# The role of alternative splicing in primate genome evolution

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
by
Lenore Pipes
August 2017

# Abstract

We present the first pipeline to systematically identify and measure changes in Ψ in alternative splicing events across different genomes without annotation. Our method identified ~15,000 one-to-one orthologous alternative splicing events across human and 4 non-human primates. We show that alternative splicing events are increasing in abundance in every human tissue relative to non-human primates. Additionally, contrary to the tissue-dominated conservation pattern of gene expression, we show that most tissues except for brain, heart, and muscle, have a species-specific splicing pattern. Using these orthologous events, we identified 3,954 significant differential splicing events in 1,807 genes between humans and non-human primates. This thesis represents is part of the ambitious goal towards quantifying all of the changes in genomic complexity that occur between primate species. We provide evidence that these changes could be part of the "missing" genomic basis for the origin of human-specific traits.

# Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university. Chapter 2 of this dissertation is from a manuscript (Pipes et al.) that is freely available at *Nucleic Acids Research*. Additionally, portions of Chapter 3 appear in a manuscript (Couger et al.) that appeared in *Concurrency and Computation Practice and Experience*, portions of which were equally shared with Brian Couger.

Lenore Pipes

# Biographical Sketch

Lenore Pipes was born and raised in Guam U.S.A. She moved to Pennsylvania when she was 18 to attend Swarthmore College. At Swarthmore College, Pipes received her B.A. with Honors in Biology and Sociology/Anthropology. While she was an undergraduate, she published papers with her summer mentors Michael L. Johnson (University of Virgnia) and Theodore Schurr (University of Pennsylvania). After college, she spent two years working as a research technician at the Wistar Institute in the laboratory of Harold Riethman. In 2011, she decided to pursue her Ph.D. in computational biology at Cornell University under the mentorships of Adam Siepel and Christopher Mason. She was awarded a graduate research fellowship from the National Science Foundation in 2013.

Outside of academics, Lenore was a professional road cyclist. She competed at the 2012 and 2015 women's road cycling world championships. She raced with numerous professional teams on the U.S. domestic circuit (including Tibco and Colavita), and also with UCI World Tour Team BePink.

# Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my advisors Adam Siepel and Christopher Mason, and my additional committee members Andrew Clark and Jeff Pleiss. I would especially like to thank Melissa Hubisz, Charles Danko, Jaaved Mohammed, Ilan Gronau, and Andre Martins for their endless insight and expert advice. I would like to thank XSEDE for their generous and seemingly infinite support in my project and especially Philip Blood who has always been quick to troubleshoot any issues I had with using their resources. This project would not be possible without their support.

# Preface

When alternative splicing was first discovered in 1977, it was considered a unique phenomenon. Now, with the sequencing of the human genome and many other genomes, it is evident that alternative splicing is not an exception, it is the rule. Since the first high-throughput studies of alternative splicing began only ~10 years ago, being able to study alternative splicing genome-wide is still in its formative years. Even less studied are the genome-wide changes in alternative splicing across species. Given the relevance of the non-human primate reference transcriptome resource (NHPRTR), the largest non-human primate RNA-Seq dataset ever created, new opportunities arose to enable a timely thesis project to study alternative splicing changes across primates. Chapter 1 discusses the biologically and bioinformatically relevant background to the dissertation. Chapter 2 is a description of the richness of the dataset, which has appeared in *Nucleic Acids Research*. Some of the computational challenges in creating isoforms directly from RNA-Seq data is discussed in Chapter 3, which also features some portions from a manuscript that appeared in *Concurrency and Computation Practice and Experience*. Chapters 4 and 5 represent unpublished material regarding the study and characterization of orthologous alternative splicing events in primates.

# Contents

*"Change alone is eternal, perpetual, immortal."*
— Arthur Schopenhauer

# Chapter 1

# Alternative Pre-mRNA Splicing and its potential for large-scale evolutionary changes

*"Introns are both frozen remnants of history and the sites of future evolution...specific recombinations between introns can bring exons together into a transcriptional unit to make special differentiation products."*

— Walter Gilbert, Why genes in pieces?, Nature 1978

## 1.1 Introduction

Alternative splicing, or the production of multiple mRNA variants from a single gene, is a fundamental regulatory crossroad between transcription and translation that affects nearly 95% of mammalian genes [1]. In 1978, Walter Gilbert proposed the now widely accepted notion that the function of alternative splicing is to increase the diversity of mRNAs expressed from the genome. Gilbert also suggested that this function has profound implications for evolution since it "can seek new solutions without destroying the old." In 2003, Modrek and Lee observed that exons that were newly created were prevalent in minor isoforms of a gene which they hypothesized was an evolutionary mechanism to allow the exon to obtain beneficial mutations without losing the benefits of the major isoform [2]. Only with the recent advent of high-throughput RNA sequencing (RNA-seq) has alternative splicing on a

transcriptome-wide level been able to be studied with single-nucleotide resolution. Two timely studies by Barbosa-Morais et al. (2012) and Merkin et al. (2012) both studying the evolution of alternative splicing in mammalian tissues proposed that the lack of conservation in alternative splicing events between species may be the driving force of species divergence. Furthermore, a controversial claim by Barbosa-Morais et al. (2012) suggested that the highest complexity in alternative splicing occurs within primates, and that the human cerebellum has more than twice the abundance of alternative splicing events than any other tissue studied. It is currently not known how the *cis*-acting splicing-regulatory "code" impacts the evolutionary divergence of splicing in closely related species. The field is at both an ideal and necessary time to understand the evolutionary changes caused by alternative splicing with a focus on primates [3].

Almost all human protein-coding genes undergo alternative splicing which means that, depending on cellular conditions, an alternative exon may be included or excluded from the mature messenger RNA (mRNA). Inside a typical human somatic cell, on average, 300,000-400,000 distinct mRNA molecules are transcribed from >10,000 genes at a time [4]. The ability to change the output of the genetic information depending on cellular states, and the ability to greatly increase the proteomic and regulatory diversity from the information content of the genome, makes alternative splicing a critical stage for regulating gene expression. The variable use of *cis*-acting RNA elements in exons and flanking introns that are recognized by *trans*-factors allows different pairs of splice sites in primary transcripts to be selected in a cell type-, condition-, or species-specific manner. The startling variability and abundance of mRNA transcripts has the ability to be the source of population-wide complexity in phenotype. In fact, it has been postulated that even under the simplest genetic systems model of alternatively spliced transcripts, the production of so many regulatory proteins can cause a cell to transition into chaos [5]. Although the mechanisms responsible for the regulation of alternative splicing have been studied in some depth for several genes, only a limited amount is known about the splicing control factors that function to regulate alternative splicing genome-wide. Yet, the availability of the genome sequences of multiple organisms has facilitated tremendous growth in the use of bioinformatics and genome-wide techniques to study alternative splicing. Although humans may have roughly the same number of protein-coding genes as *C. elegans*, alternative splicing along with other processes such as alternative use of transcription start sites, alternative polyadenylation, RNA editing, and post-translational modification (i.e., phosphorylation, ubiquitylation, and SUMOylation) together expand the proteome diversity in humans to a staggering level. Alternative splicing is just one process that

facilitates the extraordinary diversity of the functional landscape. It is still unclear as to what extent each of these processes contribute to the creation of functionally distinct proteins but it evident that alternative splicing remains one of the main drivers of this diversity in eukaryotes.

Because of the ability of alternative splicing to combinatorically create thousands of isoforms from a single gene, it had been hypothesized since its discovery that alternative splicing exhibits the capacity needed to account for organismal complexity. Ever since the first comparative genomics study in 1975 by King and Wilson between humans and chimpanzees [6], identifying the genetic underpinnings that correlate with complexity in anatomy and/or behavior has remained elusive. The sequencing of the human genome and that of model organisms surprisingly revealed that *C. elegans* and humans had roughly the same number of protein-coding genes. What exactly accounts for the vast difference in phenotypic complexity between humans and *C. elegans*? Just as one example as a difference in complexity, in the nervous system alone, humans have several billion neurons while *C. elegans* has only several hundred. However, it is clear that the number of appreciable alternatively spliced genes has greatly increased in humans compared to *C. elegans*[7] to the point where almost every gene is alternatively spliced in humans to produce isoforms with different activity, localization, stability, and/or specificity. Furthermore, alternative splicing is an ideal candidate to also cause large-scale evolutionary changes. Since the same DNA can be used to encode multiple different mRNAs, the major-isoform can be produced as normal while a minor-isoform is altered. The creation of the minor-isoform creates a tunnel in the fitness landscape that natural selection can then act upon. The most unique human-specific trait, the brain, has evolved very rapidly over only a few million years. The complexity of the human brain might be the result of mutations impacting alternative splicing of the many brain-expressed genes instead of incremental selection on mutations in many different genes. We show evidence in this thesis for the increased abundance of alternative splicing in brain relative to other tissues and evidence that the brain is under different selective pressures than other tissues.

## 1.2 Basics of Alternative Splicing

Precursor messenger RNA (pre-mRNA) splicing was a phenomemnon first discovered 40 years ago that consists of a series of biochemical reactions that function to remove introns and ligate flanking exons [8, 9]. Alternative splicing occurs within the

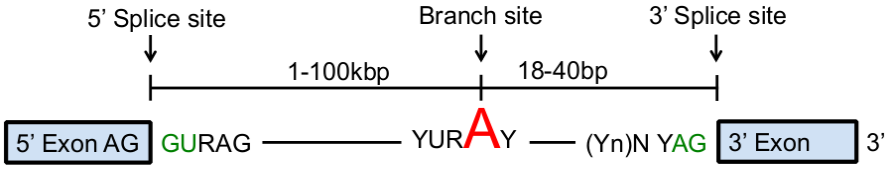


**Figure 1.1:** Schematic of canonical splice signals.

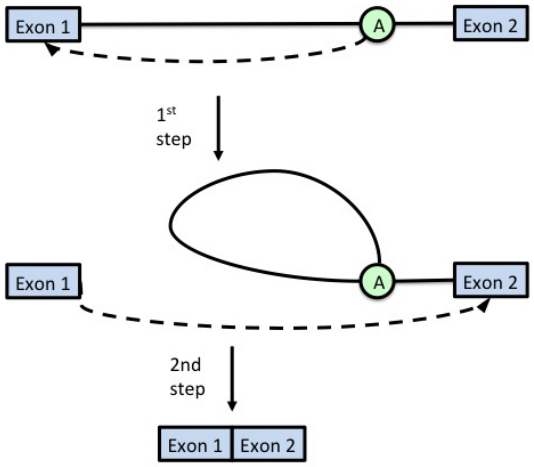


**Figure 1.2:** Schematic of the two-step splicing reaction.

context of a multimegadalton ribonucleoprotein complex known as the spliceosome. The spliceosome is comprised of five small nuclear RNAs (snRNAs) and about 200 protein components. Assembly of the spliceosome on pre-mRNA requires recognition of several *cis*-acting splicing-regulatory elements (SREs) located within the intron: the 5′ splice site GU, branch point A, polypyrimidine tract and 3′ splice site AG (**Figure 1.1**). Within the enormous machine of the spliceosome, intron excision occurs in two transesterification steps: (1) cleavage at the 5′ splice site (donor site), coupled to formation of a lariat structure in which the first nucleotide of the intron is linked via a 2′-5′ phosphodiester bond to an intronic adenosine (the branch point) in the vicinity of the 3′ splice site (acceptor site); and (2) ligation of the two exons, coupled to cleavage at the 3′ splice site (**Figure 1.2**). Many diseases are caused by point mutations that disrupt canonical splice site signals and, in consequence, disrupt the normal splicing pattern [10]. Splice site recognition is mediated by proteins (serine/arginine proteins, heterogenous nuclear ribonucleoproteins, polypyrimidine tract-binding proteins, the TIA1 RNA-binding protein, Fox proteins, Nova proteins, and more) that bind specific regulatory sequences. Combinatorial control by multiple *trans*-acting splicing regulators permits specific and differential recognition of short, degenerate signals (exonic splicing enhancers, ESEs; intronic splicing enhancers, ISEs; exonic splicing silencers,

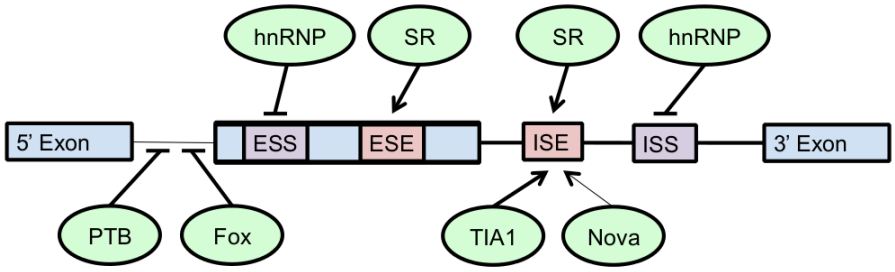


**Figure 1.3:** Schematic of *cis*-acting splicing regulatory elements. Silencers are depicted with bars and enhancers are depicted with arrows. Heterogeneous ribonucleoprotein particle (hnRNPs) are RNA complexes that bind both exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs). Serine/arginine (SR) proteins recognize exonic splicing enhancer (ESE) elements as well as intronic splicing enhancer (ISE) elements. Polypyrimidine tract binding protein (PTB) is a repressive alternative splicing regulator. Nova, Fox, and T-cell intracellular antigen 1 (TIA1) are families of RNA-binding proteins that can either promote or suppress the recognition of the splice site depending on the position of their binding site.

ESSs; and intronic splicing silencers, ISSs) and creates a situation in which variations in the concentration of a single *trans*-factor can elicit a change in the splicing pattern (**Figure 1.3**). Cartegni et al. (2002) [11] estimated that up to 50% of all mutations that lead to gene disfunction are ones that cause aberrant splicing and ~15% of inherited genetic disorders are caused by deleterious mutations the interfere with splicing. Since the cost of incorrect splicing to the cell is extremely high [12], the spliceosome has an intrinsically high degree of fidelity by efficiently pairing constitutively spliced exons separated by introns up to $10^5$ nucleotides in length [13]. Furthermore, it has been shown that the recognition of exons by the spliceosome depends equally on signals from the 3′ and 5′ splice sites [14]. Most regulation of alternative splicing occurs at the earliest stages of the spliceosome assembly pathway: between the interaction of *cis*-acting and *trans*-acting factors that either promote or repress the recognition of the canonical splicing signals. While the processing of short introns is thought to be largely dependent on the proximity of a 5′ and 3′ splice site [15], the processing of long introns is thought to require more regulatory information such as *cis*-regulatory motifs and *trans*-acting splicing factors, the chromatin landscape, and the kinetics of polymerase elongation [16–18]. In fact, a survey of the canonical splice signals determined that the information content in the signal becomes less preserved as the number of introns increases which leads to insufficiency of correct splicing in higher eukaryotes [19]. Additionally, Sun and Chasin [20] found that in large introns it is possible to identify many pseudo splice sites that represent canonical signals even closer than the actual splice sites. This degeneracy creates possibilities for splice site recognition especially

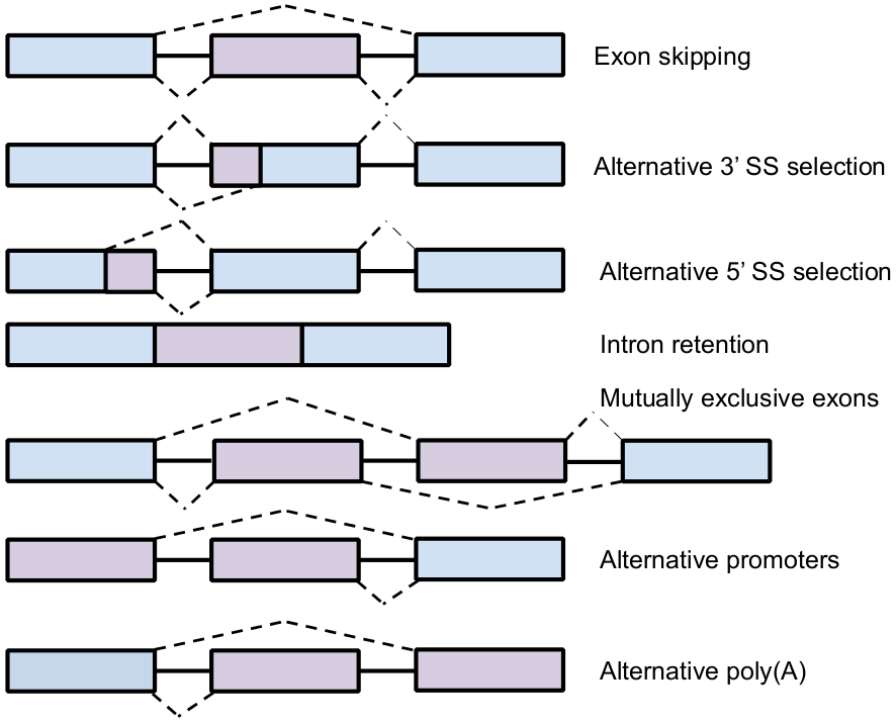


**Figure 1.4:** Schematic of common splicing patterns.

when canonical signals are weak. There are a number of features that impact splice site choice including exon and flanking intron length, splice site strength, splicing *cis*-regulatory elements, interspersed repeat content, mRNA secondary structure, and RNA editing. Exon shuffling, exonization of intronic sequences, and transition of a constitutive exon to an alternative exon are the three evolutionary mechanisms known to create an alternatively spliced exon, and different common types of alternative splicing are shown in **Figure 1.4**. In humans, exon skipping is by far the most common type of alternative splicing.

## 1.3 Identification of alternative splicing events

Starting with customized microarrays, high-throughput technology for the identification of alternative splicing events has only been around since 2001 [21]. Initial studies used multi-probe designs of microarrays to detect splicing variants and attempt to quantify their expression. Later studies improved the designs of these probes to discover novel spliced isoforms of genes and determine the tissue specificity of alternative splicing events [22–24]. Since next-generation sequencing has become avail-

able, RNA-Seq quickly became the standard form for studying alternative splicing genome-wide. RNA-Seq provides a high-throughput way to directly or indirectly sequence RNA molecules. RNA-Seq has the ability to detect unannotated alternative splicing events, is not confounded by cross-hybridization like microarrays, can relatively accurately quantify gene expression and percent-spliced in, and can provide the study of RNA editing and genetic variants with its single nucleotide resolution. Most RNA-Seq methods initially convert RNA into cDNAs which are made into a sequencing library consisting of short DNA fragments of the RNA of interest, which is then flanked by adapters. Next, the DNA library is sequenced from a single-end or both ends (paired-end) which results in the final pool of RNA-Seq reads. This results in RNA-Seq reads that represent a snapshot of the expression in a cellular sample of interest [25]. Yet there are many challenges when using RNA-Seq to study alternative splicing events. First, different library generation protocols which can also vary in sample quality, concentration, sequencing depth, and read length create batch effects that greatly influence the downstream analyses. Second, in order to study splicing it is desirable to have a large number of splice junction reads, which means that, in order to be accurate, the RNA-Seq generation can be costly because it needs to be long-read, paired-end, and at high sequencing depth. Third, the RNA-Seq reads produced are much shorter than most full-length isoforms, so these must be computationally reconstructed either *de novo* or genome-guided, and the accuracy of this is limited. In general, there are two methods, exon-centric and isoform-centric, to bioinformatically analyze alternative splicing using RNA-Seq. Exon-centric methods estimate the frequency of exon inclusion by calculating percent-spliced-in (PSI) against the global expression of the gene. Isoform-centric methods estimate the abundance of each isoform but this has large uncertainty for most full-length isoforms. Most alternative splicing studies with RNA-Seq has relied on gene annotation which has made studying splicing in unannotated or poorly annotated genomes difficult if not impossible.

## 1.4 Comparative primate transcriptomics and evolution by splicing

To date, there have been few studies that have used comparative RNA sequencing to study differences among primates. Blekhman et al. (2010) [26] sequenced liver RNA

in biological triplicate from human, chimpanzee, and rhesus macaque ( ~ 13 million 35-bp single-end reads per individual), mapped the reads to the respective reference genomes, and identified frequent lineage-specific changes in transcript expression. Brawand et al. (2011) [27] performed RNA sequencing on nine mammalian species including human, chimpanzee, gorilla, orangutan, bonobo, and macaque using cerebral cortex, cerebellum, heart, kidney, liver, and testis ( ~ 25 million reads, some with 100-bp paired-end reads, per tissue per individual). They characterized the extent of transcriptome variation between organs and species and identified putatively selectively driven expression switches that could have an impact on phenotypic changes between the species. Perry et al. (2012) [28] characterized genetic and regulatory primate variation by sequencing liver RNA from 11 NHPs ( ~ 16.4 million 76-bp paired end reads per individual) of which 7 of those NHPs had little or no published genomic resources, and *de novo* assembled the reads and performed a multispecies alignment. The Brawand et al. (2011) [27] dataset was later used by Reyes et al. (2013) [29] that identified 3,800 exons that show strong tissue-dependent usage patterns across all primate species. These three datasets (Blekhman et al., 2010 [26]; Brawand et al., 2011 [27]; and Perry et al., 2012 [28]) represented the only primate RNA-seq datasets that were publicly available before this thesis started. Merkin et al. performed RNA-seq on 9 tissues ( ~ 120 million reads per tissue per individual) from five vertebrates in biological triplicate that included rhesus macaque and Barbosa-Morais et al. (2012) performed RNA-seq from 10 vertebrate species that included human, chimpanzee, orangutan, and macaque using whole brain, forebrain cortex, cerebellum, heart, skeletal muscle, liver, kidney, and testis. Evidence from Merkin et al. (2012) and Barbosa-Morais et al. (2012) both showed that while gene expression patterns across species are highly conserved, most alternative splicing events are unexpectedly species-specific. Furthermore, they propose that the amount of isoform variability is so high between species that it may act as a driving force for speciation. The fast rate of evolution in alternative splicing might be explained by alternative splicing as a mechanism that enables cells to experiment with new versions of proteins without risking the complete loss of the original transcript isoforms. These new results are in line with results from Modrek and Lee (2003) [2] who used EST data to show that most minor splice forms between mouse and human are not conserved. This evidence supports an evolutionary model in which alternative splicing can relax negative selection pressure against large-scale changes in gene structure such as exon creation. There is strong negative selection against the introduction of a new exon into an existing gene because it is likely to disrupt the reading frame or disrupt an essential structural or functional element in the

protein product. However, if a new exon was introduced in a minor splice form, this would not interfere with the original gene product so the negative selection pressure against this event would be minimal. Thus, alternative splicing perhaps opens neutral or nearly neutral evolutionary paths for large-scale evolutionary changes like exon creation events.

## 1.5 The splicing code

Alternative splicing also introduces strong selection pressure for RNA-sequence motifs that are involved in the regulation of alternative splicing. One of the long-term goals in the splicing field is to decipher the "splicing code" which controls the splicing pattern of any primary transcript from its sequence in a wide range of cell types and conditions. Apart from the canonical splice site signals, the majority of the information required for splicing is thought to lie in the SREs. RNA-seq data has not only allowed systematic bioinformatic analyses to probe the general validity of known regulatory sequences but it has also allowed identification of new SREs. One of the most cited examples of this approach is the analysis of the splicing patterns of over 3,500 cassette-type alternative exons across 27 mouse tissues (in four different tissue systems: the central nervous system, muscle, digestive system, and whole embryos) ranging in development from embryonic stages to adult [1]. Barash et al. (2010) were able to make genome-wide predictions for different classes of tissue-specific alternative splicing events by developing a machine learning algorithm that extracted combinations of SREs (from a compendium of over 1,000 SREs). As previously mentioned, in order for the spliceosome to reliably differentiate authentic exons and splice sites from pseudo-exons and decoy splice sites, it relies on specific features of the sequence such as ISEs, ISSs, ESEs, and ESSs. More specifically, the ESEs are specific short nucleotide sequences that are targeted by trans-factors such as Serine/Argine-rich proteins which promote exon definition [30]. ESSs act as binding sites for trans-factors (such as hnRNP proteins) to help the spliceosome ignore pseudo-exons and decoy splice sites [31]. The ISEs and ISSs have similar roles and are located in the introns. While diverse, these splicing factors have some similar characteristics and binding sites for splicing factors have been identified in long stretches of RNA. For example, binding sites for polypyrimidine tract-binding protein and CELF proteins are contained in the polypyrimidine tract [32], binding sites for TIA1/TIAL1 proteins are contained in poly-U stretches [33], and binding sites for NOVA-1 are contained in clusters of YCAY

near splice sites of alternatively spliced exons [34]. The preferences for splicing factors to bind consecutive elements can be partially attributed to the modularity of their structure which typically contains several RNA recognition motifs that are involved in binding [35]. Like transcription factor binding sites, clusters of splicing factor binding sites are evolutionarily conserved. There are currently 71 experimentally validated (i.e., by CLiP-seq results) human RNA-binding splicing regulatory proteins [36] with thousands of binding sites. There are many bioinformatics tools publicly available to study or predict splice signals. They have various approaches from using blastn to align a query sequence to a database of alternative splicing events and splice signals (EuSplice from Bhasi et al., 2007 [37]) to *ab initio* prediction approaches (MHMMotif from Churbanov et al., 2006 [38]) to a method that considers the splicing factor genomic environment as wells as evolutionary conservation of the element [39].

## 1.6 Functional changes associated with splicing changes

The effects of alternative splicing on protein products can be dramatic. For example, alternative splicing in the gene encoding the Fas receptor produces varying effects on apoptosis [40], differential constitutive splicing in males and females of the *fruitless* gene in Drosophila is essential for courtship behavior [41], and ganglion-specific splicing of the TRPV1 gene in vampire bats underlies their specialized ability to detect infrared radiation [42]. A genome-wide view can be taken by analyzing DNA variants that alter splicing ratios and mapping their traits through splicing quantitative trait loci (sQTLs). Gonzalez-Porta et al. (2012) [43] estimated that $\sim$60% of total variation in transcript isoform abundance is due to transcription variation. Thus, the remaining variability can be largely due to splicing variation. The first genome-wide analyses of splicing variation were done with the Affymetrix exon array with $\sim$6 million exon-targeted probes [44–46]. In these studies, the microarray probe intensities of individual exons relative to those of the entire gene model were used to quantify exon inclusion levels and then associations with SNPs were tested to identify sQTLs. Kwan et al. (2008) [44] investigated the alternative splicing variation in lymphoblastoid cell lines derived from the CEU HapMap population, and identified marker loci linked to particular alternative splicing events. They detected both annotated and novel alternatively spliced variants, and showed that such variation among individuals is heritable and genetically controlled. Heinzen et al. (2008) [47] again used the same Affymetrix exon array to study tissue-specific alternative splicing in brain and

blood cell samples and suggested that splicing effects might have more phenotypic significance than overall changes in gene expression. With the advent of RNA-seq, we now have the ability to detect novel transcripts that are not probed on the microarray and can more accurately quantify exon inclusion levels at single nucleotide resolution. We can also precisely infer the effects of the disruption the splicing signal. For each exon of each gene, the fraction of reads mapped to that exon compared to all reads in the gene can be used as a quantitative trait. There are a few pioneering studies that have characterized transcriptome variation using RNA-seq data. Pickrell et al. (2010) [48] used low-coverage RNA-seq data from 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals and performed a linear regression of the relative exon inclusion levels against all polymorphisms within 200 kb of the gene and found 187 putative sQTLs in humans. Montgomery et al. (2010) [49] also used low-coverage RNA-seq data (37-bp paired-end sequencing; 11-22 million reads per individual) and used the exon reads counts as the phenotype and carried out Spearman correlation analysis with the genotypes.

## 1.7 Motivation

This study provides the first report of RNA-seq analysis for several primate species (i.e., Sooty Mangabey, Common Marmoset, Ring-Tailed Lemur, and Pig-tailed Macaque). This study is also the first attempt known to use >2 billion RNA-seq reads to build a transcriptome *de novo*. This is also the most extensive study profiling RNA-seq expression from 10 different tissues from 12 non-human primates, and we have already created the most comprehensive transcriptome database for these species. Since alternative splicing has not been studied in these species extensively, this will also be the first study to do so with single nucleotide resolution.

# References


1. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
2. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics* **34,** 177 (2003).
3. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–1599 (2012).
4. Cui, Y. & Irudayaraj, J. Inside single cells: quantitative analysis with advanced optics and nanomaterials. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **7,** 387–407 (2015).
5. Likhoshvai, V. A., Kogai, V. V., Fadeev, S. I. & Khlebodarova, T. M. Alternative splicing can lead to chaos. *Journal of bioinformatics and computational biology* **13,** 1540003 (2015).
6. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees (1975).
7. Tourasse, N. J., Millet, J. R. & Dupuy, D. Quantitative RNAseq Meta Analysis Of Alternative Exon Usage In C. elegans. *bioRxiv,* 134718 (2017).
8. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* **74,** 3171–3175 (1977).
9. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12,** 1–8 (1977).
10. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136,** 777–793 (2009).
11. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature reviews. Genetics* **3,** 285 (2002).

12. Jaillon, O. *et al.* Translational control of intron splicing in eukaryotes. *Nature* **451,** 359 (2008).

13. Fox-Walsh, K. L. & Hertel, K. J. Splice-site pairing is an intrinsically high fidelity process. *Proceedings of the National Academy of Sciences* **106,** 1766–1771 (2009).

14. Shepard, P. J., Choi, E.-A., Busch, A. & Hertel, K. J. Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. *Nucleic acids research* **39,** 8928–8937 (2011).

15. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences* **98,** 11193–11198 (2001).

16. Hertel, K. J. Combinatorial control of exon recognition. *Journal of Biological Chemistry* **283,** 1211–1215 (2008).

17. Yu, Y. *et al.* Dynamic regulation of alternative splicing by silencers that modulate 5âĂš splice site competition. *Cell* **135,** 1224–1236 (2008).

18. Chen, W., Luo, L. & Zhang, L. The organization of nucleosomes around splice sites. *Nucleic acids research* **38,** 2788–2798 (2010).

19. Irimia, M., Penny, D. & Roy, S. W. Coevolution of genomic intron number and splice sites. *Trends in Genetics* **23,** 321–325 (2007).

20. Sun, H. & Chasin, L. A. Multiple splicing defects in an intronic false exon. *Molecular and Cellular Biology* **20,** 6414–6425 (2000).

21. Hu, G. K. *et al.* Predicting splice variant from DNA chip expression data. *Genome Research* **11,** 1237–1245 (2001).

22. Pan, Q. *et al.* Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Molecular cell* **16,** 929–941 (2004).

23. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141–2144 (2003).

24. Stolc, V. *et al.* A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306,** 655–660 (2004).

25. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10,** 57–63 (2009).

26. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome research* **20,** 180–189 (2010).

27. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

28. Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome research* **22,** 602–610 (2012).

29. Reyes, A. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences* **110,** 15377–15382 (2013).

30. Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in biochemical sciences* **25,** 106–110 (2000).

31. Zhu, J., Mayeda, A. & Krainer, A. R. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular cell* **8,** 1351–1361 (2001).

32. Gromak, N., Matlin, A. J., Cooper, T. A. & Smith, C. W. Antagonistic regulation of $\alpha$-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein. *Rna* **9,** 443–456 (2003).

33. Aznarez, I. *et al.* A systematic analysis of intronic sequences downstream of 5' splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome research* **18,** 1247–1258 (2008).

34. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444,** 580 (2006).

35. Cléry, A., Blatter, M. & Allain, F. H. RNA recognition motifs: boring? Not quite. *Current opinion in structural biology* **18,** 290–298 (2008).

36. Giulietti, M. *et al.* SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic acids research* **41,** D125–D131 (2012).

37. Bhasi, A., Pandey, R. V., Utharasamy, S. P. & Senapathy, P. EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. *Bioinformatics* **23,** 1815–1823 (2007).

38. Churbanov, A., Rogozin, I. B., Deogun, J. S. & Ali, H. Method of predicting splice sites based on signal interactions. *Biology Direct* **1,** 10 (2006).

39. Akerman, M., David-Eden, H., Pinter, R. Y. & Mandel-Gutfreund, Y. A computational approach for genome-wide mapping of splicing factor binding sites. *Genome biology* **10,** R30 (2009).

40. Cascino, I., Fiucci, G., Papoff, G. & Ruberti, G. Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *The Journal of Immunology* **154,** 2706–2713 (1995).

41. Demir, E. & Dickson, B. J. fruitless splicing specifies male courtship behavior in Drosophila. *Cell* **121,** 785–794 (2005).

42. Gracheva, E. O. *et al.* Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature* **476,** 88 (2011).

43. Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome research* **22,** 528–538 (2012).

44. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nature genetics* **40,** 225–231 (2008).

45. Coulombe-Huntington, J., Lam, K. C., Dias, C. & Majewski, J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS genetics* **5,** e1000766 (2009).

46. Fraser, H. B. & Xie, X. Common polymorphic transcript variation in human disease. *Genome research* **19,** 567–575 (2009).

47. Heinzen, E. L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS biology* **6,** e1000001 (2008).

48. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768 (2010).

49. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464** (2010).

# Chapter 2

# The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics

from Nucleic Acids Res (2013) 41 (D1): D906-D914

# Abstract

RNA-based next-generation sequencing (RNA-Seq) provides a tremendous amount of new information regarding gene and transcript structure, expression and regulation. This is particularly true for non-coding RNAs where whole transcriptome analyses have revealed that the much of the genome is transcribed and that many non-coding transcripts have widespread functionality. However, uniform resources for raw, cleaned and processed RNA-Seq data are sparse for most organisms and this is especially true for non-human primates (NHPs). Here, we describe a large-scale RNA-Seq data and analysis infrastructure, the NHP reference transcriptome resource (http://nhprtr.org); it presently hosts data from 12 species of primates, to be expanded to 15 species/subspecies spanning great apes, old world monkeys, new world monkeys and prosimians. Data are collected for each species using pools of RNA from comparable tissues. We provide data access in advance of its deposition at NCBI, as well as browsable tracks of alignments against the human genome using the UCSC genome browser. This resource will continue to host additional RNA-Seq data, alignments and assemblies as they are generated over the coming years and provide a key resource for the annotation of NHP genomes as well as informing primate studies on evolution, reproduction, infection, immunity and pharmacology.

## 2.1 Introduction

Sequencing genomes has quickly become the scientific standard for being able to study any organism. The rapidly falling costs of sequencing from the development of massively parallel sequencing technologies have now made it possible for even

individual laboratories to undertake whole genome efforts at unprecedented resolution and scale [1]. For non-human primates (NHPs), this has resulted in genomic and transcriptomic information changing from virtually non-existent to becoming extremely expansive within the last few years [2]. Complete published draft genome sequences are now available for the chimpanzee [3], gorilla [4], baboon [5] and the Indian rhesus macaque [6], along with recently completed draft genomes for the cynomolgus macaque [7] and the Chinese rhesus macaque [8]. With the publication of each genome has come the increased power to make evolutionary and functional inferences. However, the annotation of these genomes has often lacked extensive evidence for the transcriptionally active units, again reflecting the historical high-cost and labor-intensive effort of cDNA sequencing, a problem affecting the annotation of both protein coding genes and the newly appreciated non-coding RNAs. The most recent estimates of the well-annotated human genome show more non-coding genes than protein coding genes (ENCODE)[9] and research has now confirmed the role of non-coding RNAs have in pre- and post-transcriptional gene regulation [9], developmental processes[10] and human disease [11]. However, non-coding genes have been very limited or absent in the annotation of NHP genomes and like many protein coding genes they are inferred based on the human genome[12] rather than from species-specific evidence.

NHPs provide critical biomedical models for many aspects of human health and disease and yet the genetic basis of phenotypic traits in NHPs remains poorly understood—despite the amount of genomic data now available. Therefore, the full potential of these model organisms can only be realized with a complement of genomic information that captures both the similarities and differences to human, a requirement that is equally critical to understanding primate evolution. Most notably, comparative genomics studies strongly suggest that the significant differences between modern humans and chimpanzees are likely due at least as much to changes in gene regulation as to modifications of the genes themselves, a conjecture initially proposed by King and Wilson >30 years ago [13] and reinforced by the ENCODE results that suggest functional/regulatory roles for much of the genome that is devoid of protein coding loci.

Following the 4th International Conference on Primate Genomics (Seattle, 2010), we organized a committee of investigators to assess the requirements of the research community for NHP transcriptome information; this process included representatives from many of the National Primate Research Centers, as well as experts in primate

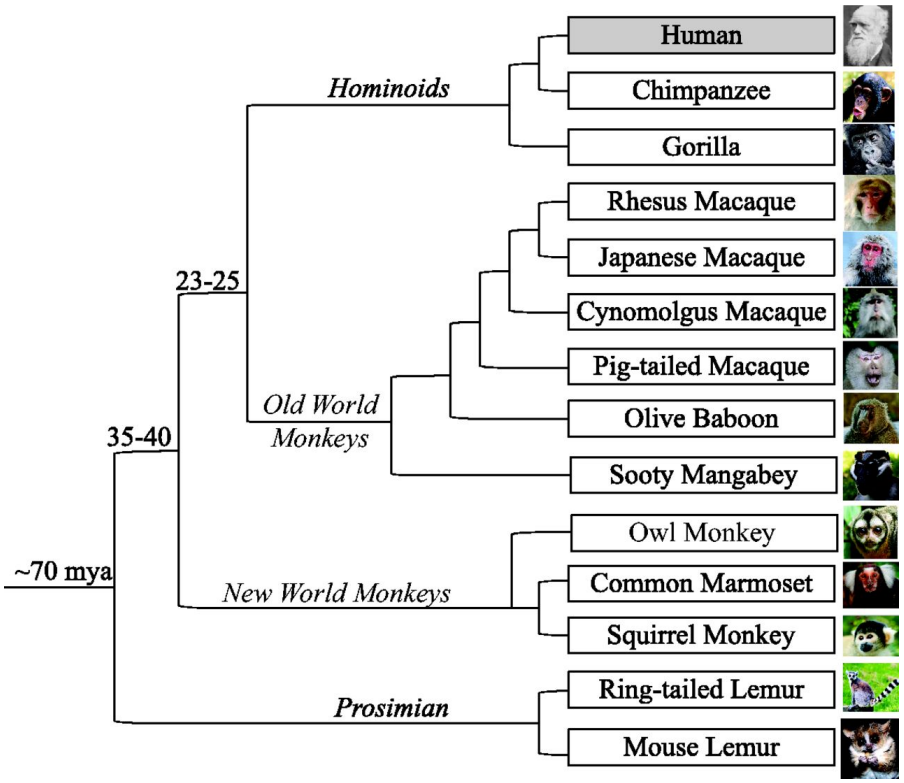


**Figure 2.1:** Species of the NHPRTR. Animals were chosen to represent large evolutionary distances, encompassing hominoid, Old World and New World Monkeys and prosimians. Two geographic subspecies were included for each of the following species: rhesus macaques (Indian-origin and Chinese-origin) and cynomolgus macaques (Mauritian-origin and Indonesian-origin).

evolution from other research organizations. Based on these discussions, 13 species of NHPs were chosen for transcriptome characterization (**Figure 2.1**), with selection emphasizing their use in important biomedical models, evolutionary diversity and the status of genome sequencing. The particular importance of NHP models for studies of AIDS pathogenesis and vaccines, respiratory disease models, metabolic disorders and neurobiology led to the inclusion of multiple Macaca species, as well as geographic subspecies for the rhesus macaque and the cynomolgus macaque due to phenotypic differences noted for these regional variants. For these 15 species/subspecies, the goal for the initial sequencing effort was to capture a maximum diversity of transcripts for any one species, thereby providing a breadth of evidence for annotating transcriptionally active regions (TARs) of the respective genomes. To accomplish this, a list of 21 relevant tissues was determined that covered the range of physical and functional compartments of the animals (cf. **Figure 2.2**) and then a centrally coordinated effort was undertaken to obtain the tissues from various institutions (see 'Materials and

Methods' section; contributing institutions are listed in 'Acknowledgments' section). For each species, RNA was isolated from the available tissues (with the exclusion of blood samples) and equal masses of RNA were combined to prepare the reference RNA sample that was used to generate the sequence data. (Blood RNA was not included in the general RNA composite due to the high abundance of hemoglobin RNA in such samples; therefore blood RNA will be the subject of a separate sequencing effort.) To improve functional genomics annotation for NHPs, we employed multiple methods of library preparation [14, 15], thereby generating RNA-based next-generation sequencing (RNA-Seq) data characterizing coding and non-coding transcripts, delineating information on strand-specificity and enabling accurate detection of antisense transcription (**Figure 2.2**). We have named this effort the 'NHP reference transcriptome resource' (NHPRTR; online at: http://nhprtr.org), intending it to provide the community with the sequence data from the composite RNA samples and with access to derived results (processed reads, alignments, assemblies) as these become available from our own efforts as well as from others who are contributing to this central resource. Though some limited amount of NHP transcriptomic data exists [16, 17], no studies or databases exist across both a large number of species and tissues, thus making the NHPRTR the most comprehensive database of primate transcriptomic information that is publicly available. Importantly, the NHPRTR is directly linked to our sample bank resource and we can provide purified RNA for the individual tissues from the species included in the resource, depending on availability.

## 2.2 Results

### 2.2.1 Summary of primary and processed data

Our current data set contains 40.5 billion 100 nt reads from 21 tissues across 13 primate organisms (**Table 2.1**), with the majority of our data coming from 100 $\times$ 100 paired-end (PE) reads from the Illumina HiSeq2000. From the home page at http://nhprtr.org (**Figure 2.3**), our resource site is designed to provide easy access to many resources, including pages that describe the overall goals of the project, its current status, contact information, external links and also the link to the data page. The data page hosts all of the raw data from the sequencing of the various species and each of their library preparations, with a file name that repre-

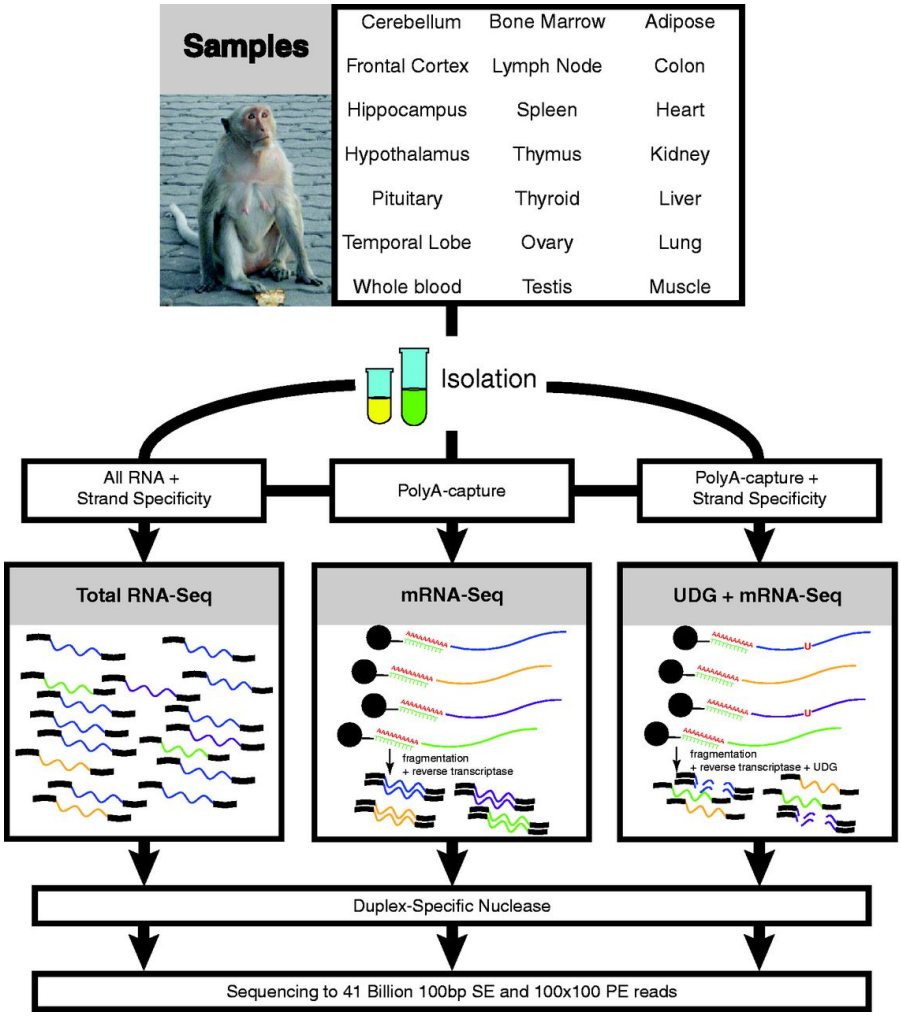


**Figure 2.2:** Tissue sources and methods for library construction and sequencing. (Top) The tissues being sequenced (top) cover 21 regions that focus on the brain, immunological and sexual tissues as well as general tissues important for pharmacogenomics. The majority of the individual tissues will eventually be sequenced as individual libraries to examine tissue-specific expression patterns. (Middle) Three different biochemical techniques were used for preparing cDNA libraries to enable the broadest examination of the transcriptome for each species. We used an RNA-ligation method for all RNA species (Total RNA-Seq), poly-A enriched cDNA synthesis (mRNA-Seq) and another version of mRNA-Seq that maintained the Watson or Crick strand of origin for the transcript by using dUNTPs during second strand synthesis (UDG). (Bottom) All cDNA libraries were then subjected to a DNA normalization step using Duplex-specific nuclease treatment, and then all samples were clustered, sequenced, and processed using standard Illumina methods and materials, generating 41 billion reads.

**Table 2.1:** Summary of current data in NHPRTR. The 40.5 billion reads span three different library preparation methods and two sequencing instruments (GAIIx and the HiSeq2000)

| **Species** | **File size (GB)** | **HiSeq2000 (2 × 100 nt paired-end reads) GAII (100 nt single-end reads)** | | | |
|---|---|---|---|---|---|
| | | **Protocol** | **Number of read pairs** | **Protocol** | **Number of reads** |
| Baboon | 973 | mRNA-seq | 955,573,799 | mRNA-seq | 71,477,607 |
| | | UDG mRNA-seq | 918,735,897 | UDG mRNA-seq | 67,763,503 |
| | | | | Total RNA-seq | 151,524,634 |
| Chimpanzee | 94.9 | Total RNA-seq | 198,954,000 | | |
| | 399.1 | UDG mRNA-seq | 836,864,082 | | |
| Cynomolgus Macaque Indochinese | 948 | mRNA-seq | 923,307,160 | mRNA-seq | 72,016,960 |
| | | UDG mRNA-seq | 894,367,594 | UDG mRNA-seq | 63,820,198 |
| Cynomolgus Macaque Mauritian | 656 | Total RNA-seq | 206,526,535 | | |
| | 422.6 | UDG mRNA-seq | 886,261,413 | | |
| Japanese Macaque | 986 | mRNA-seq | 942,269,530 | mRNA-seq | 77,740,433 |
| | | UDG mRNA-seq | 943,158,996 | UDG mRNA-seq | 72,925,864 |
| | | | | Total RNA-seq | 181,184,542 |
| Marmoset | 128.8 | Total RNA-seq | 269,969,905 | | |
| | 418.9 | UDG mRNA-seq | 878,369,246 | | |
| Mouse Lemur | 97.5 | Total RNA-seq | 204,494,231 | | |
| | 378.9 | UDG mRNA-seq | 794,659,816 | | |
| Pig-tailed Macaque | 951 | mRNA-seq | 867,009,248 | mRNA-seq | 54,292,043 |
| | | UDG mRNA-seq | 991,993,458 | UDG mRNA-seq | 54,668,320 |
| | | | | Total RNA-seq | 131,548,564 |
| Rhesus Macaque Chinese | 700 | mRNA-seq | 644,468,744 | mRNA-seq | 77,142,089 |
| | | UDG mRNA-seq | 661,177,666 | UDG mRNA-seq | 75,142,089 |
| | | | | Total RNA-seq | 121,570,595 |
| Rhesus Macaque Indian | 1331.2 | mRNA-seq | 1,716,083,364 | mRNA-seq | 84,892,037 |
| | | UDG mRNA-seq | 704,493,397 | UDG mRNA-seq | 70,346,332 |
| | | | | Total RNA-seq | 168,710,995 |
| Ring-tailed Lemur | 104.8 | Total RNA-seq | 219,647,886 | | |
| | 398.7 | UDG mRNA-seq | 835,972,568 | | |
| Sooty Mangabey | 106 | Total RNA-seq | 222,192,568 | | |
| | 424.4 | UDG mRNA-seq | 889,864,522 | | |
| Total | 9618.3 | | 18,667,116,290 | | 2,112,066,539 |
| | | | | **Total number of reads** | **39,446,299,119** |

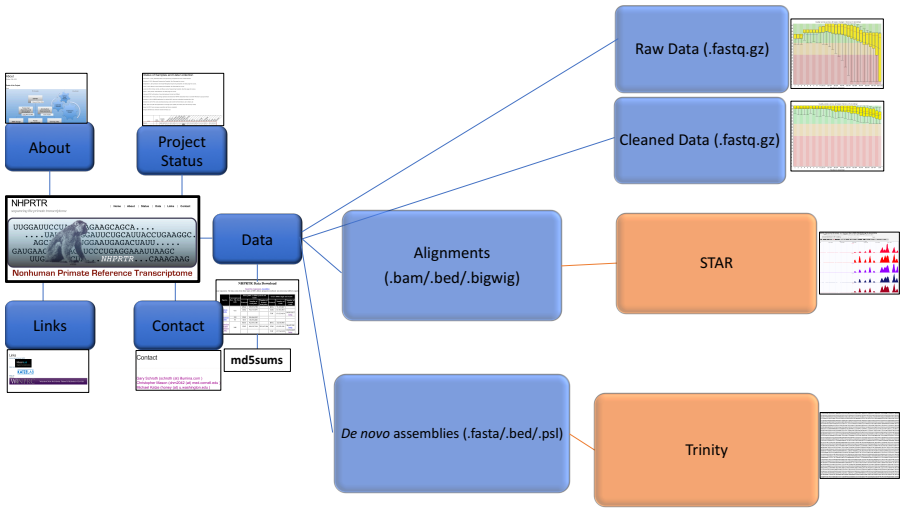


**Figure 2.3:** Organization of the NHPRTR. We have designed our database and the interface page to give users a clear sense of the goals, organization and data types present. Pages include the project background (about.html), the latest updates (status.html), connections to other sites (links.html), contact information (contact.html) and the data from the NHPRTR (data.html, including md5sums). From the data page (middle), users can access the raw data or various forms of the processed data, including cleaned data (hosted on NCBI's Sequence Read Archive), alignments (using STAR) and assemblies (Trinity). The alignments and assemblies can be viewed on our UCSC Genome Browser Mirror (http://lp364.genome-mirror.cshl.edu). The data page will continually update as data are submitted and as work is completed.

sents the provenances of the data generation. For example, the PE reads sequences from the Baboon UDG library called 'HCT20960' sequenced on lane five, appear as BAB_UDG_HCT20960_L005_R1.fastq.gz and BAB_UDG_HCT20960_L005_R2.fastq.gz. Finally, under each set of data, we have posted md5sums of each of the files, so users can readily confirm their accurate receipt of the data after download.

Our primary data analysis and quality checking have shown that our data are of very high quality (**Figures** 2.4 and 2.5), with a median Quality Score (Q-score) consistently >34 (>99.95% accuracy) across the length of the reads. Also, we used the tool Stitch [18] to check the overlap of the reads and found that the insert size of the cDNA libraries were within the expected range of 140-160 bp, since the mode of the distribution of the overlap of the PE (100 $\times$ 100) libraries was near 40-60 (**Figure** 2.6). Finally, the data distribution page also provides the output files from the FASTQC toolkit, to allow a deeper examination of the read statistics and qualities [19].

Once a species is sequenced and quality checked, the NHPRTR site also hosts a second version of the primary data. This second set is a 'cleaned version' which is

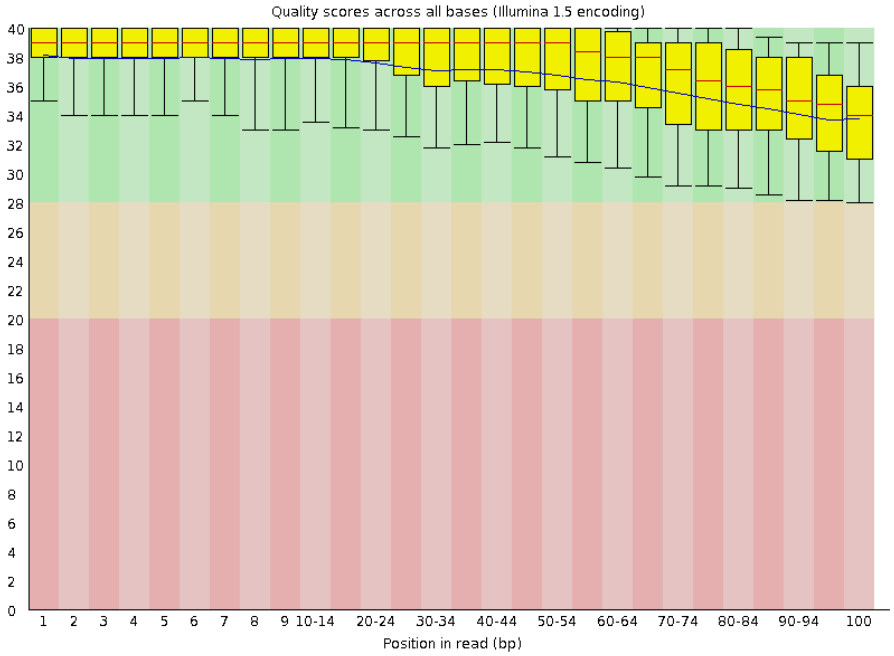


**Figure 2.4:** High accuracy base-calling across the length of the read. We used boxplots to show the distribution of quality scores from base 1-100, and we consistency observe base-calling accuracy above Q30 (>99.9%), with the median never falling below Q34 for the 1x100 runs on the GAIIX and the 2x100 runs on the HiSeq2000 . Representative plot is shown from library HCT20763 after read cleaning.

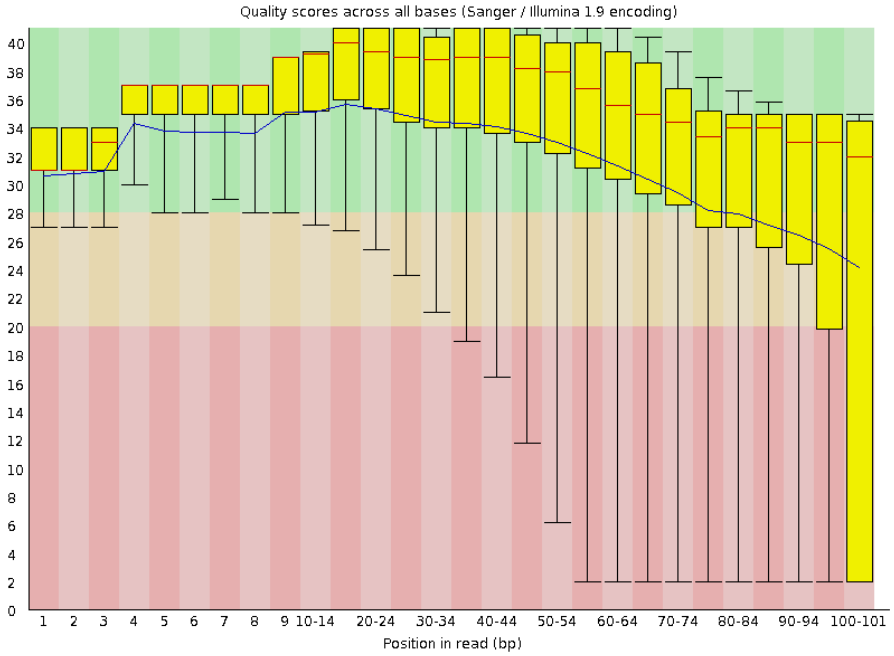


**Figure 2.5:** Quality plot is from library HCT20763 before read cleaning.

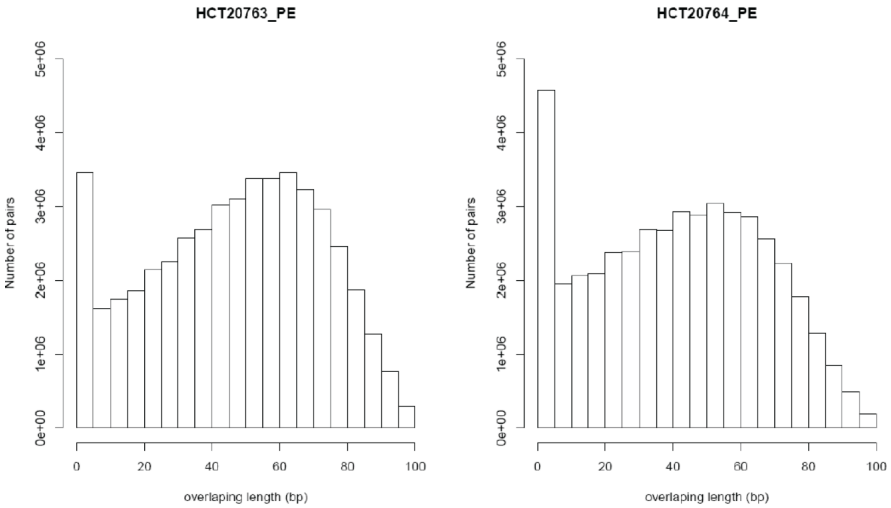


**Figure 2.6:** Bioinformatic confirmation of the estimated cDNA insert size. We used the program Stitch to estimate the overlap of bases for the 100x100 PE libraries, and we observed overlaps on average of 40-60bp, which confirms the expected insert size from the Agilent Bioanalyzer 2100 (140-160bp) on libraries 20763 and 20764.

generated for use in algorithms that are especially sensitive to sequence errors, such as *de novo* transcript assemblers and genome assemblers (**Figure** 2.7). We first trim all reads for low quality (<Q20), remove any remaining adapter sequences and any lengthy polyA/T stretches (>6 homopolymers) using Flexbar, in order to eliminate bad quality reads and the sequences from the ends of polyA tails or low complexity regions. We then align all reads to the known primate sequences for mtDNA and rRNA and exported these to a separate alignment files. We found that these steps remove between 3% and 10% of the data. These files can save significant time for researchers who want to begin with even higher quality data and who do not wish to focus on the mitochondrial or ribosomal sequences.

### 2.2.2 Alignments and data visualization

As the gene models for each species improve, it is often useful to gauge the state of these emerging data in relation to the best defined gene annotation set available– the human genome. To enable such work, the NHPRTR hosts an alignment to the human genome (hg19 and hg18) using Burrows-Wheeler Aligner (BWA) [20] and can all be readily viewed within the UCSC genome browser from a direct link on http://nhprtr.org. While we recognize that using the human genome as an alignment reference for distant phylogenetic species is not ideal, we still provide these alignments

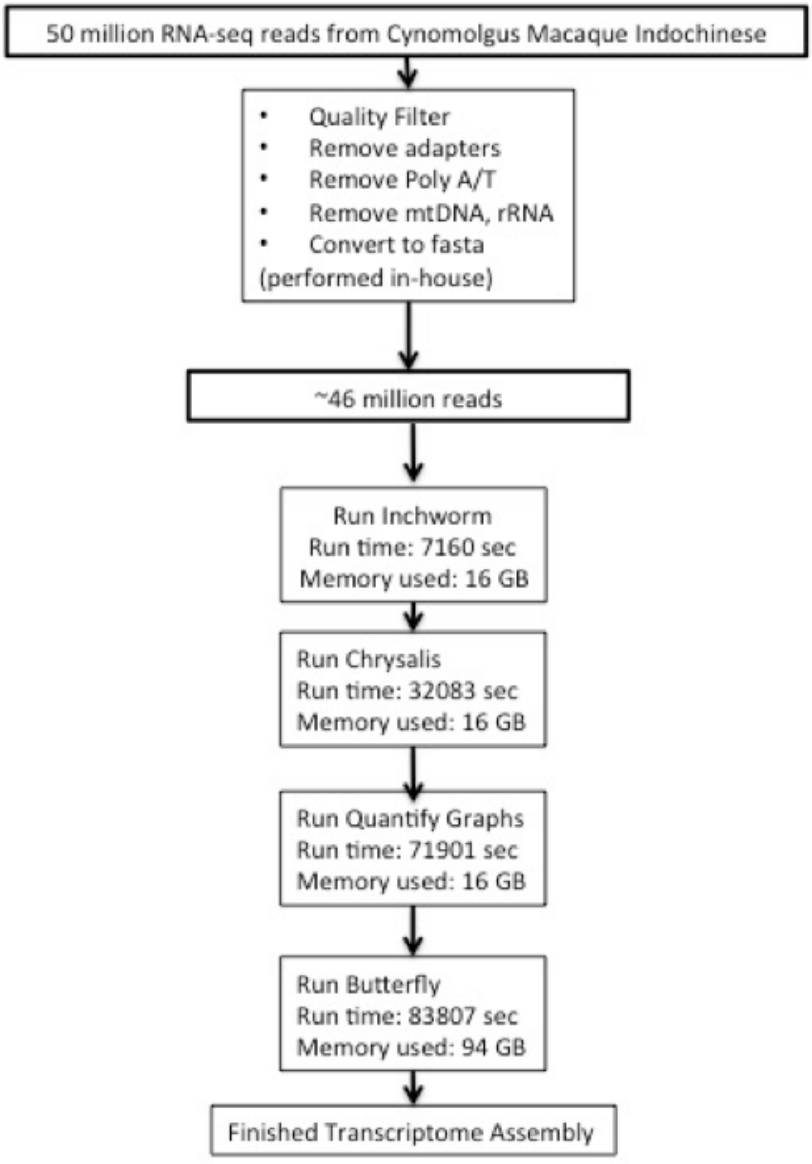


**Figure 2.7:** Pipelines for *de novo* transcript assembly. We used Trinity to test the utility of these data for *de novo* assemblers using 50M reads. We used our cleaning pipeline (methods) to filter the reads, and then we ran the three main steps of Trinity (Inchworm, Chrysalis, Butterfly), which required 120GB of RAM for a dataset of only 50M reads. These results highlighted the need to migrate these data to large-scale computing resources.

for several reasons. First, the human genome is the best annotated genome across primates and it hosts a wealth of other regulatory and functional data linked to the genomic coordinates. Second, the data can already be useful as a comparison of gene structures in expressed areas, placing genes in syntenic blocks and helping to define orthologous gene sets. Third, some species in the NHPRTR database have no genome yet sequenced. Finally, even though sequence divergence will decrease mapping rates, the alignments still provide a basic orthologous expression map across all species.

We observed that these human alignment data generated several immediate results. First, users can browse to any given human gene of interest and gauge the gene structure and rough expression level of that gene. Second, any hypothesized changes in gene structure, such as shortening, lengthening or splicing changes, can be visualized and compared to human structures. Third, the differences in the types of RNA-Seq can readily show the benefit of using multiple biochemical methods for the examination of a transcriptome (**Figure** 2.8). For instance, the detection of non-poly-adenylated transcipts such as snoRNAs or some histones can be readily seen in the Total RNA prep, whereas they are missing from the two mRNA preps.

### 2.2.3 Ongoing database work and analysis plans

As described here, this large-scale, EST-like resource of 13 species/subspecies of NHPs across 21 tissues is immensely useful for primate researchers, evolutionary biologists, immunologists and neurobiologists. With the addition of the Squirrel Monkey, Owl Monkey and other tissue-specific sequencing, we anticipate having ~100 billion reads from 15 primates when sequencing is completed in 2013. We plan to sequence individual tissues from the Indian-origin rhesus macaque from animals at different stages following SIV infection and also perform tissue-specific sequencing using different cDNA methods. Taken together, these data will create an unprecedented depth of expression and single-base resolution expression data for all of these species' tissues. Most significantly, the different types of biochemistries utilized for cDNA synthesis and RNA preparation for sequencing will create a broad, comprehensive profile (polyA and total RNA) of the transcriptome for each tissue and each species. Several ongoing analysis efforts from these data will be posted to the NHPRTR site, leveraging a variety of aligners and assemblers. First, as relevant published work in NHP transcriptomes appear [21], we will link to them on our site. Next, additional alignments from the AceView aligner [22] will be added, which will

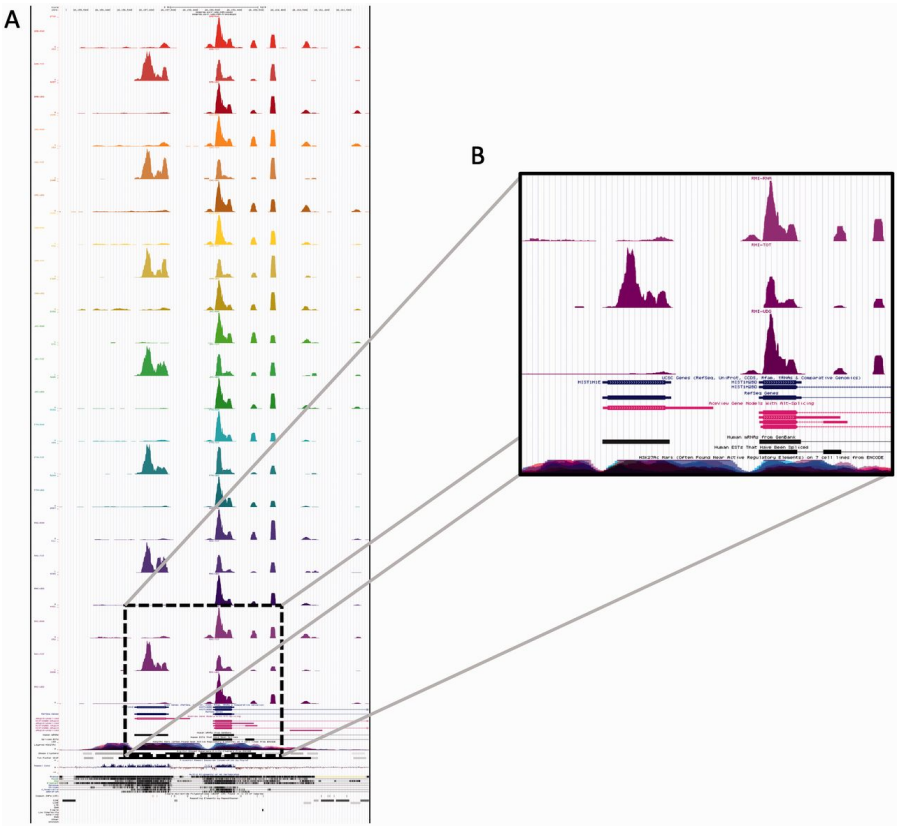


**Figure 2.8:** Browsable tracks. (**A**) We used BWA aligner to create cross-species maps of expression, based on the alignment to orthologous sequences of the human genome. Here we show the three library preparation methods (TOT, UDG, RNA), with one in each track for seven species. (**B**) The insert shows how the Total RNA preparation method (middle expression track) can more readily discern non-poly-adenylated genes, such as the histone genes.

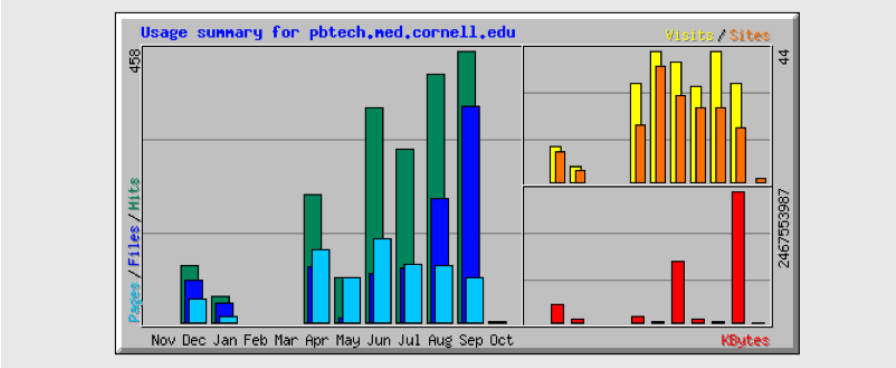


**Figure 2.9:** Web Statistics for NHPRTR Data Download Page. We used the webalizer (v2.01) program (http://www.webalizer.org/) to generate statistics about the site's download rates of the raw and processed data. We observed an increasing rate of hits and data downloads over the last year.

be very useful for defining the non-coding RNAs of the transcriptome and also the exon-intron junctions, along with the outputs from Tophat and BWA. We will also host the results of several *de novo* transcriptome assemblies as they are completed, based on the different libraries and available PE and single-end reads as described earlier. At time of submission, we are hosting preliminary *de novo* assembly results for the Mauritian cynomolgus macaque from pilot studies using Oases [23], TransAbyss [24] and Trinity [25], performed with subsets of the data (cf. Data under http://nhprtr.org). These assemblies are also linked to the reference genome for each species for species-specific browsing. Notably, our early efforts revealed that the construction of the *de novo* transcriptome assemblies can be a very memory-intensive process (**Figure 2.7**), which often required hundreds of gigabytes RAM. Thus, in an effort to help researchers utilize these data in large-scale computing environments, we are also hosting these data on the Blacklight 32TB memory node (blacklight.psc.teragrid.org) on the Extreme Science and Engineering Discovery Environment (XSEDE). We anticipate that having various means of accessing the primary, processed and assembled reads in multiple environments will ensure the broadest utilization of these data. In summary, we have designed the NHPRTR site to utilize familiar tools and formats from the genomics community and the combination of several library preparations and bioinformatic tools in the same resource have already created thousands of requests to examine and download these data (**Figures 2.9** and **2.10**). We encourage the use of the data by the community and will assist investigators in hosting their results if they wish to contribute to the resource and in reciprocity, we also have a section with links to data from other published primate RNA-Seq studies. Moreover, a main goal in

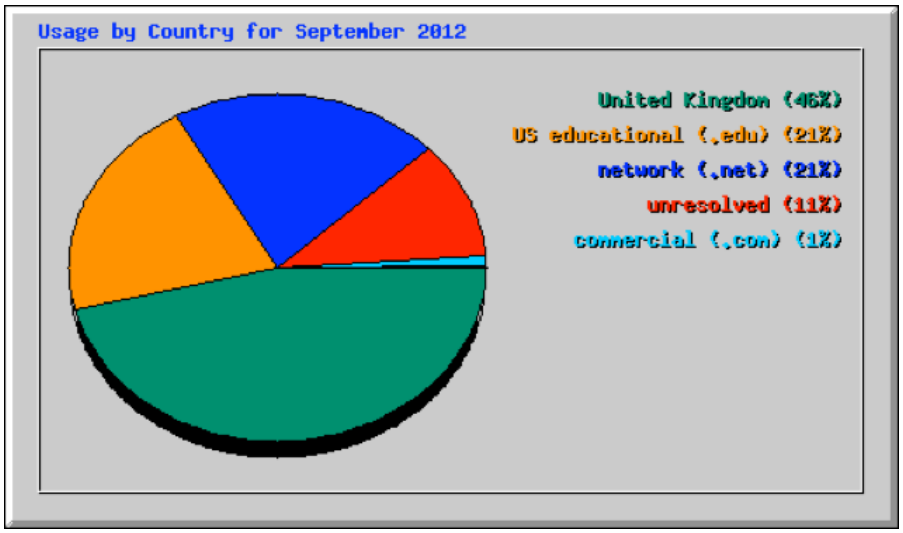


**Figure 2.10:** Web Statistics for NHPRTR Data Download Page. We observed traffic coming from sites in the United Kingdom (UK), U.S. educational sites, commercial sites (.com), and generate network sites (.net), as well as some unresolved IP addresses (11%).

generating these data was to provide a rich resource for species-specific alignment, thereby generating improved gene models for NHP genomes and this is realized by own ongoing efforts as well the gene annotation and prediction pipelines at ENSEMBL (B. Aken, personal communication). These data will also be helpful in answering a variety of questions pertaining to the complexity of transcriptome, including: new TARs, conservation/evolution of specific splicing sites, RNA editing events, UTR structures, and gene boundaries and content. In summary, the NHPRTR represents an immensely useful and timely addition to the genome sequences of these important species, a key hub for these species' RNAs and their matching transcriptomic data and an invaluable resource for genomes that will eventually be sequenced.

## 2.3 Materials and Methods

### 2.3.1 Tissue samples

Source tissues for the resource were generally obtained from animals that were being euthanized either for compassionate reasons due to failing health or as part of an existing research protocol; all veterinary procedures were approved under the local Institutional Animal Care and Use Committee (IACUC). Tissue specimens

were preserved in RNAlater®(Life Technologies) at the time of collection and frozen at −80°C. Tissues for gorilla, mouse lemur and ringtail lemur derived from frozen specimens previously collected at the time the individual animals were euthanized; these tissues were either transferred into RNAlater or homogenized in TRIzol® Reagent (Life Technologies) and frozen at −80°C. All frozen samples were shipped to the University of Washington and the RNA isolated under a standard protocol using TRIzol extraction and purification with RNeasy® columns (QIAGEN). Isolated RNA was characterized by absorbance spectroscopy to ensure the absence of contamination by protein or phenol and then analyzed by capillary electrophoresis to furnish an RNA Integrity Number (RIN) using an Agilent Bioanalyzer®. RNA concentrations for individual tissue RNA samples were based on integrated fluorescence intensity in the Bioanalyzer runs, calibrated against an RNA standard. For a species or subspecies, the reference sample combined equal masses of RNA from all the tissues. The number of available tissues varied among the species; whenever possible tissues were used from a single female individual and only was obtained from a second individual (**Figure 2.11**). The final composition of each reference sample as well as the RIN value for the individual tissue RNA components is available at the resource website (http://nhprtr.org).

### 2.3.2 Library preparations

Three different types of sequencing libraries were prepared from the reference samples. These were as follows: (i) non-directional mRNA-Seq, (ii) directional mRNA-Seq based on dUTP strand-marking and (iii) directional Total RNA-Seq, based on RNA-ligation to the initial RNA fragments which preserves their strandedness. In all the cases, the initial cDNA library was 'normalized' using a Duplex-Specific Nuclease Protocol (DSN) which removes high-abundance transcripts such as ribosomal molecules that would otherwise dominate the reads from the Total RNA-Seq libraries. The majority of sequencing for all species was done on the Illumina HiSeq2000 at Illumina or Weill Cornell Medical College (WCMC), with additional GAIIx sequencing performed at Illumina.

Supplemental Table 1. Listing of species and summary of tissues used to prepare the RNA pools used for the described RNAseq measurements. For each species equal masses of the purified tissue RNA samples were combined to generate the composite RNA pool that was used for both mRNAseq and total RNAseq.

| *Species* | Common Name | Adipose Fat | Bone Marrow | Brain cerebellum | Brain frontal cortex | Brain hippocampus | Brain hypothalamus | Brain temporal lobe | Colon | Heart | Kidney | Liver | Lung | Lymph node | Ovary | Pituitary | Skeletal Muscle | Spleen | Testis | Thymus | Thyroid |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Papio anubis* | Baboon | X | X | X | X | X | | X | X | X | X | X | X | X | X | X | X | Y | X | X |
| *Pan troglodytes* | Chimpanzee | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | X | X |
| *Macaca fasicularis* | Cynomolgus Macaque Indochinese | X | X | X | X | | | X | X | X | X | X | X | X | X | X | X | Y | X | X |
| *Macaca fasicularis* | Cynomolgus Macaque Mauritian | X | | X | X | X | | X | X | X | X | X | X | X | X | X | X | Y | X | X |
| *Gorilla gorilla gorilla* | Gorilla | X | | X' | X' | | | | X | X' | X | X | X | X' | X' | | X | X | | | |
| *Macaca fuscata* | Japanese Macaque | X | X | X | X | | X | | X | X | X | X | X | X | X | X | X | Y | X | X |
| *Callithrix jacchus* | Marmoset | X | X | | X | | | | X | X | X | X | X | X | X | X | X | | X | X |
| *Microcebus murinus* | Mouse Lemur (gray) | | | Y | Y | Y | | Y | Y | Y | Y | Y | Y | | | | Y | Y | Y | | |
| *Aotus vociferans* | Owl Monkey | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | X | X |
| *Macaca nemistrina* | Pig-tailed Macaque | X | X | X | | X | X | X | X | X | X | X | X | X | X | X | X | Y | X | X |
| *Macaca mulatta* | Rhesus Macaque Chinese | X | X | X | | | X | X | X | X | X | X | X | X | X' | X | X | Y | X | X |
| *Macaca mulatta* | Rhesus Macaque Indian | X | X | X | X | | X | | X | X | X | X | X | X | X' | X | X | Y | X | X |
| *Lemur catta* | Ringtail Lemur | | | X | X | X | X | X | X | Y | Y | Y | Y | | | | Y | Y | X | | |
| *Cercocebus atys* | Sooty Mangabey | | Y | Y | Y | Y | | | Y | Y | Y | Y | Y | Y | | Y | Y | Y | Y | Y | Y |
| *Saimiri boliviensis boliviensis* | Squirrel Monkey | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | | X |

"X" or "Y" in cell respectively indicate if donor animal was female or male. Additional donors are indicated by prime superscript.

All tissues were obtained from either tissue distribtuion programs of National Primate Research Centers or were pre-existing archival samples from research institutions.

Tissue specimens were preserved in RNAlater or had been previously homogenized in Trizol, and stored at -80˚C until the RNA was purified.

**Figure 2.11:** Listing of species and summary of tissues used to prepare the RNA pools used for the described RNA-Seq measurements. For each species equal masses of the purified tissue RNA samples were combined to generate the composite RNA pool that was used for both mRNAseq and total RNAseq. "X" or "Y" in cell respectively indicate if donor animal was female or male. Additional donors are indicated by prime superscript. All tissues were obtained from either tissue distribution programs of National Primate Research Centers or were pre-existing archival samples from research institutions. Tissue specimens were preserved in RNAlater or had been previously homogenized in Trizol, and stored at -80$^{\circ}$C until the RNA was purified.

### 2.3.3 Standard mRNA-Seq protocol

The standard mRNA-Seq library preparations were done using established Illumina methods for mRNA-Seq (Part #RS-100-0801). Briefly, poly A+ RNA is purified from 100 ng of total RNA with oligo-dT beads. Purified mRNA is then fragmented with divalent cations at elevated temperature. First strand cDNA synthesis is performed with random hexamer priming and reverse transcriptase. Second strand cDNA synthesis is performed using RNAseH and DNA PolI. Following cDNA synthesis, the double stranded products are end repaired, followed by addition of a single 'A' base and then ligation of the Illumina PE adaptors. For this study, the ligation products were purified using gel electrophoresis. The target size range for these libraries was ~250 bp on the gel such that the final library for sequencing would have cDNA inserts with sizes of ~150 bp long. Following gel purification the adapter ligated cDNA is then amplified with 15 cycles of PCR. This initial library was then subject to DSN normalization and additional rounds of PCR as described below for the Total RNA-Seq protocol.

### 2.3.4 Directional (UDG) mRNA-Seq protocol

The directional mRNA-Seq library preparations were done using the variant offered by Illumina (Part # RS-122-2303). Briefly, poly A+ RNA is purified from 100 ng of total RNA with oligo-dT beads. Purified mRNA is then fragmented with divalent cations under elevated temperature. First strand cDNA synthesis is performed with random hexamer primers and reverse transcriptase. Second strand cDNA synthesis is performed using RNAseH, dATP, dCTP, dGTP, dUTP and DNA PolI. Following cDNA synthesis, the products are end repaired, a single 'A' base is added and then the Illumina PE adaptors are ligated on to the cDNA products. The libraries are then amplified with 15 cycles of PCR as before, except in this case the strands that contain dUMP do not amplify and thus the products of the PCR process retain the original strand information. For this study, the ligation products were purified using gel electrophoresis. The target size range for these libraries was ~300 bp on the gel such that the final library for sequencing would have cDNA inserts with sizes of ~200 bp long. This initial library was then subject to DSN normalization and additional rounds of PCR as described below for the Total RNA-Seq protocol.

### 2.3.5 RNA-ligation-based directional total RNA-Seq protocol with DSN

The directional Total RNA library is constructed with a modified version of the Illumina directional mRNA-Seq sample preparation protocol, however no poly-A selection is used in a Total RNA-Seq prep. Briefly, 100 ng of total RNA is fragmented with divalent cations under elevated temperature. The ends of the fragmented RNA are treated with phosphatase to remove all 5′- and 3′-phosphate groups, followed by modification with polynucleotide kinase. This process insures that every RNA molecule contains a 5′-mono-phosphate group and a 3′-hydroxyl group. A pre-adenylated oligo is then ligated to the 3′-end of these RNA fragments, followed by the ligation of an RNA oligo to the 5′-end of the RNA. Following ligation of these adapter oligos, the RNA is reverse transcribed and amplified with 15 cycles of PCR to create the initial RNA-Seq library. Ribosomal RNA depletion from the initial RNA-Seq library is carried out following Illumina's published protocol. Briefly, 100 ng of amplified PCR products are denatured at 94°C for 5 min in 1 × hybridization buffer (50 mM HEPES, 0.5 M NaCl) followed by incubation at 68°C for 5 h; then 2 U of the DSN Enzyme (available from Evrogen) is added at 68°C for 25 min to digest double stranded DNA. Following DSN digestion the remaining undigested, single-stranded molecules are enriched with 15 more cycles of PCR.

### 2.3.6 Alignment methods

We used several alignments strategies on the data, with an initial focus on the alignment of the various species on the human genome. For an extremely conservative view of cross-species mapping, we used the BWA [20] and removed any sub-optimal matches (X0==1) and also removed any reads that were one edit distance away from mapping somewhere else in the genome (X1==0) field. These parameters reduced issues with paralogs and segmental duplications. For a broader alignment method, we used the AceView Magic aligner [22] and Tophat [26] (default settings) to generate mapping rates for each library of each species. The Magic AceView aligner uses a compressed data format for rapid processing and then uses a seed-and-extend algorithm based on sequence complexity, boundary detection for splicing, a scoring matrix for alignment and a mapping hierarchy to assign the reads to the most likely

location in the genome. Specific bash commands and shell scripts that were used in the analysis are posted online at nhprtr.org and also in Supplementary Data.

## 2.4 Funding

National Institutes of Health Office of Research Infrastructure Programs [R24RR032341]; Washington National Primate Research Center [P51RR000166 to Katze Laboratory]; [1R01NS076465-02 to Mason Laboratory]; XSEDE super-computing cluster [MCB120116]; STRIDE Center for Systems and Translational Research for Infectious Diseases at the University of Washington; National Center for Biotechnology Information (NCBI) [to J.T.-M. and D.T.-M.]; Intramural Research Program of the NIH, National Library of Medicine [partial]. Funding for open access charge: NIH and NCRR grants. Conflict of interest statement. None declared.

## 2.5 Acknowledgements

We would like to thank Illumina, Inc. for contributing almost all of the resources and reagents needed for completing the sample preps, sequencing and primary data analysis. Many people at Illumina helped create the libraries and sequencing data but we would especially like to recognize the efforts of Shujun Luo, Irina Khrebtukova, David Silva, Cindy Chen, Robin Li and Hang Pham. Tissues were obtained as research resources from the following centers: Washington National Primate Research Center, Wisconsin National Primate Research Center, Oregon National Primate Research Center, Yerkes National Primate Research Center, Southwest National Primate Research Center, the Duke University and the Duke Lemur Center; the Keeling Center for Comparative Medicine and Research, the North Carolina Zoo and Covance Inc. The Weill Cornell Medical College Epigenomics Core Facility provided support for use of their sequencing machines and technical assistance during sequencing. Finally, we would like to thank Bronwen Aken, Paul Flicek and Steve Searle from ENSEMBL for coordination of processed data also on their site.

# References


1. Perry, G. H. *et al.* A genome sequence resource for the aye-aye (Daubentonia madagascariensis), a nocturnal lemur from Madagascar. *Genome biology and evolution* **4,** 126–135 (2011).
2. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486,** 527–531 (2012).
3. Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69 (2005).
4. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483,** 169–175 (2012).
5. Babbitt, C. C., Tung, J., Wray, G. A. & Alberts, S. C. Changes in gene expression associated with reproductive maturation in wild female baboons. *Genome biology and evolution* **4,** 102–109 (2011).
6. Gibbs, R. Rhesus Macaque Genome Sequencing and Analysis Consortium: The rhesus macaque genome sequence informs biomedical and evolutionary analyses. *Science* **316,** 222–234 (2007).
7. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature biotechnology* **29,** 1019–1023 (2011).
8. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).
9. Gontan, C. *et al.* RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. *Nature* **485,** 386–390 (2012).
10. Ahfeldt, T. *et al.* Programming human pluripotent stem cells into white and brown adipocytes. *Nature cell biology* **14,** 209–219 (2012).
11. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419 (2010).
12. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22,** 1775–1789 (2012).

13. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees (1975).

14. Luo, S., Smith, G. P., Khrebtukova, I. & Schroth, G. P. Total RNA-Seq: Complete Analysis of the Transcriptome Using Illumina Sequencing-By-Synthesis Sequencing. *Tag-Based Next Generation Sequencing*, 367–381 (2012).

15. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* **7,** 709–715 (2010).

16. Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome research* **22,** 602–610 (2012).

17. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

18. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18,** 821–829 (2008).

19. Martínez-Alcántara, A. *et al.* PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* **25,** 2438–2439 (2009).

20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

21. Babbitt, C. C. *et al.* Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. *Genome Biology and Evolution* **2,** 67–79 (2010).

22. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology* **7,** S12 (2006).

23. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–1092 (2012).

24. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7,** 909–912 (2010).

25. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29,** 644–652 (2011).

26. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

# Chapter 3

# Building primate annotations *de novo* and their processing

The annotations available for non-human primate genomes are extremely poor in comparison to the human genome. For example, there are only 72 refSeq genes for gorilla. In addition to this, the genome build quality for non-human primates is extremely variable. In some species (*i.e.*, rhesus macaque), the genome is well built (and now in its 8th version) because of the wide-spread use of rhesus macaques as an animal model for disease testing. However, in other species (*i.e.*, squirrel monkey) the genome is completely in scaffolds. Additionally, for another species in our dataset (*i.e*, ring-tailed lemur), the nearest genome diverged >20 million years ago. Thus, in order to comparatively study the alternative splicing changes between species, we decided to build transcripts by *de novo* assembly. Our objectives were to create the most accurate *de novo* assembly possible with the most efficiency (time and space) and as few computational service units (cores $\times$ compute hours) as possible. We find that the best transcriptome assemblies are the result of the most input reads and are assembled without normalization. This chapter describes the pipeline for creating transcripts *de novo* and their accuracy. It describes our collaboration with XSEDE and using the Trinity software on supercomputers for billions of RNA-Seq reads as input. Portions of this chapter appears in Cougar et al. (2013).

## 3.1 Testing for the best quality *de novo* assembly

*De novo* assembly of RNA-Seq with Trinity with billions of reads require the use of computational resources that are beyond the capability of most research groups. Therefore, Trinity has made available an *in silico* normalization method, which has now become the default version of Trinity. *In silico* normalization greatly reduces the number of input reads for Trinity by probabilistically selecting reads based on median k-mer coverage value and targeted maximum coverage value. Developers of Trinity showed that normalization with just 10 million reads reconstructed full-length transcripts in *S. pombe* [1]. Since our dataset was on the order of 10 × greater than that dataset used by Haas et al. (2013), we wanted to test whether *in silico* normalization would be similar or improve our assemblies. We have found that although *in silico* normalization greatly reduces runtime by >200 × , the assemblies produced are not of the same quality as non-normalized assemblies. We compared assemblies built by normalization and non-normalization using the same 3.2 billion RNA-Seq reads from chimpanzee. **Table 3.1** shows the general statistics of the two assemblies. The

**Table 3.1:** Normalized vs. Non-normalized Trinity assembly statistics

| | **Non-normalized** | **Normalized** |
|---|---|---|
| Total length of sequence | 2,012,120,424 | 1,941,765,648 |
| Total number of sequences | 2,571,000 | 2,999,946 |
| Longest transcript | 33,952 | 36,038 |
| N25 | 3,201 | 3,362 |
| N50 | 1,350 | 1,564 |
| N75 | 532 | 599 |
| Total GC count | 866,045,003 | 823,920,621 |
| GC% | 43.04% | 43,43% |

summary statistics of the normalized and non-normalized assemblies are similar. In fact, the normalized assembly even has a higher N25, N50, and N75 than the non-normalized assembly. However, the normalized assembly has almost 500,000 more transcripts than the non-normalized assembly, but has a smaller total length of sequence. It is more helpful to identify what is actually being assembled to assess the quality of the assemblies. **Figure 3.1** shows what parts of the genes are being

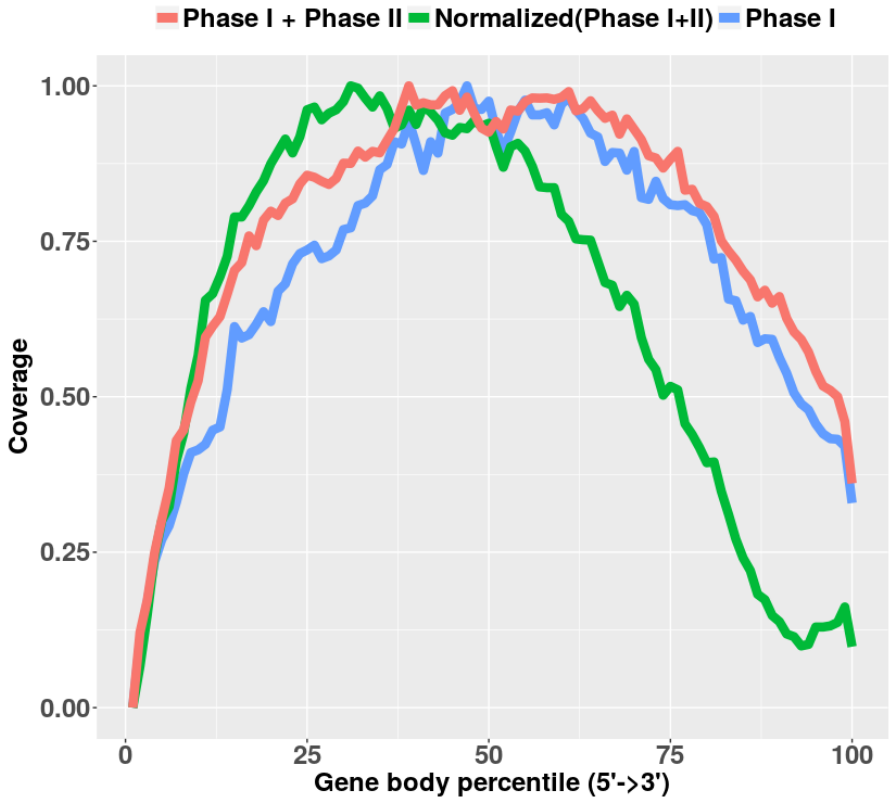


**Figure 3.1:** Comparison of contig coverage of the gene body with chimpanzee refGene annotation between Trinity assembly built with digital normalization and without. Phase I is the NHPRTR Phase I data, which includes the pooled RNA-Seq described in **Chapter 2**. Phase I + Phase II is the NHPRTR Phase I data and the NHPRTR Phase II tissue-specific RNA-Seq data. *In silico* normalization was performed on the Phase I and Phase II data.

assembled in normalized vs. non-normalized assemblies and **Figure 3.2** shows what fraction of genes are being fully assembled. We show the *de novo* assembly built with only Phase I data for scale. These results indicate a dramatic difference in the quality of the assembles being produced by each method. It shows that the normalized assembly is only able to assemble the beginning of the gene (starts to deviate from the non-normalized assembly halfway through the gene), and does a very poor job of assembling the end of the transcript. Additionally, the non-normalized assembly with Phase I and II data is able to recover the most full-length transcripts, which is important for identifying novel isoforms and accurate splice junctions. Adding the Phase II data results in a big improvement in the number of additional genes assembled. Thus, even though the normalization method greatly reduces the computational burden and runtime, we declined to use the *in silico* normalization method since we wanted to identify as many splice junctions as possible in all parts of the gene. Although the

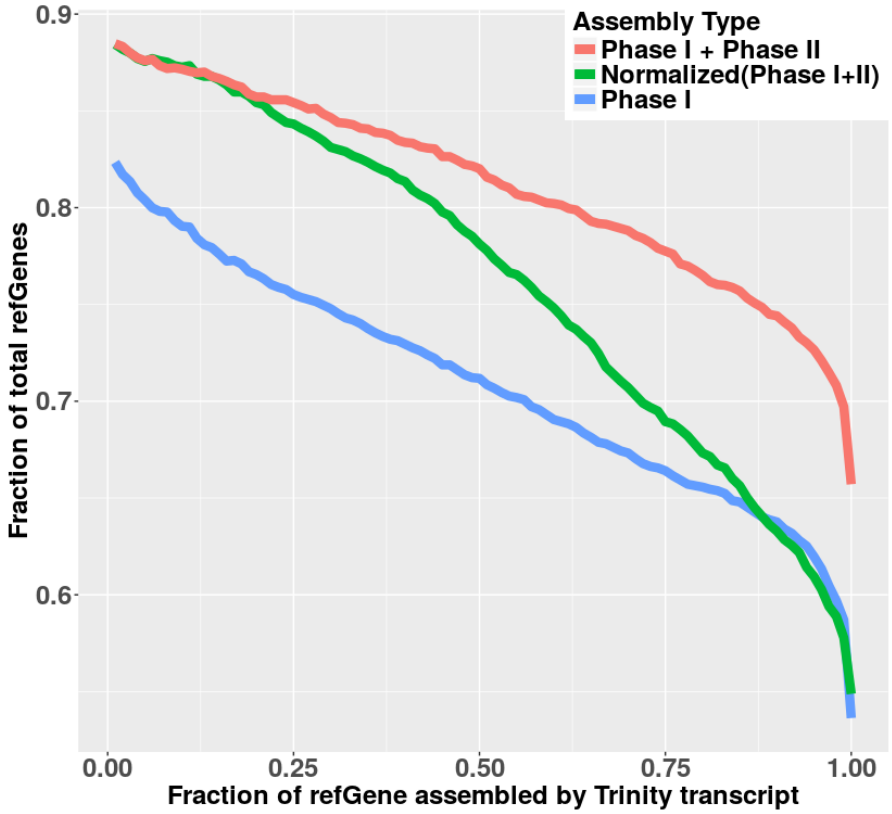


**Figure 3.2:** Comparison of fraction of Chimpanzee refGene assembled between Trinity assembly built with digital normalization and without. Trinity assemblies were aligned to panTro5 and fraction of refGene assembled was calculated by how much the contig overlapped with the refGene. Phase I is the NHPRTR Phase I data, which includes pooled RNA-Seq described in **Chapter 2**. Phase I + Phase II is the NHPRTR Phase I data and the NHPRTR Phase II tissue-specific RNA-Seq data. *In silico* normalization was performed on the Phase I and Phase II data

non-normalized method can take months at a time to assemble, we though it was necessary to produce accurate full-length transcripts in order to identify orthologous splicing events. The rest of this chapter discusses the building of *de novo* transcripts with Extreme Science and Engineering Discovery Environment (XSEDE) resources.

## 3.2 Enabling large-scale next-generation sequence assembly with Blacklight

A variety of extremely challenging biological sequence analyses were conducted on the XSEDE large shared memory resource Blacklight, using current bioinformatics tools and encompassing a wide range of scientific applications. These include genomic sequence assembly, very large metagenomic sequence assembly, transcriptome assembly, and sequencing error correction. The data sets used in these analyses included uncategorized fungal species, reference microbial data, very large soil and human gut microbiome sequence data, and primate transcriptomes, composed of both short-read and long-read sequence data. A new parallel command execution program was developed on the Blacklight resource to handle some of these analyses. These results, initially reported previously at XSEDE13 and expanded here, represent significant advances for their respective scientific communities. The breadth and depth of the results achieved demonstrate the ease of use, versatility, and unique capabilities of the Blacklight XSEDE resource for scientific analysis of genomic and transcriptomic sequence data, and the power of these resources, together with XSEDE support, in meeting the most challenging scientific problems. High-throughput, next-generation sequencing (NGS) of genomes [2, 3], transcriptomes [4], and epigenomes is currently in a phase of burgeoning growth with each passing development cycle yielding a greater than exponential return in the amount of quality sequence data generated per unit of cost (www.genome.gov/sequencingcosts). This rapid progress in data generation can currently create data sets within weeks, which are computationally intractable [5] for complete scientific analysis because of the large RAM footprint required or the volume of data to be analyzed. This limits their potential use in areas of scientific interest [6] and in translational medicine [7, 8].

The computational requirements of these data sets often exceed the capacity of personal computing systems, server-level infrastructure, distributed high performance computing, and large shared memory high performance computing systems. Hence,

many important scientific questions for which the data are or could be available either go unanswered or can only be addressed by a few research groups with a large bioinformatics infrastructure. Here we present science outcomes that highlight the ability of the cache coherent non-uniform memory access architecture of the XSEDE resource Blacklight, housed at the Pittsburgh Supercomputing Center (PSC), to allow efficient genomic analysis of data sets outside the scope of other high performance computing systems, as well as user-friendly high-throughput analysis of standard-sized to large-sized genomic data [9]. With these complementary capabilities, the XSEDE resource Blacklight extends the current technical limits of genomic and transcriptomic assembly for analyses requiring the largest shared memory systems, as well as the scope of genomic research by enabling high-throughput large shared memory analysis.

### 3.2.1 Blacklight

The Blacklight system at the PSC is an SGI Altix UV 1000 (SGI (Silicon Graphics, Inc.), Milpitas, CA 95035, USA) with two partitions, each containing 16 TB of cache coherent shared memory and 2048 cores. This means that a single application running on Blacklight can access up to 16 TB of shared memory using up to ~2000 cores. The obvious application of this system is for algorithms and problems that benefit from holding large amounts of data in RAM. However, the fast interconnect that facilitates cache coherent non-uniform memory access across the system also enables rapid communication within distributed memory applications. This dual nature of the system allows researchers to run problems across a continuum, from a single, massive shared memory application to many large shared memory applications running in parallel to fully distributed or embarrassingly parallel applications. Because the realm of genomic analysis encompasses all of these modes of computing, this flexibility makes Blacklight convenient and powerful for researchers dealing with diverse genomic analysis pipelines. In addition, because Blacklight is essentially one big system, running a single operating system, it is ideal for rapid prototyping of new serial and parallel algorithms for large data analysis.

### 3.2.2 Data generation with massive RNA-seq

In 2010, a committee of researchers set out to create a non-human primate reference transcriptome resource (NHPRTR) to help establish the genetic basis for phenotypic

differences observed between primates, including differences between humans and non-human primates (NHPs). Such a resource can provide valuable information regarding evolutionary processes, as well as insight into human health and disease from the pharmacogenomics work performed on the animal models for infectious disease and novel treatments. To provide a comprehensive resource, a committee of experts chose 13 primate species. Tissues samples were then taken from 21 tissues and next-generation sequencing of RNA (RNA-seq) was performed using three different approaches. The result was 40.5 billion 100 nucleotide reads that needed to be assembled into transcriptomes for each species and RNA-seq method used (13 species $\times$ 3 methods= 39 assemblies). Because most of these species do not have any reference genome, the transcriptomes must be assembled *de novo*. The details behind the motivation for this resource and the generation of the RNA-seq data are described in detail in the NHPRTR paper[10].

### 3.2.3 Enabling large-scale *de novo* transcriptome assembly with Trinity on Blacklight

As described in the NHPRTR paper, investigators found that assembling the nearly 2 billion reads required as input for these *de novo* transcriptome assemblies was beyond the capabilities of their local systems and even beyond the capacity of the program's initial estimates of large data inputs. At this point, they applied for an XSEDE allocation on Blacklight at the PSC, along with Extended Collaborative Support Services (ECSS) from XSEDE to help them perform these transcriptome assemblies of unprecedented size and scale using Trinity [4]. Through XSEDE's ECSS, PSC worked closely with the Trinity developers to harden Trinity on Blacklight and find the best way to run these massive assemblies.

To begin, PSC installed the latest, optimized version of Trinity contributed by the National Center for Genome Analysis Support (NCGAS) at IU, without which these assemblies would have taken several times longer to complete [11]. Even with this optimized version, challenges appeared immediately. While Blacklight had plenty of shared memory to handle the assemblies, at one point in the assembly, the Chrysalis phase [4], Trinity was creating and working on hundreds of thousands of files. Even very large assemblies of say, 600 million reads, while still producing tens of thousands of files, had no problem executing on the default Lustre filesystem, but the 2 billion read assemblies being attempted here produced too many files to be handled efficiently

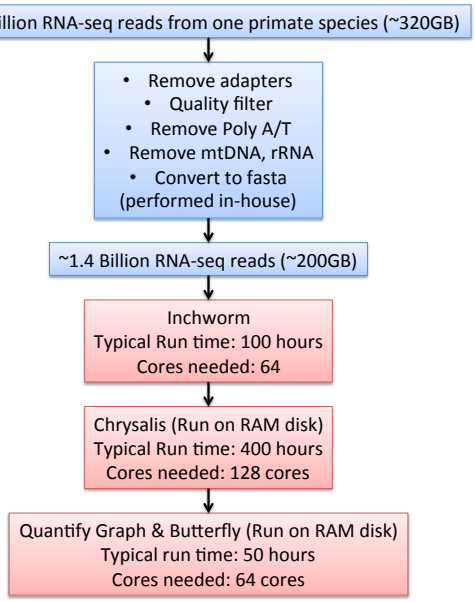


**Figure 3.3:** Walltime and core counts for various stages of the Trinity pipeline to assemble a single primate transcriptome. See Reference [4] for descriptions of the phases of Trinity.

by this filesystem. To work around this, PSC established a local filesystem attached directly to Blacklight. This alleviated the problem for a single massive Trinity assembly, but I/O-related slowdowns occurred with many massive assemblies running at once.

Finally, an ideal workflow was devised (**Figure 3.3**), utilizing Blacklight's RAM disk at the right points in the workflow to speed up the calculations, run many assemblies at once, and avoid problematic I/O issues but also avoid wastefully using RAM disk for large files where it was less beneficial. First, preprocessing of the data, a primarily serial task, was performed on the research group's local resources. The preprocessed data were then moved to Blacklight's Lustre filesystem, and the initial Inchworm stage of the assembly was performed entirely on the Lustre filesystem using 64 cores. For the Chrysalis stage, we introduced a modification to the Trinity code that allowed the Chrysalis directory to be given a different path from the rest of the Trinity working directory. This allowed the Chrysalis files to be created on RAM disk, while large files that did not need RAM disk remained on the Lustre filesystem. This phase generally required 128 cores (1 TB RAM) to provide extra memory resources to store files on the RAM disk associated with the job. After the Chrysalis phase was complete, the job script would back up the Chrysalis directory to the Lustre filesystem

but then continue to operate on those files in RAM disk for the final QuantifyGraph and Butterfly stages of Trinity. We found that using 64 threads on 64 cores for the QuantifyGraph and Butterfly stages and running from RAM disk provided optimal performance, reduces the runtime of those steps from a total of 250 h (when running from Lustre using 32 threads) to 50 h. Even after these optimizations, a significant amount of resources were needed, with a typical *de novo* assembly for one primate species with ~1.8 billion RNA-seq reads taking around 550 compute hours using 64-128 cores (35,200-70,400 service units). This *de novo* assembly method has proven successful, generating transcriptomes with a mean average size >2 kb for most RNA-seq methods used. Out of the 39 total assemblies required, 20 of the largest assemblies were performed on Blacklight over a period of a few weeks **Table 3.2**.

### 3.2.4 Characterizing the *de novo* assembled transcriptomes

In order to evaluate the accuracy of the *de novo* transcriptome assemblies, we determined the percentage that our assembly reconstituted the publicly available genome annotations (**Table 3.3**). Currently, only five NHP species from our data set have reference genomes: chimpanzee (panTro4), gorilla (gorGor3), rhesus macaque (rheMac3), marmoset (calJac3), and mouse lemur (micMur1). For species without reference genomes, we mapped to the nearest genome. For the gene models, we used annotations that were generated from native mRNA (RefSeq) and annotations that were computationally predicted (ENSEMBL). In most cases, the genes were either assembled >80% of their gene model or not assembled at all. In every NHP species with a reference genome, there was an improvement in assembling RefSeq genes compared with assembling ENSEMBL gene predictions. Most notably, we were able to recover >90% of RefSeq genes in rhesus macaque. The lack of coverage of genes in mouse lemur may represent the incompleteness of the draft genome, which lies entirely in scaffolds. The ability of Trinity to recover most of the known gene models shows that the *de novo* assemblies built for NHP species without a reference genome are good representations of the actual annotations. Because many studies are designed to be dependent on certain reference genomes, we provide transcriptome assemblies for many additional NHP species without reference genomes and also show that it is possible to accurately rebuild any species' transcriptome with high-coverage RNA-seq data. In addition to the known genes that were built by Trinity, we were interested in the assembled transcripts that were not currently annotated. We looked for evidence

**Table 3.2:** Summaries of 20 primate transcriptome assemblies completed on Blacklight.

| Species | Library | Number of input sequences | Number of contigs | Total Length (bp) | N25 (bp) | N50 (bp) | N75 (bp) | Longest contig (bp) |
|---|---|---|---|---|---|---|---|---|
| Baboon | TOT | 149,018,017 | 658,581 | 280,530,426 | 1,029 | 434 | 276 | 35,170 |
| Baboon | UDG | 1,543,247,564 | 1,131,951 | 844,127,360 | 3,534 | 1,368 | 479 | 131,395 |
| Chimpanzee | UDG | 1,465,013,009 | 987,615 | 1,433,298,968 | 6,356 | 3,806 | 1,666 | 47,873 |
| Cynomolgus Macaque (Chinese) | RNA | 1,675,591,113 | 911,282 | 864,011,637 | 4,769 | 2,316 | 706 | 59,032 |
| Cynomolgus Macaque (Chinese) | UDG | 1,629,828,914 | 990,604 | 1,055,391,869 | 5,479 | 2,822 | 875 | 122,916 |
| Cynomolgus Macaque (Mauritian) | RNA | 1,078,143,527 | 1,142,531 | 929,071,752 | 4,368 | 1,813 | 525 | 30,364 |
| Cynomolgus Macaque (Mauritian) | TOT | 166,350,980 | 526,723 | 199,771,360 | 657 | 377 | 266 | 36,976 |
| Cynomolgus Macaque (Mauritian) | UDG | 1,177,348,077 | 1,015,657 | 834,198,307 | 4,360 | 1,927 | 532 | 22,239 |
| Gorilla | UDG | 1,256,121,406 | 732,336 | 1,122,255,357 | 5,878 | 3,680 | 1,804 | 33,526 |
| Japanese Macaque (Indonesian) | RNA | 1,863,420,069 | 703,246 | 737,890,249 | 5,010 | 2,687 | 898 | 21,620 |
| Marmoset | TOT | 253,098,348 | 332,782 | 118,141,291 | 545 | 357 | 261 | 14,561 |
| Marmoset | UDG | 1,659,423,714 | 814,235 | 475,863,605 | 2,206 | 785 | 359 | 122,518 |
| Pig-tailed Macaque | RNA | 1,725,523,068 | 969,993 | 1,044,146,568 | 5,189 | 2,807 | 916 | 25,056 |
| Pig-tailed Macaque | UDG | 1,573,381,596 | 1,301,087 | 1,222,337,940 | 5,015 | 2,265 | 668 | 32,637 |
| Rhesus Macaque (Chinese) | RNA | 1,209,350,072 | 923,017 | 836,358,971 | 4,694 | 2,230 | 639 | 32,796 |
| Rhesus Macaque (Chinese) | UDG | 1,310,236,599 | 969,421 | 850,250,785 | 4,638 | 2,114 | 594 | 34,020 |
| Rhesus Macaque (Indian) | RNA | 3,200,476,713 | 703,246 | 737,890,249 | 5,010 | 2,687 | 898 | 21,620 |
| Rhesus Macaque (Indian) | UDG | 1,410,322,373 | 1,051,149 | 832,770,332 | 3,966 | 1,644 | 519 | 68,269 |
| Ring-tailed Lemur | UDG | 1,403,229,556 | 611,678 | 822,247,525 | 6,169 | 3,630 | 1,468 | 33,401 |
| Sooty Mangabey | UDG | 1,635,074,685 | 1,188,472 | 1,465,648,107 | 6,431 | 3,483 | 1,116 | 30,331 |

**Table 3.3:** Percentage of known (RefSeq) and predicted genes (ENSEMBL) covered by *de novo* assembled transcriptomes.

| Species (genome) | RefSeq genes covered >80% | Percentage of all RefSeq genes | ENSEMBL gene predictions covered >80% | Percentage of all ENSEMBL gene predictions |
|---|---|---|---|---|
| Chimpanzee (panTro4) | 1,888 | 77% | 18,928 | 68% |
| Gorilla (gorGor3) | N/A | N/A | 17,142 | 59% |
| Rhesus Macaque (rheMac3) | 5,519 | 91% | 14,290 | 57% |
| Marmoset (calJac3) | 124 | 73% | 18,247 | 56% |
| Mouse Lemur (micMur1) | N/A | N/A | 3,693 | 15% |

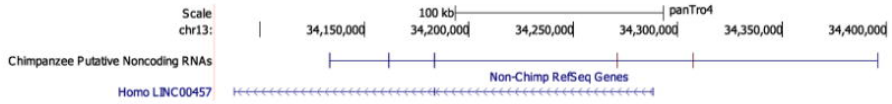


**Figure 3.4:** Putative novel noncoding RNA gene in chimpanzee on chromosome 13. This putative gene overlaps an exon from human long intergenic noncoding RNA gene 457.

of novel putative noncoding RNAs by filtering the assemblies for sequences that were in the current annotation and/or contained open reading frames that were 90 bp or longer. Noncoding RNA genes are RNA molecules that are transcribed but are not translated into proteins. Noncoding RNA genes have been implicated in many biological roles ranging from necessary components of protein translation (transfer RNAs) to major effectors of X inactivation (Xist) [12]. The abundance of long noncoding RNAs in NHP genomes remains unclear. In chimpanzee, we identified 4489 possible novel noncoding RNAs. **Figure 3.4** shows a putative novel noncoding RNA in chimpanzee that contains an exon from a human long intergenic non-protein coding RNA (LINC RNA), LINC00457. LINC00457 is primarily expressed in the human brain http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513.

### 3.2.5 Hosting a community resource

Now that the initial set of the massive *de novo* assemblies have been completed, the finished transcriptomes are being hosted on storage resources at the PSC so that the

community of researchers interested in these data can apply for an XSEDE allocation and use Blacklight or other XSEDE resources to further work with and analyze the data. The researchers working with NHPRTR are planning additional assemblies and cross-transcriptome alignments, for which the ready availability of these data on XSEDE will be most useful [10]. Lastly, because they have been encouraged by these results, the NHPRTR group of researchers has started a larger set of brain and region-specific deep RNA-sequencing of the 21 tissues across all the primates, which will create an additional 11 billion reads and an expanded resource for research in the NHPs, including evolutionary models, improved genome annotation across all primates (humans included), and improved models for infectious disease like HIV (using SIV) and AIDS.

### 3.2.6 Rapid Algorithm Development

One of the potential advantages of Blacklight's massive shared memory architecture is to enable scientists and developers to quickly prototype new parallel solutions to their research problems. As a simple example, when working with very large genomic data sets or other very large volume data, researchers often encounter circumstances where a heterogeneous group of large memory commands need to be executed expeditiously. While working with Trinity on Blacklight, a researcher and Trinity developer was able to quickly design a program to efficiently execute parallel commands that require large amounts of shared memory during the QuantifyGraph and Butterfly stages of Trinity. This program, Parafly, uses C++ and OpenMP to launch a large set of jobs with varying memory requirements, filling the need for a versatile parallel execution program within Trinity. Parafly accepts a flat file with the group of commands that a user wishes to execute for input, placing minimal requirements on the end user for operation. The program operates by loading all commands to be executed into an array data structure, assigning thread conditions to each command to be executed, then executes each command in parallel while logging the exit status of each command. Parafly was incorporated into the main Trinity code, and since then has been spun off as a separate project and extended to efficiently execute any group of tasks that require a large amount of shared memory per system thread. The Parafly resource can be found for download at http://parafly.sourceforge.net/.

### 3.2.7 Conclusion

Results have been presented comprising for high-throughput, high-memory assemblies of 20 primate transcriptomes. These advances are breaking new ground in their respective fields, and some, like the metagenome assembly and development of the NHPRTR would have been extremely difficult or impossible to do on any other system. While these diverse accomplishments highlight the power and flexibility of Blacklight's architecture for the assembly of NGS data, the research community is still becoming aware of these capabilities. As a result, Blacklight's potential to assemble the largest single organism genomes, or even larger metagenomic samples, has not yet been fully tested. As more groups engage with researchers who have benefitted from this resource, and engage with XSEDE through its ECSS and novel and innovative project programs, we expect demand to continue to grow, along with our ability to harness the full potential of available NGS data to solve the most challenging problems in computational biology.

### 3.2.8 Acknowledgments

This work was supported with funding from the National Institutes of Health (NIH), including R01HG006798, R01NS076465, R24RR032341, and the Starr Cancer Consortium grant number I7-A765 (Chris Mason); the Tri-Institutional Training Program in Computational Biology and Medicine and the National Science Foundation Graduate Research Fellowship Program under Grant No. NSF DGE-1144153 (Lenore Pipes). This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Blacklight system at the Pittsburgh Supercomputing Center (PSC).

## 3.3 Recent Assemblies

Since we saw a linear relationship in the number of input sequences used and the quality of the transcripts produced (**Figure 3.2**), we decided to pool both Phase I RNA-Seq data (presented in **Chapter 2**) and tissue-specific Phase II RNA-Seq data [13] to create a better transcriptome assembly. In addition to this, in order to have

an equivalent transcriptome assembly in humans, we decided to use GTEx RNA-Seq data [14] for tissues that were matched to the NHPRTR tissues to create a *de novo* assembly for humans. **Table 3.4** shows the summary statistics for the most recent assemblies with human and marmoset. Surprisingly, we were able to assemble the human *de novo* assemblies with GTEx RNA-Seq data with relative ease. In fact, one of our assemblies completed in only 9 days. The reason why the human assemblies finished so quickly is unclear. Typically, we perform a step-wise construction of the assemblies by separating the job into the four phases of Trinity (Inchworm, Chrysalis, QuantifyGraph, and Butterfly), and at the end of each phase, recursively copy back all of the millions of files produced in case the algorithm fails at any point or if the system fails at any point. The human assemblies did not require a step-wise construction, and ran as one single job. The human assemblies generated with >5 billion RNA-Seq reads (~2 billion reads greater than any other *de novo* assembly that we have attempted) but the data was unstranded paired-end 76bp while the NHPRTR data was stranded paired-end 100bp. We speculate that the read-length could make a difference in causing more complexity in the de Bruijn graphs, which might increase the run-time. In fact, we have attempted to create new assemblies using NHPRTR Phase I and NHPRTR Phase II data for almost all species, but currently we have only been able to complete two of these assemblies, chimpanzee and marmoset. The non-human primate assemblies are extremely difficult to complete although the exact reasons are unclear. Typically, our non-human primate assemblies require at least 1.5 months for completion if the job does not fail at some point (either system failure or algorithm failure). All non-human primate jobs have failed at some point in time because of system failure and/or algorithm failure.

## 3.4 Primate RNA-Seq Resources

On our private instance of the UCSC genome browser mirror, we have hosted all tissue-specific bigwigs and Trinity assembly annotation tracks aligned to their nearest genome. **Table 3.5** describes the genomes that are supported and the tissues that have bigwig tracks. For assemblies that are not yet available on UCSC (baboon-papAnu3, mouse lemur-micMur3, sooty mangabey-cerAty1), we have created track assembly hubs for viewing assemblies and tissue-specific data.

[1] from Brawand et al. (2011)[15]

**Table 3.4:** Summary statistics of recent *de novo* assemblies.

| Species | Number of input sequences | Number of contigs | Total Length (bp) | N25 | N50 | N75 | Longest contig (bp) |
|---|---|---|---|---|---|---|---|
| GTEx I | 5,702,593,118 | 1,647,141 | 1,702,468,593 | 72,502 | 249,964 | 671,547 | 34,636 |
| GTEx II | 5,702,593,118 | 1,127,714 | 1,167,933,578 | 54,249 | 184,459 | 469,046 | 26,509 |
| Marmoset (Phase I and II) | 3,386,750,776 | 2,207,887 | 2,404,303,203 | 115,026 | 371,976 | 906,904 | 74,067 |

**Table 3.5:** Genomes supported with Trinity assembly and RNA-Seq bigwig tracks.

| Species | Genome | Assemblies available | Tissues in bigwig tracks |
|---|---|---|---|
| Human | hg38 | 2xGTEx assemblies | cerebellum, colon, frontal cortex, pituitary gland, heart, lymphnode, kidney, liver, lung, muscle, spleen, whole blood |
| Human | hg19 | None | cerebellum, colon, frontal cortex, pituitary gland, heart, lymphnode, kidney, liver, lung, muscle, spleen, whole blood |
| Chimpanzee | panTro5 | Phase I + Phase II polyA | bone marrow, cerebellum, colon, frontal cortex, temporal lobe, pituitary gland, heart, lymphnode kidney, liver, lung, muscle, spleen, thymus |
| Chimpanzee | panTro4 | Phase I+II polyA, Phase I polyA, normalized Phase I+II polyA | bone marrow, cerebellum, colon, frontal cortex, temporal lobe, pituitary gland, heart, lymphnode kidney, liver, lung, muscle, spleen, thymus |
| Gorilla | gorGor3 | Phase I polyA | frontal cortex, cerebellum, heart, kidney, liver, testis[1] |
| Cynomolgus Chinese | macFas5, rheMac3 | Phase I polyA (mRNA standard and mRNA with UDG) | cerebellum, frontal cortex, pituitary gland, colon, kidney, liver, lung, lymphnode, spleen, thymus |
| Cynomolgus Mauritian | macFas5, rheMac3 | Phase I polyA | cerebellum, frontal cortex, pituitary gland, temporal lobe, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen, thymus |
| Pig-tailed Macaque | macNem1, rheMac3 | Phase I polyA | bone marrow, cerebellum, pituitary gland, temporal lobe, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen, thymus |
| Rhesus Macaque Indonesian | rheMac3 | Phase I polyA | none |

**Table 3.6:** Genomes supported with Trinity assembly and RNA-Seq bigwig tracks (continued).

| Species | Genome | Assemblies available | Tissues in bigwig tracks |
|---|---|---|---|
| Rhesus Macaque Chinese | rheMac3 | Phase I polyA | none |
| Japanese Macaque Indonesian | rheMac8, rheMac3 | Phase I polyA (mRNA standard) | bone marrow, cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen, thymus |
| Sooty Mangabey | cerAty1, rheMac8, rheMac3 | Phase I polyA | bone marrow, cerebellum, pituitary gland, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen, thymus |
| Baboon | papAnu2, papAnu3 | Phase I poly A | bone marrow, cerebellum, pituitary gland, temporal lobe, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen, thymus |
| Marmoset | calJac3 | Phase I+II polyA, Phase I polyA, Phase I total RNA | bone marrow, left brain, right brain, pituitary gland, colon, heart, kidney, liver, lung, lymphnode, spleen |
| Squirrel Monkey | saiBol1 | none | bone marrow, cerebellum, frontal cortex, pituitary gland, temporal lobe, colon, heart, kidney, liver, lung, lymphnode, muscle, spleen |
| Mouse Lemur | micMur2, micMur3 | Phase I polyA | cerebellum, frontal cortex, temporal lobe, colon, kidney, liver, lung, muscle, spleen |

# References


1. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols* **8** (2013).
2. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287,** 2196–2204 (2000).
3. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18,** 821–829 (2008).
4. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29,** 644–652 (2011).
5. Stein, L. D. The case for cloud computing in genome informatics. *Genome biology* **11,** 207 (2010).
6. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* **13,** 667–672 (2012).
7. Chan, J. Z., Pallen, M. J., Oppenheim, B. & Constantinidou, C. Genome sequencing in clinical microbiology. *Nature biotechnology* **30,** 1068–1071 (2012).
8. Gargis, A. S. *et al.* Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature biotechnology* **30,** 1033–1036 (2012).
9. Blood, P., Center, P. S., Couger, B. & Pipes, L. Enabling Large-scale Next-generation Sequence Assembly With Blacklight.
10. Pipes, L. *et al.* The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic acids research* **41,** D906–D914 (2012).
11. Henschel, R. *et al. Trinity RNA-Seq Assembler Performance Optimization* in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the Campus and Beyond* (ACM, Chicago, Illinois, USA, 2012), 45:1–45:8. ISBN: 978-1-4503-1602-6. doi:10.1145/2335755.2335842. <http://doi.acm.org/10.1145/2335755.2335842>.
12. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* **10,** 155–159 (2009).

13. Peng, X. *et al.* Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic acids research* **43,** D737–D742 (2014).

14. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).

15. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

# Chapter 4

# Alternative splicing expansion in humans

This next chapter describes unpublished material from recent work. It describes the development of a cross-primate skipped exon database solely based on *de novo* annotation, and displays the strength of using a *de novo* transcriptome assembly approach in analyzing poorly or even unannotated genomes.

# Abstract

We present the first study to systematically identify and measure alternative splicing directly across species with relying on annotation. Additionally, we performed the most comprehensive analysis of alternative splicing across primates to date. We identified 14,201 one-to-one orthologous skipped-exon events that relied only on *de novo* assembly of RNA-Seq data from human, chimpanzee, baboon, and cynomolgus macaque. We measured percent-spliced-in ($\Psi$) in each of these events in each tissue from each species. $\Psi$ can be interpreted as the frequency in which the cassette exon is included in the final transcript. We find that the alternative splicing has been increasing in primates leading to humans with greater relative abundance of alternative splicing events. We show that the fastest increasing abundance of alternative splicing occurs in cerebellum and heart. We show that the cerebellum has a large amount of differentiation that is driven by its low variance within humans, and show that there is an excess of high $\Psi$ values in this tissue. We show that low $\Psi$ values are propagating in tissues such as lung and spleen that have low differentiation. Additionally, we find that humans display more gains and fewer losses of alternative splicing in every tissue. We show that alternative splicing displays different amounts of conservation in different tissue. Specifically, the core splicing pattern in muscle, heart, and brain are conserved across primates whereas all other tissues are lineage-specific. Brain tissues show accelerated change compared to other tissues and we show that alternative splicing changes in the brain are more predictive of selection based on the variance in conservation scores at different levels of $\Psi$.

## 4.1 Introduction

The differential inclusion and exclusion of exonic sequences generated by alternative splicing (AS) is one of the main sources for expanding the number of proteins that can be produced by a single gene. As sequencing has increased in humans, the number of genes that encode more than one mRNA has also increased. Although it had been thought that AS was relatively rare when it was discovered, it is now widely accepted that $> 95\%$ of human genes are alternatively spliced. The evolutionary implications of expanded proteomic diversity remain unclear. AS provides a high level of evolutionary plasticity, and it is often speculated that changes in AS that are not conserved might underlie phenotypic variations between species and between individuals within species [1]. Recent examples in species-specific adaptations have provided support for this claim [2]. Because even point mutations in exons and introns can enhance or disrupt splicing control elements, it is thought that splicing patterns are constantly evolving. However, opposing viewpoints regarding the evolution of AS in tissues and the abundance of AS in primates still exist [3, 4]. In one specific but important example of the opposing viewpoints, Merkin et al. (2012) [4] show that brain, heart, and muscle AS patterns are conserved across mammals while Barbosa-Morais et al. (2012) [3] showed that all tissues had species-specific AS patterns. We sought to clarify this and many other unresolved questions on the evolution of AS in primates and provide new insight into the importance of AS in humans.

## 4.2 Methods

We identified AS events using the 8 non-human primate *de novo* transcriptome assemblies described in **Chapter 3**. We attempted to re-assemble most transcripts using both NHPRTR Phase I (pooled RNA-Seq data) [5] and Phase II (tissue-specific RNA-Seq data) [6] because we have found that increasing the amount of input RNA-Seq reads to be assembled *de novo* greatly improved the quantity and accuracy of the assembled transcripts. However, because of technical artifacts that were inherent in our RNA-Seq data, we were only reliably able to create skipped exon annotations from the transcriptome assembly. Additionally, we created an equivalent human *de novo* transcriptome assembly using tissue-matched RNA-seq data provided by the GTEx consortium [7], and used the same method described in **Chapter 3**. No other

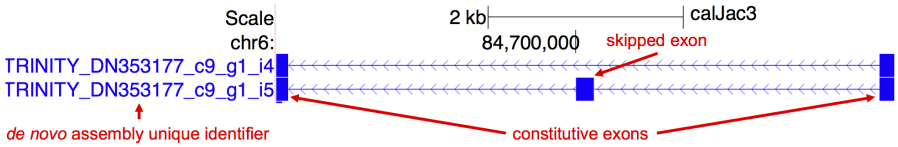


**Figure 4.1:** Example of a skipped exon event identified from Marmoset *de novo* transcriptome assembly using our method.

annotation was used for our analysis and all annotations (including human) were derived directly from RNA-seq data. Out of the four basic types of alternative splicing (alternative 5′ splice-site selection, alternative 3′ splice-site selection, cassette-exon inclusion or skipping, and intron retention), because of the intronic contamination that was present in the RNA-seq data (and carried over to the *de novo* assemblies), we filtered for high confidence exon-skipping events from the *de novo* assemblies. Exon-skipping gradually increases further up the mammalian tree and exon-skipping is the most prevalent type of AS in mammals. Out of all types of AS events conserved between human and mouse, 38.4% are exon-skipping events [8]. We identified exon-skipping events in each *de novo* assembly as an event where a cassette exon was present in between two of the same exons, and two isoforms (one with the cassette exon, one without the cassette exon, and identified by their unique ID assigned by Trinity) were fully assembled (**Figure 4.1**). Additionally, we only identified events that preserved strand in the exon-skipping structure (both within the exons and introns). We aligned all *de novo* assemblies using a derivation of STAR [9], STARlong, which is an alignment program specifically designed to align fully assembled transcripts to genomes. The STARlong alignment specifications that we used for human genome build hg38 were the following:

```
STARlong --genomeDir hg38 --readFilesIn Trinity.fasta
--outFilterMismatchNmax 5000 --outFilterMismatchNoverLmax 0.05
--sjdbOverhang 100 --seedPerReadNmax 50000
```

After alignments, strand for the transcripts was assigned directly from the strand-specific RNA-Seq raw reads where at least 90% of the raw reads were of the same strand. Since GTEx RNA-Seq was unstranded, strand assignments of skipped exon events from GTEx data was inferred from at least 90% overlap with the latest gencode annotations (version 26). **Table 4.1** describes the number of skipped-exon events identified from the *de novo* assemblies. We have found that *de novo* assemblies created

**Table 4.1:** Number of Splicing Events Identified

| Species | Genome | # of Skipped Exon Events Identified | Data used for assembly |
|---|---|---|---|
| Human | hg38 | 15,956 | 61 GTEx samples |
| Chimpanzee | panTro5 | 7,290 | NHPRTR Phase I and II |
| Baboon | papAnu2 | 3,277 | NHPRTR Phase I |
| Cynomolgus Macaque | macFas5 | 5,203 | NHPRTR Phase I (Mauritian and Indonesian subspecies) |
| Japanese Macaque Indonesian | rheMac8 | 3,161 | NHPRTR Phase I |
| Sooty Mangabey | cerAty1[1] | 6,141 | NHPRTR Phase I |
| Pig-tailed Macaque | macNem1 | 6,064 | NHPRTR Phase I |
| Marmoset | calJac3<br>calJac3 | 226<br>2,600 | NHPRTR Phase I<br>NHPRTR Phase I and II |
| Mouse Lemur | micMur2<br>micMur3 | 6,799<br>7,008 | NHPRTR Phase I<br>NHPRTR Phase I |

from more reads (~5 billion reads, when combining the Phase I and II data, as opposed to ~3 billion reads, when using Phase I data only) allowed us to identify more events. Even though ideally we would like to use *de novo* assemblies built from combined Phase I and II data only, many of our assemblies ran into complications during the build with the Trinity software we were not able to finish these jobs in time for this dissertation. The number of exon skipping events identified are not an accurate measure of the actual number of exon-skipping events present in the genome because they also depend on the number of reads used for the *de novo* assembly, the tissues used in the assembly (in some species we had missing tissues), and differences in the build quality produced by Trinity. We have found that Trinity has produced different transcriptomes of varying quality (measured against gencode annotation) from the same input dataset. Notice that using only NHPRTR Phase I data for marmoset only yielded 226 splicing events but when we ran the assembly another time using both NHPRTR Phase I and Phase II data we identified >10 $\times$ splicing events. Since we did not rely on any existing gene annotation, we were able to reliably identify exon-skipping events even in cases where the genomes were unannotated or had limited annotation. A custom skipped exon annotation was created for each primate genome build. We used UCSC reciprocal best chain files with liftOver for the latest genome builds for each primate species to map one-to-one orthologous events. For orthologous mapping of skipped exon events, flanking constitutive exons were matched by at least 90% of each exon. We used the STAR software (version 2.5.2b) with RNA-seq short-read specifications to align all of the tissue-specific RNA-seq reads to their own genome except in the case of the species Japanese Macaque (Indonesian) where only the nearest genome for Rhesus Macaque, rheMac8, was available. We identified 14,201 one-to-one orthologous exon skipping events that preserved the same structure and strand as pictured in **Figure 4.1** across human, chimpanzee, baboon, and cynomolgus macaque. We then used the software Mixture of Isoforms probablistic model for RNA-Seq (MISO) [10] to calculate percent-spliced-in ($\Psi$) in each exon-skipping event in each tissue using custom annotation created for each exon-skipping event identified from our *de novo* assemblies. MISO uses Bayesian inference to calculate the probability that an observed read was produced from an isoform which is specified by the annotation. Since we are calculating $\Psi$ using an exon-centric analysis (as opposed to an isoform-centric analysis), the $\Psi$ values calculated can be easily interpreted as the frequency in which the cassette exon is spliced-in in all isoforms of that gene. A $\Psi$ of 1 is interpreted as a constitutive exon whereas a $\Psi$ of 0 is interpreted as an exon that is never spliced-in.

[1] Unannotated genome.

To standardize our Ψ calculations for different read lengths we took into account the paired-end insert length size and standard deviations for each tissue-specific RNA-Seq sample. We calculated gene expression read counts using Bioconductor package easyRNASeq [11] with our custom skipped-exon annotation. To correct for batch effects between samples we used Bioconductor package NOISeq [12]. Our pseudogene skipped exon annotation set was manually curated from a mixture of transcribed and processed pseudogenes identified by Gencode (version 26) [13] and those exons that overlapped gencode pseudogene exons by 90%. The Ψ calculations for pseudogenes were done with the same method as the skipped-exon events identified from the assemblies. PhyloP scores [14] were taken from phyloP 7-way alignments from hg38. The assemblies used in those alignments were human, chimpanzee, rhesus, dog, mouse, rat, and opossum.

## 4.3 Results

### 4.3.1 Human cerebellum shows the most increase in AS abundance in AS events

To estimate the abundance of AS events in tissue samples we used a measure of relative AS abundance, which we define as the number of skipped exon events with a Ψ value $> 0.10$ and $< 0.9$ normalized by the total read count of that tissue RNA-Seq sample. Humans have an increased abundance in AS events at a similar rate for all tissues compared to NHPs (**Figure 4.2**). Chimpanzees show an abundance of AS events greater than macaques and baboons but less than humans for all tissues, which is in contradiction to Barbosa-Morais et al. (2012) [3] who reported that chimpanzees had greater AS abundance than humans for most tissues and except for cerebellum which displayed an increase more than double of any other tissue. Cerebellum, spleen, and colon displayed the most increase ($\beta_0 < -0.7$) in relative AS abundance vs. divergence time from human (**Table 4.2**).

**Table 4.2:** Linear regression models of log(relative AS abundance) vs. divergence time (millions of years ago) from human.

| Tissue | $\beta_0$ | $\beta_1$ | $p-value$ |
|---|---|---|---|
| Cerebellum | -0.7789087 | 10.76068 | 0.00602** |
| Frontal Cortex | -0.4737367 | 10.566 | 0.0335* |
| Pituitary Gland | -0.56529 | 10.82636 | 0.1002 |
| Colon | -0.7382033 | 10.8715 | 0.01145* |
| Heart | -0.5768427 | 10.64198 | 0.0444* |
| Kidney | -0.43059 | 10.55065 | 0.1006842 |
| Liver | -0.6620833 | 10.6155 | 0.0489* |
| Lung | -0.6399217 | 10.7223 | 0.0899 |
| Lymphnode | -0.596945 | 10.67459 | 0.0784 |
| Muscle | -0.6290173 | 10.44821 | 0.0182* |
| Spleen | -0.7784953 | 10.86707 | 0.0371* |

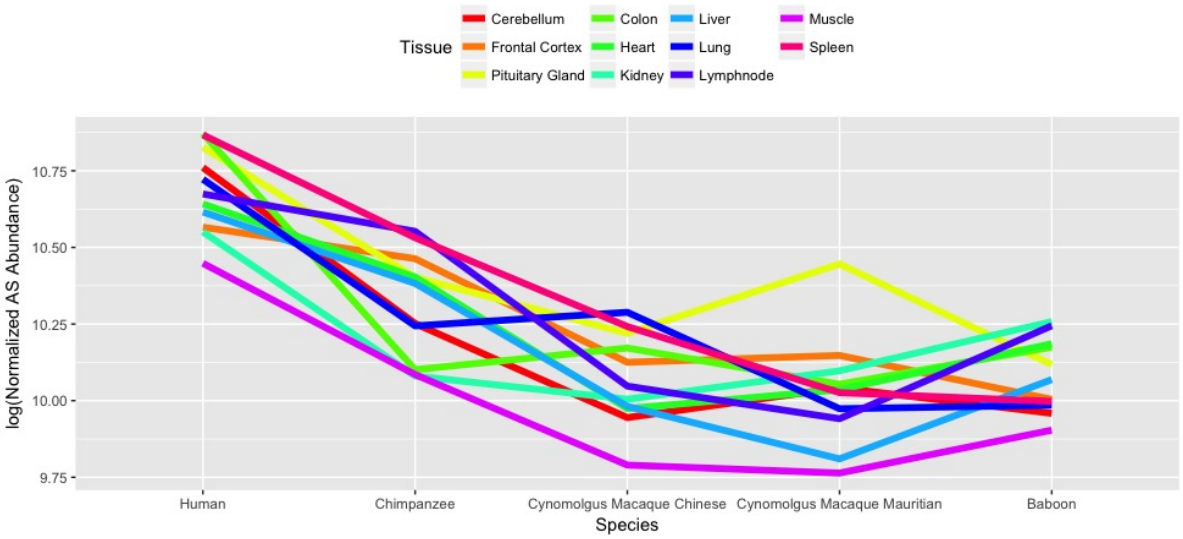


**Figure 4.2:** Relative abundance of AS events in all tissues for humans and 4 NHPs (human, chimpanzee, cynomolgus macaque chinese, cynomolgus macaque mauritian, and baboon)in 11 tissues (cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, liver, lung, lymphnode, muscle, and spleen).

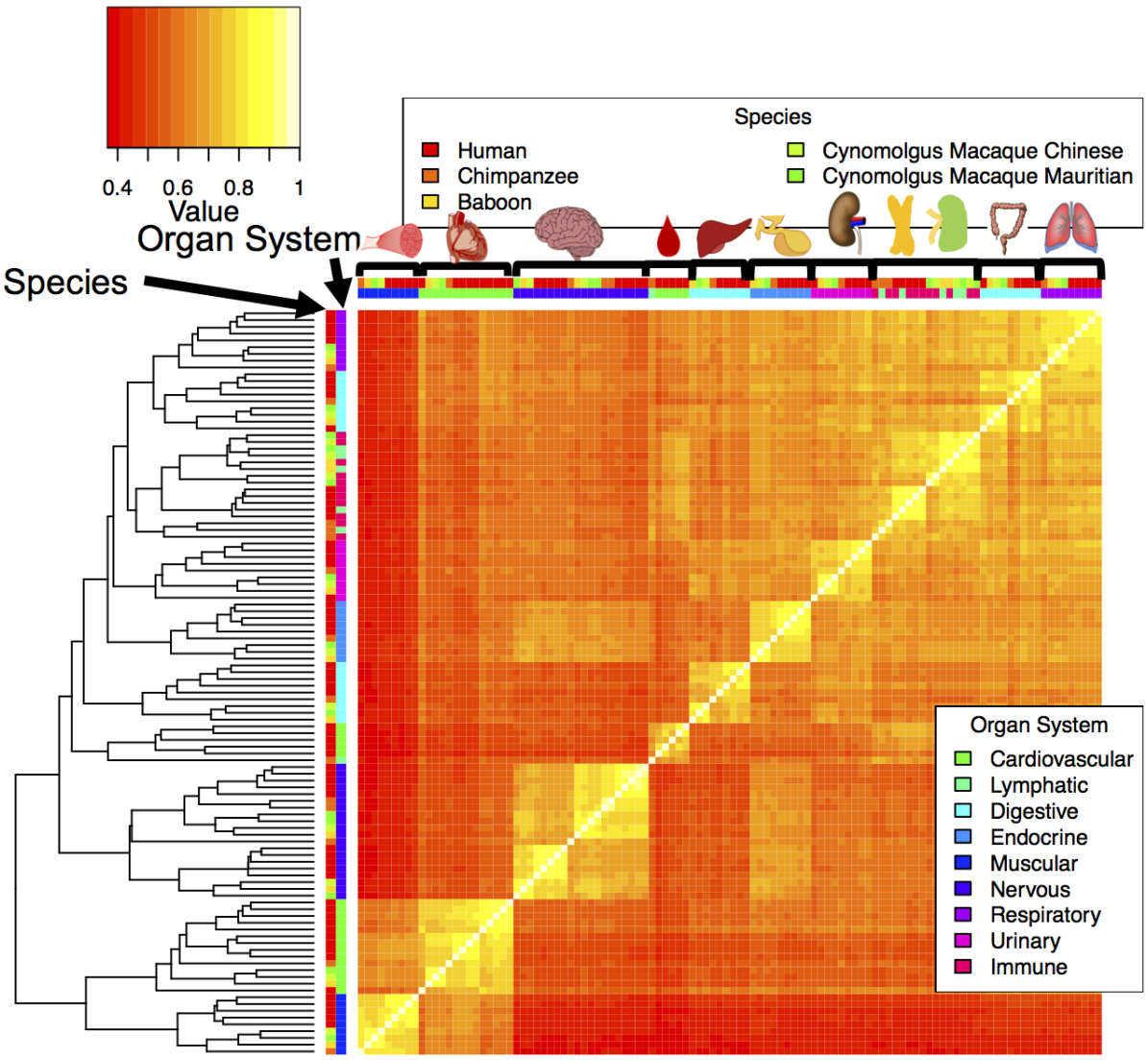


**Figure 4.3:** Heatmap of Jenson-Shannon divergence ($1 - \sqrt{\text{JSD}}$) for read counts in exon-skipping events ($n = 14,201$) in humans and 4 NHPs (Chimpanzee, Baboon, Cynomolgus Macaque Chinese, and Cynomolgus Macaque Mauritian). Tissues used for organ systems: Cardiovascular (Heart, Blood), Lymphatic (Lymphnode), Digestive (Colon, Liver), Endocrine (Pituitary Gland), Muscular (Skeletal Muscle), Nervous (Cerebellum, Frontal Cortex, Temporal Lobe), Respiratory (Lung), Urinary (Kidney), Immune (Spleen, Thymus). Clustering was performed using Euclidean distance with complete linkage.

### 4.3.2 Brain, Heart, and Muscle have a conserved splicing pattern

Since tissues have a well-described tissue-dominated conserved gene expression pattern across species, we clustered read counts calculated from our exon-skipping annotation to check for data quality (**Figure** 4.3). Although a tissue-dominated

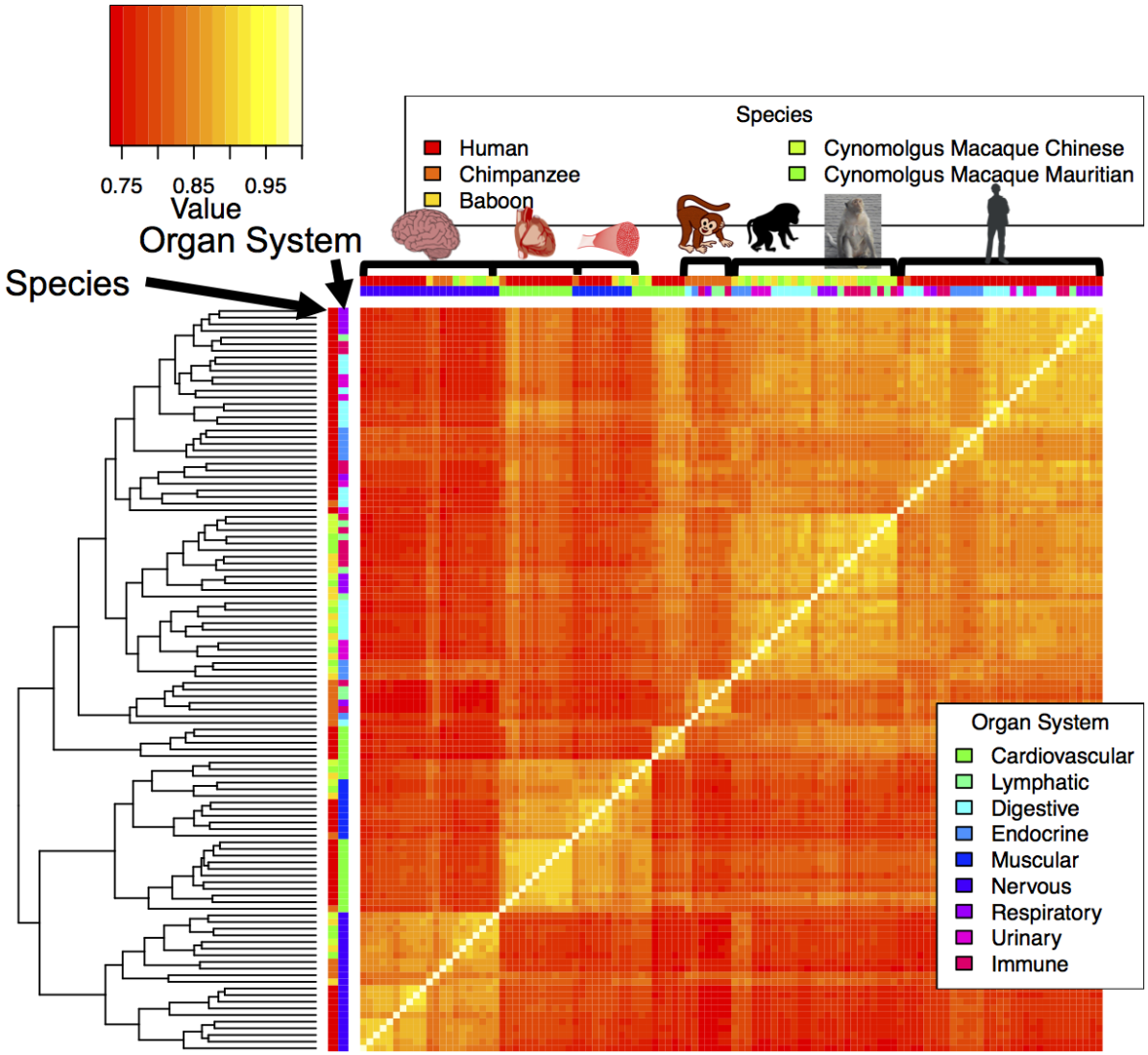


**Figure 4.4:** Heatmap of Jenson-Shannon divergence ($1 - \sqrt{\text{JSD}}$) for $\Psi$ in exon-skipping events ($n = 14,201$) for humans and 4 NHPs (Chimpanzee, Baboon, Cynomolgus Macaque Chinese, and Cynomolgus Macaque Mauritian). Tissues used for organ systems: Cardiovascular (Heart, Blood), Lymphatic (Lymphnode), Digestive (Colon, Liver), Endocrine (Pituitary Gland), Muscular (Skeletal Muscle), Nervous (Cerebellum, Frontal Cortex, Temporal Lobe), Respiratory (Lung), Urinary (Kidney), Immune (Spleen, Thymus). Clustering was performed using Euclidean distance with complete linkage.

clustering is apparent in our data for gene expression, the clustering of $\Psi$ shows a species-specific dominated clustering for most tissues except for brain, heart, and muscle (**Figure** 4.4). This provides more evidence for the brain, heart, and muscle conserved splicing pattern observed in mammals by Merkin et al. (2012) [4] which used only 489 splicing events but disagrees with the species-dominated splicing

pattern reported in humans and NHPs by Barbosa-Morais et al. (2012) [3]. It is apparent in the clustering with Ψ that certain tissues show a species-specific pattern faster than others. Even though cynomolgus macaque Mauritian and cynomolgus macaque Chinese show a similar splicing pattern for all tissues, they diverge patterns in two tissues involved in immune response (spleen and thymus). Additionally, although the divergence time between macaques and baboons (6-8 million years ago) is similar to the divergence time between chimpanzees and humans (6 million years ago), humans and chimpanzees show no conservation in splicing patterns except for tissues conserved in mammals (brain, heart, and muscle). On the other hand, baboons and macaques display a similarity for almost all other non-conserved tissues (except those involved in immune response). A closer look at the clustering of Ψ values in brain tissues reveal that although the overall splicing pattern is more conserved than other tissues, a species-dominated clustering is observed within brain tissues (**Figure 4.5**). Additionally, two distinct clusters form when $k = 2$ during hierarchical clustering between humans and NHPs. Since Ψ is a quantitative trait, we calculated $Q_{ST}$ for Ψ values. $Q_{ST}$ is an $F_{ST}$-like measure for the amount of differentiation for quantitative traits. Specifically, where "$\pi_{between}$" is defined as the amount of variance between the two groups (humans and non-human primates), $\nu_B$, and "$\pi_{within}$" is defined as the amount of variance within each group, $\nu_W$. We calculated $Q_{ST}$ for 10 tissues (**Table 4.3**). Cerebellum, kidney, and liver show the greatest amount of differentiation, but the cerebellum $Q_{ST}$ measure is driven by the least amount of within group variance. Muscle shows the least amount of differentiation with the lowest amount between group variance.

### 4.3.3 Cerebellum shows an excess of high Ψ values.

We have found that not all tissues evolve at the same rate. By comparing the bulk distributions of Ψ across tissues, we have found that some tissues such as spleen, colon, and lung show an excess of low Ψ values whereas cerebellum shows an excess of high Ψ values (**Figure 4.6**; Mann-Whitney, p-value=$5.817 \times 10^{-13}$). The tissues that display a tissue-dominated clustering such as spleen, lung, and lymphnode also display an excess of low Ψ values. For example, lung compared to muscle shows an excess of low Ψ values (**Figure 4.7**; Mann-Whitney, p-value=$3.832 \times 10^{-10}$). Cerebellum is the only tissue that shows an excess of high Ψ values.

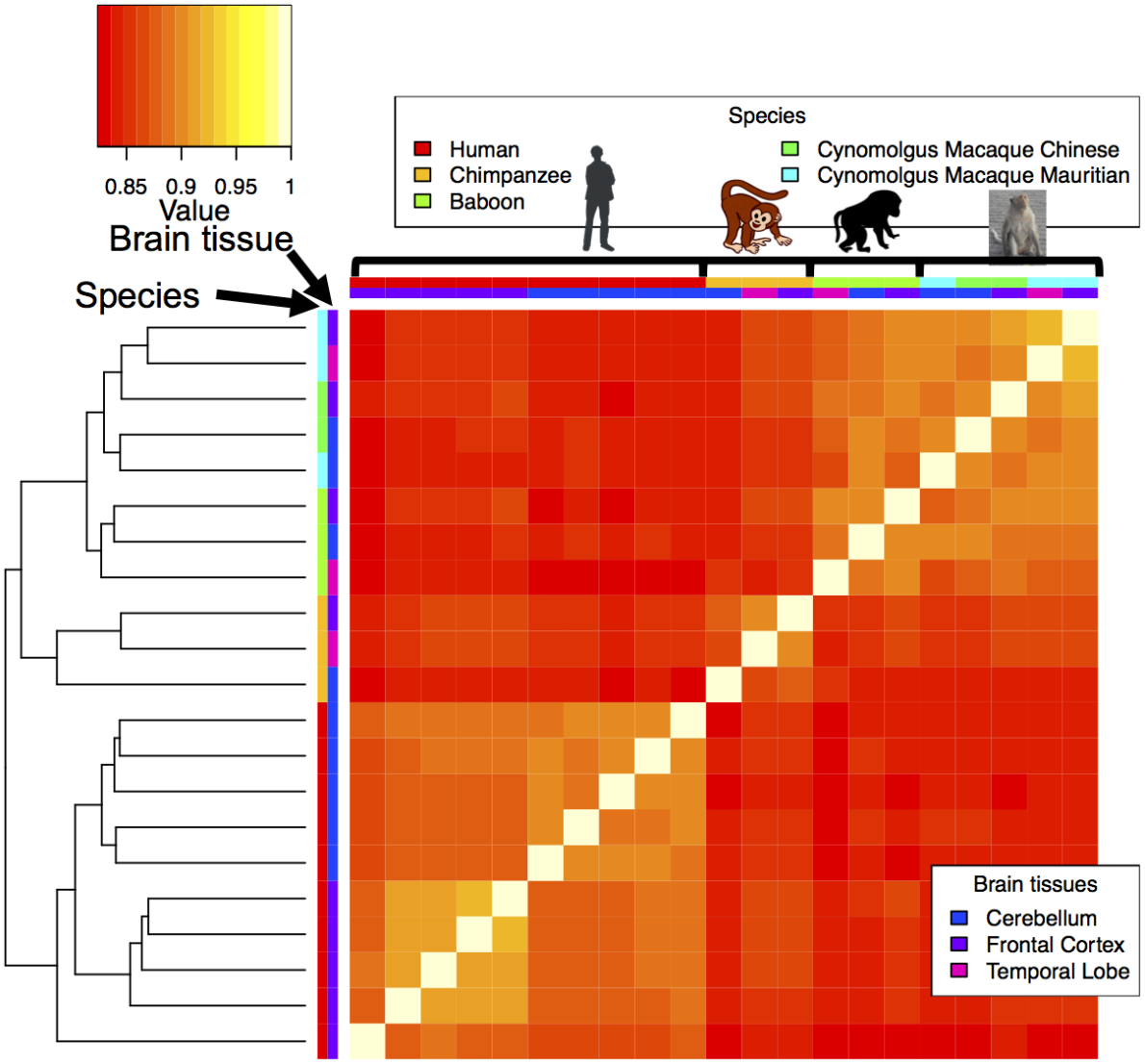


**Figure 4.5:** Heatmap of Jenson-Shannon Divergence ($1 - \sqrt{\text{JSD}}$) for $\Psi$ in exon skipping events ($n = 14,201$) in brain tissues (cerebellum, frontal cortex, and temporal lobe) for humans and 4 NHPs (chimpanzees, baboon, cynomolgus macaque chinese, cynomolgus macaque mauritian). Clustering was performed using Euclidean distance with complete linkage.

**Table 4.3:** $Q_{ST}$ measurements between humans and non-human primates for differentiation in $\Psi$ values for 10 tissues. $v_B$ is the amount of variance between the two groups (humans and non-human primates) and $v_W$ is the amount of variance within each group.

| Tissue | $Q_{ST}$ | $v_B$ | $v_W$ |
|---|---|---|---|
| Kidney | 0.04155683 | 0.01062531 | 0.1225281 |
| Liver | 0.03331782 | 0.008369649 | 0.1214184 |
| Cerebellum | 0.02116257 | 0.005071102 | 0.1172775 |
| Frontal Cortex | 0.01541406 | 0.003782551 | 0.1208068 |
| Pituitary Gland | 0.01334571 | 0.003117966 | 0.1152563 |
| Heart | 0.0132788 | 0.003330524 | 0.1237423 |
| Spleen | 0.01251225 | 0.003084136 | 0.1217025 |
| Colon | 0.008418166 | 0.002076753 | 0.1223111 |
| Lung | 0.007827337 | 0.001947248 | 0.123414 |
| Muscle | 0.005002717 | 0.00124231 | 0.1235423 |

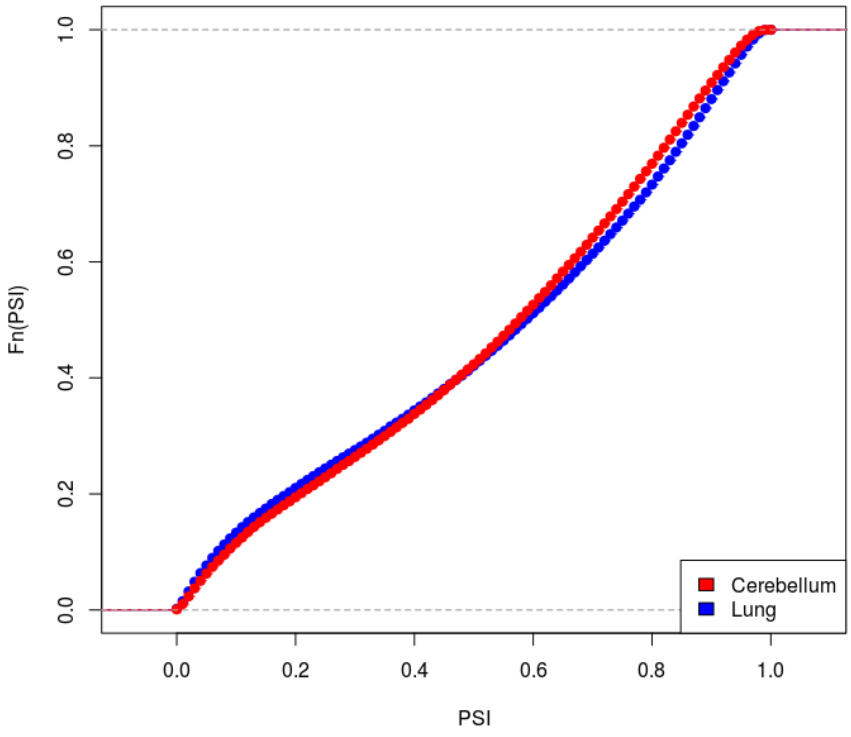


**Figure 4.6:** Empirical cdf plot of cerebellum and lung.

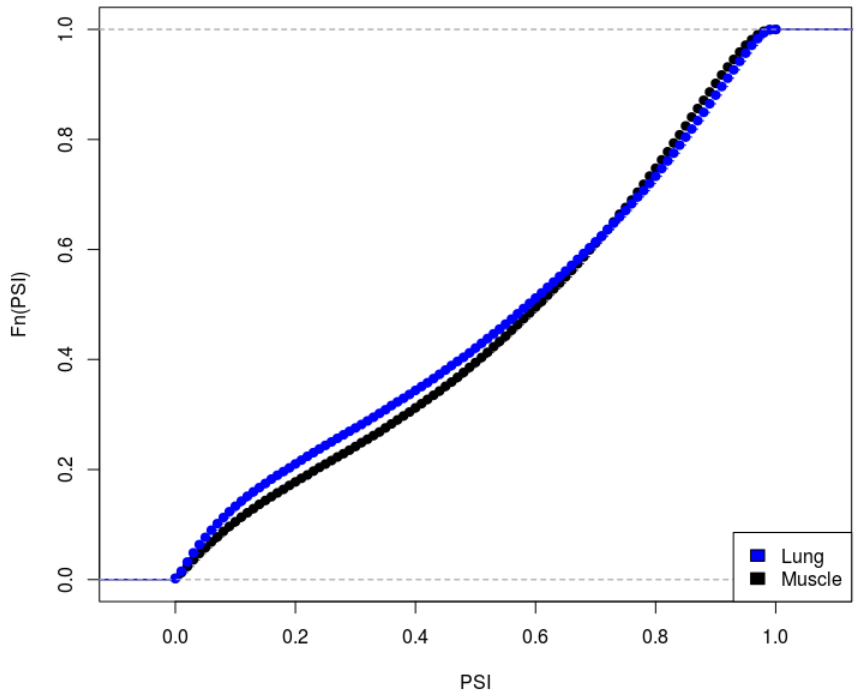


**Figure 4.7:** Empirical cdf plot of muscle and lung.

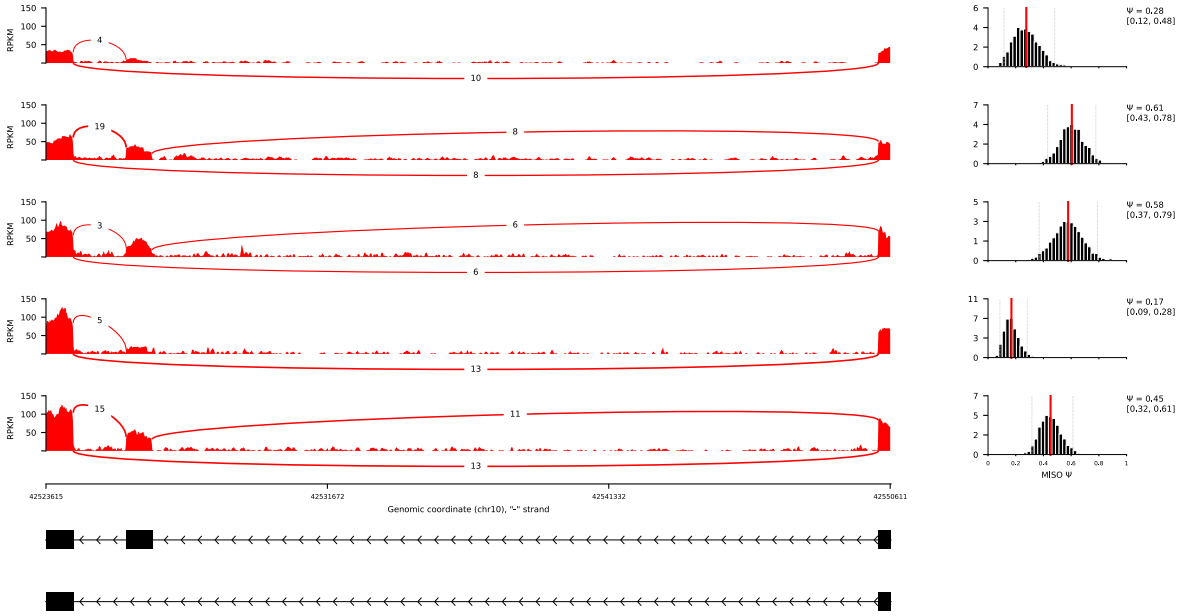


**Figure 4.8:** Sashimi plot in liver for a pseudogene example: zinc finger protein 37B, pseudogene (ZNF37BP, ENSG00000234420).

### 4.3.4 Skipped-exon events in pseudogenes are conserved in brain

We curated a set of skipped-exon events in known human transcribed and processed pseudogenes to serve as a neutrally evolving dataset in which to compare our skipped-exon dataset identified from genes. Many pseudogenes show a non-conserved pattern in both expression and Ψ values in the same tissue (**Figure 4.8**). In one example from the zinc finger protein 37B pseudogene, Ψ can range from 0.17 to 0.61 in the same human tissue. But upon looking at the overall splicing pattern in all human tissues, a conservation of Ψ values in brain was observed (**Figure 4.9**)

### 4.3.5 Differences in conservation scores surrounding skipped-exon in different tissues

We calculated the mean conservation scores (phyloP) 60bp upstream and downstream of the skipped exon as well as 25bp within the skipped exon. The meta plot of these mean phyloP scores surrounding skipped exon events reveal a difference in conservation between high Ψ values (Ψ>0.9) and low Ψ values (Ψ<0.1) (Mann-Whitney Ψ<0.1 vs. Ψ>0.9 p-value=0.04903; Mann-Whitney 0.45>=Ψ<=0.55 vs. Ψ>0.9

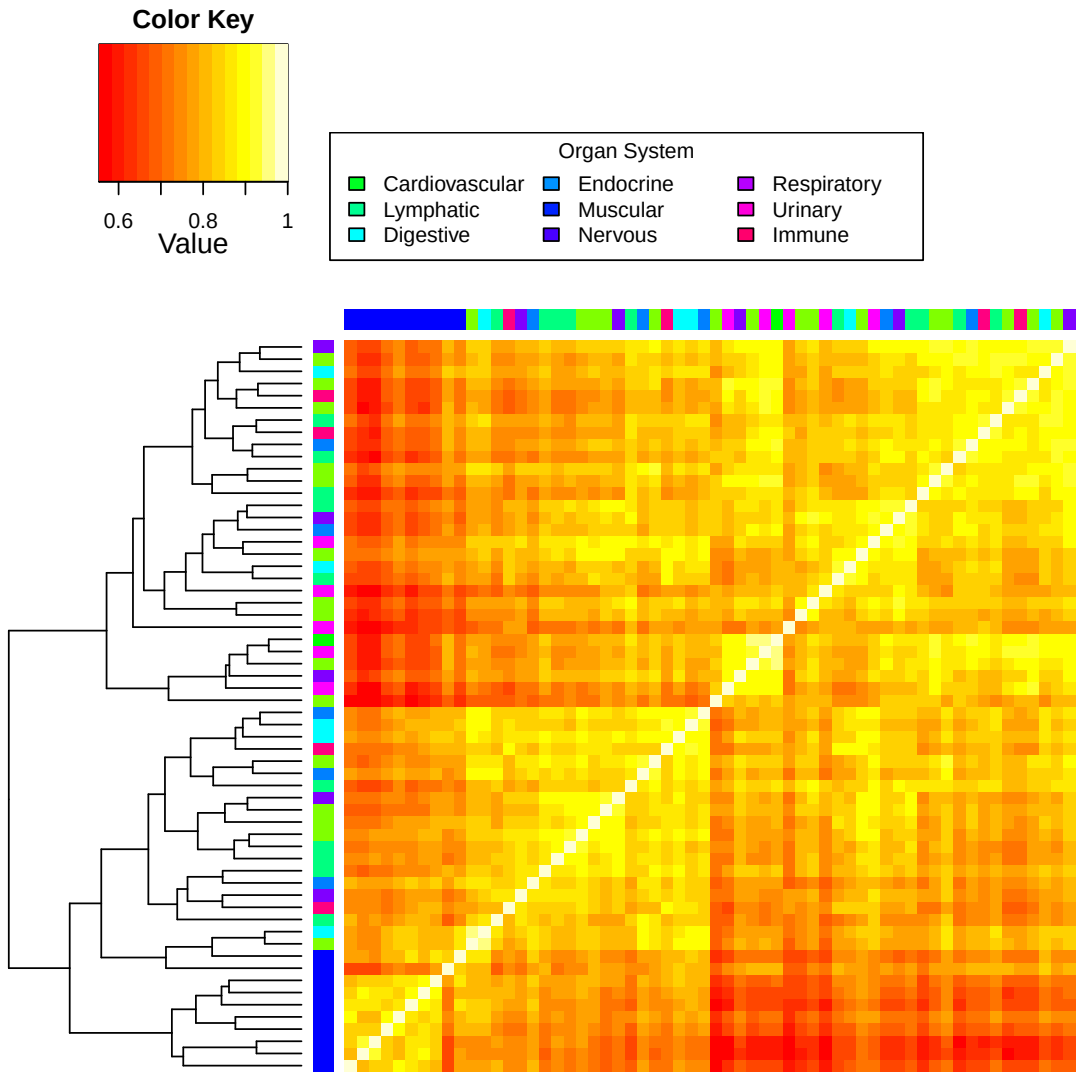


**Figure 4.9:** Heatmap of Jenson-Shannon Divergence ($1 - \sqrt{\text{JSD}}$) for $\Psi$ in exon skipping events in human pseudogenes. Tissues used for organ systems: Cardiovascular (Heart, Blood), Lymphatic (Lymphnode), Digestive (Colon, Liver), Endocrine (Pituitary Gland), Muscular (Skeletal Muscle), Nervous (Cerebellum, Frontal Cortex, Temporal Lobe), Respiratory (Lung), Urinary (Kidney), Immune (Spleen, Thymus). Clustering was performed using Euclidean distance with complete linkage.

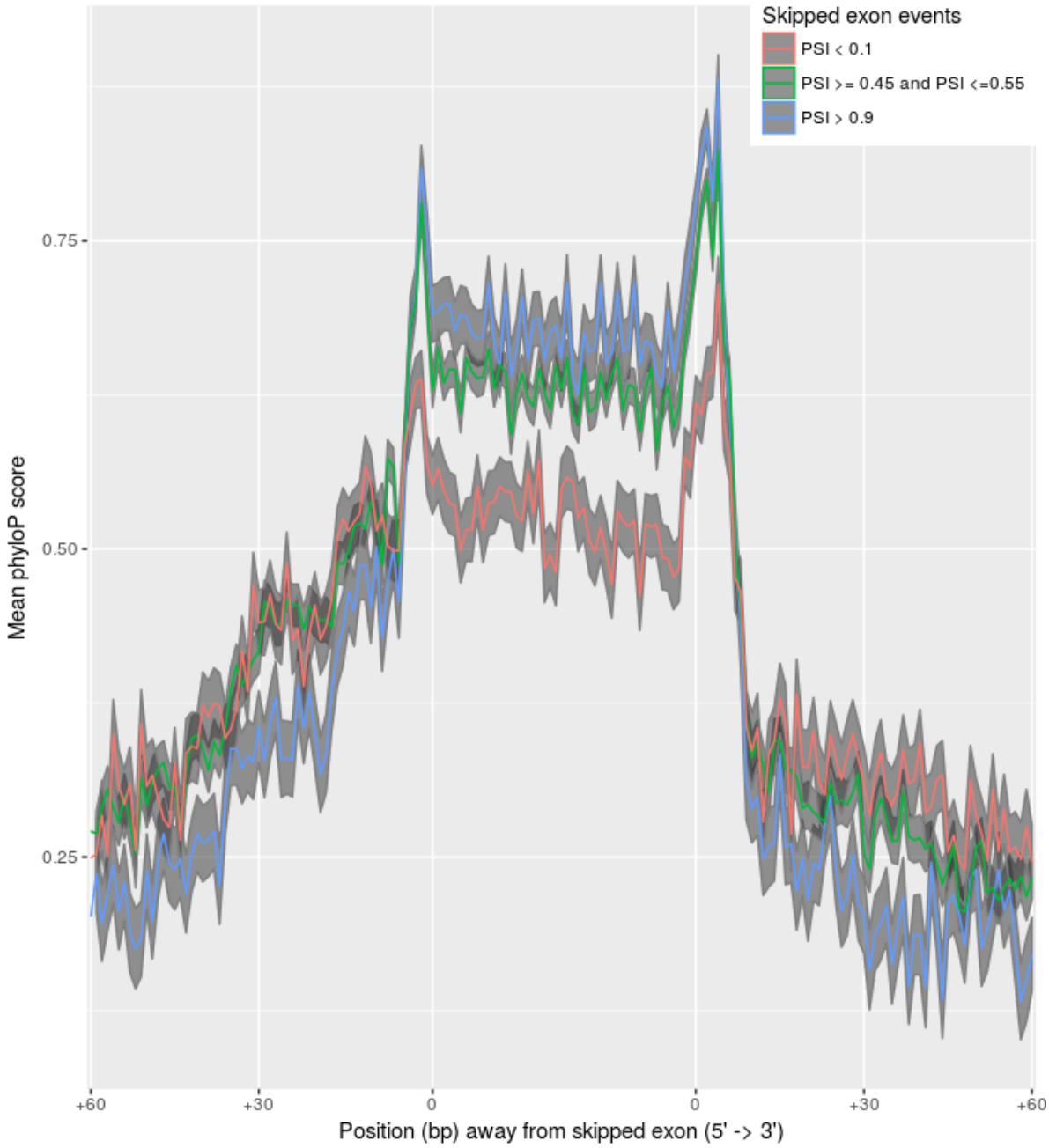


**Figure 4.10:** Meta plot of mean phyloP conservation scores (mean $\pm$ SEM) in human cerebellum surrounding skipped exon events (60bp upstream of the exon, 25bp into the exon for both exon boundaries, and 60bp downstream of the exon) at different $\Psi$ values ($\Psi > 0.9$,$\Psi >= 0.45 and \Psi <= 0.55$, and $\Psi < 0.1$).

p-value=0.04171; Mann-Whitney 0.45>=Ψ<=0.55 vs. Ψ<0.1 p-value=0.1681; **Figure 4.10**). For high Ψ values, there is less conservation in the bases in the flanking introns and high conservation at the splice-site and within the exon. For low Ψ values, the opposite is observed, there is less conservation at the splice-sites and within the exon

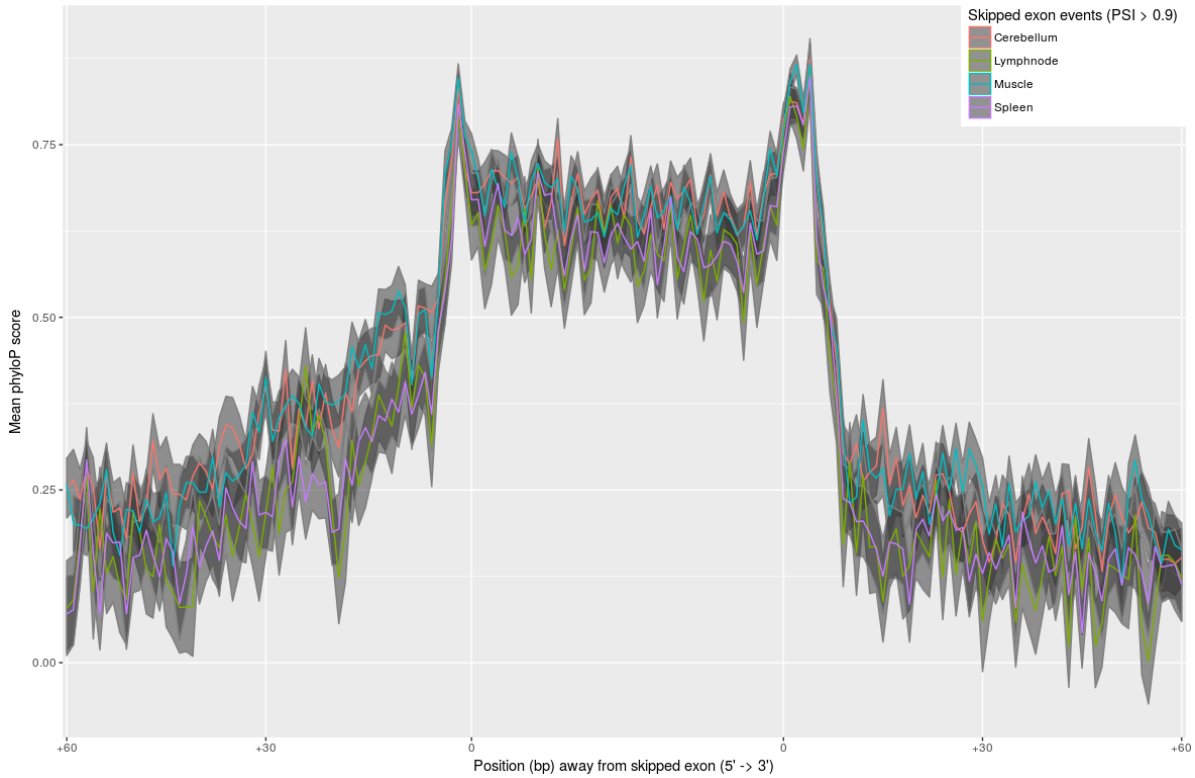


**Figure 4.11:** Mean $\pm$ SEM phyloP conservation scores at high $\Psi$ values ($\Psi > 0.9$) surrounding skipped exons (60bp upstream of the exon, 25bp into the exon for both exon boundaries, and 60bp downstream of the exon) for cerebellum, lymphnode, muscle, and spleen.

but greater conservation in the flanking introns. There is also more conservation in the upstream intron than the downstream intron at all $\Psi$ values. These observations are consistent with the known importance of GT-AG splice site recognition in constitutive exons, and the observations that alternative exons depend more on splice control elements (enhancers and silencers) for the alternative exon to be recognized by the splicing machinery and that the control elements upstream of the exon may be more important for exon recognition. The conservation scores surrounding skipped exon at high $\Psi$ ($\Psi > 0.9$) are similar for all tissues (**Figure 4.11**, Mann-Whitney test not significant). However, for low $\Psi$ values ($\Psi < 0.1$), there is a significant difference between brain tissues and other tissues in conservation scores within and surrounding skipped exons (**Figure 4.12**, Mann-Whitney test between cerebellum and muscle p-value=$3.71 \text{x} 10^{-4}$, Mann-Whitney test between cerebellum and frontal cortex not significant, and Mann-Whitney test between muscle and spleen not significant).

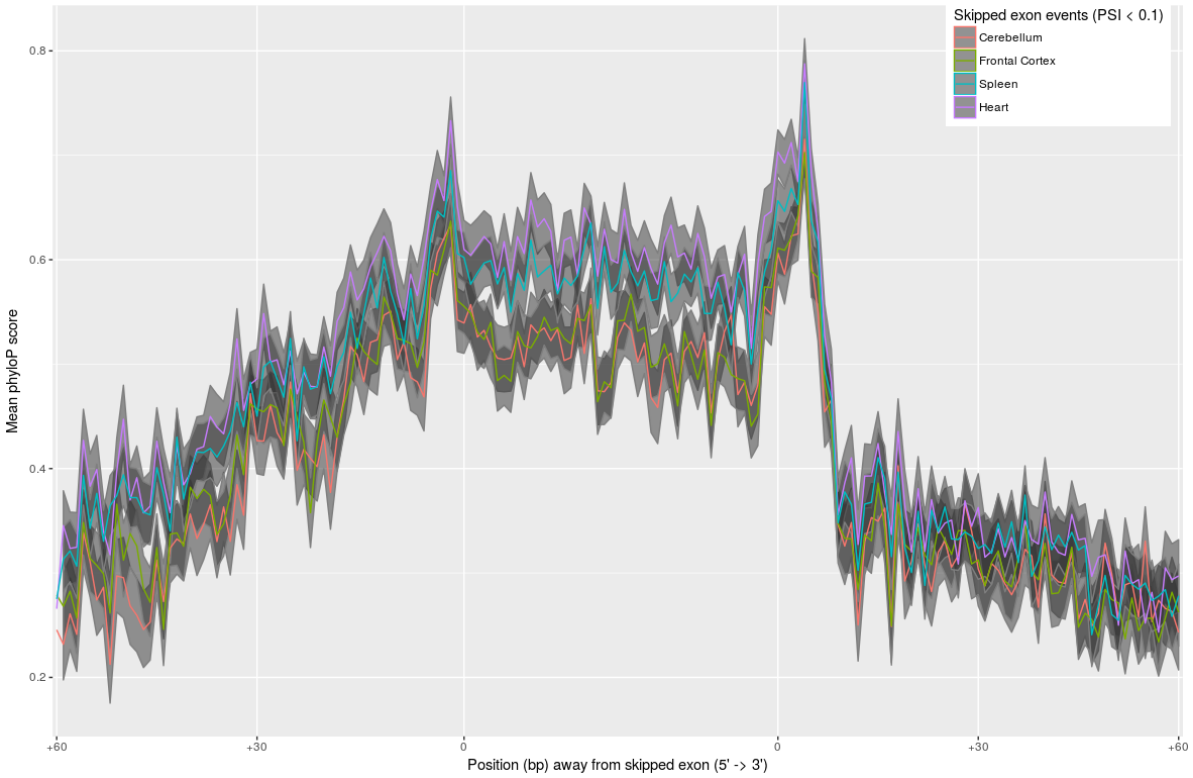


**Figure 4.12:** Mean $\pm$ SEM phyloP conservation scores at low $\Psi$ values ($\Psi < 0.1$) surrounding skipped exons (60bp upstream of the exon, 25bp into the exon for both exon boundaries, and 60bp downstream of the exon) for cerebellum, frontal cortex, spleen, and heart.

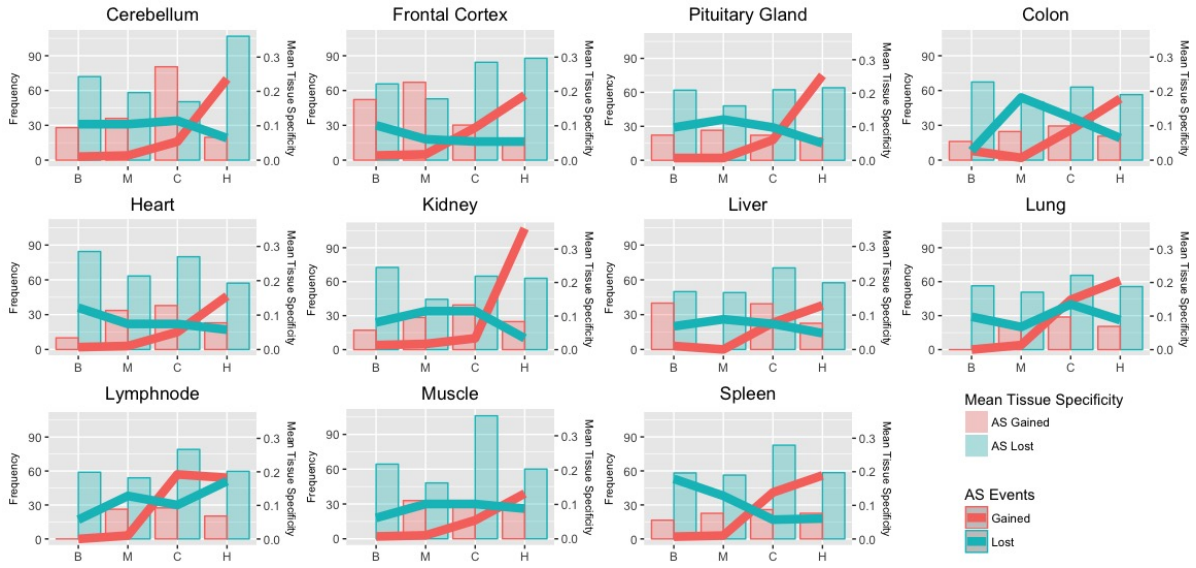


**Figure 4.13:** Gains and losses of AS events (line plot) for 11 tissues (cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, liver, lung, lymphnode, muscle, and spleen) and their mean tissue specificity (bar plot). All gained AS events (both line and bar plot) are in red and all lost AS events (both line and bar plot) are in blue. B=baboon, M=cynomolgus macaque, C=chimpanzee, H=human.

### 4.3.6 Humans have fewer loss of AS events and more gain of AS events in all tissues.

We analyzed the frequency of AS events gained and lost in 11 tissues and their mean tissue specificity (**Figure** 4.13). We defined an AS event gained as $\Psi <= 0.9$ in 1 species with all other species having a $\Psi$ value of $\Psi >= 0.95$. We defined an AS event lost as $\Psi >= 0.95$ in 1 species with all other species having a $\Psi$ value of $\Psi <= 0.9$. We used a measure of mean tissue specificity described by Yanai et al. (2004) [15]. All tissues (except for lymphnode) show an increase in AS events gained in humans with kidney, cerebellum, and frontal cortex gaining the most AS events. AS events that were lost tended to have greater tissue specificity than AS events that were lost in all species. Skeletal muscle showed the least changes in AS events gained or lost across species. Additionally, tissues that had high $Q_{ST}$ values such as cerebellum and kidney also had more gains and less loss of AS events while tissues that had low $Q_{ST}$ values such as muscle also had fewer changes in AS events. Surprisingly, both baboon and macaques displayed only a limited amount of AS events gained (all tissues for baboon and macaque had fewer than 15 gained events). AS events that were lost in humans

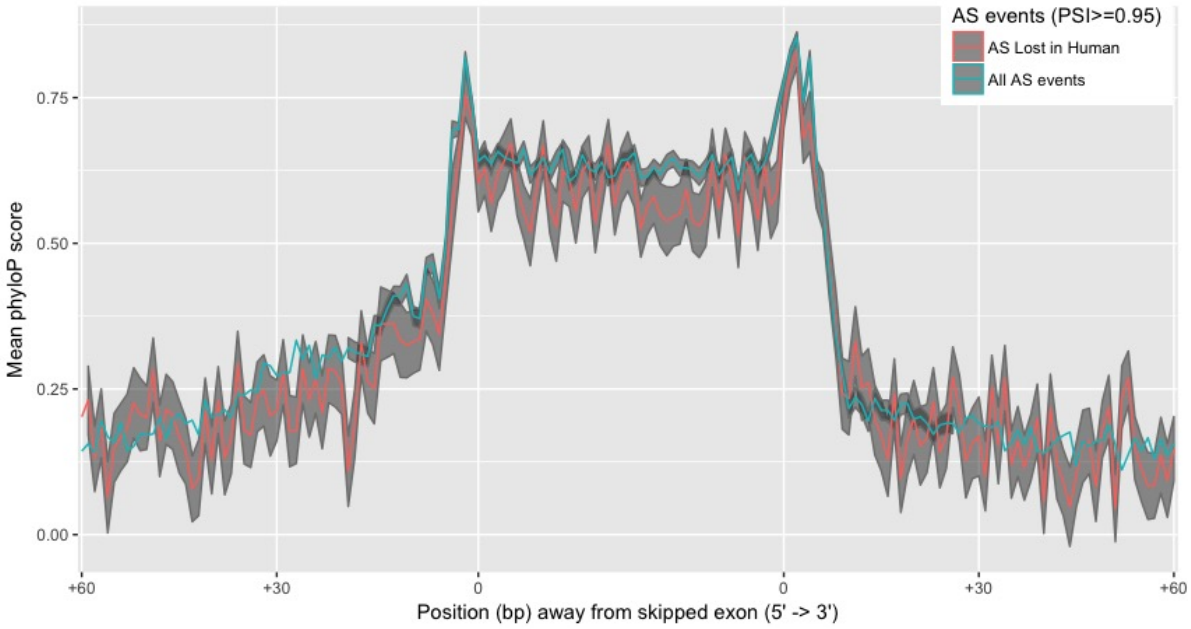


**Figure 4.14:** Mean phyloP conservation profile for AS events lost in human compared to all AS events of high Ψ.

show a significantly lower profile in conservation surrounding the skipped exon than AS events that share the same high level of Ψ values ($\Psi >= 0.95$; **Figure 4.14**);

## 4.4 Discussion

We present the most comprehensive evolutionary analysis of alternative splicing in mammals to date. By not relying on existing annotation and generating custom annotations directly for each species and for each of the tissues we had RNA-Seq data for, we were able to generate an orthologous splicing database that was almost 30 × the scope of previous analyses [4]. The results provide overwhelming support to show that not all tissues evolve through AS at a similar rate. Even though it was hypothesized in 2012 that AS evolves at rate two- or three- times faster than gene expression, it has been inconclusive whether AS truly species-specific or if AS is conserved in certain tissues in mammals. From our clustering (**Figure 4.4**), $Q_{ST}$ (**Table 4.3**), and AS gains and losses (**Figure 4.13**) analyses, it is apparent that skeletal muscle is the most conserved tissue in terms of AS. Skeletal muscle exhibits both a tissue-dominated clustering as well as the lowest $Q_{ST}$, and the fewest AS gains and losses out of all of the tissues studied. Muscle also does not display an excess of low Ψ

values in its distribution. As for the faster evolving tissues, lymphnode, spleen, and thymus are the first tissues to exhibit a species-specific clustering and all show an excess of low Ψ values in their distributions. Furthermore, at high Ψ values (Ψ>0.9), these tissues have significantly lower mean conservation scores than muscle at high Ψ values (**Figure 4.11**). Since lymphnode, spleen, and thymus are all part of the immune system, the faster change in Ψ values might be a reflection of the important role AS plays in species-specific immunity against pathogens. Although brain tissues show a tissue-dominated clustering pattern, there are many important distinctions from other tissues that exhibit the same pattern (i.e., muscle). The tissue-dominated clustering pattern might be more of a reflection of the importance of a conserved core set of genes to exhibit such a pattern. Specifically, **Figure 4.10** does not show a significant difference in mean conservation scores surrounding the skipped exon between high Ψ values (Ψ>0.9) and intermediate Ψ values in cerebellum. Brain tissues also show the greatest range in mean conservation scores surrounding the skipped exon. At high Ψ values (Ψ>0.9), cerebellum and frontal cortex exhibit similar mean conservation scores but at low Ψ values (Ψ<0.1), their scores are significantly lower than any other tissues. This diverse range in conservation scores (as opposed to spleen) might mean that AS in brain is more predictive of selection. This notion along with the increasing abundance and differentiation of AS events in human brain (and in cerebellum especially) provide good evidence that this expansion of AS events could explain part of the phenotypic complexity exhibited between humans and non-human primates. Furthermore, cerebellum is the only tissue that exhibits an excess of higher Ψ values in its distribution. Although we attempted to use pseudogenes as a neutral evolutionary dataset in which to estimate the rate of AS change in tissues, we surprisingly observed a conservation of AS of pseudogenes in brain. Additionally, we saw an elevation of mean conservation scores within the skipped exon of pseudogenes compared to the surrounding introns. Since processed pseudogenes were not randomly spliced as expected, this could be evidence of a functional role for pseudogenes in the brain. Additionally, it is especially remarkable that baboons and macaques both show such few gains of alternative splicing events, which might be representative of the increased reliance of alternative splicing going towards the human lineage as a mechanism to generate new isoforms.

## 4.5 Concluding Remarks

Alternative splicing has been long proposed to underlie species-specific morphological adaptations in mammals because a small change in the DNA could give rise to novel combinations of existing genes. However, to date, only a handful of examples have been shown to support this theory. Similarly, studies analyzing the changes in the DNA underlying novel phenotypes between humans and chimpanzees has also only been shown for limited examples. This is the largest and most comprehensive study for analyzing alternative splicing in primates, and we have provided more evidence to support previous reports of splicing conservation in brain and muscle and have disputed other reports that all tissues show lineage-specific splicing patterns. Although it is unknown whether AS expansion reflects a functional expansion of the transcriptome, we provide systematic evidence that AS provides a large source of diversity that ultimately could underlie the observed phenotypic diversity between humans and non-human primates. In the following chapter, we go into detail about a few of the specific examples of changes that we have found in humans.

# References


1. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457 (2010).
2. Gracheva, E. O. *et al.* Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature* **476,** 88 (2011).
3. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–1593 (2012).
4. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–1599 (2012).
5. Pipes, L. *et al.* The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic acids research* **41,** D906–D914 (2012).
6. Peng, X. *et al.* Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic acids research* **43,** D737–D742 (2014).
7. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).
8. Kim, E., Goren, A. & Ast, G. Alternative splicing: current perspectives. *Bioessays* **30,** 38–47 (2008).
9. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
10. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7,** 1009–1015 (2010).
11. Delhomme, N., Padioleau, I., Furlong, E. E. & Steinmetz, L. M. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* **28,** 2532–2533 (2012).
12. Tarazona, S., García, F., Ferrer, A., Dopazo, J. & Conesa, A. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet. journal* **17,** pp–18 (2012).
13. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome biology* **13,** R51 (2012).

14. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20,** 110–121 (2010).

15. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2004).

# Chapter 5

# Human-Specific Alternative Splicing

This next chapter describes unpublished material from recent work. It describes the human-specific alternative splicing events identified by the skipped-exon database described in **Chapter 4** through the use of Distributions of Isoforms in Single Cell Omics (DISCO) which was developed by Mason lab member and fellow Tri-Institutional Computational Biology and Medicine student Priyanka Vijay.

# Abstract

We present the first study to systematically identify human-specific isoforms. We sought to identify the human-specific alternative splicing events that are significantly different between humans and non-human primates. We used the skipped-exon database of 14,201 one-to-one orthologous alternative splicing events in human, chimpanzee, baboon, and cynomolgus macaque described in **Chapter 4**. We identified 3,954 significant differential splicing events in 1,807 genes. Heart and cerebellum showed the most significant changes between humans and non-human primates. We identified the underlying causes that might explain some of our top differential splicing results: IFI16, MLH3, HERC2P2, and LRRC75A-AS1, a highly expressed sno-lncRNA. We show that small changes in the DNA can cause large changes in isoforms to the point where exons are either completely gained or completely lost. Our GO analyses show that the human-specific splicing events are enriched for pathways and biological functions involving morphogenesis and immune functions. We reveal that GO analyses of the top differential splicing results are important in many highly conserved pathways. Our top GO canonical pathway in every tissue was integrin signaling, which is a main contributer of morphogenesis. Additionally, GO analysis of significantly different splicing brain tissues are involved in morphological adaptations and development of the central nervous system, and heart is enriched for the morphogenesis and development of the heart.

## 5.1 Introduction

The rapid recent accumulation of genomic, transcriptomic, and proteomic data has led to a wealth of resources to allow for the study of the molecular characteristics that could best explain the observed phenotypic differences between species. A large part of the evidence for the basis of organismal complexity in the primate lineage leading to humans is still missing. In **Chapter 4**, we described a database that used transcriptomic data to allow for the analysis of alternative splicing changes across primate species. We were interested to identify specific examples that are present in the overall splicing pattern changes and to see whether they reveal insight into whether or not they might underlie morphological adaptations. To date, only a handful of examples of human-specific isoforms have been described and most are not identified with a systematic approach. By utilizing our skipped exon database, we were able to show that the differences in alternative splicing in humans is enriched for morphogenesis and development of highly species-specific tissues such as brain and muscle. Additionally, our top differentially spliced genes could all be accounted for by changes in the DNA surrounding or directly adjacent to the splice sites which either hindered or enhanced splicing of the alternative exon. All of our differential splicing results presented here show a "switch-like" effect where an exon is completely spliced-in or completely spliced-out.

## 5.2 Methods

We used the skipped-exon splicing database described in **Chapter 4**. We used Distributions of Isoforms in Single Cell Omics (DISCO) to identify the significantly different splicing between humans and non-human primates in each skipped exon event for 10 tissues (cerebellum, frontal cortex, colon, heart, kidney, liver, lung, muscle, pituitary gland, and spleen). DISCO uses non-parametric statistical testing with multiple testing correction to test differences in isoform abundance relative to all isoforms specified by annotation supplied to MISO [1]. We identified 3,954 total significant differential splicing events in 1,807 genes using the t-test with significance level $q - value < 0.05$. Splicing control element motifs were identified with human splicing finder 3.0 [2]. We used dbSNP to identify polymorphic mutations in humans. Gene ontology analysis was performed with Ingenuity Pathway Analysis (IPA) version 36601845 Software (Ingenuity Systems, Redwood City, CA, USA; www.ingenuity.com).

We assigned genes to our custom splicing events if they overlapped our event by at least 95%. 13 of the significantly different splicing events could not be assigned to known genes. We used dbSNP to identify polymorphisms in our genes of interest [3]. Ortholog sequences for multiple alignments were downloaded from ENSEMBL, t-Coffee was used for the multiple alignment, and PAML was used to calculate dN/dS ratios. Human variation analyses were performed with 1000 Genomes Phase III data. Extended haplotype homozygosity analysis was performed using Selscan.

## 5.3 Results

The significant differential splicing results using DISCO are described in **Table 5.1**. Cerebellum, heart, and frontal cortex showed the most significant changes in splicing.

**Table 5.1:** DISCO significant results (Humans vs. Non-Human Primates).

| **Tissue** | **Number of significant genes** | **Number of genes expressed** | **(# of significant genes)/(# of genes expressed)** |
|---|---|---|---|
| Heart | 950 | 11,155 | 0.0851636 |
| Cerebellum | 594 | 11,057 | 0.05372162 |
| Frontal Cortex | 472 | 11,576 | 0.04077402 |
| Pituitary Gland | 445 | 11,786 | 0.03775666 |
| Spleen | 313 | 10,372 | 0.0301774 |
| Kidney | 297 | 11,099 | 0.02675917 |
| Muscle | 241 | 9,572 | 0.0251776 |
| Colon | 238 | 11,620 | 0.020418193 |
| Liver | 216 | 8,914 | 0.0243155 |
| Lung | 188 | 12,069 | 0.0155771 |

We identified the top 10 differentially spliced genes in all tissues in **Table 5.2**. Many of our top results were significant in multiple tissues. The low p-values in heart are a reflection of the conservation of splicing in that tissue and because we had double the number of human samples for that tissue to include in the analysis.

**Table 5.2:** Top 10 differentially spliced genes between humans and non-human primates.

| Gene Name (Full name) | Tissue | p-value | Type of Gene | GO annotation(s) | Related pathway(s) | Associated diseases | other tissues <0.05 |
|---|---|---|---|---|---|---|---|
| IFI16 (Interferon Gamma Inducible Protein 16) | Heart | 9.023e-16 | Protein coding | core promoter binding, transcription factor binding | Innate Immune System, Cytosolic sensors of pathogen-associated DNA | Herpes Simplex, Diffuse Cutaneous Systemic Sclerosis, Psoriasis, Insulin-dependent mellitus, Sjogren syndrome | Cerebellum (6.391e-09), Frontal Cortex (2.546e-08), Pituitary Gland (3.737e-06), Colon (2.644e-08), Kidney (1.945e-10), Liver (3.104e-09), Lung (1.221e-05), Muscle (4.381e-09), Spleen (1.731e-06) |
| MLH3 (MutL Homolog 3) | Heart | 1.094e-10 | Protein coding | centromeric DNA binding, chromatin binding, single-stranded DNA binding | Meiosis, Mismatch repair, DNA damage response | Colorectal Cancer, Hereditary Nonpolyposis Type 7, Endometrial Cancer, Aspermatogenesis, Migraines | Cerebellum (6.813e-09), Frontal Cortex (7.949e-07), Pituitary Gland (4.974e-10), Colon (2.761e-06), Kidney (2.274e-05), Liver (1.743e-06), Lung (2.606e-07), Muscle (3.996e-07), Spleen (1.270e-06) |
| ZMYND11 (Zinc Finger MYND-Type Containing 11) | Heart | 2.848e-10 | Protein coding | transcription corepressor activity, methylated histone binding | Toll-like Receptor Signaling Pathway | Mental retardation, Autosomal Dominant 30, Intellectual Disability-Expressive Aphasia-Facial Dysmorphism Syndorme | Cerebellum (0.00087949), Frontal Cortex (0.00069571), Colon (0.00079323), Kidney (0.0002200803), Lung (3.8731e-05), Muscle (0.000117198) |
| HERC2P2 (Hect domain and RLD 2 pseudogene 2) | Cerebellum | 3.581e-10 | Pseudogene | | Regulated by megakaryocytes | Prader-Willi/Angelman, Autism Spectrum Disorder, Schizophrenia | Frontal Cortex (2.117e-07), Pituitary Gland (6.2896e-08), Colon (0.00013493), Heart (5.357e-08), Kidney (4.247e-07), Liver (1.7408e-06), Lung (7.4045e-05), Muscle (2.0463e-08), Spleen (2.775e-05) |
| SLC45A4 (Solute Carrier Family 45 member 4) | Pituitary Gland | 7.421e-09 | Protein coding | sucrose:hydrogen transporter, symporter | Regulated by LHX1, Regulates sucrose | Schizophrenia, Bipolar Disorder | Cerebellum (2.734e-05), Frontal Cortex (7.224e-06), Colon (2.347e-06), Kidney (0.0008904), Lung (0.0004198) |

**Table 5.3:** Top 10 differentially spliced genes continued.

| Gene Name (Full name) | Tissue | p-value | Type of Gene | GO annotation(s) | Related pathway(s) | Associated diseases | other tissues <0.05 |
|---|---|---|---|---|---|---|---|
| C16orf46 (Chromosome 16 Open Reading Frame 46) | Pituitary Gland | 7.753e-09 | Protein coding | | | Conduct disorder, Attention Deficit Hyperactivity Disorder | Cerebellum (6.770e-07), Frontal Cortex (0.0010152), Heart (0.0008751), Kidney (0.0001396695) |
| PDLIM5 (PDZ and LIM Domain 5) | Muscle | 1.005e-08 | Protein coding | actin binding, protein kinase C binding, cadherin binding | Protein-protein interactions at synapses, transmission across chemical synapses, cell growth involved in cardiac muscle cell development | Nephrogenic Adenofibroma, Nail-Patella Syndrome, Dilated cardiomyopathy | |
| TMEM175 (Transmembrane Protein 175) | Heart | 1.862e-08 | Protein coding | potassium ion leak channel activity | | Parkinson's disease | Spleen (0.000300833) |
| LRRC75A-AS1 (LRRC75A Antisense RNA 1) | Heart | 2.348e-08 | RNA gene (non-coding) | protein binding | Binds APP | | Cerebellum (0.0001756), Frontal Cortex (0.0008629), Pituitary Gland (0.0002002), Colon (7.8534e-05), Kidney (0.00062155), Lung (0.00031105), Muscle (6.729198e-06), Spleen (0.0003099472) |
| PLEKHA1 (Pleckstrin Homology Domain Containing A1) | Frontal Cortex | 3.393e-08 | Protein coding | PDZ-domain binding, phosphatidylinositol-3,4 bisphosphate binding | Class I PI3K signaling events, iCos-iCosL Pathway in T-Helper Cell | Age-Related Macular Degeneration, Eye Disease | Cerebellum (6.097e-07), Pituitary Gland (8.679e-07), Colon (0.002863), Heart (2.3351e-07), Liver (4.5432e-05), Muscle (5.65305e-08) |

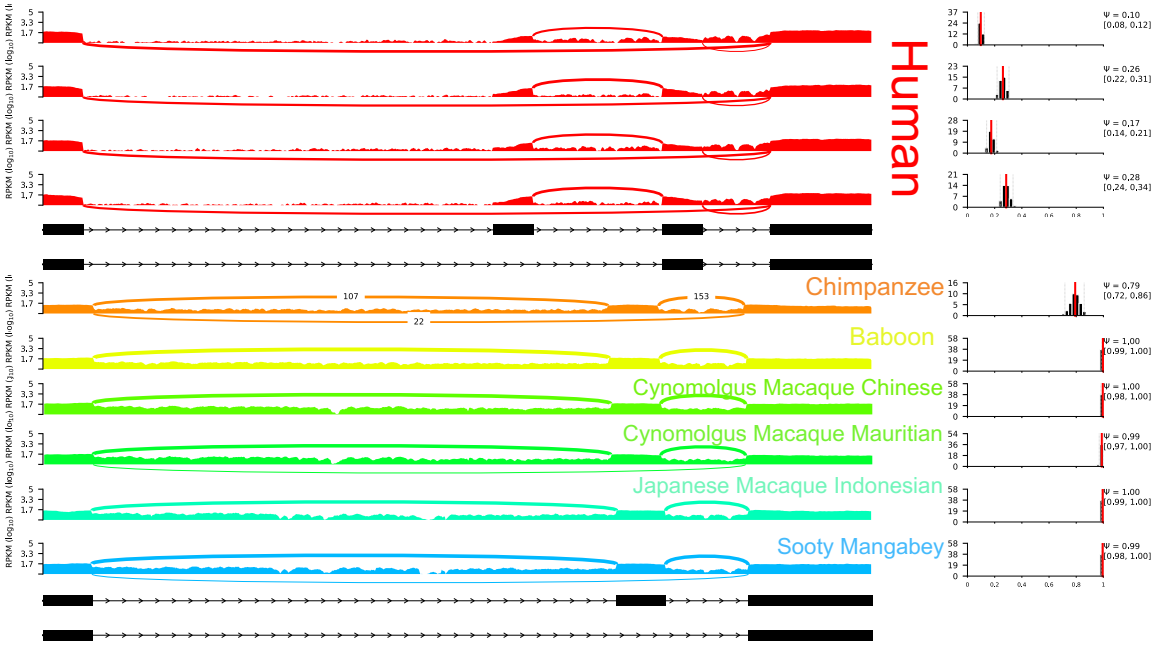


**Figure 5.1:** Sashimi plot for exons 6, 7, and 8 in IFI16 in spleen for six species. All species used their latest genome build while Sooty Mangabey and Japanese Macaque Indonesian RNA-Seq reads were aligned to rheMac8.

### 5.3.1 Gene duplication within IFI16 creates a human-specific exon.

Our top differentially spliced gene was IFI16 (in heart, p-value=9.023x$10^{-16}$). It was also significantly differentially spliced in cerebellum, frontal cortex, pituitary gland, colon, kidney, liver, lung, muscle, and spleen. Sashimi plots of spleen revealed that non-human primates are missing an exon upstream of the alternatively spliced exon, and include the alternative exon in most transcripts (**Figure** 5.1). All species that have a greater divergence time from human than chimpanzee almost always splice-in the alternative exon ($0.99 >= \Psi <= 1$), while chimpanzee has a decreased $\Psi$ value of 0.79. Marmoset, squirrel monkey, and mouse lemur also show high values of $\Psi$ ($\Psi > 0.95$) in all tissues (not shown). IFI16 encodes a nucleic acid sensor that is essential for directly or indirectly mediating the responses against viruses and bacteria. It plays a central role in the immune response to herpesviruses. Previous studies have shown that IFI16 carries a polymorphic segmental duplication of exon 7 (**Figure** 5.2; [4]). Exon 7 carries a 56-amino acid motif which encodes the central "hinge" domain of the molecule. The exon 7 duplication results in differing the size of the hinge domain, which separates two conserved copies of the hematopoietic interferon-inducible protein (HIN-200) domains [5]. Differential splicing of IFI16 in

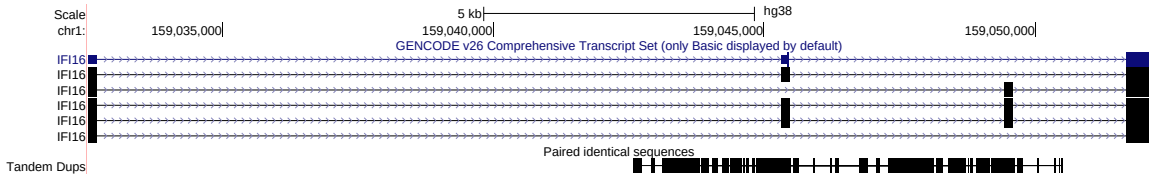


**Figure 5.2:** Humans carry a polymorphic segmental duplication of exon 7 in IFI16.

humans has been shown to generate three protein isoforms from exon 7 which either result in exon 7 being completely spliced-out or one or both copies of exon 7 being spliced-in [5]. Furthermore, Johnstone et al. (1998) [6] reported that the addition of the hinge region changes the functional properties of IFI16 and its ability to act as a transcriptional repressor. They speculated that the three splice variants possibly change the conformation of the molecule, and it has also been observed that the three isoforms have different functional properties [7]. The exon 7 segmental duplication is polymorphic in humans and has been reported as a risk variant for celiac disease and rheumatoid arthritis [8]. Additionally, Kimkong et al. (2009) [9] observed an upregulation of the isoform that results from one copy of exon 7 being spliced-in in the inflammatory disease systemic lupus erythematosus. Interestingly, Cagliani et al. (2014) [4] reported the presence of a duplicated exon 7 in gorilla and orangutan but single copies of exon 7 in chimpanzees, cynomolgus macaque, and vervet. Numerous studies have reported that IFI16 is under strong positive selection [4, 10]. Even though there are low and high levels of nucleotide diversity throughout IFI16, the duplicated exon 7 is flanked by low levels of nucleotide diversity (**Figure 5.3**). Cagliani et al. (2014) have reported that IFI16 is under positive selection in primates, but they observe that the exon 7 segmental duplication is neutrally evolving or is subject to recent or weak selection. We sampled a SNP from the duplicated exon (rs199769901) to test for recent positive selection using 1000 genomes data. We observed a strong signal of positive selection from a preservation of the extended haplotype homozygosity (**Figure 5.4**).

### 5.3.2 Significant change in MLH3, a member of a conserved family of genes involved in the mismatch repair system

Mutant L homolog 3 (MLH3) is a member of a family of evolutionarily conserved proteins that has roles in both DNA mismatch repair and meiosis. MLH3 has been proposed to be involved in the repair of insertion-deletion errors at microsatellite

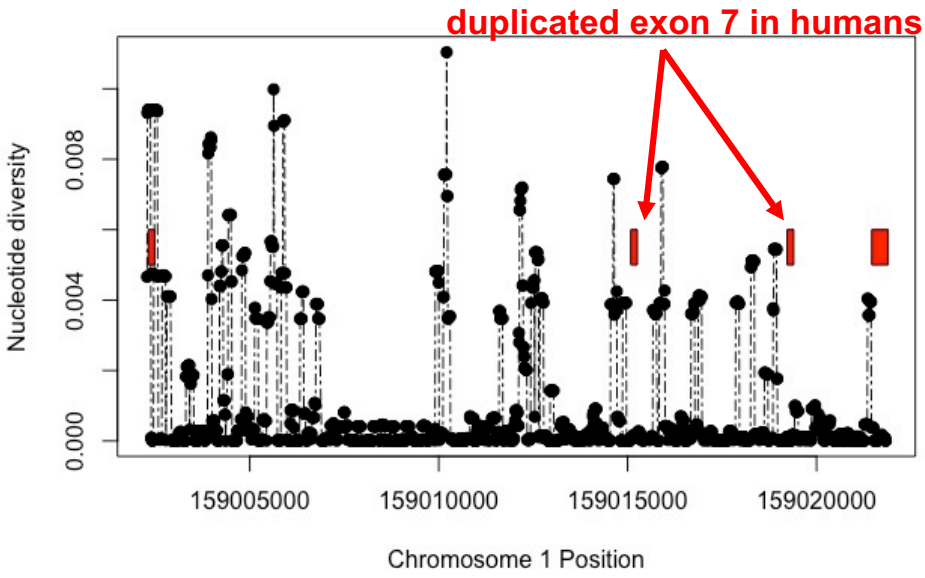


**Figure 5.3:** Nucleotide diversity in IFI16. Red boxes are locations of exons.

repeat sequences [11]. MLH3-deficient mice result in meiotic arrest and aneuploidy [12]. Additionally, mismatch repair genes have been postulated to play a role in human fertility both through spermatogenesis and female infertility. Polymorphisms in this gene have been associated with an increased risk for female infertility [13]. We observed a significant difference in $\Psi$ in humans for exon 3 of MLH3 in cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, liver, lung, muscle, and spleen (**Figure** 5.5). Kansal et al. (2015) [14], identified an infant with MLH3 variants (one heterozygous missense mutation in exon 3) which displayed severe developmental delay along with many tumors in the cerebellum, brainstem, lumbar spine, and in additional tissues. Multiple alignments with 17 primates near the 5′ end of the skipped exon revealed 3 fixed human-specific mutations (dbSNP reveals no polymorphisms identified) within 25bp of the splice site. A search for intronic splicing elements in non-human primates revealed that one of the fixed mutations falls within a motif of a binding site for splicing silencer hnRNP A1 (Score of 97.14/100; **Figure** 5.6). The disruption of an hnRNP A1 binding site in such proximity to the splice site as well as the other fixed changes in an otherwise highly conserved region might be responsible for the dramatic change observed for exon 3 inclusion of this gene in humans. Since MLH3 plays such an important role in fundamental biological processes, it is surprising that we observe this change in exon inclusion in humans in

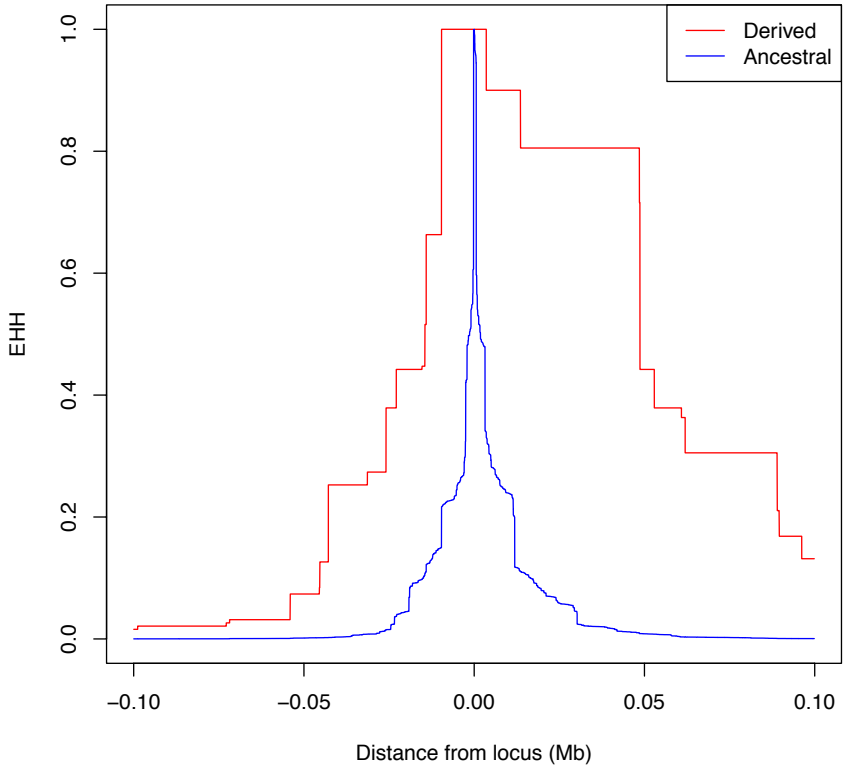


**Figure 5.4:** Extended Haplotype Homozygosity (EHH) plot of rs199769901.

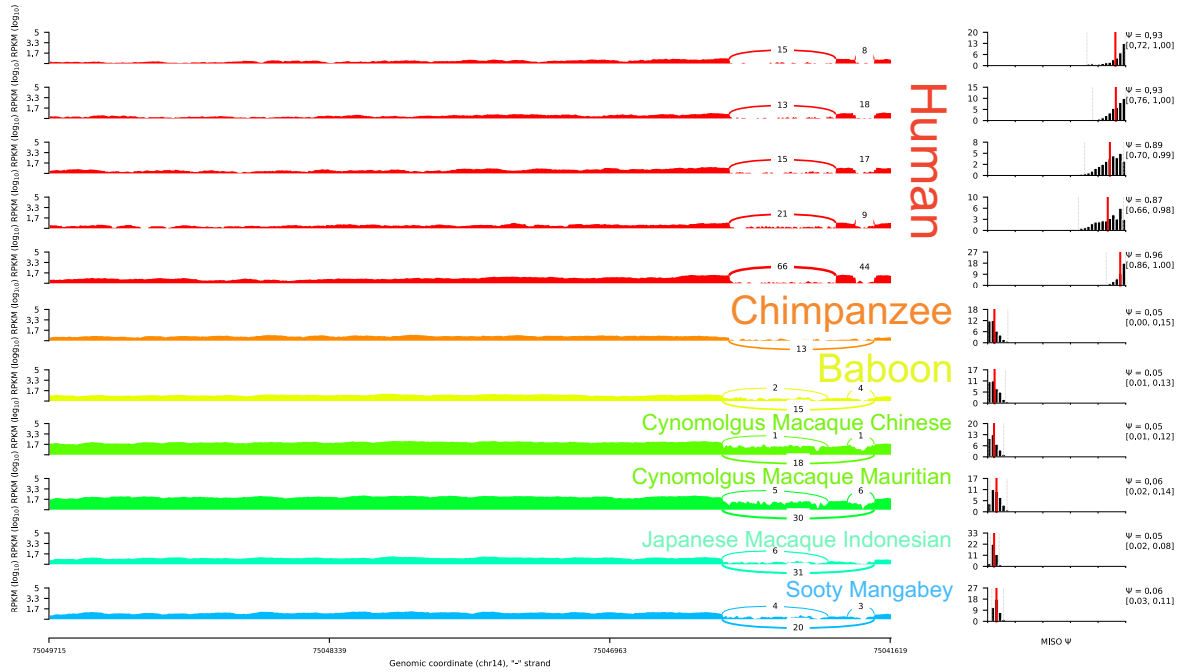


**Figure 5.5:** Sashimi plot for MLH3 in pituitary gland.

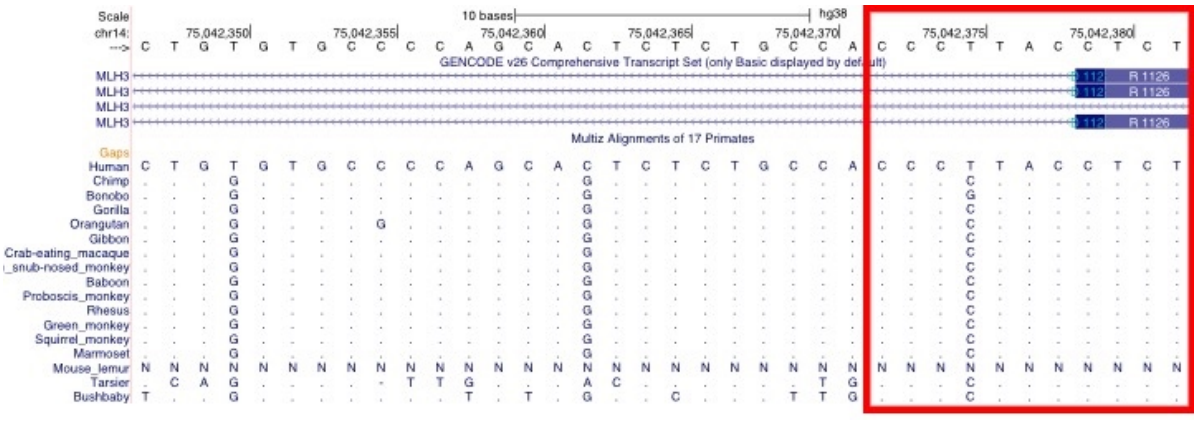


**Figure 5.6:** Multiple alignment with 17 primates for 5′ region of alternative exon for MLH3. Red box indicates where the splicing control element binding site motif was identified.

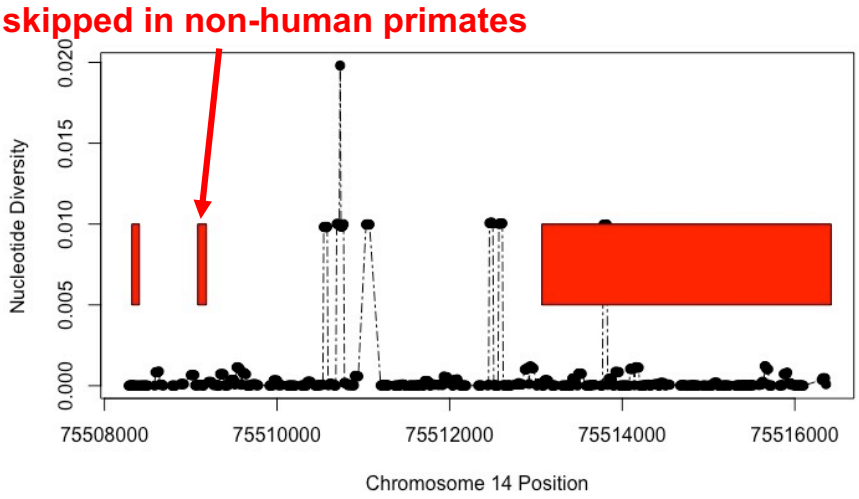


**Figure 5.7:** Nucleotide diversity in MLH3.

every tissue tested. We also observed a dN/dS ratio of 0.5628 in the coding region of MLH3 using multiple alignments of MLH3 with human and 7 non-human primates (orangutan, marmoset, mouse lemur, gibbon, chimpanzee, rhesus macaque, and baboon). Weis et al. (2008) analyzed evolution of MLH3 found an iHS <2, dN/dS of 0.6 in the ORF, sequence divergence of 1.0, 1.7, 1.6, and 2.1 in the 3'UTR, 5'UTR, promoter, and introns, respectively. We also found a dN/dS ratio of 0.5628 in MLH3 using 8 primates. Although dN/dS ratios are low within this gene, the skipped exon is in a region of very low polymorphism (**Figure** 5.7). Humans have a specific set of disease susceptibilities compared to other species. In particular, because of the environment of the modern lifestyle and the extension of life expectancy in humans, humans are more susceptible to obesity, diabetes, cardiovascular disease, cancer, and neurodegenerative disease. Humans have a particularly high rate of spontaneous neoplasms, and this could be attributed to differences in the human cellular response to DNA damage.

### 5.3.3 The birth of a new human exon: HERC2P2

The hect domain and RLD 2 pseudogene 2 (HERC2P2) is highly expressed in all tissues and during human fetal brain development tissues (Brainspan data). It is one

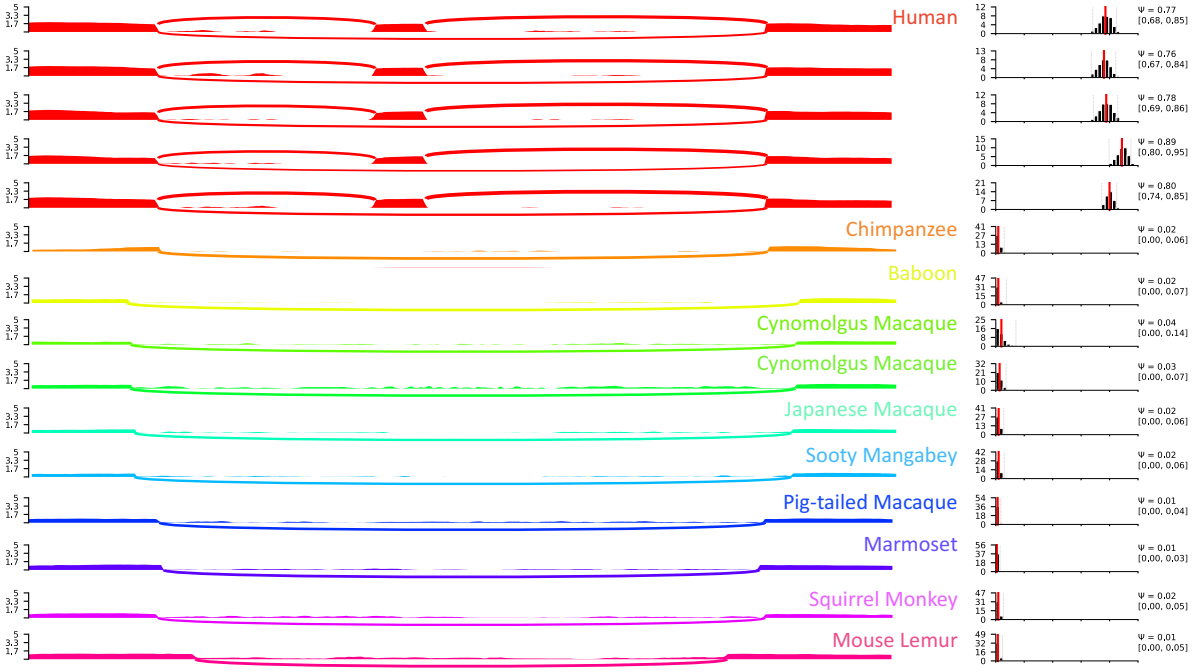


**Figure 5.8:** Sashimi plot of HERC2P2 for 11 species in cerebellum.

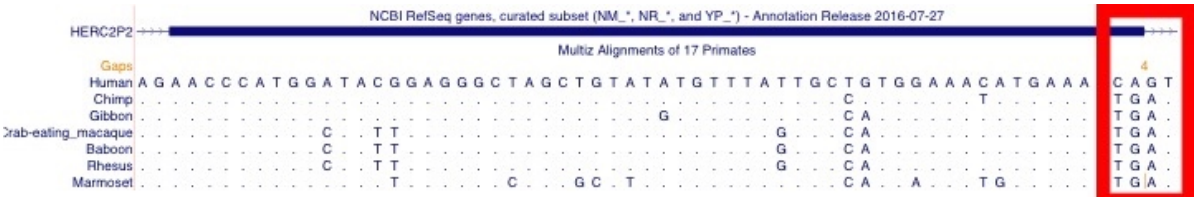


**Figure 5.9:** Multiple alignment of HERC2P2 for 17 primate species. Red box indicates the location of the 5′ splice site creation.

of the few transcribed and processed pseudogenes that has a conserved expression pattern in primates. Even though HERC2P2 is named a pseudogene, it has been functionally validated as a lncRNA [17]. We found that it exhibited one of the strongest shifts in $\Psi$ when testing for differential splicing (**Figure 5.8**). The inclusion of the skipped exon was significantly different in cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, liver, lung, muscle, and spleen. All tissues showed a high inclusion of the exon ($\Psi > 0.75$). None of the non-human primates showed expression at this skipped exon. The multiple alignment of 17 primates revealed three fixed nucleotide changes at the 5′ end of HERC2P2 to create a new splicing site (**Figure 5.9**). The fitCons integrated signal score for the fraction of sites under selection was 0.29 in the skipped exon indicating that the pseudogene has functional potential [15]. Furthermore, the skipped exon is in a region of very low heterozygosity further

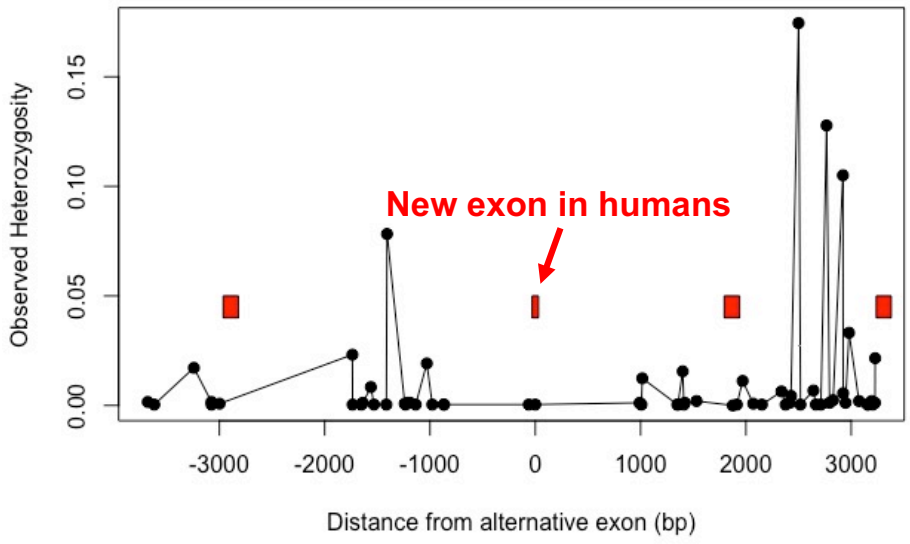


**Figure 5.10:** Low level of polymophisms within 1kb of alternative exon of HERC2P2.

indicating that this is in a conserved region (**Figure 5.10**). The deletion of HERC2P2 has been implicated in many neuropsychiatric and developmental disorders such as Prader-Willi/Angelman Syndrome, schizophrenia and Autism spectrum disorder. HERC2P2 is contained in a well described microdeletion of 15q11 that has also been detected in behavioral and learning problems as well as neurological syndromes such as epilepsy or spastic paraplegia. It has been hypothesized that this 15q11 region is functionally related to the nervous system [16]. Additionally, our analysis from **Chapter 4** indicated that transcribed and processed pseudogenes have higher conservation in brain tissues. It was also upregulated in transcriptional profiling of myocyte enhancer factor 2 (MEF2), a transcription factor that is highly expressed in brain and is fundamental for neuronal survival and synaptic plasticity, in neural progenitor cells [18]. Interestingly, HERC2P2 was one of the most significant genes out of >47,000 genes tested that displayed modulation to exposure to a electromagnetic field in human epidermal keratinocytes [19].

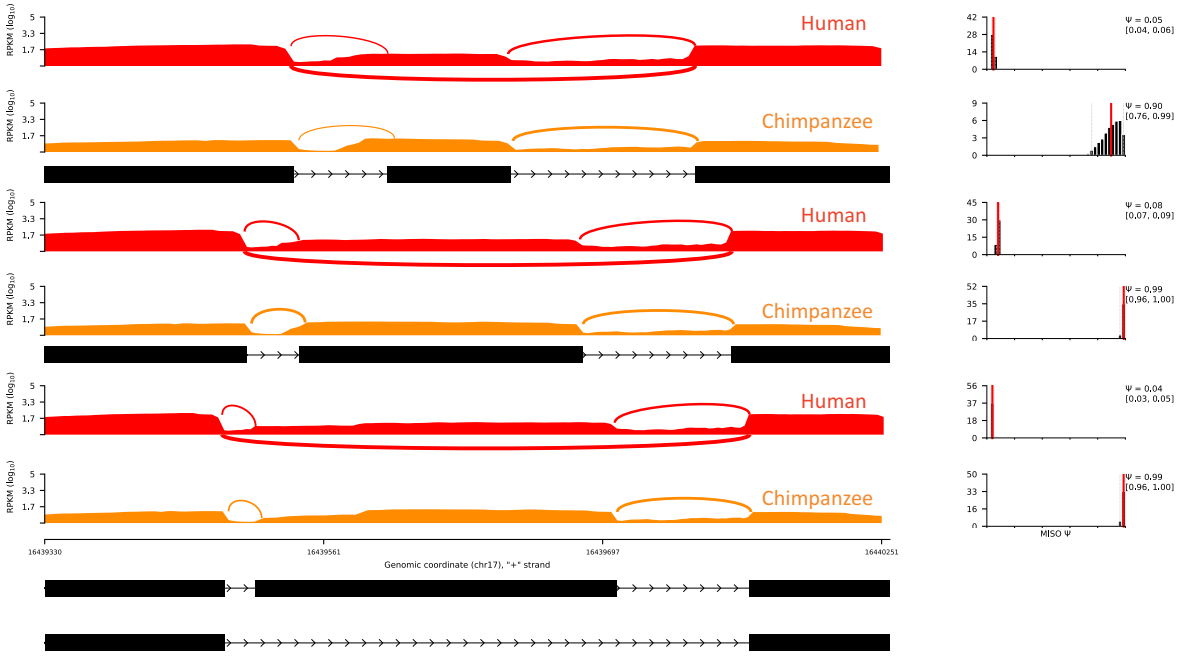


**Figure 5.11:** Sashimi plot of the 4 isoforms in LRRC75A-AS1 in frontal cortex for humans (red) and chimpanzee (orange).

### 5.3.4 Significant decrease in inclusion of exon in lncRNA LRRC75A-AS1.

We observed a highly significant decrease in Ψ in humans in the highly expressed lncRNA, LRRC75A-AS1. LRRC75A-AS1 is a host gene for two snoRNAs, SNORD49A and SNORD49B, that alter the 5′ inclusion of exon 3 to produce 4 different isoforms. All 4 isoforms show significantly different exon inclusion measures between humans and non-human primates (**Figure** 5.11). The inclusion of exon 3 was significantly decreased in human cerebellum, frontal cortex, pituitary gland, colon, heart, kidney, lung, muscle, and spleen. Exon 3 overlaps with SNORD49B (**Figure** 5.12). A fixed difference adjacent to the 5′ splice site in humans creates a putative hnRNP A1 binding site (score of 74.05) was also observed. hnRNP A1 is an inhibitor of splice site recognition. Additionally, the integrated fitCons score for the alternative exon is 0.12324 indicating that a fraction of the sites are under selection, and LINSIGHT scores were high in the snoRNAs. LRRC75A-AS1 is located in a region that is amplified in human high grade osteosarcomas and has been shown to be significantly upregulated in patients with rheumatoid arthritis and in triple negative breast cancer tissues [20–22]. It is also a parent gene for a specific class of intron-derived lncRNAs (sno-

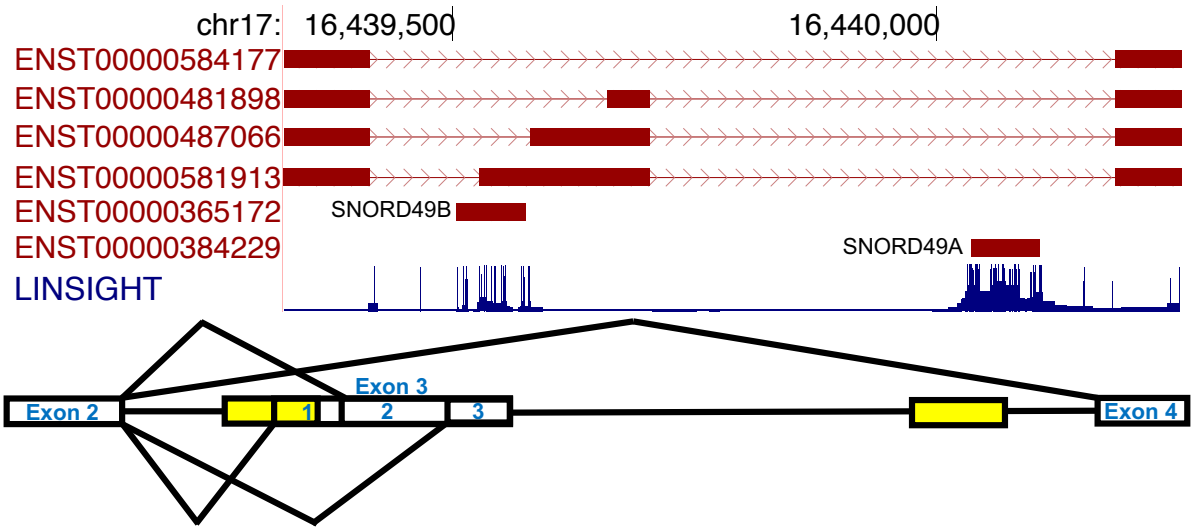


**Figure 5.12:** The different isoforms of LRRC75A-AS1 as a consequence of SNORD49B and SNORD49A and LINSIGHT scores. Yellow indicates location of snoRNAs.

lncRNAs) that depend on snoRNA machinery at both ends for their processing. No sno-lncRNAs have been identified in mouse, and only 19 sno-lncRNAs have been identified in the human genome [23]. Furthermore Zhang et al. (2014) [23] have found that emergence of sno-lncRNAs is often co-encoded with the alternative splicing of their host genes which leads to a species-specific expression. We observe this lineage-specific change between humans and NHPs. Lykke-Andersen et al. (2014) [24] have validated that the 5′ splice site of the skipped exon generates 4 alternative 3′ splice sites that effect the expression of two snoRNAs, SNORD49B and SNORD49A. Since the inclusion/exclusion of the skipped exon is extremely different in humans, this difference in alternative splicing is potentially causing differences in the expression of these snoRNAs.

### 5.3.5 Gene ontology enrichment in differentially spliced genes

We performed a gene ontology (GO) enrichment analysis of the genes that showed significantly different splicing between humans and non-human primates. We analyzed each tissue separately for their enrichment of biofunctions (**Figure 5.13**). A majority of the tissues (except tissues that we identified to be more differentiated in

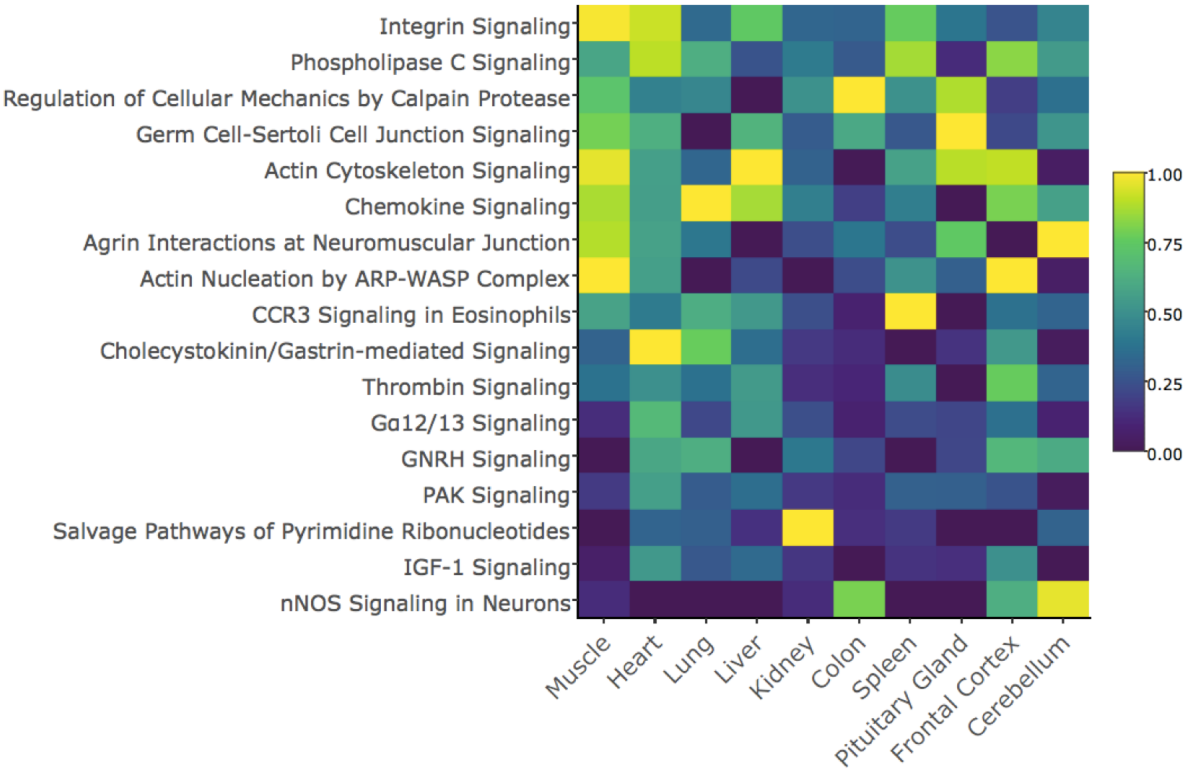


**Figure 5.13:** Heatmap of enriched biofunctions for significantly different splicing. log(p-values) had a threshold of 6 and were normalized.

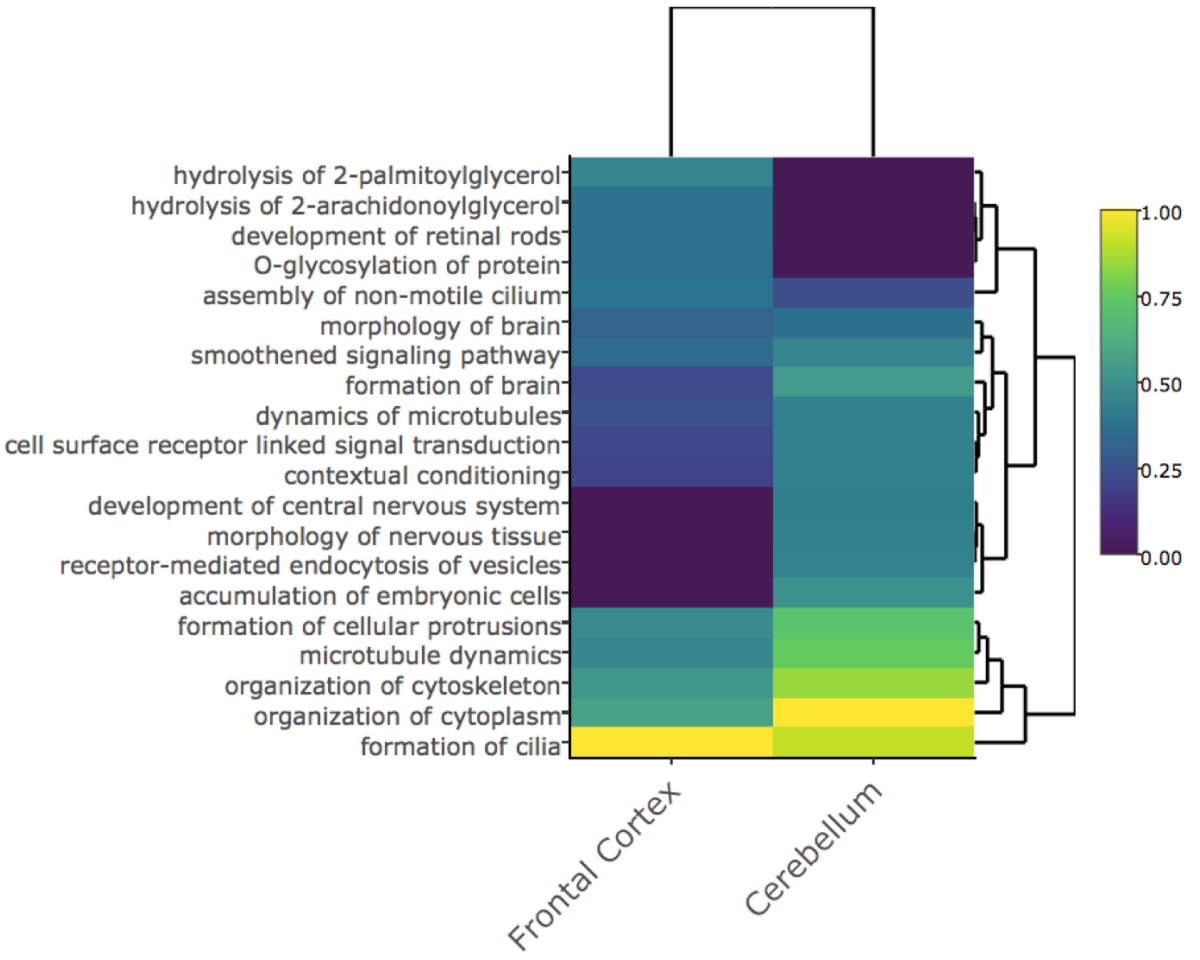


**Figure 5.14:** Heatmap of enriched biological functions for significantly different splicing in cerebellum and frontal cortex. log(p-values) had a threshold of 1.2 and were normalized.

**Chapter 4**: colon, spleen, and lung) showed an enrichment for many highly conserved functions like the organization of the cytoplasm and cytoskeleton, formation of cilia, and microtubule dynamics. Furthermore, even in such a splicing conserved tissue as muscle, there was an enrichment for biological functions involved in the aggregation of myoblasts and the formation of muscle. Biological functions in the heart had a surprising number of biological functions related to the morphogenesis of the heart, specifically, morphology of the heart, QT interval of the heart, and morphology of cardiac muscle. The GO enrichment analysis in cerebellum and frontal cortex also showed enrichment for morphological biological functions related to those tissues (**Figure 5.14**). Along with the biological functions also enriched in other tissues, brain tissues show an enrichment of biofunctions directly related to their formation such as morphology of nervous system, morphology of brain, and formation of brain.

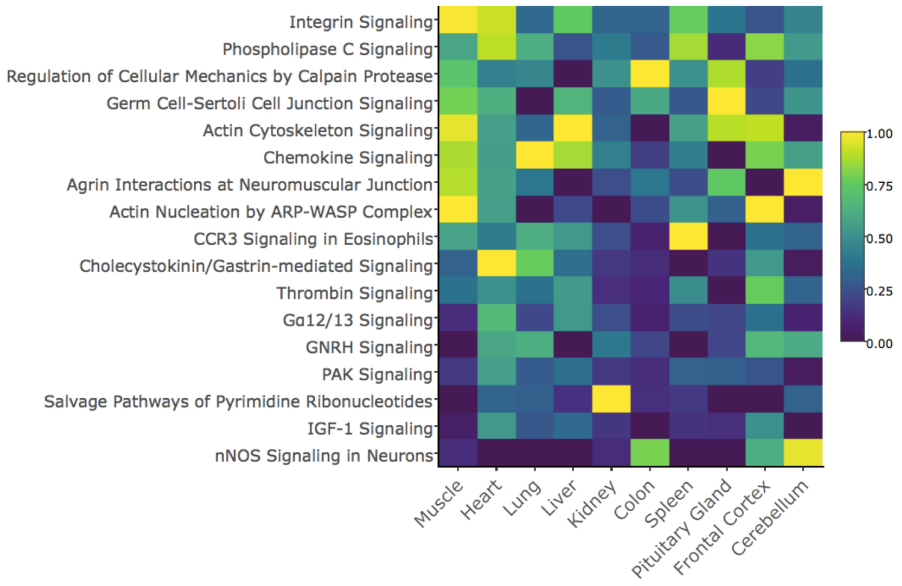


**Figure 5.15:** Heatmap of enriched pathways for differential splicing results.

Both cerebellum and frontal cortex are enriched for formation of cilia, which play an important role in neurogenesis, and neural migration and differentiation. GO pathway analysis revealed many highly conserved signaling pathways (**Figure 5.15**). Our most significant canonical pathway, which was significant in every tissue, integrin signaling, is a surprising result. Integrins are cell surface glycoproteins involved in cell-cell interactions and cell-extracellular matrix interactions. Integrin-mediated signaling has evolved and expanded for 1.5 billion years as the number of interactions has increased due to such processes like gene duplication and alternative splicing that has allowed for integrin-mediated specificity as the developmental complexity of organisms has increased. Integrins contribute to the majority of morphogenetic events that in a developing, multicellular organism [25, 26]. The phospholipase C pathway has also been associated with evolutionary alterations of nucleosome affinities to core promoters in the evolutionary lineage [27]. It is a key signal transduction pathway involved in many sensory stimuli. Given the significant loss of olfactory receptors in humans, changes in this pathway could be related [28]. Also surprising is the significance of enrichment of neuronal nitric oxide synthase (nNOS) signaling in neurons in cerebellum tissue. nNOS is a key regulator of affective behavior, and increasing nNOS in brain has resulted in anxiety even in conditions where the environment is enriched [29].

## 5.4 Concluding remarks

The number of differences in splicing that we observed in muscle and frontal cortex is consistent with differences observed in human muscle and frontal cortex metabolomes [30]. Bozek et al. (2014) observed that the metabolite divergence in human muscle and frontal cortex was much higher than any other tissues, and they posited that it might be a reflection of the specialized endurance capacity of the cardiovascular system and might account for the complexity of the brain. It is not surprising that many of our top differentially spliced genes impact almost all of the tissues that we analyzed. In our specific examples, we have identified mutations that have become fixed in humans, and have either destroyed or created new splice sites or binding sites for splicing control elements. At least two of our top differentially spliced genes have been potentially impacted by hnRNP A1s which have been described as the "Swiss Army knife" of gene expression [31]. Interestingly, in the case of our identification of a new exon emerging in HERC2P2, we observed that Denisovans do not have all of the fixed nucleotide changes that humans have [32]. All of our results seem to be human-specific, and many are in highly conserved genes that are highly expressed in most tissues. It is possible that these switch-like AS events in pivotal genes that we have identified may play a role in rewiring entire programs of gene regulation that can impact the observed differences between humans and NHPs. We welcome further experimental validation, functional interpretation, and regulatory studies from our outcome of our differentially spliced events analysis.

# References


1. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7,** 1009–1015 (2010).
2. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* **37,** e67–e67 (2009).
3. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29,** 308–311 (2001).
4. Cagliani, R. *et al.* Ancient and recent selective pressures shaped genetic diversity at AIM2-like nucleic acid sensors. *Genome biology and evolution* **6,** 830–845 (2014).
5. Gariglio, M. *et al.* Immunohistochemical expression analysis of the human interferon-inducible gene IFI16, a member of the HIN200 family, not restricted to hematopoietic cells. *Journal of interferon & cytokine research* **22,** 815–821 (2002).
6. Johnstone, R. W., Kerry, J. A. & Trapani, J. A. The human interferon-inducible protein, IFI 16, is a repressor of transcription. *Journal of Biological Chemistry* **273,** 17172–17177 (1998).
7. Berry, A *et al.* Interferon-inducible factor 16 is a novel modulator of glucocorticoid action. *The FASEB Journal* **24,** 1700–1713 (2010).
8. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS genetics* **7,** e1002004 (2011).
9. Kimkong, I, Avihingsanon, Y & Hirankarn, N. Expression profile of HIN200 in leukocytes and renal biopsy of SLE patients by real-time RT-PCR. *Lupus* **18,** 1066–1072 (2009).
10. Van der Lee, R., Wiel, L., van Dam, T. J. & Huynen, M. A. Genome-scale detection of positive selection in 9 primates predicts human-virus evolutionary conflicts. *bioRxiv,* 131680 (2017).
11. Lipkin, S. M. *et al.* MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. *Nature genetics* **24,** 27 (2000).
12. Lipkin, S. M. *et al.* Meiotic arrest and aneuploidy in MLH3-deficient mice. *Nature genetics* **31,** 385 (2002).

13. Pashaiefar, H. *et al.* Analysis of MLH3 C2531T polymorphism in Iranian women with unexplained infertility. *Iranian journal of reproductive medicine* **11,** 19 (2013).

14. Kansal, R. *et al.* An infant with MLH3 variants, FOXG1-duplication and multiple, benign cranial and spinal tumors: A clinical exome sequencing study. *Genes, Chromosomes and Cancer* **55,** 131–142 (2016).

15. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics* **47,** 276–283 (2015).

16. Elert-Dobkowska, E. *et al.* Familial 15q11. 2 Microdeletions are not Fully Penetrant in Two Cases with Hereditary Spastic Paraplegia and Dysmorphic Features. *Journal of Genetic Syndromes & Gene Therapy* **5,** 1 (2014).

17. Niazi, F. & Valadkhan, S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3âĂš UTRs. *Rna* **18,** 825–843 (2012).

18. Chan, S. F. *et al.* Transcriptional profiling of MEF2-regulated genes in human neural progenitor cells derived from embryonic stem cells. *Genomics data* **3,** 24–27 (2015).

19. Roux, D. *et al.* Human keratinocytes in culture exhibit no response when exposed to short duration, low amplitude, high frequency (900 MHz) electromagnetic fields in a reverberation chamber. *Bioelectromagnetics* **32,** 302–311 (2011).

20. Both, J., Wu, T., ten Asbroek, A. L., Baas, F. & Hulsebos, T. J. Oncogenic Properties of Candidate Oncogenes in Chromosome Region 17p11. 2p12 in Human Osteosarcoma. *Cytogenetic and genome research* **150,** 52–59 (2016).

21. Zhang, Y. *et al.* Long noncoding RNA expression profile in fibroblast-like synoviocytes from patients with rheumatoid arthritis. *Arthritis research & therapy* **18,** 227 (2016).

22. Lv, M. *et al.* LncRNAs as new biomarkers to differentiate triple negative breast cancer from non-triple negative breast cancer. *Oncotarget* **7,** 13047 (2016).

23. Zhang, X.-O. *et al.* Species-specific alternative splicing leads to unique expression of sno-lncRNAs. *BMC genomics* **15,** 287 (2014).

24. Lykke-Andersen, S. *et al.* Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes & development* **28,** 2498–2517 (2014).

25. Brown, N. H., Gregory, S. L. & Martin-Bermudo, M. D. Integrins as mediators of morphogenesis in Drosophila. *Developmental biology* **223,** 1–16 (2000).

26. Gumbiner, B. M. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell* **84,** 345–357 (1996).

27. Gunbin, K. V. *et al.* Evolution of Brain Active Gene Promoters in Human Lineage Towards the Increased Plasticity of Gene Regulation. *Molecular Neurobiology,* 1–34 (2017).

28. Gilad, Y., Man, O., Pääbo, S. & Lancet, D. Human specific loss of olfactory receptor genes. *Proceedings of the National Academy of Sciences* **100,** 3324–3327 (2003).

29. Workman, J. L., Fonken, L. K., Gusfa, J., Kassouf, K. M. & Nelson, R. J. Post-weaning environmental enrichment alters affective responses and interacts with behavioral testing to alter nNOS immunoreactivity. *Pharmacology Biochemistry and Behavior* **100,** 25–32 (2011).

30. Bozek, K. *et al.* Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. *PLoS biology* **12,** e1001871 (2014).

31. Jean-Philippe, J., Paz, S. & Caputi, M. hnRNP A1: the Swiss army knife of gene expression. *International journal of molecular sciences* **14,** 18999–19024 (2013).

32. Prüfer, K. *et al.* The complete genome sequence of a Neandertal from the Altai Mountains. *Nature* **505,** 43 (2014).

# Conclusion

*"Evolution can seek new solutions without destroying the old…the extra material is scattered in the genome, to be called into action at any time"*

— Walter Gilbert, Why genes in pieces?, Nature 1978

Out of over 200 primates currently existing on earth today, in an evolutionary blink of the eye humans have become the most dominant, adapting to all climates and even changing the environment in a potentially irreversible way. Since the first comparative genomics study by King and Wilson over 40 years ago, researchers have been searching for the missing genetic basis of the origins of human traits. Humans are different from other great apes by many biological characteristics (bipedalism, brain size, hairlessness, opposable thumbs, adolescence, gestation time) and many behavioral traits (tool-making, language, social grouping, and symbolic thought), yet still only a few specific examples (like human-specific Siglec genes and FOXP2) that underlie these characteristics have been identified. Many expected that the sequencing of the human and chimpanzee genomes would reveal these differences. Similarly, many expected that comparing gene expression data through microarrays and now RNA-Seq would reveal these differences. Yet, sequencing of the chimpanzee genome revealed that the two genomes only differed by 1.2% in mostly non-coding DNA and comparative gene expression studies revealed that tissue gene expression is highly conserved even across as much as 300 million years. Although many important and compelling human-specific discoveries have been made (for example loss of olfactory receptors, loss of androgen receptor enhancers and changes in KITLG pigmentation genes), there remain many more questions than answers. Alternative splicing is an ideal candidate that has the ability to create these differences. In fact, for most tissues, overall splicing patterns are species-specific (**Figure 4.4**). Because of short-read sequencing technology and poor annotation in non-human primates, the ability to comparatively study differences in isoforms, has been inherently difficult. With short-read technology, the accuracy in which to assign abundance to actual longer full-length isoforms (not exons) remains extremely poor if not impossible. Only longer and more accurate sequencing reads can solve this problem. The obstacles in data processing, annotation creation, identifying skipped exon events, and orthologous database creation in this project should not be overlooked. However, being able to finally find and study these splicing differences in humans and non-human primates has been rewarding. In this

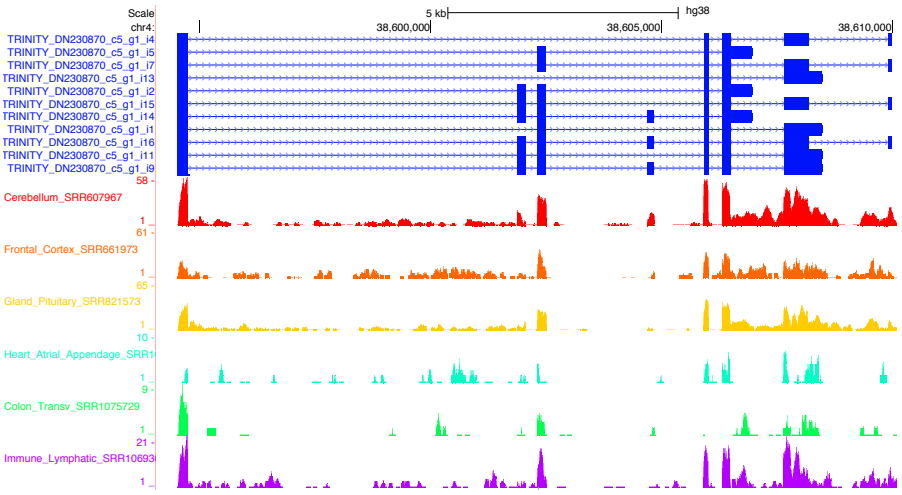


**Figure 5.16:** A new human-specific gene discovered by *de novo* assembly.

thesis, we present a methodology that can systematically study alternative splicing comparatively and our results show that splicing can indeed produce the diversity that can account for the specialized human adaptations. Functionally proving what each of these splicing changes do will be much more difficult. Currently, the vast majority of alternative splicing events have not been functionally characterized on any level. The future of the study of alternative splicing will be in the integration of systems-level approaches to predict "the splicing code" more accurately. These studies will be empowered by incorporating regulatory networks of multiple splicing factors, epigenetic influences, the epitranscriptome, and the kinetics of splicing. In addition to studying the comparative splicing of genes, we have observed many different species-specific highly expressed genes that have yet to be characterized. **Figure 5.16** shows just one example of a human gene that we assembled that is not present in any current annotation and is highly expressed in many different tissues, and is not present in any of our non-human primate assemblies. We have not yet characterized all of the genes even in the best annotated mammalian genome, ours. This just shows the extent to which the complexity of the human genome still needs to be explored.

# List of figures

# List of tables