth et al., 2015; Zheng et al., 2015, 2016a; Straniero et al., 2017).

The importance of 3’UTRs to ceRNA interactivity and normal cellular processes was recently demonstrated. In breast cancer tumours it was revealed that transcripts predicted to function as ceRNAs for tumour suppressor genes are enriched for 3’UTR shortening through aberrant alternative polyadenylation (Xia et al., 2014b; Park et al., 2018). In this manner tumour suppressor genes are repressed in trans, contributing to tumour development. Notably, nine genes involved in a ceRNA network with the PTEN tumour suppressor gene were predicted to be trans-targets of 3’UTR shortening (Park et al., 2018). While not investigated in this study, we predict that PTENP1 is also a target of this 3’UTR shortening in tumour cells, as selective loss of the PTENP1 locus has been exhibited in human cancers (Poliseno et al., 2010). These findings further hint at a ceRNA role for expressed pseudogenes that have retained their 3’UTRs, and highlights the potential consequences of dysregulated expression. However, we find no evidence to suggest tumour suppressor genes more frequently have pseudogenes acting as ceRNAs, indicating that pseudogenes of both oncogenes (e.g. BRAF) and tumour suppressor genes (e.g. PTEN) may exhibit this function.

In general, pseudogene 3’UTRs appear to be under greater constraint relative to their CDS regions. However, separating these pseudogenes by biotype, focusing on either processed or duplicated pseudogenes, reveals opposing trends.
Processed pseudogenes exhibit greater constraint in their 3'UTRs, whereas duplicated pseudogenes have more constrained CDS regions. As mentioned, one possibility is that pseudogenes which have been annotated as duplicated pseudogenes in human are in fact still protein-coding orthologs in other mammalian species, resulting in a stronger signature of constraint in the CDS regions of these elements. Given their mode of formation this is a more likely scenario to occur for duplicated rather than processed pseudogenes. In contrast to processed pseudogenes - which arise via retrotransposition of an RNA intermediary - duplicated pseudogenes are a direct copy of genomic DNA. Thus, the regulatory elements and introns present in the parent gene should be duplicated also, and expression patterns between the two elements are more likely to be correlated (Guschanski et al., 2017).

Processed pseudogenes are more likely to be retained and possess a 3'UTR than duplicated pseudogenes, which suggests that expressed processed pseudogenes may more frequently have a ceRNA role. This observation is interesting, as one might assume that duplicated pseudogenes are more likely to function as ceRNAs given their probable sharing or copying of the same regulatory elements as their parents. By contrast, processed pseudogenes are often considered dead-on-arrival, and by necessity would have to either co-opt existing neighbouring regulatory elements or make use of random *de novo* elements. It is therefore surprising that we find signatures of constraint consistent with ceRNA function for processed pseudogenes specifically. However, processed pseudogenes may be more easily maintained as ceRNAs given they are shorter and do not possess introns. Duplicated pseudogene ceRNAs would have to avoid mutations in their intron splice sites, lest they become large, unwieldy transcripts that may be detrimental to miRNA regulation due to unintentional off-target binding effects. Additionally, a recent study has demonstrated that *de novo* prokaryotic promoters are
easily evolved from random sequences after successive generations (Yona et al., 2018), suggesting this could also be the case for some eukaryotic promoters. Processed pseudogenes may have acquired expression in this fashion, and beneficial expressed pseudogenes and their new promoters may subsequently have become fixed in mammalian lineages. The majority of currently known pseudogene ceRNAs are processed; however, *GBAP1* is an example of a duplicated pseudogene with proven ceRNA interaction with its parent gene, *GBA* (Straniero et al., 2017).

While it is not yet clear to what extent correlation of pseudogene and parent gene expression levels is an indication of ceRNA activity, such expression correlation may help identify pseudogene ceRNAs. On the other hand, it is also possible that such correlation need not be evident for miRNA competition to exist. There are several factors to consider, such as which mode of miRNA-mediated repression these transcripts are undergoing, either translational repression and/or transcriptional degradation. Either the pseudogene or parent gene may be degraded more rapidly than than the other, which may lead to apparent expression level disparity. Furthermore, the parent gene or pseudogene may be involved in competition between other ceRNAs that the other is not involved in using non-shared MREs. Pseudogenes could also act as ceRNAs for multiple parent genes from the same gene family, if their 3'UTRs are broadly conserved. While we do find increased sequence constraint in pseudogenes with positive parent-correlated expression levels, such correlation alone should not be be relied upon as predictive of ceRNA activity.

Our observation of increased constraint on the 3'UTR of *BRAFP1* is mirrored genome-wide, and we find that pseudogenes with 3'UTRs are more likely to be expressed and retained. These observations are suggestive of a widespread ceRNA role for expressed pseudogenes, often regarded as non-functional. While we focus here on pseudogenes as potential sources of ceRNAs it remains possible that some
expression or evolutionary trends we observe could also be attributed to pseudogenes with other functional roles, including as sources of siRNAs or piRNAs (Muro et al., 2010; Siomi et al., 2011; Chan et al., 2013). Also, our expression analyses were limited to human GTEx data, and investigation of pseudogene expression in other lineages may exhibit stronger or weaker correlations between expression and 3'UTR retention. Interestingly, we find tentative evidence for recent positive selection of some pseudogene loci in human, with a significant proportion of these also possessing 3'UTRs. However, more research is necessary to elucidate the potentially complex ceRNA crosstalk between these pseudogenes and their parents. Detailed molecular experiments involving expression knockdown and/or artificial mutation of predicted conserved MREs mediating cross-talk would be of great benefit to understanding this new RNA language.
Chapter 5

Testing gene duplication as a mechanism to restore dosage of X-linked genes.

5.1 Introduction

Maintenance of gene dosage balance and dosage compensation are important factors shaping genome content and organisation (Ohno, 1967; McLysaght, 2008; Potrzebowski et al., 2010; Zhang et al., 2010a; Hurst et al., 2015; Graves, 2016). Gene duplication is a common occurrence throughout evolution, and is one of the main processes by which new material is generated for later genetic innovation (Stephens, 1951; Nei, 1969; Ohno, 1970; Conrad and Antonarakis, 2007; Conant and Wolfe, 2008). Many genes are able to tolerate some changes in dosage; however, for a subset of genes there are phenotypic effects associated with copy number variation (Rice and McLysaght, 2017a). For instance, at the individual level, many human diseases, such as cancers, heart disease, and neuropsychiatric disorders, are linked with changes in copy number - be they duplication or deletion events (Gardiner, 2004; Stefansson et al., 2008; Craddock et al., 2010; Cooper et al., 2011; Glessner et al., 2014). Haploinsufficiency, promiscuous off-target interactions, and upset stoichiometry of metabolic pathways or protein complex
members, are all possible deleterious dosage changes (Fisher and Scambler, 1994; Vavouri et al., 2009; Karayiorgou et al., 2010; Cardarelli et al., 2011; Chen et al., 2012; Rice and McLysaght, 2017a). Furthermore, it was recently shown that human pathogenic copy number variant regions are enriched for genes with greater evolutionary copy number conservation across mammals (Rice and McLysaght, 2017b). Thus, it has become clear that for many genes dosage balance is an important evolutionary constraint.

An interesting case of expression evolution and resolution of dosage imbalance, known as dosage compensation, is seen with the evolution of sex chromosomes (Ohno, 1967). In therian mammals, the X and Y chromosomes originally evolved from a pair of autosomes, which began to diverge around 150 to 180 mya, after the therians diverged from monotremes (Charlesworth, 1996; Graves, 1996; Rice, 1996; Lahn and Page, 1999; Potrzebowski et al., 2008; Veyrunes et al., 2008). The current configuration of therian sex chromosomes arose principally through a number of inversion events on the Y chromosome, proceeding from the distal long arm to the distal short arm, which resulted in suppression of X-Y recombination. These events probably marked the beginning of nucleotide sequence divergence and gradual degeneration of the Y chromosome (Lahn and Page, 1999; Lemaitre et al., 2010). This process eventually resulted in the loss of most genes on the Y (Charlesworth and Charlesworth, 2000). These inversions occurred at least four times throughout therian evolution, forming distinct evolutionary strata which can be seen along the X, with the most recent stratum dating to less than 50 mya (Lahn and Page, 1999; Pandey et al., 2013). Currently, only the small pseudoautosomal regions (PARs) at the extremities of the X and Y chromosomes are capable of recombining during male meiosis, and high sequence identity between these regions is maintained (Mangs and Morris, 2007).

With the formation of the sex chromosomes and degeneration of Y-linked
Duplication as mechanism to restore dosage of X-linked genes

genes came the problem of gene expression level imbalance between the X chromosome and autosomes in males. As they have only one X chromosome, male gene expression on the X for genes whose gametolog was lost from the Y would be two-fold reduced (0.5 X/autosome expression ratio) compared to autosomal genes. Ohno suggested that a global, chromosome-wide upregulation of X-linked genes would restore dosage balance between X-linked and autosomal genes (Ohno, 1967). However, while this upregulation would resolve this dosage imbalance in males, it would also extend the problem to females, in that female X-linked gene expression levels would be two-fold increased relative to autosomal genes. It was thus proposed that global upregulation of X-linked genes occurred first during sex chromosome evolution, to restore X:autosome(A) dosage balance in males, with subsequent evolution of X chromosome inactivation (XCI) in order to restore female X-linked gene expression to correct levels (Ohno, 1967).

Whereas XCI (also known as Lyonisation) has been well documented (Lyon, 1961; Deng et al., 2014; Galupa and Heard, 2015; Graves, 2016), global upregulation of X-linked genes has not been found. Based on RNA-seq data, one study claimed that median expression of X-linked genes was half that of autosomal genes (X/A ratio of ~0.5), i.e. what would be expected in the case of no expression upregulation (Xiong et al., 2010; Casci, 2011). However, the X chromosome contains an abundance of testis-specific genes that are not expressed in adult somatic tissues. After removal of non-expressed genes X:A median expression ratios significantly increased to ~0.7 in the majority of tissues examined (Pessia et al., 2012). This was dependent on level of expression, with more highly expressed genes coming closer to an overall 1:1 X:A ratio. These observations suggest that there is a mix of upregulated (X/A ratio of 1, i.e. dosage compensated) and non-upregulated (X/A ratio of 0.5, i.e. no compensation) genes on the X. Thus, it is likely that global upregulation of X-linked genes did not occur after sex chromo-
some evolution in therians, suggesting localised mechanisms may resolve dosage imbalances arising from these events, possibly on a gene-by-gene basis.

Dosage-sensitive genes are expected to be affected most during sex chromosome evolution (Mank et al., 2011). There are several possible mechanisms to restore correct X:A ratios for dosage-sensitive genes on the X: (1) become upregulated to match expression from a diploid locus; (2) relocate to an autosome; (3) retain a functional Y gametolog; and (4) become duplicated on the X. There is evidence to support the first three mechanisms. Genes encoding protein-complex members have previously been shown to exhibit dosage-sensitive properties due to constraints on the stoichiometry of their members (Papp et al., 2003; Blanc, 2004; Seoighe and Gehring, 2004; Barker et al., 2008). Accordingly, it was shown that genes associated with large protein complexes (≥ 7 complex members) are upregulated to roughly ancestral levels to match expression of their autosomal counterparts (Pessia et al., 2012). It was also shown that there was an increased rate of genes relocating from the X chromosome to the autosomes via retrotransposition, suggesting the X is a non-optimal environment for these genes (Emerson et al., 2004; Wang, 2004; Khil et al., 2005; Gurbich and Bachtrog, 2008; Potrzebowski et al., 2008). Furthermore, highly dosage-sensitive genes - for which haploinsufficiency would be deleterious - are enriched for retention of a functional homolog on the Y chromosome (Bellott et al., 2014).

In general, evidence for upregulation of X-linked genes to ancestral levels remains limited. While upregulation may restore expression levels to adequate levels for some dosage-sensitive X-linked genes, this solution may not be feasible for others (Vicoso and Charlesworth, 2009; Pessia et al., 2012; Hurst et al., 2015). Here we investigate the fourth possible mechanism for restoring dosage balance: duplication on the X. In particular, we expect that for certain genes upregulation will not be available if the locus was already at maximal expression, i.e. ances-
trally highly expressed genes. For instance, in cases where the transcriptional rate is high it would presumably be optimal for genes to exist in a diploid state such that transcription can occur from two loci simultaneously, allowing for greater net transcriptional output. The traffic jam, or “weak X”, model posits that for genes transcribed from a single locus - which is the case for haploid-expressed genes on the X - high rates of transcription may lead to transcriptional traffic jams (Hurst et al., 2015). This model expands on the notion that there is a physical upper limit to the rate of transcription that is possible from a single locus (Vicoso and Charlesworth, 2009). Indeed, there is evidence to suggest that X-linked genes have maximal expression levels three times lower on average than autosomal genes, and that highly expressed X-linked genes may have relocated to an autosome (Hurst et al., 2015).

Alternatively, it is plausible that highly expressed X-linked genes have duplicated in tandem on the X chromosome - resulting in a pseudo-diploid state - to match expression levels of autosomal-interacting partners. In this scenario, these highly expressed pseudo-diploid genes would resolve dosage imbalances, avoiding the limitations imposed by maximal transcription rates. An abundance of X-linked gene duplication after sex chromosome differentiation provides initial support for this hypothesis, however many of these genes exhibit sex-biased expression (Potrzebowski et al., 2010; Zhang et al., 2010a).

Based on observations of maximal expression level constraint of X-linked genes (Hurst et al., 2015), we propose that the most highly expressed genes on the X chromosome are more likely to have duplicated (i.e., are more duplicable) after X chromosome formation. Here we examine the relationship between X-linked gene duplication status and expression level. We find that maximal expression levels of X-linked genes that have duplicated in tandem after sex chromosome formation (here referred to as “young X-linked duplicates”) are lower than singleton
levels, and also have lower levels than X-linked genes that duplicated on the X before sex chromosome formation (“old X-linked duplicates”). In addition, young X-linked duplicates have lower maximal expression levels than young autosomal duplicates. Separating duplicates into expression level categories we find no enrichment of young X-linked duplicates among the most highly expressed genes compared to young autosomal duplicates. Overall, these data suggest that tandem duplication is not a general mechanism for highly expressed X-linked genes to escape constraints on maximal expression level.

5.2 Materials and Methods

5.2.1 Gene duplication status and stratification.

We obtained human protein-coding gene annotations from Ensembl (GRCh38) with their associated paralogs and duplication timing inferred from Ensembl gene trees (see Chapter 2) (Aken et al., 2017). Genes were then filtered to include only duplications where both paralogs are present on the same chromosome, i.e. presumed tandem duplications. In addition, these tandem duplications were further stratified into pre- and post-X-formation (i.e., old and young) duplicates for both X-linked genes and autosomal genes, based on their inferred duplication timing. Duplications occurring on the Therian branch or in younger taxa (i.e., < 160 mya) such as Eutheria, Primates, Simiiformes, Catarrhini, Homo/Pan/Gorilla, and Human (see Figure 2.2), are considered young X-linked duplicates or young autosomal duplicates (Kumar et al., 2017), with others classified as “old”. These gene sets are non-overlapping such that if a gene has duplication events both before and after X formation it is placed in the young duplicate set and excluded from the old duplicate set. For each gene we also calculated a duplicability score, which is the number of duplication events occurring within that gene family.
Duplication as mechanism to restore dosage of X-linked genes

(number of events = number of genes in family − 1).

5.2.2 Expression data.

RNA-seq expression data for each gene was obtained from the GTEx Consortium dataset, comprising 11,688 samples across 53 tissues from 714 donors (GTEx Consortium, 2015). Maximal expression levels in transcripts per million (TPM) for each gene were calculated by taking the median expression value across all samples in its most highly expressed tissue. Genes with 0 TPM were excluded from the analysis. For some genes their Ensembl IDs were not present in GTEx, as GTEx uses an older Ensembl annotation build and some genes have since had their IDs updated. In these cases HGNC gene names were extracted from Ensembl and used to obtain associated expression levels. Maximal net expression for paralogs with a single duplication post- or pre-sex chromosome formation was determined by summing median expression values of the paralog pair for each tissue and taking the highest value across tissues.

5.3 Results

5.3.1 Tandem duplications on the X chromosome and autosomes.

We aim to test the hypothesis that some (highly-expressed) X-linked dosage-sensitive genes regained ancestral dosage levels by tandem gene duplication on the X. Here, we use tandem duplication to mean duplicate genes which are present on the same chromosome, regardless of intergenic distance. Using paralogous tandem duplicate relationships inferred from Ensembl gene trees, we found an overall total of 1,124 duplicate pairs on the X chromosome, which include 382 unique
**Table 5.1** | X chromosome and autosomal tandem duplicate and singleton genes.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Expressed*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X chromosome</td>
<td>Autosomes</td>
</tr>
<tr>
<td>Young Duplicates</td>
<td>244</td>
<td>2955</td>
</tr>
<tr>
<td>Old Duplicates</td>
<td>138</td>
<td>3277</td>
</tr>
<tr>
<td>Singletons</td>
<td>125</td>
<td>5244</td>
</tr>
</tbody>
</table>

* Median TPM > 0 in the most highly expressed tissue per gene.

genes comprising 130 gene families. Additionally, we found 11,825 duplicate pairs on the autosomes, which include 6,232 unique genes from 2,533 gene families. We excluded non-tandem duplicate pairs from these analyses. Furthermore, we identified 125 and 5,244 singletons on the X chromosome and autosomes, respectively (Table 5.1). Next, we stratified duplicate pairs into non-overlapping gene sets according to whether the duplication event occurred before or after X chromosome formation, inferred from gene trees (see Materials and Methods). These sets include 244 young X-linked and 138 old X-linked duplicates, as well as 2,955 young autosomal and 3,277 old autosomal duplicates (Table 5.1).

### 5.3.2 More frequent tandem duplications on the X relative to autosomes after sex chromosome formation.

It was previously shown that there has been a burst of new retroduplicated genes (retrogenes) both on and off of the X chromosome after sex chromosome evolution (Emerson et al., 2004; Potrzebowski et al., 2008, 2010). However, whether there is a greater frequency of DNA-based tandem gene duplications on the X was, to our knowledge, not reported. Our hypothesis predicts an increase in the number of tandem duplication events - either DNA-based or retroduplication based - on the X chromosome relative to the autosomes after sex chromosome formation, at least for highly expressed genes experiencing transcriptional traffic.
Figure 5.1 | Post sex chromosome formation duplication event distributions.

(A,B) Density plots of the number of tandem duplication events for each gene family on the X chromosome and autosomes, after sex chromosome formation. A subset (i.e., the distribution between 3 and 32 duplication events) of the overall distribution shown in (A) is given in (B). (C,D) Histograms showing the number of young tandem duplication events per 10 genes separated by chromosome for all gene families (C) or for families with 10 or fewer duplication events (D). The number of duplication events = number in gene family − 1. The autosomal mean number of duplications per 10 genes and the number of X chromosome duplications per 10 genes are also shown.
jams. We analysed the number of tandem duplication events per chromosome, scaled by the number of genes on each chromosome. We find that, since sex chromosome formation, the X chromosome has experienced more tandem duplication events than any of the autosomes except for chromosome 19 (Figure 5.1C). The X chromosome has 2.18 duplications per 10 genes, whereas the autosomes have, on average, 0.92 duplications.

However, these data may be biased by the inclusion of large gene families with many duplication events. For instance, chromosome 19 harbours a family of transcriptional repressors, KRAB-zinc finger genes, which have undergone a large expansion in primate lineages (Hamilton et al., 2006; Lukic et al., 2014). This family is likely the source of the large number of tandem duplications we observe for chromosome 19 (Figure 5.1C), which has also been observed previously (Grimwood et al., 2004). Additionally, the melanoma-associated antigen (MAGE) gene family appears to also have undergone a large recent expansion on the X chromosome (Katsura and Satta, 2011). Of the 37 MAGE family members currently identified in human, 33 are present on the X arising from a single ancestral copy in eutherians. Many MAGE members are specifically expressed in reproductive tissues, and are often aberrantly expressed in different cancer types (Weon and Potts, 2015).

We examined the distribution of the number of duplication events per gene family to determine an optimal threshold at which to exclude large gene families (Figure 5.1A,B). A threshold of ten or fewer duplication events appears to retain the majority of gene families while excluding those that are significantly larger and obviously unusual. After excluding larger gene families we find that the X chromosome has the largest number of tandem duplications per 10 genes, though chromosomes 19 and 11 also appear to have more duplication events than average, suggesting these chromosomes may harbour additional gene families having
undergone recent expansion (Figure 5.1D).

Overall we find that the X chromosome exhibits a larger number of tandem duplication events after sex chromosome formation, relative to most of the other autosomes. This finding is consistent with the model of X-linked gene duplication as a means to restore expression levels.

### 5.3.3 Young X-linked duplicates have lower expression levels than other duplicate categories.

If highly expressed X-linked genes have duplicated after sex chromosome formation to restore pseudo-diploid levels then we might expect to see higher expression overall of young duplicate genes relative to singletons and old duplicates. We determined maximal expression levels for each gene by identifying their median TPM level in their most highly expressed tissue. Genes with 0 median TPM were excluded, as these genes are likely expressed in tissues or at developmental stages not interrogated in the GTEx database (Table 5.1).

We examined the overall distribution of maximal expression levels for young and old duplicates and singletons for both the X chromosome and autosomes (Figure 5.2A). Contrary to our expectations we find that young X-linked duplicate genes have overall lower expression levels than old X-linked duplicates and singletons (Bonferroni-corrected $P = 7.6 \times 10^{-8}$ & $P = 2.4 \times 10^{-9}$, respectively, Mann-Whitney U tests). Furthermore, young X-linked duplicates also exhibit lower expression than young autosomal duplicates (Bonferroni-corrected $P = 0.0088$, Mann-Whitney U test). In case median TPM does not accurately reflect maximal expression levels we also investigated this relationship using the 90th percentile TPM values in their most highly expressed tissue. We did not use the overall highest TPM values, as for many genes these values likely represent unreliable outliers. While the individual expression distributions shifted towards
Figure 5.2 | Expression distributions of tandem duplicate and singleton genes.
(Continued on the next page.)
Figure 5.2 | Expression distributions of tandem duplicate and singleton genes.

(A) Raincloud plots showing the distributions and box plots of young and old duplicate and singleton genes for the X chromosome and autosomes. Maximal median expression for each gene is the median TPM value across samples in its most highly expressed tissue. (B) Raincloud plots showing expression distributions for genes using maximal 90th percentile expression values - which is the 90th percentile TPM value across samples in its most highly expressed tissue. Black dots denote the distribution means. The numbers of genes in each category are indicated. Bonferroni-corrected $P$-values were calculated by Mann-Whitney U test.

Higher values by an order of magnitude, we did not observe any difference in between-distribution trends using the 90th percentile TPM values (Figure 5.2B). Thus, these results do not support the hypothesis of higher duplicability for the most highly expressed young X-linked genes.

5.3.4 Highest expressed genes are not enriched for young X-linked duplicates.

Even though overall expression levels of young X-linked duplicates are lower than other gene sets we may still see an enrichment of these duplicates when categorising genes by expression level. Our hypothesis predicts an enrichment of young X-linked duplicates amongst the most highly expressed genes relative to young autosomal duplicates and old X-linked duplicates. We separated young X-linked duplicate genes into 5 expression categories (Figure 5.3A). We then compared the proportions of young X-linked genes and young autosomal genes across expression categories, as we expect a greater proportion of young X-linked duplicates in the highest expression category if duplication is driven by transcriptional traffic jams (Figure 5.3B). However, we find no enrichment of young X-linked duplicates in this category, but we do find that the lowest expression category is enriched for these duplicates (Bonferroni-corrected $P = 0.046, \chi^2$-test). Furthermore, we
Duplication as mechanism to restore dosage of X-linked genes

Figure 5.3 | Tandem duplicate genes by expression categories.
(A) Histogram showing the number of young X-linked duplicate genes falling into each expression category. (B) Histogram of young X-linked and autosomal duplicates by expression category. (C) Histogram of young X-linked and autosomal duplicates and old X-linked and autosomal duplicates by expression category. Red and blue stars indicate enrichment or depletion of duplicates for that category based on significantly different standard residuals. Bonferroni-corrected $P$-values were calculated by $\chi^2$-test.

compared the proportions of old duplicates also, and we find that old autosomal duplicates are enriched within the highest expression category, whereas young duplicates, both for the X chromosome and autosomes, are depleted (Bonferroni-corrected $P = 0$, $\chi^2$-test; Figure 5.3C).

Additionally, we examined whether we see a correlation between gene duplicability and expression level for these genes, excluding large gene families as before. Among young autosomal duplicates we find an overall negative association between maximal expression and duplicability ($\rho = -0.23$, $P = 4.14 \times 10^{-22}$, Spearman rank-order correlation test; Figure 5.4C). We did not find any association for young or old X-linked duplicates or old autosomal duplicates (Figure 5.4A,B,D). However, while not significant we do observe an overall upwards trend of duplicability to maximal median expression for young X-linked duplicate genes (Figure 5.4A). In summary, these data suggest that there is no increase in X-linked duplicates among highly expressed genes after sex chromosome forma-
Duplication as mechanism to restore dosage of X-linked genes

5.3.5 Net expression of X-linked and autosomal duplicate pairs.

If X-linked genes have duplicated in tandem due to expression constraints it may be the case that net maximal expression is more important than individual maximal expression. Young X-linked duplicates may have lower expression per gene...
Duplication as mechanism to restore dosage of X-linked genes than singletons or old X-linked duplicates, but a greater overall net expression (Hurst et al., 2015). Thus, we might expect to see higher net expression of young X-linked duplicates relative to net expression of old X-linked duplicates. For simplicity - and to avoid comparing net expression of gene families of different sizes - we restricted these analyses to genes with only one duplication event either before or after sex chromosome formation, comprising 13 young X-linked and 47 old X-linked duplicate pairs, and 246 young and 883 old autosomal duplicate pairs. Taking the highest net median expression across duplicate pairs we find no significant difference in net expression between young X-linked pairs and old X-linked or young autosomal duplicates (Bonferroni-corrected \( P = 0.19 \) & \( P = 0.93 \), respectively, Mann-Whitney U tests). This analysis may be limited by the small sample size for young X-linked duplicates; however, the mean and median net expression values for this set of duplicates were lower than for each of the other sets.

We also explored whether each gene in each young X-linked duplicate pair are most highly expressed in the same tissue as their counterpart. High expression for one duplicate and low or no expression for the other in a given tissue suggests sub-functionalisation by expression divergence, whereas if these genes have duplicated to resolve dosage imbalance we expect expression of both genes in the same tissues. We found that 8 (~61.5%) of the 13 young X-linked duplicates are most highly expressed in the same tissues, compared to 7 (~14.9%) old X-linked duplicates, 111 (~45.1%) young autosomal duplicates, and 126 (~14.3%) old autosomal duplicates. While sample sizes for X-linked genes are small these results suggest that tandem duplication may be a mechanism for restoring ancestral expression levels or resolving dosage imbalance, though perhaps not as a result of transcriptional traffic jams.
5.3.6 Constrained X-linked gene duplication after sex chromosome formation.

Dosage-sensitive X-linked genes are more likely to have experienced dosage imbalances after sex chromosome formation (Mank et al., 2011; Pessia et al., 2012). Any selection for restoration of dosage should have been most acute for these genes. We wished to test whether, after these events, there was an increase in duplication of X-linked genes that were constrained prior to sex chromosome formation. It is expected that copy number constrained genes are more likely to be dosage-sensitive (Rice and McLysaght, 2017b). Here, we use “constrained” to denote genes which have no duplication events between the base of the vertebrates and the origin of therian mammals - a period of ~313 million years - whereas unconstrained (i.e., duplicable) genes have at least one duplication event on lineages leading to human during this time.

Among all X-linked genes (n = 716), we find 251 genes (our unconstrained set) with at least one duplication event - tandem or non-tandem - after the divergence of vertebrate lineages (~473 mya) but before the divergence of therians (~160 mya). Additionally, we find 465 X-linked genes that had no duplication events during this period, i.e. our constrained set. Similarly, we find 5,313 and 8,435 unconstrained and constrained autosomal genes, respectively. After removing large gene families with > 10 members as before, we asked whether X-linked genes genes from our constrained set were more likely to duplicate on the X (in tandem) compared to those from our unconstrained set after X chromosome formation. We find a significant increase in X-linked duplication of the constrained genes relative to unconstrained genes (Bonferroni-corrected $P$-value = $1.39 \times 10^{-15}$, $\chi^2$-test; Figure 5.5A). Additionally, we find that constrained X-linked genes were significantly more likely to duplicate after X formation than...
Duplicated Not duplicated
Post-X gene status

<table>
<thead>
<tr>
<th></th>
<th>Constrained</th>
<th>Unconstrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-linked</td>
<td>140</td>
<td>1611</td>
</tr>
<tr>
<td>Autosomal</td>
<td>88</td>
<td>862</td>
</tr>
</tbody>
</table>

Figure 5.5 | X-linked and autosomal dosage-constrained genes.

(A) Barplot showing the proportions of X-linked constrained and unconstrained genes which have duplicated in tandem after X chromosome formation. Numbers in each category are indicated. (B) Barplot showing the proportions of X-linked and autosomal pre-X constrained genes which have duplicated after X formation. (C) Barplot showing the proportions of X-linked and autosomal pre-X unconstrained genes which have duplicated after X formation. Bonferroni-corrected $P$-values were calculated by $\chi^2$-test.

Constrained autosomal genes (Bonferroni-corrected $P$-value = $5.50 \times 10^{-10}$, $\chi^2$-test; Figure 5.5B). By contrast, the same test for unconstrained X-linked and autosomal genes shows no increase in post-X tandem duplication for unconstrained X-linked genes relative to autosomal genes (Bonferroni-corrected $P$-value = 0.37, $\chi^2$-test; Figure 5.5C). These results are overall consistent with previous reports of an increase in duplication of X-linked genes after X formation (Emerson et al., 2004; Potrzebowski et al., 2008, 2010; Zhang et al., 2010a), though we show here that this is the case specifically for tandem duplication of comparably copy number constrained genes. This indicates that post-X duplication could potentially represent a route to resolving dosage imbalances for constrained X-linked genes irrespective of expression level.
5.4 Discussion

Sex chromosome evolution in therian mammals has led to the utilisation of multiple mechanisms to resolve the gene dosage imbalance caused by these events, including XCI, upregulation of dosage-sensitive genes from large protein complexes, and relocation of retrogenes on to and off of the X chromosome (Lyon, 1961; Ohno, 1967; Emerson et al., 2004; Potrzebowski et al., 2008, 2010; Pessia et al., 2012). Given these observations it is clear that expression conditions were not optimal, or even tolerable, for many dosage-sensitive genes on the X chromosome, which subsequently found themselves as haploid in males and haploid-expressed in females, with some exceptions (Berletch et al., 2011). There has been evidence to suggest that, as would be expected, dosage imbalance is more likely to have been problematic for dosage-sensitive genes, such as protein-complex members (Blanc, 2004; Seoighe and Gehring, 2004; Barker et al., 2008; Mank et al., 2011; Pessia et al., 2012). Conversely, less dosage-sensitive genes would not necessarily have required a return to ancestral levels after sex chromosome formation. Indeed, several lines of evidence indicate that many genes are not restored to diploid-like levels (Xiong et al., 2010; Casci, 2011; Deng et al., 2011; Pessia et al., 2012)

Here, we explored an additional mechanism by which genes could have recapitulated ancestral levels: tandem duplication. In particular, we examined whether this is more important for highly expressed X-linked genes as these may have no opportunity to increase expression from a single allele due to transcriptional “traffic jams”. It has previously been shown that haploid-restricted maximal expression levels have shaped gene content on the therian X (Hurst et al., 2015). Specifically, the X chromosome is not an optimal location for highly expressed genes because of a physical upper limit on the amount of transcription possible from a single promoter for haploid-expressed genes, resulting in transcriptional
Duplication as mechanism to restore dosage of X-linked genes

traffic jams (Vicoso and Charlesworth, 2009). Expression of X-linked genes in human is three times lower than autosomally expressed genes, and retrogenes which have relocated to an autosome from the X chromosome have higher maximal expression levels than retrogenes moving in the opposite direction (Hurst et al., 2015). Additionally, X-linked genes tend to be more tissue-specific, which may be explainable by the traffic jam model if broadly expressed genes are associated with high maximal expression (Lercher et al., 2002; Deng et al., 2011).

Our results suggest there has been an increase in tandem duplications on the X since sex chromosome formation. This initially lends support to our hypothesis, as the rate of occurrence of tandem duplications should be higher on the X relative to the autosomes if X-linked genes duplicated to resolve expression constraints, avoiding transcriptional traffic jams. However, we found that these young X-linked tandem duplicates are expressed at lower maximal levels than both old X-linked duplicates and singletons. Furthermore, young X-linked duplicates have lower maximal levels than young tandem autosomal duplicates, consistent with previous reports of overall lower maximal expression of X-linked genes (Hurst et al., 2015). There was also no enrichment of young X-linked duplicates amongst the most highly expressed genes compared to young autosomal duplicates, though there is a small enrichment of the former amongst lowly expressed genes. These observations are possibly explained by the duplication bias of lowly expressed genes and/or retention of non-essential genes that are less likely to be deleterious (Woods et al., 2013; Hurst et al., 2015).

Net maximal expression is not higher for young X-linked duplicates, though for 8 of the 13 duplicate pairs we examined the maximally expressed tissue is the same for each gene, initially suggesting tandem duplication might be compensating for dosage imbalance for these genes. However, 6 of these pairs are maximally expressed in testis at relatively low levels, suggesting these are male-biased gene
DUPPLICATION AS MECHANISM TO RESTORE DOSAGE OF X-LINKED GENES 145

duplications. This result is consistent with a study showing recent bursts of male-biased gene gain on the mammalian X as a result of sexual antagonism (Zhang et al., 2010a). One of the remaining two duplicate pairs with the same maximally expressed tissues (ENSG00000186310 and ENSG00000186462) are expressed at relatively high levels in the cerebral hemisphere of the brain and are involved in nucleosome complex assembly. The other pair (ENSG00000172465 and ENSG00000185222) are transcriptional elongation factors most highly expressed in the ovary. Both of these pairs duplicated on the eutherian branch, shortly after X chromosome formation, suggesting it is possible these duplications occurred due to constraints on expression level. However, considering we find only two duplicate pairs such as these it is difficult to say whether this is the result of such constraints.

Overall, our results do not support the notion of tandem duplication on the X chromosome as a mechanism for restoring dosage balance for highly expressed genes with constrained maximal expression. However, this study is limited to expression and gene duplication data in human only; it would be of interest to explore the relationship between maximal expression and gene duplication in other therian mammals and in other taxa which have independent sex chromosome origins. Furthermore, a comparison between net expression levels for duplicate pairs and/or gene families and their estimated ancestral single copy expression levels could also reveal whether tandem duplication plays a role in dosage compensation for X-linked genes. It is important to note that a relationship between duplicability and high maximal expression is only expected under the traffic jam model, where duplication may provide a route to avoiding transcriptional traffic jams. Another model is simply that any X-linked gene - not only highly expressed genes - may have duplicated to recapitulate ancestral expression. For example, if mutation biases are such that gene duplication events are more common than reg-
ulatory sequence mutations that increase expression level, then restoring ancestral expression by gene duplication may be the more readily available evolutionary route. Supporting this hypothesis we found evidence that X-linked genes that were relatively constrained in copy number before sex chromosome formation were more likely to have duplicated after these events compared to unconstrained X-linked genes and constrained autosomal genes. These initial results suggest a model of gene duplication of dosage-sensitive genes as a means to restore dosage balance, regardless of expression level. It will be of interest to examine these constrained X-linked duplicates for further signs of dosage compensation.

Dosage imbalance during sex chromosome evolution in therian mammals is certainly an interesting case, with several mechanisms involved in dosage compensation. It remains possible that additional, potentially more subtle, mechanisms are also involved. Uncovering and studying dosage compensation mechanisms should aid in our overall understanding of dosage-sensitivity, both for population-based copy number variation, and copy number conservation during genome evolution.
Chapter 6

Conclusions

Aside from some well documented cases, the majority of phenotypic differences observed between different species cannot be explained solely by nucleotide substitutions within protein-coding genes (Enard et al., 2002; Preuss, 2012). It is argued that changes in gene regulatory elements, rather than amino acid substitutions, are the main contributors to the emergence of phenotypic differences (King and Wilson, 1975; Franchini and Pollard, 2017). Importantly, regulatory elements are also conserved between species (Hon et al., 2017). Myriad processes and elements are involved in regulating expression of coding and non-coding genes, the amounts of which can vary at the cellular, tissue, individual, population, and species levels. Unravelling the complexities of these regulatory processes and the evolutionary pressures exacted on genes due to dosage constraints is crucially important for understanding how different species evolve from their common ancestors, and how various diseases resulting from aberrant expression can form.

Expansion of RNA sequencing efforts over recent years has enabled the discovery and annotation of numerous lncRNAs in the human genome, though only a fraction have known functions (Carninci et al., 2005; Derrien et al., 2012; Djebali et al., 2012; Iyer et al., 2015; Palazzo and Lee, 2015; Hon et al., 2017; Franchini and Pollard, 2017). The ceRNA hypothesis has been proposed as a mechanism of gene expression regulation which would ascribe a role to many lncRNAs and expressed pseudogenes with currently unknown functions (Seitz, 2009; Poliseno
In this thesis I examined expression patterns and evolutionary metrics associated with pseudogenes to provide insight into whether ceRNA functionality is common among these elements under normal physiological conditions. I reasoned that expressed pseudogenes in particular make excellent candidates for ceRNAs competing with their parent genes, given they automatically share homology, and therefore MREs, upon pseudogene formation.

In Chapter 3 I performed an evolutionary case study of BRAFP1, a previously described pseudogene ceRNA in human (Karreth et al., 2015), which I found to have originated at the base of the Catarrhini primate lineage. By contrast, Braf-rs1 is a species-specific pseudogene of Braf in mouse, but the fact that both BRAFP1 and Braf-rs1 exhibit ceRNA activity despite independent origins is especially interesting. Employing number of substitutions as an evolutionary metric I found that the 3’UTR of BRAFP1 is under greater sequence constraint as compared to its CDS region, and exhibits similar levels of constraint as the 3’UTR of BRAF, its parent gene. Given that the majority of currently known MREs occur within the 3’UTRs of coding genes, increased constraint of this region in BRAFP1 is consistent with a ceRNA role. The most effective ceRNAs are those that share multiple different MREs with their targets. I found that BRAFP1 and BRAF have good conservation of multiple shared MREs in both their CDS regions and 3’UTRs. However, their 3’UTRs have a higher MRE density, which is thought to be optimal for effective miRNA binding. This likely also accounts for the increased levels of constraint exhibited by the pseudogene 3’UTR.

Additionally, I explored expression patterns of BRAF and BRAFP1 across human tissues and in some non-human primate samples. For ceRNAs to interact they must be expressed concurrently, and I found BRAFP1 expression alongside BRAF in several adult somatic tissues in human. I also found initial evidence
for *BRAFP1* expression in other Catarrine species, though the sample sizes were small. While *BRAFP1* does not appear to be expressed in adult human brain samples I found that some brain tissues at certain developmental stages exhibit high expression, indicating that this pseudogene may have an important role during embryonic brain development. As the function of *BRAF* as a MEK kinase is to promote cellular proliferation, movement, and differentiation, it is presumably required to be at high levels during embryogenesis. Given the role of miR-9 in brain development and how *BRAF* and *BRAFP1* have a well conserved binding site for this miRNA, expression of the pseudogene during these stages could aid in upregulating *BRAF* to the required levels by sponging *BRAF*-targeting miRNAs which may be required to repress other genes (Radhakrishnan and Alwin Prem Anand, 2016). Overall, I provided evidence of a ceRNA role for *BRAFP1* using evolutionary metrics, which can subsequently be used to identify additional pseudogene ceRNAs.

Chapter 4 expands on my analyses of *BRAFP1* by exploring expression and evolutionary trends of human pseudogenes on a genome-wide scale. Based on my expression cut-offs, many more pseudogenes than were previously estimated are seemingly expressed in at least one tissue. The majority of expressed pseudogenes are more tissue-specific than their parents, consistent with previous reports. Similarly to *BRAFP1*, I found that the 3'UTRs of pseudogenes genome-wide are under stronger sequence constraint than their CDS regions, though this appears to be more often the case for processed pseudogenes. This finding suggests that processed pseudogenes may be more likely to exhibit ceRNA functionality, despite their presumed “dead-on-arrival” status upon formation. Other evolutionary patterns I investigated also support the pseudogene ceRNA hypothesis, such as increased retention of expressed pseudogenes with recognizable 3'UTRs, and greater constraint of expressed pseudogenes as well as those with parent-
correlated expression. I also provide tentative evidence that some pseudogenes may be under recent positive selection in human populations, and my set of top candidates is significantly enriched for pseudogenes with identifiable 3'UTRs. From my findings it appears that pseudogene 3'UTRs in particular are frequently constrained throughout evolution, which is consistent with a widespread ceRNA role for many human pseudogenes.

Mammalian sex chromosome evolution required the balancing of gene dosage to account for expression differences between autosomal and X-linked genes. Multiple mechanisms are thought to have facilitated restoration of dosage balance, but global upregulation of genes on the X chromosome has not been observed. It has been proposed that the X chromosome is an unfavourable environment for high maximally expressed genes due to the traffic jam model, which posits that there is an upper limit on the amount of transcription that can occur from a single locus, and that this constraint means that the diploid expression of a highly expressed gene cannot be recapitulated from just one allele (Hurst et al., 2015). In Chapter 5 I examined the hypothesis that tandem gene duplication on the X chromosome after sex chromosome formation is an alternative mechanism that could restore adequate dosage levels for high maximally expressed genes. Focusing on tandem duplications that occur after sex chromosome formation I found that these duplicate genes do not have either higher individual expression or higher net expression than tandem duplicates on the autosomes or X-linked duplicates that arose prior to sex chromosome formation. The data presented here do not currently support the hypothesis of increased duplicability for high maximally expressed genes. However, I found that X-linked genes constrained in copy number prior to sex chromosome formation were more likely to duplicate on the X after these events relative to autosomal genes, suggesting gene duplication may provide a route to restoring dosage balance for dosage-sensitive genes
irrespective of expression level.

In summary, this thesis explores two potential mechanisms for optimising gene dosage. Tandem duplication on the X chromosome does not appear to be a general mechanism for restoring dosage balance for highly expressed genes, though additional future study may reveal whether this is the case for more lowly expressed dosage-sensitive genes. The evolutionary metrics employed here in Chapters 3 and 4 provide a clear picture that many pseudogenes likely represent novel post-transcriptional regulators of gene expression through ceRNA action and thus the copy number of such pseudogenes is expected to be constrained, though the true extent of their functionality is as yet unknown. These observations highlight the usefulness of evolutionary metrics in identifying functional genomic elements with potentially important implications for human health and disease.

6.1 Future directions.

I provided evidence that many human pseudogenes are under sequence constraint that is consistent with ceRNA functionality. However, my results are generally suggestive only and are limited by currently available public sequence data, including the quality of current genome builds and expression datasets. For instance, better quality genomes with higher coverage and more reliable assemblies would enable more accurate identification of human pseudogene orthologs and calculation of evolutionary metrics. My analyses in Chapter 4 are largely based on whole genome alignments across 20 mammalian species. Genome alignments which include a larger number of species are available, though the genome builds for many of these species are not of high quality. Improving large whole genome alignments such as these would enable pseudogene identification in a greater number of species to provide a more robust exploration of patterns of pseudogene
Expansion of RNA-seq datasets is likely to continue in the coming decades, which should allow researchers to pinpoint tissues in which different elements, including pseudogenes, are expressed. A greater number of samples from different tissues and life stages, for instance during embryonic development, would facilitate enhanced study of tissue-specific elements. In particular, more robust miRNA expression quantification would help researchers investigate the relationship between miRNA and ceRNA expression levels. Additionally, expanded resources for RNA-seq data from tissues in other organisms would be greatly beneficial, as evolutionary expression trends could be more easily unravelled.

In Chapter 5 I examined tandem duplication of X-linked genes as a mechanism of dosage compensation. These analyses were limited to human expression data at current levels, whereas the hypothesis predicts that highly expressed genes duplicate to restore ancestral levels. More robust expression data across mammalian species would further facilitate calculation of ancestral expression levels for many of these genes. I propose comparing estimated ancestral expression levels of highly expressed tandem duplicates to current net expression levels. If the hypothesis I examined here is correct we expect highly expressed duplicates with duplication events occurring after sex chromosome formation to have restored their overall net expression levels, via tandem duplication, to those resembling ancestral singleton levels. The timing of these duplication events should correlate with the timing of evolutionary stratum formation, as it is believed that sex chromosome divergence occurred in steps. This mechanism of dosage balance may be more common specifically for highly expressed dosage-sensitive genes, as conditions on the X chromosome after sex chromosome formation are predicted to be sub-optimal for these genes. Alternatively, less highly expressed dosage-sensitive genes may also duplicate to restore dosage balance, if this is an easier end route
than upregulation to achieving higher expression.

Dosage-sensitive genes on the X chromosome are more likely to have retained a Y gametolog after sex chromosome formation (Bellott et al., 2014). It follows then that these genes on the X will have been less likely to duplicate compared to others, which could be further investigated. Additionally, genes which escape X chromosome inactivation, and thus are expressed from two loci in females, may be under evolutionary pressures to be transcribed at higher levels than can be provided by a single locus, and may be enriched for genes which have retained a Y gametolog in males.

In Chapter 4 I focused on human-centric pseudogenes, as pseudogene annotation is lacking in other species. I propose annotating pseudogene orthologs in other mammalian species, potentially using our method of identification, to enable further study. Additionally, further lineage-specific pseudogenes could also be identified and annotated using currently available pipelines for automated pseudogene identification. Pseudogenes could then be investigated with focus on other mammalian species to determine whether the evolutionary trends observed here for humans remain consistent.

My investigation of pseudogene loci in human populations suggests that at least several pseudogenes may be under recent positive selection. However, these results are limited by the pre-computed statistics obtained from dbPSHP, which likely contain many false positives. I attempted to overcome this limitation by requiring loci to exceed thresholds from multiple statistics, though this may still result in some false positives. A more thorough review of pseudogene loci identified as potentially under positive selection would be beneficial, as identification of such elements could aid in understanding how some novel human-specific traits arose. Furthermore, positive selection could also be investigated for pseudogenes in other species such as mouse. If pseudogenes throughout different lineages are
experiencing recent positive selection this could indicate that pseudogene ele-
ments, acting as ceRNAs or otherwise, are continually being co-opted into gene
regulatory networks, contrary to the long-held belief that pseudogenes are func-
tionless evolutionary relics.

The conclusions I present here are based on in silico inferences using evolution-
ary metrics on sequence data as well as large-scale RNA-seq expression datasets.
Further investigation of pseudogenes with potential ceRNA functions using in vitro or in vivo experiments should provide additional evidence for pseudogene
ceRNA functionality under normal physiological conditions. However, given the
potentially large number of pseudogenes to investigate, identifying the specific
function of each pseudogene and the effects of aberrant expression is likely to
be laborious. This process would be aided by the knowledge that pseudogene
ceRNAs will frequently compete with their parent genes, and so any functional
information about these genes should hint at potential implications of altering
pseudogene regulation and expression. Furthermore, to confirm a ceRNA role
for pseudogenes, predicted MREs could be mutated and the resulting effects ex-
amined, in a similar manner to studies on previously characterised pseudogene
ceRNAs. Pseudogenes functioning as ceRNAs should have reduced or abrogated
efficacy upon MRE mutation.

Clearly, pseudogene ceRNAs represent a class of potentially important post-
transcriptional regulators of gene expression, with implications for human molec-
ular evolution, health, and disease. Improvements to and expansion of genomic
and RNA-seq datasets, coupled with in vitro and in vivo molecular experiments,
will aid in more accurately identifying pseudogenes with physiologically relevant
ceRNA functions.

The work in this thesis illustrates the importance of taking a broad and inclu-
sive view of evolution: the expression of some genes is mediated by expression of
their pseudogenes; the sequence conservation of pseudogenes only makes sense in
the context of effects on other loci; and the expression and duplication of X-linked
genes needs to be understood in the context of the special biology and genetics of
the sex chromosomes. With this thesis I have advanced our knowledge and un-
derstanding of the interdependence of various elements of the genome and shown
how evolutionary analysis can unlock biological secrets.
Bibliography


Caygill EE, Johnston LA. 2008. Temporal Regulation of Metamorphic Processes


BIBLIOGRAPHY


Graves JAM, Peichel CL. 2010. Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome Biol.* 11.


Korneev SA, Park Jh, Shea MO. 1999. Neuronal Expression of Neural Nitric Ox-


Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes


of pseudogene PTENP1 promotes malignant behaviours and is associated with the poor survival of patients with HNSCC. *Sci Rep.* 7.


Lukic S, Nicolas JC, Levine AJ. 2014. The diversity of zinc-finger genes on


Nanda I, Shan Z, Schartl M, Burt DW, Koehler M, Nothwang G, Grutzner F,


Rambaut A. 2016. FigTree v1.4.3. *tree.bio.ed.ac.uk/software/figtree/*


Rothenfluh H, Blanden R, Steele E. 1995. Evolution of V genes: DNA se-


