Investigating the Role of Gene Duplication in Ribosomal Protein Evolution and Testing a Model of Duplicate Gene Retention in Mammals

Asav Dharia

University of Connecticut - Storrs, asavdharia7@gmail.com

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation


https://opencommons.uconn.edu/dissertations/971
Since Susumu Ohno’s seminal work in 1970, gene duplication has been widely recognized as the origin of multi-gene families and a major mechanism of evolutionary change. Understanding forces that govern the evolution of gene families through retention or loss of duplicated genes has been the subject of much inquiry and debate. The key challenge in this debate is accounting for retention of duplicate genes when, in the absence of some countervailing selective pressure leading to their retention, population genetics predicts that the overwhelming majority of duplicated genes should be lost.

In an attempt to investigate the generation and retention of duplicate genes in mammals, the Nelson lab undertook annotation of duplication events in five mammalian genomes. We classified each event by duplication mechanism and duplicate gene fate. This led to two important and unexpected findings: First, half of all conserved duplicates are generated by RNA-based duplication (Retroduplication) events; second, ribosomal protein genes constitute one of the largest classes of conserved duplicated genes in mammals with majority of these duplicates being RNA-based.

The work in this thesis begins with identifying and characterizing all gene duplicates of mammalian ribosomal protein gene (RPG) families. We found an unexpected large amount of intact retroduplicates (RTs) which cannot be readily
explained by Ohno’s classic gene duplication trajectories. Hence, we propose a novel
gene duplication model, Duplication Purification and Inactivation (DPI) that would be able
to account for this phenomenon and ultimately serve in conjunction with other established
models. Specifically, we hypothesize that dominant negative phenotypes prevent fixation
of missense mutations in duplicated genes, thereby extending the survival of intact copies
in the genome.

Together, this thesis work provides a comprehensive history of ribosomal protein
evolution in mammals, comprises a body of evidence that meets or exceeds that
available for any other model of duplicate retention, and establishes the impact of forces
that could influence the fate of every gene duplication event. Thus, the work described
here has the potential to provide one of the most rigorously tested and widely applicable
models of duplicate gene retention since Ohno first articulated the problem in the 1970’s.
Investigating the Role of Gene Duplication in Ribosomal Protein Evolution and Testing a Model of Duplicate Gene Retention in Mammals

Asav P. Dharia

B.S., University of Massachusetts Boston

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy At the University of Connecticut 2015
Approval Page

Doctor of Philosophy Dissertation

Investigating the Role of Gene Duplication in Ribosomal Protein Evolution and Testing a Model of Duplicate Gene Retention in Mammals

Presented by

Asav P. Dharia, B.S.

Major Advisor

________________________________________________
Dr. Craig Nelson

Associate Advisor

________________________________________________
Dr. Victoria Robinson

Associate Advisor

________________________________________________
Dr. Peter Gogarten

Associate Advisor

________________________________________________
Dr. Charles Giardina

Associate Advisor

________________________________________________
Dr. Leighton Core

University of Connecticut

2015
Acknowledgements

I would like to thank my mom Mayuna and dad Piyush for their loving support throughout my life. They are the best parents someone can ask for. Being thousands of miles away, it is amazing how close I feel to them every day. I would also like to thank my wife Jeenal for her love and support during all these years. I would not be writing my thesis without that little pocket-sized dynamite. And finally I would like to thank my uncle Mayur and aunt Shetal (my second parents) and my sister Mansi for their enthusiasm in my endeavors as well as their strong support in tough times. Without my uncle, I would not be where I am today. He has been a constant strength that has been always present in my life. I would also want to thank my parents-in-law for being so understandable and supportive. All in all, it is impossible to imagine how I would have done this without my family. I would also like to thank my lovely neighbor Mr. Bhupendra Patel who believed in me and spent countless hours helping me apply to universities in America and helping my parents deal with the fact that I was leaving home and going thousands of miles away at the age of 17 for the sake of a good education. I also want to thank all my friends for being my life outside of work.

Thank you, Craig, for allowing me to study with freedom and independence and always being there for critical guidance and insight. You have been a great support and a great friend whenever it was much needed. My ability to think like a scientist is all thanks to you. As you say, “You come in here with a skull full of mush; you leave thinking like a scientist.” I hope I have done that and I hope I can keep in touch throughout my career. You have always been two steps ahead of everybody else in terms of your scientific thought process and I really hope you continue to succeed in all your projects and endeavors. You know I am just a phone call away if you need me.

Thanks to my committee for being there when I needed crucial inputs and suggestion. Thanks to Dr. Vikki Robinson for teaching me all my ribosomal biology and helping at every step in my experiments. I would want to thank the rest of my committee as well for their continued support.

Thanks to Fred, Ajay, Albert, Caroline, Ed, Jay and all other graduate students who made lab life such a breeze and aided me at every step during my graduate career. And thanks for your patience during our long lab discussions. You have all become life-long friends for me. Thank you to my undergraduate students; Matt, Amanda, Steven, and Erika for keeping lab life fun and being flexible with my experiments.
### Table of Contents

**Chapter 1 Introduction** .................................................................................................................. 1

- Gene Duplication and Genome Evolution ......................................................................................... 1
- Scales, Mechanisms and Outcomes of Gene Duplication ................................................................. 2
- Retroduplication in the Mammalian Genome ...................................................................................... 3
- Known Examples of Retrotransposed Duplicates ............................................................................. 4
- The Mammalian Ribosome and Ribosomal Proteins ........................................................................ 5
- Known Examples of Mammalian Ribosomal Protein Duplicates ...................................................... 5
- Previous Work and Implications ......................................................................................................... 6

**Chapter 2 Tempo and Mode of Gene Duplicates in RP Families** ................................................ 8

- Pipeline to Reconstruct Mammalian RP Gene Families ................................................................. 8
- Ribosomal Dataset .......................................................................................................................... 8
- Gene Order Helps Illuminate Gene Family Evolution ...................................................................... 9
- Ribosomal Protein Family Member Analyses .................................................................................. 11
- Fate of RP Duplications over time ................................................................................................. 14
- Analysis of Selective Pressure Acting on all RP Gene Duplicates .................................................. 16
- EST Analysis for human and mouse RP duplicate genes ............................................................... 21
- Conclusion ...................................................................................................................................... 23

**Chapter 3 Current Status of Gene Duplication Models in Light of Duplication Mechanisms**  ................................................................................................................................. 26
Molecular Mechanisms of Gene Duplication ................................................................. 26

Outcomes and Models of Gene Duplication .................................................................. 31

Do Existing Gene Duplication Outcomes and Models Account for the Retention of Large
Numbers of Retroduplicates? .......................................................................................... 36

Examples of Gene Duplicates in light of Mechanisms and Models......................... 39

Retention of RT-RPs cannot be readily explained by current models....................... 43

Chapter 4 Introducing and Testing the DPI Model to Explain Retention of Previously
Unaccounted Gene Duplicates ......................................................................................... 44

DPI Model Explains the Impact of Dominant-negative Effects on Duplicate Evolutionary
Trajectory ......................................................................................................................... 44

  *In silico* and *in vitro* predictions and tests for the DPI Model ...................................... 46

Testing Key *In Silico* Predictions of the DPI Model ...................................................... 51

Comparative Analyses of DN and Non-DN Gene Family Sizes ................................. 51

Selective Pressure Analyses of DN and Non-DN Gene Families ............................... 52

Testing Key *In Vitro* Predictions of the DPI Model ...................................................... 54

Ribosomal protein gene candidates and CRISPR-Cas9 sgRNA designs ................. 54

Efficient CRISPR-Cas9 mediated gene targeted knockouts of source and duplicate
genes of human RPS15 and RPS26 ribosomal proteins ............................................. 57

CRISPR/Cas9 Gene Knockouts of ribosomal protein source genes and their gene
duplicates ......................................................................................................................... 59
Appendix 1 .........................................................................................................................86
Appendix 2 ..........................................................................................................................123
Appendix 3 ..........................................................................................................................134
Authoring Papers ...............................................................................................................134
Chapter 8 References ........................................................................................................136
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Pipeline for ribosomal protein family analyses</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Diagram illustrating the computation of the maximum number of unique homologous matches</td>
<td>10</td>
</tr>
<tr>
<td>Figure 3</td>
<td>RP gene duplicates in 8 mammalian genomes</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Example of the inferred evolutionary history for duplications of the ribosomal protein gene Rpl36a</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Observed frequencies for RNA-mediated duplicates are much higher than expected frequencies in RP families</td>
<td>14</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Ribosomal Protein Family duplication events based on age</td>
<td>15</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Selective Pressures on Ribosomal Protein Gene Duplicates</td>
<td>17</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Gene Tree for RPL10A showing PAML branch specific omega values leading up to a clade.</td>
<td>18</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Scatterplots for pair-wise and branch-wise Ka against Ks values show that both methods capture the strong selective pressure acting on the RP gene duplicates</td>
<td>19</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Human/Mouse EST Counts and evolutionary selective pressure</td>
<td>23</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Common Mechanisms of Gene Duplication.</td>
<td>28</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Mechanisms in Light of Gene Duplication Models.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Illustration of the Duplication, Purification and Inactivation (DPI) Model</td>
<td>45</td>
</tr>
<tr>
<td>Figure 14</td>
<td>In Silico Pipeline to test DPI predictions</td>
<td>48</td>
</tr>
<tr>
<td>Figure 15</td>
<td>In Vitro Pipeline to test DPI predictions</td>
<td>50</td>
</tr>
<tr>
<td>Figure 16</td>
<td>DN and non-DN gene family analyses</td>
<td>52</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Selective Pressures on DN and non-DN gene families</td>
<td>53</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Ribosomal protein gene candidates to test the DPI model</td>
<td>55</td>
</tr>
<tr>
<td>Figure 19</td>
<td>qRT-PCR gene expression testing for ribosomal source and duplicate protein gene candidates.</td>
<td>56</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Representative sequences of RPS15 and RPS26 source and duplicate locus targeted by the Cas9n for CRISPR knockouts</td>
<td>58</td>
</tr>
<tr>
<td>Figure 21</td>
<td>CRISPR CAS9 aided knockouts of Source RPS15 and RPS26 genes</td>
<td>60</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Induced expression of the duplicate gene can rescue source gene function, but at a fitness cost arising due to missense mutations</td>
<td>64</td>
</tr>
<tr>
<td>Figure 23</td>
<td>RPS15 and RPS26 duplicate protein missense mutation illustrations</td>
<td>65</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Population Genetics representation of the DPI (Duplication, Purification and Inactivation) model.</td>
<td>68</td>
</tr>
</tbody>
</table>

List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Likelihood Ratio Test statistics (LRT) for random site models</td>
<td>20</td>
</tr>
<tr>
<td>Table 2</td>
<td>Gene duplication Mechanism in light of Models. Expanded version of Figure 12</td>
<td>42</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

Gene Duplication and Genome Evolution

The duplication of genes and genomes was postulated to be an important process for evolution of functional and organismal diversity long before we entered the genome sequencing era. Many of the groundbreaking intellectual concepts for this hypothesis come from the work of Susumu Ohno. In a seminal publication of the book *Evolution by Gene Duplication* (1970), Susumu Ohno claimed that gene duplication was the mechanism by which raw material required for the diversification of gene function was created. Since then, a lot of advances have been made in understanding gene duplication and it is now well regarded that gene duplication events play a seminal role in giving rise to complex gene families like HOX, WNT, FGF, etc. that execute many of the most important functions in vertebrates and plants. Recent data reveal that lineage specific expansion and contraction of gene families is more rapid than previously appreciated and creates major differences in gene family size between closely related mammalian genomes. These differences are likely to have been a major contributor to the divergence of mammalian lineages and human evolution. For this reason, understanding the forces that lead to the retention or loss of duplicated genes and that govern the evolution of gene families is fundamentally important to understanding genome evolution.
Gene duplication occurs at all genomic scales, ranging in size from a single gene to a whole genome and occurs and varies greatly in frequency. Depending on the nature of the duplication, these events can either positively, negatively or neutrally affect a given species. For example, duplications involving a single gene or a set of genes can be associated with enrichment for some essential functions, while large scale duplication events can be associated with important evolutionary transitions or major leaps in development and adaptive radiation of species. There are four major mechanisms by which DNA can be duplicated: 1) whole-genome duplication (WGD), 2) tandem Duplication (unequal crossing over), 3) DNA transposition, and 4) retrotransposition. Of these mechanisms, WGD, tandem duplication, and duplicative transposition are DNA-mediated duplication events, while retrotransposition is RNA-mediated. Duplicative transpositions and tandem duplications are what most biologists envision when thinking about the origin of gene families. They are DNA-mediated processes that preserve varying amounts of the source gene’s intron-exon structure and often, depending on the scale of the duplication, varying amounts of intergenic regulatory DNA flanking the duplicated gene increasing their likelihood of them being functional. In contrast, retrotransposition is a process whereby a spliced mRNA transcript is reverse-transcribed into DNA and randomly re-integrated into the genome, creating an intronless copy, devoid of the promoter and enhancer elements of the source gene. Such retrogenes have traditionally been regarded as non-functional however; recent studies have shown that rampant retrotransposition can create genes that function as protein-coding genes and small RNA’s. The recent rise in
importance of studying retrogenes can be credited to the findings of *Fgf4* and *c1orf37-dup* that have been classified as functional retrogenes in mammals\(^\text{20,21}\).

All duplicate genes are expected to be initially carried by a single member of a population and are hence highly vulnerable to stochastic loss early in their evolutionary history. To be successful in the long term (i.e. – fixed and retained in the population) a duplicate gene must drift towards fixation, and then having arisen to high frequency, the selective forces for its maintenance must be sufficiently large to prevent its subsequent loss by degenerative mutation\(^\text{22}\). Hence, understanding why duplicates are not lost has a fundamental bearing on genome evolution. Susumu Ohno recognized that duplicate genes may be retained in a genome as a result of three distinct evolutionary trajectories: conservation of function for the modulation of gene dosage, subfunctionalization for partitioning multiple functions to distinct gene copies, and neofunctionalization generating novel functions in the newly duplicated gene\(^1\). Since 1970 several useful extensions of these models have been proposed including: MDN (Mutation During Nonfunctionality)\(^1,23\), DDC (Duplication Degeneration and Complementation)\(^24\), EAC (Escape from Adaptive Conflict)\(^25\), and IAD\(^8\) (Innovation, Amplification, Divergence) amongst others.

**Retroduplication in the Mammalian Genome**

Previously, we reported that conserved retroduplicates are widespread in mammals, representing half of all gene duplicates under purifying selective pressure\(^2,26\). In addition, we noted that individual gene families have a strong tendency to evolve via DNA-mediated or RNA-mediated duplication, but not both. Developmentally important
classes of genes, such as transcription factors, which often require large amounts of regulatory information to function properly, tend to evolve through DNA-mediated events. However, gene families involved in metabolic processes, such as protein synthesis, evolve primarily through RNA-mediated duplication. In fact, we reported that ribosomal protein (RP) genes are the largest class of conserved, retroduplicated genes in mammals\(^2\). While it is not surprising that the highly abundant ribosomal protein transcripts appear to be more frequently captured by retroviral reverse transcriptase than less abundant transcripts, it is intriguing that the slowly-evolving, highly-conserved ribosomal proteins have hundreds of intact duplicates in the genome.

**Known Examples of Retrotransposed Duplicates**

One of the most prominent examples of a RNA-mediated duplicate is *Fgf4*, a retrogene associated with the breed-defining chondrodysplasia in domestic dogs\(^20\). A human-specific example is the *C1orf37*-duplicate, derived through retrotransposition after divergence of human from chimp expressed selectively in certain human tissues, such as brain. It is suggested to encode a novel transmembrane protein\(^21\). Similar examples include TRMT12 retrogene\(^27\), IMP3 gene\(^28,29\) and other such retrogenes (see\(^30–32\)). The majority of the aforementioned RT genes follow the convention that most retrogenes are in a state of “relaxed” selection. The molecular evolution of retrogenes is selectively neutral, allowing them to freely mutate, giving them a chance to be inactivated or positively selected, while the source genes remain subjected to purifying selection\(^33,34\).
The Mammalian Ribosome and Ribosomal Proteins

The ribosome is an ancient molecular machine that is responsible for production of protein in all living cells. The mammalian ribosome consists of 79 RPs and four rRNAs. RPs play a central role in protein synthesis, are expressed at high levels, and evolve very slowly showing strong conservation across the three domains of life. Proper ribosomal biogenesis requires equimolar production of all RPs and rRNAs. These transcriptional regulatory constraints have been extensively elucidated in various studies, along with the evidence that different ribosomal protein promoters exhibit equipotent strength. Additionally, a strict copy number constraint is also observed as the majority of full length RP’s have been shown to be single copy genes.

Due to the necessity of protein synthesis in any living cell, and the complexity of ribosome structure and assembly, it is perhaps unsurprising that mutations in ribosomal genes almost inevitably lead to pathological conditions such as Minutes in Drosophila and Diamond-Blackfan anemia (DBA) in humans. Despite, or perhaps because of, their stringent conservation, the evolution of RPs in vertebrates is relatively understudied.

Known Examples of Mammalian Ribosomal Protein Duplicates

One of the most recently published DNA-mediated ribosomal protein (DD-RP) gene duplicates is RPL22L1, paralog of ribosomal protein RPL22. These mouse paralogs play essential, distinct, and antagonistic roles in hematopoietic development. Another known rodent-specific RT-RP duplicate is Rps23rg1, a gene originating from a retrotransposition of s23 mRNA that encodes proteins that decrease Alzheimer's β-
amyloid level and tau phosphorylation\textsuperscript{44}. There is also evidence for a ubiquitously expressed RT-RP duplicate, Rpl36al, and testis-specific RPL10L duplicate that have been implicated in compensation for the reduced dosage of X-linked RP genes\textsuperscript{45}. As mentioned in a previous section, RPL3L is a DD-RP duplicate that has been observed to be highly expressed in a group of tissues where the source gene RPL3 has very little expression, exhibiting a potential functional role\textsuperscript{37}.

**Previous Work and Implications**

The project began in our lab when Dr. Jin Jun and Dr. Craig Nelson uncovered the importance of determining ancestral orthology between genes in distinct genomes and contrasted the definition of ancestral orthology with functional orthology\textsuperscript{2}. They presented a new approach to identify ancestral orthologs by utilizing non-coding characters of constituent genes. Specifically, using local synteny information and intron content to identify orthologous genes in mammalian genomes. This method was able to distinguish retrotransposed duplicates from ancestral orthologs in cases where other methods like Inparanoid\textsuperscript{46} failed to do so. Leveraging the fact that local synteny was able to accurately determine orthology, identify the duplication mechanism, and infer a gene family history Dr. Jun applied these methods to measure the relative contribution of DNA- and RNA-based gene duplication events to new functional genes in five mammalian genomes. This effort yielded two very important and unexpected findings. First, half of all duplicate genes under purifying selective pressure are generated through RNA-based duplication mechanisms (retroduplication)\textsuperscript{3,26,47} and secondly, ribosomal protein genes constitute the largest class of conserved retroduplicated
genes\textsuperscript{3}. These results were unexpected because retroduplication has long been considered to give rise almost exclusively to pseudogenes, and only very rarely to functional genes. Moreover, ribosomal protein genes comprise an old and slowly evolving protein family\textsuperscript{35,36}, not the kind of gene family one would predict to be actively evolving through the conservation of a large number of recent retroduplicates. More importantly, we found that none of the existing models for the retention of duplicated genes can readily account for these observations. Taken together, these findings suggest that the current conceptual framework for understanding the evolution and retention of duplicated genes is missing important knowledge including models that adequately account for the retention of half of all conserved duplicates observed in mammalian genomes.
Chapter 2 Tempo and Mode of Gene Duplicates in RP Families

Pipeline to Reconstruct Mammalian RP Gene Families

Ribosomal Dataset

Seventy-six ribosomal protein (RP) sequences from nine species [human, chimp, monkey, mouse, rat, dog, cow, opossum, and chicken (outgroup)] were manually collected from Ensembl 62. Three RPs were excluded due to annotation issues. When a single gene encoded multiple transcripts, the longest was used. These protein sequences served as seed sequences, or input, to the pipeline (Figure 1).

Figure 1 - **Pipeline for ribosomal protein family analyses.** Protein sequence for all source ribosomal proteins were collected manually from Ensembl62. These were input to tBLASTn against whole genomes to capture all putative duplicates. The resulting duplicates were processed by Pseudopipe to determine the mechanism of duplication (DNA or RNA) and the fate of the duplicate (intact or pseudogenized). We then utilized our in-house pipeline steps of hierarchical clustering by local synteny in order to build our gene family trees after filtering false-positives and redundant entries. Final gene family analyses were conducted in 2 steps: 1) calculating the selective pressures on all gene duplicates using the Nei-Gojobori
method against the species- and family-specific seed protein via an exon-based reconstruction, and 2) checking for expression signatures via EST analyses using the UCSC genome browser EST track for both human and mouse.

**Gene Order Helps Illuminate Gene Family Evolution**

Reconstructing phylogenetic trees informs our understanding of the evolutionary history of gene families. Using tree reconciliation between a species tree and gene tree, we can identify duplication and loss events on the tree. Additionally, by distinguishing two duplication mechanisms, DNA-mediated and RNA-mediated, no just identifying duplication events, we can sometimes place duplication events in an appropriate phylogenetic context. While DNA-mediated events are generally very straightforward to place because of the discernable gene orders, RNA-mediated events can be tricky. Local synteny can precisely help us place both DNA and RNA-mediated events before or after speciation. Local synteny information, while very reliable for clear DNA mediated duplication events, often is able to resolve the order of iterative DNA-mediated duplication events in large gene families as well. Even when the coding sequence of the RNA duplicated drifts apart, pre-speciation RNA genes often retain local synteny. Conversely, when RNA duplicates are young enough to be indistinguishable by coding sequence comparison, synteny can discriminate between pre-speciation duplications, and independent RNA duplication events in parallel lineages.

**Local Synteny Driven Orthology Definition**

The accurate determination of orthology is central to comparative genomics. Pinpointing the origin of new genes, understanding the evolution of new gene families,
and assessing the impact of gene and genome duplication events all require the accurate assignment of orthology between genes in distinct genomes. In complex genomes with large gene families this task requires differentiating between genes that have diverged through a speciation event (orthologs) and those derived through duplication events (paralogs). Determination of orthology and paralogy is especially challenging in mammalian species. Very large gene families, high rates of gene duplication and loss, mechanisms of gene duplication, and high rates of retrotransposition all combine to make the determination of orthology between mammalian genes difficult.

We define the local synteny of two genes as the maximum number of unique homologous matches between their six neighboring genes (three upstream and three downstream immediate neighbors for each gene. Figure 2 below shows us the diagram illustrating the computation of the maximum number of unique homologous matches. Homology between two neighboring genes is defined as Blastp E-value <1e-5. The homologous matches do not need to be between the genes with the same orientations as we can see in the figure below.

![Diagram illustrating the computation of the maximum number of unique homologous matches](image)

Figure 2 - Diagram illustrating the computation of the maximum number of unique homologous matches. We counted the homologous matches between 3 neighboring genes (shown as filled arrows with corresponding gene orientations) on each side of the two genes of interest (GOI, shown as two black boxes). Homology between neighboring genes (shown as line between genes) is defined as Blastp E-value <1e-5. The homologous matches do not need to be between the genes with the same
orientations (1) or on the same strand (2). Also they do not need to be co-linear (3). When there are many-to-many homologous matches, I choose the maximum unique matches (4). The number of maximum unique homologous matches in this case is 5.

Ribosomal Protein Family Member Analyses

The first step of our pipeline identified all detectable duplicates of RP genes in eight mammalian genomes. RP families included 14,552 gene duplicates in the eight genomes analyzed: human, chimp, monkey, mouse, rat, dog, cow, and opossum (Figure 3A). Although data in figure 3A include duplicates with shared ancestry, the counts for each species represent the number of duplicate genes present in each extant species. To determine if sequencing coverage had a significant impact on our detection of RP gene duplicates, we compared the depth of sequence coverage in each species’ genome to the number of duplications recovered in that species. We found no significant correlation between the number of duplications and genome coverage (Pearson’s $r = -0.353$, $p = 0.391$, Figure 3B). We also tested for bias in duplication types in each species and found no species-specific bias in duplication mechanisms. As we found significant association between species ($p = 6.07e^{-17}$, two-way chi square test, Figure3B), all species were grouped for subsequent analyses. Next we assessed the fate of each duplicate. Of the 14,552 duplication events detected, only 28 of these gene duplications are DNA-mediated (DD) events; the remainder (99.8%) are RNA-mediated (RT) duplications. Approximately 88% of RNA-mediated duplications are pseudogenes (12,800 duplicates), while 12% are intact (1724 duplicates, Figure 3C). We also examined every ribosomal protein gene’s duplication history and evolutionary trajectory in the context of the encompassing species tree.
Figure 3 - **RP gene duplicates in 8 mammalian genomes.**

A) Distribution of duplication events in 8 mammalian genomes. B) Assessment of coverage or species-specific bias in ribosomal protein gene duplicates. C) Representation of DNA and RNA-mediated duplications in RP gene families. Abbreviations: Hs, *Homo sapiens* (human); Pt, *Pan troglodytes* (chimpanzee); Mmul, *Macaca mulatta* (Rhesus macaque); Mm, *Mus musculus* (house mouse); Rn, *Rattus norvegicus* (Norway rat); Bt, *Bos taurus* (cattle); Cf, *Canis familiaris* (dog); Md, *Monodelphis domestica* (gray short-tailed opossum); Gg, *gallus gallus* (chicken).
An example of the resulting information is shown in Figure 4, for the ribosomal protein gene RPL36A. All 74 ribosomal protein gene family history trees are attached in Appendix 1. Hereafter, all the intact RNA-mediated ribosomal protein gene duplicates will be referred to as RT-RPs, intact DNA-mediated copies as DD-RPs and RNA-mediated pseudogenes as RΨ-RPs.

Figure 4 – Example of the inferred evolutionary history for duplications of the ribosomal protein gene Rpl36a. Grey outlined tube tree represents the species tree that includes 8 mammals and chicken. Source intron-bearing gene (in blue). RT-RPs (clear triangles), RΨ-RPs (grey triangles). An RT-RP duplicate generated from one of these events, Rpl36al (in red, at the base of the mammalian lineage on the branch between LCA with opossum and the other mammals) is conserved in all descendent species. All the 74 ribosomal protein gene family history trees are attached in Appendix 1.
Leveraging previously published data by Jun et al; we observed a clear overrepresentation of RT-RPs among 8,872 gene families analyzed as shown in Figure 5 below.

Figure 5 - **Observed frequencies for RNA-mediated duplicates are much higher than expected frequencies in RP families.** Observed frequencies and expected frequencies (in brackets) shown for each speciation node for 5 mammalian genomes. The values were generated using data table created in Jun et al\(^2\). Observed frequencies for RP-RTs were derived from ribosomal families (Number of RP-RTs/Total Number of Duplicates in RPs) and expected frequencies for intact retroduplicates were derived from 8872 non-ribosomal gene families (Number of RTs in non-RP gene families/Total Number of duplicates in non-RP gene families). All diversification times are from Ureta-Vidal et al\(^{50}\).

**Fate of RP Duplications over time**

In the second step of our pipeline, we determined the probable location of each RP duplication event in evolutionary history of these eight species, and distinguished between RNA- and DNA-mediated duplication events (Figure 6). Based on our methodology, Figure 6 clearly shows that the majority of detectable duplications have occurred during recent mammalian evolution: 100 million years ago (MYA) or more recently. However, a significant number of duplications date between 100–300 MYA.
Figure 6 - Ribosomal Protein Family duplication events based on age. All RP gene duplication events are displayed for 8 mammalian species. The bar charts at all speciation nodes show events classified by fate of duplication. The duplication counts on the bar charts are log normalized. RT-RPs are shown in red and RΨ-RPs in green. DD-RPs are not shown due to a very small sample size. The numbers above the bar charts represent the total number of gene duplication events at that speciation node. Age is marked at the bottom of the tree in millions of years (age estimates from [55,103]).

The majority of RP gene duplications older than 90 MYA result in RNA-mediated pseudogenes (RΨ'-RPs) (190), though some events (25) are intact RNA-mediated duplications (RT-RPs), and a very small number (4) are linked to intact DNA-mediated duplicates (DD-RPs) [data not shown for DD-RPS due to small sample size]. It is important to note that many of the more ancient duplications detected represent
incomplete clades; therefore we infer a considerable amount of gene loss. However, our inability to detect these genes may also be due to loss of synteny or other limitations of our pipeline. The majority of duplicates (N=13,588) observed in our dataset are young (91 MYA or younger). However, a few RT-RPs and DD-RPs have been conserved in all (or most) of the 8 mammalian species analyzed (see the base of the tree in Figure 6).

Analysis of Selective Pressure Acting on all RP Gene Duplicates

To gain insight into the forces shaping the fate of these RP gene duplicates, nonsynonymous/synonymous substitution rates were evaluated using pair-wise and branch-wise methods (see Methods section and 17,51,52). For the pair-wise method, we observe that DD-RP dups and RT-RP dups have mean Ka/Ks values of 0.166 (95% CI 0.083, 0.248) and 0.295 (95% CI 0.285, 0.305) suggesting that they are under strong purifying selective pressure. RΨ-RP’s were under relatively less purifying selective pressure with a mean value of 0.455 (95%CI 0.453, 0.458) (Figure 7A). In order to avoid false positives with Ka/Ks > 1, we did not include cases that had a very low Ka and Ks values. Calculation of pairwise DNA sequence distances reveals that the mean sequence distance for DD-RP duplicates was 0.091 (95% CI 0.062, 0.118), for RT-RP duplicates was 0.0059 (95% CI 0.062, 0.118) and for RΨ-RPs was 0.172 (95% CI 0.169, 0.173). This corroborates the evidence from the Ka/Ks analysis suggesting that these sequences are under strong selective pressure (Figure 7). Next we compared selective pressures on all RT-RP duplicates of various ages in each lineage. (DD-RPs were not included in this analysis due to the very small dataset.) Box-Whisker plots (Figure 7B and 7C) showed that RT-RP duplicates at all speciation nodes, irrespective
of their age or lineage, are under strong selective pressures, as determined by Ka/Ks values. However, chimp (Pt) values seem to be an exception, likely due to a small sample size (Figure 7B). The trends appeared similar for all RΨ-RPs as the median Ka/Ks values are similar (~0.45) for all ages.

Figure 7 – Selective Pressures on Ribosomal Protein Gene Duplicates. A) Mean Ka/Ks ratios were calculated for all classes (DD-RPs, RT-RPs and RΨ-RPs) of RP gene duplicates using the Nei Gojobori method. Results were then filtered based on p-values (< 0.1) and the fraction of the source gene represented by each duplicate (> 65%). Error bars represent 95% confidence interval. B) Box and whisker plots for RT-RPs (blue) and RΨ-RPs (green) were generated for inner speciation nodes and C) Extant Species. DD-RPs were not included in the analyses due to small sample size (N=3).
While pairwise Ka/Ks calculations are computationally rapid and provide a good screen for selective pressure, especially within a gene family, for added support we wanted to cross-check our estimates of selective pressure using branch-specific omega values. To do this we used PAML to calculate branch-specific omega values for a sub-sample of 28 RP gene families. An example RP gene tree with all PAML branch-specific omega values is shown in Figure 8 below.

Figure 8 - Gene Tree for RPL10A showing PAML branch specific omega values leading up to a clade. An abridged Gene tree of RPL10A generated by parsimony-based syntenic method (see Methods). The branch specific omega values are listed at each node in purple. Ka/Ks values are represented at all leaves in green. The RT-RP duplicates and their omega values are highlighted in red.
Using this approach we obtained $\text{Ka/Ks}$ values for RT-RP duplicates (mean = 0.162, CI 0.137, 0.188), and for the $\Psi$-RPs (mean = 0.357, CI 0.347, 0.367). As previously mentioned, to avoid false positives with $\text{Ka/Ks} \gg 1$, we excluded cases with very low Ka and Ks values. As observed in Figure 9, both pairwise and PAML-based estimation methods confirm the strong purifying selective pressure acting on RT-RPs ($\text{Ka/Ks} < 0.3$) and a slightly lower pressure on $\Psi$-RPs ($\text{Ka/Ks} < 0.5$).

Figure 9 – Scatterplots for pair-wise and branch-wise Ka against Ks values show that both methods capture the strong selective pressure acting on the RP gene duplicates. Plot of Ka against Ks for RP families with branch-wise and pair-wise methods. A) Distribution for all 76 RP families using the pair-wise selective pressure calculation method. Red dots represent RT-RPs and green dots represents $\Psi$-RPs. The black solid line represents $\text{Ka} = \text{Ks}$ and the red & green line are the best line of fits for the distribution of RT-RPs and $\Psi$-RPs respectively. B) Distribution for 28 RP families analyzed by codeml program in PAML. C) Distribution of the aforementioned 28 families from PAML analysis using the pair-wise method.
As evolutionary pressure is often time dependent, we also plotted Ka against Ks estimated by both pair-wise (Figure 9A) and branch-wise (Figure 9B) methods. As expected the branch-wise method estimates higher divergence, as seen by the large distribution of Ks values (Figure 9B) compared to pairwise method. The influence of strong purifying selection over time is readily observed in Ka values for both methods as the data points of RP-RTs are compressed near the origin relative to RΨ-RPs, which have a much wider distribution (Figure 9). In order to further confirm the nature of the selective pressure acting on our RP-RTs and RΨ-RPs, we also used different codon-substitution models developed by Nielsen and Yang\textsuperscript{53} and Yang et al.\textsuperscript{54}. Random-site models M0, M1a and M2a which assume variation in $\omega$ among sites but not among lineages were fitted to our data. The models used, parameter estimates and log-likelihood values are provided in the Appendix 2 (Table A). Table 1 below shows the results of the LRT tests for these models.

<table>
<thead>
<tr>
<th>Genes</th>
<th>M1a$^a$ vs. M2a$^b$</th>
<th>M1a$^a$ vs. fix_omega=1$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>2</td>
</tr>
<tr>
<td>RPL28</td>
<td></td>
<td>150.153</td>
</tr>
<tr>
<td></td>
<td>P-Value</td>
<td>1</td>
</tr>
<tr>
<td>RPL14</td>
<td>2Δ$\psi$</td>
<td>-70.308</td>
</tr>
<tr>
<td></td>
<td>P-Value</td>
<td>1</td>
</tr>
<tr>
<td>RPS16</td>
<td>2Δ$\psi$</td>
<td>1024.072</td>
</tr>
<tr>
<td></td>
<td>P-Value</td>
<td>0</td>
</tr>
<tr>
<td>RPS18</td>
<td>2Δ$\psi$</td>
<td>-1269.931</td>
</tr>
<tr>
<td></td>
<td>P-Value</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ Alternative model; $^b$ null model; $2\Delta \psi = 2(\psi_1 - \psi_0)$, df degrees of freedom
We applied the simplest of site-based models M0 [30], which assume a uniform ω ratio for all codons, to four random ribosomal genes namely Rps16, Rps18, Rpl14 and Rpl28. The estimated single ω value for each of these trees ranges from 0.22 to 0.35 (Appendix 2: Table A). These values can be interpreted as an average of all lineages in the tree and over all sites in the protein. The low ω range obtained indicates a strong action of purifying selection in the evolution of ribosomal gene duplicates studied. To test if branch-specific omegas are statistically justified, we compared Model M1a (nearly neutral), which constrains Ka/Ks≤1 but not positive selection (Ka/Ks>1) and M2a which allows for positive selective pressure. This comparison leads us to reject the nearly neutral model as seen in Table 1. Our final comparison was model M1a vs. M0 with a fix omega = 1 and we find out that model M1a fits our data better (p-value < 0.0001). These results confirm that purifying selection is the predominant force acting in the evolution of ribosomal protein genes. Hence it further validates the Ka/Ks values obtained from both pair-wise and branch-wise methods.

EST Analysis for human and mouse RP duplicate genes

EST data for human and mouse were mined in the final step of our analysis pipeline. Using very stringent constraints (see Methods), we found evidence of expression for approximately 8% of all human and mouse duplicates. It should be noted that in order to avoid false positives resulting from the strong sequence similarity between source and duplicate genes, a large number of EST matches were filtered out, suggesting that our estimates of active transcription are likely underestimates. The majority of EST data results from duplicates arising along younger portions of the mammalian lineage (younger primate or rodent lineage or the mouse, hominoid and
human branches (Figure 10A)). For example, 320 out of 335 ESTs are either primate-specific or mouse-specific.

Figure 10 - Human/Mouse EST Counts and evolutionary selective pressure. A) Counts of human and mouse genes with EST (green) versus NO EST (blue) at all speciation nodes were calculated. B) Ka/Ks values were calculated for RT-RPs and RY-RPs for gene duplicates with (green) and without expression (blue). C) Pairwise distances for RP-RTs and RY-RPs with and without expression.
Finally, we compared the selective pressure on all RT-RP duplicate genes with evidence of expression (in the form of EST matches) to those without matching EST data. DD-RPs were not analyzed due to the small sample size (N=3). Expressed RT-RP duplicates exhibit significantly higher levels of purifying selection than their non-transcribed counterparts (mean Ka/Ks value of 0.12 (95% CI 0.09, 0.15) compared to 0.24 (95% CI 0.22, 0.26) respectively (Figure 10B)). However, no similar difference in selective pressure is observed between expressed RΨ-RPs, whose mean Ka/Ks value is 0.36 (95% CI 0.31, 0.41), and their non-transcribed counterparts (mean Ka/Ks value of 0.35 (95% CI 0.34, 0.36)) (Figure 10B). Similarly, pairwise sequence distances for these duplicated genes show that expressed RT-RP duplicates (mean pairwise distance of 0.02 (95% CI 0.01, 0.03)) have diverged less than intact non-transcribed duplicates (pairwise distance of 0.06 (95% CI 0.05, 0.07)) (Figure 11C). However, just as in the Ka/Ks analysis above, expressed RΨ-RPs have diverged less than non-transcribed pseudogenized duplicates (0.09 (95% CI 0.06, 0.13) compared to 0.16 (95% CI 0.15, 0.17) (Figure 10C). It is interesting to note that mean pairwise sequence distances are lower for pseudogenes with ESTs, suggesting purifying selective pressure prior to the pseudogenization event.

Conclusion

Here we provide a near-comprehensive study of ribosomal protein gene sequence evolution, duplication, and loss in eight mammalian species. We find that these highly-conserved and highly-expressed genes are, not unexpectedly, frequently
duplicated by retrotransposition, and comprise the largest such class of genes in mammalian genomes. It is quite clear that RNA-mediated RP duplicates (14,524 out of 14,552 events) dominate RP gene families. There is rare evidence of an old DNA duplicate, RPL3L (see the first gene family tree in Appendix 1) that has been retained for function (see^{37,55}). However, the presence of only a very few such old duplicates and a complete absence of recent DD-RP duplicates, implies selection against the retention of DNA-mediated RP duplications. Negative selection against DNA-duplicates combined with the abundance of ribosomal protein gene mRNAs, and the observation that reverse transcription and transposition are more efficient on short GC-poor sequences like the ribosomal mRNAs^{56,57}, likely explain the almost complete dominance of retroduplication events in the evolution of the mammalian ribosomal protein genes.

Less expectedly, we also find that many of these retrotransposed RP duplicates are under strong purifying selective pressure (N=1,724), and that this pressure is greatest amongst transcribed RP retroduplicates, regardless of whether these duplicates have been pseudogenized or retain intact coding regions. As gene duplicates are often found to be under relaxed selective pressures^{33,49,58}, the strength of selective pressure we observe across RT-RP duplicates was unexpected. It was not immediately obvious to us why so many duplicates are under selective pressure when the source ribosomal genes exist almost exclusively in single copy, when DNA-mediated duplications appear to be selected against, where RP transcript levels are tightly regulated for optimal fitness, and the duplications are occurring over a timeframe where ribosomal evolution is thought to be almost stationary. Indeed, we expected RT-RP duplicates to be evolving neutrally for exactly these reasons.
The precise combination of forces enabling the retention of duplicated genes in complex genomes leading to the formation of gene families has been a subject of much study\textsuperscript{1,3}. Several interesting studies have focused on the fate of ribosomal protein duplicates in non-mammalian lineages. RP duplicate fate after WGD events have been closely studied in yeasts and plants\textsuperscript{59,60}. RP duplicates have been shown to be retained to maintain gene dosage after WGD\textsuperscript{13,60–62}. But these retention events are not expected to affect the relative stoichiometry between RPs. However, the primary mode of duplication observed in the present study is RNA-mediated, small-scale duplications, which could result in severe stoichiometric imbalance. Additionally, it has been implied that RP duplicates after WGD’s can be selected for defined functions like increasing levels of gene expression and divergence of gene function\textsuperscript{63}. But evidence for this is not readily apparent in mammalian RT-RPs.
Chapter 3 Current Status of Gene Duplication Models in Light of Duplication Mechanisms

In Chapter 2, we annotate all putative gene duplication events in 79 ribosomal protein gene families for 8 mammalian species, we find more than 1000 old and young highly conserved RNA duplicates. We also discovered that all these duplicates are under strong purifying selective pressure and a large number of these retrotransposed RP duplicates exhibit signatures of gene expression. We decided to take a much deeper look into different gene duplication models to observe whether the retention of this duplicates can be explained in light of various gene duplication mechanisms. When we started evaluating the current literature, we discovered that general models for retention of duplicate genes have not been evaluated for applicability to RT-copies, and our knowledge of the evolutionary trajectory of majority of RNA-based duplications is sparse. In order to fully understand the impact of RNA-based gene duplication on mammalian genome evolution, we evaluate and test every gene duplication mechanism and model that account for the evolutionary retention for bulk of these genes.

Molecular Mechanisms of Gene Duplication

From an evolutionary perspective, there is a clear distinction between the initial establishment of a gene after it duplicates and the secondary modification it undergoes due to mutation and natural selection. Basically, all new genes must arise from the accidental duplications of preexisting genes or parts thereof, which implies an initial state of a single copy in a single member of the population. Thus, understanding the
processes that enable the expansion versus contraction of gene number requires an in depth understanding of the molecular mechanisms that give rise to duplication events and models that can explain the dynamics of newly arisen genes.

There are 4 major mechanisms by which DNA is duplicated: 1) unequal crossing over 2) duplicative (DNA) transposition, 3) retrotransposition, and 4) whole genome duplications. Figure 11 below, adapted and edited from Zhang review paper\textsuperscript{64}, provides an overview of all these gene duplication mechanisms.

![Common Mechanisms of Gene Duplication](image)

**Figure 11 – Common Mechanisms of Gene Duplication.** A) Unequal Crossing Over. B) DNA Transposition. C) Retrotransposition, and D) Whole Genome Duplication.

Although contributions made by each type of duplication to any single genome or species is generally unknown, estimates of the frequency of each can be made based largely on the locations of paralogs (gene duplicates) across the genome.
Unequal Crossing Over

Unequal crossing over occurs when two homologous sequences on different chromosomes misalign during recombination. Depending on the position of crossing over, the duplicated region can contain part of a gene, an entire gene, or several genes. In the latter two cases, introns, if present in the original genes, will also be present in the duplicated genes. Crossing over in a bivalent carrying a duplication in one of the two chromosomes may lead to different consequences\(^6^5\). If the duplicated segment pairs with its homologous segment in the other chromosome in complete disregard of other homologous segments then the unequal crossing over produces duplication of other segments\(^6^5\). If the duplicated segments are present in reverse order of the original segments or if duplication is present on the other arm then the pairing and crossing over forms dicentric along with acentric fragments. If the duplicated segments are on another, nonhomologous chromosome, crossing over with this duplicated region will produce two interchange chromosomes\(^6^5,6^6\). Estimates from Arabidopsis thaliana, Mus musculus, Rattus norvegicus, Homo sapiens, and Saccharomyces cerevisiae put the number of tandemly arrayed duplicates between 10% and 20% of all genes, though the exact meaning of “tandem” varies in each paper\(^6^7–6^9\).

DNA Transposition

Duplicative transposition of DNA sequences can be accomplished by 1 of 2 main pathways: nonallelic homologous recombination (NAHR) or nonhomologous end joining (NHEJ; reviewed in\(^7^0\)). The difference in the 2 pathways is largely based on whether homologous sequences are used as a template during double-strand break repair, and
this difference can also be used to infer the mechanism by which individual genes are duplicated (unequal crossing-over is a form of NAHR, albeit involving closely linked sequences). Bailey et al. found an enrichment of transposable elements at the junctions of interchromosomally duplicated sequences in humans\textsuperscript{71}, a pattern also recently found in Drosophila melanogaster\textsuperscript{72}. Recombination between these non-allelic homologous sequences can result in the duplication of the intervening sequences, which can then lead in turn to more duplications because of pairing between the new paralogs\textsuperscript{71}. But other studies in humans have also found multiple cases with no repetitive DNA or long stretches of homologous sequence at duplication breakpoints, suggesting the action of NHEJ\textsuperscript{73}. Due to the relatively low proportion of duplicated sequences arranged in tandem in the human genome, it has been proposed that duplicative transposition (of one mechanism or another) is the major mode of duplication in humans\textsuperscript{11}.

\textit{Retrotransposition}

Retrotransposition occurs by mobile genetic elements that copy themselves into other regions of the genome, using RNA intermediates and reverse transcriptase. Primarily, retrotransposition occurs when a message RNA (mRNA) is retrotranscribed to complementary DNA (cDNA) and then inserted into the genome. There are several molecular features of retroposition: lack of introns and regulatory sequences of a gene, presence of a poly-A sequence, and presence of flanking short direct repeats\textsuperscript{15}. Because promoter and regulatory sequences of a gene are not transcribed and hence not duplicated by retroposition, the resulting duplicate often lacks necessary elements for transcription and thus immediately becomes a pseudogene. Nevertheless, several
retroposition-mediated duplicate genes are expressed, probably because of the chance
insertion of cDNA into a genomic location that is downstream of a promoter sequence.
Recent studies have found that retrogenes that are integrated near other coding regions
or even in introns of expressed coding sequences are much more likely to be expressed
than those that are integrated far from coding sequences\textsuperscript{74}.

\textit{Whole Genome Duplications}

Whole-genome duplication is an evolutionary process whereby two or more
genomes are brought together into the same nucleus, usually by hybridization followed
by chromosome doubling. This results in new gene copies of every gene in a genome
and, obviously, all the flanking regulatory sequences. Though every gene is duplicated,
only 10–30\% of all genes are maintained in the genome for very long periods\textsuperscript{12,75}. The
type or function of genes maintained after polyploidization appears to differ from those
duplicated by smaller scale mechanisms: Many of the genes kept after whole-genome
duplications exhibit dosage effects (reviewed in\textsuperscript{76,77}). Though an excess of duplicates in
these categories have not necessarily held up in studies of additional taxa (e.g.,\textsuperscript{59,78}),
polyplody events are likely to have had a large impact on genome evolution and gene
duplication overall.
Outcomes and Models of Gene Duplication

**Neofunctionalization**

One of the more notable mechanisms for the preservation of a pair of gene duplicates is the process of neofunctionalization whereby one copy acquire a beneficial mutation that results in a new function. Models of neofunctionalization via gene duplication generally assume that new beneficial functions are acquired at the expense of essential ancestral functions, the unspoken reasoning being that selectively advantageous mutations with no negative pleiotropic effects on wild-type fitness should have had no barriers to fixation prior to duplication. Under this reasoning, the temporary phase of redundancy provided by gene duplication is thought to release one copy from prior selective constraints thereby enabling it to taken on a previously adaptive feature. A very interesting example of neofunctionalization involves the evolution of insecticide resistance in the mosquito *Culex pipiens*. Specifically, the acetyl-cholinesterase enzyme in this species normally plays a role in the central nervous system, but a mutant allele at the locus also confers resistance to organophosphate insecticides. Another well-studied examples of neofunctionalization is *GLUD2*, a duplicate glutamate dehydrogenase gene in humans and apes, that is important for glutamate detoxification after neuron firing, appears to have gained expression in the brain and testes after human-Old World monkey split and it also shows signs of directional selection on its protein sequence.

There are two most common neofunctionalization models namely, Dykhuizen–Hartl model and Adaptation model. Dykhuizen–Hartl model for gene duplication proposes that none of the mutations at the redundant locus are fixed by selection.
Instead, mutations accumulate due to drift and at some later point in time there is a change in environment such that the new version of the duplicated gene is advantageous to the organism. The important feature of this model is that none of the newly arising mutations at the redundant locus ever have a fitness advantage over another segregating allele before they are fixed\textsuperscript{82}. The Adaptation model posits that a new function occurs by the adaptive fixation of mutations at one of the duplicated loci\textsuperscript{83}. However, it is not clear if this model specifies whether the first “illicit” mutation is fixed by selection or whether only subsequent mutations are\textsuperscript{1,58}.

**Subfunctionalization**

The general opinion in the gene duplication literature is that protein neofunctionalization happens on such a large time scale that most duplicates are lost or degenerated through mutations before the occurrence of neofunctionalization\textsuperscript{84}. This brings up an interesting question whether there are other mechanisms at work that preserve gene duplicates for a long enough time for them to acquire novel functions. One of the potential answer to this question can be provided by the outcome subfunctionalization. It hypothesizes that after duplication, the functionality of an ancestral protein can get partitioned over the duplicates, so that both copies are needed to perform the complete ancestral function. These functions may comprise expression domains, for example, expression in multiple tissues; protein operations, for example, functions carried out by different active sites of the same peptide or any other genetic function\textsuperscript{24}. Additionally, this outcome doesn’t require the action of selective forces contrasting to most of the other models and outcomes. A classic example for this
outcome is the engrailed gene in zebrafish. Engrailed-1 and Engrailed-1b are pair of transcription factor genes in zebrafish generated by a chromosomal segmental duplication (discussed in^{24}). Subfunctionalization can be observed as engrailed-1 is expressed in the pectoral appendage bud, whereas engrailed-1b is expressed in a specific set of neurons in the hindbrain/spinal cord.

There are several models that fall under the sunfunctionalization, namely, segregation avoidance, specialization model, gene sharing model and the most famous and highly discussed Duplication-Degeneration-Complementation (DDC) model. Segregation avoidance claims that if balancing selection is occurring at a single-copy locus via heterozygote advantage, then homozygotes will be produced every generation regardless of the strength of selection^{1,85}. Specialization model proposed by Hughes et al. suggest that if the original gene was performing two functions that could not be independently improved, then after duplication each gene copy can be driven by positive selection to specialize — that is, to improve one of the two functions^{23}. Gene sharing model is also derived from the similar principal with the only difference between the specialization and gene sharing models revolves around the number of distinct functions carried out by the ancestral protein^{86}. The quintessential model that has defined subfunctionalization is the Duplication-Degeneration-Complementation (DDC) model. According to the DDC model, degenerative mutations can occur neutrally in both copies as long at the duplicates as a pair retain all ancestral functionality^{24}. After duplication, purifying selection is expected in both genes, but its intensity may be relaxed compared with the pre-duplication phase. At the end of the fate determination
phase, the original function will be partitioned by the two genes in terms of expression or protein function in a neutral manner without the involvement of positive selection\textsuperscript{24,87}.

\textbf{Gene Conservation}

Ohno addressed that this outcome shows that maintenance of a gene duplicate was to simply increase the number of genes coding for a protein. In this case, both loci maintain the original functions of the gene, and hence is known as "gene conservation"\textsuperscript{64}. Multiple authors have also recently proposed that this is a major force in duplicate gene retention\textsuperscript{87–89}. A pertinent example for gene conservation was studied by Perry et al.\textsuperscript{90} discussing the variation in the number of duplicates of the salivary amylase gene (AMY1) among humans. They found that human populations that consume starch-rich diets had on average more copies of AMY1 per individual and that this translated into higher protein levels and enhanced ability to break down starches.

There are two proposed models in this category to explain why these duplicates would maintain the original functions, namely Redundancy model and Dosage model. Redundancy model specifies that a second gene could provide functional redundancy if the original locus was disabled by mutation and the Dosage model posits that there is an advantage to producing more of a gene. While it is certainly true that increased levels of protein production can be achieved by increasing gene expression, duplicating a gene can potentially have an equivalent effect\textsuperscript{1,91}. 
Pseudogenization

It is generally not advantageous for species to carry two identical genes. Duplication of a gene produces functional redundancy. Pseudogenization, the process by which a functional gene becomes a pseudogene, usually occurs in the first few million years after duplication if the duplicated gene is not under any selection\textsuperscript{92}. The two major forces of pseudogenization are mutation and deletion, where pseudogenization occur through promoter mutation, nonsense mutation or missense mutation in coding region, or loss of exon splicing junction. Mutations that disrupt structure and function of one of the two duplicate genes are not deleterious and are not removed by selection. Gradually, the copy of the gene that accumulates mutations becomes a pseudogene, which is either unexpressed or non-functional\textsuperscript{64}. After a long time, pseudogenes will either be deleted from the genome or become so diverged from the source genes such that they are no longer identifiable. Humans and mice have similar numbers of members of the olfactory receptor gene family (~1000 genes) but the proportion of pseudogenes is >60\% in humans and only 20\% in mice. This may be due to reduced use of olfaction since the origin of hominoids, which can be compensated by other sensory mechanisms, such as better vision\textsuperscript{65}. Occasionally pseudogenes may also serve some functions. In chicken, there is only one functional gene (VH1) encoding the heavy chain variable region of immunoglobulins, and immunoglobulin diversity is generated by gene conversion of the VH1 gene by many duplicated variable region pseudogenes that occur on its 5’ side\textsuperscript{93}. 
Do Existing Gene Duplication Outcomes and Models Account for the Retention of Large Numbers of Retroduplicates?

Population genetics suggests that duplicates should be lost long before adaptive forces can fix them in the population\(^{64,94,95}\). To be successful in the long term, a duplicate gene must first drift toward fixation, and then, once it has risen to a high frequency, the selective forces for its maintenance must be sufficiently large to prevent its subsequent loss by degenerative mutation\(^{22}\). Many models have been forwarded that attempt to explain this apparent paradox and provide scenarios within which duplicated genes will be retained at the levels observed in many genomes (for an excellent review see\(^{58}\)). In an attempt to understand the origin of the widespread selective pressure we observe on mammalian ribosomal protein retroduplicates, we focus this discussion on the ability of current models to account for this phenomenon.

Neofunctionalization

After assessing the current literature for existing retrogenes and ribosomal gene duplicates, we wanted to evaluate the gene duplication models. Gene duplication models for neofunctionalization, namely, the Dykhuizen-Hartl model, the Adaptation model, and the Adaptive Radiation model, predict that the rate of evolution after gene duplication will be accelerated in the duplicated copy and constrained in the original gene\(^{58,64,87}\). However, these models fail to account for thousands of ribosomal retrogenes in our dataset which demonstrates that rather than experiencing neutral selection, the new copy is under stringent purifying selection. Moreover, while some extra-ribosomal functions for divergent RP duplicates has been observed these events
appear to be very rare\textsuperscript{96}. Therefore, neofunctionalization models appear unlikely to account for the very large number of conserved ribosomal protein gene retroduplicates in mammalian genomes.

**Subfunctionalization**

Subfunctionalization, and its most cited model, DDC does appear to account for the retention of some number of gene duplicates\textsuperscript{24,58}. DDC postulates that the genetic drift and accumulation of mutations will cause the loss of specific subfunctions from each copy of the duplicated genes. Once one copy has lost an essential function, selection on that function in the other duplicate will be reasserted. Eventually the two copies preserve largely non-overlapping complementary functions and both must be maintained by selection\textsuperscript{24}. This division of function can result from changes in the regulatory regions or the coding regions of duplicated genes, and is most often envisioned as a driving force for the divergence of gene expression [for example see\textsuperscript{62}]

However, DDC seems an improbable model for retention of the ribosomal protein retrogenes due to the fact that rather than appearing to drift, the coding regions of these duplicates are under strong purifying selective pressure. Large numbers of degenerative mutations in the coding regions are not observed until after pseudogenization. Also, because RT-RP duplicates do not carry any regulatory information, the most likely scenario for DDC, the evolution of complementary regulatory regions is unlikely. In addition, EST signatures retrieved from our pipeline and a review of existing literature\textsuperscript{97} suggests that retroduplicates typically have a much narrower expression profile compared to the ubiquitous expression patterning of their source
genes, while the source genes never seem to lose ubiquitous expression, as would be expected under DDC. Hence, division of function in such a manner seems improbable for ubiquitously expressed ribosomal source genes. Other subfunctionalization models like EAC\textsuperscript{25}, and specialization and gene sharing\textsuperscript{98} require neutral selection on the duplicate copy\textsuperscript{87} and are not consistent with the purifying selective pressure we observe.

\textit{Gene conservation}

Gene conservation is another outcome that can be used to explain the retention of retrogenes. The primary gene conservation model that has been employed to explain gene retention is the dosage model, which posits that gene duplicates are retained in order to produce more of the same gene product\textsuperscript{1,99}. In comparison, the dosage compensation model states that the gene duplicates can compensate for the activity of the source gene\textsuperscript{64}. The RP genes are under strict transcriptional regulatory control to maintain equimolar ratio of ribosomal constituents\textsuperscript{100–102}, and changes in ribosomal protein levels, including overexpression, are often highly deleterious\textsuperscript{99,103}. This point is confirmed by DeSmet et al. 2013 paper\textsuperscript{104} as they suggest that retention of small scale duplications (SSDs) will result in the stoichiometric imbalance among protein complexes and that the dosage balance hypothesis would work for a WGD as relative ratios among subunits can be flawlessly maintained, which would not be the case with SSDs. Similar conclusions were drawn for SSDs, suggesting that they would be selected against in a highly connected protein network\textsuperscript{105}. This suggests that retroduplications that alter gene dosage would be selected against, not favored.
Another very important piece of evidence that argues strongly against the retention of RP-RT duplicates by dosage is the study conducted by Kittler et al and Gilsdorf et al.\textsuperscript{106,107}. In this study 34 ribosomal retrogenes (highly conserved old, new, intact and pseudogenized candidates) were knocked-down with no detectable phenotypic defects. However, knock-down of each of 70 source RP genes had drastic phenotypic defects on the cells, with no evidence of retrogenes compensating for the loss of source gene products (data obtained and analyzed from\textsuperscript{107}). Previous work done on \textit{Paramecium tetraurelia}\textsuperscript{75,99} discusses about dosage compensation affecting the short term retention rate of duplicate genes after WGD’s, while maintaining stoichiometry. While they correctly predict selection against the retention of non-balanced duplicates, they do not predict the knock-down results obtained in mammalian RT-RPs discussed earlier.

\textbf{Examples of Gene Duplicates in light of Mechanisms and Models}

As there are 4 different mechanisms by which gene duplication occurs and many different models to explain the retention of gene duplicates, it was important to conduct a systematic study to compare each model and mechanism uniquely to find whether any models successfully explain the retention of highly conserved retroduplicates. Hence we analyzed every gene duplication model to see whether there were existing examples in literature that can explain the retention of certain duplicates and the mechanism by which they duplicate (see Figure 12 and Table 2- for convenience, the references used for the figure and the table 2 are listed separately in the Appendix 2: Table B). In Figure 12, each duplication model is presented in the context of which
Figure 12 – **Mechanisms in Light of Gene Duplication Models.** The four columns in this figure represent the mechanisms by which gene duplication takes place, first 3 columns representing DNA-mediated duplication mechanisms and the last one represents RNA-mediated ones. The rows in this table represent all the standard gene duplications models to explain certain gene duplicate fates. In green, are listed all the known examples when a particular model is tested against the mechanism (For reading convenience, all references for this table are provided in the Appendix 2). We have also categorized probable events (blue text) plausible but unlikely events (orange text). Some of these categories boxes don’t have annotated examples, but can be confidently placed in a category due to literature findings. Improbable events when a model is tested against a mechanism are represented in red text.

duplication mechanisms it takes into account. Boxes with green text represent “Known Examples”, when the model was tested against a mechanism. For example, a particularly interesting example of gene conservation has been uncovered in yeast by Conant and Wolfe\(^{108}\), who hypothesized that retention of specific glycolytic genes after WGD’s in yeast has caused an increased glycolytic flux that gave post-WGD yeast species a growth advantage by increasing their glucose fermentation speed. Hence, this example would mark the explanation of WGD duplication mechanism by gene conservation model, specifically the dosage-balance model. All the other examples

<table>
<thead>
<tr>
<th>DNA-mediated duplication</th>
<th>RNA-mediated duplication</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WGD</strong></td>
<td><strong>Tandem</strong></td>
</tr>
<tr>
<td>Redundancy</td>
<td><strong>KNOWN EXAMPLE(^1)</strong></td>
</tr>
<tr>
<td>Dosage</td>
<td><strong>IMPROBABLE</strong></td>
</tr>
<tr>
<td>Dosage balance effect</td>
<td><strong>KNOWN EXAMPLE(^4)</strong></td>
</tr>
<tr>
<td>DDC on regulatory regions</td>
<td><strong>KNOWN EXAMPLE(^5,6)</strong></td>
</tr>
<tr>
<td>DDC on coding regions</td>
<td><strong>PROBABLE</strong></td>
</tr>
<tr>
<td>Gene sharing</td>
<td><strong>KNOWN EXAMPLE(^7,10)</strong></td>
</tr>
<tr>
<td>NEO-functionalization</td>
<td><strong>KNOWN EXAMPLE(^12)</strong></td>
</tr>
</tbody>
</table>

\* Models that have no superscripted reference belong to ref\(^17\) and ref\(^18\)
(boxes with green text) were reasoned in the same manner for Figure 12 and the expanded table with more examples and scenarios of retention can be seen in Table 2 below. Published examples were used wherever possible and if examples were not possible, we try to reason the possibility of a mechanism to model fit based on the literature. Boxes with blue text represent “Probable” cases and boxes with orange text represent all the “Plausible but unlikely” cases for a particular mechanism and duplication outcome. Boxes with red text represents “Improbable” cases that may not be explained by that particular model. In this figure, we consider all the models for which we found at least one known example. As it is clearly seen in the table, all the mechanisms except for retrotransposition have been represented by one or more gene duplication model. A clear distinction can be observed when it comes to retrotransposition as none of the models within any outcomes, apart from neofunctionalization, can explain retention of RNA-mediated gene duplicates. The known examples for retrotransposed genes that can be explained by neofunctionalization models required the gene duplicate to be free of any selective pressure and be freely evolving to acquiring beneficial mutations. We do not observe freely evolving duplicates as our ribosomal protein gene duplicates are clearly under strong selective pressure and their retention cannot be explained by any of the neofunctionalization models. So after our intensive study of gene duplication mechanisms and existing models, the highly conserved retrotransposed gene duplicates that we discover do not seem to fit any gene duplication model that is currently in literature.
Table 2: Gene duplication Mechanism in light of Models. Expanded version of Figure 12.

<table>
<thead>
<tr>
<th>MODELS</th>
<th>Whole Genome Duplications</th>
<th>Unequal crossing-over (tandem)</th>
<th>Duplicative (DNA) transposition</th>
<th>Retrotransposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redundancy (Hahn 2009; Gu et al. 2003)</td>
<td>The loss of function in one copy can be compensated by the other copy (Yeast)</td>
<td>Potentially all the genes by WGDs went through this stage.</td>
<td>Probable due to the tight linkage.</td>
<td>Hard to get symmetric functions by RT.</td>
</tr>
<tr>
<td>Dosage (Hahn 2009)</td>
<td>There is an advantage to producing more of a gene.</td>
<td>Cannot have more copies on one specific gene if the whole genome was duplicated.</td>
<td>Duplicated antimicrobial genes evolve rapidly in copy number (Baskin et al. 2007) [Fly]. More Alu1 duplicates in the human population than consume starch-rich diets (Perry et al. 2007) [Human].</td>
<td>No examples were found to show this yet. Plausible if regulatory regions duplicate along with the duplicated regions, otherwise unlikely.</td>
</tr>
<tr>
<td>Dosage balance effect (Papp et al. 2003)</td>
<td>Each gene’s expression is expected to remain in the same stoichiometric relationship with all of its interacting partners (Yeast).</td>
<td>Metabolic genes appeared more retained than other genes after the recent WGD (Gout et al. 2009) [Eg: Glycolytic genes in yeast causing increased glycolytic flux (Conant 2007)].</td>
<td>Possible if the genes are highly linked as balancing gets harder as the gene linkage decreases.</td>
<td>Possible if the genes are highly linked as balancing gets harder as the gene linkage decreases.</td>
</tr>
<tr>
<td>DOC (Force et al. 1999)</td>
<td>The duplication-degeneration-complementation (DOC) model predicts that (1) degenerative mutations in regulatory elements can increase rather than reduce the probability of duplicate gene preservation and (2) the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions rather than the evolution of new functions. Qualitative vs quantitative models.</td>
<td>Enlarged genes [Zebrafish] “We conclude that the two gene pairs en1/2 en1/2 deceived rapidly in copy number (Baskin et al. 2007) [Fly]. More Alu1 duplicates in the human population than consume starch-rich diets (Perry et al. 2007) [Human].”</td>
<td>Probably hard to have the separate regulatory regions for each copy in a array, based on the following quote from Force et al. 1999 that, “We focus on duplications of entire chromosomes or genomes rather than tandem gene duplications because we wish to exclude for now complications caused by uncertainty about the extent of the original duplication and local homogenization events caused by unequal crossing over or gene conversions”.</td>
<td>No examples were found to show this yet. However, it seems reasonable to claim that 2 gene pairs that have duplicated by DNA-transposition can end up performing common functions if the coding and regulatory regions duplicate accurately.</td>
</tr>
<tr>
<td>Gene sharing (Piatigorsky 2007)</td>
<td>Requires proteins with multiple distinct functions. (vs. Specialization “the products of the duplicated genes... acquire kinetic properties which are markedly different from each other” e.g. hemoglobin (Hahn 2009)).</td>
<td>Galactokinase encoded by GAL4 in yeast (Yeast)</td>
<td>No examples were found to show this yet. However, tandem duplicated genes could potentially work in this fashion.</td>
<td>Crystallin [Bride] (Piatigorsky, 1989, 1991, 1997). Might be from WGD, but not by TD or RT.</td>
</tr>
<tr>
<td>Neofunctionalization</td>
<td>Vertebrate lineages like salmonid fish, zebrafish and Xenopus have duplication due to genome-wide duplications. (Davison et al. 2004) Specific example: Mammal Red Vistacha Rat (Davison et al. 2004, Galardo et al. 1999).</td>
<td>OCP and EDU genes were produced by tandem gene duplication from an EDN-like ancestral gene after divergence of CW and NW monkeys but before separation of hominoids and CW monkeys (Zhang et al. 1993).</td>
<td>Three different immunoglobulin chain genes getting duplicated (presumably residing on syntenic chromosome regions. (Davison et al. 2004) and for specific example: human immunoglobulin vs. genes (Zimmer et al 1990).</td>
<td>x - if neutrally evolving y - if under purifying selection</td>
</tr>
</tbody>
</table>

- o - known example
- oo - probable
- x - improbable
Retention of RT-RPs cannot be readily explained by current models

As discussed in the previous sections, it is clear that existing models do not adequately (if at all) account for the retention of retroduplicated genes. Due to the fact that retroduplicates like RP-RTs (discussed in chapter 2) are very abundant, are highly conserved, lack source gene regulatory regions, and because changes in ribosomal gene dosage are strongly selected against, the retention of the RT-RP duplicates is not readily explained using current models of Ohno’s three trajectories of dosage, subfunctionalization, and neofunctionalization. Additionally, we have also shown that in mammalian genomes, RNA duplications comprise half of all conserved duplicated genes. Thus, any set of models that hopes to explain the bulk of duplication events contributing to mammalian genome evolution must account for RNA-based duplicates. All the models that are found in the literature tend to explain the terminal retention of duplicates and the potential consequence of the duplicate once fixed in the population. However, it is very important to describe the forces that could influence the early evolutionary trajectory of a large percentage of all gene duplication events. Based on the above observations, it is clear that there is a need to introduce a new model that can explain conserved retroduplicated gene duplicates that persist in the genome and don’t seem to neofunctionalize, subfunctionalize or be retained for gene dosage, and can become inactivated even after millions of years of purifying selection.
DPI Model Explains the Impact of Dominant-negative Effects on Duplicate Evolutionary Trajectory

One factor not fully explored in most existing models for the retention of duplicated genes is the potential for dominant-negative effects of missense mutations on cellular processes. In an attempt to explain the retention of large numbers of duplicates that we discovered earlier as studied in Chapter 2, we devised a model we call the Duplication, Purification, and Inactivation (DPI) model of duplicate gene fate. This model posits that gene duplicates are retained in the genome not because they contribute to the fitness of the organism, but rather because missense mutations in these genes is not tolerated due to dominant negative effects of missense alleles. Mutant proteins can act in a dominant-negative fashion in a wide variety of ways\textsuperscript{78,109,110}, and this model could account for the strong purifying selective pressure we observe on duplicated genes. The acquisition of dominant-negative mutations in duplicates may represent a threat to the viability of an organism \textit{via} expression alone (see Figure 13 for an illustration). Thus, these gene copies will remain under purifying selection until they are inactivated (pseudogenization or transcriptional silencing) or acquire a fate determining mutation. We suggest that dominant negative phenotypes may exert an immediate and strong purifying selective pressure upon any duplicated gene, with this pressure varying directly with the potential for the gene product to act in a dominant negative fashion\textsuperscript{78} (Figure 13).
Illustration of the DPI model

Figure 13: Illustration of the Duplication, Purification and Inactivation (DPI) Model: A Model for Retention of Duplicate Genes through Purifying Selective Pressure Against Dominant Negative Alleles. The model states that gene duplicates are retained in the genome not because they contribute to the fitness of the organism, but rather because missense mutations in these genes are not tolerated due to dominant negative effects of missense alleles. In the top panel, a wild-type ribosomal protein (WT RP) and its epitopes (orange and blue rectangle) are interacting with the eukaryotic ribosome represented in blue and yellow, with mRNA represented as a single-strand in dark blue. When ribosomal proteins are all present in equal stoichiometric amounts, normal translational activity as well as normal functionality is observed. At this point, the ribosomal protein gene duplicate (RP-DUP) would not interfere if it is constrained by strong selective pressure. In the bottom panel, along with the WT RP, a dominant-negative version of a ribosomal protein gene duplicate (DN RP-DUP) with a missense mutated domain in the epitope (magenta circle) is shown to be interacting with the ribosome. DPI model claims that when the WT RP interacts with the ribosome, normal translational activity and function would be observed, while interaction of the DN RP-DUP with the ribosome results in deficient translational activity leading to absence of ribosome function.
Another aspect of the DPI model is that a dominant-negative mechanism does not require complementation, neofunctionalization, subfunctionalization, ubiquitous expression, or a selective advantage for the new copy. In a recent study of flowering plants, De Smet et al., postulate a very similar idea that dominant negative model constrains genes to be maintained as single copies to avoid non-specific interactions\textsuperscript{104}. Strongly conserved multiprotein complexes like the ribosome are the most commonly observed context for dominant negative phenotypes, but dominant negative phenotypes are not restricted to such multi-protein complexes, in fact, they are widespread\textsuperscript{110}. Because selection against dominant negative alleles acts immediately upon newly duplicated genes, and serves to maintain gene products in a very restricted portion of protein conformational space, it likely facilitates the retention of duplicates by many of the models described above by increasing the half-life of functional alleles in the population and the exploration of the small local region of allowable variation in protein conformation. Liberles and coworkers have proposed similar models in the context of negative pleiotropy\textsuperscript{105,111–113}, and have reached parallel conclusions on the impact of these bottlenecks in sequence space during evolution. To gain support for these models of gene family evolution, it will be important to functionally test the predictions of these models in experimentally tractable systems.

\textit{In silico and in vitro predictions and tests for the DPI Model}

The DPI model for the retention of duplicate genes by dominant negative forces makes a number of explicit, testable in vitro and in silico predictions.
**In silico predictions**

Because of the bioinformatics investments we have made in lab over the past few years, we are in an excellent position to test some very important *in vitro* predictions. We have collections of gene families annotated by copy number and mechanism of duplication. As seen in Chapter 2, we have computationally tested the ribosomal protein genes and their duplicates for selective pressures and signatures of the expression. Now we would like to expand out of ribosomal families and show the universal applicability of the DPI model and test all annotated dominant-negative source genes and their duplicates. We would like to test whether dominant-negative effects could have an effect on the sizes of gene families and we would also like to determine the selective pressures that are observed on these genes and their duplicates to further support the DPI model. Some of the key *in silico* predictions are as follows:

1) Any gene that can act as a dominant-negative (DN) will be part of a relatively smaller gene family than its non-DN counterparts;

2) Purifying selection will be stronger on gene families with known dominant-negative genes; and,

3) Purifying selection will be enforced regardless of the number of copies of a dominant-negative genes in a given gene family.

In order to address these predictions of the DPI model, we have created a pipeline as seen in Figure 14 below.
Figure 14: *In Silico Pipeline to test DPI predictions*. Protein sequence annotations for all gene families were collected using the ENSEMBL62 database. These gene families were annotated as dominant-negative (DN) or non dominant-negative (non DN) families using the annotations from the Online Mendelian Inheritance in Man (OMIM) database. Guilt-by-association approach was utilized to pick out DN gene families based on presence of atleast one annotated DN member. These genes were cross-referenced to Pseudopipe data to capture all the putative gene duplicates. We then utilized our in-house pipeline steps of hierarchical clustering by local synteny in order to build our gene family trees after filtering false-positives and redundant entries. Final gene family analyses were separated in 2 steps: 1) Analysis of DN and non-DN Gene Family sizes, and; 2) calculating the selective pressures on all DN and non-DN gene families using PAML branch-wise method.
**In vitro predictions**

A very unique feature about the DPI model is that some of the key predictions that it makes can be tested directly *in vitro*, a very important aspect that seems to be missing from all the current more theoretical-heavy gene duplication models. We would like to test the gene duplicates whose retention can be explained in the realms of the DPI model a little further using techniques that can help us dissect the potential role of gene duplicates with respect to their source genes. *In vitro* predictions we will test for the DPI model are:

1) There will be no fitness penalty incurred through inactivation of the conserved duplicate gene copy, and

2) Induction of the duplicate gene can rescue the function of the source gene, however missense mutations in the duplicate copy will have a high probability of exerting dominant negative effects.

We can test these predictions for candidate ribosomal protein gene duplicates identified through our previous studies. The phenotypes of known ribosomal mutations consistently include several characters readily assayed in individual cells. We have designed a series of experiments that will allow us to rigorously these gene duplicates using precise genetic modifications of HEK293T cells. For the first prediction, we will employ the use of the CRISPR/Cas9n system to knockout source and duplicate genes respectively (Pipeline A on the left in Figure 15). To test the second prediction, we have designed a complementation-rescue system with which we will induce the expression of the duplicate gene using the PiggyBac cumate induction system and knockout the source genes using the aforementioned CRISPR/Cas9n system (Pipeline B on the right...
in Figure 15). See Materials and Methods section for more information on the protocol steps used in Figure 15.

**CRISPR Knockout and Complementation-Rescue Pipelines to test *In Vitro* predictions of the DPI model**

A) Schematic of the CRISPR-CAS9 pipeline to knockout source and duplicate ribosomal protein genes. B) Schematic of the Complementation Pipeline for Ribosomal Protein Candidates RPS15 and RPS26. See Materials and Methods for more information on the experimental setup and details.
Testing Key *In Silico* Predictions of the DPI Model

Comparative Analyses of DN and Non-DN Gene Family Sizes

The first *in-silico* test for the DPI model involved testing whether gene families that can act in a dominant-negative (DN) fashion will maintain low family member counts compared to non-DN gene families due to the constraints on DN genes to be maintained as single copies to avoid any and all non-specific interactions.

The member count distributions of both DN (n=493) and Non-DN gene families (n=11945) were compared and their log transformed distributions are shown in Figure 16 B and C. We clearly observe that DN-gene families has fewer members in the family compared to non-DN gene families. The moments data for DN and non-DN families is also shown in Figure 16A. We found a statistically significant difference between the two datasets and also found that the distribution of DN gene families lies below that of Non-DN gene families as performing a one sided Kolmogrov-Smirnov (KS) gave us a p-value << 0.01, confirming the smaller gene family size for DN gene families (Figure 16D). D-statistic and p-value for all alternative hypotheses testing from the KS test are also provided in Figure 16D.
Gene Family Sizes

A

Moments data for gene family members

1. DN Gene Families
   • Mean: 0.803
   • Variance: 1.093
   • Skewness: 1.767
   • Kurtosis: 3.328

2. Non-DN Gene Families
   • Mean: 2.028
   • Variance: 4.380
   • Skewness: 10.311
   • Kurtosis: 27.189

B

Distribution of DN Gene Family Members

C

Distribution of Non-DN Gene Family Members

D

Kolmogrov-Smirnov Test

1. Alternative hypothesis: Two Sided KS test:
   • D = 0.39008, p-value < 2.2e-16

2. Alternative hypothesis: the CDF of x (DN Gene families) lies below that of y (Non-DN gene families)
   • D^+ = 0.39008, p-value < 2.2e-16

3. Alternative hypothesis: the CDF of x (DN Gene families) lies above that of y (Non-DN gene families)

Figure 16: DN and non-DN gene family analyses. Gene Families that can act in a dominant-negative (DN) fashion have smaller family sizes than non-DN gene families. A) In the top panel, moments data is shown for DN and non-DN families. B and C) In the middle panel, DN and non-DN families are shown on X-axis respectively and log-transformed values for gene members in each family on Y-axis. D) Results from one and two-sided non-parametric Kolmogrov-Smirnov test for DN and non-DN families are shown.

Selective Pressure Analyses of DN and Non-DN Gene Families

Since we observe relatively smaller family sizes for genes that are associated with DN traits, we now wanted to take a closer look at the selective pressure constraints
acting on DN and non-DN genes. For our second *in silico* test of the DPI model, we calculated Ka/Ks ratios for all gene families in DN (n=472) and Non-DN gene family datasets (n=9100) using CODEML program in PAML (see methods section) (Figure 17A). For the DN gene families, we observe a mean dN/dS value of 0.12 (95% CI 0.0128, 0.016) and 0.18 (95% CI 0.018, 0.019) for Non-DN gene families. We also observed that there was significant difference between these two distributions (bootstrapped p-values <0.01 and n=1000). To test whether this selective constraint is an artifact of family member size, we binned distributions of DN gene families based on their member sizes (Figure 17B). We statistically confirm that member sizes do not influence change in Ka/Ks ratios (post-hoc HSD test pair-wise pvalues <0.01).

**Mean selective pressures**

![Box and Whisker plots depicting tree-wise Ka/Ks ratios calculated for DN and non-DN families using the codeml program in PAML. B) Box and whisker plots show selective pressures calculated for all DN gene families binned based on the number of members in each gene family.](image)

Figure 17: **Selective Pressures on DN and non-DN gene families.** A) Box and Whisker plots depicting tree-wise Ka/Ks ratios calculated for DN and non-DN families using the codeml program in PAML. B) Box and whisker plots show selective pressures calculated for all DN gene families binned based on the number of members in each gene family.
Testing Key In Vitro Predictions of the DPI Model

Ribosomal protein gene candidates and CRISPR-Cas9n sgRNA designs

To find appropriate ribosomal protein gene candidates to test the DPI model, we examined the source and duplicate gene dataset we had generated previously\(^6\). By mining our duplicate gene list and evaluating the selective pressure and expression data, we selected RPS15 and RPS26 source genes (as annotated by ENSEMBL 62) and their respective retrotransposed gene duplicates as candidates. RPS15 gene duplicate is located on the negative strand of human chromosome 2:172373783-172374226 and RPS26 gene duplicate is located on the negative strand of human chromosome 4: 114135205-114135549. We based our selection of the candidate gene selections based on their expression, integral function in the ribosome and the possibility of designing unique CRISPR gRNA sequences. Structural locations in the 80S ribosome and their amino acid sequence alignments for source and duplicate RPS15 and RPS26 can be observed in Figure 18. Rps15 and RPS26 source sequences from here on will always be depicted in light green and light purple respectively and their duplicate genes in dark green and dark purple respectively.
Figure 18: **Ribosomal protein gene candidates to test the DPI model**  

A) Schematic showing the locations of human ribosomal protein genes RPS15 and RPS26 in the global map of the 80S human ribosome. This is a view looking down the 80S ribosome from the 40S head/60S central protuberance. For the 60S subunit, protein is shown in dark blue and rRNA in slate blue. For the 40S subunit, protein is shown in yellow and rRNA in pale yellow. The black arrow marks the path of the mRNA and the grey structure represents the tRNA. RPS15 and RPS26 source genes are shown in Purple and green respectively.  

B) Protein sequence alignment of RPS15 (top panel) and RPS26 (bottom panel) source and duplicate amino acid sequences. The differences are highlighted in yellow. RPS15 source gene sequence is represented in light green and the RPS15 duplicate gene in dark green. RPS26 source gene sequence is represented in light purple and the RPS15 duplicate gene in dark purple.
Both the proteins are found in the 40S subunit and are located on the surface of the ribosome near the A-site and have been implicated in the mRNA binding. It is involved in the 40S maturation and seems to be a determinant of pre-ribosomal export from the nucleus (See Figure 18A). For our candidates, we wanted to make sure that the EST tags that were associated with them were reliable as for the DPI model to play a role, it is important that the gene duplicates are expressed in the cells of interest. Hence, we used qRT-PCR to validate expression of the source and duplicate genes and confirmed their transcript-level expression in our experimental cell line HEK293T as illustrated in Figure 19.

**Figure 19: qRT-PCR gene expression testing for ribosomal source and duplicate protein gene candidates.** Representative genes showing gene expression in HEK293T cells. Messenger RNA levels for GAPDH, RPS15 source and duplicate gene and RPS26 source and duplicate gene are shown in the graph. Expression values for all probesets were log2 transformed. Housekeeping gene GAPDH is represented in grey, RPS15 source gene in light green, RPS15 duplicate gene in dark green and RPS26 source gene sequence in light purple and RPS15 duplicate gene in dark purple respectively.
Efficient CRISPR-Cas9 mediated gene targeted knockouts of source and duplicate genes of human RPS15 and RPS26 ribosomal proteins.

While previous studies have shown siRNA knockdowns of RPS15 and RPS26 source genes lead to defects in 40S production, ribosome biogenesis defects and reduced levels of 18S and 18E rRNA106,114–116, we wanted to confirm the phenotypic effects on mammalian cells. To this end, we employed the widely used CRISPR-Cas9 system to precisely knock-out either the parental or duplicate ribosomal genes and assess the impact on cellular fitness. Due to the high similarity between the parental and duplicate genes we used the double-nickase CRISPR system which relies on two specific targeting events to delete the region of interest117. Using this system we avoid off-target effects and ensure targeting of the correct gene copy. In this system, a pair of plasmids each encoding a Cas9 (D10A) nickase mutant (Cas9n) are directed to distinct, genomic loci by a target-specific guide RNA. Each Cas9n/sgRNA complex creates one nick in the DNA strand that is complementary to the guide RNA117. The double nick created by the pair of Cas9n/sgRNA complexes mimics a DSB and results in end resection and non-homologous end joining of the DNA fragments. Thus, the use of paired-guide RNAs allows for increased specificity of Cas9-mediated gene editing, while maintaining a high level of efficiency117,118.

For RPS15 and RPS26 source genes, we used a pair of CRISPR guide RNAs that targeted the exon1 and intron2 junction, hence avoiding any off-targeting in the intronless retroduplicate gene candidates. For the duplicate gene candidates, we targeted the 5’ UTR region and exon 1 regions as the UTRs were unique for the duplicate genes and hence off-targeting of source genes could be avoided. The
advantage in doing so was that the region of the gene gets knocked out only if both the guides are identified by the Cas9n mutant protein and hence helping us avoid any false-positive knockouts. RPS15 and RPS26 source and duplicate CRISPR gRNA targeting regions and guide design strategy can be seen in Figure 20 and all the guides used for the knockouts can be seen in Appendix 2. The stepwise experimental strategy was previously presented in Figure 16A.

**RPS15 Source Gene sgRNA Design**

5’ GGACCAAGCGATCTTCTCTGAGATCCGGGCAAGATGTTGAGTGTGCTGATT TGGCGC 3’

3’ CCTGTTTGGTAGAGAACACTCTTA GGCCGTCTACACTCACAACGCCTAACCCGG 5’

**RPS15 Duplicate Gene sgRNA Design**

5’ CCAGTATGCTCTCTCTGGAGATCAGGGAAGGCAGAGAAGACGAAGAAGCGG 3’

3’ GGGTATACGACGAGAAAGACTCTTGGCCGTCTACCGCTCTCTCTCGCTCTTCCACCCG 5’

**RPS26 Source Gene sgRNA Design**

5’ TCTGCTCTCTCCGCTGCCCTCCAAGATGCTGGAGTCTTCTCTTGGTGGAGGGTG 3’

3’ AGAGGAGAGAGCCAGGCAAGGAGGTTTCTACCAGCAAGAAGACGCACCACCTCCAC 5’

**RPS26 Duplicate Gene sgRNA Design**

5’ GGTCCGTGCTCTCCAAGATGCAAAAGAAAAGGAAGAACAATGGTCCTGCGGAAAAGG 3’

3’ CCAAGGAGAGAGAGGTCTACTCTTTTCTTTTCTTTGGTTACCAGACCGGTTTTTCC 5’

Figure 20: **Representative sequences of RPS15 and RPS26 source and duplicate locus targeted by the Cas9n for CRISPR knockouts.** RPS15 and RPS26 source and duplicate gene sgRNA target sites are indicated in blue and Protospacer adjacent motif (PAMs) are underlined in red respectively. Exon-Intron junctions were targeted for source genes and UTR-Exon region were targeted for duplicate genes. See Appendix 2 (Table C) for all the sgRNA sequences that were used. Colored bars on the left depict the candidate for which the sgRNA designs were made.
CRISPR/Cas9 Gene Knockouts of ribosomal protein source genes and their gene duplicates

**Cell Line Selection**

The DPI model predicts that many stabilized RT-duplicates will be dispensable for the organism, but that when the source genes are knocked-out, it will exert dominant negative effects on the cell. The shared phenotypes of ribosomal mutations include reduced protein synthesis, reduced proliferation rate, impaired ability to compete with wild-type cells and lethality, all of which can be readily assayed in tissue culture. Ribosomal protein synthesis is universally required for all cells, but the most stringent requirement occurs in rapidly dividing cells, such as HEK293T cells. HEK293T cells are rapidly dividing, amenable to transfections and highly efficient and faithful in translation and processing of proteins that can be readily cultured and genetically modified. For these reasons, they represent an excellent system in which to test the DPI model. Hence HEK293T cells were used to perform source and duplicate gene knockouts using the CRISPR sgRNA pairs (see all gRNA sequences in Appendix 2: Table C).

**CRISPR-Cas9 knockouts of gene candidates**

In order to confirm the cellular growth defects that knocking out a ribosomal protein source gene would cause and to determine the effects that would incur due to the knockout of ribosomal protein duplicate genes, we used the double-nickase CRISPR-Cas9 system protocol as previously mentioned\textsuperscript{117,118}. After designing and cloning appropriate gRNA sequences into the CRISPR plasmid, these plasmids were transfected into HEK293T cells and kept under antibiotic selection. CRISPR knockouts
of the source ribosomal protein genes RPS15 and RPS26 causes HEK293T cells to crenate and lose adherence suggesting cell death and reduced cell viability in comparison to Empty vector HEK293T cells (Figure 21). Although some cell death was observed at 24 hours post-transfection, it was most evident at 72 hours-post transfection.

**CRISPR/Cas9 experiments show that source gene knockouts affect cell viability, but duplicate gene knockouts do not**

**CV absorbance at 550 nm RP source and duplicate gene knockouts**

![CV absorbance graph]

Figure 21: **CRISPR Cas9 aided knockouts of Source RPS15 and RPS26 genes exhibit drastic phenotypic defects while knockouts of duplicate genes do not affect the cell phenotype.** A) Bar graph depicting crystal violet absorbance values for wildtype HEK293T cells along with HEK293T cells that were transfected with CRISPR sgRNA guides to knockout RPS15 and RPS26 riboprotein source and duplicate genes. RPS15 source gene is represented in light green bar and RPS15 duplicate gene in dark green. RPS26 source gene sequence is represented in light purple and the RPS15 duplicate gene in dark purple. Values are means of ±SD (n=3). ** P-value < 0.01. B) Representative microscopy images of the targeted cells are depicted below the bar graphs. See Materials and Methods for more information on cell culture, CRISPR designs and transfection methods.
CRISPR knockouts of RPS15 and RPS26 duplicate ribosomal protein genes showed no signs of crenation or loss of adherence and the cell viability and phenotype was comparable to empty vector and wild-type HEK293T cells. These source and duplicate gene knockouts results were further confirmed with photomicrograph images at day 3 (Figure 21 and Appendix 2: Figure D) along with quantitative measurements of HEK293T cell growth using the crystal violet staining to measure the absorbance of cell lysis products as an indirect measurement of cell growth and viability (Figure 21).

We also confirmed that the CRISPR deletions we intended were being accurately created in the cells by using the Surveyor Nuclease assays that cleaves DNA with high specificity at sites with base-substitution mismatches (data not shown). Additionally, we also clonally isolated the source and duplicate gene knockout cells that had survived after the knockouts were pursued, by isolating single cells through Fluorescent-activated cell sorting (FACS). Once the single cells expanded into colonies, we extracted DNA, performed a round of PCR to amplify the source and duplicate gene sequence respectively and then pursued DNA sequencing analyses to analyze the candidate DNA sequences of the cells. On performing sequence alignments of the source genes (see Appendix 2: Figure B), we show that the cells surviving selection did not have alterations in their coding sequences, and thus were escapers that were not edited or only received one of the CRISPR gRNA pairs (potentially only got one of the gRNA sequences out of the two). On performing sequence alignments for duplicate genes, we show that the cells were transfected and were in fact edited accurately, but have survived the duplicate gene knockouts with both our sgRNAs. This means that their knockout did not affect cell viability in this case.
Interestingly, we found all the above data to be concurrent with a major study pursued at the Max Planck Institute (Dresden, Germany) where similar results were observed for a much larger set of source and duplicate genes, albeit the data was collected through knockdowns, instead of knockout studies. In this study 34 ribosomal retrogens (highly conserved old, new, intact and pseudogenized candidates) were knocked-down with no detectable phenotypic defects. However, knock-down of each of 70 parental RP genes had drastic phenotypic defects on the cells, with no evidence of retrogens compensating for the loss of parental gene products (data obtained and analyzed from\textsuperscript{106,107}). So in summation, Figure 21 shows that on knocking out source ribosomal protein genes, we observe drastic phenotypic defects on cells and cell numbers dwindle down at a very fast rate. No such effects were observed when the duplicate genes were knocked out. This could mean that duplicate genes, with their endogenous level of expression, cannot rescue the function of the source genes, when knocked out are dispensable for the organism when knocked out.

**Complementation Rescue of the source RP gene function by its gene duplicate**

Having shown that duplicate gene knockouts do not exhibit the growth and viability defects observed for parental gene knockout, we assessed whether over-expression (induction) of the duplicate gene could compensate for the loss of parental gene function. As the dominant negative mechanism of the DPI model requires that duplicate genes compete for function in the ribosome, the duplicate gene should be able to at least partially compensate for loss of the parental gene. The severe defects seen in parental gene knockouts indicate that the endogenous levels of duplicate genes are
not sufficient for compensation. Hence, we overexpress (Induce) the duplicate genes in parental knockouts to assess their ability to rescue the source function in the ribosome.

The genetic complementation-rescue system we have developed leverages a dual hygromycin-puromycin antibiotic selection method (See Materials and Methods for more details) to select for cells containing both the CRISPR-Cas9 knockout (puromycin) as well as a cumate-inducible piggyback vector (hygromycin) used to overexpress the gene of interest. The experiment is designed in to address whether the induction of the duplicate gene in cells that will undergo a CRISPR knockout of the respective source gene, rescue the source gene depletion and hence keep the cells viable avoiding crenation and cell death. The pipeline for this experimental setup was shown previously in Figure 16B. We also induced the expression of the source gene coding sequences in HEK293T cells that were also subject to source gene knockouts as a positive control.

Figure 22 shows us that when we induce the duplicate gene’s expression in cells that consequently have the source genes knocked out, there is definitely a rescue that can be observed (see absorbance graphs in 22A and C and cell images in 22 B and D) as significantly lower amount of cell crenation and death can be observed when compared to source gene knockouts with no rescue (see Figure 21). However it is important to note that the rate of rescue is slightly lesser compared to the positive control. This indicates that while the duplicate gene is able to rescue the function of the source gene at some level, but even the conservative amino acid differences between the source and the duplicate gene is not allowing a full complementation like the positive control.
Figure 22: **Induced expression of the duplicate gene can rescue source gene function, but at a fitness cost arising due to missense mutations.** A) Bar graph depicting crystal violet absorbance values for HEK293T cells that were induced with source and duplicate cDNA respectively, followed by CRISPR/Cas9 transfections to knockout riboprotein source gene. Induced RPS15 source gene with source gene knockout (positive control) is represented in light green bar with white diagonal lines followed by non-induced RPS15 source gene with source gene knockout in light green. Induced RPS15 duplicate gene with source gene knockout is show in the next bar colored in dark green with white diagonal lines followed by non-induced RPS15 duplicate gene with source gene knockout in dark green. Exact same representation is shown for RPS26 source and duplicate gene in C. Values are means of ±SD (n=3). ** P-value < 0.01. B and D) Representative microscopy images of the targeted cells are depicted below the bar graphs for RPS15 in B and RPS26 in D. See Materials and Methods for more information on cell culture, CRISPR designs and transfection methods.
The protein structures depicting the conservative amino acid changes are shown in Figure 23. As observed in Figure 23, the amino acid substitutions (shown in red) for both RPS15 and RPS26 occur in exposed regions and do not seem to interact directly with other ribosomal proteins, while the amino acids do not change at all for regions that seem to be directly interacting with other ribosomal units. Additionally, the insertion of 3 AA for the RPS15 duplicate protein (Figure 18B) is seen to be occurring in

Figure 23: **RPS15 and RPS26 duplicate protein missense mutation illustrations.** A) RPS15 duplicate protein shown in green with the insertion of 3 AA pointed out in red. The right panel depicts the position of the mutational region with respect to the ribosome. B) RPS26 duplicate protein shown in purple with the substitution of 3 AA at different locations pointed out with red balls. The right panel depicts the position of the mutational region with respect to the ribosome.
an exposed loop that does not interact with any other unit of the ribosome. Similarly, substitutions in the RPS26 duplicate protein (seen in Figure 18B) are also extremely conservative in terms of amino acid size, charge and hydrophobicity. All these factors suggest that the rescue of the source phenotype occurs due to the highly conserved nature of the duplicate protein genes. However, the incomplete nature of this rescue and the partial cell viability defects are observed due to the few missense mutations (although conservative) that have accrued in the duplicate genes. In cases where the amino acid substitutions results in drastic changes in the duplicate protein, the DPI model would posit that such a protein would be eliminated from the population due to its potential dominant-negative nature. This might be one of the reasons why we only recover gene duplicates that are highly conserved in terms of its sequence with respect to its source gene.

The cells which were not induced by the duplicate genes and had the source gene knocked out survive presumably due to the same reasons as the knockout experiments discussed previously, as they only received one of the CRISPR gRNA pairs (potentially only got one of the gRNA sequences out of the two) and thus were not edited correctly. We confirmed cell viability using calcein-red AM for live cell staining. These images (Figure G) along with other replicates of the RPS15 and RPS26 complementation experiments (Figure E and F) are included in Appendix 2.
Discussion:

Based on our studies described in Chapter 2 and 3, we discovered thousands of retroduplicates in the mammalian genome whose retention cannot be easily explained by the current models described in the literature. We believe that the gene duplication model that we present in this chapter, not only explains the retention of several previously unexplained gene duplicates, but also acts as a precursor to all the major gene duplication models out there. The DPI models posits that gene duplicates are retained in the genome not because they contribute to the fitness of the organism, but rather because missense mutations in these genes are not tolerated due to dominant negative effects of missense alleles. To solidify the applicability of the model, we test several predictions that can be made for the model. The unique aspect of testing DPI model is that unlike other gene duplication models, that are almost exclusively theoretical, we designed in-vitro experiments to prove the model along with the in-silico testing.

We first wanted to test whether dominant-negative effects play a role in limiting the number of gene duplicates in a family. We confirmed that notion as we observe that DN gene families are relatively smaller in size than non-DN gene families. According to the DPI model, this occurrence is potentially an attempt by a gene family to reduce its chances of ending up with fewer members that are subject to dominant-negative forces. We also show that DN gene families are under stronger selective pressures compared to non-DN gene families, further exhibiting the notion that gene duplicates are avoiding all non-synonymous substitutions, which in turn extend their half-lives and their chances of being retained in the genome, but not necessarily from a functional standpoint.
Furthermore, we test the DPI Model by \textit{in vitro} experimentation specifically testing source genes and their gene duplicates using the CRISPR-Cas9 gene knockout system. We pursued knockout experiments where we independently target the source gene and duplicate genes to show that that knocking out the source gene impair the functions of cells drastically and that the loss of a conserved duplicate has little to no phenotypic consequence. Next, we analyzed whether the gene duplicate’s induction in cells where the source gene was knocked out, could potentially rescue the function of the cells. Our results showed that the duplicate gene did recover the phenotype of the source genes, but it comes with a fitness cost incurred due to the rare missense mutations present in the duplicate genes. However, the amount of phenotypic recovery was due to the fact that the duplicate gene candidates had conservative mutations, as the DPI model would predict little to no recovery if the missense mutations incurring were drastic. Perhaps, the gene duplicates with the drastic phenotypes have already been eliminated from the population according to the realms of the DPI model.

Through a series of \textit{in silico} and \textit{in vitro} tests, we successfully explain the retention of thousands of previously unexplained gene duplicates that are seen to be held under strong selective pressure and were not explained by any of the existing gene duplication models. The obvious next challenge is to apply the principles of population genetics in order to integrate the DPI model with current models of duplicate gene retention. We attempt to begin this modeling using the foundation laid by Innan and Kondrashov in their excellent review of gene duplication that focuses our attention on the key phases that lead to the stable preservation of a duplicated gene, namely, origin through mutation, a fixation phase when it segregates in the population and a
preservation phase when the fixed change is maintained. Another factor to be exclusively considered for gene duplication is that it is important to study each duplication event from the moment of its emergence.

Figure 24: Population Genetics representation of the DPI (Duplication, Purification and Inactivation) model. Two cases have been represented in this figure where the dominant negative effect is considered to be applicable 100% of the time in the top panel and 0% in the bottom panel. In the pre-duplication phase, the single-copy genotype (A) is fixed in the population; when a duplicate arises, the fixation phase begins. The duplicate is most likely to be lost to drift but can also achieve fixation. After the duplicated genotype (A–A) is fixed, the fate-determination phase begins and continues until the fixation of a fate-determining mutation. (S) represents synonymous mutations while (N) represents non-synonymous mutations. dN/dS ratio trends are represented by the panel in green. DPI model explains how you can retain a duplicate for a longer amount of time where it stays in the fate-determination phase for a very long time and is under strong purifying selective pressure while doing so.
In Figure 24, a preliminary population genetics representation of DPI model shows that a duplicate gene, while accepting synonymous substitutions at a neutral rate, does not allow non-synonymous substitutions to accrue in its attempts to avoid potential dominant-negative mutational effects. This apprehension in acquiring non-synonymous mutations allows the duplicate gene to be retained in the population for a comparatively longer time while maintaining the strong purifying selective pressure, before acquiring a fate determining mutation, thus increasing their probability of achieving any of the eventual fates, pseudogenization, gene conservation, subfunctionalization or neofunctionalization.

So in conclusion, the DPI model successfully explains the retention of thousands of highly conserved gene duplicates that might not be functionally important to the organism, however, missense mutations in these gene duplicates can exert severe dominant-negative effects on the source gene. The DPI model addresses the missing evolutionary gaps in the field of gene duplication, as it focuses on the immediate fate of the gene duplicate and acts as a precursor to all the established gene duplicate models that deal with a terminal fate of these duplicates.
Chapter 5 Conclusions and Future Directions

Final Synthesis and Conclusions

Since we began studying gene duplication dynamics in our lab, the understanding of mechanisms of gene duplication has always remained uncertain till date. This is due to several reasons like changes in gene copy number in genomes, difficulty to measure formation rates and the facts that mechanism vary with positions in the genome. Emphasis have always been placed on extensive DNA-mediated duplication studies, while RNA-mediated duplications have mainly been ignored due to connotations of these duplicates being evolutionary dead ends and representing mainly the junk parts of the genome. However, four of our lab publications as well as several other works over the years, have shown that perhaps we have ignored studying a major duplication mechanism in gene duplication. We began this project with an intensive study of gene duplicates in 5 different mammalian species and discovered that conserved retroduplicates are widespread in mammals, representing half of all gene duplicates under purifying selective pressure and that ribosomal protein genes constitute one of the largest classes of conserved retroduplicated genes.

We decided to investigate these findings further with an expanded set of mammals and we conducted a large study in which we annotated every single gene duplicate of 79 ribosomal protein genes and annotated them with associated selective pressures and expression data. In this study, we found that ribosomal proteins gave rise to thousands of intact retroduplicates that are strongly conserved and exhibit signatures of expression. It was interesting and unexpected to find thousands of highly
conserved gene duplicates originating from slowly-evolving, highly conserved ribosomal proteins that could potentially face severe stoichiometry issues due to the imbalance in core RP ratios (Figures 3, 6 and 7). We also discovered that none of the existing gene duplication models could explain the retention of these gene duplicates (Figure 12) and hence we devised the DPI model to explain retention of thousands of these unexplained duplicates. The model posits that intact gene duplicates that have are strong selective pressure might not necessarily contribute to the fitness of the organism, however, missense mutations in these gene duplicates can exert dominant-negative effects on the source gene accruing phenotypic defects in cells (See Figure 13 for illustration). We designed a specific set of predictions for the DPI model and tested them using *in silico* and *in vitro* experiments. This was a very unique point about the DPI model as it is one of the first gene duplication models that was experimentally tested in lab and not exclusively based on theoretical predictions. Through our *in silico* tests we show that DN gene families limit themselves to smaller sizes to avoid any detriment due to missense mutations and that effect was confirmed as DN gene family members were under stronger selective pressures compared to non-DN gene family members (Figures 16 and 17). For our *in vitro* tests, by designing precise gene knockouts via the CRISPR-Cas9 system, we successfully proved the first aspect of the model that gene duplicates could be knocked out of cells without any detriment to the fitness, quite contrary to a source gene knockout, that showed severe impairment in cells (Figure 21). So we concluded that gene duplicates with its endogenous expression profile could not rescue the phenotype of its source genes. To study these highly conserved intact gene duplicates a little further, we designed a
complementation rescue experiment where we induced expression of gene duplicates and consequently knocked out the source genes. We concluded that on induction of gene duplicate’s expression, we successfully recover the phenotype of the source genes and that the duplicate gene was able to keep the cells viable even in absence of the source copy, but at a slightly lesser extent due to the fitness cost accrued due to the missense mutations (Figure 22). Interestingly, the missense changes we discover are very conservative changes in terms of size, charge, hydrophobicity as well as from a structural perspective (Figure 23). The DPI model predicts this outcome as allowance of major missense changes would result in a near-lethal phenotype due to the dominant-negative effects incurred and hence an observable defect is observed as the mutational changes are not as drastic. So in conclusion, we have developed a gene duplication model that explains retention of previously unexplained gene duplicates that have much longer half-lives in the population (see Figure 24 for illustration).

The DPI model works as a precursor model to all the other models as it focuses on the immediate fate of a gene after it duplicates, while the latter focus on terminal fates of the duplicates. Together, these results provide a comprehensive history of ribosomal protein evolution in mammals, comprise a body of evidence that meets or exceeds that available for any other model of duplicate retention, and establish the impact of forces that could influence the fate of every gene duplication event.
Future Studies

The central goal of this thesis was to present and test a gene duplication model that can explain gene duplicates that do not seem to fit any of the current models that exist. We pursued several *in silico* studies and *in vitro* knockout and rescue studies with the source ribosomal protein genes and their gene duplicates to definitively test the DPI model for gene duplication. However, there are several future studies that need to be pursued to show the applicability and further exhibit the functionality of the model. Future studies on the population genetics front to integrate DPI model with existing models as well as wet-lab experiments to further explore the retention of thousands of gene duplicates, are critical for the field of evolutionary genetics.

On the basis of our experimental results, we see that ribosomal protein gene duplicates provide partial compensation of the function of source genes. This leads us to the hypothesis that while the retroduplicate ribosomal protein is getting incorporated in the ribosome but cannot fully function like the source protein potentially due to the cost incurred due to the rare missense mutations in the retroduplicate proteins. So as the next step to understanding ribosomal biogenesis and solidify the basis for all gene duplicates following the DPI model, we need to confirm whether ribosomal retroduplicate proteins get incorporated in the ribosome. The strategy would be to clone the duplicate ribosomal gene cDNA into a pFLAG-CMV5a vector (Sigma Aldrich) and transfecting this plasmid into HEK293T cells. This vector encodes a flag-tag onto the duplicate protein that will allow us to distinguish it from the source copy. Next step would be to use a protocol similar to the ribosomal incorporation assay published by Das et al. (2013) in order to collect several fractions using the upward displacement
method\textsuperscript{119}. Light RNP fractions, 40S, 60S, and 80S, and heavy polyribosome fractions can be collected by monitoring through the continuous UV absorption profile at $A_{254}$. Total protein from each fraction can be precipitated using trichloroacetic acid (TCA). The TCA precipitates were subjected to SDS-PAGE followed by immunoblot analysis with anti-Flag antibody (Sigma-Aldrich, St. Louis, MO). Co-sedimentation of source and recombinant duplicate (detected by immunoblotting) with ribosomal fractions can then be used as an indication of ribosome incorporation of the protein.

After confirming that the duplicate protein gets incorporated in the ribosome, it would be very interesting to test whether the incomplete rescue that we observe in our complementation-rescue experimentation is only due to the defects in protein synthesis. Hence, we would measure protein synthesis directly in the source and duplicate genes. We would test for protein synthesis in WT HEK293T cells and in HEK293T cells in which we induce the duplicate gene and knockout the source gene. If we can label newly synthesized proteins of source and duplicate genes and study them spatially, it will help us understand whether knocking out these genes actually hastens the protein degradation process. Using a system such as the Click-iT Plus technology (ThermoFisher) provide us with a way to label nascent proteins with fluorescent tags that can help us track synthesis and degradation of proteins. This will provide us with a strong quantitative measurement of protein synthesis along with our robust gene-level studies.

The ribosome incorporation and protein synthesis assays will give us a clear idea whether there are any other factors involved in the incomplete rescue of the source function by the retroduplicate gene. At this point, we can test another aspect of
the DPI model to see whether missense mutations that incurred in the duplicate genes are the primary reason for the reduced rescue of the source gene function. We can pursue this by creating a more sensitized experiment by designing difference missense versions of our source genes and duplicate genes (varying the severity of missense changes) and pursue the same set of experiments as Chapter 4 and the ones stated above. We have seen rescue of the source gene phenotype partially by the duplicate genes that have conservative missense mutations, as missense mutants that have non-conservative amino acid mutations are potentially lethal according to the DPI model due to the severe dominant-negative effects in action. By creating these different missense versions of retroduplicate genes, we can further gain a more thorough understanding of their potential dominant-negative effects on the source genes. Additionally, it will give us insight into determining the role of specific protein residues. A very simple and efficient way to probe protein structure and function in such a manner is to use a technique such as alanine-stretch screening mutagenesis. We can use an approach similar to Lefevre et al. (1997), where they are able to rapidly scan a whole protein sequence in search of secondary structures or to characterize the functional role of a stretch of residues\textsuperscript{120}.

There are several \textit{in silico} studies that can be pursued to enhance the impact and applicability of the DPI model as moving forward it will be important to experimentally verify the DPI model outside of the ribosomal protein genes. Testing the DPI model in a larger set will help us extend its application to nearly every single gene duplication event that occurs in a genome. The DN calls based on experimental data found in the OMIM database serve as ideal candidates for testing with
experiments like those described in Chapter 4. The first step would be to run the DN candidates through our tree reconstruction pipeline as described in Chapter 2. After running the pipeline, we will have all possible duplicate genes associated with our DN member, their genomic locations, and selective pressures acting on them, as well as EST tags signifying signatures of expression. Source gene and pertaining duplicates can then be tested by in silico experiments described in Chapter 4 and previously in this section. Example DN genes that we can use for testing our DPI model include the Fibrillin 1 (FBN1), SRY-BOX 10 (SOX10) and Ikaros Family Zinc Finger 1 (IKZF1) genes to name a few.

Analyses of all the ribosomal genes in Chapter 2 and other DN gene families discussed above are providing us with an invaluable dataset. Not only can we test the intact duplicate genes in vitro, we have access a large set of pseudogene data. Pseudogenes have long been labeled as “junk” DNA, failed copies of genes that arise during the evolution of genomes. However, recent results are challenging this moniker; indeed, some pseudogenes appear to harbor the potential to regulate their protein-coding cousins\textsuperscript{121–123}. We have a big opportunity to learn more about gene duplicate evolution as we can closely analyze pseudogenes, their sequences and the selective pressures that have acted on them. The DPI model predicts that pseudogenes that belong to dominant-negative gene families would tolerate nonsense mutations at a higher than expected frequency, compared to missense mutations. This hypothesis (if correct) should bias the expected ratio of these mutations in duplicated genes within a DN gene family and we should observe that the ratio of nonsense to missense mutations is higher than expected under neutral
sequence evolution. For this prediction, we can calculate frequencies of missense mutations and non-sense mutations in both DN and non-DN pseudogenes by using a modified version of a conventional Nei-Gojobori\textsuperscript{124} implementation.

The final challenge will be to build on our preliminary population genetics model (shown in Figure 24) and develop a more robust population genetic model to explain the retention of all previously unexplained gene duplicates in mammalian genomes. We would like to model this gene duplicates and probability of achieving all the fates, namely pseudogenization, gene conservation, subfunctionalization and neofunctionalization, using various iterations and simulations of population size, mutation rate and selection. The main goal is to precisely set up simulations to test the predicted half-life of gene duplicates under different regimes of purifying selective pressure. As shown in Figure 24, we have begun establishing a population genetics model, however a more in-depth study is required. We can begin by setting up the basic assumptions of the model and enhancing its applicability by following the principle of the Hughes (1993) gene duplication model\textsuperscript{125}. This model emphasizes purifying selection in the early stages of duplicates genes, motivated by several observations that are not fully consistent with the strict model of Ohno. It is important to study this model as some of these inconsistencies match up with the predictions of the DPI model, especially the study in which they found evidence for purifying selection against amino acid changes in both copies of most of the duplicated genes created in the tetraploidization event of the \textit{Xenopus laevis} frog species\textsuperscript{125}. 

78
Chapter 6 Materials and Methods

Ribosomal Dataset

Seventy-six ribosomal protein (RP) sequences from nine species [human, chimp, monkey, mouse, rat, dog, cow, opossum, and chicken (outgroup)] were manually collected from Ensembl 62 [51]. Three RPs were excluded due to annotation issues. When a single gene encoded multiple transcripts, the longest was used. These protein sequences served as seed sequences, or input, to the pipeline (Figure 1).

Extraction of Gene Family Members

RP seed sequences were submitted to tBLASTn against donor genomes to capture as many putative duplicates of the seed gene as possible. Each resulting putative duplicate was processed using Pseudopipe [52] to determine the mechanism of duplication (DNA- or RNA-mediated) and the fate of the duplicate (intact or pseudogene). The default Pseudopipe filters for tBLASTn hits (E-value cutoff ≤ 10^{-4} and identity and identity ≥ 40%) were used to define putative duplicates. Ambiguous duplicates, where the duplication mechanism was not confirmed, were resolved using an intron comparison algorithm [18], which compares intron/exon structure within a group while accounting for exon fusions and large insertions in exonic regions. These methods generated a set of RP superfamilies that consist of both protein-coding genes and related pseudogenes.
Identification of Duplications and Phylogenetic Analysis

Orthologous and paralogous relationships were determined using local synteny and a hierarchical clustering algorithm described in Jun et al, 2008 [18,19]. A local synteny score was assigned to all gene pairs based on the homology of genes (three upstream and three downstream) neighboring the two query genes. Pairwise synteny measures were obtained for all members of a gene family. The output generated based on these scores was used to construct phylogenetic trees in Newick format, representing the history of duplication in each family. Parsimony [53] was used to assign each inferred duplication event to a specific branch of the species tree [54,55]. ‘Tube’-style phylogenetic trees for 74 mammalian RP genes were used to illustrate the history of DNA/RNA-mediated duplications across various evolutionary time periods (ancient vs. lineage specific) (See Appendix 1 for all trees).

Conservation and EST Analyses

Using exon-based reconstruction and the Nei Gojobori method, Ka/Ks ratios for all members of a gene family were calculated against the seed proteins. The putative exon-intron structures of duplicates were generated with an in-house algorithm, using these seed proteins. Results were then filtered based on p-values (< 0.1) and the fraction of the source gene represented by each duplicate (> 65%). Pairwise distances using ClustalW were also calculated as an added metric to evaluate sequence identity and account for all nucleotide level substitutions. Additionally, we also determined branch-wise omega values for 28 ribosomal protein families with following parameters,
model=2 & Nsite=0, using codeML in PAML 4.7 [56]. In order to confirm the selective
pressures, standard codon models M0, M1a, M2a were fitted to the data set with
codeML. We used likelihood ratio tests (LRT) to determine the relative fit of the
hierarchically nested models. Log likelihood ratio test statistic is $2\Delta\ell = 2(\ell_1 - \ell_0)$, where
$\ell_1$ is the log-likelihood of the model corresponding to the alternative hypothesis and $\ell_0$
represents the log-likelihood corresponding to the model used as null hypothesis. These
values were compared with a chi-squared distribution in which the difference between
the number of parameters of both models provides the degrees of freedom (df) [57,58].
Log likelihood values and parameter estimates are detailed in the results section (and
Appendix 2: Table A). In order to determine if duplicates were actively transcribed,
human and mouse expressed sequence tags (EST) were mined from the UCSC
genome browser EST. ESTs that mapped to multiple locations that showed less than
95% identity or 95% fraction length were discarded. Additionally, EST presence &
absence calls were also made using data mined from Bgee database for annotated
duplicates in our dataset [59].

Data Collection for DN and Non-DN Negative Gene Families

Gene family definitions were obtained from Ensembl for four mammalian species
Human, Monkey, Rat and Dog. Only gene families containing members in all four
species were selected. Since, gene families are groups of homologous genes that are
likely to have highly similar functions, guilt by association principle was used to assign
an entire family as Dominant Negative (DN) gene family if at least one member was
annotated as such in OMIM database. This resulted in 465 DN and 9362 non dominant-negative (non DN) gene families. Putative duplicates and pseudogenes were added to the gene family to create a super-gene family. The gene family evolutionary trees were constructed using a previously developed workflow which leverages local synteny followed by hierarchical clustering to build gene trees (see refs in papers, our paper). Single omega values were obtained from PAML CODEML package for all gene trees in DN and background datasets. See Figure 14 in Chapter 4 for entire in-silico pipeline.

DN and NON-DN Selective Pressure Calculations

Gene trees were obtained from the pipeline described in Figure 15 of Chapter 4 and basic premise of the selective pressure calculations was adopted from our previous publication\textsuperscript{6}. Multiple sequence alignment (MSA) for each gene family was done using MACSE (reference) to account for proper alignment of full length genes with pseudogene members. The gene tree and MSA were then used as input for CODEML package in PAML to generate single omega value for each gene family.

Cell Culture

HEK293T (human embryonic kidney 293T) cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% FBS (fetal bovine serum), 100 units/ml penicillin and 100 µg/ml streptomycin with 5% CO\textsubscript{2} incubation.
Construction of Plasmids

**CRISPR vector and sgRNA design:** For all our genome engineering experiments, we followed the CRISPR-Cas9 system protocol provided by the Zhang Lab at the Broad institute (Ran et al. 2013). The pSpCas9n (D10A) expression vector (Addgene plasmid # 48141) carrying a codon-optimized Cas9 gene was purchased from Addgene (http://www.addgene.org) for increased targeting specificity (Vector map is shown in Appendix 2: Figure B). The construction of the gRNA expression vector was based on the method shown on the Addgene website (http://www.addgene.org/static/data/85/85/e19394c4-5e76-11e2-a7c4-003048dd6500.pdf). Source and duplicate gene sgRNAs sequences were designed using CRISPR design tools CHOPCHOP and DNA2.0. Specific care was taken to make sure that the target sequence was immediately preceded by a 5’ NGG PAM sequence, and the 20-nt guide sequence base pairs with the opposite strand to mediate Cas9 cleavage at ~3bp upstream of the PAM. All the sgRNA designs and sequences are listed in Appendix 2 (Table B) and more details about their design strategies are discussed in the Results section. These sequences were cloned into the Cas9n vector using the protocol from Feng Zhang’s lab. These recombinant plasmids were amplified in Escherichia coli and finally purified using a QIAGEN Miniprep Kit.

**Inducible transgene expression vectors:** Expression vectors for candidate ribosomal protein RPS15 and RPS26 source and duplicate genes were constructed by inserting each cDNA into a piggyBac cumate switch inducible vector (System Biosciences, Mountain View, CA, USA; Vector map is shown in Appendix 2: Figure C). All of the plasmid vectors were confirmed by DNA sequencing.
Transfections and Antibiotic Selection

CRISPR/Cas9n transfections for gene knockout experiments

Cells were transfected using Lipofectamine 2000 (Life Technologies) at 80%–90% confluency following the manufacturer’s recommended protocol. A total of 500 ng Cas9 plasmid and 100 ng of U6-sgRNA PCR product was transfected. CRISPR positive cells were selected by puromycin selection. These cells were then confirmed by Sanger DNA sequencing.

Clonal isolations of cell lines and functional testing

Isolation of clonal cell lines were achieved by FACS using GFP selection, followed by an expansion period to establish a new clonal cell line. Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit and gene-specific primers were used to amplify the DNA sequences. Targeted genome modifications were detected by Sanger DNA sequencing.

Inducible transgene expression and subsequent CRISPR/Cas9n transfections for the complementation-rescue system

Cells were seeded into 6-well plates (Corning) at a density of 20,000 cells/well, 24 hr prior to transfections. HEK293T cells were transfected with plasmid vectors using the Lipofectamine 2000 reagent (Invitrogen) according to the supplier’s recommendations. In brief, the cells were plated on to 6-well dishes and transfected with 2 μg of plasmid DNA mixed with 5 μL of Lipofectamine 2000 for the experiments. To establish stable inducible cell lines, positively transposed cells were selected using puromycin (2
μg/mL). Since the inducible piggyBac vector features a tight cumate switch combined with an EF1-CymR repressor-T2A-Puro cassette for the establishment of stable cell lines, cumate solution (System Biosciences) was added to the puromycin-selected cells for inducing transgene expression. Polyclonal cultures were generated by continued selection under puromycin selection. For the complementation experiments, CRISPR pSpCas9n transfections were performed in the same manner as described above. Full pipeline of the complementation experiment can be seen in the Figure 7 (also see Results section for more detail). Additionally, for the Complementation rescue experiment, we replaced the antibiotic selection from puromycin to hygromycin in the Cas9n vector in order to select cells using dual selection of hygromycin (Cas9n vector) and puromycin (piggyBac vector).

**Cell Viability Tests and Data Analyses**

For crystal violet dye cell quantification assay, cells were washed twice with 1X PBS and fixed with 20% methanol solution for half an hour at room temperature. Cells were stained with 0.5% crystal violet solution for 30 min at 37 C. Stained cells were washed with water until a clear background was visible. Crystal violet absorbance was determined using a microplate reader at 550nm (Varioskan Flash, Thermo Scientific, USA). For live cell stain and viability quantification, fluorogenic esterase substrate Calcein red-orange AM was passively loaded into viable cells. Cells were washed twice with 1X PBS and 5 ml 0.5 mM working Calcein solution was added to 1 ml serum free media (1:200, 2.5 mM) and were incubated at 37 C for 1 hour before cell imaging.
Chapter 7 Appendices

Appendix 1

74 RP gene trees with all annotated duplication events. For legend, see figure 4.
Appendix 2

Table A: Log-likelihood and parameter estimates generated from random-site models for RP genes. $P =$ number of free parameters for each model, $\ell =$ log-likelihood value for each model.

RPL28:

<table>
<thead>
<tr>
<th>Models</th>
<th>$p$</th>
<th>$\ell$</th>
<th>Estimates of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: one ratio</td>
<td>1</td>
<td>-14664.881</td>
<td>$\omega=0.223$</td>
</tr>
<tr>
<td>M1a: nearly neutral</td>
<td>2</td>
<td>-15243.079</td>
<td>$p_0=0.0001, p_1=0.9999, \omega_0=0.0001, \omega_1=1.0000$</td>
</tr>
<tr>
<td>M2a: positive selection</td>
<td>4</td>
<td>-15168.003</td>
<td>$p_0=0.0000, p_1=0.83519, p_2=0.16481, \omega_0=0, \omega_1=1, \omega_3=3.39940$</td>
</tr>
<tr>
<td>M0, omega=1: fixed omega</td>
<td>4</td>
<td>-15848.007</td>
<td>None</td>
</tr>
</tbody>
</table>

RPL14:

<table>
<thead>
<tr>
<th>Models</th>
<th>$p$</th>
<th>$\ell$</th>
<th>Estimates of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: one ratio</td>
<td>1</td>
<td>-11320.898</td>
<td>$\omega=0.348$</td>
</tr>
<tr>
<td>M1a: nearly neutral</td>
<td>2</td>
<td>-10997.396</td>
<td>$p_0=0.7266, p_1=0.27331, \omega_0=0.1868, \omega_1=1.0000$</td>
</tr>
<tr>
<td>M2a: positive selection</td>
<td>4</td>
<td>-10962.242</td>
<td>$p_0=0.71909, p_1=0.17451, p_2=0.10640, \omega_0=0.20926, \omega_1=1, \omega_3=2.54602$</td>
</tr>
<tr>
<td>M0, omega=1: fixed omega</td>
<td>4</td>
<td>-12018.938</td>
<td>None</td>
</tr>
</tbody>
</table>
### RPS16:

<table>
<thead>
<tr>
<th>Models</th>
<th>p</th>
<th>ℓ</th>
<th>Estimates of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0:one ratio</td>
<td>1</td>
<td>-14320.835</td>
<td>omega=0.286</td>
</tr>
<tr>
<td>M1a:nearly neutral</td>
<td>2</td>
<td>-14134.842</td>
<td>p₀=0.76, p₁=0.24,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>omega₀=0.226, omega₁=1</td>
</tr>
<tr>
<td>M2a:positive selection</td>
<td>4</td>
<td>-14646.878</td>
<td>p₀=0, p₁=0.17451, p₂=0.10640</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>omega₀=0, omega₁=1, omega₃=4.65</td>
</tr>
<tr>
<td>M0, omega=1:fixed omega</td>
<td>4</td>
<td>-15340.056</td>
<td>None</td>
</tr>
</tbody>
</table>

### RPS18:

<table>
<thead>
<tr>
<th>Models</th>
<th>p</th>
<th>ℓ</th>
<th>Estimates of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0:one ratio</td>
<td>1</td>
<td>-14320.835</td>
<td>omega=0.328</td>
</tr>
<tr>
<td>M1a:nearly neutral</td>
<td>2</td>
<td>-14134.842</td>
<td>p₀=0.0001, p₁=0.9999,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>omega₀=0, omega₁=1.0000</td>
</tr>
<tr>
<td>M2a:positive selection</td>
<td>4</td>
<td>-14646.878</td>
<td>p₀=0.8292, p₁=0.11455,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p₂=0.05563,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>omega₀=0.288, omega₁=1, omega₃=2.19489</td>
</tr>
<tr>
<td>M0, omega=1:fixed omega</td>
<td>4</td>
<td>-176000.284</td>
<td>None</td>
</tr>
</tbody>
</table>
Table B: References for Figure 14 and Table 2.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Details</th>
</tr>
</thead>
</table>
Table C: List of sgRNA Pairs Used with Cas9 Nickase to Identify the Optimal Target Site Spacing for Double Nicking for RPS15 and RPS26 source and duplicate genes. Related to Figure 20.

<table>
<thead>
<tr>
<th>CRISPR sgRNA Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S15_Source_G1Top</td>
<td>CACCGATCCTCAGAAGAGATCGCTT</td>
</tr>
<tr>
<td>S15_Source_G1Bot</td>
<td>AAACAAGCGATCTTTCTTGAGGATC</td>
</tr>
<tr>
<td>S15_Source_G2Top</td>
<td>CACCGAGATGGTGAGTGTTGCGATT</td>
</tr>
<tr>
<td>S15_Source_G2Bot</td>
<td>AAACAAATCGCAACACTCACCATCTC</td>
</tr>
<tr>
<td>S15_Duplicate_G1Top</td>
<td>CACCGAAGTAGAGCAGAAGAAGAAG</td>
</tr>
<tr>
<td>S15_Duplicate_G1Bot</td>
<td>AAACCTTCTTCTCTGCTCTACTTC</td>
</tr>
<tr>
<td>S15_Duplicate_G2Top</td>
<td>CACCGTCCTCAGAAGAGGACAGCATCAC</td>
</tr>
<tr>
<td>S15_Duplicate_G2Bot</td>
<td>AAACGTATGCTGCTCTCTGAGGAC</td>
</tr>
<tr>
<td>S26_Source_G1Top</td>
<td>CACCGTGAGGGCAAGGACGGAG</td>
</tr>
<tr>
<td>S26_Source_G1Bot</td>
<td>AAACCTCTCCGGTGCTGCTCCAC</td>
</tr>
<tr>
<td>S26_Source_G2Top</td>
<td>CACCGGTGAGTCTTCTCTGGTG</td>
</tr>
<tr>
<td>S26_Source_G2Bot</td>
<td>AAACCACAAGCAGAAGAAGACTCACC</td>
</tr>
<tr>
<td>S26_Duplicate_G1Top</td>
<td>CACCGTCTTTGTCATCTTGGAGGCA</td>
</tr>
<tr>
<td>S26_Duplicate_G1Bot</td>
<td>AAACTGCCTCCAAGATGACAAAGAC</td>
</tr>
<tr>
<td>S26_Duplicate_G2Top</td>
<td>CACCGGAACAGTGCTGTCGGCAAA</td>
</tr>
<tr>
<td>S26_Duplicate_G2Bot</td>
<td>AAACTTTTGGACCGACCATTGTTCC</td>
</tr>
</tbody>
</table>
Figure A: Sequencing Alignments of clonally isolated HEK293T cells in which source and duplicate genes were knocked out respectively. The alignments were performed to observe the whether the surviving cells received the deletions/manipulations correctly or were escapers. For example, the cells in which RPS15 source genes were knocked out and survived did not receive the deletions correctly and hence were viable, as seen in part (A) of the figure. The cells in which RPS26 duplicate genes were knocked out and survived had received the knockout correctly and were still viable, as seen in part (B) of the figure.
Figure B: Vector PX462 is the CRISPR-Cas9n vector for our knockout experiments, as mentioned in Materials and Methods.
Figure C: Vector PBQM812A-1 is the cumate inducible PiggyBac vector for our complementation rescue experiments, as mentioned in Materials and Methods.
Figure D: Replicates of RPS15 and RPS26 CRISPR-Cas9n knockout experiments.
Figure E: Replicates of RPS15 at two timepoints for the Complementation-rescue experiments.

**RPS15 COMPLEMENTATION EXPERIMENTS**

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPS15_Source_Ind</td>
<td>RPS15_Source_NotInd</td>
</tr>
<tr>
<td>RPS15_Duplicate_Ind</td>
<td>RPS15_Duplicate_NotInd</td>
</tr>
</tbody>
</table>

Day 1 and Day 3 images show replicates of RPS15, with S15 Source Gene Knockout indicating the experimental conditions.
Figure F: Replicates of RPS15 at two timepoints for the Complementation-rescue experiments.

**RPS26 COMPLEMENTATION EXPERIMENTS**

Day 1
- RPS26_Source_Ind
- RPS26_Source_NotInd
- S26 Source Gene Knockout

Day 3

Day 1
- RPS26_Duplicate_Ind
- RPS26_Duplicate_NotInd
- S26 Source Gene Knockout

Day 3
Figure G: Calcein live cell imaging examples to confirm viability
1) Tempo and Mode of Gene Duplication in Mammalian Ribosomal Protein Evolution


Abstract: Gene duplication has been widely recognized as a major driver of evolutionary change and organismal complexity through the generation of multi-gene families. Therefore, understanding the forces that govern the evolution of gene families through the retention or loss of duplicated genes is fundamentally important in our efforts to study genome evolution. Previous work from our lab has shown that ribosomal protein (RP) genes constitute one of the largest classes of conserved duplicated genes in mammals. This result was surprising due to the fact that ribosomal protein genes evolve slowly and transcript levels are very tightly regulated. In our present study, we identified and characterized all RP duplicates in eight mammalian genomes in order to investigate the tempo and mode of ribosomal protein family evolution. We show that a sizable number of duplicates are transcriptionally active and are very highly conserved. Furthermore, we conclude that existing gene duplication models do not readily account for the preservation of a very large number of intact retroduplicated ribosomal protein (RT-RP) genes observed in mammalian genomes. We suggest that selection against dominant-negative mutations may underlie the unexpected retention and conservation of duplicated RP genes, and may shape the fate of newly duplicated genes, regardless of duplication mechanism.
2) SCLD: a stem cell lineage database for the annotation of cell types and developmental lineages


Abstract: Stem cell biology has experienced explosive growth over the past decade as researchers attempt to generate therapeutically relevant cell types in the laboratory. Recapitulation of endogenous developmental trajectories is a dominant paradigm in the design of directed differentiation protocols, and attempts to guide stem cell differentiation are often based explicitly on knowledge of in vivo development. Therefore, when designing protocols, stem cell biologists rely heavily upon information including (i) cell type-specific gene expression profiles, (ii) anatomical and developmental relationships between cells and tissues and (iii) signals important for progression from progenitors to target cell types. Here, we present the Stem Cell Lineage Database (SCLD) (http://scld.mcb.uconn.edu) that aims to unify this information into a single resource where users can easily store and access information about cell type gene expression, cell lineage maps and stem cell differentiation protocols for both human and mouse stem cells and endogenous developmental lineages. By establishing the SCLD, we provide scientists with a centralized location to organize access and share data, dispute and resolve contentious relationships between cell types and within lineages, uncover discriminating cell type marker panels and design directed differentiation protocols.
Chapter 8 References


