Transcription of human-specific duplicate genes

Max L. Dougherty

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Evan E. Eichler, Chair

Marshall S. Horwitz

Kelley Harris

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Transcription of human-specific duplicate genes

Max L. Dougherty

Chair of the Supervisory Committee:
Evan E. Eichler, PhD
Genome Sciences

In this work, I set out to characterize new genes contained specifically within the human genome but absent from any other, including our closest evolutionary cousins. Our interest in these genes is twofold: First, gene duplication is a fundamental process by which evolutionary innovations at the organismal level arise; and second, we believe that variation in these new genes is an important and unappreciated source of phenotypic differences between individuals, both normal and pathogenic.

The ability to interpret observed variation in a gene, such as to determine if a mutation in that gene is likely to be impactful, requires quality annotation. This includes an understanding of gene structure: how the gene is transcribed, processed, and which base pairs code for protein or have other specific roles. The major challenge presented by this category of genes is that they are highly

identical to other parts of the genome, and as such, most methods of investigation struggle to tell them apart. As will be expanded upon, this annotation is currently absent or insufficient for genes that are specific to the human genome. Combined with evolutionary and expression analysis, solving this annotation problem enables us to understand how new genes are created in the human genome at the very earliest stages. Combined with surveys of natural occurring and pathogenic variation, it enables us to understand which new genes are functional and which are not, and among those with function, which harbor deleterious variants that can cause disease.

With these goals in mind, I set out to solve the annotation of human-specific duplicate genes most promising for functional status. I performed a close study of one such gene, *HYDIN2*, where I present an evolutionary analysis that gives insight into gene creation and associated disease mechanism, present a survey of naturally occurring variation, and describe the complex transcriptional pattern of a gene that serves as a case study in how duplication and rearrangement of genome segments can lead to rapid gene innovation. Next, I present a technique to more rapidly and rigorously study the transcription of any recently created duplicate gene. I apply this technique to the body of human-specific duplicate genes as well as other expanded gene families. I show that by improving upon current annotations we gain insight into the structural history, expression pattern, and functional status of such genes. I conclude with logical next steps and promising future directions. Ultimately, this work increments our understanding of how gene duplication leads to evolutionary innovation specifically in human, the functional impact of these species-specific differences, and how variation in these genes can contribute to disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Forsan et haec olim meminisse iuvabit
–Aeneas, The Aeneid

ACKNOWLEDGEMENTS

I am indebted to many people who have provided me with support and assistance during the completion of my PhD, including those that have instructed me directly, led by example, sparked insightful conversation, or simply provided encouragement.

I am grateful to my thesis advisor Evan Eichler, who accepted me into his lab merely because I was interested and has fostered an environment in which I have learned an enormous amount. His example as a leader, a manager, a hard worker, and most of all as a scientific mind, will always be with me. I am particularly grateful for the close attention he shows graduate students like myself, and his patience and persistence in working to help me become a better scientist.

I thank the members of the Eichler lab that were instrumental in getting me situated in lab. Xander Nuttle guided me through my first steps as a computational biologist and, along with other graduate students Peter Sudmant and Michael Duyzend, created a welcoming and supportive environment. I would also like to thank Megan Dennis, Fereydoun Hormozdiari, Mark Chaisson, and Stuart Cantsilieris for their generosity with me and their example as scientists. I thank Jason Underwood for his collaboration and friendship. Finally, I would like to thank Tonia Brown for the innumerable ways she has provided assistance during this time.

I also owe a debt of gratitude to my past scientific mentors, including Jeff Lichtman and Ted Betley of Harvard University; Max Essex, Raabya Rossenkhan, and Rebecca Mitchell of the Botswana-Harvard Partnership; and Eric Liao and George Kamel of Massachusetts General Hospital.

I am thankful for my thesis committee members—Marshall Horwitz, Kelley Harris, Mike MacCoss, Cole Trapnell, and Josh Akey—for their attention and guidance. And I am also grateful to Brian Giebel for his frequent assistance.

I would like to acknowledge Molly Jackson, my clinical mentor, whose example as a physician, a mentor, and human being I will think of frequently as I return to medical school, and throughout my medical career.

An essential part of this experience has been the friendships that have been created or strengthened during my time here. I have been very fortunate to have friends that I can learn from, and I have learned so much, quite a bit about science, and quite a bit more about other things. In particular I would like to thank Greg Findlay, Andrew Bogaard, Jacob Baudin, Molly Gasperini, and Kiana Mohajeri for all I've gained through knowing them.

Finally, I would not be here were it not for the selfless and unwavering support of my family. I also owe my parents, in particular my father, for my appreciation of the value of curiosity. And I thank them for allowing me to go wherever my interests have taken me.

Chapter 1. INTRODUCTION

Genes found in structurally complex regions of the human genome are reservoirs for unexplored variation. Such regions are dynamic, showing variation among human populations and between humans and our closest primate relatives. But this dynamism derives from a susceptibility to rearrangements that can result in pathogenic structural variants. The main defining feature of these regions is the presence of large, highly identical blocks of duplicated sequence. Such duplications have long been a thorn in the side of complete assembly and annotation of the human genome. Persistent advances in sequencing technology and assembly methods, however, have improved their accessibility. Human-specific duplicate genes are a product of these same processes and represent dramatic changes to the genome that have occurred over a relatively short period of time. Their role in underlying human-specific traits, once theoretical, is now being increasingly explored and frequently validated. And we are just beginning to be able to better assess their roles in disease, particularly diseases of neurodevelopment. In this introduction, I will begin by discussing the many ways that impactful genomic differences between humans and other species have been sought out, followed by how duplicate genes have been thought to create evolutionary novelty, and finally I will describe what human-specific duplicate genes are and why we believe they are strong candidates for playing a role in both evolution and disease.

1.1 IDENTIFYING HUMAN-SPECIFIC DIFFERENCES

Two of the more surprising findings of modern genetics are perhaps most surprising because of our anthropocentric tendencies. In the presentation of the initial draft sequence of the human genome, it was posited that it contained 30,000–40,000 protein-coding genes, despairingly “only

about twice as many as in worm or fly" (Lander et al., 2001). Later, of course, that estimate was further reduced (Ezkurdia et al., 2014).

The other surprising finding was the result of a comparative biochemical study carried out in 1975 that sought to measure the genetic distance between human and our closest evolutionary branch, the chimpanzee, using mostly protein polymorphisms (King & Wilson, 1975). What was surprising about this was how little difference they found. After the sequencing of the chimpanzee genome, we can now say that 29% of proteins are identical between the two species and that on average proteins differ by about two amino acids (Chimpanzee Sequencing and Analysis Consortium, 2005).

These observations have led to many alternative explanations as to what differences in the genomes of the two species are responsible for our apparent phenotypic differences. The most striking phenotypic difference, and the one most closely tied to our sense of self, is our increased cognitive capacity, and all that has followed as a result. Thus, a focus on genes that are responsible for differences in the human brain specifically is a prominent feature of such comparative studies.

Despite the small number of protein sequence differences between human and chimpanzee, it may be that some of these differences are particularly impactful, though identifying those specific impactful differences from sequence alone may be challenging, as only a small fraction of such differences appear to be under positive selection (Chimpanzee Sequencing and Analysis Consortium, 2005). One such difference may be found in *FOXP2*, a highly conserved transcription factor, mutations in which have been associated with specific language impairments and which contains two amino acid substitutions on the human lineage specifically, despite very

high conservation throughout mammals (Enard et al., 2002). Similar examples, however, are few.

Alternatively, it may be that regulatory differences between otherwise highly similar proteins—when, where, and to what degree they are expressed—are behind the observed differences (Carroll, 2008; King & Wilson, 1975). Efforts to identify key human-specific regulatory changes follow a similar logic; genomic sequences are identified that are highly conserved among other species but contain an excess of differences in human and designated human-accelerated regions (HARs) (Pollard et al., 2006). One such HAR was found to be an enhancer of *FZD8*, and in transgenic mice, *FZD8* under the control of the human (relative to the chimpanzee) form of the enhancer, drove accelerated cell cycle of neural progenitor cells and resulted in increased brain size (Boyd et al., 2015). However, most HARs, the majority of which are intergenic, still have unknown consequences (Levchenko et al., 2018).

Finally, another proposed source of phenotypic differences is gene loss, due to the frequency of spontaneous loss-of-function mutations and their ability to produce rapid phenotypic change (Olson, 1999). One appealing example of this is the gene *MYH16*, a form of myosin expressed in jaw muscles that contains a human-specific frameshift mutation, which is believed to be responsible for the reduction of type II muscle fibers in human and slighter jaw musculature, leading to a weaker jaw, and potentially relieving the cranium of the structural requirements that constrain cranial size (Stedman et al., 2004).

While each of these sources of human-specific variation has yielded interesting anecdotes of specific, likely impactful changes, the contribution of each type to the totality of human-specific phenotypic differences remains unclear. Here we choose to focus on gene duplication as a major but underexplored source of this variation.

1.2 Duplicate Genes and Evolutionary Novelty

Most prominently associated with the notion that gene duplication is the primary mechanism by which evolutionary novelty arises is Susumu Ohno, who proposed that purifying selection is such a dominant force that it is rare for novel functions to arise spontaneously from accumulated mutations. However, following duplication of a gene, the redundancy provided by the extra copy creates a permissive state where one gene is free to accumulate previously forbidden mutations, while the other can maintain the ancestral role (Ohno, 1970). The strongest advocate for the importance of gene duplication, Ohno went so far as to say that the evolution of complex multicellular organisms could not have occurred without gene duplication, popularizing the idea, though earlier related speculations had arisen following observations of duplicated sequence in *Drosophila* (Bridges, 1936; Stephens, 1951).

A different model, in which the multiple functions of an ancestral gene are partitioned between the duplicate copies, was proposed after the observation of a single-copy gene in chicken, whose embryonic expression pattern was recapitulated by the sum of the expression pattern of the two duplicate copies in zebrafish (Force et al., 1999). Called the duplication-degeneration-complementation model, it holds that separate, degenerative mutations in regulatory sequence result in the complementary loss of expression for either copy and allow each new copy to further specialize in their new role. A thematically similar partitioning has been observed in an alternatively spliced transcription factor found at single copy in the human genome; following duplication in fish, each duplicate has taken over one of the two isoforms in human (Dermitzakis & Clark, 2001). The various expected outcomes of gene duplication are summarized in **Figure 1.1**.

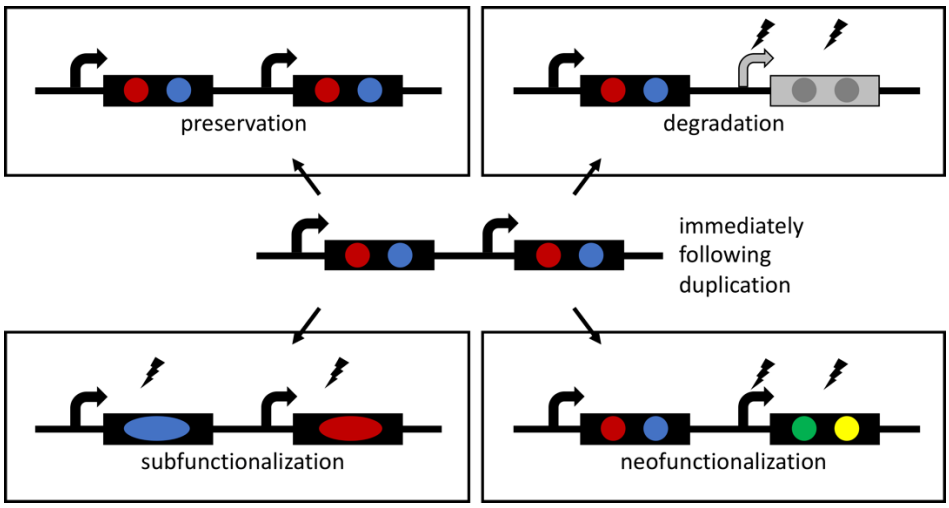


Figure 1.1. Potential fates of duplicate genes.

Shown is a schematic of a gene immediately following duplication, where the bent arrows represent the promoter and the black rectangles represent the gene body, while colored circles represent gene functions and lightning bolts represent mutations. Starting from the upper left, the simplest fate is preservation of function, which results in an effective increase in dosage. The most common fate for a duplicate gene is believed to be degradation, where the gene is silenced by mutations that affect its expression and/or protein-coding capacity. Alternatively, the new duplicate pair may undergo subfunctionalization, where each copy specializes in one of the functions of the ancestral gene. Finally, a duplicate gene can undergo mutations that confer entirely novel function, known as neofunctionalization.

In general, these models focus on the gradual accumulation of mutations following the duplication event, occurring over timescales much longer than human–chimpanzee divergence time of ~6 million years. It is possible that immediate consequences of gene duplication may be more relevant at this timescale, such as the case of the *Drosophila* gene *jingwei*, which was created following a gene duplication that itself was subject to an in-frame retrogene insertion ~2 million years ago (Long & Langley, 1993). Immediate change such as this bypasses the slow accumulation of mutations as proposed by the earlier models of gene duplication.

That the immediate consequences of gene duplication (the state following the duplication event but before subsequent mutations have accumulated) may be important for the most recently duplicated genes is supported by the evolutionary history of the duplicate gene *SRGAP2C*, created ~2.4 million years ago in the human genome (Dennis et al., 2012). *SRGAP2C*

is a truncated copy of its ancestral gene and, consequently, produces a truncated protein. Specifically, compared to the ancestral protein, SRGAP2C has retained the ability to dimerize (with both itself and the ancestral protein) but lost its effector domains. Thus, by interacting with the ancestral protein, this duplicate gene acts in a dominant negative fashion to antagonize it (Charrier et al., 2012). This provides a glimpse into how partial gene duplication can significantly change duplicate genes at the time of their birth, thus leading immediately to novel functions for the duplicate copy. It is worth noting, however, that subsequent amino acid substitutions appear to have strengthened its effectiveness (Charrier et al., 2012; Sporny et al., 2017). The importance of the immediacy of change appears to be a recurring theme in the study of human-specific duplicate genes, which will be expanded upon in Chapters 2 and 3.

1.3 The Challenges and the Promise of Human-Specific Duplicate Genes

Human-specific duplicate (HSD) genes are genes that have been created by duplication of genomic segments containing genes in whole or in part, where the duplication event has occurred on the human lineage, that is, after the evolutionary divergence from chimpanzee approximately 6 million years ago (Dennis et al., 2017; Sporny et al., 2017). There are estimated to be about 218 HSD genes (Dennis et al., 2017; Sudmant et al., 2010), ranging in size from 5 kbp to 362 kbp, and in age from 5.3 to 0.3 million years, though there are likely more yet to be discovered. HSD genes are currently a frontier in our understanding of two types of variation: the variation that exists between the human species and its closest evolutionary cousin, the chimpanzee, and the variation that exists between individuals and populations within the human species. The former addresses essential questions about human origin, identity, and uniqueness. The latter

addresses currently unexplored variation that contributes to phenotypic differences between individuals, including pathogenic variants.

While the premise behind our interest in HSD genes is that regions of the genome that are most rapidly changing are likely to contain important differences between the genomes of species, varied studies of human-specific aspects of brain development have converged upon this group of genes (Charrier et al., 2012; Florio et al., 2015, 2018; Ju et al., 2016; Bhaduri et al., 2018). HSD gene families are enriched for roles in neuronal cell death and neurological disease (Sudmant et al., 2010). This, combined with the fact that species-specific gene duplications have frequently been found to have been drivers of distinctive species-specific traits (Charrier et al., 2012; Chen et al., 2008; Dennis et al., 2012; Duda & Palumbi, 1999; Florio et al., 2015; Ju et al., 2016; Yim et al., 2014), draws our interest toward these genes as potential drivers of that most distinctive human-specific feature of all: human cognition. However, it is worth noting that HSDs are a diverse set of genes. For instance, some HSD gene families have roles related to immunity and contain variants implicated in autoimmune or immune deficiency-related conditions (Roy et al., 2017; Zhao et al., 2017).

Recently duplicated genes in the human genome were initially identified via two strategies: 1) array comparative genomic hybridization (CGH) using bacterial artificial chromosome (BAC) microarrays combined with detection of increased read depth over individual BACs from whole-genome shotgun sequencing (Sharp et al., 2005; Bailey et al., 2002) and 2) the use of cDNA-based array CGH to identify transcribed genes that have undergone copy number change in primates (Fortna et al., 2004; Dumas et al., 2007).

With the availability of whole-genome sequencing data from diverse humans and nonhuman primates, regions containing HSD genes can be identified by a specific read-depth signature,

where increased read depth is observed in particular genomic segments in humans, but not in chimpanzee and other closely related primate species (Sudmant et al., 2010). However, the complex histories of rearrangements that characterize these regions can make them difficult to detect (Dennis et al., 2017). Part of the reason underlying this is that HSD genes, and indeed all segmental duplications, are not distributed randomly but rather cluster, a reflection of the propensity for duplicate sequence to catalyze further duplications (Bailey et al., 2002; Jiang et al., 2007).

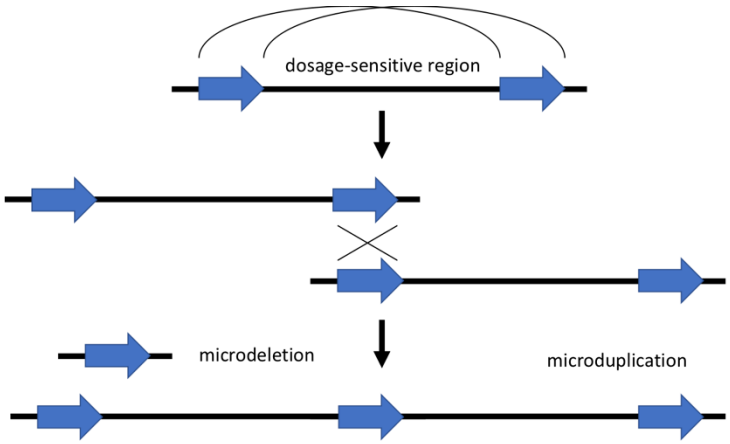


Figure 1.2. Large duplications predispose to further genomic rearrangements.

Large genomic segments of high identity, represented here by block arrows, predispose the genome to further rearrangements by serving as templates for non-allelic homologous recombination. Unequal crossing over can result in deletions or duplications of intervening sequence, as well as other structural variants (not shown). This architecture is responsible in large part for many of the genomic regions whose recurrent rearrangements are associated with neurodevelopmental disease (see also Figure 1.3).

Puzzlingly, one certain consequence of the presence of HSDs throughout the human genome in this clustered fashion is increased genomic instability (**Figure 1.2**). Specifically, the presence of long stretches of highly identical sequence dispersed through the genome, with intervening, gene-containing, non-duplicate sequence, promotes unequal crossing over, mediated by non-allelic homologous recombination (Lupski & Stankiewicz, 2007). Indeed, this is reflected in the extensive overlap between regions of the genome that harbor human-specific duplications and

regions that are recurrently rearranged with pathogenic consequences, including common microdeletion and microduplication syndromes (**Figure 1.3**; Dennis et al., 2017). This observation, that the duplications that give birth to HSD genes also predispose the genome to pathogenic rearrangement, shows that they are in fact a double-edged sword, and has led to the hypothesis that the fitness gain provided by the creation of these new genes outweighs the fitness cost of the increased frequency of congenital disease caused by the genomic architecture these duplications engender (Marques-Bonet et al., 2009).

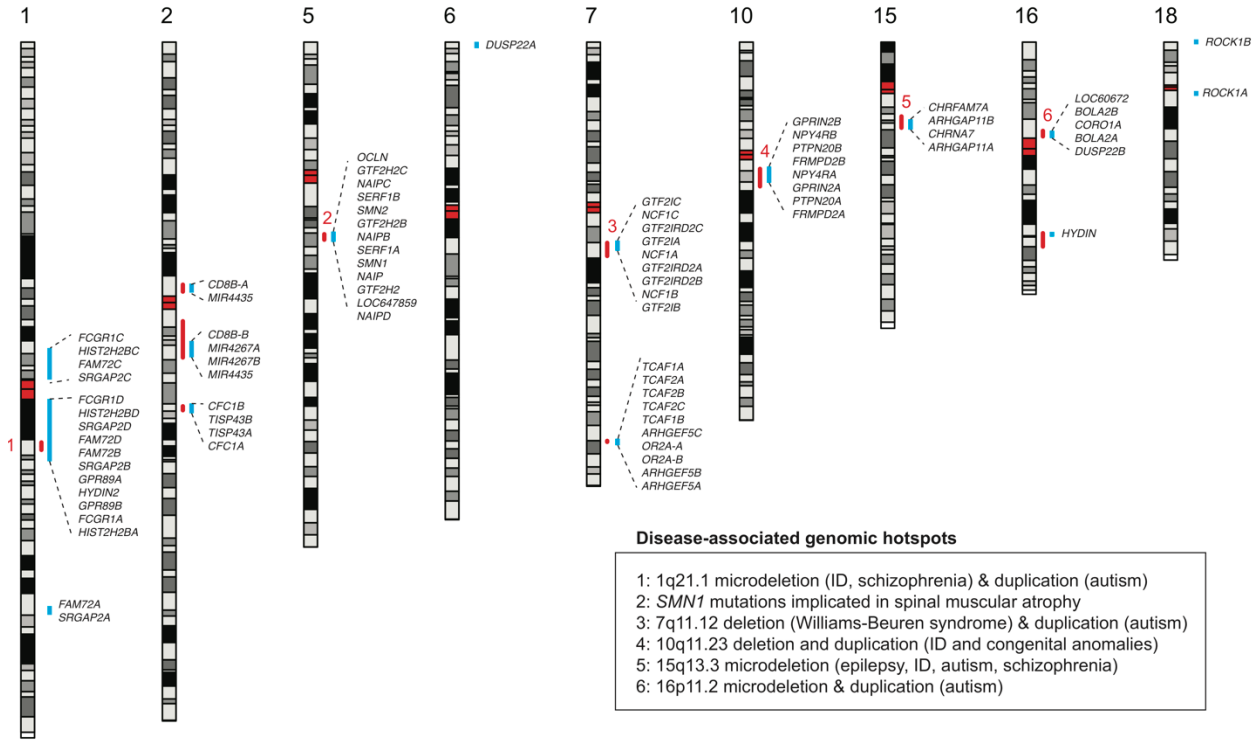


Figure 1.3. Extensive overlap between disease-causing and gene-creating genomic rearrangements.

The locations of large, gene-containing HSDs are highlighted (blue lines) across nine different human autosomes. Many of these HSDs overlap known disease-implicated genomic hotspots (red lines) prone to recurrent copy number variation associated with developmental delay. The genomic hotspots labelled with red numbers (1–6) have significant associations with specific disorders including epilepsy, autism, schizophrenia and intellectual disability (ID). Adapted from Dennis et al., 2017.

What follows then is to understand the roles of each one of these HSD genes. Which are functional, and which are subject to purifying, neutral, or positive selection? What specific evolutionary processes created these new genes? What is the spectrum of natural variation in HSD genes and do they contain pathogenic variation as well? And finally, what differences in phenotype are derived from the presence of these genes in the human genome? To begin to answer these questions requires first and foremost accurate sequence data, at both the genomic and transcript level. This work takes up at a point when many HSDs have been correctly sequenced and assembled, but their transcription is still poorly understood, therefore most of this work focuses on the study of HSD gene transcription.

1.4 Topics in this Dissertation

Dissecting the role of HSDs in human evolution and disease requires understanding their evolutionary history, the nature of their transcription, and correct annotations that allow us to interpret variation. The work described here seeks to advance the field of HSDs in these three areas. Because of the inherent difficulties in studying large, nearly identical duplications, these genes tend to be less well characterized (Dougherty, 2018 *submitted*).

Chapter 2 describes a focused study of a particularly intriguing HSD gene (as well as the largest), *HYDIN2*. Through careful determination of the evolutionary history of that region, we show how *HYDIN2* is a large chimeric gene, created through fusion transcription between multiple segmental duplication blocks that came together in a stepwise fashion. We show how this amalgam of genomic duplications led *HYDIN2* to gain a new promoter that drives a pattern of expression quite different from the ancestral gene, *HYDIN* (a theme that will be further expanded upon in Chapter 3), including dramatically increased expression in the developing brain. It was readily apparent that the gene annotation for *HYDIN2* was inadequate, and so we

used RT-PCR and long-read sequencing to determine, for the first time correct gene models for *HYDIN2* that were required for its study. We explored its role as a potential cause of microcephaly/macrocephaly in chromosome 1q21.1 reciprocal rearrangement syndromes, and used comparative analyses, as well as queried natural variation in coding sequence, to study selection acting on the gene.

Lessons learned from this project then led to the development of the more comprehensive study of HSD gene transcription described in Chapter 3. Here, we developed a new technique, also based on long-read sequencing of RNA, but using probe-based enrichment to target genes of interest and study their transcription in both a broader and less biased way. We present a close study of the transcription of 19 HSD gene families. We find that the themes learned in Chapter 2 are repeated throughout HSD gene families, and that many of the duplicate genes are in fact incomplete copies of the original, and yet still the vast majority transcribed. We show how the extent of the initial duplication event plays a large part in determining the evolutionary course of duplicate genes, particularly with respect to expression. We find numerous cases of missing or incomplete gene annotations, including conserved coding sequence, indicating the utility of our method.

Finally, in Chapter 4 I discuss future directions, specifically the experiments that the knowledge gained from this work allows. This includes 1) the extension of our method to the more complex gene family known as *morpheus*, which has undergone dramatic copy number expansion in primate genomes (Johnson et al., 2001); 2) a long-read-based genotyping method to assign ambiguous variants identified through large-scale short-read-based genotyping to the correct paralog; and 3) the use of single-cell RNA-sequencing methods to identify HSDs critical to neurodevelopment.

Chapters 2 and 3, respectively, have been modified from manuscripts published and under review as of May 2018.

Chapter 2. THE BIRTH OF A HUMAN-SPECIFIC NEURAL GENE BY INCOMPLETE DUPLICATION AND GENE FUSION

Chapter 2 is adapted with minimal modification from:

> Dougherty, M. L., Nuttle, X., Penn, O., Nelson, B. J., Huddleston, J., Baker, C., et al. (2017). The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biology*, *18*(1), 49. http://doi.org/10.1186/s13059-017-1163-9
>
> First authorship is shared between MLD and XN.

2.1 ABSTRACT

Gene innovation by duplication is a fundamental evolutionary process but is difficult to study in humans due to the large size, high sequence identity, and mosaic nature of segmental duplication blocks. The human-specific gene hydrocephalus-inducing 2, *HYDIN2*, was generated by a 364 kbp duplication of 79 internal exons of the large ciliary gene *HYDIN* from chromosome 16q22.2 to chromosome 1q21.1. Because the *HYDIN2* locus lacks the ancestral promoter and seven terminal exons of the progenitor gene, we sought to characterize transcription at this locus by coupling RT-PCR and long-read sequencing. 5'-RACE indicates a transcription start site for *HYDIN2* outside of the duplication, and we observe fusion transcripts spanning both the 5' and 3' breakpoints. We observe extensive splicing diversity leading to the formation of altered open reading frames (ORFs) that appear to be under relaxed selection. We show that *HYDIN2* adopted a new promoter that drives an altered pattern of expression, with highest levels in neural tissues. We estimate that the *HYDIN* duplication occurred ~3.2 million years ago and find that it is nearly fixed (99.9%) for diploid copy number in contemporary humans. Examination of 73

chromosome 1q21 rearrangement patients reveals that *HYDIN2* is deleted or duplicated in most cases. Together, these data support a model of rapid gene innovation by fusion of incomplete segmental duplications, altered tissue expression, and potential sub-functionalization or neo-functionalization of *HYDIN2* early in the evolution of the *Homo* lineage.

2.2 BACKGROUND

Gene duplication has long been hypothesized to be an important source of evolutionary innovation (Ohno 1970). The great ape lineage that includes humans has experienced a surge of interspersed segmental duplications over the last 10-15 million years (Marques-Bonet et al. 2009; Sudmant et al. 2013). While large, highly identical duplications sensitize the human genome to recurrent rearrangements associated with disease (Lupski 1998; Stankiewicz et al. 2004; Bailey et al. 2002; Sharp et al. 2005), they also have the potential to drive the emergence of novel duplicate genes and functions (Dennis and Eichler 2016). The identification of functional duplicate genes, however, is difficult, as these duplications are typically large, highly identical, and clustered into complex mosaic structures juxtaposing sequence blocks of diverse origin (Bailey and Eichler 2006). Furthermore, duplicated regions of the genome are frequently misassembled and are the source of extensive copy number variation in human populations (Sudmant et al. 2015a; Sudmant et al. 2015b).

Considerable attention has been focused on the identification of duplicate genes that have emerged since the human–chimpanzee divergence because of their potential to contribute to human-specific traits (Bailey et al. 2002; Fortna et al. 2004; Sudmant et al. 2010). Already two such genes, *SRGAP2C* and *ARHGAP11B*, have been functionally characterized by heterologous expression studies in mouse suggesting potential roles of the duplicates in increasing dendritic spine density (Charrier et al. 2012) and expanding the number of cortical neurons (Florio et al.

2015), respectively. Both duplicate genes are nearly fixed for copy number in human populations but absent from all nonhuman primates (Florio et al. 2015; Dennis et al. 2012; Dennis et al. 2017; Anotonacci et al. 2014). In addition, both genes carry only a subset of the exons of the ancestral gene due to incomplete segmental duplication of the progenitor locus, supporting the hypothesis that truncation may facilitate neofunctionalization (Dennis and Eichler 2016; Dennis et al. 2012).

Chromosome 1q21.1 is one of the largest regions of human-specific segmental duplication blocks in the human genome (Bailey et al. 2002; O'Bleness et al. 2014), and, as such, is a reservoir for human-specific transcripts and genes (Sudmant et al. 2010; O'Bleness et al. 2014). The presence of large blocks of directly oriented, highly identical duplicated sequence renders this region genetically unstable (Sharp et al. 2005). Specifically, recurrent deletions and duplications occur at this locus and have been associated with cognitive and motor impairment, articulation abnormalities, and hypotonia (Mefford et al. 2008; Brunetti-Pierrri et al. 2008; Bernier et al. 2015). Duplication carriers show an increased prevalence of autism spectrum disorder (ASD) and macrocephaly, while deletion carriers show an increased prevalence of microcephaly.

The human-specific gene *HYDIN2* was previously mapped to this region of chromosome 1q21 (Doggett et al. 2006), but because it was contained within an assembly gap in GRCh37/hg19, its role in 1q21 rearrangement syndromes remained uncertain. It was postulated that *HYDIN2* might contribute to the reciprocal macrocephaly/microcephaly phenotype (Brunetti-Pierri et al. 2008), because mutation of the ancestral *HYDIN* gene leads to hydrocephalus in the mouse (Davy et al. 2003). In humans, however, recessive mutations in *HYDIN* were found associated with primary ciliary dyskinesia without hydrocephalus (Olbrich et al. 2012). As such, whether dosage of *HYDIN2* plays a role in the 1q21.1 microdeletion/microduplication phenotype has remained unanswered.

With the initial discovery of the *HYDIN* duplication (Doggett et al. 2006), two features of the derived locus were noted. First, although the gene duplication was truncated and did not include the promoter, transcription was observed at *HYDIN2*. Second, *HYDIN2* transcripts appeared to derive primarily from neuronal sources, in contrast to the ciliated tissues from which *HYDIN* transcripts were originally cloned. We sought to investigate the origin and significance of *HYDIN2* transcription by reconstructing the evolutionary history of this locus, assessing patterns of human genetic variation in both normal and disease populations, and exploring transcript diversity in human tissues.

2.3 RESULTS

2.3.1 *Molecular evolution and breakpoint analyses*

The ancestral *HYDIN* gene is notable for its large size—its canonical gene structure occupies 423 kbp of genomic sequence and by homology to mouse is predicted to produce a 15,179 bp transcript encoding a 5,121 amino acid protein (**Figure 2.1a**). Comparative sequencing of both human *HYDIN* paralogs as well as the putative integration or acceptor site in chimpanzee shows that the duplicated sequence is 364 kbp long and shares 99.4% nucleotide identity with its ancestral paralog. The duplication includes 79 coding exons but excludes the sole promoter, as well as the canonical polyadenylation site, though shorter isoforms of *HYDIN* with earlier polyadenylation sites are recorded. For transcription to occur, this gene segment must have acquired a new promoter and at least one novel polyadenylation site, potentially from flanking sequences.

The ancestral site on chromosome 16q22.2 is flanked by unique sequence, with no indication of a predisposition to rearrangement. In contrast, the acceptor site occurs within a large segmental duplication block on chromosome 1q21 that has been subject to extensive duplication and

rearrangement over the last 25 million years of evolution (Bailey et al. 2002; Fortna et al. 2004; Dennis et al. 2012; O'Bleness et al. 2012). This includes the hyper-expanded core duplicon for chromosome 1 that carries members of the neuroblastoma breakpoint family (*NBPF*) and its associated DUF1220 protein domain (Vanderpoele et al. 2005; Jiang et al. 2007; Popesco et al. 2006). Copies of *NBPF* map 36 kbp downstream of the insertion site and represent the nearest protein-coding gene (**Figure 2.1a**, see also Appendix A: Figure S1). We refined the breakpoint of the duplication integration by targeted sequencing of chimpanzee bacterial artificial chromosome (BAC) clones. A comparison with the high-quality chimpanzee sequence indicates that the insertion at chromosome 1q21.1 occurred at the boundary between the LTR and LINE repeats, with concomitant loss of 841 bp of LINE sequence (**Figure 2.1b**) in the human lineage.

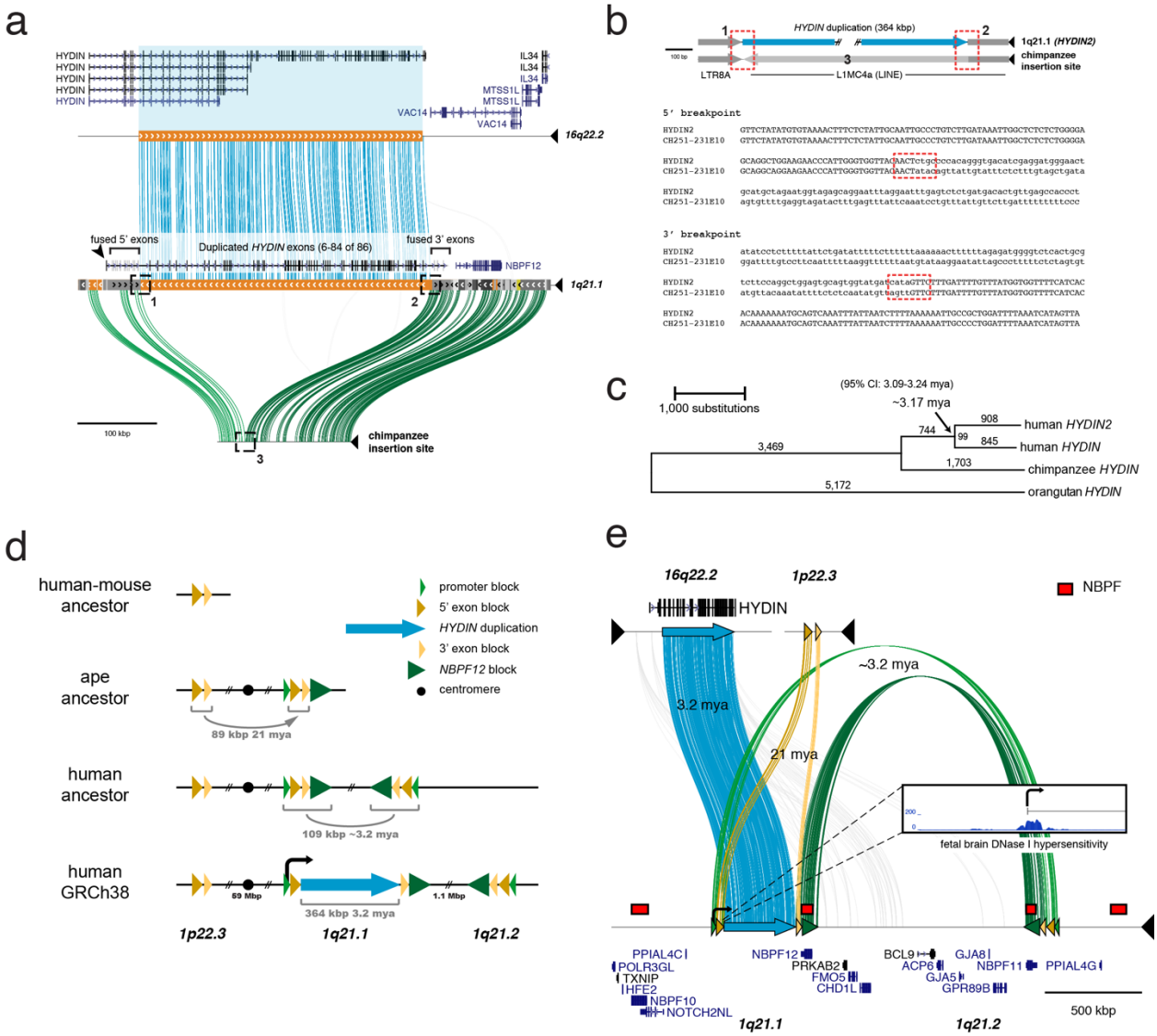


Figure 2.1. *HYDIN* duplication and evolution.

a) Comparison of human genomic sequence from the donor locus on chromosome 16q22.2 (top) to the acceptor locus on chromosome 1q21.1 (middle) shows a 364 kbp duplication (blue lines) including exons 6-84 of the 86-exon ancestral gene *HYDIN*, visualized using Miropeats. Connecting lines indicate nearly identical segments (s = 1,000). The orthologous insertion site (bottom) prior to insertion based on sequencing of chimpanzee BAC (CH251-231E10) is shown with homology indicated (green lines). Human gene annotation (GENCODE) as well as the location of the exons found in fusion transcripts, breakpoints (dashed boxes) and acquired novel promoter (arrowhead) are depicted. **b)** Sequence alignment of the 5' and 3' breakpoints (dashed red box) shows that the duplication integrated at the boundary of an LTR and LINE repeat with the concomitant loss of 841 bp of LINE sequence based on analysis of the chimpanzee orthologous sequence. Uppercase bases are beyond the breakpoint while lowercase bases indicate a break in homology. Numbers are as indicated in Figure 2.1a. **c)** A neighbor-joining phylogenetic tree based on a 315,349bp MSA using the human paralogs as well as orthologous chimpanzee and orangutan sequences. Based on the genetic distance (Kimura-2 parameter) and assuming a human–chimpanzee divergence of 6 mya, we estimate the duplication occurred ~3.17 mya (95% CI: 3.09-3.24 mya, bootstrap method). **d)** The model depicts the simplified evolutionary history of the *HYDIN2* genomic locus as a series of juxtaposed segmental duplications that contributed novel exons and regulatory machinery. Human-mouse comparative sequence analysis (GRCh38/GRCm38) shows that the 5' and 3' exon blocks (yellow arrows) originated as a single ~89 kbp segmental duplication mapping to human chromosome

1p22.3. It was duplicated in the common ape ancestor (~21 mya) from chromosome 1p22.3 to chromosome 1q21.1 in close proximity to the *NBPF* core duplicon and the promoter-containing segment of *HYDIN2* (green arrows). Approximately 3.2 mya, an inverted duplication (109 kbp) occurred with *NBPF* cores defining the breakpoints at chromosome 1q21.1 and 1q21.2. This was followed by the insertion of the 364 kbp *HYDIN* segmental duplication from chromosome 16q22.2. See Appendix A: Figure S2 for phylogenetic analyses. **e)** Miropeats (s = 800) schematic shows the genomic organization of the segmental duplications and surrounding gene annotation. This includes the 364 kbp *HYDIN* segment (blue), the 89 kbp exon-containing segment from chromosome 1p22.3 (yellow), and the larger, 109 kbp, inverted segmental duplication, shared with between chromosome 1q21.1 and chromosome 1q21.2 (green and yellow). Inset shows DHS data for fetal brain in the ~14 kbp surrounding the first exon of *HYDIN2*. The new promoter (bent arrow) corresponds to a peak of chromatin accessibility. (See also Appendix A: Figure S6, Appendix A: Table S8).

To estimate the evolutionary age of the *HYDIN* duplication, we generated high-quality sequence data for the two human *HYDIN* paralogs and built a multiple sequence alignment (MSA) using orthologous sequences from chimpanzee (panTro4) and orangutan (ponAbe2) over the 364 kbp duplicated region. Phylogenetic analysis predicts that the duplication occurred approximately 3.17 million years ago (mya; 95% CI: 3.09-3.24 mya, bootstrap method) (**Figure 2.1c**)—a period corresponding to the transition between australopithecines and the genus *Homo*. The derived *HYDIN2* duplication has inserted into evolutionarily older segmental duplications shared among great apes (summarized in **Figure 2.1d**). Immediately flanking the large central segment are two blocks, referred to here as the 5' exon block (41 kbp) and the 3' exon block (24 kbp), because they provide exons that form fusion transcripts by joining with exons from within the *HYDIN2* duplication. These two blocks were formerly a single segment that was bisected by the insertion of the duplication from chromosome 16.

The segment composed of the 5' exon block and 3' exon block maps to two other locations in the human genome in addition to chromosome 1q21.1 (**Figure 2.1d**): an inversely oriented copy mapping 1.1 Mbp telomerically at chromosome 1q21.2 and another mapping at chromosome 1p22.3. Only the chromosome 1p22.3 locus shares conserved synteny with mouse, where it is found as a single copy and likely represents the ancestral locus. The other two copies on chromosome 1q21 are both associated with *NBPF* and form a larger homologous segment of ~109

kbp in length. Phylogenetic reconstruction predicts that the first duplication occurred ~21 mya (95% CI: 20.6-21.6 mya, bootstrap method; Appendix A: Figure S2a), placing the event at the root of ape lineage after divergence from the Old World monkeys (Glazko et al. 2002). This is consistent with our observation that this segment is found at single copy in New World and Old World monkeys and at two copies in gibbon and orangutan.

The second duplication, from chromosome 1q21.1 to chromosome 1q21.2, shows greater sequence identity (99.6%) than the *HYDIN* duplication (99.4%). Although phylogenetic timing predicts a more recent origin for this duplication (2.31 mya; 95% CI: 2.17-2.45 mya; Appendix A: Figure S2b) than for the *HYDIN* duplication, it must necessarily have occurred prior to or together with the insertion of the *HYDIN* segment from chromosome 16 in order for it to have been disrupted by the insertion. It is likely that interlocus gene conversion between human chromosome 1q21.1 and chromosome 1q21.2 copies distorts the timing of this duplication. Longer stretches of conserved synteny between sequenced nonhuman primate clones and human chromosome 1q21.1 than between the clones and human chromosome 1q21.2 indicate that chromosome 1q21.1 is the more likely ancestral locus. This second duplication extends further upstream to include a segment that provides the new promoter (26 kbp) and further downstream to the segment that includes *NBPF12*. Altogether, the *HYDIN* segment is sandwiched by two layers of duplications—the first 21 million years old, and the other 3.2 million years old—providing the genomic substrates for novel *HYDIN2* flanking exons and regulatory elements (**Figure 2.1e**).

2.3.2 *Copy number diversity and gene conversion in human populations*

We assessed *HYDIN* copy number variation across diverse human populations, archaic hominins, and nonhuman primate genomes by applying whole-genome shotgun sequence detection and singly unique nucleotide k-mer (SUNK) analysis to obtain aggregate copy number estimates

(Sudmant et al. 2010). In contemporary human populations, we observe a narrow distribution of *HYDIN* copy number, centered on individuals having two diploid copies of each paralog for a total of four aggregate copies (**Figure 2.2a**). As expected from our phylogenetic analysis, all nonhuman primates have just two diploid copies of *HYDIN*. Neanderthal and Denisova are believed to have diverged from modern humans approximately 700,000 years ago (Reich et al. 2010; Meyer et al. 2012; Prüfer et al. 2014). Consistent with this, we observe the duplication in their genomes as well as in three archaic human genomes (~7,000-45,000 years old), though with a greater variability, possibly due to the relatively lower coverage and/or quality of these genomes (Fu et al. 2014; Lazaridis et al. 2014). In total, our analysis of 2,401 human genomes from the 1000 Genomes Project (1KG) and the Human Genome Diversity Project (HGDP; Appendix A: Figure S3), humans.

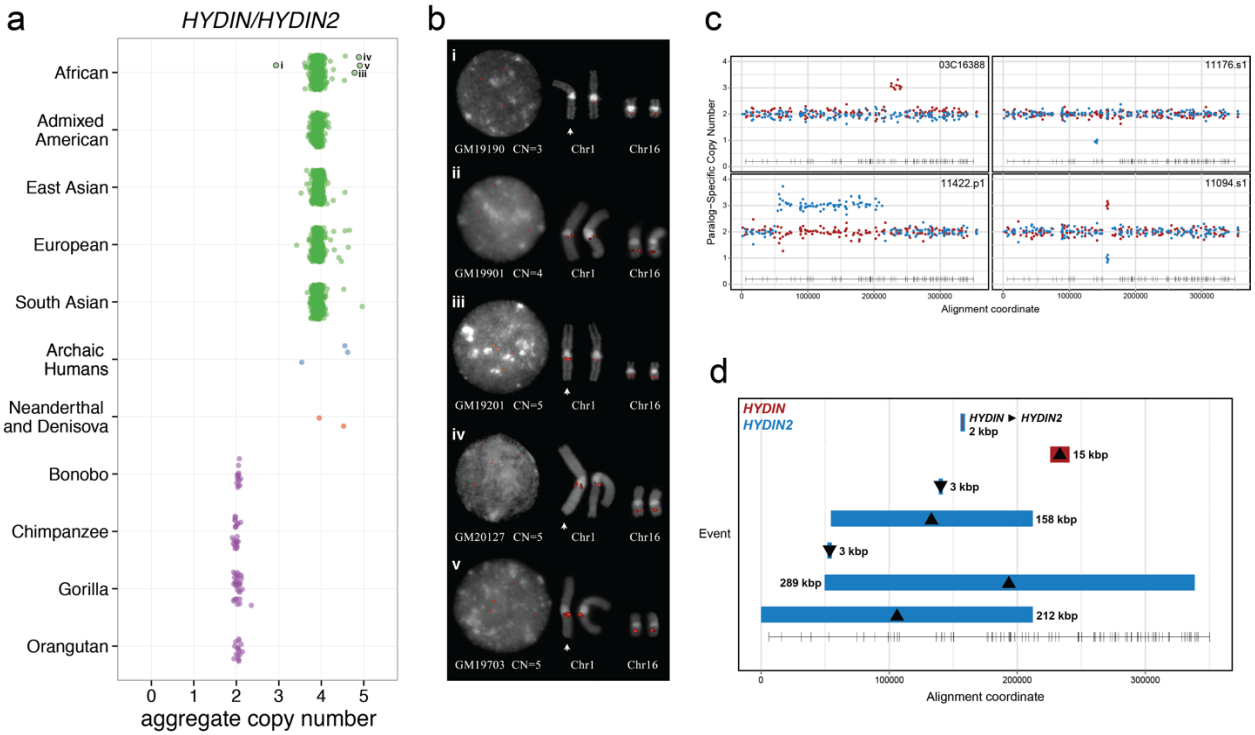


Figure 2.2. *HYDIN* copy number diversity in humans and great apes.

a) Diploid aggregate copy number for *HYDIN* loci based on genome sequencing data from 1KG (611 Africans, 285 admixed Americans, 400 East Asians, 376 Europeans, 451 South Asians), as well as archaic genomes and great apes

(14 bonobos, 23 chimpanzees, 32 gorillas, 17 orangutans). Copy number was estimated using average sequence read depth across the 364 kbp segmental duplication as described previously and represents the aggregate diploid copy number for *HYDIN* and *HYDIN2*. **b)** FISH analysis of metaphase and interphase chromosomal preparations from four human outliers (enumerated in panel a) and one control (aggregate cn = 4 copies) confirms rare duplications (aggregate cn=5 copies) and losses (aggregate cn=3 copies) of *HYDIN* restricted to *HYDIN2* on chromosome 1q21.1. **c)** MIP-based genotyping of paralog-specific copy number identifies partial duplications (top and bottom left panels), deletions (top right panel) and putative interlocus gene conversion events (bottom right panel). Each point estimates paralog-specific copy number (red, *HYDIN*; blue, *HYDIN2*) based on sequencing read depth over SUNs that distinguish *HYDIN* paralogs. 153 MIPs were used for genotyping, and events were detected by an automated caller. Also shown is the canonical *HYDIN* gene model (bottom of each plot). **d)** Summary of *HYDIN* internal structural variation and interlocus gene conversion events based on MIP genotyping of 6,055 humans. Duplications (up arrows), deletions (down arrows) and the sole interlocus gene conversion event (horizontal arrow) are colored according to locus (red, *HYDIN*; blue, *HYDIN2*) and their spatial extent shown with respect to exonic structure (bottom of plot).

For five human samples that showed copy number variation, we performed fluorescent *in situ* hybridization (FISH) on chromosomal metaphase spreads for validation and cytogenetic characterization (**Figure 2.2b**). In all instances, rare copy number variation is restricted to the duplicate copy, *HYDIN2*, on chromosome 1q21.1. We designed an orthogonal method to assay *HYDIN* paralog-specific copy number at a finer scale by designing molecular inversion probe (MIP) assays (Nuttle et al. 2013) to >153 nucleotide differences that distinguish the *HYDIN* paralogs. We genotyped 6,055 DNA samples from controls as well as patients with neurodevelopmental and autism spectrum disorders (Appendix A: Table S1). Other than patients with the 1q21 microdeletion/microduplication syndrome, copy number variation of *HYDIN2* was rare. The assay confirmed rare deletions and duplications of *HYDIN2* (described above) as well as rare copy number variants in *HYDIN* previously detected by array comparative genomic hybridization (CGH) of autism patients (Girirajan et al. 2013). We discovered additional rare internal duplications and deletions, ranging in size from 3 kbp to 289 kbp, affecting both *HYDIN* paralogs (**Figure 2.2c-d**; see also Appendix A: Figure S4 and Appendix A: Table S2). Remarkably, two control individuals showed clear signatures of interlocus gene conversion over a common ~2 kbp region. This event is copy number neutral but clearly shows that *HYDIN* has served as the donor for the conversion of sequence to *HYDIN2*. Such events are rare (2/2,981 or 0.00067) but

indicate nonreciprocal sequence exchange has occurred between *HYDIN* paralogs on chromosomes 1 and 16.

In summary, we observe no heterozygous deletions in *HYDIN* in the 1000 Genomes Project or the HGDP genome samples. Pathogenic mutations in the ancestral copy of *HYDIN* have been observed only in the homozygous state in consanguineous families (Olbrich et al. 2012). The *HYDIN* duplications (2) and deletions (1) that have been observed in autism cases are large and include other genes (Girirajan et al. 2013). *HYDIN2* copy number variation occurs rarely and is largely restricted to individuals carrying chromosome 1q21.1 rearrangements. We have observed no individuals with a loss of both copies of *HYDIN2*.

2.3.3 *HYDIN2 fusion transcripts*

We investigated the gene structure of *HYDIN2* by first considering an alignment of the theoretical open reading frame (ORF) based on the ancestral *HYDIN* gene model (**Figure 2.3**). If all duplicated codons were maintained as in *HYDIN*, the theoretical alignment would yield 21 synonymous and 32 nonsynonymous differences between the shared sequence, as well as three deletions (**Figure 2.3c**). The latter includes: a 2,095 bp deletion (with intronic sequence) on *HYDIN2* that eliminates the splice acceptor for exon 42, a 15 bp in-frame deletion in exon 46, and a 1 bp deletion in exon 69. Due to the deletion of exon 42 and the frameshift in exon 69, a premature stop codon is predicted. As a result of the incomplete gene structure with respect to *HYDIN*, *HYDIN2* is annotated as a pseudogene by both RefSeq and GENCODE.

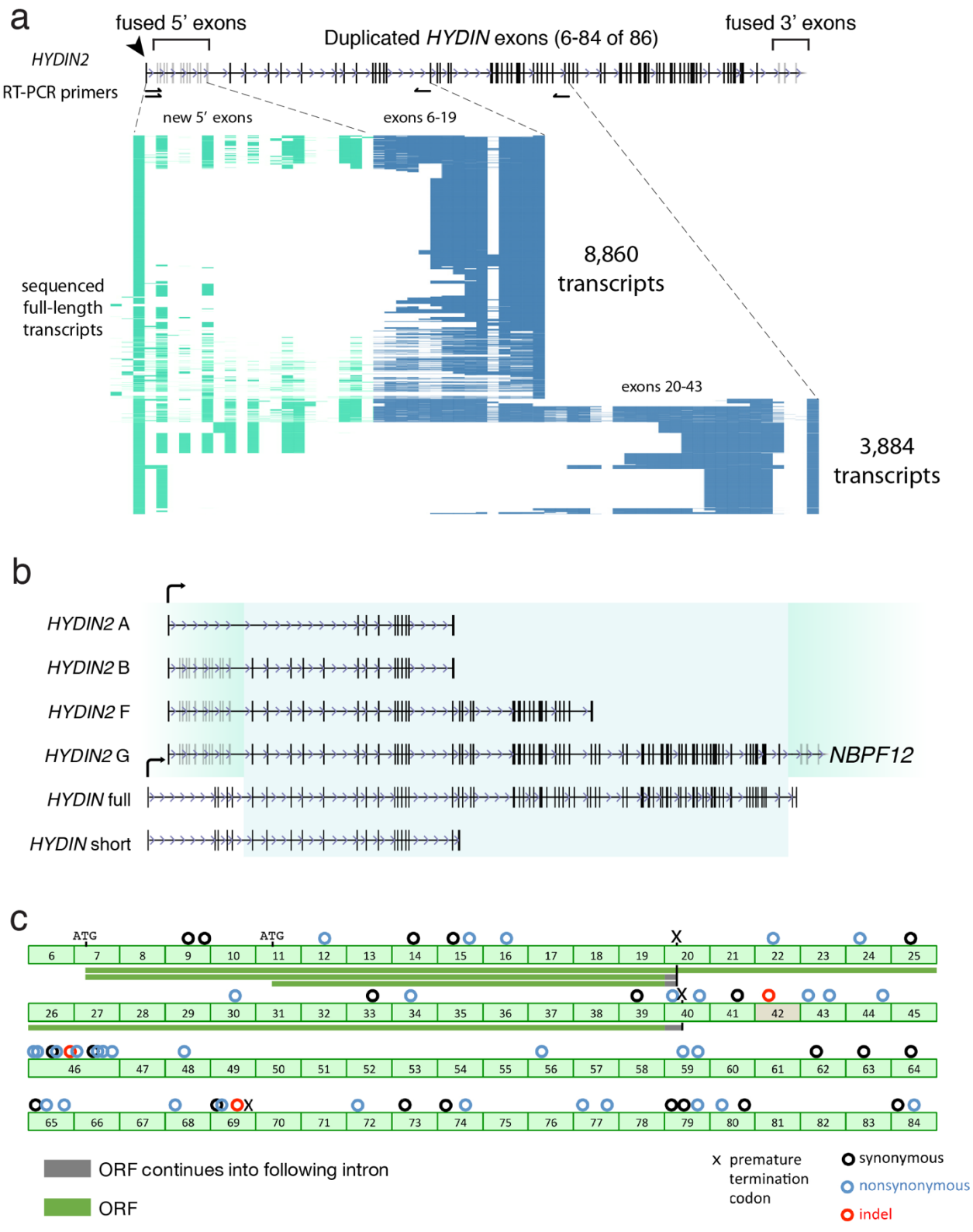


Figure 2.3. *HYDIN2* transcript diversity and ORF potential.

a) Long-range RT-PCR amplicons spanning the first exon identified by 5'-RACE to putative terminal exons 19 and 43 (primer pairs shown as half arrows) were targeted for long-read single-molecule sequencing. 12,744 amplicons were characterized, and isoform content visualized for each product (row), with exons (columns) colored based on whether they are part of the canonical *HYDIN* gene structure (blue) or exapted from flanking sequence (green). **b)** *HYDIN* and *HYDIN2* transcript isoforms based on long-read sequencing of RT-PCR products. Exons corresponding to the duplicated segment (blue shading) and flanking sequences for *HYDIN* (white) and *HYDIN2* (green). Three *HYDIN2* isoforms were identified (isoforms A, B, and F) and an isoform that spans the segmental duplication on both sides (isoform G) was constructed from multiple overlapping reads. The full-length, canonical *HYDIN* (ENST00000393567.6) and its shorter isoform (ENST00000321489.9) are shown. Exons in gray are subject to alternative splicing. **c)** Predicted ORFs for *HYDIN2* (green bars) are shown with respect to *HYDIN* gene structure. Coding differences are indicated above exons, numbered with respect to the canonical isoform of the ancestral gene. Circles indicate synonymous (black), nonsynonymous (blue), and indel (red) differences. Note: a 2,095 bp *HYDIN2* deletion eliminates part of the intron 41 and exon 42, including the splice acceptor for exon 42. Exon 42 is skipped and exon 41 is rarely observed in *HYDIN2* transcripts. Productive *HYDIN2* transcripts are unlikely to continue past exon 42. The three longest ORFs are predicted to be 1,852 aa (Isoform F, exons 7-39), 668 aa (Isoform B; exons 7-19) and 467 aa (Isoform A, exons 11-19); only Isoform A lacks multiple exons 5' to the ORF. ORF extensions into the intron of terminal exons are indicated (gray).

We identified mapped expressed sequence tags (ESTs), both within and spanning the *HYDIN* duplication breakpoints, as evidence that transcription might still be occurring, potentially by way of gene fusion to neighboring sequence. To identify the new transcriptional start site, we performed 5' rapid amplification of cDNA ends (5' RACE), using fetal brain RNA as starting material. We identified a site 55 kbp upstream, which we define as the *HYDIN2* promoter. The high density of spliced ESTs at this location supports this as a site of transcription initiation; we also observe a robust fetal brain DNase1 hypersensitivity peak (**Figure 2.1e**, Appendix A: Figure S6) consistent with its function as an alternate promoter. Although duplication of *HYDIN* excluded the canonical polyadenylation site of the longest *HYDIN* isoform, shorter isoforms of *HYDIN* that terminate at exons 15, 19, and 20 are also annotated. We identified spliced ESTs supporting exon 19 as an alternative site of polyadenylation in *HYDIN*. We further investigated alternative polyadenylation through 3' RACE and identified a potential polyadenylation site at exon 43 of *HYDIN2*.

We amplified by RT-PCR the putative full-length transcripts that spanned from the new *HYDIN2* promoter to both the polyadenylation site at exon 19 and the polyadenylation site at

exon 43, using fetal brain RNA as a starting material (**Figure 2.3a**). We also designed a series of smaller RT-PCR products extending into the fused 3' exons. Products were sequenced with single-molecule, real-time (SMRT) sequencing technology, which allowed us to resolve patterns of splicing, although we note that abundance of different isoforms cannot be taken as evidence of mRNA expression levels due to the preferential loading bias for smaller isoforms inherent to the sequencing technology. We identified fusion transcripts with both upstream and downstream segments beyond the *HYDIN* duplication breakpoints. Upstream, these transcripts begin at the promoter identified by 5' RACE and continue into the *HYDIN* duplication. We observe thirteen 5'-fused exons, with considerable diversity in alternative splicing. The transcripts that begin with these exons continue into the *HYDIN* duplication and then generally follow the canonical pattern of *HYDIN* splicing. Downstream, we also identify fusion transcripts between the *HYDIN* duplicate exons and the 3' exon block. These fusion transcripts continue into the neighboring gene *NBPF12*, a gene that itself undergoes highly variable splicing, with annotated transcripts ranging from 1,831 bp to 7,061 bp in length.

Transcripts that begin at the new promoter and continue into the duplicated *HYDIN* exons are predicted to initiate translation consistent with the *HYDIN* ORF at the ATG located in exon 7 or an alternate ATG mapping to exon 11 (**Figure 2.3c**). Based on sequencing of transcripts, we predict ORFs of 467 and 668 amino acids (transcript termination at exon 19) and an ORF of 1,852 (transcript termination at exon 43). These transcripts are designated *HYDIN2* isoforms A, B, and F, respectively (**Figure 2.3b**). If translated, these products would represent truncated HYDIN proteins (green bars in **Figure 2.3c**). The 5' and 3' fusion exons (located in *NBPF12)* do not extend ORFs beyond the *HYDIN* duplication. We do not find evidence of internal alternative sites of transcription initiation, whether by ESTs, cap analysis gene expression (CAGE), or 5'

RACE, although RACE has not been performed exhaustively throughout the 364 kbp duplicated segment.

2.3.4 *Expression analysis*

Our results suggest that *HYDIN2* acquired a new promoter and we hypothesize that the novel promoter is responsible for the expression differences between the paralogs. We took advantage of a 15 bp in-frame exonic deletion that distinguishes *HYDIN2* from *HYDIN* to investigate the transcript abundance in different tissues (**Figure 2.4a**). By designing RT-PCR primers that flanked this deletion, we inferred relative expression of the *HYDIN* paralogs from the differences in signal intensity. Next, we leveraged RNA-seq reads that could be mapped to SUNK differences between *HYDIN* and *HYDIN2* to quantify relative expression levels across a panel of tissues using data from the GTEx project (The GTEx Consortium, 2013) (**Figure 2.4b**). The latter analysis allowed us to compare expression among the different isoforms characterized by cDNA sequencing as well as quantify differences between paralogs.

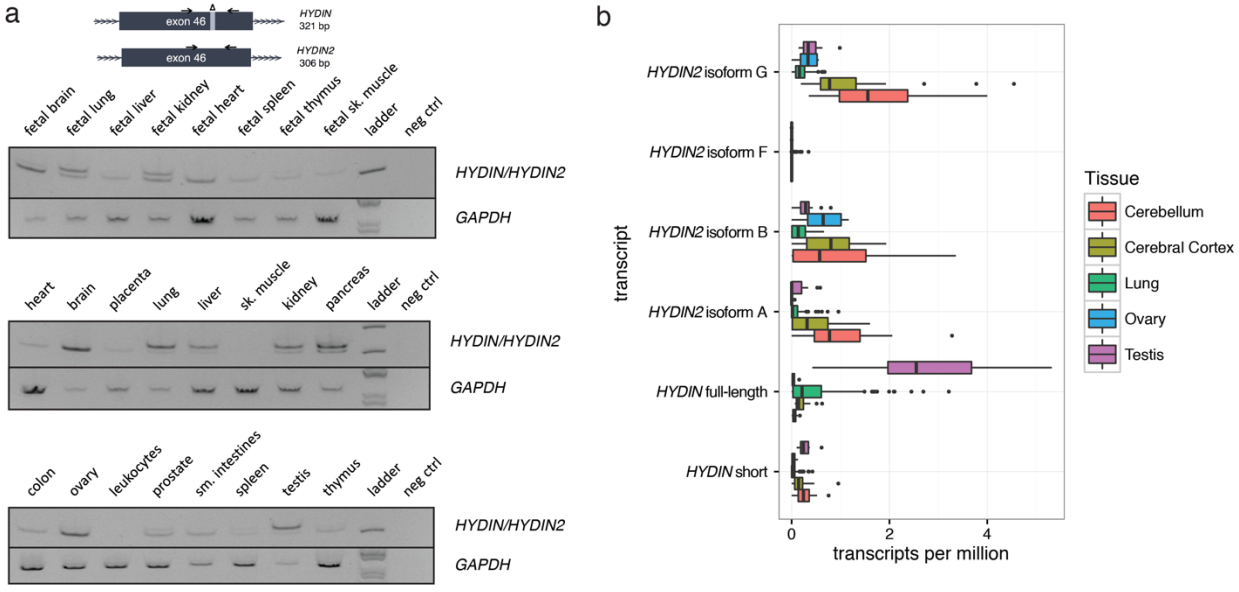


Figure 2.4. Tissue-specific expression of *HYDIN/HYDIN2* isoforms.

a) RT-PCR analysis over a 15bp deletion in exon 46 of *HYDIN2* compares the relative abundance of *HYDIN* (top band, 321 bp) and *HYDIN2* (bottom band, 306 bp) mRNA in different adult and fetal tissues. Images have been inverted and brightness adjusted for clarity. Adult brain and ovary show the highest levels of *HYDIN2* while *HYDIN* is expressed predominantly in lung, pancreas and testis. In fetal tissues, *HYDIN2* is expressed more ubiquitously. **b)** RNA-seq reads from various tissues (The GTEx Consortium, 2013) containing SUNKs (k = 30) were used to estimate *HYDIN* and *HYDIN2* expression. Full-length *HYDIN* is expressed most highly in the lung and testis, while all isoforms of *HYDIN2* are more highly expressed in brain tissues. Isoform F is not expressed at a level that is likely to be to be significant. Boxplots indicate median and interquartile range (IQR) with outliers shown beyond 1.5 x IQR.

We observe that *HYDIN* is most highly expressed in ciliated tissues, such as the lung and testis, consistent with its known function as a ciliary structural protein (Lechtreck et al. 2007). By comparison, *HYDIN2* expression is higher in the brain, while also prominent in the ovary. Interestingly, *HYDIN2* appears more broadly expressed in fetal tissues with the strongest signal observed in fetal brain. It should be noted that the RT-PCR assay measures expression of exon 46 alone, which is not included in some shorter RefSeq isoforms of *HYDIN*. Comparison of this assay with SUNK-based mapping of the GTEx data is in close agreement. This is also consistent with the observation that cDNA clones from *HYDIN2* were predominantly derived from neuronal sources (Doggett et al. 2006). With the exception of *HYDIN2* isoform F, all three isoforms show moderate levels of expression in cerebellum and cerebral cortex. The longest isoform, *HYDIN2* isoform G, with both 5' and 3' fused exons shows a slightly higher level of expression, although we note only that only *HYDIN2* isoform A produces a transcript without a large number of 5' untranslated exons due to exon skipping from the first exon to exon 11 of the canonical gene model where an alternate ATG start codon exists.

The predicted *HYDIN2* promoter overlaps a DNase I hypersensitivity (DHS) peak (**Figure 2.1e**, Appendix A: Figure S6) in fetal brain. A number of fetal brain cDNA sequences have been mapped to the duplicated locus of the promoter-containing block on chromosome 1q21.2 (Ota et al. 2004; Harrow et al. 2012). Notably, at the ancestral chromosome 1p22.3 locus (**Figure 2.1d**), where the 5' and 3' exon blocks sit in the absence of the promoter, there are neither annotated

transcripts nor spliced ESTs. Within the DHS peak a smaller, higher copy repeat is found, approximately 1,113 bp long. A simple BLAT search reveals that this segment has propagated throughout chromosome 1, with ten locations that contain the full-length repeat at a level of identity of 90% or above (Appendix A: Table S3). Many are clustered at chromosome 1q21 and are found in association with, though not transcriptionally joined to, the *NBPF* core duplicon. Eight out of ten produce spliced ESTs, many from neuronal tissues, further supporting this segment as contributing to the expression pattern of *HYDIN2*.

2.3.5 *Coding variation and selection in HYDIN and HYDIN2*

We investigated whether there was evidence of selection acting on *HYDIN2* by comparing the ratio of nonsynonymous to synonymous changes between paralogs, using nonhuman primate *HYDIN* as an outgroup (Appendix A: Table S4). We observe moderately strong purifying selection acting on ancestral *HYDIN* throughout the primate lineage, with an average dN/dS value of 0.29. *HYDIN2*, in contrast, shows an elevated pairwise dN/dS value in comparison with nonhuman primates (e.g., 0.39 for human *HYDIN2* and chimpanzee *HYDIN* vs. 0.29 for human *HYDIN* and chimpanzee *HYDIN*). Branch-based estimates of dN/dS that can detect adaptive evolution after gene duplication (Bielawski et al. 2003) show a similar trend with a consistently elevated dN/dS value for human *HYDIN2* when compared to human *HYDIN*. This result holds if we restrict our analysis to only those exons predicted to be part of the *HYDIN2* ORF (*HYDIN2* isoform A). Although none of these differences achieve statistical significance due to the limited number of mutational differences occurring within the human lineage, it is interesting that all three mutational changes that occurred within the *HYDIN2* ORF result in amino acid changes while only synonymous changes (n=2) occurred in the corresponding portion of ancestral *HYDIN*.

As an alternative approach, we sought to characterize and compare deleterious coding variation between *HYDIN* and *HYDIN2* among 3,484 probands and 2,629 healthy controls from families with autism (O'Roak et al. 2014; Boyle et al. 2014). We targeted all coding exons with at least five nucleotides of flanking sequence based on the canonical *HYDIN* gene structure using MIPs. We identified all likely gene-disruptive (LGD) variants (frameshift, stop-gain, stop-loss or splice-site mutations) and successfully assigned 39% of such deleterious variants to either *HYDIN* or *HYDIN2*, made possible by singly unique nucleotides (SUNs) contained within the target sequence of the MIP (**Table 2.1**). No common (>1% allele frequency) LGD variants were observed for either paralog. Considering the canonical ancestral gene structure, we initially observed a lower number of LGD mutations for *HYDIN* (n=2) when compared to *HYDIN2* (n=10) in controls. Interestingly, if we restrict our analysis to the most likely ORF model (*HYDIN2* isoform A), only two *HYDIN2* LGD mutations remain and at an allele frequency comparable to what has been observed for the functional *HYDIN*. Assuming that all unassigned LGD mutations originate from *HYDIN2*, all higher frequency LGD mutations (>0.15% frequency) correspond to ancestral exonic sequence mapping outside of the *HYDIN2* ORF gene model (Appendix A: Table S5).

Table 2.1.: Likely gene disruptive events detected in *HYDIN*/*HYDIN2* by MIP-based sequencing of exons in cases and controls.

						Cases (N=3483)			Controls (N=2629)		
Paralog*	**Variant**	**Exon**	**Intron**	**Protein position**	**Amino acid**	**N**	**Freq.**	**Number genotyped****	**N**	**Freq.**	**Number genotyped****
Cases only											
HYDIN2	splice_donor	-	66/85	-	-	1	0.03%	3431	0	0.00%	2604
HYDIN2	splice_acceptor	-	53/85	-	-	1	0.03%	3432	0	0.00%	2599
HYDIN	stop_gained	48/86	-	2690	Q/*	1	0.03%	3433	0	0.00%	2603
HYDIN2	stop_gained	46/86	-	2540	R/*	1	0.03%	3425	0	0.00%	2598
Controls only											
HYDIN2	stop_gained	80/86	-	4563	W/*	0	0.00%	3429	1	0.04%	2599
HYDIN2	frameshift	48/86	-	2680	G/X	0	0.00%	3427	1	0.04%	2598

HYDIN2	splice_donor	-	42/85	-	-	0	0.00%	3428	1	0.04%	2599
HYDIN	splice_donor	-	29/85	-	-	0	0.00%	3434	1	0.04%	2603
HYDIN	stop_gained	11/86	-	1330	R/*	0	0.00%	3429	1	0.04%	2599
Found in both cases and controls											
HYDIN2	splice_acceptor	-	67/85	-	-	11	0.32%	3430	6	0.23%	2601
HYDIN2	splice_acceptor	-	54/85	-	-	1	0.03%	3426	1	0.04%	2595
HYDIN2	frameshift	46/86	-	2485	A/X	1	0.03%	3428	1	0.04%	2600
HYDIN2	frameshift	41/86	-	2115-2116	VI/VSX	11	0.32%	3427	10	0.38%	2598
HYDIN2	splice_donor	-	28/85	-	-	1	0.03%	3430	1	0.04%	2600
HYDIN2	frameshift	19/86	-	2531-2532	A/X	1	0.03%	3434	1	0.04%	2607
HYDIN2	splice_acceptor	-	14/85	-	-	4	0.12%	3429	2	0.08%	2605

*Paralog determined by presence of SUN on variant-containing MIP reads, variants identified by MIP reads that did not intersect a SUN could not be assigned; Variants in *HYDIN2* are annotated with the exon numbering scheme from *HYDIN*

**Number of samples successfully genotyped for this variant (Freebayes)

We genotyped 3,483 probands and 2,629 healthy controls from families with autism using a MIP-based genotyping assay that targeted coding exons and at least 5 flanking intronic nucleotides. LGD variants (frameshift, stop-gain, stop-loss, and splice-site) were called using Freebayes. Only variants that could be definitively assigned to *HYDIN* or *HYDIN2* based on the presence of an identifying SUN are shown. Variants include those seen only in cases, seen only in controls, and those seen in both cases and controls. Most of the variants seen in *HYDIN2* occur outside of the putative coding sequence.

2.3.6 *HYDIN2 and the chromosome 1q21 microdeletion/microduplication syndrome*

Recurrent microdeletions and microduplications at chromosome 1q21 have been associated with a variety of neurodevelopmental phenotypes, including microcephaly and macrocephaly. Loss or gain of *HYDIN2* has been hypothesized to underlie these head circumference phenotypes in light of its expression in brain, its inclusion in the typical rearrangement interval, and the association of homozygous losses of *HYDIN* in mouse with hydrocephalus (Brunetti-Pierri et al. 2008; Davy et al. 2003; Doggett et al. 2006). To explore this hypothesis, we leveraged our MIP assay (**Figure 2.5a**) to genotype *HYDIN* paralog-specific copy number in 73 individuals carrying a chromosome 1q21 rearrangement, corresponding to 45 independent rearrangement events (15 duplications and 30 deletions). MIP data revealed that *HYDIN2* is usually, but not always, affected by chromosome 1q21 deletions and duplications (**Figure 2.5b**, Appendix A: Figure S5). Overall, we observe that 87% of duplications (13 of 15) and 93% of deletions (28 of 30)

examined include *HYDIN2*. Targeted array CGH on a subset of patients validated our whole-genome-shotgun- and MIP-based results in every instance, confirming inclusion or exclusion of *HYDIN2* among both 1q21 microduplications and microdeletions (**Figure 2.5c**, Appendix A: Figure S5).

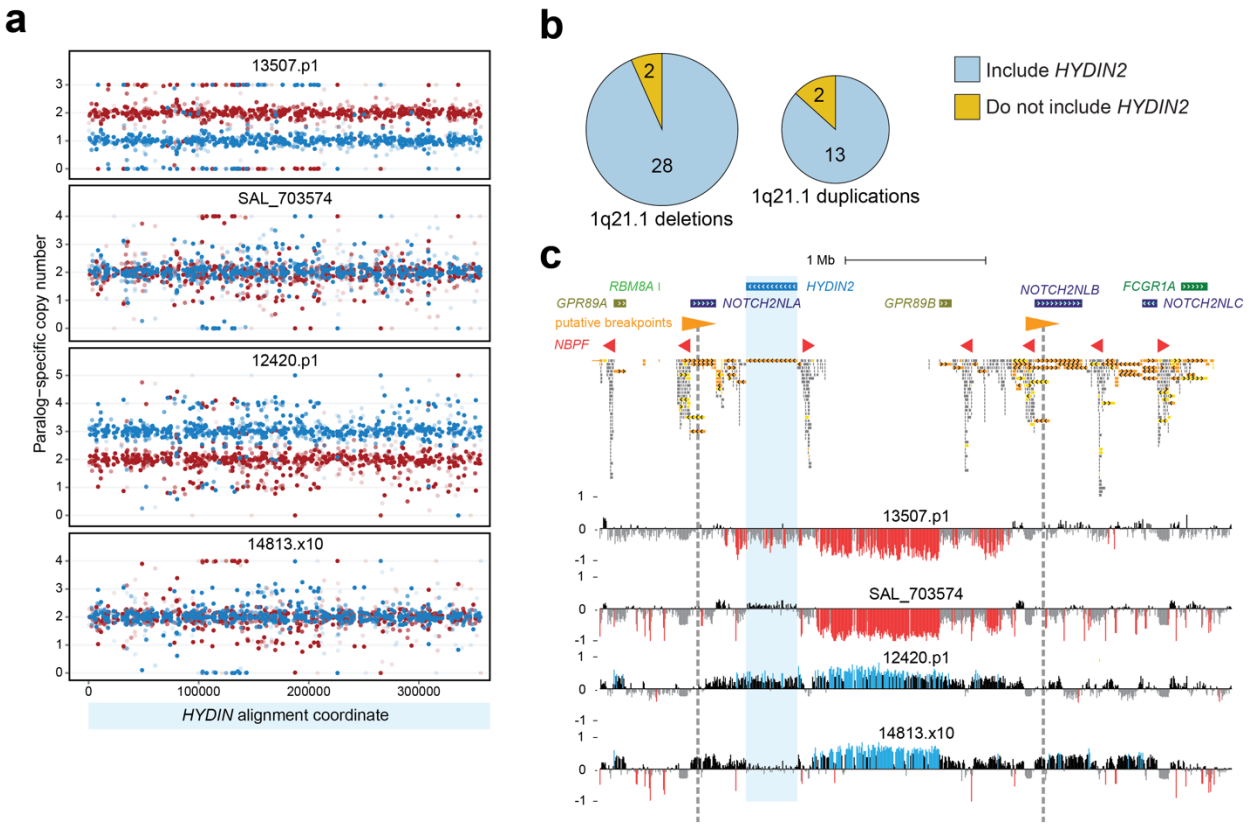


Figure 2.5. *HYDIN2* and chromosome 1q21 rearrangement breakpoint variability.

We genotyped 73 patients carrying either the chromosome 1q21 microdeletion (n = 48) or microduplication (n = 25) for *HYDIN2* copy number using MIPs. **a)** Patients were genotyped using 717 MIPs targeting variants that distinguish *HYDIN* paralogs. Points show *HYDIN* paralog-specific copy number estimates (red, *HYDIN*; blue, *HYDIN2*) for two microdeletion (13507.p1 and SAL_703574) and two microduplication (12420.p1 and 14813.x10) patients. Patients 13507.p1 and 12420.p1 show deletion and duplication of *HYDIN2*, respectively, while SAL_703574 and 14813x10 do not. **b)** Summary of results across 45 independent microdeletion and microduplication events from 73 individuals based on MIP sequencing and analysis. ~91% of 1q21 rearrangements examined include *HYDIN2*.**c)** Array CGH results confirm 1q21 rearrangements in the samples in panel a and copy number changes of *HYDIN2* (blue shading) only in patients 13507.p1 and 12420.p1. Note: log2 hybridization signal intensity (y-axis) values are depressed when compared to unique sequence due to duplicated nature of sequence (red = deletion signal; blue = duplication signal). Results are shown for a 4.5 Mbp region at chromosome 1q21 (GRCh38 chr1:145,500,001-150,000,000) with genes and segmental duplications annotated (orange = 99% sequence identity or above; yellow = 98%–99%; gray = 90%–98%). Orange triangles indicate high-identity, directly oriented *NOTCH2NL-NBPF* duplications, with putative

breakpoints of the canonical 1q21 rearrangement shown as vertical gray dashed lines. Shown below are the locations of *NBPF* core duplicons.

We also performed comprehensive segmental duplication analysis on the GRCh38 reference haplotype to identify directly oriented duplication pairs with high sequence identity that may confer susceptibility to nonallelic homologous recombination. For typical 1q21 rearrangements, those that include *HYDIN2*, the most likely candidates are large (247 kbp), highly identical (99.7%) segments that include truncated *NOTCH2* duplications (*NOTCH2NL*) adjacent to members of the *NBPF* gene family (**Figure 2.5c**, Appendix A: Figure S5b). For atypical rearrangements, our data are consistent with multiple possible breakpoint locations. Alternatively, these events may have occurred on a still undescribed haplotype structure or originated through a non-recurrent mechanism.

We examined the phenotype of atypical carriers without copy number variation in *HYDIN2*. All three atypical 1q21 microduplications excluding *HYDIN2* present with macrocephaly, and similarly all three patients with atypical 1q21 microdeletions excluding *HYDIN2* exhibit microcephaly (Appendix A: Table S6). These observations suggest that loss or gain of a genomic *HYDIN2* copy is not necessary for chromosome 1q21 rearrangement patients to manifest head circumference abnormalities. We cannot rule out the possibilities that these individuals harbor disruptive point mutations in *HYDIN2* or that atypical chromosome 1q21 rearrangements dysregulate *HYDIN2* expression. Further studies will be necessary to elucidate the potential role of *HYDIN2* in brain size and in other aspects of chromosome 1q21 rearrangement phenotypes.

2.4 DISCUSSION

HYDIN2 is a human-specific gene that emerged ~3 mya, created by incomplete duplication of the ancestral gene *HYDIN*. Young duplicate genes can rapidly evolve essential functions (Chen

et al. 2013); however, unless increased gene dosage is itself beneficial, long-term maintenance of a duplicate gene typically requires mutational change leading to functional innovation or subfunctionalization (Lynch and Conery 2000). Incomplete gene duplication provides one mechanism for rapid functional change because the duplicate differs in structure from its progenitor, with potentially profound functional consequences beyond a simple dosage increase. Such a mechanism has been postulated in the case of the human-specific incomplete duplications of *SRGAP2*, where the truncated granddaughter paralog, *SRGAP2C*, has been shown to antagonize the ancestral copy, *SRGAP2A* (Dennis et al. 2012; Charrier et al. 2012).

Interspersed duplications such as *HYDIN* have an additional mechanism promoting functional divergence, namely, duplicate copies in new genomic locations are subjected to asymmetric rates of mutation (Jun et al. 2009). Because the *HYDIN* duplication excluded the promoter, naively one would expect the duplicate copy to have been silenced. In contrast, our analysis shows that *HYDIN2* is actively transcribed, more highly expressed than the ancestral paralog in many tissues. The acquisition of a novel promoter effectively created a fusion gene. Thus, the partial duplication was "rescued" by its juxtaposition with active regulatory sequence at chromosome 1q21.1.

The original function of this novel neuronal promoter is not clear, but 2 Mbp away at chromosome 1q21.2 sits an earlier duplication of the *HYDIN2* flanking sequences without the *HYDIN2* insertion. Here, we observe robust transcription, as evidenced by a number of GENCODE transcripts—all classified as long noncoding RNA and most of which derive from a fetal brain cDNA library (Ota et al. 2004). At *HYDIN2*, this promoter has driven an altered expression pattern, with widespread expression in fetal tissues, decreased expression in testis and

lung, and increased expression in brain tissues (cerebellum and cerebral cortex) and ovary being the most prominent changes.

Long-read sequencing of cDNA from both the fetal and adult brain reveals an extraordinary diversity of *HYDIN2* isoforms, including the presence of additional 5' and 3' exons within transcripts spanning the duplication junctions. Although we can confirm expression of at least three distinct *HYDIN2* isoforms, we favor isoform A as the most likely protein-encoding transcript for several reasons. There is no evidence for a premature termination codon; deleterious coding mutations in the human population are rare and the other isoforms carry an unusually large number of untranslated exons. The presence of abnormally long untranslated regions (UTRs) or exon junctions downstream of a premature termination codon usually indicates strong signatures for nonsense mediated decay of mRNA (Kurosaki et al. 2013). Similarly, a large number of 5' noncoding exons is thought to impede translational efficiency (Kozak 1989; Kozak 1991).

In the case of *HYDIN2* isoform A, intervening untranslated exons are skipped, resulting in a putative 467 amino acid protein with relatively short 5' and 3' UTRs. Although we know little regarding the function of *HYDIN2*, it is noteworthy that *HYDIN* has a structural role in motile cilia (Lechtreck et al. 2007; Dawe et al. 2007; Lechtreck et al. 2008). Its expression in lung and testis is consistent with the observation that recessive mutations in *HYDIN* cause primary ciliary dyskinesia, with the primary phenotypes being chronic respiratory infections and male infertility in humans (Olbrich et al. 2012). It is interesting that in addition to these deficiencies, mouse mutants in *hy3* (*Hydin*$^{-/-}$) show a more severe phenotype, developing lethal hydrocephalus due to impaired ciliary motility and fluid flow in the developing brain. One possible explanation for this phenotypic discrepancy between mutant mice and humans lacking functional *HYDIN* may be that

human *HYDIN* paralogs have undergone subfunctionalization. In particular, the neuronally expressed *HYDIN2* may have assumed some of the ancestral gene's function during human brain development.

Our copy number analyses reveal that the *HYDIN2* duplication has largely fixed in the human population. In fact, *HYDIN2* shows the lowest degree of copy number variation in the normal population when compared to other human-specific duplications (Sudmant et al. 2010; Dennis et al. 2016). While it was speculated that *HYDIN2* played an important role in the reciprocal macrocephaly/microcephaly phenotype associated with 1q21.1 duplications/deletions (Brunetti-Pierri et al. 2008), we have identified rearrangement patients with head size abnormalities lacking altered genomic dosage of *HYDIN2*. It is plausible that altered expression or point mutations effectively disrupt *HYDIN2* in these individuals, or, alternatively, that *HYDIN2* dysfunction does not contribute to their head circumference phenotypes. Distinguishing these possibilities and determining whether *HYDIN2* plays an important role in neurodevelopment more broadly will require further functional studies complemented by large-scale genotyping in various neurodevelopmental disease cohorts and relevant, well-phenotyped controls.

In this study we characterize the evolutionary history, transcriptional landscape, and potential clinical impact of the human-specific duplicate gene *HYDIN2*. We show that *HYDIN2* was generated by the juxtaposition of multiple segmental duplications culminating with the partial duplication of *HYDIN* ~3.2 million years ago. We identify a new promoter that "rescued" the truncated gene duplicate and drives a neuronal pattern of expression. We show that long-read sequencing can be used to understand a previously intractable large and complexly spliced gene, and identify transcribed unannotated ORFs. We show that the reciprocal macro/microcephaly phenotypes associated with chromosome 1q21 rearrangements can occur without *HYDIN2* copy

number changes. Ultimately we provide a clear example of how juxtaposition of transcriptionally active segmental duplications can lead to the birth of a new gene.

2.5 METHODS

2.5.1 *FISH*

Metaphase and interphase spreads were prepared from lymphoblastoid human cell lines (GM19190, GM19901, GM19201, GM20127, GM19703; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using fosmid clone WIBR2-3823N03, directly labeled by nick-translation with Cy3-dUTP (PerkinElmer) as previously described (Antonacci et al. 2010) with minor modifications. Briefly, 300 ng of labeled probe was used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 mg sonicated salmon sperm DNA, in a volume of 10 µl. Posthybridization washing was at 60°C in 0.1xSSC (three times, high stringency). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI and Cy3 fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

2.5.2 *Sequencing and assembly of large-insert clones (BACs)*

We searched for discordant BAC-end mappings that spanned the *HYDIN2* insertion site at chromosome 1q21.1 in libraries CH251 (chimpanzee) and CH276 (orangutan). One clone was identified for chimpanzee (CH251-231E10) and one in orangutan (CH276-57C3). DNA was isolated from these clones and SMRTbell libraries were prepped and sequenced on the Pacific Biosciences RSII. Inserts were assembled using HGAP and Quiver-polished as previously

described (Huddleston et al. 2014).

2.5.3 *Phylogenetic analysis*

Orthologous *HYDIN* sequences in chimpanzee and orangutan were identified using BLAT (Kent 2002a) in the UCSC Genome Browser (Kent 2002b). The human chromosome 16 *HYDIN* shared sequence was used as a query against panTro3 and ponAbe2, respectively. An MSA of these nonhuman primates as well as the duplicated sequences from chromosomes 1 and 16 (both from CH17) was created using ClustalW (Thompson et al. 2012). An unrooted phylogenetic tree was constructed in MEGA6 (Tamura et al. 2013) using the neighbor-joining method (Saitou et al. 1987) with complete-deletion option, yielding a total of 315,349 positions. Genetic distances were computed under the Kimura two-parameter model (Kimura 1980) with standard error estimates (Felsenstein 1985) (N=500 bootstrap replicates). A Tajima's relative rate test using chimpanzee as the outgroup failed to reject the hypothesis that both human *HYDIN* paralogs are evolving at the same rate (p=0.21). Thus the timing of the duplication event is estimated by taking the average evolutionary distance between the two *HYDIN* paralogs as a ratio of the total distance from chimpanzee *HYDIN*. This yields a timing estimate of 3.17 mya assuming the divergence took place 6 mya. 95% confidence interval was estimated by the bootstrapping method.

2.5.4 *Copy number genotyping*

Aggregate and paralog-specific copy number estimates of *HYDIN/HYDIN2* were determined using previously described methods (Sudmant et al. 2010). Raw sequences from 236 human individuals from HGDP (Sudmant et al. 2015a), 2,143 human individuals through Phase 3 of 1KG (Sudmant et al. 2015b), 86 nonhuman primate individuals from the Great Ape Genome

Project [including bonobos (n=14), chimpanzees (n=23), gorillas (n=32), and orangutans (n=17)] (Prado-Martinez et al. 2013), a Denisovan individual (Meyer et al. 2012), a Neanderthal individual (Prüfer et al. 2014), and 3 archaic hominids (Fu et al. 2014; Lazaridis et al. 2014) were mapped to the human reference genome using mrsFAST (Hach et al. 2010). For aggregate copy number estimates (**Figure 2.2a**) GRCh37 was used. For paralog-specific copy number estimates (Appendix A: Figure S3) GRCh38 was used, as recent correction to the *HYDIN2* locus in the human genome made possible paralog-specific copy number estimates.

2.5.5 *Expression quantification*

Kallisto (v. 0.42.4) (Bray et al. 2016) was used to estimate the expression levels of nine transcripts detected and putative *HYDIN2* isoforms (see Appendix A: Additional Data 1). We added the *HYDIN2* sequences to the GENCODE reference transcriptome (release 25) (Harrow et al. 2012) and generated a new index using kallisto. Transcripts per million values were then calculated using kallisto with default parameters for all of the GTEx RNAseq samples (dbGaP version phs000424.v3.p1) from the following tissues: cerebellum (38 samples), cerebral cortex (31 samples), lung (133 samples), ovary (6 samples), and testis (15 samples).

To experimentally determine relative expression of *HYDIN* and *HYDIN2* in various tissues, we took advantage of a 15 bp deletion in exon 46 of *HYDIN*. Identical flanking sites were chosen for priming, so that relative expression of transcripts containing exon 46 could be measured in a single reaction. Expected band sizes were 321 bp for *HYDIN* and 306 bp for *HYDIN2*. 5 µL of cDNA from various adult and fetal tissues normalized for expression level was used as template (Cloneech Human MTC Panel I, Human MTC Panel II, Human Fetal MTC Panel I [obtained from spontaneously aborted fetuses, ages 16-40 weeks]) and PCR was performed as per manufacturer's instructions, with GAPDH as a positive control. Reactions were monitored by the

level of SYBR Green I (Invitrogen) fluorescence using Bio-Rad MiniOpticon Real-Time PCR System. Reactions proceeded for 31 cycles of amplification (19 for GAPDH reactions) and PCR products were visualized by 20 minutes of electrophoresis using E-Gel EX Agarose Gels (4%; Invitrogen).

2.5.6 *RACE experiments*

5' and 3' RACE were performed using the FirstChoice RLM-RACE Kit (Ambion) as per manufacturer's instructions using the nested protocol on poly-A^+ RNA derived from fetal brain (Clontech). Spliced ESTs in the *HYDIN2* locus were taken as evidence of active transcription and pileups of ESTs with shared edges were taken as evidence of potential sites of transcription initiation and termination and were chosen as targets for RACE. The ancestral paralog was also targeted as a positive control. PCR products including secondary bands were gel extracted and purified using the QIAquick Gel Extraction Kit (Qiagen) and capillary sequenced. In cases where RACE products did not include unique sequence, PCR products were purified using the QIAquick PCR Purification Kit (Qiagen) and cloned using the TOPO XL PCR Cloning Kit with OneShot TOP10 Chemically Competent *E. coli* (Invitrogen). Cells were streaked onto agar plates containing 50 µg/mL kanamycin and incubated overnight at 37° C. Individual colonies were picked and subjected to colony PCR, where the insert was amplified using standard M13 forward and reverse primers. The PCR products were purified and sequenced as before. Unique nucleotide differences were used to infer the paralog of origin. Primers used can be found in Appendix A: Table S7.

2.5.7 *PacBio cDNA sequencing*

cDNA was synthesized from poly-A^+ RNA using oligo(dT) priming either SuperScript II or

SuperScript III reverse transcriptase (Invitrogen) with the following modification: an extra 50 minutes was added to the reverse transcription incubation time following the addition of 1 µL of enzyme was added. 1 µL of from the cDNA synthesis reaction was used as template for nested PCR with Kapa HiFi PCR Kit. PCR amplicons were purified using magnetic beads (Agencourt AMPure XP, 1X concentration). Library preparation for PacBio sequencing of PCR amplicons was performed using standard and approved reagents and protocols. Two SMRT cells were sequenced: the first, which included amplicons spanning from new *HYDIN2* promoter to exon 19 was run using P5C3 chemistry, the second, which included amplicons spanning from the new *HYDIN2* promoter to exon 43, as well as intermediate fragments from exon 19 through beyond the 3' duplication breakpoint, was run using P6C4 chemistry. Primers used can be found in Appendix A: Table S7.

2.5.8 *DNase I hypersensitivity (DHS) at HYDIN2 promoter*

Chromatin accessibility, as measured by DHS, was assessed for evidence of regulatory activity of the new *HYDIN2* promoter (John et al. 2013; Thurman et al. 2012). Because standard DHS analysis pipelines discard multiply mapping reads, and the *HYDIN2* promoter sits in duplicated space, reads corresponding to DHS sites in fetal brain were remapped to a repeat-masked GRCh38 using mrsFAST-Ultra (version 3.3.11) (Hach et al. 2014) before being used to determine cut counts. Sample information including GEO accession numbers are shown in Appendix A: Table S8.

2.5.9 *MIP exon sequencing*

Human reference sequence (GRCh37) of coding exons from the ancestral paralog only (+/- 5 bp) was used as input to design single-molecule MIPs (Hiatt et al. 2013) using MIPgen (Boyle et al.

2014). Each MIP was designed to capture 112 bp of genomic sequence and included 40 bp unique to the target region (split between a ligation and an extension arm of the MIP), a universal 30 bp backbone, and a degenerate 8 bp molecular tag included on the extension arm. A total of 240 MIPs were designed to cover *HYDIN*. MIP phosphorylation, capture, and barcoding were performed as previously described (O'Roak et al. 2012). Briefly, oligos were pooled together at equal concentrations (100 uM), phosphorylated, and an 800:1 excess of oligos was used for the genomic DNA capture (100 ng). Capture reactions were incubated at 60°C for 18 hours. Finished libraries were pooled together and sequenced using either MiSeq (2 x 150 bp) or HiSeq2000 (2 x 101 bp). Probe sequences can be found in Appendix A: Table S9.

We used the MIPgen data analysis pipeline to map and filter reads in fastq format to a minimal human reference containing only the region containing the ancestral *HYDIN* paralog (chr16:70821397-71282326) included in our MIP design with the remainder of the genome, including *HYDIN2*, masked out. This masking ensured reads mapped to only the ancestral paralog for proper variant annotation. Discovery variant calling was performed across the entire ASD cohort per pooled sequence set containing up to 384 samples using FreeBayes (https://github.com/ekg/freebayes) with the following command: freebayes -b <sorted_bams> -f <masked_reference> -t <targeted_regions> -F 0.07 -C 2 -n 4. We removed any variants with the following feature: trinucleotide or homopolymer repeat, read depth ≤10, quality score ≤20, or with no alleles using previously described methods (Coe et al. 2014). The resulting variant set was annotated using the Ensembl Variant Effect Predictor (VEP) (Cunningham et al. 2015) using the canonical transcript for each gene. Subsequently, for the ASD study, the complete list of coding variants was used to separately genotype cases and controls to assess overall frequency of

events in each cohort: freebayes -b <sorted_bams> -f <masked_reference> -s <sample_list> -@ <variant_vcf> --only-use-input-alleles -F 0.07 -C 2 -n 4 --min-coverage 10.

2.5.10 *Tests for selection*

The full-length ancestral *HYDIN* sequence (NM_001270974) was used as a BLAST query to obtain sequenced *HYDIN* transcripts from other primates. All codons from *HYDIN2* that could align with *HYDIN* were used. An MSA was generated using MAFFT and manually edited for obvious alignment errors. Bases aligning to the 15,366 nt ORF from human *HYDIN* were selected and a neighbor-joining tree was generated using MEGA. The alignment and tree were input into CODEML (Yang 2007) and dN/dS values (omega) were estimated using the Nei-Gojobori method with pairwise deletion (Nei and Gojobori 1986). Branch-based tests were performed by allowing additional branches to vary in their dN/dS parameter and comparing the log-likelihood to the nested model. P-values were calculated by performing a Chi-square test (df=2) on twice the difference between the log-likelihood values for different models considered.

2.5.11 *HYDIN paralog-specific copy number genotyping using MIPs*

HYDIN paralog-specific copy number was genotyped using a previously described method (Nuttle et al. 2013) with single-molecule MIPs (Hiatt et al. 2013). Briefly, MIPs were designed to SUNs distinguishing *HYDIN* paralogs (Appendix A: Table S10). MIP capture, library preparation, massively parallel sequencing, and data analysis allowed quantification of reads derived from each *HYDIN* paralog over each MIP target for each individual. These data were input to a program that output paralog-specific copy number calls and detected duplications, deletions, and interlocus gene conversion events. MIP data for each individual was visualized by plotting paralog-specific *HYDIN* copy number point estimates across the spatial extent of

sequence shared between paralogs. These estimates were calculated at each MIP target by multiplying paralog-specific *HYDIN* read count relative frequencies by corresponding aggregate *HYDIN* copy number estimates called by the genotyping program. The algorithmic details of this program have been previously described (Nuttle et al. 2013). In this case, the program considered 25 possible hidden underlying *HYDIN* paralog-specific copy number states, where both *HYDIN* and *HYDIN2* were allowed to possibly have copy numbers ranging from 0 to 4 (5*5 = 25 combinations). To enable detection of internal events, highest scoring paths through likelihood-based graphs allowing 0, 1, and 2 transitions between copy number states were considered, with the same biologically motivated restrictions on permitted transitions as previously detailed. Prior probabilities were set to reflect the observation that most humans have two copies of both *HYDIN* paralogs, with log-likelihoods of -15, -7.5, 0, -7.5, and -15 assigned to initial single-paralog copy number states of 0, 1, 2, 3, and 4, respectively. Probe sequences can be found in Appendix A: Table S10.

2.5.12 *Array CGH*

Array CGH was performed as previously described (Sudmant et al. 2010; Dennis et al. 2012) using a custom microarray (Agilent) with dense probe coverage across the chromosome 1q21 region.

2.6 NOTES

Ethics approval and consent to participate
The human samples included in this study did not meet the U.S. federal definitions for human subjects research. All samples were publicly available or encoded, with no individual identifiers available to the study authors. Samples were collected at respective institutions after receiving informed consent and approval by the appropriate institutional review boards. There are no new health risks to participants. Samples that fall within this category include probands with autism and their parents from the SSC, AGRE, and TASC cohorts and individuals from representative human populations from the 1000 Genomes Project.

Availability of data and materials

Raw sequencing data from PacBio RT-PCR experiments can be found at the Sequence Read Archive (SRA) under BioProject PRJNA359986. Mapped MIP-based sequencing data can be found under SRA BioProject PRJNA356308. BAC sequences can be accessed under GenBank accession numbersAC275446 (chimp) and AC212876 (orangutan).

Funding

This work was supported, in part, by the US National Institutes of Health (1R01HG002385 to E.E.E. and R00NS083627 to M.Y.D) and the UW Medical Scientist Training Program (M.L.D. and M.H.D.). This study was supported by a grant from the Italian Ministry of University and Research-MIUR (Project "Futuro in ricerca" 2010 RBFR103CE3). O.P. is a recipient of a Human Frontier Science Program postdoctoral fellowship. X.N. was supported by a US National Science Foundation Graduate Research Fellowship under grant DGE-1256082. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Authors' contributions

Conception and design: MLD, XN, MYD, EEE; Acquisition of data: MLD, XN, CB, LH, MHD, MV, FA, RS, MYD; Data analysis: MLD, XN, OP, BJN, JH, MYD; Manuscript: MLD, XN, EEE. MLD and XN contributed equally to this work.

Acknowledgements

We would like to thank K. Munson and M. Malig for experimental support with the sequencing of BAC clones and K. Penewit for experimental support on the MIP-based sequencing assay. We would also like to thank H. Mefford for helpful discussions and support on analyzing the 1q21 patients. We also thank T. Brown for critical review of the manuscript.

Chapter 3. TRANSCRIPTIONAL FATES OF HUMAN-SPECIFIC SEGMENTAL DUPLICATIONS

Chapter 3 is adapted from an unpublished manuscript under review as of May, 2018:

> Dougherty, M. L., Underwood, J. G., Nelson, B. J., Tseng, E., Munson, K. M., Penn, O., et al. (2018). Transcriptional fates of human-specific segmental duplications. *Submitted*.
>
> First authorship is shared between MLD and JGU.

3.1 ABSTRACT

High-quality sequence assembly and accurate gene annotation are critical to understanding gene evolution but are often complicated in regions of segmental duplications (SDs). Although such regions are particularly important for gene innovation, a basic understanding (e.g., gene structure and protein-coding potential) is still incomplete, incorrect, or lacking for many duplicate genes. We developed a method to yield full-length transcript information and confidently distinguish between nearly identical genes/paralogs. We used biotinylated probes to enrich for full-length cDNA from duplicated regions, which were then amplified, size-fractionated, and sequenced using single-molecule, long-read sequencing technology, permitting us to distinguish between highly identical genes by virtue of multiple paralogous sequence variants. We examined 19 gene families as expressed in developing and adult human brain, selected for their high sequence identity (average >99%) and overlap with human-specific SDs. We characterized the transcriptional differences between related paralogs to better understand the birth-death process of duplicate genes and particularly how the process leads to gene innovation. In 48% of the cases, we find that the expressed duplicates have changed substantially from their ancestral models due to novel sites of transcription initiation, splicing, and polyadenylation, as well as fusion transcripts that connect

duplication-derived exons with neighboring genes. This transcriptional diversity occurs early during evolution likely in in the absence of selection. We detect unannotated open-reading frames in genes currently annotated as pseudogenes, while relegating other duplicates to pseudogene status. Our method significantly improves gene annotation, specifically defining full-length transcripts, isoforms, and open-reading frames for new genes in highly identical SDs. The approach will be more broadly applicable to genes in structurally complex regions of other genomes where the duplication process creates novel genes important for adaptive traits.

3.2 BACKGROUND

Genomic duplication is one of the primary forces by which novel genes evolve within species (Ohno 1970). Numerous studies have shown that recently duplicated sequences often provide the substrates for positive selection and the emergence of gene innovations important for species adaptation (Yim et al. 2014; Duda and Palumbi 1999; Chen et al. 2008; Dennis et al. 2012; Charrier et al. 2012; Florio et al. 2015; Ju et al. 2016). Among apes, for example, novel human-specific genes (e.g., *SRGAP2C*, *ARHGAP11B*, *TBC1D3* and *BOLA2B*) have recently been identified and implicated in promoting progenitor cell proliferation, altering neuronal spine density, increasing excitatory/inhibitory synaptic density, and affecting iron homeostasis early in development (Dennis et al. 2012; Florio et al. 2015, 2016; Ju et al. 2016; Nuttle et al. 2016). Notably, the extent of the duplication with respect to the ancestral transcriptional unit appears to play an important role in determining the potential outcomes for duplicate genes (**Figure 3.1**). While a complete gene duplication might alter dosage, a truncated protein or altered expression pattern points to an entirely different fate for a recently duplicated gene. Despite such examples, our understanding of the birth and death of duplicate genes near their point of inception has been limited by poor gene annotation in these regions. Indeed many of these genes were absent from the annotation of the

human genome until recently. Incomplete annotation further limits our ability to delineate paralog-specific expression patterns that could inform the functional significance of novel genes.

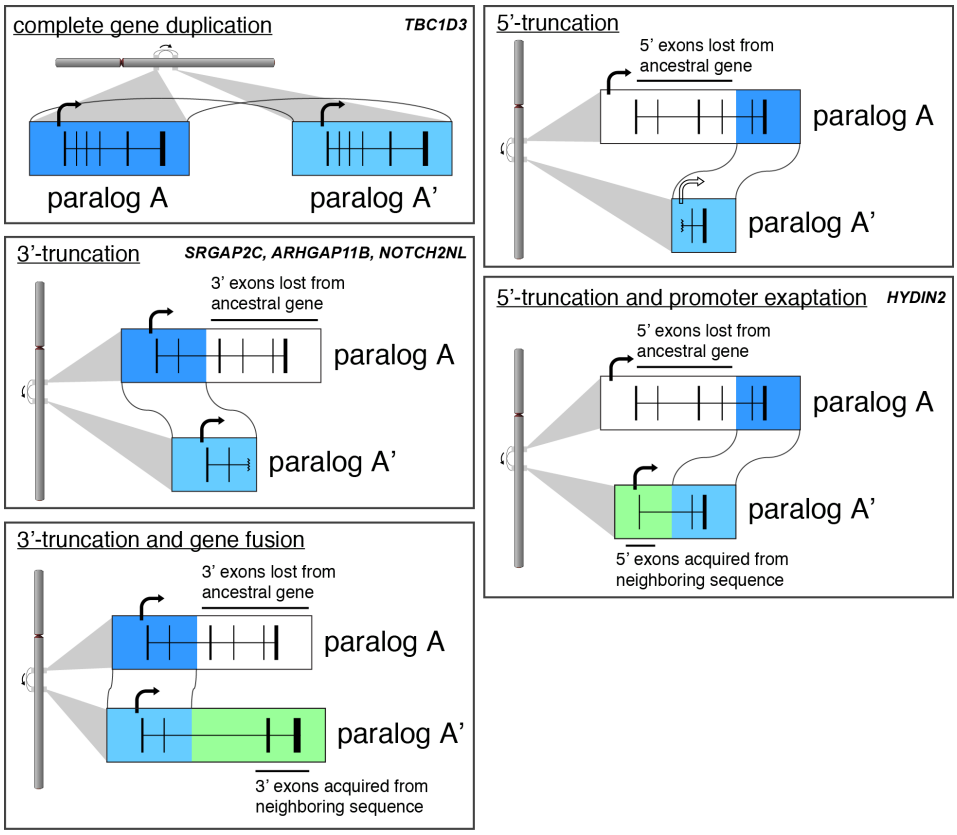


Figure 3.1. Possible transcriptional fates.

For a complete gene duplication, a new copy is created that is most likely to maintain the isoform structure of the ancestor. Incomplete duplications result in only a portion of the ancestral gene being duplicated. This can lead to a truncated duplicate gene or a fusion transcript, where additional exons are acquired from flanking sequence. For 3' truncations, transcription may persist until a polyadenylation signal or a new exon is encountered. For 5' truncations, a promoter unlike that of the ancestral gene must be used if such duplicates are to be transcribed. Specific examples of known human-specific genes by type are indicated.

Duplicated genomic segments of high sequence identity (>90%) (also known as segmental duplications, or SDs) pose particular challenges for gene annotation because: 1) they are enriched in assembly gaps (Alkan et al. 2011), 2) they are more prone to copy number polymorphism among individuals of the same species (Sudmant et al. 2015), and 3) different paralogs are difficult to distinguish because of their high sequence identity. Standard short-read RNA-sequencing (RNA-

seq) data are generally insufficient for characterizing high-identity duplicate genes because most reads do not map uniquely and the data yield almost no information regarding overall transcript structure (i.e., splice isoforms). Paralogous sequence variants (PSVs) are too sparse even within paired-end short-read sequence data to accurately reconstruct paralog-specific isoforms. As a result of such mappability and annotation challenges, these regions are typically excluded from large-scale RNA-seq expression analyses or disease association studies. For example, in a recent analysis by Lan and Pritchard, ~50% of all recent duplicate genes ($ds < 0.1$) were either filtered or deemed unassayable using short-read sequence data (Lan and Pritchard 2016). Similarly, studies that attempt to identify recurrent de novo mutations associated with disease typically exclude such gene models as targets (Iossifov et al. 2014). This already difficult problem is made even harder in organisms like mammals, which display complex patterns of transcription initiation, alternative splicing, intron retention, and polyadenylation (Steijger et al. 2013; Nilsen and Graveley 2010; Barbosa-Morais et al. 2012).

Recent advances in long-read RNA-seq provide the possibility for full-length transcript sequencing obviating the need for transcript assembly. Even among the most recently duplicated regions, long reads would contain a sufficient number of PSVs to be assigned to their respective paralogs with confidence. Long-read transcriptomics, thus, presents a simple solution, although low levels of expression may lead to some duplicate genes being missed by whole-transcriptome RNA-seq. To overcome these limitations, we develop a method that combines advances in long-read, full-length cDNA sequencing with target enrichment to study the transcription of highly identical duplicate genes. We target gene families that have expanded in the human genome following the evolutionary divergence from chimpanzee (~6-7 million years ago), since we hypothesize that their degree of sequence identity (>98.4%) would make them most susceptible to

incomplete or incorrect annotation (Dennis et al. 2017). We use full-length reads from long-read (Pacific Biosciences or PacBio) sequencing technology to generate *ab initio* transcript and gene annotations, then compare these models to current annotation standards (RefSeq (O'Leary et al. 2016) and GENCODE (Harrow et al. 2012)) to demonstrate improved annotation.

These new transcript models allow us to more accurately assess short-read sequencing data in order to explore expression differences among paralogous transcripts. Our analysis identifies new protein-encoding gene models, reclassifies other loci as likely pseudogenes, resurrects predicted pseudogenes back to gene status, and corrects other previously unrecognized annotation errors. More importantly, the analysis provides insight into the diverse and dynamic transcriptional fates of duplicated loci, including their potential to acquire novel promoters and form fusion genes with patterns of expression that sometimes differ from those of ancestral loci. The approach developed here will be more generally applicable to the investigation of gene innovations by duplication as more high-quality genomes emerge in the near future.

3.3 RESULTS

3.3.1 *Targeted capture and sequencing of duplicate gene transcripts*

In order to study the transcription of recently duplicated genes, we sought an approach that met the following criteria: 1) sequence reads would be sufficiently long to carry at least one distinguishing PSV; 2) data would originate from full-length cDNA molecules, representing complete transcripts; and 3) sequence reads would be sufficiently abundant to capture the diversity of major isoforms for any given duplicated locus. The first goal is largely met by application of PacBio sequencing technology. For the second, we employed a widely used strategy based on reverse transcriptase (RT) template switching, which enriches for full-length cDNA molecules

(Zhu et al. 2001). Finally, to focus on duplicate genes, we designed a complementary oligonucleotide capture panel to enrich for cDNA originating from paralogous loci.

We selected gene families found within and near human-specific duplications (HSDs) (Dennis et al. 2017) as targets for probe design (Appendix B: Table S1). We generated two panels of targeting probes: HSD1 (515 probes, Appendix B: Table S2a), representing duplicate loci where there was no evidence of gene disruption (Dennis et al. 2017), and HSD2 (271 probes, Appendix B: Table S2b), representing duplicate loci likely to be polymorphic and enriched for pseudogenes. We also included nine neurodevelopmental genes from single-copy loci to serve as controls for evaluating expression and splicing errors of unique regions as part of our annotation procedure. Probes were designed to exonic sequence within the duplicated portion of the ancestral gene. We used RNA derived from both developing and adult whole brain (pooled from multiple individuals) for cDNA synthesis because previously described HSDs are enriched for roles in the structure and function of the brain (Sudmant et al. 2010; Fortna et al. 2004). We monitored chimeric molecule formation during PCR by implementing a dual barcoding strategy (**Figure 3.2a**) in which 1 of 96 barcodes is appended during first-strand cDNA synthesis to the 3' end of the molecule, and same barcode is appended during second-strand synthesis to the 5' end of the molecule (Appendix B: Table S3 for sequence composition). This "barcode concordant" mode allows us to detect chimeric molecules by the presence of discordant barcodes on the ends of a single cDNA read (Appendix B: Figure S1a), and we estimate the frequency of chimeric molecules identified by mismatched barcodes to be ~1.2% (Appendix B: Figure S1c). We also tested a "barcode discordant" mode, where the relationship between the 5' and 3' barcode is random, in which the pairing of barcodes (96 x 96 arrangements) can be used as a pseudo-unique molecular identifier to monitor for PCR duplicates in low-complexity libraries (Appendix B: Figure S1b). Additionally, we performed

post-capture size selection of libraries using electrophoresis-based fractionation (SageELF, see Methods) to enrich for larger cDNA.

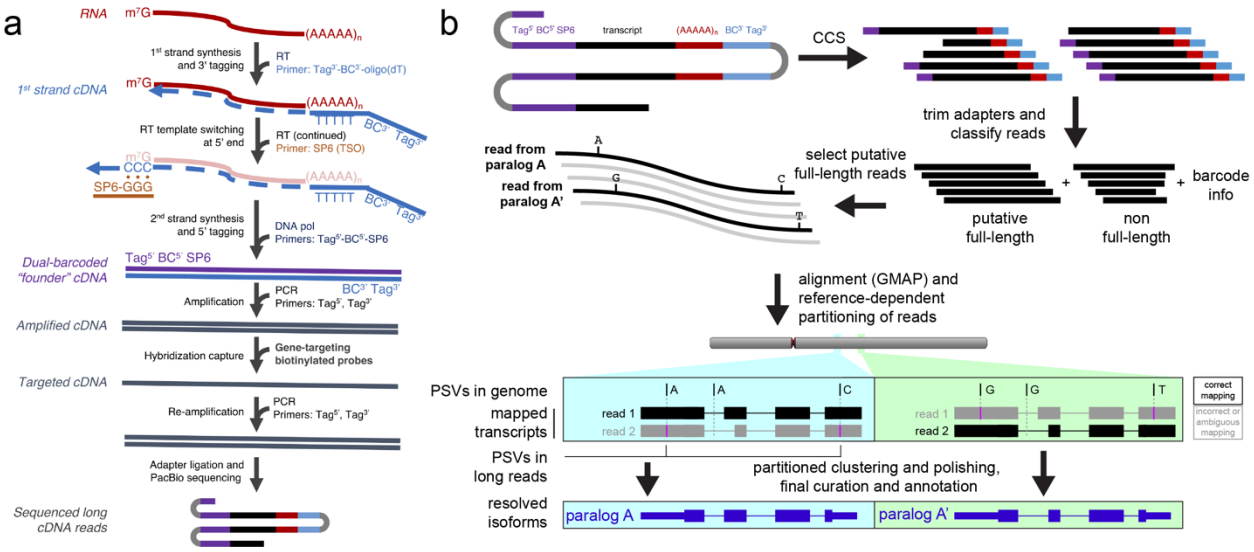


Figure 3.2. Transcript capture and long-read sequencing for resolution of nearly identical duplicate genes.

a) PolyA$^+$ RNA is converted to first-strand cDNA by reverse transcriptase (RT) using a specialized oligo(dT) primer containing the 3' barcode (BC) and an outer sequence tag for later amplification. Template-independent cDNA synthesis extends the 3' end of the cDNA with oligo-dC. RT extends the cDNA by pairing to a template switch oligo (TSO, SP6 sequence) with 3' rG bases. Second-strand synthesis is carried out with DNA polymerase and a primer directed toward the SP6 sequence, containing the 5' barcode and the other outer tag. After ssDNA depletion (not shown), the recovered ds-cDNA founder molecules are amplified before biotinylated probes designed to genes of interest are used for hybridization capture. A final PCR step on the target-enriched cDNA generates double-stranded molecules for long-read sequencing. **b)** As part of a modified Iso-Seq workflow, sequences are first error-corrected through circular consensus sequence (CCS) generation. Then for each read, the sequences flanking the transcripts are identified and trimmed. If such sequences are present on both ends, reads are designated as putative full-length (pFL). pFL reads are mapped to the human reference (GRCh38) where the presence of multiple PSVs along the long read promotes accurate mapping even in the presence of sequencing errors. To avoid confounding paralogs, confidently mapped reads (MAPQ > 40) are partitioned into genomic segments before the Iso-Seq cluster step is performed.

Using our method, we sequenced a total of 40 SMRT cells, including unenriched whole-transcriptome controls (n = 4), HSD1-enriched cDNA (n = 30), and HSD2-enriched cDNA (n = 6) on the PacBio RS II (Appendix B: Table S4) sequencing platform. In circular consensus sequence (CCS) generation, multiple passes of the polymerase around a covalently closed sequencing molecule are used for consensus-based correction (**Figure 3.2b**). In total, 1.4 million CCS reads were generated, divided between developing and adult brain cDNA sources (Appendix B: Table

S5). As expected, longer CCS reads show lower read accuracy due to fewer full passes of the sequencing polymerase (Appendix B: Figure S2). Because the primary error modality in PacBio sequencing involves indels, these errors are unlikely to be mistaken as PSVs and, as such, do not significantly interfere with paralog assignability. Of the CCS reads generated for our HSD1 panel, 82% (adult brain) and 77% (developing brain) were designated by the PacBio Iso-Seq analysis pipeline as full-length due to the presence of the expected barcode, primer sequence, and polyA tail. Since some of these do not represent truly full-length isoforms due to possible 5' RNA degradation, we refer to them as *putative* full-length (pFL) reads. pFL reads mapped to the human reference genome (GRCh38) using GMAP (v 2015-07-23) were used for further analysis (**Figure 3.2b**).

Mapping of these pFL reads revealed an on-target rate of 65% (adult brain) and 62% (developing brain) in HSD1-enriched cDNA. We estimate that this approach enriched for target genes by >250-fold (Appendix B: Table S5). Similar results were achieved for the second probe panel, HSD2. Out of the original set of 39 duplicate gene families screened, we focused on 19 for a more detailed analysis. These were chosen by the following criteria: 1) they contain at least one known protein-coding gene and the gene model contains multiple exons; 2) the corresponding genomic loci are correctly assembled and present in more than one copy in the latest build of the human reference genome (GRCh38); 3) the ancestral gene was expressed sufficiently in our data to make inferences about transcriptional differences between the ancestor and duplicates; and 4) the gene was confirmed to have been generated through segmental duplication (as opposed to retrotransposition). The final set includes gene families corresponding to *SRGAP2*, *NOTCH2*, *ARHGEF5*, *ARHGAP11*, *PTPN20*, *FRMPD2*, *CHRNA7*, *GTF2I*, *GTF2IRD2*, *ROCK1*, *CORO1*, *HYDIN*, *FAM72*, *SLX1B*, *GPR89*, *FCGR1*, *NFC1*, *CD8B*, and *BOLA2* (Appendix B: Table S6).

3.3.2 *Classification of duplication events*

We initially classified each HSD as complete or incomplete depending on whether the SD event in the genome carries the entire transcriptional unit of the ancestral gene or a merely a truncated portion (**Figure 3.3a**, see also **Figure 3.1**). Of the 19 gene families (or 12 non-ancestral paralogs), eight are "complete" and these tend to be those with smaller ancestral genes. Note that some gene families (e.g., *GTF2IRD2*) contain both "complete" and "incomplete" duplicates, hence the sum of the two categories exceeding the total count. We further classify the incomplete HSD gene families (n = 12, 19 non-ancestral paralogs) by what portion of the gene body is truncated relative to the ancestral gene. We categorize duplicates as 3' truncations and 5' truncations. 3'-truncated paralogs retain ancestral transcription start site (TSS) but lack some downstream exons (e.g., *SRGAP2C* (Dennis et al. 2012; Charrier et al. 2012), *ARHGAP11B* (Florio et al. 2015, 2016)), while 5'-truncated paralogs have lost upstream exons and their ancestral promoter (e.g., *CHRFAM7A* (Gault et al. 1998)).

We took advantage of the full-length cDNA sequences to classify the consequences of the SD with respect to transcript or isoform structure of the duplicate genes. Truncated transcripts are simply shortened versions of the ancestral transcript while fusion transcripts are linked to upstream or downstream sequence through splicing to a gene segment homologous to another annotated gene. By this metric, recently duplicated genes show a range of transcript models. Based on counts of pFL reads, we classified truncated genes as predominantly truncated (<20% pFL reads belonging to this gene demonstrate bridging transcription), predominantly fusion (>80% pFL reads demonstrate bridging transcription), or both (**Figure 3.3a**, see also Appendix B: Table S7). We distinguished exaptation events (inclusion of a novel exon or promoter) from gene fusions when bridging exons are themselves not known homologs to any other gene. In this study, promoter

exaptation rescues the transcriptional activity of two 5'-truncated genes, *ROCK1P* and *HYDIN2*, the latter confirming earlier observations (Dougherty et al. 2017). Only two of the 5'-truncated gene duplications examined (*GTF2IRD2P1* and *CORO1B*) have lost expression in brain as a consequence of promoter loss; thus, of the 31 duplicate paralogs analyzed, 29 retain expression.

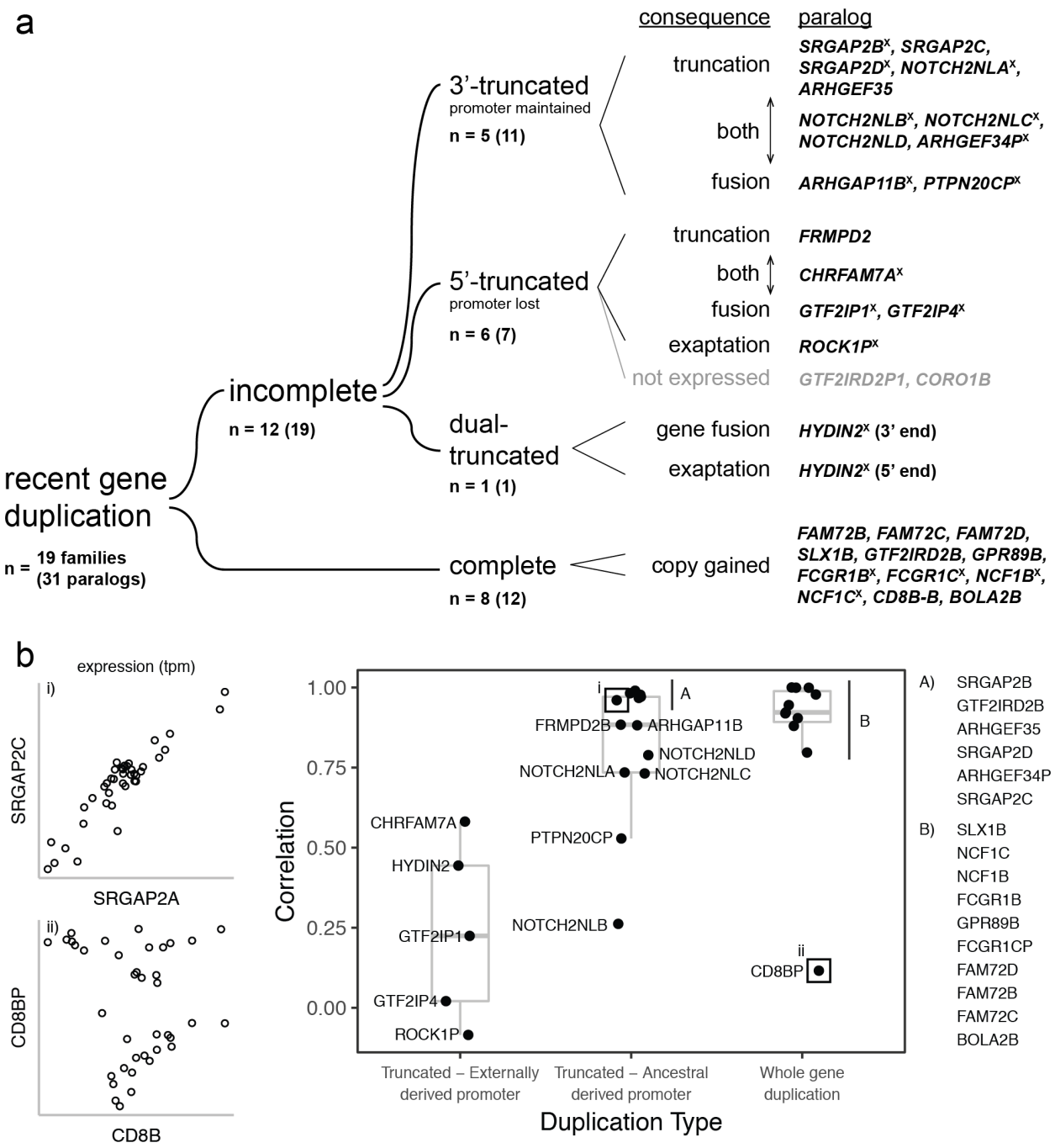


Figure 3.3. Transcriptional fates of human-specific duplicate genes and expression correlation between ancestral and duplicate gene copies.

a) We classify 19 gene families (31 duplicate paralogs in GRCh38) by the transcriptional characteristics of the duplicate genes. In 8 of 19 gene families and 12 of 31 paralogs, the duplication includes the complete gene (with respect to the canonical isoform). More common are incomplete gene duplications, of which 5 of 12 gene families and 11 of 19 paralogs are 3' truncated (whereby the ancestral promoter is maintained in the duplicate gene) while 6 of 12 gene families and 7 of 19 paralogs are 5' truncated (whereby the ancestral promoter is lost). The outcomes of such truncated duplications can be simply shortened versions of the ancestral gene ("truncation") or transcript fusion with adjacent sequence ("fusion"), and often both are observed. For 5' truncations, we also observe the phenomenon of exaptation of upstream exons and regulatory elements, which provide a new promoter for what would presumably be

otherwise transcriptionally silent genes (2 of 19 gene families). Note that duplicates of *GTF2IRD2* are classified as both complete and incomplete. **b)** We estimated expression similarity between ancestral and duplicate copies by calculating the pairwise correlation of the median expression levels across GTEx tissues. Duplicate genes whose promoter was included in the human-specific SD show expression patterns that are more similar to their ancestors than those that acquire it from new sequence.

Excluding minor (<2% of isoforms) products for each duplicate gene, we finally characterized the protein-coding potential of sequenced duplicate gene isoforms. While open-reading frames (ORFs) more than 100 amino acids in length could be found across duplicate gene families, we specifically asked whether isoforms were present in which the entirety of the ancestral ORF was intact (complete gene duplications) or the entirety of the duplicated portion of the ancestral ORF was intact (partial gene duplications). Overall, among the 29 expressed duplicate paralogs, the integrity of the duplicated portion of the ORF has been compromised in 17 (58%), by either acquired frameshift mutations, changes in splicing, or multiple events. The 12 duplicate genes with "intact" ORFs include *SRGAP2C*, *NOTCH2NLD*, *ARHGEF35*, *FRMPD2B*, *FAM72B*, *FAM72C*, *FAM72D*, *SLX1B*, *GTF2IRD2B*, *GPR89B*, *CD8BP*, and *BOLA2B*. The relationship between ORF length, integrity, and gene function is a complex one as ORF-disrupting mutations may in some cases confer critical functional activity (Florio et al. 2016).

3.3.3 *Frequent transcript fusion observed in 3'-truncated HSD genes*

Approximately, one-third of the duplicate paralogs are 3' truncations of the ancestral gene and all show evidence of transcription (**Figure 3.3a**). Included in this set are gene innovations (e.g., *SRGAP2C*, *ARHGAP11B*) recently implicated in cortical expansion and increased dendrite density of the human brain (Florio et al. 2015, 2016; Dennis et al. 2012; Charrier et al. 2012). Since such duplicates retain the 5'-proximal regulatory sequence of the ancestral locus, the pattern of expression, as expected, is highly correlated with that of the ancestral gene (**Figure 3.3b**). We find that "fusion" transcripts are common, linking the duplicate gene segment with exons from

downstream sequence, though they rarely alter the ORF. In some cases, these fusion transcripts represent major isoforms. The relative abundance varies by gene family and paralog (Appendix B: Table S7). For example, *SRGAP2C* transcripts are predominantly truncations (5% fusion), while for *SRGAP2B* the proportion of fusions increases (14%). Among *NOTCH2NL* paralogs, the proportion of fusion transcripts ranges widely from 16-49%, exclusively with adjacent members of the *NBPF* gene family, in which copy number variation has been associated with cranial size (Dumas et al. 2012). Only a small fraction (<2%) of such *NBPF* fusions, however, maintain an ORF.

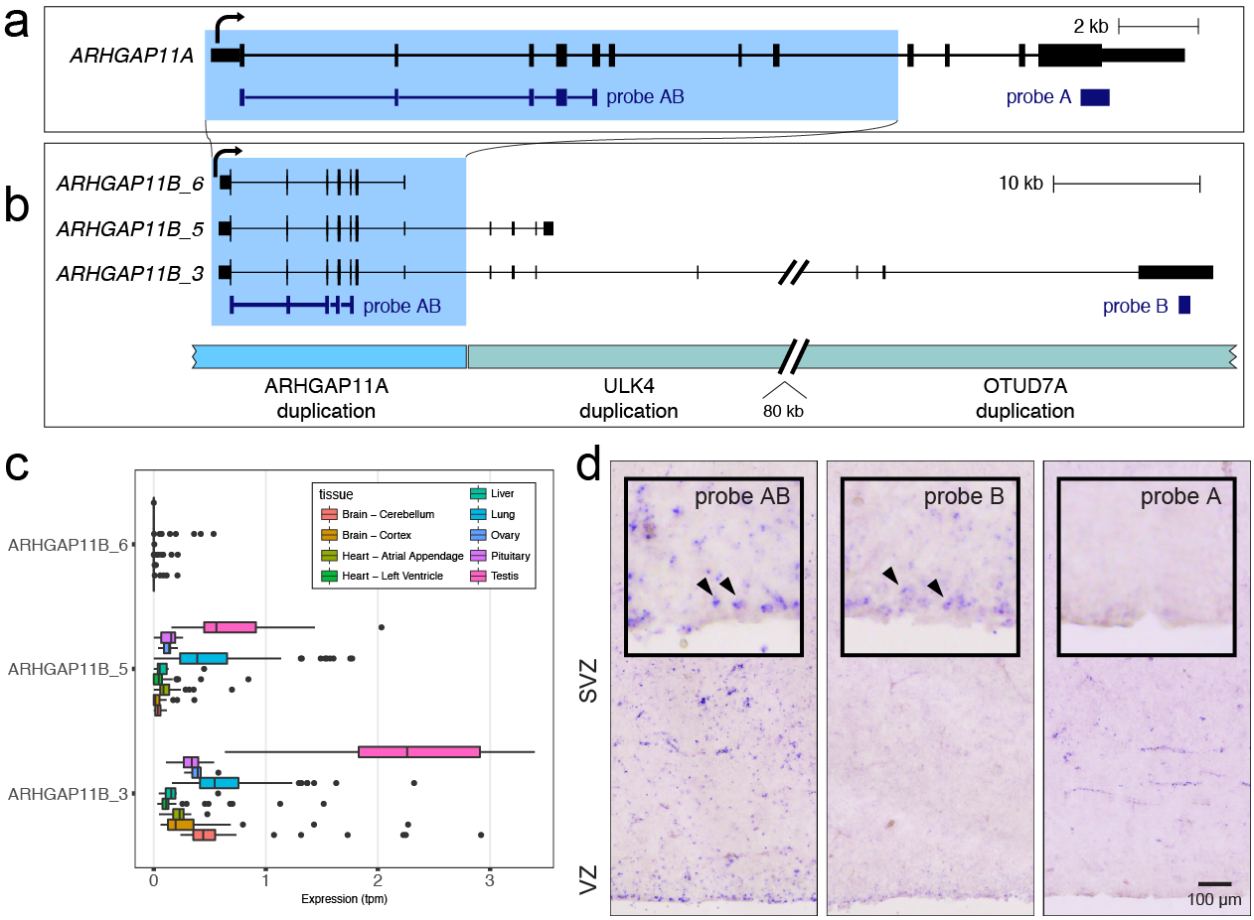


Figure 3.4. Identification of a longer fusion isoform of *ARHGAP11B* expressed in dividing radial glia.

a) Partial duplication of *ARHGAP11A* resulted in *ARHGAP11B*. **b)** We identified three isoforms of *ARHGAP11B* based on full-length transcript sequencing and these are shown in the context of SDs. The "long" isoform

(ARHGAP11B_3) extends deeply into adjacent duplications; a "medium" isoform (ARHGAP11B_5) has four additional exons beyond the duplication shared with *ARHGAP11A*; a "short" isoform (AHRGAP11B_6), consists entirely of sequence shared with *ARHGAP11A*. **c)** Expression estimates for the three isoforms in select tissues support the prominence of the long isoform but limited evidence for expression of the short isoform. **d)** *In situ* hybridization performed on sections of developing cortical brain (gestational week 18) indicates expression in cells along the ventricle of the ventricular zone (VZ, arrowheads, magnified inset), where radial glia undergo mitosis, consistent with long form of *ARHGAP11B* expressed specific to dividing ventricular radial glia, but missing from outer radial glia. Probe targets are shown in panel a. Note that probe B is not predicted to hybridize to *OTUD7A* itself.

ARHGAP11B has been implicated in basal progenitor amplification and neocortical expansion (Florio et al. 2015, 2016). The key isoform studied by Florio et al. is a truncated form of the ancestral locus, *ARHGAP11A*, with a short, modified C-terminus due to an acquired splice-site mutation (**Figure 3.4a**). While we observe this specific *ARHGAP11B* isoform (isoform "ARHGAP11_6"), we also observe prominent longer isoforms that initiate at the same shared ancestral promoter but differ dramatically in their downstream exons (**Figure 3.4b**). Continuing beyond the annotated polyadenylation site, these longer isoforms extend downstream into other SDs, including duplications of *ULK4* (isoform "ARHGAP11B_5") and *OTUD7A* (isoform "ARHGAP11B_3"). Expression estimates that include these new isoforms of *ARHGAP11B* suggest greater abundance in adult brain tissues (**Figure 3.4c**). We designed probes that would detect expression in aggregate (AB), of *ARHGAP11A* specifically (A), and of the newly discovered longest isoform of *ARHGAP11B* (B) and performed *in situ* hybridization on developing human brain (**Figure 3.4d**). We find that the longer isoform is expressed specifically along the ventricle where radial glia undergo mitosis. The staining is not as strong with probe B as with probe AB, indicating that this isoform is not exclusively responsible for *ARHGAP11B* expression in these cells. However, it can be said that with current annotations alone, the picture of *ARHGAP11B* activity in these key neural progenitor cells is incomplete.

3.3.4 *Promoter loss and retention contribute to duplicate gene expression patterns*

Surprisingly, we also find evidence of transcription for the majority (5/7 paralogs) of HSDs associated with 5' truncations (**Figure 3.3a**). Since the TSS was lost during the duplication, transcription necessitates the acquisition of a novel TSS. Similar to the 3'-truncated HSDs, some duplicates encode primarily truncated transcripts, their TSS derived from an internal promoter (e.g. *FRMPD2*) while others represent fusion events, deriving their TSS from new upstream sequence. The latter is the case for the partial duplicates of *GTF2I* (*GTF2IP1* and *GTF2IP4*) whose promoter and first exon originate from a duplication of the adjacent *GATSL2*.

ROCK1P is derived from the four terminal exons and a portion of the 5th exon of the 33-exon serine/threonine kinase, *ROCK1*, which duplicated to the telomeric end of chromosome 18, adjacent to a 5 kbp satellite repeat from which the TSS was acquired (Appendix B: Figure S3). Our capture-based sequencing approach identifies two predominant TSSs (TSS1 and TSS2) from the adjacent upstream sequence, which we refer to as the "promoter block" (Appendix B: Figure S3). This promoter block is primarily composed of beta and LSAU satellite repeat sequence; TSS1 maps within a beta satellite repeat ~900 bp upstream from the *ROCK1* duplication break point, and TSS2, which contains the microRNA MIR8078, is found ~200 bp upstream from the LSAU3 breakpoint. TSS2, however, provides an alternate first exon, with a novel translation initiation codon and a potential short ORF of 216 amino acids (158 shared with *ROCK1*). Based on GTEx data, we estimate that highest expression of *ROCK1P* is in the testis, consistent with the tissue expression of MIR8078 (NR_107045.1, miRBase (Meunier et al. 2013; Kozomara and Griffiths-Jones 2011)).

We hypothesized that this may indicate a general trend, that when a new promoter is acquired by a 5'-truncated duplication, it would direct the expression of the new gene fusion, as was

observed in *HYDIN2* (Dougherty et al. 2017). We divided the duplicate genes into three categories based on the nature of the duplication: truncated with a different promoter (i.e., loss of ancestral TSS), truncated with the same promoter (usually loss of the ancestral polyadenylation site), and whole-gene duplication. We then measured the median expression level of the duplicate gene and ancestral gene in available tissues from GTEx and measured the correlation as a proxy for preservation of expression pattern (**Figure 3.3b**). We find that when the duplicate gene maintains the same promoter, the expression correlation coefficient is almost always quite high, while when a new promoter is acquired, expression correlation is variable.

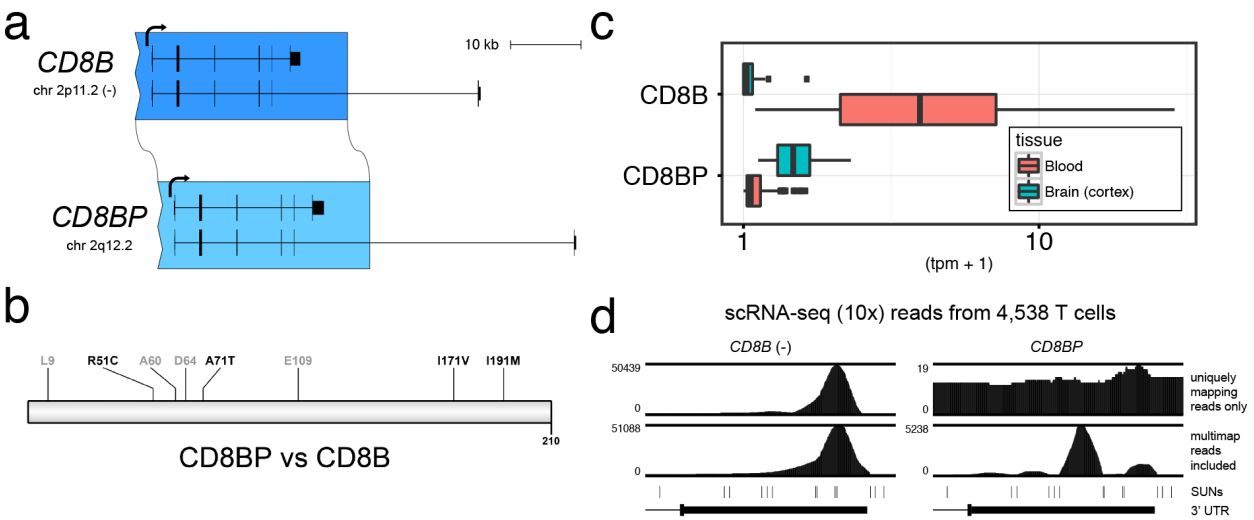


Figure 3.5. Cortical expression of *CD8BP* and maintenance of a complete ORF.

a) *CD8B* was duplicated in full (canonical isoform) from the p-arm to the q-arm of chromosome 2, generating *CD8BP*. We identify expressed transcripts from *CD8BP* that resemble those of *CD8B* and maintain the 210 aa ORF but with four amino acid replacements (**b**). **c)** *CD8BP* has all but lost its ancestral expression in blood and instead has gained expression in the brain (cortex). **d)** Single-cell RNA-seq data from T cells (10x Genomics) supports that *CD8BP* expression has been lost specifically in T cells likely due to the duplication excluding a tissue-specific enhancer (see text).

A notable exception to this rule is the case of *CD8B* and its human-specific duplicate *CD8BP*, which includes the ancestral promoter but has a markedly different derived expression (**Figure 3.3b**). Together with CD8A, CD8B forms a heterodimer that serves as a coreceptor for the T cell

receptor and is the defining marker of CD8+ T cells, which respond to intracellular antigens such as those found in virally infected or cancerous cells. *CD8BP* is the consequence of a whole-gene duplication (with respect to the major isoform) across the centromere of chromosome 2 (**Figure 3.5a**). Despite its pseudogene annotation, we find that the 210 amino acid ORF of *CD8B* is maintained in *CD8BP* with four substitutions (**Figure 3.5b**). However, similarity in expression between the paralogs is among the lowest in the pairwise comparisons we measured ($\rho = 0.10$ for correlation across tissues). The most dramatic tissue-specific changes include a near total loss of expressed *CD8BP* in whole blood and a substantial gain in brain tissues, including cortex (**Figure 3.5c**). We confirmed that this loss of expression can be attributed specifically to T cells by examining single-cell RNA-seq from 4,538 T cells derived from a healthy donor generated using the 10x Genomics platform (data obtained from https://support.10xgenomics.com/single-cell/datasets/t_4k), which generates sequence reads from the 3' end of the transcript (**Figure 3.5d**). While most reads map equally to both paralogs, when strict mapping criteria (MAPQ > 40) are applied, 99.9% of reads map preferentially to the 3' untranslated region (UTR) of *CD8B*, confirming that *CD8BP* is not expressed in circulating T cells. Therefore, it is unlikely that *CD8BP* expression defines a subtype of T cells, but rather it is expressed in entirely different cell types.

We hypothesize that the loss of cis-regulatory elements at the ancestral locus may be partly responsible for the expression change. An array of enhancers has been defined in the ~100 kbp region that includes *CD8B* and its partner *CD8A* (Kieffer et al. 1997, 2002), and a complex combinatorial relationship between these enhancers is thought to direct cell type and stage appropriate expression (Kioussis and Ellmeier 2002). However, only two of the six enhancers are included in the extent of the duplication and, as such, the loss of the four defined enhancers as well as other elements may be responsible. Interestingly, this is also one of the least copy number

variable duplicate sequences in the human population—a property that appears to associate with functional duplicate copies (Dennis et al. 2017).

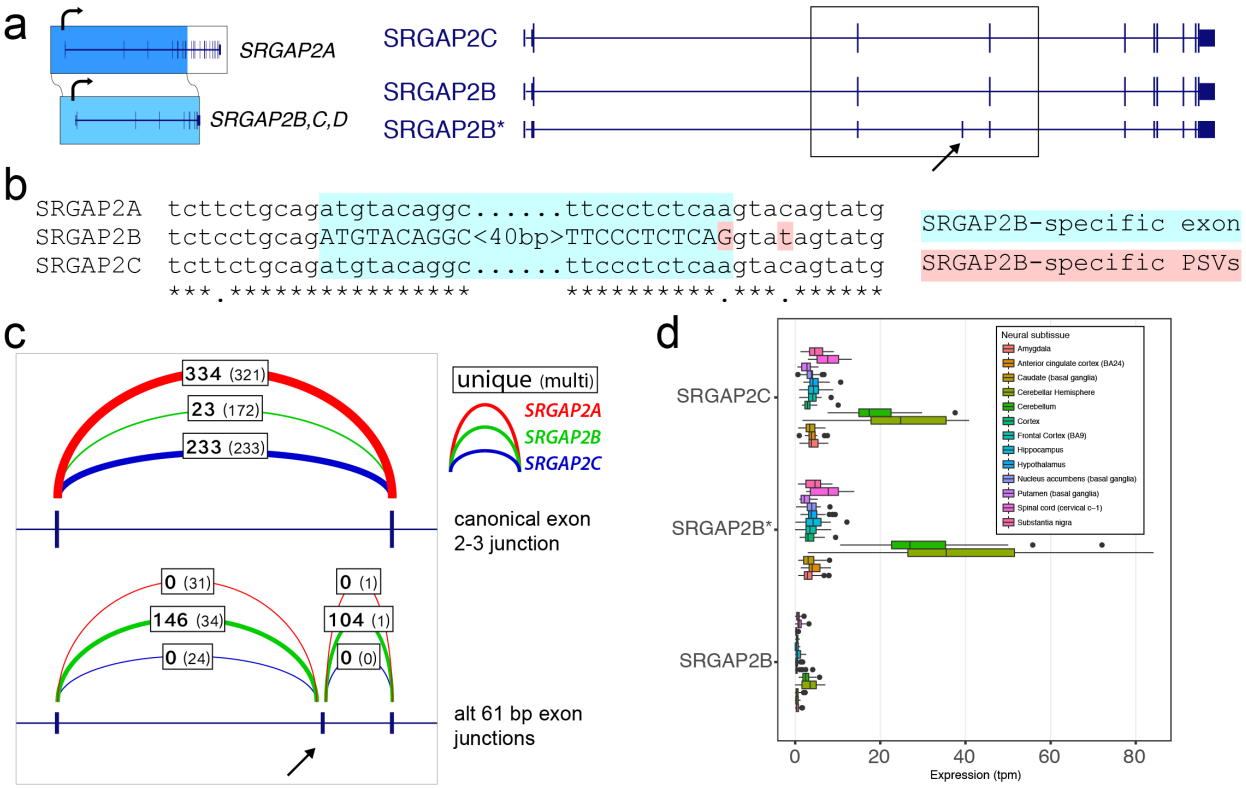


Figure 3.6. Inclusion of a 61 bp exon and premature stop codon in *SRGAP2B*.

a) *SRGAP2B* and *SRGAP2C* are duplicate copies of *SRGAP2A*, created by an initial 3'-truncated duplication ~3 million years ago. From the long-read capture data, we identify a major isoform containing an additional 61 bp exon (arrow) specific to *SRGAP2B*. **b)** Alignment of the 61 bp exon (highlighted in blue) and flanking sequence identifies a key nucleotide change in the -1 position of the splice donor (highlighted in red). This A-to-G transition is predicted to substantially increase the strength of the splice donor signal. **c)** Intron-spanning reads from RNA-seq performed on cerebral cortex (GTEx) confirm that this additional exon is frequently included in *SRGAP2B* transcripts, while rarely included if at all in *SRGAP2A* and *SRGAP2C*. **d)** Expression estimates for isoforms shown in (a) generated using Kallisto with data from GTEx corroborates that the isoform of *SRGAP2B* that includes the 61 bp exon is the predominant one in brain tissues. These subsequent mutational changes likely non-functionalized *SRGAP2B* leading to the fixation of the granddaughter duplicate *SRGAP2C* in the human population.

3.3.5 *Splicing as an indicator of selection acting on duplicate genes*

Full-length isoform characterization accompanied by expression analysis facilitates identification of shifts in the predominant isoforms between duplicate paralogs. We observed major differences

in splicing for three gene families: *ARGHAP11*, *SRGAP2* and *FCGR1*. For example, in contrast to the highly uniform splicing of *SRGAP2C*, we identify two major isoforms of *SRGAP2B*. The more common isoform includes a 61 bp exon not observed among the other paralogs, leading to premature truncation of the otherwise highly homologous ORF (**Figure 3.6a**). The splice donor of this exon, *SRGAP2B*, contains two distinguishing nucleotide variants, most importantly an A-to-G transition at the -1 position with respect to the 5' splice site (**Figure 3.6b**). These *SRGAP2B* mutations increase the strength of this cryptic splice donor (MaxENT score 0.24 *SRGAP2C*, 8.73 *SRGAP2B*) (Yeo and Burge 2004) making the ORF-truncating transcript the predominant form. Counts of intron-spanning reads from short-read RNA-seq data from the human brain (cortex, GTEx (GTEx Consortium 2013)) corroborates that this frameshifting exon is a feature unique to *SRGAP2B* (**Figure 3.6c**), and transcript-wide expression estimates concur (**Figure 3.6d**). This difference helps explain why *SRGAP2B* is copy number polymorphic, why this particular paralog's transcript is subject to nonsense-mediated decay, and why *SRGAP2C* ultimately replaced the older duplicate *SRGAP2B* as the functional and fixed copy in the human species (Dennis et al. 2012).

Similarly, splicing patterns differ between *FCGR1A* and *FCGR1B*, despite their shared 99.0% nucleotide identity at the genomic level (Appendix B: Figure S4a). Most of this difference involves the penultimate exon, which is constitutive in *FCGR1A*, but a cassette exon of varying length in *FCGR1B* likely a result of a 4 bp deletion at the splice donor site. Using Shannon's entropy of normalized isoform abundance as a metric of increased isoform diversity (Ritchie et al. 2008), we find that entropy for *FCGR1B* (3.81 bits) is much higher than that of *FCGR1A* (1.92 bits) ($p = 1.3e-7$, Kolmogorov-Smirnov test). For example, ~80% of sequence reads come from two major isoforms of *FCRG1A*, in contrast to *FCRG1B* where 80% of sequence reads are distributed among 12 isoforms (Appendix B: Figure S4b). While in some cases (e.g., *FCGR1B*, *SRGAP2B*) disruptive

splicing mutations appear to be associated with relaxed selection, in others such as *ARHGAP11B* they are thought to be the key mutational event for neofunctionalization of the duplicate (Florio et al. 2016).

3.3.6 *Exon exaptation and novel gene annotations*

Our analysis of the *GTF2IRD2* gene family, associated with Williams-Beuren syndrome, (**Figure 3.7a**) identifies two novel isoforms. The first is an out-of-frame fusion with *STAG3L2*, a high-copy pseudogene upstream of *GTF2IRD2* that we estimate accounts for 33% of *GTF2IRD2* transcripts in brain. The second novel isoform contains a distinct first exon that is derived from the DNA-binding domain of the Tigger7 DNA transposon and adds 162 N-terminal amino acids (**Figure 3.7b**). Comparative sequence analysis shows that this novel N-terminus has been conserved throughout primate evolution and has been subjected to purifying selection ($dN/dS = 0.019$, $p < 0.01$) (**Figure 3.7c**). A similar phenomenon can be observed in the ancestral *GTF2I*, although inclusion of the Tigger7 repeat is associated with a much less abundant isoform (**Figure 3.7d**, Appendix B: Table S8). Taken together, this analysis provides strong evidence that the human-specific gene *GTF2IRD2* (as well as *GTF2IRD2B* and *GTF2I*) has a currently unannotated isoform that encodes a protein with a distinct N-terminal domain derived from Tigger7, of the TcMar-Tigger DNA transposon family of repeats, and that this protein-coding sequence is under significant purifying selection throughout the primate phylogeny.

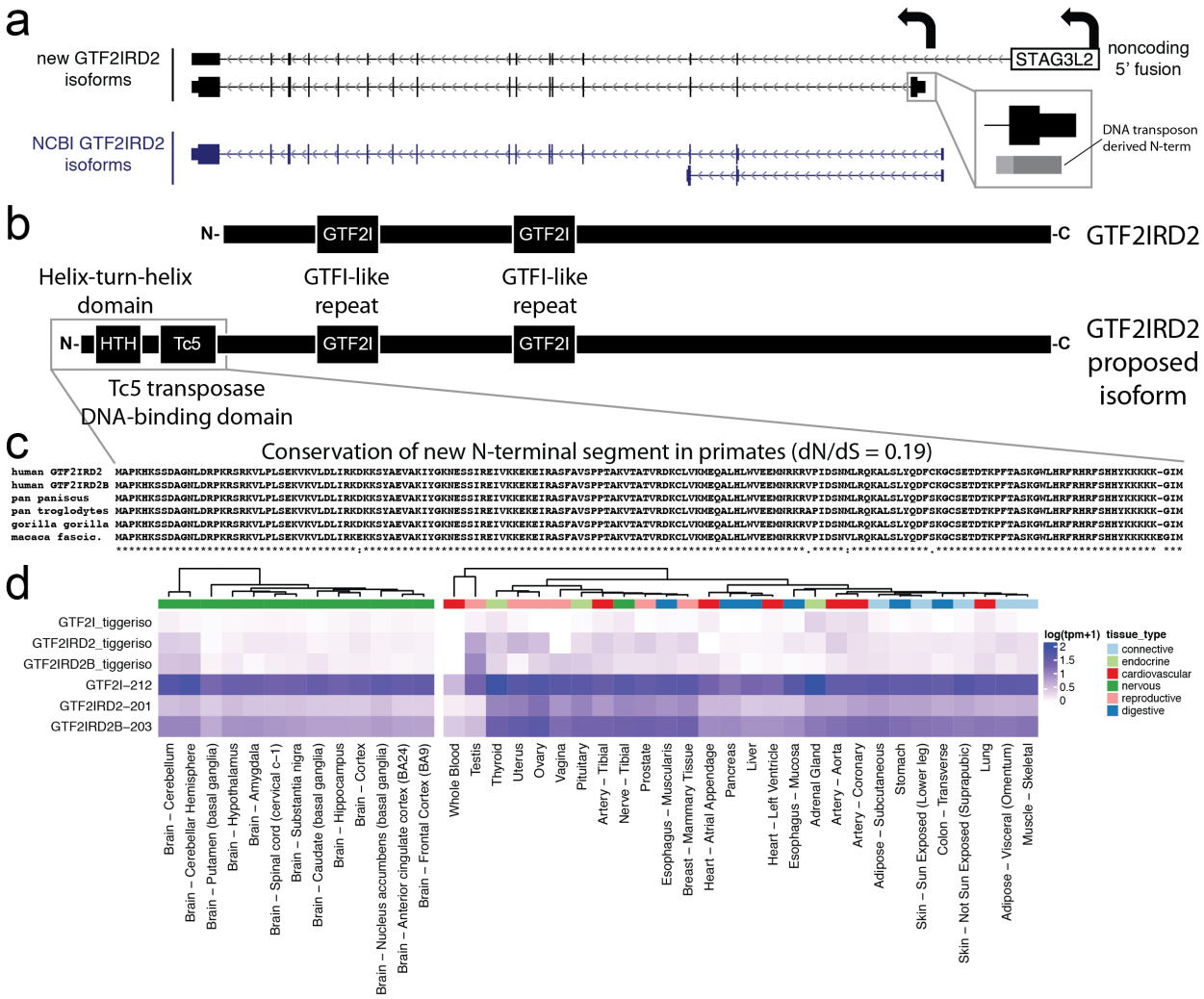


Figure 3.7. Discovery of novel N-terminal segment DNA-binding domains for *GTF2IRD2* and *GTF2I*.

a) Two classes of novel isoforms were identified for the human-specific duplicate gene *GTF2IRD2* (transcribed from right to left in this view), shown above the current NCBI gene annotations. The upper isoform consists of an out-of-frame gene fusion to upstream gene *STAG3L2* while the lower isoform includes an alternative first exon, derived from a DNA transposon (TcMar-Tigger family), yielding a new N-terminal segment. **b)** The proposed *GTF2IRD2* isoform contains two additional N-terminal DNA-binding domains (helix-turn-helix domain and a Tc5 transposase DNA-binding domain) derived from the DNA transposon Tigger7. **c)** A multiple sequence alignment of the newly identified N-terminus predicted protein sequence shows conservation among primates. **d)** A heatmap of expression levels estimated using Kallisto across tissues (GTEx) shows that the canonical form is more broadly expressed than the transposon-containing isoform with the exception of the testis.

CHRFAM7A is a human-specific fusion gene that has been associated with neuropsychiatric disease (Flomen et al. 2006; Rozycka et al. 2013; Casey et al. 2012) and is thought to interact in a dominant negative fashion with the normally homopentameric *CHRNA7* to

decrease its efficiency as an ion channel (Araud et al. 2011). Current annotations have the longest ORF spanning the boundary between *CHRNA7* and *FAM7A* duplications; however, this is disrupted by a 2 bp polymorphism common in European populations (Appendix B: Figure S5). All current models of *CHRFAM7A* have a shortened ORF that initiates at this exon but also contain multiple upstream exons, which results in either a multi-exon 5' UTR or a complicated annotation with shorter upstream ORFs. We identify a new isoform of *CHRFAM7A* where transcription is initiated at this exon and, as a result, would place the 2 bp deletion polymorphism within a short 5' UTR. This isoform appears more likely to result in a translated product capable of interacting with the ancestral *CHRNA7*.

3.4 DISCUSSION

We have developed a capture-based method to target the duplicated regions of genomes and enrich for the recovery and sequencing of full-length transcripts. This method can be systematically applied to duplicated regions once they are defined within a genome to characterize the intron/exon structure and to identify those duplicates most likely to maintain an ORF. This is important because duplicate genes are often associated with evolution of novel genes important for species adaptations (Yim et al. 2014; Sulak et al. 2016), but such innovations occur in a background where the most frequent evolutionary fate is pseudogenization (Lynch and Conery 2000). Thus, the method allows an investigator to quickly focus on those loci with the potential to encode proteins. For most genomes, annotation of the evolutionarily youngest duplicates lags behind because such transcripts are difficult to discern among paralogous loci using short-read RNA-seq data and homology to the ancestral gene structure is used to guide transcript models (Li et al. 2015). Not surprisingly, duplicates are often the last genes in genomes to be discovered and correctly annotated (Church et al. 2009; Sudmant et al. 2010; Dennis et al. 2012) and are more often

excluded from genome-wide analyses. The recovery of full-length transcripts coupled with long-read sequencing provides a less biased approach for investigation of transcriptional fate among recent paralogs and may be broadly applied to any genome where duplicated regions have been characterized. Once defined, these paralog-specific transcript models allow us to better interrogate expression differences using available short-read data.

An important finding of our study of HSDs is that most of the duplicates are transcriptionally active despite the fact that only portions of the ancestral genes are duplicated. In our study, 94% (29/31) of HSD paralogs show evidence of transcription even though 30 of these genes are incomplete (**Figure 3.3**) with respect to ancestral structure. This is especially surprising for 5' truncations where the promoter has been lost as part of the duplication event. Of seven such events, five showed evidence of transcription, although these events were more likely to show differential expression patterns when compared to 3' truncations, which showed similar spatial temporal expression patterns to the ancestral gene. Overall, 18% (3/17) of the HSD genes show diverged patterns of expression and such rapid changes in expression patterns might be expected for SDs when compared to whole-genome duplication events. Studies of a recent whole-genome duplication in the common carp (Li et al. 2015), for example, indicate that 92.5% of the genes show some evidence of co-expression.

Expression dissociation between paralogs is sometimes taken as evidence of neofunctionalization or subfunctionalization (Lan and Pritchard 2016). Our data suggest, however, that expression dissociation can occur much more rapidly because of two SD properties: the first being that HSDs are most likely to be incomplete (i.e., truncated with respect to the ancestral gene model); second, such duplications are interspersed, preferentially duplicated to regions enriched for other incomplete duplications (Dennis et al. 2017). As a result, HSDs are likely to be juxtaposed

beside other incomplete duplications providing the raw material for regulatory (e.g., *CD8BP*) and exonic exaptations that quickly alter the transcript model and the expression profile of the new duplicate without the need for purifying selection (Hahn 2009). In other cases, such as the *GTF2IRD/GTF2I* gene family, which has been associated with hypersociability in both humans and dogs (vonHoldt et al. 2017), we have identified entirely novel protein-encoding DNA-binding domains derived from an ancient DNA transposon.

In 48% (16/33) of HSD paralogs, the gene models have changed more substantially from the ancestral gene as a result of 5' extensions, 3' extensions, and gene fusions. Among HSDs, partial gene duplication is the predominant mode and there is evidence that such transcripts have the potential to encode truncated proteins that, lacking protein domains, function differently than their ancestor. This is the model for the human-specific duplicate genes *CHRFAM7A* (Araud et al. 2011), *SRGAP2C* (Charrier et al. 2012), and *ARHGAP11B* (Florio et al. 2016)—the latter two fixed for copy number and associated within neuronal spine maturation and cortical neuron expansion, respectively. In the case of *SRGAP2C* and *CHRFAM7A*, the truncated duplicate acts antagonistically inhibiting the function of the ancestral protein, and, thus, by definition may have been partially functional at birth acting in a dominant negative manner. However, for both *SRGAP2C* and *ARHGAP11B*, additional mutations occurred subsequent to the duplication event—missense changes in the case of *SRGAP2C* (Sporny et al. 2017) and a splice-site mutation in *ARHGAP11B* (Florio et al. 2016)—that apparently refined (e.g., *SRGAP2C*) or even activated a new function (e.g., *ARHGAP11B*). These differences are confirmed in the full-length transcripts that were generated, although our analysis predicts additional novel isoforms whose functions have not yet been investigated.

A key difference from previous studies on gene duplications is our focus on the most recent events and therefore the most identical duplications. The majority of duplicate genes are thought to become pseudogenes, with an estimated half-life of four million years (Lynch and Conery 2000, 2003), older than most HSDs (Dennis et al. 2017). Therefore, the duplication events that are the focus of this study include genes that are transient, neutral, or near-neutral sequence ultimately destined to be lost in the absence of selective pressure. Our results suggest that changes in the exon-intron structure are common and are among some the earliest events that occur during the birth-death process, likely orthogonal to the action of selection (Hahn 2009). Thus, transcriptional divergence from the ancestral gene appears to be the most common fate, and this occurs soon after or even at the time of the duplication event itself. We hypothesize such rapid changes in the gene structure and transcriptional landscape facilitate the emergence of new function. In the case of humans, a small number of these duplicates appear to be undergoing the first step of a multi-stage process where the duplicates subsequently fix in copy number (Dennis et al. 2017) and maintain an altered ORF ultimately leading to neofunctionalization and subfunctionalization events. Among these, are novel genes thought to be important in neuroadaptive traits critical for the development of the human species (Dennis et al. 2012; Charrier et al. 2012; Florio et al. 2015, 2016; Fiddes et al. 2018; Suzuki et al. 2018).

3.5 METHODS

3.5.1 *Probe design*

Biotinylated oligonucleotide probes (see Appendix B: Table S2 for sequence) were designed preferentially to constitutive exons and coding sequence within the duplicated portion of ancestral genes as well as putative exons where annotation was absent or questionable. Repeat-masked sequence was avoided. Because of the high homology between paralogs, probes designed to exons

of the ancestral gene were presumed to hybridize successfully to the duplicated gene as well. Probes were synthesized on the sense strand resulting in 515 total probes for the HSD1 panel and 271 for the HSD2 panel.

3.5.2 *cDNA synthesis, library preparation, enrichment, and sequencing*

Double-stranded cDNA was synthesized by a modified version of the standard Iso-Seq template preparation (https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Isoform-Sequencing-Iso-Seq-Analysis-using-the-Clontech-SMARTer-PCR-cDNA-Synthesis-Kit-and-SageELF-Size-Selection-System.pdf) protocol that incorporates a barcode/molecular identifier at the end of each strand. This helps facilitate deconvolution of PCR duplicate sequences versus unique founder molecules.

We synthesized specialized poly-dT oligonucleotides to prime first-strand cDNA synthesis (Integrated DNA Technologies [IDT]) with the following configuration:

5'-
AAGCAGTGGTATCAACGCAGAGT(BC16bp)TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTV
N-3',

where the sequence BC16bp encodes one of 96 of the 16 bp barcodes, V=(A,G,C), and N=(any base).

We synthesized oligonucleotides for second-strand synthesis with the following configuration:

5'-AAGCAGTGGTATCAACGCAGAGT(BC16bp)ATACGATTTAGGTGACACTATAGG-3'

where the sequence BC16bp encodes one of 96 of the 16 bp barcodes. The template switch oligonucleotide, a chimeric RNA-DNA sequence, was synthesized: AAGCAGTGGTATCAACGCAGAGTACATrGrGrG

For cDNA amplification, the 5' flanking sequence in the first- and second-strand oligonucleotides was utilized for PCR: /5Phos/AAGCAGTGGTATCAACGCAGAGT.

PolyA RNA (20 ng) from pooled human adult brain (Clontech cat #636102) or developing brain (Clontech cat #636106) was reverse-transcribed in a 10 uL reaction containing 50 mM Tris-HCl (pH 8.3 at 25°C), 75 mM KCl, 3 mM MgCl2, 10 mM DTT, 0.5 mM dNTPs, 100U of Maxima RNaseH- RT (ThermoFisher), 5 mM SP6 template switch oligo, and 10 pmols barcoded oligo-dT primer. For experiments with concordant primers, a single barcoded primer was used for each of 96 parallel reactions. For experiments with discordant primers (a form of molecular indexing), an equimolar mix of all 96 barcodes was used. Reactions were incubated as follows: 45°C for 1 hr, 55°C for 30 min, 45°C for 30 min, 85°C for 5 min.

After the heat kill step, the first-strand cDNA was purified by precipitation on magnetic beads (1X AMPure PB; PacBio). The recovered material was subsequently carried into a 50 uL second-stranding reaction in 1X Takara LA Taq HS buffer (Clontech), 200 uM dNTPs, 2.5U of Takara LA Taq HS (Clontech), and 0.5 uM of barcoded SP6 second-stranding oligo. This oligo binds at the 3' ends of the first-strand cDNA at the SP6 sequence added from the template switch. For experiments with concordant primers, a single barcoded primer was used for each of 96 parallel reactions. For experiments with discordant primers (a form of molecular indexing), an equimolar mix of all 96 barcodes was used. The second-stranding reactions were incubated as follows: 95°C for 1 min, 65°C for 10 min.

The second-stranding reaction was immediately stopped by depletion of primers by Exonuclease I (NEB; 10U) and dNTPs by alkaline phosphatase (rSAP: NEB; 1U) at 37°C for 20 min. The double-stranded cDNA ("founder molecules") were purified by precipitation on magnetic beads (0.5X AMPure PB; PacBio).

The double-stranded cDNA (20% of founder molecule reaction) was amplified by a 100 uL PCR reaction in 1X Takara LA Taq HS buffer (Clontech), 250 mM dNTPs, 5U of Takara LA Taq HS (Clontech), and 0.5 mM of the PCR primer. Reactions were incubated as follows: 95°C for 1 min, 95°C for 30 sec, 68°C for 30 sec, 72°C for 10 min, 72°C for 10 min, with the underlined steps for 12 cycles.

Amplified double-stranded cDNA was purified by precipitation on magnetic beads (0.4-0.6X AMPure PB; PacBio). In some cases, the cDNA was size-fractionated by an automated gel electrophoresis and recovery instrument (SageELF, Sage Sciences). Size fractions were then assayed on a Bioanalyzer High Sensitivity chip and amplified in batches (~1-2 kbp, 2-3 kbp, 3-4 kbp, 4-6 kbp) with the same conditions as the prior PCR. 1-3 kbp fractions were run through 5 cycles, while larger fractions required 8-10 cycles.

Biotinylated probes (~120 nucleotides) were synthesized (IDT) corresponding to known and putative exons of target genes (sense strand). Custom blocker oligonucleotides were synthesized (xGen blockers; IDT) to match the first- and second-strand oligonucleotides with 16 deoxyinosines in place of the barcodes.

Enrichment was carried out on various size fractions (1 ug each) using the hybridization and wash reagents (xGen Lockdown Reagents; IDT) according to manufacturer instructions. (https://www.pacb.com/wp-content/uploads/Unsupported-Protocol-Full-length-cDNA-Target-Sequence-Capture-Using-IDT-xGen-Lockdown-Probes.pdf) The final step involves resuspending the streptavidin beads holding the immobilized enriched sample in PCR conditions (same PCR primer as prior; Kapa HiFi Hot Start polymerase/buffer). PCR was carried out according to xGen Lockdown instructions but with longer extension time of 5 min. Amplification reactions were purified by precipitations on magnetic beads (0.5X AMPure PB, PacBio) and assayed both by

fluorometer (Qubit, ThermoFisher) for dsDNA concentration and Bioanalyzer (DNA12000 chip, Agilent) for size.

Final cDNA was purified by precipitation on magnetic beads (0.5X AMPure PB; PacBio) and single-molecule, real-time (SMRT) sequencing libraries were prepared according to manufacturer guidelines (SMRTbell Template Prep Kit 1.0, PacBio). Final libraries were purified by two sequential precipitations on magnetic beads (2 x 0.5X AMPure PB, PacBio) and assayed both by fluorometer (Qubit, ThermoFisher) for dsDNA concentration and Bioanalyzer (DNA 12000 chip, Agilent) for size. SMRT sequencing was performed using the P6-C4 chemistry on the PacBio RS II instrument with 6-hour movies.

3.5.3 *Gene model determination from long-read RNA-seq data*

A modified version of the Iso-Seq bioinformatics incorporating ToFU (Transcript isOforms: Full-length and Unassembled; Gordon *et al.* 2015 *PLoS One*) was used for processing the long-read RNA-seq data (all of which is available at https://github.com/EichlerLab/isoseq_pipeline). For each sequencing molecule, an intra-molecular CCS read was generated using CCS2 (deviations from default parameters include "–minLength=100 –maxLength=10000 –minPasses=1"). The CCS reads were then classified as pFL if the expected terminal sequences and a polyA tract were observed.

Reads were then mapped to the human reference genome (GRCh38) using GMAP (v 2015-07-23) and mapped pFL reads were used for further analysis. To take advantage of the higher read qualities available through the Iso-Seq "clustering" pipeline (in which multiple reads are used to generate a consensus isoform), but to avoid the potential for confounding by reads from separate paralogs being merged in the same cluster, we performed this step in partitioned genomic regions (as described in Kronenberg et al., 2018). Regions are split to ensure that no pair of SD "mates" is

found in any one region. Only confidently mapped reads (MAPQ > 40) are input into the clustering step, and clustering is performed separately in each region, generating consensus isoforms without contamination from other paralogs (see Appendix B: Figure S6).

We further performed the "collapse" step whereby consensus isoforms are mapped and the mappings are used to remove redundant isoforms. In practice, we found that this generated a greater number of "non-redundant isoforms" for most paralogs than would be expected and that many appeared to be fragments of isoforms, especially for genes with longer transcripts for which we were less likely to have captured full-length isoforms on single reads. Therefore, it was necessary to merge more than one fragment isoform from the final output of this modified Iso-Seq pipeline to generate a full-length gene model.

Finally, newly determined isoforms were assessed for support by other data sources. This includes reads of 5' ends of RNA molecules generated with cap analysis of gene expression (CAGE) data from the FANTOM5 consortium (Lizio et al. 2015), 3' ends of polyadenylated RNA molecules (polyA-seq) from Leslie, Mayr, and colleagues (Lianoglou et al. 2013) remapped to GRCh38, and various PacBio RNA-seq datasets, including: H1 human embryonic stem cell (GEO accession GSM1254204), Iso-Seq whole transcriptome from human brain with Alzheimer's disease (https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/ accessed Dec 2016), whole transcriptome from human brain, liver, and heart (http://datasets.pacb.com.s3.amazonaws.com/2014/Iso-seq_Human_Tissues/list.html accessed Dec 2016), and Iso-Seq whole transcriptome generated from MCF7 human breast cancer cell line (https://github.com/PacificBiosciences/DevNet/wiki/IsoSeq-Human-MCF7-Transcriptome accessed Dec 2016). ANGEL (https://github.com/PacificBiosciences/ANGEL) was used to

identify ORFs. Genomic adenine homopolymers that could lead to spurious oligo-dT priming were identified to avoid incorrect 3' end annotation.

3.5.4 *Secondary analysis of long-read RNA-seq data*

Proportion of fusion (duplication-spanning) versus truncation reads: For each duplicate gene investigated, a constitutive exon close to the breakpoint was chosen as an anchor point, and mapped pFL reads (MAPQ > 40) including that exon were selected in order to mitigate non-full-length reads from inflating the "truncated" count. Counts of reads that contain spliced exons beyond the duplication breakpoint were calculated using BEDTools (Quinlan and Hall 2010). Based on this proportion, a duplicate gene was designated as "primarily truncation" (<0.2), "primarily fusion" (>0.8), or "both" (0.2–0.8).

Splicing disorder: The number of supporting pFL reads for *FCGR1A* and *FCGR1B* isoforms output by the "collapse" step of the modified Iso-Seq pipeline (minimum 2 pFL reads) were used as an approximation of relative isoform abundance. Shannon's Entropy was calculated using the isoform abundances for each paralog with the entropy package (v1.2.1) (Hausser and Strimmer 2008 https://arxiv.org/abs/0811.3579).

3.5.5 *Illumina RNA-seq*

~5 ng of polyA$^+$ RNA pooled adult human brain (Clontech cat #636102) or developing brain (Clontech cat #636106) was used as input for the TruSeq Stranded mRNA-seq kit (Illumina) with parameters set for ~150 bp insert-size libraries. Final purified libraries were assayed both by fluorometer (Qubit, ThermoFisher) for dsDNA concentration and Bioanalyzer (DNA12000 chip, Agilent) for size. Sequencing-by-synthesis (SBS) was performed on a HiSeq 2500 with 2x125 bp reads. Reads were demultiplexed using deML (Renaud et al. 2015) yielding 61 M and 72 M reads

for developing and adult brain, respectively, and trimmed of adapter and low-quality sequence using Trimmomatic (Bolger et al. 2014) following quality control (QC) by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed 12 July 2016). Reads were mapped to GRCh38 using STAR (Dobin et al. 2013) and further QC was performed using QoRTs (Hartley and Mullikin 2015). Counts of uniquely mapping and multi-mapping reads were taken from the output of STAR mapping ("SJ.out.tab" file).

3.5.6 *Tissue-specific expression estimates*

RNA-seq data from the Genotype-Tissue Expression GTEx project (dbGaP version phs000424.v3.p1) was used to generate tissue-specific expression estimates for gene models with Kallisto (version 0.42.4) (Bray et al. 2016). New gene models were added to a fasta file of the reference transcriptome (GENCODE v25 "Transcript sequences" file ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25/gencode.v25.transcripts.fa.gz). Redundant reference transcriptome sequences were removed (e.g., current fragments of *CD8BP* were removed after our putatively corrected gene models were added). This custom transcriptome was indexed and the Kallisto quantification algorithm was run using default parameters on each of the GTEx samples. Results in the form of transcripts per million were analyzed in R with the aid of dplyr (https://github.com/tidyverse/dplyr) and plotted using ggplot2 (https://github.com/tidyverse/ggplot2) and the ComplexHeatmap package (Gu et al. 2016). Pearson correlation coefficients for duplicate-ancestral gene pairs were generated in R based on median expression levels in each tissue, for tissues that had at least five samples.

3.5.7 *Tests for purifying selection*

We tested the hypothesis that the Tigger7-derived coding sequence was under purifying selection (dN/dS < 1) using CODEML (Yang 2007) by comparing two evolutionary models, one in which dN/dS (omega) is fixed at 1, and one in which it is a free parameter. We used a Chi-squared test (1 d.f.) with twice the difference in log-likelihood as the test statistic, with a significance threshold of $p < 0.05$, to test if the higher parameter model was a statistically significantly better fit.

3.5.8 *Tissue samples and in situ hybridization*

De-identified primary cortical tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations approved by the Human Gamete, Embryo and Stem Cell Research Committee (Institutional Review Board) at the University of California, San Francisco. Tissue specimens were fixed overnight in 4% paraformaldehyde, dehydrated in 30% (w/v) sucrose, and embedded optimal cutting temperature solution (Tissue-Tek). Frozen tissue blocks were sectioned at 20 μm thickness using a Leica freezing microtome.

Probes used for RNA *in situ hybridization* were synthesized within a pCI-NEO vector flanked by XhoI-NotI cloning sites to the antisense strand of target mRNAs (Promega): probe ARHGAP11AB-e1e2e3e4 ("AB") was designed to bases 146 to 805 of NM_001039841.1; probe ARHGAP11A-e12-utr ("A") to bases 3286 to 4005 of NM_014783.4; probe ARHGAP11B-07utr ("B") to bases 2983 to 3688 of NR_038253.1. Digoxigenin labeled RNA probes for *in situ* hybridization were generated by *in vitro* transcription using T7 RNA Polymerase (Roche) in the presence of DIG-RNA Labeling Mix (Roche). In situ hybridization was performed according to a previously described protocol and NBT/BCIP was used to develop alkaline phosphatase conjugated to the sheep antibody against DIG (Roche) (Wallace and Raff 1999). Images were

collected with a Leica DMI 4000B microscope using a Leica DFC295 camera. Images were uniformly adjusted for brightness and contrast for clarity.

3.6 NOTES

Data availability

All RNA-seq data will be deposited at the Sequence Read Archive (SRA). The custom Iso-Seq pipeline used can be found at https://github.com/EichlerLab/isoseq_pipeline. Secondary processing of sequencing data was performed using custom scripts written in Python and R, which are available upon request.

Acknowledgements

We thank Z. Kronenberg for computational support and T. Brown for assistance editing the manuscript. We thank M.Y. Dennis for helpful comments on the manuscript. This work was supported by a training award from the National Human Genome Research Institute (F30HG009478 to M.L.D.) and, in part, by US National Institutes of Health (NIH) grant HG002385 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Chapter 4. FUTURE DIRECTIONS IN DETERMINING HSD GENE FUNCTION

4.1 INTRODUCTION

In this chapter I will discuss what I believe are the next frontiers in the study of human-specific duplicate genes. Ultimately, the work described in Chapters 2 and 3 are intermediate steps along the way to determining gene function. Determining gene models, reconstructing the evolutionary history of genes, and correcting annotations all fit into the larger goal of determining which paralogs are functional, what functions they have, and what the impact is of previously inaccessible naturally occurring and pathogenic variation.

These questions are inherently complicated by the fact that we are interested in highly identical genes, and most every aspect of their study requires special experimental and analysis techniques to account for this. I will discuss three avenues of ongoing investigation that I see as the next steps in which this work is to be taken.

First, I will describe an application of the method described in Chapter 3 to core duplicon genes. Core duplicons are interspersed, highly duplicated segments of the genome that are mediators of continued rearrangements throughout the great ape lineages. Many are genic, and one in particular, a gene family known as *morpheus*, also called *NPIP*, is of particular interest for being under positive selection in the great ape lineage (Johnson et al., 2001). Application of the method described in Chapter 3 to this gene family reveals which of the ~20 paralogs are transcribed and which are silent.

Second, I will describe a new approach for loss-of-function genotyping of HSD genes for the purpose of determining functional status as well as disease association. I will discuss how previous genotyping approaches based on short-read methods failed in the majority of instances to determine paralog identity (Dennis et al., 2017), but how our long-read approach allows for such assignment, and how this improvement in variant interpretation improves our understanding of the functional roles of HSD paralogs.

Third, I will discuss how single-cell RNA-seq techniques can provide more specific insights into the function of HSD genes, in particular in the developing brain. In the search for genes underlying human-specific aspects of brain development, multiple groups have converged on HSD genes as the result of experiments that have focused on the neural stem cells whose divisions are thought to determine the size of the cortex (Florio et al., 2018; Bhaduri, 2018 *submitted*). I will show data that suggest genes specifically upregulated in human in these cells are enriched for HSDs and recent duplicate genes in general.

The work presented in this chapter represents early stages of avenues of investigation that I hope will be expanded upon in the future. Section 1 includes work I performed with Jason G. Underwood; Section 2 includes work I performed with Tianyun Wang and Alexandra Lewis; Section 3 includes work performed in collaboration with Aparna Bhaduri and Alex A. Pollen.

4.2 Transcriptional Study of the *MORPHEUS* Gene Family

While HSD genes are generally found at relatively low (2-5) copy number, there are recently duplicated gene families at much higher copy in the human genome as well (Sudmant, 2010). These gene families have grown throughout the primate lineage and thus are not human-specific. Their expansions follow a particular pattern, where an inner ("core") sequence that appears to drive the duplication is shared but the flanking sequence may differ, and accordingly have been labeled

"core duplicons" (Jiang et al., 2007). These core duplicons are frequently found at the breakpoints of structural variants and appear to be a major predisposing force to genomic rearrangement (Johnson et al., 2006).

One such core duplicon found throughout chromosome 16 contains a gene family known as *morpheus*, also called nuclear-pore-interacting protein (*NPIP*), which has expanded to over 20 copies in human and chimpanzee and shows a strong signal of positive selection (Johnson et al., 2001). In particular, two exons of the gene (exon 2 and exon 4) show a much higher degree of amino acid divergence between paralogs than the surrounding intronic sequence. Little is known about the functional role of these genes, though there is some weak evidence of immune-related function (Bekpen et al., 2017).

We sought to determine if the method described in Chapter 3 could be applied to such high-copy core duplicon gene families as *NPIP*. We designed 74 oligonucleotide probes targeting both subfamilies of paralogs (known as *NPIPA* and *NPIPB*) and used these probes for an enrichment experiment as described in Chapter 3. We used cDNA from the complete hydatidiform mole (CHM1), which has a haploid genome and thus is devoid of allelic diversity (Fan, 2002). We sequenced one PacBio SMRT cell to determine the efficacy of our enrichment strategy.

By mapping the long-read RNA-seq data, we see the approximate distribution of captured reads (**Figure 4.1**). Here, a density plot of mapped reads is displayed, with the probe-enriched data on top, above the mappings of the probes in gray, and below are two RT-PCR sequencing experiments also using RNA from CHM1 as starting material (each is a single PacBio SMRT cell) for comparison. We sequenced a total of 19,125 reads, with an on-target rate of 54%. It is apparent from comparing the density and distribution of reads from the probe-enriched experiment that a better coverage across *NPIP* paralogs on chromosome 16 was achieved.

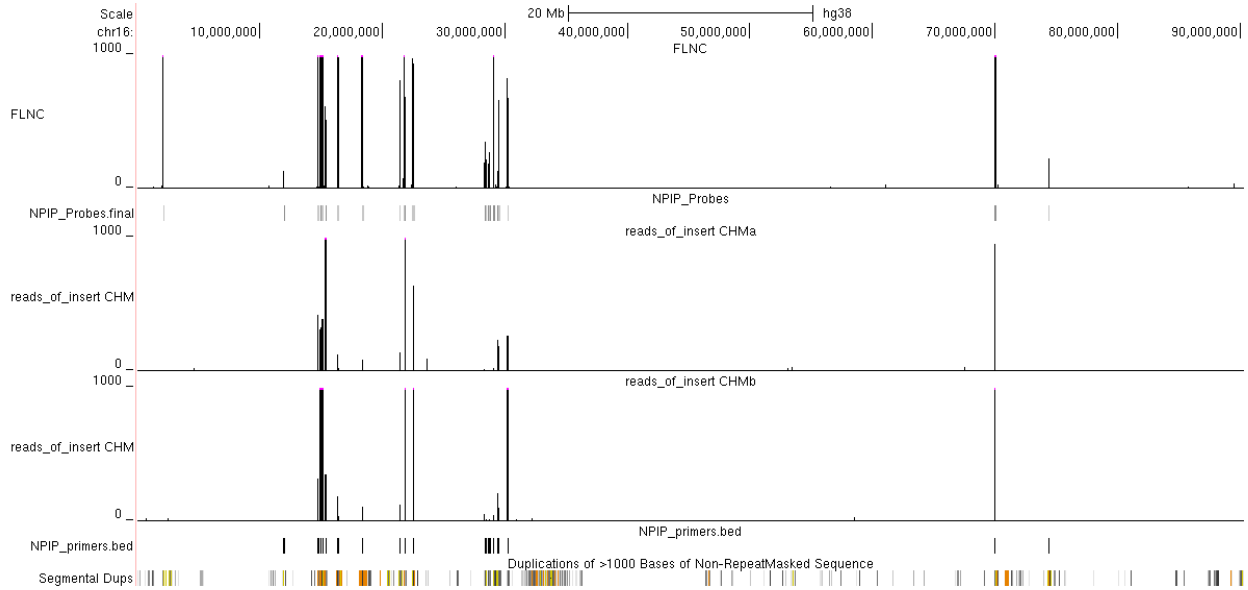


Figure 4.1. *NPIP* coverage across chromosome 16

Genome browser screenshot spanning the entirety of chromosome 16 shows improved coverage across *NPIP* paralogs with probe-based capture (top) versus RT-PCR experiments (middle and bottom) targeting similar regions. For reference, locations of designed *NPIP* probes as well as RT-PCR primers are also show, as well as the segmental duplication track.

We then sought to quantify expression by counting mapped reads at each paralog (**Table 4.1**). We first counted confidently mapped reads (those with a mapping quality score >40). These numbers tell us which paralogs are expressed with certainty; however, the sum of these reads (4,102) indicates that many reads are being excluded from these counts due to ambiguous mapping. Thus, we also generate counts that include multiple mappings of reads. In this scenario, a read may be counted as belonging to more than one paralog. These two counting strategies yield well-correlated results (Pearson correlation coefficient = 0.80) and point to *NPIPA1*, *NPIPA6*, *NPIPA9*, and *NPIPP1* as the most highly expressed paralogs.

Table 4.1. Long-read RNA-seq read counts of *NPIP* paralogs.

Gene	Counts (uniquely mapping)	Counts (multi-mapping)
NPIPA1	733	5021
NPIPA2	23	332
NPIPA3	17	764
NPIPA5	15	901
NPIPA6	731	5332
NPIPA7	3	1182
NPIPA8	0	1160
NPIPA9	673	5292
NPIPB10	9	117
NPIPB11	34	236
NPIPB12	54	1049
NPIPB13	53	1136
NPIPB14	292	423
NPIPB15	0	236
NPIPB1	0	275
NPIPB2	0	211
NPIPB3	71	907
NPIPB4	158	936
NPIPB5	110	1196
NPIPB6	58	499
NPIPB7	56	293
NPIPB8	5	320
NPIPB9	80	477
NPIPP1	927	1546
Total	4102	29841

Reads from the pilot enrichment experiment were mapped to GRCh38 using GMAP, and the number of reads belonging to each paralog was calculated in two ways. In the first, only high-confident, uniquely mapping reads were counted. In the second, multi-mapping was allowed.

This result shows that the probe-based capture methods described in Chapter 3 can be applied to high-copy core duplicon gene families such as *NPIP*. We present for the first time an estimate of the relative abundance of specific *NPIP* paralogs across the entire gene family. Using a minimum cutoff of 5 reads, we can detect with confidence transcription at 19 out of 24 paralogs in CHM1. However, this may be an undercount, as this count excludes ambiguously mapping reads. For example, predicted transcripts for paralogs *NPIPA7* and *NPIPA8* are 100% identical, and approximately 1,200 reads map to both copies equally.

Work is underway to develop a metric to define confidence in paralog assignment, and thus to have a more precise estimate of paralog abundance from the data. We also are expanding our scope to other core duplicon gene families (Dennis et al., 2017; Jiang et al., 2007). With further refinement of these methods, I believe it will be possible to thoroughly characterize the transcription of core duplicon gene families as we have done for HSD genes.

4.3 Long-Read-Based Loss-of-Function Genotyping for Determining the Functional Status of Duplicate Paralogs

The most common fate for a recently duplicated gene is pseudogenization and eventual loss (Lynch & Conery, 2000). Therefore, one of the challenges inherent to studying the most recently duplicated genes is that selection has not yet had time to purge nonfunctional copies from the genome, as the average half-life of a duplicate gene is believed to be about 4 million years (Lynch & Conery, 2000), which is on par with the age of HSD genes (Dennis et al., 2017). We cannot take for granted that the presence of a duplicate gene, its fixation, or its ordered transcriptional activity is an indicator of functional status.

However, understanding which paralogs are functional and which are not is crucial to deciphering the roles of HSD genes in evolution and disease. For example, a frameshift variant in one paralog may be pathogenic, but the same variant in another nonfunctional paralog would be expected to have no effect at all. This can lead to confusion over the pathogenicity of variants. For instance, a loss-of-function mutation (c.579 G>A) in the HSD gene family *NCF1*, which has three copies, is causative of chronic granulomatous disease; however, this mutation appeared to be at high frequency in the Ashkenazi Jewish population (De Boer et al., 2018; Wolach et al., 2018). These seemingly contradictory findings turn out to result from the fact that pathogenic variants reside on the ancestral copy, but the harmless high-frequency variant is found on a duplicate,

nonfunctional copy. However, the standard genotyping methods failed to distinguish these two states. Thus, when genotyping recently duplicated genes, it is easy for the confounding of multiple paralogs to affect the interpretation of results.

As part of a recent study of HSD genes (Dennis et al., 2017), we performed short-read-based genotyping across 30 HSD gene families in 658 individuals from the 1000 Genomes Project (1KG) using molecular inversion probes (MIPs) (Hiatt, 2013). Variants were called by mapping reads exclusively to the ancestral paralog to take advantage of the generally more complete gene annotation. This yielded 96 loss-of-function (LoF) variants, though only 33 (34%) could be definitively assigned to a paralog, similar to the rate we found assignable when performing the same manner of genotyping in *HYDIN2* (see Chapter 2). High-frequency LoF variants in diverse population controls have the potential to inform us of which paralogs are tolerant of such mutations, while a dearth of such mutations informs us of which paralogs are potentially functional. This is analogous to querying a candidate disease gene in a large-scale genotyping database to assess its tolerance to LoF mutation (Lek et al., 2016; Samocha et al., 2014).

The remaining 63 LoF variants in HSD genes that could be assigned to a gene family, but not a specific paralog, hold potentially valuable information about which paralogs are functional and which are not, but that information is inaccessible in the current data. In particular, there are 10 such variants (**Table 4.2**) present at over 2% frequency in control populations. The lack of paralog assignability derives from the fact that 112 bp MIP-based sequences used to genotype were not long enough to reliably capture paralogous sequence variants (PSVs), sequence differences that distinguish duplicate paralogs. However, a longer flanking region around the LoF variant would have a better chance of capturing PSVs. So, we sought to re-genotype these variants in a subset of individuals known to harbor them, but this time by generating a 2 kbp amplicon. We then subjected

these 2 kbp amplicons to long-read (PacBio) sequencing so that we could accurately "phase" the variants and determine to which paralog they belonged.

Table 4.2. Common unassigned LoF variants from MIP-based sequencing.

Variant Designation	Coordinates	Reference	Consequence	Gene family	Pop frequency	Paralog
ARHGEF_var1	chr7:144059763-144059786	GRCh37	stop_gained	*ARHGEF5*	0.534	Unknown
ARHGEF_var2	chr7:144062632-144062632	GRCh37	stop_gained	*ARHGEF5*	0.452	Unknown
ARHGEF_var3	chr7:144068253-144068253	GRCh37	splice_acceptor	*ARHGEF5*	0.293	Unknown
FRMPD2_var1	chr10:49379223-49379230	GRCh37	frameshift	*FRMPD2*	0.023	Unknown
GTF2H2C_var1	chr5:68863605-68863605	GRCh37	splice_donor	*GTF2H2C*	0.023	Unknown
GTF2H2C_var2	chr7:74211003-74211003	GRCh37	stop_lost	*GTF2IRD2*	0.942	Unknown
NAIP_var1	chr5:70270055-70270055	GRCh37	stop_gained	*NAIP*	0.021	Unknown
NAIP_var2	chr5:70281265-70281267	GRCh37	frameshift	*NAIP*	0.167	Unknown
PTPN20B_var1	chr10:48751834-48751836	GRCh37	frameshift	*PTPN20B*	0.024	Unknown
TCAF2_var1	chr7:143417403-143417410	GRCh37	frameshift	*TCAF2*	0.147	Unknown

10 LoF variants with allele frequencies in control populations over 2% described in Dennis et al., 2017. Common LoF variation in control populations indicates a tolerance to mutation and can help distinguish functional from nonfunctional duplicate paralogs.

Primers were designed to degenerate sequence such that all paralogs would be amplified by PCR. Amplicons were designed to be ~2 kbp in length, and it was ensured that at least two PSVs were located in the amplified region to ensure paralog distinguishability. For each unspecified variant, 10 individuals who were positive for that variant from the MIP-based sequencing were selected and their genomic DNA was used for amplification. PCR products were purified and pooled at approximately equimolar concentrations, with a barcoding strategy to distinguish individuals. All were sequenced in a single PacBio SMRT cell. Reads were demultiplexed and mapped to the reference genome using BLASR (Chaisson et al., 2012). Each paralog was searched for having the heterozygous variant specified, and paralog assignment was made by manual assessment using the UCSC Genome Browser (**Table 4.3**).

Table 4.3. Long-read-based genotyping results of unassigned HSD LoF variants.

		Individual										
Gene family	**Paralog**	1	2	3	4	5	6	7	8	9	10	**Result**
ARHGEF5	*ARHGEF5*	+	-	+	+	-	+	+	+	+	+	solved - ARHGEF5
	ARHGEF34P	-	i.d.	-	-	-	-	-	i.d.	-	-	
	ARHGEF35	-	-	-	-	-	-	-	-	-	-	
ARHGEF5	*ARHGEF5*	-	-	-	+	-	+	-	-	+	-	solved - ARHGEF34P
	ARHGEF34P	i.d.	-	+	+	+	+	-	+	+	i.d.	
	ARHGEF35	paralog does not contain variant										
ARHGEF5	*ARHGEF5*	-	-	+	-	-	-	i.d.	-	-	-	solved - ARHGEF34P
	ARHGEF34P	+	+	+	+	+	+	+	+	+	+	
	ARHGEF35	paralog does not contain variant										
GTF2H2C	*GTF2H2C*	-	-	-	-	-	-	i.d.	-	+	-	solved - GTF2H2
	GTF2H2B	i.d.	i.d.	i.d.	-	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	
	GTF2H2	-	-	+	+	+	+	i.d.	-	+	+	
GTF2IRD2	*GTF2IRD2*	-	-	-	-	-	-	-	-	-	-	unsolved
	GTF2IRD2B	+?	-	-	-	i.d.	-	-	-	-	+?	
	GTF2IRD2P1	-	-	-	-	-	-	-	-	-	-	
NAIP	*NAIP*	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	i.d.	solved - NAIP_un4
	NAIP_un	paralog does not contain variant										
	NAIP_un2	-	-	-	-	-	-	-	-	-	i.d.	
	NAIP_un4	+	-	+	+	+	+	+	+	+	i.d.	
PTPN20	*PTPN20*	-	-	-	-	-	-	-	-	-	-	solved - PTPN20B
	PTPN20B	-	-	+	+	+	+	+	+	+	-	
TCAF2	*TCAF2*	-	-	-	-	i.d.	i.d.	-	+?	-	-	solved - TCAF2A
	TCAF1P1	paralog does not contain variant										
	TCAF2A_un	-	+	i.d.	+	i.d.	i.d.	+	+	-	+	
	TCAF2A_un2	-	+	i.d.	+	i.d.	i.d.	+	+	-	+	
	TCAF2P1	paralog does not contain variant										
	RP11-61L23.2	-	+?	-	-	-	-	-	+	-	-	
	TCAF1	paralog does not contain variant										
NAIP	*NAIP*	-	-	+	-	i.d.	+	-	+	n.d.	n.d.	unsolved
	NAIP_un	paralog does not contain variant								n.d.	n.d.	
	NAIP_un2	i.d.	-	+	-	i.d.	+	-	+	n.d.	n.d.	
	NAIP_un4	paralog does not contain variant								n.d.	n.d.	
FRMPD2	*FRMPD2*	-?	Y	Y	-?	-?	-?	-?	-?	n.d.	n.d.	unsolved
	FRMPD2B	-?	-?	-?	-?	-?	-?	-?	-?	n.d.	n.d.	

Out of 10 variants examined in 10 individuals (with the exception of two variants sampled in eight individuals due to the constraints of the 96-well format), seven could be assigned with at least mild confidence to a particular paralog. Show for each variant is the result of manually assessing the presence (+) or absence (-) of the variant in each paralog. Question marks indicate a low-certainty call, while i.d. indicates insufficient data (< 10 reads) and n.d. indicates the variant was not queried in the individual. Paralog names with the "un" suffix refer to unannotated regions of homology to the ancestral gene.

Out of the 10 variants, seven could be assigned to a particular paralog, while three could not be assigned through manual inspection of read mappings alone. We find that the LoF variant belongs often but not always to what is believed to be a pseudogene paralog. Two of the three *ARHGEF5* gene family variants were assigned to *ARHGEF34P*, though one is assigned to *ARHGEF5*, the ancestral gene. For the *NAIP* gene family the variant belongs to a duplicate copy

that is not part of any annotated gene, while for *PTPN20*, the variant is strongly indicated to be in the duplicate pseudogene, *PTPN20CP*. However, the precise details of paralog-specific gene annotations are important. For example, the *TCAF* gene family frameshift variant is assigned to *TCAF2A*; however, at this locus it does not overlap an exon, as the gene model is distinct. Thus, we would not conclude that *TCAF2A* is tolerant to LoF variation. Finally, the *GTF2H2* gene family variant has been assigned to *GTF2H2* itself, not *GTF2H2B*, which appears to be a pseudogenized duplication. This may be permitted due to the redundancy created by *GTF2H2C*, which maintains the open-reading frame of *GTF2H2*. It is perhaps surprising that a common (53% allele frequency) LoF variant has been assigned to *ARHGEF5*. While this variant directly disrupts the ATG start codon, it may be tolerated due to alternate translation at the next ATG codon, which is found 81 bp downstream.

To summarize this pilot experiment on assigning LoF variants in HSD gene families to specific paralogs, we used a long-read genotyping approach, generating 2 kbp sequences (instead of 112 bp) and were able to assign 7 out of 10 of the variants to specific paralogs. Current work is underway to develop an automated approach to paralog assignment employing an algorithm originally designed for phasing of long reads for *de novo* genome assembly over duplicated sequence (Chaisson et al. 2017).

Looking ahead, the same probe-based enrichment can be applied concurrently to genomic DNA containing genes of interest. *De novo* assembly of a suspected disease region using probe-based enrichment followed by long-read sequencing enabled the identification of a previously undetected structural variant causing X-linked dystonia-parkinsonism (Aneichyk et al. 2018). With the complementary abilities to assemble duplicated regions and capture full-length transcripts,

previously inaccessible variation in duplicate genes that manifests in altered transcription can be assessed.

4.4 Single-Cell RNA-Sequencing Implicates HSD Genes in Neural Progenitor Cell Function

The most salient distinguishing feature of the human species is the behavioral and cultural differences attributed to increased intelligence, and the expansion of the cortex—in particular the neocortex—is believed to be the basis for higher-order cognition and complex behavior (Sousa et al., 2017). There is no simple relationship between brain size and cognitive abilities between species. However, humans have the largest brains among extant primates (Stephan et al., 1981). Humans are also believed to have the most cortical neurons among extant primates (Herculano-Houzel et al., 2015). Therefore, great interest surrounds the genetic changes responsible for these observed differences.

The neural stem cells whose proliferation gives rise to the cortex are a natural focal point of attention for the question of what expression-related changes may relate to the difference in brain size and neuron count. Single-cell RNA-sequencing (scRNA-seq) allows for the determination of cell-type identity and the examination of cell-type-specific expression patterns. The brain is a heterogeneous tissue, and as such, differences in gene expression of one particular cell type may be lost in bulk tissue expression measurements. The neural stems cells of the neocortex are known as radial glia, for their spoke-like morphology as observed in a cross section of the developing neocortex. Of particular importance are the outer radial glia, which have a distinct cellular identity and generate the majority of cortical neurons (Pollen et al., 2015).

Developments in single-cell techniques have improved our ability to focus on key cell types, such as radial glia, and have yielded new insights into the specific expression differences that set

apart both this cell type from others as well as this cell type in humans from this cell type in nonhuman primates. Many techniques have been applied to isolate cells of interest. These include laser capture microdissection of specific cortical zones (Fietz et al., 2012; Miller et al., 2014), cell types isolated by fluorescence-activated cell sorting (Florio et al., 2015a; M. B. Johnson et al., 2015), and microdissection followed by scRNA-seq (Pollen et al., 2015). A recent meta-analysis of the aforementioned five datasets that sought to identify genes preferentially expressed in progenitor cells of the human neocortex and found that 15 human-specific genes fit this expression pattern (Florio et al., 2018).

Therefore, recent species- and clade-specific genes appear to play an important role in expression that distinguishes the neural progenitor cells from more mature neural cell types. We were also interested in expression differences of neural progenitor cells in the developing human cortex as compared to those in developing nonhuman primate cortex. However, comparative studies of human and nonhuman primate brain development suffer from a lack of primary developing brain tissue, especially that of nonhuman primates like chimpanzee. Organoid models derived from induced pluripotent stem cells (iPSCs) (Clevers, 2016) offer an alternative approach. Cerebral organoids have great potential for application to scRNA-seq and recapitulate many aspects of early brain development (Pollen et al., 2015).

Our collaborators Aparna Bhaduri and Alex A. Pollen at the University of San Francisco have recently undertaken an effort to identify such differences (Bhaduri, 2018 *submitted*). They have developed a cerebral organoid model derived from iPSCs and compared expression between organoids derived from 10 human and 8 chimpanzee individuals. Specifically, scRNA-seq performed on dissociated cells from such cerebral organoids was used to define orthologous cell types and compare expression between the species. In parallel, they performed scRNA-seq on

dissociated cells from primary tissue from 48 human and 6 macaque primary samples of developing cortex. Using a likelihood ratio test, they identified 200 differentially expressed genes that showed increased expression in human in both the human–chimpanzee and human–macaque comparison.

Inspection of this list revealed that nine HSD genes were present. We tested whether the derived gain-of-expression genes in human were enriched for HSD genes (Dennis et al., 2017) using a one-sided Fisher's exact test as well as by permuting identity (HSD or not) and recording how often eight or more HSDs were up-regulated by chance. Note that a gain-of-expression gene (*DDAH1*), previously designated as an HSD gene (Dennis et al., 2017), was not counted for the purposes of this analysis because the duplication is intronic, but conservatively the list of all HSDs was not changed.

Because the classification of HSD genes is conservative and does not include genes that have expanded earlier in the primate lineage or independently in both humans and other primates, we also tested whether the same gene list was enriched for genes overlapping recent (>99% identity) segmental duplications. Note that this should also include many of the core duplicon genes (Jiang et al., 2007). We used a one-sided Fisher's exact test to test for enrichment, as well as 10,000 permutations in which segmental duplications were shuffled across the genome and the proportion of overlapping genes that were gain-of-expression was recorded; in no permutation was this proportion greater than observed.

And so, we found that gain-of-expression genes in human are enriched for HSD genes ($p = 1.3 \times 10^{-5}$ by Fisher's exact test; $p = 1.6 \times 10^{-5}$ by permutation test) as well as genes overlapping recent segmental duplications (>99% identity) in human ($p = 7.0 \times 10^{-11}$ by Fisher's exact test; $p < 10^{-5}$ by permutation test), supporting a role of young duplicate genes in human-specific aspects of

cortical development. Such genes are prime candidates for comparative functional studies to determine what specifically their contributions are to development.

4.5 Concluding Remarks

This work has included careful examination of gene duplication as a mechanism for the evolution of novelty in human. While ultra-expanded gene families such as *NPIP* consist of whole-gene duplications, one striking finding for HSD genes is just how many (19 out of 31) duplicate paralogs we characterized are incomplete gene duplications. As a model for rapid evolution of novelty, incomplete gene duplication is inconsistent with the classic model of Ohno whereby duplicate copies accumulate mutations gradually over time (Ohno 1970) as well as more recent findings that tandem duplicate genes in mammals are slow to diverge and rarely subfunctionalize (Lan and Pritchard 2016).

The findings presented here are distinguished from such earlier work on gene duplication by 1) the timescale of our analysis—we focus on the most recently duplicate genes now that we have methods that make their transcription accessible and 2) the interspersed nature of the duplicate genes studied here. Not one of the duplicate genes in **Figure 3.3** is a tandem gene duplication. This is consistent with the much higher proportion of interspersed (versus tandem) duplications found in great ape genomes (Marques-Bonet et al., 2009). The frequency of fusion transcripts that arise from such interspersed duplications speaks to both the transcriptional promiscuity of recently duplicated segments as well as modularity of genomic segments when placed into new contexts.

It is likely no coincidence that the most promising candidates among HSDs for roles in shaping human neurodevelopment are 3'-truncated genes (Dennis et al., 2012; Charrier et al., 2012; Florio et al., 2015; Fiddes et al., 2018; Suzuki et al., 2018). Maintenance of the promoter supports maintenance of the ancestral expression pattern, allowing interaction of the new duplicate gene

with its ancestor or other ancestral interacting partners, and a truncated protein copy immediately presents an opportunity for neofunctionalization. The role of the frequent fusion transcripts seen from "run-on" transcription remains to be elucidated.

While some genes discussed have the support of experimental evidence supporting function (Charrier et al., 2012; Florio et al., 2015; Ju et al., 2016; Araud et al. 2011), for most, such evidence is still lacking. Discussion of duplicate genes has long included the notion of the "potential" for future, beneficial mutations (Ohno, 1970), which runs the risk of anthropomorphizing the genome somewhat, so as to see "intent", where of course, the genome has no intent, nor foresight (Doolittle and Brunet, 2017; Brenner, 1998). Is it meaningful to discuss evolutionary innovation in such cases, or are we reading too much into genomic spasms that are currently tolerated but will eventually be lost for lack of selective pressures?

It is worth noting that in two of the best functionally studied HSD genes, there were intermediate stages where the duplicate gene (in both cases 3'-truncated) may have persisted with absent or at least weaker functionality than what is believed to be the active form today. For instance, in a separate event following its birth through duplication, *ARHGAP11B* underwent a single point mutation that appears to have conferred the key ability to promote neural progenitor activity (Florio et al. 2016). *SRGAP2C* has undergone five amino acid substitutions since its birth by gene duplication without which its key ability to antagonize the ancestral *SRGAP2* through heterodimerization is much weaker (Sporny et al., 2017). While it is no surprise that subsequent mutations should refine the selectively advantageous activities of new genes, it is worth pointing out that in both of these instances, there appears to have been an intermediate state where the new duplicate gene function was not yet fully realized.

Therefore, though we must be cautious to avoid the inference of "intent", and while not all recently duplicate genes may be functional, it would seem that they provide fodder for further mutational changes that may make them so. But then, how do we go on to make that critical assignment of functionality? For even experiments based on ectopic expression of HSD genes run the risk missing the subtleties of context, including the spatiotemporal details of expression patterns, and the contemporary features of the early genomes where these changes initially took place.

I believe the strongest evidence for HSD gene function will come from studies of naturally occurring human variation. Genes (paralogs specifically) that satisfy the following conditions can be said to have strong evidence of function: 1) expression of specific paralogs in relevant tissue and cell types; 2) a depletion of naturally occurring loss-of-function variation in healthy population controls versus what would be expected by chance; and 3) an association between deleterious variation in HSD genes and disease.

To close, I see this work as a required intermediate step towards investigating these three conditions for HSD genes. As discussed in Chapters 2 and 3, correct and complete gene annotation is essential for interpreting sequence variation and producing accurate expression estimates. Looking forward, I see scRNA-seq as a key tool for further refining the expression of HSD genes to the most relevant cell types, especially in highly heterogeneous brain tissue. In Chapter 3 we distinguished expressed from non-expressed paralogs, and in Chapter 4, I presented a way forward for addressing this question with a gene family with ~20 paralogs. I would expect that variation in some of these paralogs is impactful, perhaps even pathogenic, but certainly not all are the same. In Chapter 4, I also presented a way forward for genotyping LoF variation in HSD genes in a paralog-specific manner. It is important to note that copy number difference in HSD genes may be

an even more relevant form of variation than single and multi-nucleotide variants. Careful assessment of all forms of variation in HSD paralogs will ultimately tell us the most about what role these new genes play in making us who we are.

BIBLIOGRAPHY

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61–65.

Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, Collins RL, Domingo A, Currall B, Stortchevoi A, et al. 2018. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**: 897–909.e21.

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M, et al. 2014. Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* **46**: 1293–1302.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**: 745–750.

Araud T, Graw S, Berger R, Lee M, Neveu E, Bertrand D, Leonard S. 2011. The chimeric gene CHRFAM7A, a partial duplication of the CHRNA7 gene, is a dominant negative regulator of α7*nAChR function. *Biochem Pharmacol* **82**: 904–914.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**: 552–564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587–1593.

Bekpen C, Baker C, Hebert MD, Bahar Sahin H, Johnson ME, Celik A, Mullikin JC, Program NCS, Eichler EE. 2017. Functional Characterization of the Morpheus Gene Family. http://dx.doi.org/10.1101/116087.

Bernier R, Steinman KJ, Reilly B, Wallace AS, Sherr EH, Pojman N, Mefford HC, Gerdts J, Earl R, Hanson E, et al. 2016. Clinical phenotype of the recurrent 1q21.1 copy-number variant. *Genet Med* **18**: 341–349.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics* **3**: 201–212.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Boyd JL, Skove SL, Rouanet JP, Pilaz L-J, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr Biol* **25**: 772–779.

Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. 2014. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**: 2670–2672.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Brenner S. 1998. Refuge of spandrels. *Curr Biol* **8**: R669.

Bridges CB. 1936. The bar "gene" a duplication. *Science* **83**: 210–211.

Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* **40**: 1466–1471.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**: 25–36.

Casey JP, Magalhaes T, Conroy JM, Regan R, Shah N, Anney R, Shields DC, Abrahams BS, Almeida J, Bacchelli E, et al. 2012. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet* **131**: 565–579.

Chaisson MJ, Mukherjee S, Kannan S, Eichler EE. 2017. Resolving multicopy duplications using polyploid phasing. *Res Comput Mol Biol* **10229**: 117–133.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923–935.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14**: 645–660.

Chen Z, Cheng C-HC, Zhang J, Cao L, Chen L, Zhou L, Jin Y, Ye H, Deng C, Dai Z, et al. 2008a. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences* **105**: 12944–12949.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**: e1000112.

Clevers H. 2016. Modeling Development and Disease with Organoids. *Cell* **165**: 1586–1597.

Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**: 1063–1071.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43**: D662–9.

Davy BE, Robinson ML. 2003. Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in Hydin, a large novel gene. *Hum Mol Genet* **12**: 1163–1170.

Dawe HR, Shaw MK, Farr H, Gull K. 2007. The hydrocephalus inducing gene product, Hydin, positions axonemal central pair microtubules. *BMC Biol* **5**: 33.

De Boer M, Gavrieli R, van Leeuwen K, Wolf HR, Dushnitzki M, Bar-Yosef Y, Bar-Ziv A, Behar D, Lipitz S, Miller TE, et al. 2018. A false-carrier state for the c.579G>A mutation in the NCF1 gene in Ashkenazi Jews. *J Med Genet* **55**: 166–172.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* **41**: 44–52.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol* **1**: 69.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

Dermitzakis ET, Clark AG. 2001. Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol* **18**: 557–562.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, Berbari NS, Davy BE, Robinson ML, Rudd MK, Weber JL, et al. 2006. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* **88**: 762–771.

Doolittle WF, Brunet TDP. 2017. On causal roles and selected effects: our genome is mostly junk. *BMC Biol* **15**: 116.

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, et al. 2017. The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol* **18**: 49.

Duda TF, Palumbi SR. 1999b. Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod Conus. *Proceedings of the National Academy of Sciences* **96**: 6820–6823.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* **17**: 1266–1277.

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, et al. 2012. DUF1220-domain copy number implicated in human brain-size pathology and evolution. *Am J Hum Genet* **91**: 444–454.

Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**: 869–872.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. 2014. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet* **23**: 5866–5878.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**: 1356–1369.e22.

Fietz SA, Lachmann R, Brandl H, Kircher M, Samusik N, Schröder R, Lakshmanaperumal N, Henry I, Vogt J, Riehn A, et al. 2012. Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *Proc Natl Acad Sci U S A* **109**: 11836–11841.

Flomen RH, Collier DA, Osborne S, Munro J, Breen G, St Clair D, Makoff AJ. 2006. Association study of CHRFAM7A copy number and 2 bp deletion polymorphisms with schizophrenia and bipolar affective disorder. *Am J Med Genet B Neuropsychiatr Genet* **141B**: 571–575.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, et al. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**: 1465–1470.

Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M. 2018. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *Elife* **7**. http://dx.doi.org/10.7554/eLife.32332.

Florio M, Namba T, Pääbo S, Hiller M, Huttner WB. 2016a. A single splice site mutation in human-specific causes basal progenitor amplification. *Sci Adv* **2**: e1601941.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al. 2004. Lineage-Specific Gene Duplication and Loss in Human and Great Ape Evolution. *PLoS Biol* **2**: e207.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**: 445–449.

Gault J, Robinson M, Berger R, Drebing C, Logel J, Hopkins J, Moore T, Jacobs S, Meriwether J, Choi MJ, et al. 1998. Genomic organization and partial duplication of the human alpha7 neuronal nicotinic acetylcholine receptor gene (CHRNA7). *Genomics* **52**: 173–185.

Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD, Mark K, Vu TH, Alkan C, Cheng Z, et al. 2013. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* **92**: 221–237.

Glazko GV. 2003. Estimation of Divergence Times for Major Lineages of Primate Species. *Mol Biol Evol* **20**: 424–434.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**: 2847–2849.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577.

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2014. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* **42**: W494–500.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* **100**: 605–617.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Hartley SW, Mullikin JC. 2015. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics* **16**: 224.

Herculano-Houzel S, Catania K, Manger PR, Kaas JH. 2015. Mammalian Brains Are Made of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship with Body Mass. *Biotechnol Bioprocess Eng* **86**: 145–163.

Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23**: 843–854.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**: 688–696.

Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, et al. 2014. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**: 216–221.

Jian-Bing Fan, Urvashi Surti, Patricia Taillon-Miller, Linda Hsie, Giulia C. Kennedy, Lori Hoffner, Thomas Ryder, David G. Mutch, and Pui-Yan Kwok. 2002. Paternal Origins of Complete Hydatidiform Moles Proven by Whole Genome Single-Nucleotide Polymorphism Haplotyping. *Genomics* **79**: 58–62.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007b. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**: 1361–1368.

Johnson MB, Wang PP, Atabay KD, Murphy EA, Doan RN, Hecht JL, Walsh CA. 2015. Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat Neurosci* **18**: 637–646.

Johnson ME, National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, Green ED, Eichler EE. 2006.

Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**: 17626–17631.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001b. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514–519.

John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP, Thurman RE, et al. 2013. Genome-scale mapping of DNase I hypersensitivity. *Curr Protoc Mol Biol* **Chapter 27**: Unit 21.27.

Jun J, Ryvkin P, Hemphill E, Nelson C. 2009. Duplication mechanism and disruptions in flanking regions determine the fate of Mammalian gene duplicates. *J Comput Biol* **16**: 1253–1266.

Ju X-C, Hou Q-Q, Sheng A-L, Wu K-Y, Zhou Y, Jin Y, Wen T, Yang Z, Wang X, Luo Z-G. 2016c. The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *Elife* **5**. http://dx.doi.org/10.7554/eLife.18197.

Kent WJ. 2002a. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kent WJ. 2002b. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Kieffer LJ, Greally JM, Landres I, Nag S, Nakajima Y, Kohwi-Shigematsu T, Kavathas PB. 2002. Identification of a candidate regulatory region in the human CD8 gene complex by colocalization of DNase I hypersensitive sites and matrix attachment regions which bind SATB1 and GATA-3. *J Immunol* **168**: 3915–3922.

Kieffer LJ, Yan L, Hanke JH, Kavathas PB. 1997. Appropriate developmental expression of human CD8 beta in transgenic mice. *J Immunol* **159**: 4907–4912.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111–120.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kioussis D, Ellmeier W. 2002. Decision making in the immune system: Chromatin and CD4, CD8A and CD8B gene expression during thymic differentiation. *Nat Rev Immunol* **2**: 909–919.

Kozak M. 1991. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* **266**: 19867–19870.

Kozak M. 1989. The scanning model for translation: an update. *J Cell Biol* **108**: 229–241.

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**: D152–7.

Kurosaki T, Maquat LE. 2013. Rules that govern UPF1 binding to mRNA 3' UTRs. *Proc Natl Acad Sci U S A* **110**: 3357–3362.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**: 1009–1013.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**: 409–413.

Lechtreck K-F, Delmotte P, Robinson ML, Sanderson MJ, Witman GB. 2008. Mutations in Hydin impair ciliary motility in mice. *J Cell Biol* **180**: 633–643.

Lechtreck K-F, Witman GB. 2007. Chlamydomonas reinhardtii hydin is a central pair protein required for flagellar motility. *J Cell Biol* **176**: 473–482.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Levchenko A, Kanapin A, Samsonova A, Gainetdinov RR. 2018. Human Accelerated Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their Relevance for Brain Development. *Genome Biol Evol* **10**: 166–188.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**: 2380–2396.

Li J-T, Hou G-Y, Kong X-F, Li C-Y, Zeng J-M, Li H-D, Xiao G-B, Li X-M, Sun X-W. 2015. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). *Sci Rep* **5**: 8199.

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**: 22.

Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* **260**: 91–95.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**: 417–422.

Lupski JR, Stankiewicz PT. 2007. *Genomic Disorders: The Genomic Basis of Disease*. Springer Science & Business Media.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302**: 1401–1404.

Marques-Bonet T, Girirajan S, Eichler EE. 2009a. The origins and impact of primate segmental duplications. *Trends Genet* **25**: 443–454.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009b. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* **359**: 1685–1699.

Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. *Genome Res* **23**: 34–45.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**: 222–226.

Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* **508**: 199–206.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–426.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463.

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, et al. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**: 205–209.

Nuttle X, Huddleston J, O'Roak BJ, Antonacci F, Fichera M, Romano C, Shendure J, Eichler EE. 2013. Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat Methods* **10**: 903–909.

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak ACY, Lai YYY, Lin C, Chu C, Graves T, et al. 2014. Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* **15**: 387.

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM. 2012. Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* **13**: 853–866.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.

Olbrich H, Schmidts M, Werner C, Onoufriadis A, Loges NT, Raidt J, Banki NF, Shoemark A, Burgoyne T, Al Turki S, et al. 2012. Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *Am J Hum Genet* **91**: 672–684.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–45.

Olson MV. 1999. When Less Is More: Gene Loss as an Engine of Evolutionary Change. *Am J Hum Genet* **64**: 18–23.

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, et al. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**: 1619–1622.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**: 40–45.

Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**: 615–619.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* **2**: e168.

Pollen AA, Nowakowski TJ, Chen J, Retallack H, Sandoval-Espinosa C, Nicholas CR, Shuga J, Liu SJ, Oldham MC, Diaz A, et al. 2015. Molecular identity of human outer radial glia during cortical development. *Cell* **163**: 55–67.

Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM. 2006. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313**: 1304–1307.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–475.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43–49.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–1060.

Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J. 2015. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31**: 770–772.

Ritchie W, Granjeaud S, Puthier D, Gautheret D. 2008. Entropy measures quantify global splicing disorders in cancer. *PLoS Comput Biol* **4**: e1000011.

Roy AL. 2017. Pathophysiology of TFII-I: Old Guard Wearing New Hats. *Trends Mol Med* **23**: 501–511.

Rozycka A, Dorszewska J, Steinborn B, Lianeri M, Winczewska-Wiktor A, Sniezawska A, Wisniewska K, Jagodzinski PP. 2013. Association study of the 2-bp deletion polymorphism in exon 6 of the CHRFAM7A gene with idiopathic generalized epilepsy. *DNA Cell Biol* **32**: 640–647.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**: 944–950.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al. 2005. Segmental Duplications and Copy-Number Variation in the Human Genome. *Am J Hum Genet* **77**: 78–88.

Sousa AMM, Meyer KA, Santpere G, Gulden FO, Sestan N. 2017. Evolution of the Human Nervous System Function, Structure, and Development. *Cell* **170**: 226–247.

Sporny M, Guez-Haddad J, Kreusch A, Shakartzi S, Neznansky A, Cross A, Isupov MN, Qualmann B, Kessels MM, Opatowsky Y. 2017b. Structural History of Human SRGAP2 Proteins. *Mol Biol Evol* **34**: 1463–1478.

Sporny M, Guez-Haddad J, Kreusch A, Shakartzi S, Neznansky A, Cross A, Isupov MN, Qualmann B, Kessels MM, Opatowsky Y. 2017c. Structural History of Human SRGAP2 Proteins. *Mol Biol Evol* **34**: 1463–1478.

Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res* **14**: 2209–2220.

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428**: 415–418.

Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**: 1177–1184.

Stephan H, Frahm H, Baron G. 1981. New and Revised Data on Volumes of Brain Structures in Insectivores and Primates. *Fprc* **35**: 1–29.

Stephens SG. 1951. Possible Significance of Duplication in Evolution. In *Advances in Genetics*, pp. 247–265.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010b. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015a. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015b. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP, Emes RD, Lynch VJ. 2016. TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *Elife* **5**. http://dx.doi.org/10.7554/elife.11994.

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, et al. 2018. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* **173**: 1370–1384.e16.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F. 2005. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* **22**: 2265–2274.

vonHoldt BM, Shuldiner E, Koch IJ, Kartzinel RY, Hogan A, Brubaker L, Wanser S, Stahler D, Wynne CDL, Ostrander EA, et al. 2017. Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Sci Adv* **3**: e1700398.

Wallace VA, Raff MC. 1999. A role for Sonic hedgehog in axon-to-astrocyte signalling in the rodent optic nerve. *Development* **126**: 2901–2909.

Wolach B, Gavrieli R, de Boer M, van Leeuwen K, Wolach O, Grisaru-Soen G, Broides A, Etzioni A, Somech R, Roos D. 2018. Analysis of Chronic Granulomatous Disease in the Kavkazi Population in Israel Reveals Phenotypic Heterogeneity in Patients with the Same NCF1 mutation (c.579G>A). *J Clin Immunol* **38**: 193–203.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

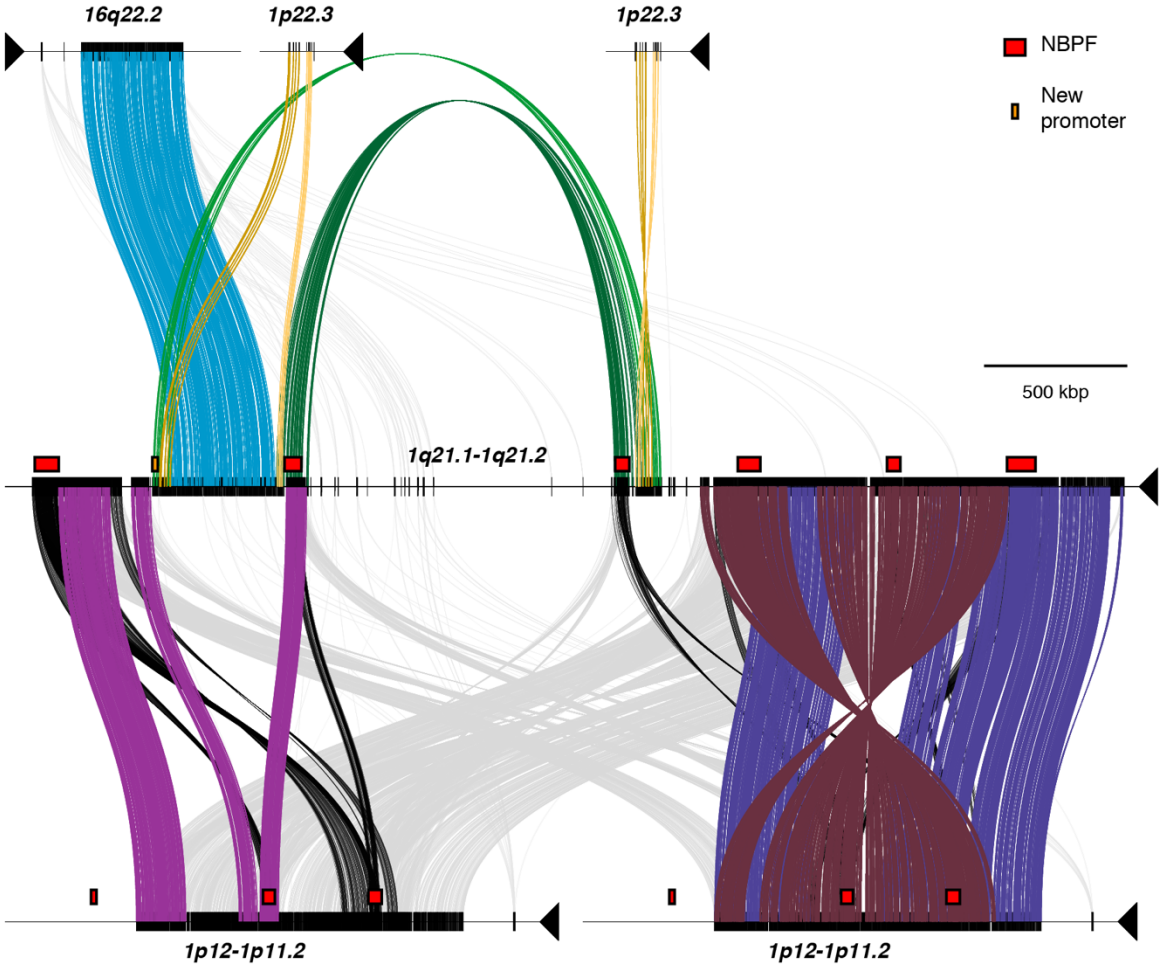
Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–394.

Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon KK, et al. 2014b. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet* **46**: 88–92.

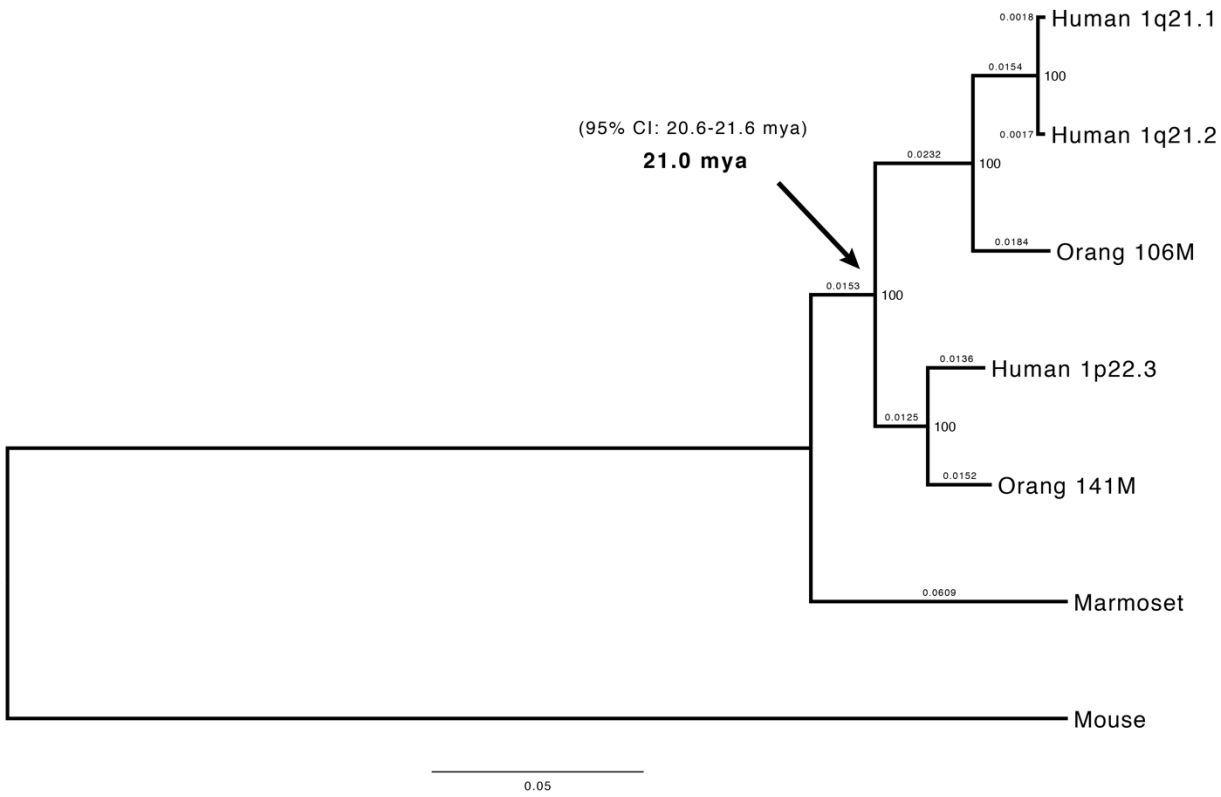
Zhao J, Ma J, Deng Y, Kelly JA, Kim K, Bang S-Y, Lee H-S, Li Q-Z, Wakeland EK, Qiu R, et al. 2017. A missense variant in NCF1 is associated with susceptibility to multiple autoimmune diseases. *Nat Genet* **49**: 433–437.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**: 892–897.

APPENDIX A: ADDITIONAL MATERIAL FOR CHAPTER 2



Supplementary Figure 1. *HYDIN* and associated chromosome 1 duplications. The segmental duplications that led to the formation of *HYDIN2* are seen in the context of larger genomic rearrangements on chromosome 1. Duplications are visualized using Miropeats (s = 800; Parsons, 1995). The central sequence represents chromosome 1q21.1-1q21.2 (chr1:146061572-150028860) with the *HYDIN* duplication in light blue. Also highlighted is the new *HYDIN2* promoter. The homology with chromosome 16q22.2 (chr16:70611384-71368670) and chromosome 1p22.3 (chr1:87315068-87609100) shown above the 1q21 region is the same as in Figure 1e. Shown below the 1q21 region is additional homology to chromosome 1p12-1p11.2 (chr1:119530046-121401465). This region is shown twice for clarity. Locations of the core duplicon gene *NBPF*, by GENCODE annotation, are highlighted as red boxes and are found at or near the breakpoints of most observed rearrangements, including the palindromic duplication shown at the bottom right (also described in O'Bleness, 2014).



Supplementary Figure 2A. Timing of first duplication of the 5' and 3' segments flanking *HYDIN2*. We estimate that the segmental duplication (65 kbp) that was bisected by the *HYDIN2* duplication underwent an earlier duplication from chromosome 1p22 to chromosome 1q21 approximately 21.0 mya (95% CI: 20.6-21.6 mya). This is consistent with the observation that this segment is at haploid single-copy in marmoset, macaque, and baboon, and at haploid copy number 2 in gibbon and orangutan.

Sequences. Sequences homologous to the three human loci were identified in primates and in mouse using BLAT and the UCSC Genome Browser. A single homologous locus was identified in mouse, which was used to root the tree. One homologous locus was identified in marmoset, representing New World monkeys, and one homologous locus was identified in both the rhesus macaque and baboon, representing Old World monkeys. This suggests the duplication occurred after the divergence of apes from Old World monkeys. Within apes, two homologous loci were identified in gibbon and orangutan each, suggesting that the duplication occurred within the common ancestor of hominoids (lesser and greater apes). Multiple homologous loci of varying lengths were identified in chimpanzee, suggesting this region has undergone further rearrangement and amplification in that lineage.

Alignment and tree building. A 30,872 bp multiple sequence alignment (MSA) was generated using MAFFT with sequences deriving from the single locus in mouse and marmoset, the two loci in orangutan, and the three loci in human. The phylogenetic tree was inferred using the maximum-likelihood method with distances estimated using the Tamura-Nei model and tested with 50 bootstrap replicates. Branches are labeled with the number of substitutions per site and nodes are labeled with bootstrap support.

Timing the duplication. The two loci on human 1q21 pass the relative rate test ($p = 0.65$, outgroup orangutan), as do either with their ortholog in orangutan ($p = 0.89$ for 1q21.1 and $p =$

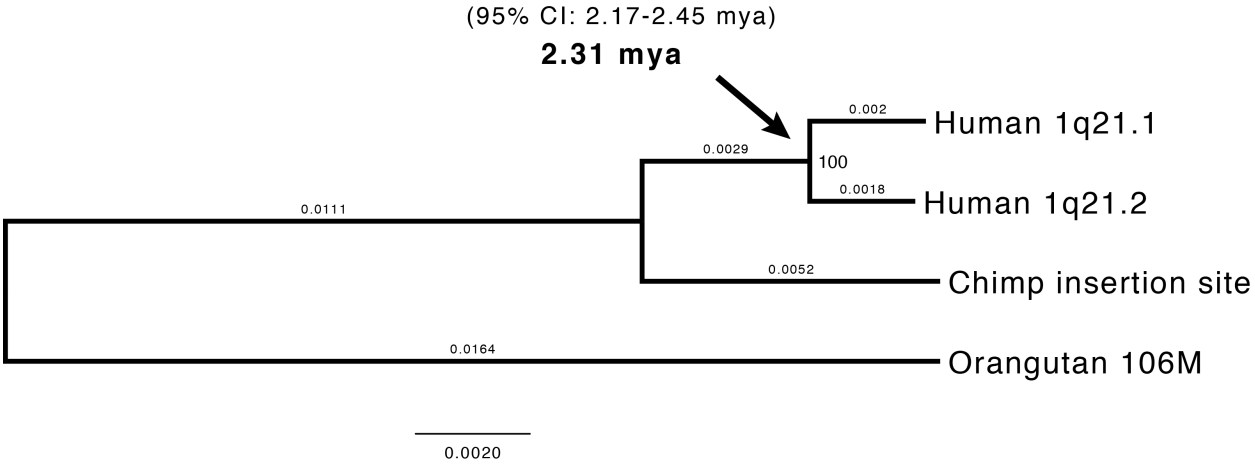
0.61 for 1q21.2, outgroup marmoset). The locus on human 1p22.3 and its closest orangutan ortholog also pass the relative rate test ($p = 0.25$). However, the 1q and 1p clades appear to be evolving at different rates ($p < 0.00001$ for both, outgroup marmoset), and neither clade passes the relative rate test with marmoset ($p = 0.004$ for 1q and $p < 0.00001$ for 1p, outgroup mouse). In marmoset, the derived sequence sits on the long arm of chromosome 7, consistent with known translocations that differentiate the species (Sherlock, 1996). All in all, we cannot assume the same local rate of neutral substitution between any pair in these three clades (marmoset, 1p clade, 1q clade).
We use the 1p22.3 clade to estimate the timing of the duplication since that region is syntenic to the original sequence in marmoset. These estimates are based on the proportion of genetic distance since divergence from marmoset that occurred since the duplication of our segment of interest.

clade 1p
$D_{\text{human 1p22.3 branch}} = 0.013632$
$D_{\text{orangutan 141M branch}} = 0.015181$
$\text{Average}(D_{\text{orangutan 141M branch}}, D_{\text{human 1p22.3 branch}}) = 0.014407$
$D_{\text{1p until human-orang split}} = 0.012466$
$D_{\text{from divergence from marmoset to duplication}} = 0.015257$
(Distance after duplication)/(Total distance since divergence from marmoset) = (0.014407 + 0.012466)/(0.014407 + 0.012466 + 0.015257) = 0.63785

Time of human–marmoset divergence (Glazko and Nei 2002) = 33 mya (range 32-36)

Estimate based on *clade 1p*
0.63785 * 33 mya = **21.0 mya** (95% CI: 20.6-21.6 mya)

We calculated 95% confidence intervals around our duplication timing estimate above using branch length error estimates and the following approach. First, for each branch in the tree, we set the branch length to a randomly chosen value between the actual branch length minus the branch length error (or zero if that value is negative) and the actual branch length plus the branch length error, inclusive. Second, we recomputed the timing estimate above using the same calculations as for the original tree except using the modified branch length values. For these calculations, we assumed a human–marmoset divergence of 33 mya. Third, we repeated the above two steps until we obtained one million modified trees and corresponding timing estimates. Finally, we sorted the estimates and reported the 25,000th and 975,000th sorted timing estimate values as the 95% confidence interval around the corresponding timing point estimate: 20.6-21.6 mya. Note that this error is less than the error that results from uncertainty in the timing estimate for human–marmoset divergence (Glazko and Nei, 2003).
This range of estimates, which places the duplication just after the divergence of apes and Old World monkeys (23 mya, range 21-25), is consistent with where we place the duplication in evolutionary history based on its presence and absence in reference genomes. It is important to note, however, that nonhuman primate reference genomes may be incomplete, especially in these duplicated regions.

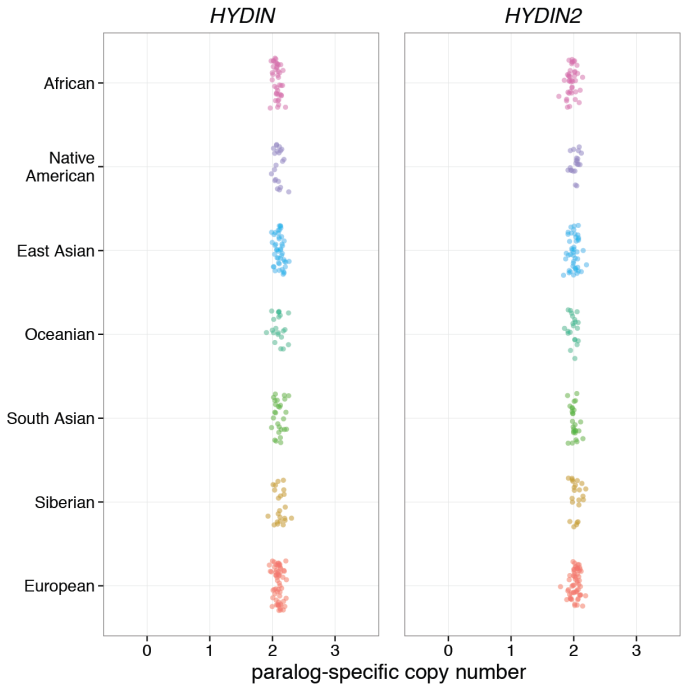


Supplementary Figure 2B. Timing of second duplication of the 5' and 3' segments flanking *HYDIN2*. To time the more recent duplication of the sequence immediately flanking *HYDIN2*, which occurred between 1q21.1 and 1q21.2, a new MSA was generated with sequence from the two human loci, the homologous chimpanzee locus, and the homologous orangutan locus. The sequences for the two human loci and the orangutan locus are the same as used in Figure S2a. Because this region appears to have undergone further duplication in chimpanzee (data not shown), we identified and sequenced a BAC containing homologous chimpanzee sequence (CH251-231E10). A 209,299 bp MSA was generated using MAFFT and manually edited for obvious alignment errors. The phylogenetic tree was inferred using the maximum-likelihood method with distances estimated using the Kimura 2-paramter model and tested with 50 bootstrap replicates. Branches are labeled with the number of substitutions per site and nodes are labeled with bootstrap support.

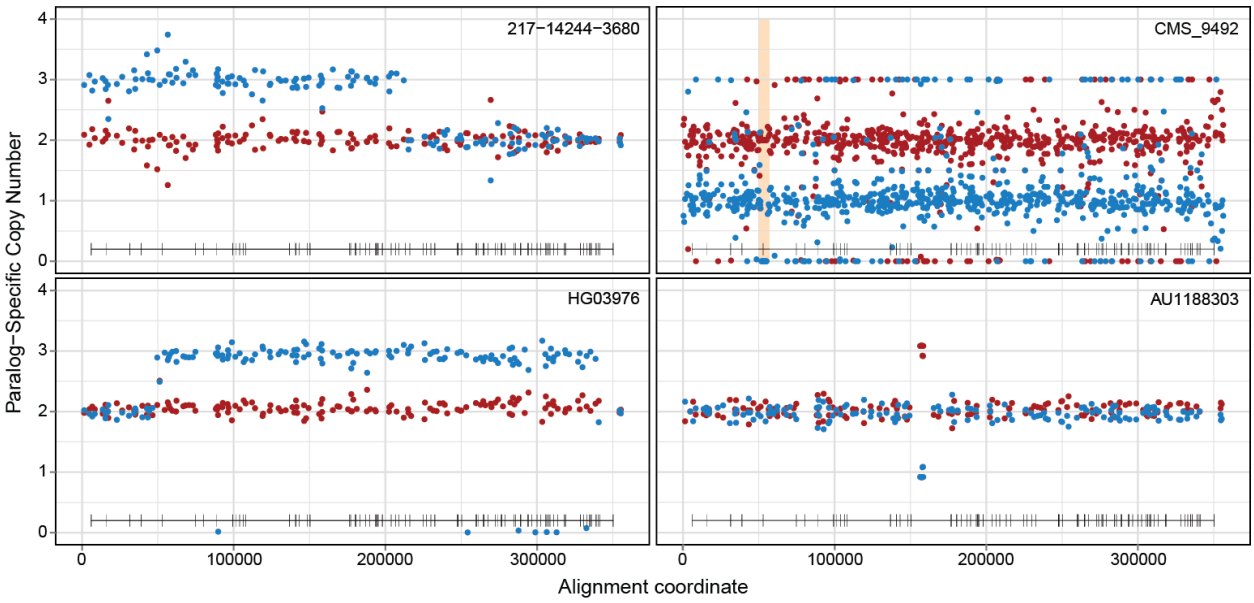
Timing the human-specific duplication. We similarly estimate the timing of the human-specific duplication of the segments that flank *HYDIN2* using orangutan as the outgroup, and a divergence time of 6 mya (Glazko and Nei, 2003), based on the proportion of genetic distance since divergence from chimpanzee that occurred since the duplication of our segment of interest. The human branches pass the relative rate test ($p = 0.18$).

$D_{\text{average human 1q21.1/1q21.2 branch}} = (0.002032 + 0.001846)/2 = 0.001939$
$D_{\text{1q human until 1q21.1/1q21.2 split}} = 0.002932$
$D_{\text{1q human total}} = (0.002032 + 0.001846)/2 + 0.002932 = 0.004871$
(Distance after duplication)/(Total evolutionary distance since divergence from chimpanzee) = 0.001939/(0.004871 + 0.005213) = 0.1923
0.1923 * 2 * 6 mya = **2.31 mya** (95% CI: 2.17-2.45 mya)*

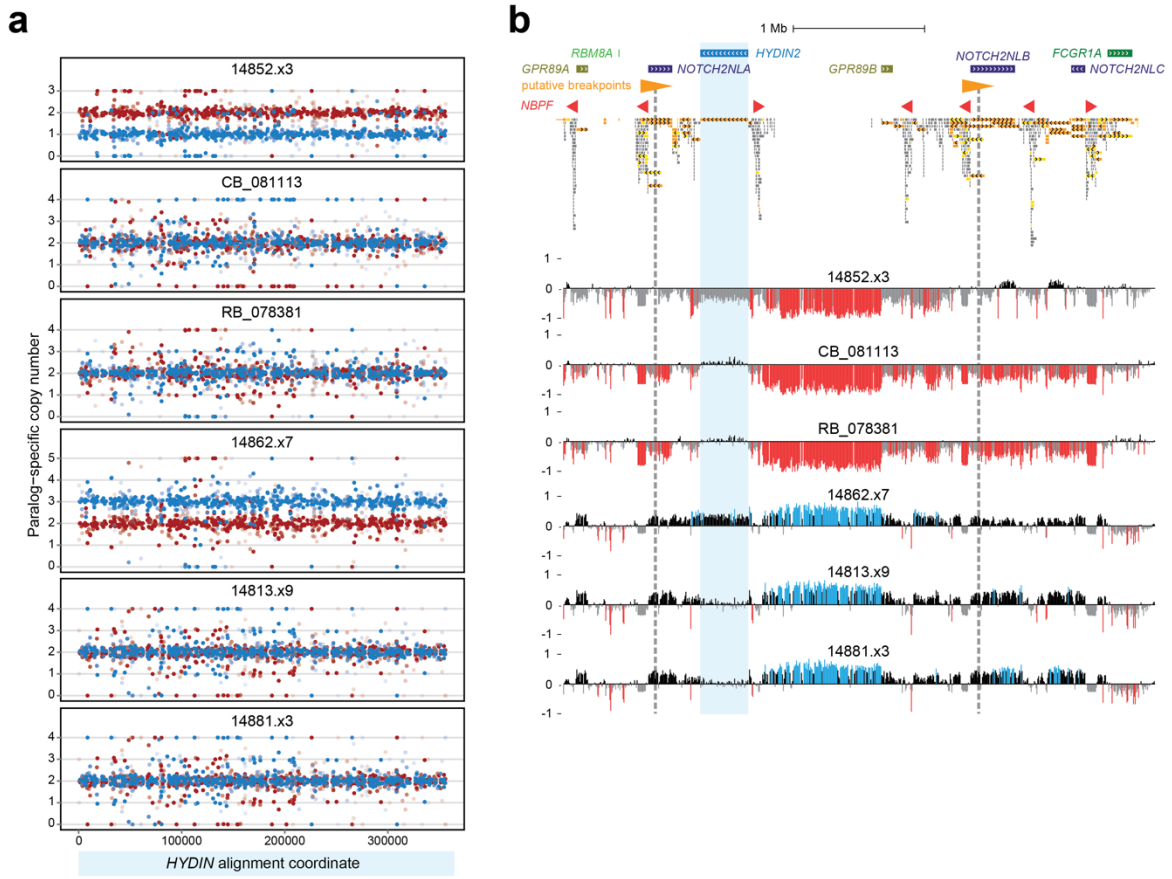
This estimate is more recent than the estimate of the *HYDIN2* duplication, however it must necessarily have preceded the insertion of the duplicated *HYDIN2* sequence. We think it most likely that interlocus gene conversion is responsible for reducing the genetic divergence between these segments and distorting this timing estimate.



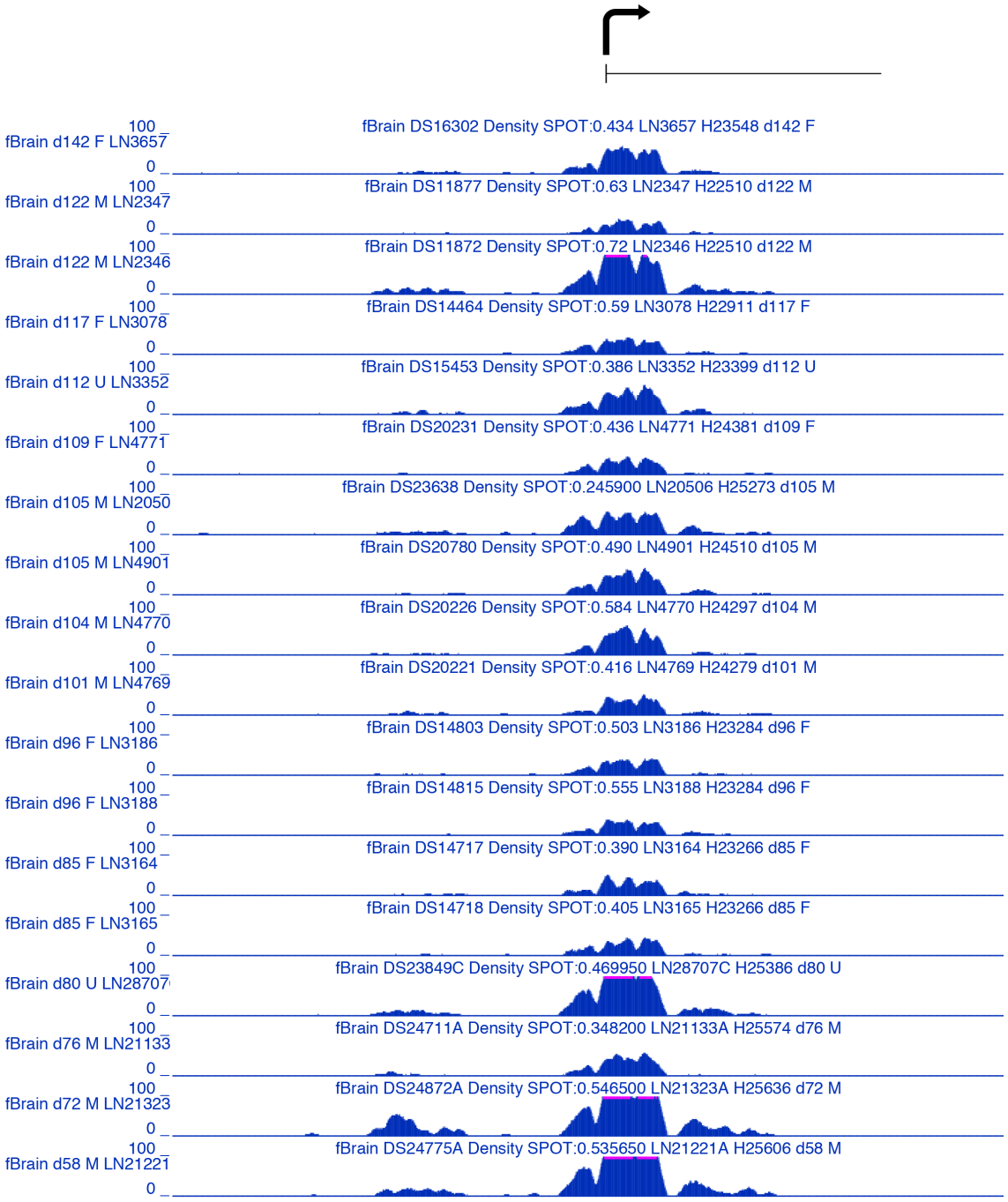
Supplementary Figure 3. Paralog-specific copy number estimates for 236 individuals from the Human Genome Diversity Project (HGDP). Whole-genome sequencing data from the HGDP were mapped to the *HYDIN* segmental duplication, and SUNK-based read depth was used to assess paralog-specific diploid copy number. Both paralogs are found at copy number 2 in all individuals shown, including 41 Africans, 21 Native Americans, 45 East Asians, 21 Oceanians, 27 South Asians, 22 Siberians, and 59 Europeans.



Supplementary Figure 4. *HYDIN* internal structural variation and interlocus gene conversion. Examples of structural variation within *HYDIN* paralogs and a putative interlocus gene conversion event. Each point shows a paralog-specific copy number estimate (red, *HYDIN*; blue, *HYDIN2*) based on sequencing data corresponding to a single MIP targeting sequence that distinguishes *HYDIN* paralogs. Sequencing reads were analyzed to compute paralog-specific read count relative frequencies for each MIP, which were multiplied by the aggregate estimated copy number at each target site to infer paralog-specific copy number. Shown are an ~212 kbp duplication affecting *HYDIN2* (upper left), an ~289 kbp duplication affecting *HYDIN2* (lower left), a putative ~3 kbp deletion affecting *HYDIN2* in a 1q21 microdeletion patient (orange highlight, upper right), and an ~2 kbp interlocus gene conversion event (lower right). Note that the putative *HYDIN2* deletion shown might instead reflect interlocus gene conversion—all reads for 11 consecutive MIPs over the highlighted interval mapped to *HYDIN*, consistent with zero copies of *HYDIN2* and either two or three copies of *HYDIN*. Also note that the interlocus gene conversion event identified in AU1188303 is the same as that discovered in 11094.s1 (**Figure 2C**, bottom right), indicating this event likely segregates at very low frequency. 153 MIPs were used for genotyping all individuals shown except CMS_9492, a 1q21 microdeletion patient genotyped with 717 MIPs. All events shown were detected by an automated caller. Duplicated exons based on the canonical *HYDIN* gene model are indicated at the bottom of each plot.



Supplementary Figure 5. 1q21 rearrangement breakpoint variability. a) 717 MIPs targeting regions that distinguish *HYDIN* paralogs were employed to genotype *HYDIN* paralog-specific copy number in 48 1q21 microdeletion and 25 1q21 microduplication patients. Points show *HYDIN* paralog-specific copy number estimates (red, *HYDIN*; blue, *HYDIN2*) for three microdeletion patients (14852.x3, CB_081113, and RB_078381) and three microduplication patients (14862.x7, 14813.x9, and 14881.x3). These estimates were calculated as the product of the paralog-specific read count relative frequency for a particular MIP and the aggregate estimated copy number at the corresponding target site. The MIP results indicate that 1q21 rearrangements do not always include *HYDIN2*. **b)** The segmental duplication organization of a 4.5 Mbp region at chromosome 1q21 (GRCh38 chr1:145,500,001-150,000,000) is shown along with array CGH profiles for the individuals in panel a. Thin colored boxes indicate sequences duplicated between this region and another genomic locus, with colors corresponding to sequence identity (orange = 99% or above, yellow = 98%–99%, gray = 90%–98%) and markings showing orientation (right-pointing, directly oriented; left-pointing, inversely oriented) between duplication pairs. Thick colored boxes highlight locations of several duplicated genes. Orange triangles indicate high-identity, directly oriented *NOTCH2NL-NBPF* duplications, with putative breakpoints of the canonical 1q21 rearrangement shown as vertical gray dashed lines. Shown below are the locations of *NBPF* core duplicons. Array CGH confirms 1q21 rearrangements in these individuals, and array data over *HYDIN2* (blue highlight) indicate a loss or gain in some (14852.x3 and 14862.x7) but not others (all other individuals shown), validating the MIP data shown in panel a.



Supplementary Figure 6. The *HYDIN2* promoter corresponds to a peak of chromatin accessibility in fetal brain. Reads indicating sites of chromatin accessibility as determined by sensitivity to DNase I digestion from various fetal brain time points (day 58 – day 142) were mapped using mrsFAST-Ultra (Hach, 2014) to allow for measurement over duplicate sequence. Shown is a ~14 kbp region (chr1:146479496-146493419) surrounding the acquired promoter and first exon of *HYDIN2*, with a peak is visible at all time points. See Appendix A: Table S8 for sample information.

Supplementary Table 1. MIP-based copy-number genotyping of *HYDIN2*.

Number of individuals

Cohort	Individuals genotyped	Cases or controls	*HYDIN2* deletions (CN = 1)	*HYDIN2* normal (CN = 2)	*HYDIN2* duplications (CN = 3)	*HYDIN2* complex event
SVIP 16p	336	both	0	336	0	0
16p triplications	*1*		*0*	*1*	*0*	*0*
16p duplications	*70*		*0*	*70*	*0*	*0*
16p normals	*187*		*0*	*187*	*0*	*0*
16p deletions	*78*		*0*	*78*	*0*	*0*
SVIP 1q21	120	both	45	52	22	1
1q21 duplications	*25*		*0*	*3*	*22*	*0*
1q21 normals	*46*		*0*	*46*	*0*	*0*
1q21 deletions	*49*		*45*	*3*	*0*	*1*
HapMap/1KG	1082	controls	1	1076	4	1
AGRE Probands	941	cases	0	941	0	0
SSC Probands	787	cases	0	786	0	1
TASC Probands	890	cases	0	888	1	1
TASC Siblings	123	controls	0	123	0	0
AGRE Siblings	638	controls	0	637	0	1
SSC Siblings	1133	controls	0	1130	1	2
SSC Parents	2	controls	0	2	0	0
Cell Lines	3	controls	0	3	0	0
All Combined	6055	both	46	5974	28	7
Cases (excluding SVIP)	*2618*	*cases*	*0*	*2615*	*1*	*2*
Controls (excluding SVIP)	*2981*	*controls*	*1*	*2971*	*5*	*4*
SVIP cohorts	*456*	*both*	*45*	*388*	*22*	*1*

Number of unrelated chromosomes

Cohort	Chromosomes genotyped	Cases or controls	*HYDIN2* deletions (CN = 1)	*HYDIN2* normal (CN = 2)	*HYDIN2* duplications (CN = 3)	*HYDIN2* complex event
HapMap/1KG_YRI	117/117	controls	1/1	115/115	1/1	0/0
HapMap/1KG_GBR	185/186	controls	0/0	185/186	0/0	0/0
HapMap/1KG_FIN	190/190	controls	0/0	190/190	0/0	0/0
HapMap/1KG_CHS	4/4	controls	0/0	4/4	0/0	0/0
HapMap/1KG_CHB	96/96	controls	0/0	96/96	0/0	0/0
HapMap/1KG_PUR	4/4	controls	0/0	4/4	0/0	0/0
HapMap/1KG_IBS	190/190	controls	0/0	190/190	0/0	0/0
HapMap/1KG_KHV	10/10	controls	0/0	10/10	0/0	0/0
HapMap/1KG_CDX	2/2	controls	0/0	2/2	0/0	0/0
HapMap/1KG_GWD	10/10	controls	0/0	10/10	0/0	0/0
HapMap/1KG_PJL	2/2	controls	0/0	2/2	0/0	0/0
HapMap/1KG_ESN	10/10	controls	0/0	10/10	0/0	0/0
HapMap/1KG_MSL	6/6	controls	0/0	6/6	0/0	0/0
HapMap/1KG_ITU	8/8	controls	0/0	7/7	0/0	1/1
HapMap/1KG_CEU	237/243	controls	0/0	237/243	0/0	0/0
HapMap/1KG_JPT	102/102	controls	0/0	102/102	0/0	0/0
HapMap/1KG_LWK	140/151	controls	0/0	140/151	0/0	0/0
HapMap/1KG_ASW	115/131	controls	0/0	113/129	2/2	0/0
HapMap/1KG_MXL	112/115	controls	0/0	112/115	0/0	0/0
HapMap/1KG_TSI	174/174	controls	0/0	174/174	0/0	0/0
HapMap/1KG_GIH	2/2	controls	0/0	2/2	0/0	0/0
HapMap/1KG_MKK	150/150	controls	0/0	150/150	0/0	0/0
HapMap/1Kg_all_combined	1866/1903	controls	1/1	1861/1898	3/3	1/1

*First number (before slash) is minimum number of unrelated chromosomes, second number (after slash) is maximum number of unrelated chromosomes.

Supplementary Table 2. *HYDIN* duplication, deletion, and interlocus gene conversion events.

Variant*	Carrier(s)	Carrier status(es)	Called by	Alignment location	Genic location**	Minimum size	Number of supporting MIPs	Transmission observed?†	Previously known?
HYDIN internal duplication	03C16388	case	automated caller	225785-240797	intron 42 to intron 45	15 kbp	9	no	no
HYDIN2 internal deletion	CMS_9492	case (chromosome 1q21 deletion)	automated caller	51702-55192	intron 9 to intron 10	3 kbp	11	no	no
HYDIN2 internal deletion	SSC 11176.s1	control	automated caller	138752-141883	intron 19 to intron 20	3 kbp	5	no	no
HYDIN2 internal duplication	217-14244-3680	case	automated caller	1-212135	intron 5 to intron 39	212 kbp	92	no	no
HYDIN2 internal duplication	HG03976	control	automated caller	49583-338591	intron 9 to intron 80	289 kbp	129	no	no
HYDIN2 internal duplication	SSC 11422.p1	case	automated caller	54502-212135	intron 10 to intron 39	158 kbp	70	no	no
HYDIN -> HYDIN2 conversion	SSC 11094.s1, AU1188303	control, control	automated caller	156380-158432	intron 23	2 kbp	5	no	no
HYDIN deletion	SSC Si1295	case	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	no	yes
HYDIN duplication	05C47106	case	automated caller	1-357274	intron 5 to intron 84	357 kbp	153	no	yes
HYDIN duplication	Ag449	case	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	no	yes
HYDIN2 deletion	NA19190	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	no	yes
HYDIN2 duplication	NA19201	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	yes	yes
HYDIN2 duplication	NA19703	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	yes	yes
HYDIN2 duplication	NA19705	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	153	yes	yes
HYDIN2 duplication	NA20127	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	272	yes	yes
HYDIN2 duplication	215-13135-1523	case	automated caller	1-357274	intron 5 to intron 84	357 kbp	153	yes	yes
HYDIN2 duplication	SSC 12325.s2	control	automated caller	1-357274	intron 5 to intron 84	357 kbp	153	yes	yes

*For interlocus gene conversion variants, the conversion donor is listed before the arrow and the conversion acceptor after the arrow.

**Exons and introns are numbered according to the *HYDIN* gene model, with exons 6-84 shared between *HYDIN* and *HYDIN2*.

†Did we observe at least one instance of transmission of this variant within a trio? Note that for most individuals, DNA from parents and/or children was not available, so not observing transmission does not necessarily indicate the variant is *de novo*.

Supplementary Table 3: Locations of other copies of the *HYDIN2* promoter-associated duplication, their relationship to *NBPF*, and evidence for transcription

BLAT result	chr	start	end	% identity	associated *NBPF* member	distance to EV15 promoter (kbp)	distance to CM promoter (kbp)	distance to *NBPF* 3' UTR	orientation (promoter/ NBPF)	relationship of promoter to *NBPF*	spliced ESTs*	tissue source(s) of spliced EST(s)**
1	chr1	146427435	147030676	100.0%	*NBPF12*	451	472		+/+	upstream	yes	hippocampus (3), brain (1), fetal brain (1)
2	chr1	148088687	148263630	99.5%	*NBPF11*	98	118		-/-	upstream	yes	fetal brain (1)
3	chr1	145239245	145540744	91.5%	*NBPF20*	49	69		+/-	upstream	yes	hippocampus (1)
4	chr1	148002800	148304299	91.4%	*NBPF11*			52	+/-	downstream	yes	hippocampus (1), NE lung carcinoid (2), carcinoid (2)
5	chr1	145973900	146728039	90.8%	*NBPF10*		222		-/-	upstream	yes	cerebellum (2)
6	chr1	148471458	148681704	90.2%	*NBPF14*			2.1	-/-	downstream	yes	pooled tissues (1)
7	chr1	144393920	144469916	90.3%	*NBPF15*			2.1	-/-	downstream	no	
8	chr1	145239245	145540744	90.2%	*NBPF20*			2.1	-/-	downstream	no	
9	chr1	149354590	149691092	90.6%	*NBPF19*			2.1	-/-	downstream	yes	germinal center B-cell (1), embryonic 1st PA (1)
10	chr1	120795255	120894554	90.0%	*NBPF26*			2.8	+/+	downstream	yes	brain (1), spleen (1)

*Are spliced human ESTs present with 5' ends contained within the promoter?

**When >5 spliced ESTs were present, only tissue sources for 5 were recorded, chosen by order of display on the UCSC genome browser. NE: neuroendocrine; PA: pharyngeal arch.

Supplementary Table 4. Pairwise dN/dS values for *HYDIN* in primates.

Species/Paralog	Human duplicate	Human ancestral	Chimpanzee	Gorilla	Gibbon	Macaque	Marmoset
Human duplicate	-						
Human ancestral	0.47	-					
Chimpanzee	0.39	0.29	-				
Gorilla	0.28	0.23	0.24	-			
Gibbon	0.29	0.28	0.28	0.26	-		
Macaque	0.25	0.24	0.25	0.23	0.26	-	
Marmoset	0.31	0.30	0.30	0.30	0.31	0.31	-

Supplementary Table 5: Likely gene disruptive events detected in *HYDIN/HYDIN2* by MIP-based sequencing of exons.

Paralog*	Variant	Exon	Intron	Protein position	Amino acid	Samples	Frequency	Number genotyped**
HYDIN	splice_donor	-	29/85	-	-	1	0.04%	2603
HYDIN	stop_gained	11/86	11/86	1330	R/*	1	0.04%	2599
HYDIN2	splice_acceptor	-	14/85	-	-	2	0.08%	2605
HYDIN2	frameshift	19/86	-	2531-2532	A/X	1	0.04%	2607
HYDIN2	splice_donor	-	28/85	-	-	1	0.04%	2600
HYDIN2	frameshift	41/86	-	2115-2116	VI/VSX	10	0.38%	2598
HYDIN2	splice_donor	-	42/85	-	-	1	0.04%	2599
HYDIN2	frameshift	46/86	-	2485	A/X	1	0.04%	2600
HYDIN2	frameshift	48/86	-	2680	G/X	1	0.04%	2598
HYDIN2	splice_acceptor	-	54/85	-	-	1	0.04%	2595
HYDIN2	splice_acceptor	-	67/85	-	-	6	0.23%	2601
HYDIN2	stop_gained	80/86	-	4563	W/*	1	0.04%	2599
unknown	splice_donor	-	8/85	-	-	1	0.04%	2601
unknown	frameshift	17/86	-	766-768	LVL/X	1	0.04%	2597
unknown	splice_donor	-	31/85	-	-	1	0.04%	2595
unknown	frameshift	35/86	-	1771	P/X	24	0.92%	2595
unknown	frameshift	37/86	-	1885	N/X	1	0.04%	2603
unknown	splice_donor	-	43/85	-	-	3	0.12%	2602
unknown	stop_gained	45/86	-	2352	R/*	1	0.04%	2597
unknown	splice_donor	-	50/85	-	-	1	0.04%	2598
unknown	frameshift	51/86	-	2876	T/NX	1	0.04%	2604
unknown	stop_gained	52/86	-	2928	Q/*	1	0.04%	2595
unknown	splice_acceptor	-	52/85	-	-	5	0.19%	2594
unknown	splice_donor	-	53/85	-	-	2	0.08%	2601
unknown	stop_gained	62/86	-	3504	E/*	1	0.04%	2604
unknown	stop_gained	70/86	-	3970	R/*	1	0.04%	2604
unknown	splice_donor	-	75/86	-	-	1	0.04%	2602
unknown	splice_acceptor	-	83/85	-	-	1	0.04%	2598
unknown	stop_gained	84/86	-	4846	Q/*	1	0.04%	2602

*Paralog determined by presence of SUN on variant-containing MIP reads, variants identified by MIP reads that did not intersect a SUN could not be assigned; Variants in *HYDIN2* are annotated with the exon numbering scheme from *HYDIN*

**Number of samples successfully genotyped for this variant (Freebayes)

Supplementary Table 6. Phenotypes for patients having atypical chromosome 1q21 rearrangements.

Individual	Cohort	1q21 CNV	*HYDIN2* copy number (MIP)	Phenotype	Notes
14813.x10	SVIP	duplication	2	The patient is a 6-year-old Caucasian male. Facial features include macrocephaly, mild bilateral down slanting palpebral fissures and small ears with thickened/overfolded helices and large lobes that are posteriorly rotated. Additional facial features include mild plagiocephaly, small mouth, and tented upper lip. Physical examination reveals left 5th and right 4th-5th clinodactyly of fingers as well as 4th and 5th bilateral clinodactyly of toes. Patient has an anterior hair whorl, a hypopigmented spot on his right abdomen and a Café-au-lait colored macule on his left inner knee. Examination also reveals lordosis and double sacral dimples with gluteal cleft at midline (shallow, but easily visualized). Patient currently has a significantly above average head circumference measurement (72 months: HC = 56.1, z = 2.68). Patient has a significantly above average height and weight, but has a normal BMI of 21.4 (72 months: height = 127.5 cm, z = 2.36 weight = 34.74 kg, z = 2.83). Patient was diagnosed with Autistic Disorder (confirmed with ADOS, ADI, and clinical judgment using DSM-IV criteria) as well as Mild Mental Retardation (based on diagnostic history, cognitive and adaptive assessment, parent report of symptoms and clinical judgment). He uses mostly single words with occasional phrases. Patient shows autism-related impairments in language and social communication, including odd intonation/pitch, echolalia, stereotyped language, limited spontaneous use of gestures and eye contact, limited range of facial expressions, limited response to social smile and name, difficulties with joint attention, initiation of social interaction and pretend play. Finger flapping and repetitive use of play objects are also noted. Patient's cognitive abilities fall in the Very Low range (Mullen Verbal IQ = 38, DAS-II Nonverbal IQ = 51) Adaptive abilities fall in the Low range (Vineland Adaptive Composite = 64). Patient used his first single words at 54 months of age. Earlier in his history, he did spontaneously sing/babble several simple songs, but mother reports loss of babbling/singing between ages 2 and 3. Abnormalities were first noted in his development at 24 months of age. Patient has low receptive and expressive communication (Vineland Communication Domain Standard Score = 63; Mullen Receptive Language Subscale Age Equivalent = 28 months, Mullen Expressive Language Subscale Age Equivalent = 27 months). Patient is right-hand dominant. Patient evidences moderate impairments in gross motor functioning and manual dexterity, per parent report (Vineland Motor Skills Domain Standard Score = 72). Patient's parent endorses concerns with attention problems and hyperactivity. Patient was born at 39 weeks gestation via cesarean section due to mother's anal fissures, but no other labor complications were reported. Following birth, patient experienced hyperbilirubinemia but did not receive treatment. Patient was diagnosed with intermittent constipation at 3 months of age, which has not been resolved. At 19 months of age, patient was diagnosed with heart murmurs and has a history of chronic pneumonia. In addition, patient was diagnosed with Tics at 2 years of age. Patient has strabismus, and was initially treated with glasses (later unnecessary). Parent reports that the patient is excessively clumsy and uncoordinated, but no other neurological diagnoses have been made.	
14813.x9	SVIP	duplication	2	The patient is a 37-year-old Caucasian male. Physical examination reveals pupillary hippus bilaterally, as well as an irregular bordered light brown macule on the left cheek. Patient has an irregular bordered light brown macule on his right upper back, as well as an oval café au-lait spot on his left buttock. Patient has wide-spaced 1st-2nd toes bilaterally, and very mild left concave thoracic scoliosis. Patient also has mild plagiocephaly. Additionally, he has a mild left deviation of the gluteal cleft at the top, but no dimple. Patient has a head circumference of 60.0cm. Patient has a height of 177.5cm and a weight of 122.92kg, with a BMI indicative of obesity. Patient has a diagnosis of Depressive Disorder Not Otherwise Specified (confirmed with rating scales, clinical judgment using DSM-IV criteria) for which he takes Sertraline. Patient does not meet diagnostic criteria for ASD or other related disorders, but is observed to have a flat affect and flat/monotone speech. He self-reports some social difficulties, few close friendships, and has moderately low scores on Vineland Socialization scale (Standard Score = 85). As a child he notes he was shy, stayed to himself and did not talk much. Patient's cognitive abilities fall in the Average range (WASI Verbal IQ = 111, Nonverbal IQ = 97, Full Scale IQ = 105). Besides the aforementioned scores on Socialization, his adaptive abilities fall in the Adequate range (Adaptive Composite = 90). Academic skills are Average to High Average (WIAT-III Reading Comprehension = 120; Sentence Composition = 92; Word Reading = 114; Numerical Operations = 106). Patient is right-hand dominant. Patient has Low Average fine motor coordination (Purdue Pegboard T scores, Dominant = 40.35, Non-dominant = 49.07, Both Hands = 37.68). Patient was born vaginally at full-term (exact gestational weeks unknown). No pregnancy or labor complications were reported. Patient wears glasses to correct vision to normal. Patient has macrocephaly, but no other neurological problems are noted. Patient has a structural intestinal malrotation. He has had several surgeries, including gallbladder removal, an appendectomy and an orchiopexy (for undescended testicles in childhood). Patient also has severe sleep apnea, and uses a CPAP machine.	father of 14813.x10
14881.x3	SVIP	duplication	2	Patient is a 36-year old male. Physical examination reveals presence of a café au lait (location not specified). Patient also exhibits rapid and fast postural tremor. Patient currently meets criteria for macrocephaly (435 months: HC = 59.2 cm, z = 2.86). Patient has an above average height, an above average weight (435 months: height = 186.4 cm, z = 1.34, weight = 91 kg, z = 1.40) and a BMI indicative of being overweight (BMI = 26.2). Patient does not meet diagnostic criteria for ASD or other psychiatric disorders. Patient's cognitive abilities fall in the Average to Above Average ranges (WASI Verbal IQ = 107, Nonverbal IQ = 120, Full Scale IQ = 115). His adaptive abilities fall in the Average range (Adaptive Composite = 107). Patient is right-hand dominant and has Low Average fine motor coordination (Purdue Pegboard T scores, Dominant = 40, Non-dominant = 38, Both Hands = 38). Patient was born vaginally at 41 weeks gestation. While labor complications were endorsed, no specifics were obtained. Patient was diagnosed with recurrent otitis media (> 8 total occurrences) at 8 years of age, eczema at 10 years of age, and hypertension at 10 years of age. Additionally, patient was diagnosed with heart problems at 28 years of age (no specifics available) and intermittent pneumonia at 32 years of age, which has resolved. Patient was diagnosed with head injury/loss of consciousness at 10 years of age and macrocephaly at 15 years of age, but no other neurological problems are noted. No gastrointestinal conditions or sleep problems were reported	
CB_081113	Manchester	deletion	2	Patient has similar problems and facial features as proband RB_078381. Learning disability is milder but did not do well at school. Short with a small head size on 3rd centile.	father of RB_078381
RB_078381	Manchester	deletion	2	Born at term weighing 3.16 kg. Poor growth; now short stature 0.4th – 2nd centile. Microcephalic with OFC below 3rd centile. Heart murmur investigated by echocardiogram and found to be innocent. Dysmorphic facial features; long columella, prominent incisors with wide gap between front teeth. Broadish thumbs and great toes, prominent fetal pads on fingers. Learning disability with IQ of 65 (mild MR), attends special school. Poor concentration. Normal echocardiogram. Exophoria. Normal neurologic examination.	
SAL_703574	Salisbury	deletion	2	Referred ectrodactyly, ectodermal dysplasia, clefting syndrome; re-referred for array CGH with ectrodactyly on left hand, absent left foot. Mild global delay. Height, weight, and OFC all <0.4th percentile for age. Epicanthic folds, mild micrognathia, high palate, bifid uvula, turricephaly. Talipes of the right foot. Normal heart. Duane anomaly. Truncal hypotonia. Also have photos.	

Supplementary Table 7. Primers used in RACE and RT-PCR experiments.

a) RACE primers grouped by target region (GSP: gene-specific primer; 5' GSP refers to positive control primer)

5' RACE

target name	outer GSP	inner GSP	5' GSP
4.1	TTTGTGCCTGCAGAAATGCCA	GCCTCTGGGGTAGAAGAAATGT	GCGATTAGAGCGGTTCAAACAA
4.2	ATGGCAGCTCCATAGAGAGA	ACAAACACCTTTTCACCTGTGT	ATGTGGGAGAGTCCATGCAA
4.3	AATGTGCCCCCAAATTGAGTTC	TGGAGCAAAGGGCTTATCTGAA	GACGATTGAACCCAATGAAGGC
4.4	GTGTTTTTGCCAGTCTCTGCTC	TGAGGGTGTAAGCATTCGGTTT	GCTGTTCAGATGGGATTGGTCA

3' RACE

target name	outer GSP	inner GSP
6.1	CTAGCCATAACTTGGTTGCATT	CTTGGTTGCATTCTCTAATCCG
6.2	GGCTTTGAGTTCAAGGTTCTGAC	AGTTCAAGGTTCTGACTGACCA
6.3	AACAGAGGCGAAACACAAATCC	CCTGGAGGAAAAGGCCTTGAA
6.4	CTTTGCCACGGTGTCCTTCA	CGCAGATCATGCAGAACTACCA
6.5	CCAAGAACCGGAAAGGCATC	GGCATCGCCATTATCATTCACG
6.6	TATAGGGCCTGTGATGATCTGA	AGAGGAGAAGGATGAGACTGATGA
6.7	TCTGGGAATGGAGGAAGAGACT	CCTCAAGCTGTCCTGCCTTC
6.8	GGTGACCTTCTCCATCATCGTG	GATAACCCAGCCTTCACCATTC

b) Primers used in nested targeted RT-PCR of *HYDIN2* transcripts

Targeted RT-PCR

	outer fwd	inner fwd	outer rev	inner rev
SMRT cell 1	CATCCCTTCCAGCCTCAG	GGCCTTGGAAAATCGGAGC	CAGTGTGTCATTCAGGTTGG	TGGGTTCAATCGTCCAAAGG
	AGGACTACTGAGCAAGGC	GGCCTTGGAAAATCGGAGC	TTGAGTTCTGGAGCAAAGGG	TGGGTTCAATCGTCCAAAGG
SMRT cell 2	CATCCCTTCCAGCCTCAG	GGCCTTGGAAAATCGGAGC	CTTCATGGCTGCGTAATCC	TCCGAGCAAAGAGAGTGTCG
	CCCTTTGGACGATTGAACC	AGCAGTACCTATCGGATTCC	CTTCATGGCTGCGTAATCC	TCCGAGCAAAGAGAGTGTCG
	CCAAGAACCGGAAAGGCATC	GGCATCGCCATTATCATTCACG	AGAAGACTCCTAGCTCATCC	CATTGTTCTTGAGGATGAGAGG
	GCAACGTGGGAAAGATCACC	TGTGCATTGCCAGTCATTCC	AGTTCCATCATCTTCCAGCC	TGTATAGGCCTGAGAGCTGC
	GCAACGTGGGAAAGATCACC	TGTGCATTGCCAGTCATTCC	AAGCTCCAGACATCCTCAGG	TCACCAGGAACCTTTGCC

c) Primers used for tissue expression measurements of *HYDIN* and *HYDIN2*

Expression RT-PCR

GAPDH_fwd	TGAAGGTCGGAGTCAACGGATTTGGT
GAPDH_rev	CATGTGGGCCATGAGGTCCACCAC
ex46_fwd1	CATGTACTGGGACCGGAAGC
ex46_rev2	CCTTCAAAGTCTGGTGTCTGG

Supplementary Table 8. Fetal brain DNase I hypersensitivity samples with GEO accession numbers.

Sample ID	Tissue	Timepoint	Sex	Sequencing*	Project	Accession
DS16302	fetal brain	d142	F	se	Roadmap	GSM665819
DS11877	fetal brain	d122	M	se	Roadmap	GSM595913
DS11872	fetal brain	d122	M	se	Roadmap	GSM723021
DS14464	fetal brain	d117	F	se	Roadmap	GSM595920
DS15453	fetal brain	d112	U	se	Roadmap	GSM665804
DS20231	fetal brain	d109	F	se	Roadmap	GSM878652
DS23638	fetal brain	d105	M	pe	ENCODE	ENCBS493ZWQ
DS20780	fetal brain	d105	M	se	Roadmap	GSM1027328
DS20226	fetal brain	d104	M	se	Roadmap	GSM878651
DS20221	fetal brain	d101	M	se	Roadmap	GSM878650
DS14803	fetal brain	d96	F	se	Roadmap	GSM595926
DS14815	fetal brain	d96	F	se	Roadmap	GSM595928
DS14717	fetal brain	d85	F	se	Roadmap	GSM595922
DS14718	fetal brain	d85	F	se	Roadmap	GSM595923
DS23849C	fetal brain	d80	U	pe	ENCODE	ENCBS694XIX
DS24711A	fetal brain	d76	M	pe	ENCODE	ENCBS980LUR
DS24872A	fetal brain	d72	M	pe	ENCODE	ENCBS489VFT
DS24775A	fetal brain	d58	M	pe	ENCODE	ENCBS852UJL
DS23813A**	fetal brain	d56	U	pe	ENCODE	ENCBS539WGT

*pe:paired end, se:single end

**shown in Figure 1e

Supplementary Table 9. Molecular Inversion Probes (MIPs) used for *HYDIN* exon sequencing.

MIP sequence
cagatgggtttccttcaaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatacaggtaaccctggt
ccaccaagctgactgtgagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtgaccttctccatcatc
caaagaggggatgatatactctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaaggctgggttatcc
atccgagccgggtacagcataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagtcttattcgagcccag
caaacaaactcagctcacgtgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttggtctcacccaggt
gaacaggcttttccggaagtgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttcatgtgagggaaggg
ccagactgtccttggcagcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgaatttcagcccctga
caaggctccatagcaggttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaagttagtcttccgaaggt
gcagtgagaccccatctctaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctgagtgtagcagca
actcctacaatcaaggcctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctataacttcagcgatatc
gttctccaacttgatggaggctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacagctctcagggagcc
gccctactcggtgaccttctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaggtgacaaatgagttct
gcctccttctctttttagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctgaaactcacattgtag
ggtccggcttctctggtttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtggtgcctttgctatgta
gcatcactatgaagggccttgatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtacagctggcacttctc
ctttggggagctcagaagtcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctttcatctccccaccc
ccaatatctatcgtgaagtgccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtaaagaatgctgtgtg
actcaggcccctcccagtgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactatgccaggaaaccaatg
gcaaaacaagccctatgagatcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggtgcgctccaagc
gtctccagggccagaagtacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgattccaggtctggttgg
cctcggtgggatggtaggtcacttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccccgggcagcagcta
gctacatccagggaggcagtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaaaaaatttgagcctcat
gctactgagaacaagagactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaggtaatatagccttct
gaatatgttcctggtccagtgagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagggacagcagcctct
ggacccgtggtgtatcagacgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagaagcgtgtccctcc
ccagagtcagcagctgagggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgaacacttcctcagagaa
ggtaggacagctcccaacttcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaggagcccctctgagc
gggcaggttgtgaaaagaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcaccagagtagacatata
gttattggaatctccaatgtgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaggcctcccgactgtc
cctcgagaaagtatcaactatcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttacgcctcttcctctt
gaaatggatagccgggctcacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttggaggctctggaagc
ccttcaccagctacaactttgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttaagagccctctcct
attccaagtactgcagcaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttggacaatgttaaatgga
gccaccctgtcttttcaaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaattcaactttatcaacac
gcatttgtaacatgttgatgaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctggaagctgaaggtga
catttaatgtcacagggtggacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaaggtcttcttacacc
caaggccgagggctacactaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggtaatctgctgttttgtt
aacccctgctcctccttgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggatgcaaacaaaatgagac
caggacaactcccgctattctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtctgctgagagactcca
gctcgggaattaggaaggtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccaactctaaatacatacc
cctgctggtaggcaaaactacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagaagtgtcttttaccac
cagaaggagtaggtgctattggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaagccttgggttgagtg
gatctctgaagaaattgaaagtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggggccaggttctaaga
gccacttatgtagtccgagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaatcaatgtaaacctaag
caacccagagctccgagggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccttatcctctcaggattc
agcagaggctgggtctgtcttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctccaggtggcaggttg
agtagaggctgggtctgtcttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctccaggtggcaggttg
gaagggctgtggggaaccctcgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgtggcgagcacatttt
gtggagccctcttcgggaatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatcacccactccctgtct
cctgcaaatcgatcagggagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggccctgggttttact
aagatacctcaaaggcagtcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacagtatagatgacatgca
attctcgcttcttcacttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaattcaacaaatgcacga
caaggacaaaaatgtaagccacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgataggaggtaaccaagt
gcacatgacattatgctgtaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctctcactttgttctcac
aagcaaaggggtatttaaagcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacattcagacacaggg
actgatttgaagctgcagttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaagagcttttccatcta
ccaatgtggatttcgcttcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtttttctgtccctgtgc
acataatcctggagagcttgcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtgcttgatgtcccca
gcatgcacacagtcaagtgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgtgccaaggacatagtg
atattctagagaagggaaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggtacatctgacttcat
gcattatagctgtgtccaatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttaggactaagggattcc
agcttcacagtcacaaatcacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctccttaaggcagca
gtgatgtcatcctcatagccctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacttcactgctgtttct
ggacaacatccatggactggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgatgtcgttttccact
gctgcaaagagaagagagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcatattggttgttgatcac
gaaagctcacacagcctccttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcaccactggcctctag
cagaaggaaagcagggatgagtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagcaccctccccaca
ctgcatgttgacctgcaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaaaaaagtgaatactagg
catattggttatgaagaactggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctccctttcagcctcat
aaacccctgctcctctttaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgacaggctacaatcttc
gaaggacaccgtggcaaaggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgacactccctacagagtgt
ccccgcagatcatgcagaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaggcactgacacactct

gaagatgaacttgttctcatcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggggccattgggagc
gccaagccaaggctcgtttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgccttgatatttgaagag
actctggattctcacaaaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcactggtacatatctgg
atatcgatggctataaactcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaccaggtctcagatga
cactttgctagctgaagcctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccattccttccggaggac
gtgacagccagactatgggggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtcaacgttgatgacctg
gaagctctcactttctcgggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagctaattgggagttggc
caaaactggctcttccagagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccttgcaagaaaacagac
gagtgccacagatcaaagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatggaatgagttaagtaaa
aagcaccaccttcaccatagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatatttcagaaggaggtg
cctattttcatcattgacatgaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaattgggatgctggca
gttggactgagatcatggaatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctccctttggagctct
catgcaaaaaaggaagtgaagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttgtctttcagctttc
ccctgtttagccactacgtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgttgaaatccctacagag
ctcctgcagagaccaggggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttgatggccttacatctga
aagcagcccctgcaattgaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgaatttcaacaggcact
aacctggcagggaaagggaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccaaacatcctgtccttg
gcagaaaatcttcttggtgttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacctacaggggaatacac
agggatacagtgaaatcatcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgataaccaagtctgtgga
gtacggcctgcacctgtacttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggaatccagggtagttc
gttttaggttcaggggttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatccgccaggccacaggcag
gatgcagcagacaatgctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcctgacatgatctagaact
cagagccagccatcttccaattaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccttggcatccttggt
ccaccattaaggaagtgttcaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagccttgccctggtgac
ggaggcatccttctattttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttttccacctgcattctc
catcttccgctgaggaaatacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtctgaaatctgctttaa
caaatgaggtcatctttaagaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagcagggaatattggg
gtcaaagttcccgaactcatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtcaacctttctttta
atcctaggaacttgctgccagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNataaacctgtgtggtcat
ggaagtcctgtcctgttggggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgcacctgcaacgttac
ggcagctctcctgtcttctcttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagggaggggaatctgg
acagctgggaaatggttccagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagtgctgtgacagtca
agatacagacctggacaacttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaaggggggcaaacaga
cagaagacagagatggccacatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctcgaagacttggtct
ggtcgtctgtcatggtcaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagaataaactgctcatgt
atcccaccatccgaagatataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcctagacatcttgggtct
gtttaattgtctgcctttgcaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgggatgggtggtccg
gcccaggtccttctccttcttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctgacttgagacttccc
acaccagactttgaaggcttgagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcggagcgggagcg
cgctggtcgtcgggctcatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagggctcgcagcttctccag
cgcaaggaccgggagagagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgaaggatgtccagaacat
gtcctcttggtcctggacaagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcagctggactcctt
caaaggaccccgacaagcaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaaaatatctgttagggaa
gggtagtaaattgcataaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccttctcagacattgttt
cctcttccttttgtcgctccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctgcttctgtctcctgg
aaacggagggtcaaaaaaggaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcatgggggcccctga
ccttctctttctcaagagctccctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctggccagcccttggc
gtctccaaaacatggatgaggaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccacactaagacccta
ccttcagcaggcagaggaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtttcacttaagcccaatg
cattggcagccgggagcatatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtggccatagtaactgtc
actgggactaacactgagccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatggtgcgtgaatgccaa
gggctgatgagctgtgtgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggagttgttgagatacaagg
gtttaatttctggggccaccagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcggtagagccctgatt
ggaaagagtgttcgtgggagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaattcaccctctccttat
caatagacatataatgtgaaccacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccttgccctatgacgt
atggagtcgatgctcaggcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaacccaagagaatgggct
gtgctggaagctgtggccaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttctctctgggacactga
ccagttccccatattgaatgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcactgatggccatttgga
caggccacagtacatttatatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctaaaattttcttaggaa
ggatttttttctctaatattttaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgaacattcaagcag
cagggaaactgagcacagacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNataacatcaacagcttttc
ccgtaaagtggtattaacgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaacatcagctgtgtggtc
ggtcttaacctcccctcagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacacagctattaatgtata
gtaaactctgcggcttccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcattgacatttctgcagaag
cagctgcaaatttaaaatatatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttcttggccaggcgg
cttcagagcattgacagccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgccatttaaaaaaaacagc
caggttggatgtgacagagatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcttggagagaactagtat
ccctgttggtcgagtccaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgttttcatcctttgctac
gtactcgtacagttctgggggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattcagaggggcaaaggc
gagaatctggcacaggagaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggctatgattcctacaa
gtgggtctgctcactgatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttctccccagggttgtggg
gtgggtctgctcactgatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttctccccagggttgcggg
aagtagcagtggccccagatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaacagtttggtttgacctc
ggagacgaggccgaggtgatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaaaccctggtgtttcagat

cattacatgtcatgtaaaacaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggctctagaccttgc
cgcgtagtttctgtcttaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagccatatttctttaaag
gaattaaagccaaagacacggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggaaatgtgggttgtg
aagtccacttttccacgagagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaatgggccacagctc
gccacaattcagtgtggacagtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaccccatgacacatga
aatatttccgtttcacaatgaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccatcagcaagcagagc
gaagtgagatttgacccacaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctttctctagagcagct
aggatgatgtagccaaagtccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaagtgcaataggaaag
ggcgaagtccgaacccacatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaaaaaaaaaatcagactc
catagctttggactataagtcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctgaaacaagcacagc
cctagaacatcagaaaacaacccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttaggggtggtattgtc
gaacatttcatacttttcatttgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgtgcaggctggagct
gaatcaagccaggaaaaacacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaccttaccaggagctg
ccttttaaagacctcaggtactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggaactactccagtgtg
gtttccagatacagatcgcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaacctgtactgtagcac
atactcacctgcttcggagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttactgtgggactaaataa
cacgaccaggattacaagtgtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgactgaccaccagtcttc
gagtgactccatcttgggatgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagggagaaggttgtctgg
cagagctttagcttgtgcaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaaaatctctacccgatgg
gtgaagtggaagaaggacccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtacagaatgactttctctg
aggtagcccagcaccctgaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggcaaaatatcaaacacct
gctctggcttgagtatattctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgctcagttagaaagtaa
ggaatccctacagacctgtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattttctgtcaattgatga
cctaggaagtttaatgaaatgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgggaaggaccctga
gtccttgtggatgaagatgccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaatccccctacgtt
gatggactccatctgggacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaagaaaagaaaagctacta
gccagcccaccagcaatcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgcaaaaactgatgtctc
gatgcagccaaaatccagctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgagagttctgtgctggg
gaacgatactgagctcattcgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggtggcagaagaaattc
gctacaaacaagataatggggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctatctgagggtgttcc
gtttgacccttcctacagaaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaaatacctggctgtcc
caacaagttcacgggcagggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcttcctttatcacaccc
gtgagactgataaatccctgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtttcctctttttctgttt
caccaagctcttcccccttctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctgaccctagtcgaag
gcacctctcatccagctctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatagttgttccccaaatgc
gatgaaacacggggctctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagtagagatttatttagga
gtacccaggccaggcaattgatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggccaagctgagtaaga
gctggaaaagacgaggcaggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcatgtcccttcttaaccc
actttaagttgatcaacaagggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattgttgctcaccatttt
cctgaacaggaatccgataggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctcagcatatgtaacata
ccggaactggttccactattgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcatggttcctccagaa
gttctatagatgtcaaagccctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtcagtgccagttgaa
attgggatggacgtagataactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagctgtgtttcatcatg
aagtggacttcgggaatatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttttattgaaggaatggct
gagatgaccccgctgggggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgcagttctgaaaatggac
caagcagcaccatccacataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttcaaaataagcaaatacgc
gtattgaccaggtggagggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatactgaagaaggagaaa
gaggtggactttgggcactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatgtgacactgacaaatgt
gaggtggactttgggcactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatgtgacactgacaaaggt
ccagctcgtatttctgcacagtgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgggaggcagcccat
ccagctcgtatttctgtacagtgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgggaggcagcccat
cagcaaggtattgctccccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctttttcgaagaccaacttg
atatttcttccttggtccaagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatttgaaggctcagctg
cctgactgtggcaccattcgccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactgcgtatccctgggga
gagaaataaaagccccccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacaatatgaaatgctttta
atgacacagcctctggaaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtaaaccttcctccttc
gttccagctctgcactttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggtatggtttttgtactta
atagcttggactccacttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgttagtttctacattcat
cctgggaaactttgaagaagagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatcgatgctctcttca
gtactgtttaacagcaacaaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagttggaggggtcat
cccatgccttccccctttgattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaacaaaccccaaaagccaa
gattcacttcaactttgaattgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtaacgtacatgaggg
accttcctgaaaaacatgaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtaacctgagtttacca
gccaagctctatcaacagacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgtaataacacttacaagc
aagatgttctcggagtaaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatttcctgtatggtcag
gttctgtcccgaacctttgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtacccacatgaatctcac
agatatgtaggttttctcgatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtataggtgatttttgtta
ggccaatcagcgaactataacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcttgaagtgtgtttgtac
gactctcccacattaagagttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctcatctgacctttctt
gcaactggaagtggagtttgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaactcctgtctccttt
aggaaaatcgagaatggctcgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacaagagctaaaggcatg
ccacttgtcctgtcaaatacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatcataccctccttg
gatgagaggcattttgcaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgatgggtacaataaactt
actttcttccacaactttcaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccaatttcttgccattc

cagccccaaagatattggccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtagttgatggatttcagt
caattcctgaaaacttcagttgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaccccagtgaacata
atcaggcattattccagcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatatttcccatgtttagttt
ggaactctgagggtgtaagctagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagtacgacagagagca
gaaggaaatgtccctgaccacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaattatcctgaaagctg
aatcctttgaacatattgaccaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaactgtcattctcct
caaagcaaggttttgccaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtatttgttctaagacat
gtgtgagtccagaacagataaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaacccaggacaggtaat
ggacccatatgagcctctgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccatatttcacaggctgg
acctttagtcccggctacttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttccccagcacattggg
ccatgcacagctaatcttggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagagtctgactctgtcac
gaaaactgacacattttacatgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcttgagcctgagaag
agccgcccgcactctccatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacaatctctgcgaggacct
ggcggcggggtctgtgagttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgcctcctccaatcatcg
gggctcgtaggtttgaattcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatttgtttaataaatgagcg
actccgaagcaggtgagtattcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggctttgagttcaaggt
caggactgacagcatcaccttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggagatggttctgaggagt
ccatagagccaagaggtccctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaaagtatgtaaccagaa
aaggtagttgggttgagaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggataaaaatggtatgct
gccccaggcaggtatgagaatatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtgggaggaagggtgc

Supplementary Table 10. Molecular Inversion Probes (MIPs) used for *HYDIN* paralog-specific copy number genotyping.

MIP sequence	MIP set(s)*
cgctttttcttaggcgatcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacttatgctatctagacatg	*HYDIN* full set, *HYDIN* reduced set
cgcctctgttttctccctacaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctgagccctgcattcc	*HYDIN* full set, *HYDIN* reduced set
gccagtgatgtggtctatctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctagagtgtagttcagag	*HYDIN* full set, *HYDIN* reduced set
aggctgtaggattgaaagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattctctcatattgagcagc	*HYDIN* full set, *HYDIN* reduced set
gttgctctaagatcgcttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaacctctatagttaaaag	*HYDIN* full set, *HYDIN* reduced set
acagcccaggggtggtcagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggacacagacagagggtg	*HYDIN* full set, *HYDIN* reduced set
acagggtctaatctcttgatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaggacttgcagttcta	*HYDIN* full set, *HYDIN* reduced set
ccatgctgtttctaaggcttcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaaagtgcacagaggt	*HYDIN* full set, *HYDIN* reduced set
agacttctctagcctgggttctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaccctggattgctggg	*HYDIN* full set, *HYDIN* reduced set
gcagaggtacaagcaggtgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaacaaacccaagttactct	*HYDIN* full set, *HYDIN* reduced set
gatcaggtgttagttgtggcaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttggcaacacagtggtc	*HYDIN* full set, *HYDIN* reduced set
ccatgtagagataatgaaatgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaatgagagccaacctg	*HYDIN* full set, *HYDIN* reduced set
atttaccccaaatcatgttcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttcccaatatccagcaa	*HYDIN* full set, *HYDIN* reduced set
gaactatagagtagattccaaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcctgactgtcagtg	*HYDIN* full set, *HYDIN* reduced set
acaacatagtgcatttaacccaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcaacagggctgccat	*HYDIN* full set, *HYDIN* reduced set
caacctatttagccaccctatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagcgtaagcataccattt	*HYDIN* full set, *HYDIN* reduced set
atccatgtgattacaagtatttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcccagttctccga	*HYDIN* full set, *HYDIN* reduced set
gtgagctgggcctaatccacatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaacattccccaggtgg	*HYDIN* full set, *HYDIN* reduced set
aaatttagcctcagattcagaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacaaagcaaagggaac	*HYDIN* full set, *HYDIN* reduced set
acactcagcatgccctgacccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttcgatgttgctcatcgtt	*HYDIN* full set, *HYDIN* reduced set
cccttgatgagaaagagaaatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttcctgaagtacccgta	*HYDIN* full set, *HYDIN* reduced set
caattgcagttctgaaaatggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaccggtattttaccat	*HYDIN* full set, *HYDIN* reduced set
aacacagcgagggccatgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagagagaaagaacacagga	*HYDIN* full set, *HYDIN* reduced set
ggaagggctgacattcaccctatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctgacccctctggctg	*HYDIN* full set, *HYDIN* reduced set
actctgtcttctccttggggaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagagtgaaaattgtggct	*HYDIN* full set, *HYDIN* reduced set
agtgtgggcttgagagtctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctggtcacatagttcttttc	*HYDIN* full set, *HYDIN* reduced set
gcccccaagagccaactagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgattaatgtcagtcacatc	*HYDIN* full set, *HYDIN* reduced set
acttaccaggcatgatgtcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccataaactaccaagtaa	*HYDIN* full set, *HYDIN* reduced set
cgagtatcacatcaattgcctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaccatatccttcctcag	*HYDIN* full set, *HYDIN* reduced set
caactctgattccctgcttttattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcactctgggctgcctc	*HYDIN* full set, *HYDIN* reduced set
gcatcctcagttcaacaaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggaagcaaaaagaacttt	*HYDIN* full set, *HYDIN* reduced set
caatgcctaaaaaagccaactcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaagcccagtggaaaga	*HYDIN* full set, *HYDIN* reduced set
ggaggggagagaccaccatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgttatcaagctgatcagt	*HYDIN* full set, *HYDIN* reduced set
actgattgtgggagaaaatatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgttctgagaggtacaa	*HYDIN* full set, *HYDIN* reduced set
agtgatgttcgtaaaaagggctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccagtatcagtggagg	*HYDIN* full set, *HYDIN* reduced set
ccagcccatctgtttccatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggaagtattctgtggtat	*HYDIN* full set, *HYDIN* reduced set
attccgaacctcagctgagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagcccttattttggcatta	*HYDIN* full set, *HYDIN* reduced set
agatgagcccaggaggtgaggtatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacagtgtgtggggtg	*HYDIN* full set, *HYDIN* reduced set
gaggtgatcagccagtgtttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccttatggacctgaagtg	*HYDIN* full set, *HYDIN* reduced set
gccctcatctgctcctcttttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctggcaggtactcagaaac	*HYDIN* full set, *HYDIN* reduced set
catttactcacctggttgatgaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagggctgctcggagag	*HYDIN* full set, *HYDIN* reduced set
cagaatgagccaggaggaatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgtttatgtcctgtgc	*HYDIN* full set, *HYDIN* reduced set
ggagaagagaaccggcttccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccattctcagaaaagaagta	*HYDIN* full set, *HYDIN* reduced set
ggtccactctctgctgtcagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatgagtttgcttcagtgg	*HYDIN* full set, *HYDIN* reduced set
gctgcaaatagtctttcttttaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtgggtggtcaggtga	*HYDIN* full set, *HYDIN* reduced set
gtttgttccctaacagatattttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttgcttgtcggggtc	*HYDIN* full set, *HYDIN* reduced set
gtatttgtcaggaaatgcctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctccaccttccaaatcat	*HYDIN* full set, *HYDIN* reduced set
gttcaaagaggcacatgtttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagggaatttgctgatga	*HYDIN* full set, *HYDIN* reduced set
ggtacaagtcatccttattcccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagccaccatgccactg	*HYDIN* full set, *HYDIN* reduced set
gttgttcagattgaaaatatcatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaaaacgttccctgcc	*HYDIN* full set, *HYDIN* reduced set
gggaatagcgtgtctgggaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcgttcaggtaactggaat	*HYDIN* full set, *HYDIN* reduced set
atttgtgtatcagaatcacgtttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgaggtgcgtgtacta	*HYDIN* full set, *HYDIN* reduced set
ggtttctgctccttgtaactccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttctaaagggcggactt	*HYDIN* full set, *HYDIN* reduced set
agtcttccatatcattgtcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaggagagtcagtccta	*HYDIN* full set, *HYDIN* reduced set
ggaagccttccttgactctcttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccttcattctcgcttc	*HYDIN* full set, *HYDIN* reduced set
gtggttttggtcctggctgcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggttttgcaaagctgac	*HYDIN* full set, *HYDIN* reduced set
cccatttggctctgcagacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatctataatctcccacaag	*HYDIN* full set, *HYDIN* reduced set
agggcttctctggtagatgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatcttgaagacagaaataaa	*HYDIN* full set, *HYDIN* reduced set
caattcttttgggatttggttgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagtttcctgcccaggg	*HYDIN* full set, *HYDIN* reduced set
ggggtcctgggcatgtcaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattctcaatgctccatattc	*HYDIN* full set, *HYDIN* reduced set
gttggtagtcaaatatccttcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcagatgatgttggt	*HYDIN* full set, *HYDIN* reduced set
gtacaaaggcttcaacaatggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtccaaactccaaaagt	*HYDIN* full set, *HYDIN* reduced set
ccagagtgagtgtggaattgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaaaatctcagggcacc	*HYDIN* full set, *HYDIN* reduced set
ccctttaactaatgaactcttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagtgggcacttgaggc	*HYDIN* full set, *HYDIN* reduced set
gctgtgtgaccttaagtacacgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagaacagaggcaact	*HYDIN* full set, *HYDIN* reduced set
atcaggtagggaactctgcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcatcagttagctagttct	*HYDIN* full set, *HYDIN* reduced set
gtttgctcttttctttttccttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacattccttggctcat	*HYDIN* full set, *HYDIN* reduced set
ccggagacaaggcctcacaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaagaaaaactcataccatt	*HYDIN* full set, *HYDIN* reduced set
aagcttttatcattccccccctcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcacaaaggaaatgc	*HYDIN* full set, *HYDIN* reduced set
ggaaagggatagcaccctcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagacttaaaatatgtatgc	*HYDIN* full set, *HYDIN* reduced set
gtcctcaaacactgagattgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacttaaacaatgcaacaat	*HYDIN* full set, *HYDIN* reduced set

gcatttggaatgacacccataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctttcactatagcagat *HYDIN* full set, *HYDIN* reduced set
aaatgtgtacggcaaaaggttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtagcattaaaatgcaccca *HYDIN* full set, *HYDIN* reduced set
ccttaaagcccatgccaagtctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccatgctgaccactctg *HYDIN* full set, *HYDIN* reduced set
gttccaggccaagtgtggacatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgaagatttacgtggtac *HYDIN* full set, *HYDIN* reduced set
atcatgtctgttttgtccatcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaacaaagctgctatttct *HYDIN* full set, *HYDIN* reduced set
cccctcccacaaattttagtcctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctggaggccctctt *HYDIN* full set, *HYDIN* reduced set
ggactaagatgaaagtattgaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcacttgcctcaaagg *HYDIN* full set, *HYDIN* reduced set
gcttttacctgaattggttattgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagatagcatctcccag *HYDIN* full set, *HYDIN* reduced set
gtccattcctttataggcagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttttgccttcatactcact *HYDIN* full set, *HYDIN* reduced set
gcacaggcttgggtcagggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatatgtaggtggcaaaatta *HYDIN* full set, *HYDIN* reduced set
gatagggagactatcctgcatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaagcacagaagtggcc *HYDIN* full set, *HYDIN* reduced set
catgagatttaacccacgtctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctatgttgctacttttgc *HYDIN* full set, *HYDIN* reduced set
gtgcactataactgttggctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtttatgtttttacgtgc *HYDIN* full set, *HYDIN* reduced set
gaaacaacacaaatgtccaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacatcctgtatgtcctgg *HYDIN* full set, *HYDIN* reduced set
atggaaaatcaaagggccaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaagattcaatgtcatcg *HYDIN* full set, *HYDIN* reduced set
aacagtttatatgaaagcatggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctagtccagtcctctct *HYDIN* full set, *HYDIN* reduced set
aaacatcaccaaagtgcagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatgtccttgtgctggact *HYDIN* full set, *HYDIN* reduced set
ggtcaagttttctgccctaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagagcaatatttgaagaga *HYDIN* full set, *HYDIN* reduced set
agctgcacataaggctaggaaataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccagcagaaaggagg *HYDIN* full set, *HYDIN* reduced set
gccttatgtgcagcttggttgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcccagatttgctaagt *HYDIN* full set, *HYDIN* reduced set
cattgaatcatgcttgctgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccacatataaaacaagccc *HYDIN* full set, *HYDIN* reduced set
acaatcaatagtgcaaactgaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctagaaaagaaggtgt *HYDIN* full set, *HYDIN* reduced set
caagctgcctctactggtttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaaggtatagcttcaggaac *HYDIN* full set, *HYDIN* reduced set
ggaaagctgttgggggaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagttgtcttgtctccttata *HYDIN* full set, *HYDIN* reduced set
ggcacattcttgtaaagatgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaacatgggggtatggtt *HYDIN* full set, *HYDIN* reduced set
attctgcctgtcacaaccccataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagctgatttgggacct *HYDIN* full set, *HYDIN* reduced set
acctaagtttgtcaggaggttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaccaggattttggtgac *HYDIN* full set, *HYDIN* reduced set
actcaaagagaaggggtcatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccattcataagccacgg *HYDIN* full set, *HYDIN* reduced set
gtcactgggtggacagttctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaatagaatggcctctct *HYDIN* full set, *HYDIN* reduced set
cagtgtcataagacaaagtctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggactactgaaaagtgg *HYDIN* full set, *HYDIN* reduced set
actcccaaattggtggctgttttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtgctcctcctagcca *HYDIN* full set, *HYDIN* reduced set
acatttttatttttcttattgggccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgactcggacaccagc *HYDIN* full set, *HYDIN* reduced set
ggaaacagaatcaggggccacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgagctcaagtatttgcg *HYDIN* full set, *HYDIN* reduced set
gccttttaaatcaatgaaacaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatctaggagcttggct *HYDIN* full set, *HYDIN* reduced set
acatggttagagtggacagtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgatagttccacagatgtg *HYDIN* full set, *HYDIN* reduced set
gttctttttcttcaggcctggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagcattgtactggccac *HYDIN* full set, *HYDIN* reduced set
ggctgggaagaaatgtgggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtattaatttttaagccccct *HYDIN* full set, *HYDIN* reduced set
cacaaggcaatactatgaaattcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacctctgcacttgcat *HYDIN* full set, *HYDIN* reduced set
caaaacagaaaaaaaattcctgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctcccctcctgaggt *HYDIN* full set, *HYDIN* reduced set
gcctatgaatcacctgggagttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtttgctgactcatcac *HYDIN* full set, *HYDIN* reduced set
atccatcgtcctgtacctgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggagagttagtaacttcag *HYDIN* full set, *HYDIN* reduced set
ggggccgcaaaccattgtaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttcattccctttactgct *HYDIN* full set, *HYDIN* reduced set
gggcactctcaggaacatctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggagagtcttttcattag *HYDIN* full set, *HYDIN* reduced set
gggcaactccaaagggccaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaattgacccatttattta *HYDIN* full set, *HYDIN* reduced set
gtaacatatttgcaggtggaaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctctgactccaaccct *HYDIN* full set, *HYDIN* reduced set
ccctgtccagcagggctctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctacagtggggaaaattaac *HYDIN* full set, *HYDIN* reduced set
gtggagcttacagggcttctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatttgaacaaggtgttcc *HYDIN* full set, *HYDIN* reduced set
gccttactcagtgttactcaaagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagatgctcccaagca *HYDIN* full set, *HYDIN* reduced set
atgcaggagagtcagtcagagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttactcagaacctgtgac *HYDIN* full set, *HYDIN* reduced set
agtaagcatggatatgtacaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagttgttcacccatcga *HYDIN* full set, *HYDIN* reduced set
gaatgaagattggtgggtctttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagagatttaggctgc *HYDIN* full set, *HYDIN* reduced set
gtctttgagtttccagactgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttaagagataacaggcttt *HYDIN* full set, *HYDIN* reduced set
gtcttccatgtggctgttgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttactcaagttttgcttgt *HYDIN* full set, *HYDIN* reduced set
ggagtttacagcccagttgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaatagatggaaaaactctg *HYDIN* full set, *HYDIN* reduced set
gctgatggttatctctcttctggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaggctctgtcagaggg *HYDIN* full set, *HYDIN* reduced set
ggagtctgaagatagggcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttataatactggcttggta *HYDIN* full set, *HYDIN* reduced set
gagcagggcagggtctagtctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctcctcttgagccttaat *HYDIN* full set, *HYDIN* reduced set
caccaaatttgggaaagttccatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccctgaagcttagaga *HYDIN* full set, *HYDIN* reduced set
aagtaggcactcagcagaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacaaatacccttttttccaa *HYDIN* full set, *HYDIN* reduced set
gggttgagctcagaattcttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacttgaactgattctttga *HYDIN* full set, *HYDIN* reduced set
gtcctggtgctcctggcgtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatacacgaaatcaccagta *HYDIN* full set, *HYDIN* reduced set
ataaatggaaaatcgactttctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccttgaatgacagcaga *HYDIN* full set, *HYDIN* reduced set
gtgtgagtgttctgctgcactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgcattgacaggactt *HYDIN* full set, *HYDIN* reduced set
gtgtgtcactcaaaggcccataaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaacctccagcaatctc *HYDIN* full set, *HYDIN* reduced set
gtgcagcttcacactcccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttctaccttctttcttgca *HYDIN* full set, *HYDIN* reduced set
gaggtcctcaggccccagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattcaggataccaagagagg *HYDIN* full set, *HYDIN* reduced set
agaatgaaaaatccttttaaggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacatgttctcctccacc *HYDIN* full set, *HYDIN* reduced set
gcaattttcatttgcttctgccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttctactcatctttgcc *HYDIN* full set, *HYDIN* reduced set
actgtttggagggctcaggagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgtctttatcagcaacg *HYDIN* full set, *HYDIN* reduced set
gaagggagcccacttacaggtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagtggtccatggtgctg *HYDIN* full set, *HYDIN* reduced set
gcaacaactgatgacaaaataaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcttatgctcagccca *HYDIN* full set, *HYDIN* reduced set
ccagggacatgtggagtaagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgacatgtgggcaggtcag *HYDIN* full set, *HYDIN* reduced set
ccaccaagtatcttttaaacctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcaaggtcatcaagag *HYDIN* full set, *HYDIN* reduced set
attgttccccaaaatcggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttatccttagtgtccttt *HYDIN* full set, *HYDIN* reduced set

agtgcgtttgctatggcaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatgtttgtctcctccaaa	*HYDIN* full set, *HYDIN* reduced set
caatgtagggcctgattacagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgggcatccaatcag	*HYDIN* full set, *HYDIN* reduced set
atccagactcctctctgcctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctcatttaatcctcacaa	*HYDIN* full set, *HYDIN* reduced set
agcgtgcaggatagaaataagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttacgggtaccttctatg	*HYDIN* full set, *HYDIN* reduced set
gcagagtggaaatcgggcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaataaccattcgataaatgt	*HYDIN* full set, *HYDIN* reduced set
aaactcagtgaaaaggtaacccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaaagcaatttcaggct	*HYDIN* full set, *HYDIN* reduced set
gtgccaatatcagagcatcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtttttacttaaaaggc	*HYDIN* full set, *HYDIN* reduced set
gtcctggctgacaacttgattccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaggcccgcatagtgag	*HYDIN* full set, *HYDIN* reduced set
ccttactcctcctacttctcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgagacaggagaatagc	*HYDIN* full set
gaattaattgtaagccagcagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggtggtgtcatttacag	*HYDIN* full set
catgagatgacacagatgtggactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcccaaggctgaaagc	*HYDIN* full set
catgcactacactcactggggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttaccgactttccttcct	*HYDIN* full set
gctcacccccctcacctcttcagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtggtctggtgccccca	*HYDIN* full set
gagcatgaaaggatgcctcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcctgttagcctgggaaa	*HYDIN* full set
gctttgggtctacccagtcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctgcctatactctttgc	*HYDIN* full set
ccttcaagcccatgtattagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatgcttcatgtgtctgt	*HYDIN* full set
agttttgatgtgaaatacagcattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccagctttgcaactg	*HYDIN* full set
gttacaatggcaacagaaaactaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccctcaagcctccaga	*HYDIN* full set
gagggatagagaggtcacatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgattcactggatgagccca	*HYDIN* full set
gtgcagcctgattcctaaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgctcctatgagaatcta	*HYDIN* full set
gcaaatcccacaccttcataaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctgacttaatgactcatt	*HYDIN* full set
ccagttgagggcacatcctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttatttttgtgcatgctca	*HYDIN* full set
caggtatggaccactacacccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttttttttgagacatggt	*HYDIN* full set
gggcttgtttgccctcaggtacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtagacatgctggcccag	*HYDIN* full set
agaaaacataaagccattatttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccatcagcgttctca	*HYDIN* full set
gagaatttggagtcaaaagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagttagaatctctaagga	*HYDIN* full set
aaggcaaagtaacccttaccacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgtaaccaccagaactt	*HYDIN* full set
cccagactcaaagtcacctcttaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagcactgcacctgctt	*HYDIN* full set
gacctctggagccaaccgaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatctcctctcctttcactt	*HYDIN* full set
acaatgtctgcatcccaaattaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgggtttcctatttagga	*HYDIN* full set
ccccataaggcagccttctctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttgtgttcacccagtgag	*HYDIN* full set
gatgccctctctcctgtgaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcctgtagttcctctac	*HYDIN* full set
gtgggagagtccatgcaactggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtctgcagaactagaa	*HYDIN* full set
gagaacattctcatctgaccttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtgtacactgtactcag	*HYDIN* full set
catgagagttctagttgcttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctagcttaagaaaaggggg	*HYDIN* full set
gtgaagcaactagaactctcatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaagcacctgagaaggc	*HYDIN* full set
actataacccttccagaagtaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccatattccagaacaag	*HYDIN* full set
aaaaatttccttctatctctaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctgtgcattggaagta	*HYDIN* full set
agctataagcactctaaatcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggcttctgattagttct	*HYDIN* full set
caccacccctggctgaatgattaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacacgccatcacaccca	*HYDIN* full set
gctcaacctccttaattgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcttaaacaaacttgattt	*HYDIN* full set
acttctgctacttcctattctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagggagttaaagaaacg	*HYDIN* full set
gttttctctttggtttaagccaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctggtcttgaactcccg	*HYDIN* full set
gagggaaaattcagtaacttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttactctatgcctcatct	*HYDIN* full set
cctagattttgaagacaaaaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctaacttggttctggt	*HYDIN* full set
gattggcatcaatgtggaatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtctcaagctccttgacc	*HYDIN* full set
atcaaaatggcccaacatgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaccacttatgcctaaaat	*HYDIN* full set
cccttgctctttctattatttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcagccacagtggatgg	*HYDIN* full set
gcacgtcgtagactctttaccgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaggaacaatggagagaa	*HYDIN* full set
actggtatccttttttccctctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttcaaggacaagaattcac	*HYDIN* full set
gggcaaacttgtctgtacttgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctttcttgacccctgat	*HYDIN* full set
ggcgttattctcttgatgtctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagcagaatgttctttggt	*HYDIN* full set
gcattattgaagcgatcttagagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggtattcctggggct	*HYDIN* full set
cctccccgtgggaagaagttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattcgtttccagcagtttac	*HYDIN* full set
cagagacttggcatgggctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcagagaaattaatctct	*HYDIN* full set
catttgtaacattgatcacagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctaccacctgtgtca	*HYDIN* full set
cccatgtccacacttggcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagtaatagtccctcatctt	*HYDIN* full set
agggctggcctgctgttatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaagatgagggactattac	*HYDIN* full set
gttttaaggtgaaaaacaaagtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgaatcccctggtgc	*HYDIN* full set
acagatcacagaattagtatgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcactagagccttgatt	*HYDIN* full set
gaccaccccagggtggtggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgagtattgaataggaaag	*HYDIN* full set
gcactcttcaaaaaactcatcagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcaccaggggctcca	*HYDIN* full set
gtttcttcaggtggactcttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaatgagaagaactgcata	*HYDIN* full set
gccattgacttatgcaggctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaaaaaacagcaatagga	*HYDIN* full set
ggatgtctccaagggaaaagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaggttcttctctcaag	*HYDIN* full set
ggttcatgcggtccaagactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaaataggtcaagcccaga	*HYDIN* full set
ccaatgagcttctccttatgtttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaggctgatgactgc	*HYDIN* full set
gcgatctgaactgaagattggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacagacttcttggcaca	*HYDIN* full set
ggaaacattctctgtctccttttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaccacaacatggcagt	*HYDIN* full set
gtctctggtcttgacctatttaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggaagagacttggacca	*HYDIN* full set
gttggagaatgtttccaattggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctaacccattgtatgtgg	*HYDIN* full set
gtcaaataagcttaagaaatactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagggaaggagctacct	*HYDIN* full set
ccactggggtccatgttctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttccttgcaaagagggag	*HYDIN* full set
gctgttccaggtcagcagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagaaacacaaatgaaattc	*HYDIN* full set

caaatggcttgtttaagcttataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctccctgacctcaagaa *HYDIN* full set
aggtctggaaagaaataagaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaaaagggtgatgaaga *HYDIN* full set
gggtcattactccctttctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaccttctgccatgattat *HYDIN* full set
aagaactgccaagttgaagttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtttctgtcaagtttttc *HYDIN* full set
catcttagttatcactgacagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacaactgctcgtctag *HYDIN* full set
accttcaccacaccaggataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatagctgaaaatatttgtt *HYDIN* full set
aaagaattgtcaaaatggccttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccgcgtcccccattc *HYDIN* full set
cacagaaggatgctttcaatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcactgagcccgagagata *HYDIN* full set
gccagttgagtctagccatgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagagtattaacccaccaaa *HYDIN* full set
aagtcctaagcactttgatgtaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtgtagtcagtagcag *HYDIN* full set
gtgtctatataaatacagcccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcactgtctgcccaatg *HYDIN* full set
gaaacgtggtgcctccaccaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaatgaaggaggcacaagg *HYDIN* full set
gactgtctggagctggagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaattggctctagatttttat *HYDIN* full set
aatctataaacagtcccgccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgatcgatgttatttttgt *HYDIN* full set
gtactgggaacacattgggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgggctgtatttttataga *HYDIN* full set
atgaatctggagacattgtgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcacaatggccaagctgt *HYDIN* full set
gtatgatatgttgggtacttgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtggttaccaggcatc *HYDIN* full set
gtcaccatgcactacattaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgattactgcagtcaaggaa *HYDIN* full set
ccctctcacttcttcctataacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggcacccagagcatg *HYDIN* full set
cacagatcagcacgactggggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaaactttaaagtgcttt *HYDIN* full set
gattcagttacctccacatagtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaccttcttcacaaggc *HYDIN* full set
caggacttctcaatggacctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaatccaggctcatcttat *HYDIN* full set
caggctggtcattttattttgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctatgtttcccaggacca *HYDIN* full set
ggaaaccttttctgcattttaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttccaggttgaaggaga *HYDIN* full set
cactgcagagccaaataggttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttggtgcctataactaggt *HYDIN* full set
aatcttgttctttgacaacaaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggctgcggagagcta *HYDIN* full set
gtgaagcactgaactttagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgtcttgaacaggttcaa *HYDIN* full set
atctgtacaacaaaaacaaaactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgagaacacacggaca *HYDIN* full set
attccgtctttgctattgtgactaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctccagcttcatcc *HYDIN* full set
cctacctactactctagctttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatggaaaggcttttcatg *HYDIN* full set
gtctactttcctacctgcaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaacatgttttcaggtac *HYDIN* full set
aaagggtgaatgttgccttataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaaagatggctacatag *HYDIN* full set
atttgtttttccttcggcatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatttgtgcatctattctg *HYDIN* full set
acatttacgaaggtcaactgtatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaggagttcaaggctg *HYDIN* full set
caattgttcctcctacctttgtcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctcttgcctcagcct *HYDIN* full set
accaggaaatgcaaaccatggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaaaacatagagcagtcct *HYDIN* full set
acacctaggaaagcaaggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggaattagcacaaccattc *HYDIN* full set
ccttacttatctttttggcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctgaacttcagtctct *HYDIN* full set
gccggagggagttatcttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttcttgggagagttgagaaa *HYDIN* full set
accagccagggacttgggagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaaaaactggtggaaacct *HYDIN* full set
agatctcagctctttgggaaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaagtgcattatcacgac *HYDIN* full set
actgtgatcacctgaaataccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagactaacaacagtcaca *HYDIN* full set
attgcagttccctcggtgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaaagaatgtcttttgttt *HYDIN* full set
attcggaggggcagacggattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtactgtttgaagtatca *HYDIN* full set
gccaacatgccctggtaattttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagacagagtctcactctg *HYDIN* full set
acagtaaaagaggaagtgcattcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgcttggccacccc *HYDIN* full set
cccacaaaaattagttgggtgtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgactacgtgtggtggc *HYDIN* full set
caatgaagggtgaagcctggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacagagaaactaagtagat *HYDIN* full set
gaggaagagagtaattctagtgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaacgcggaagctga *HYDIN* full set
cctggcttccaaagacaccacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcatgtaaagagctcatg *HYDIN* full set
cagattctactggatggatgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgagaattacctaaacct *HYDIN* full set
ccctttgttgctgttaaacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttcatgttaaagaggaga *HYDIN* full set
cagtcccaaggaaggcatcattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtcagaggcccattgac *HYDIN* full set
atcctgtcttggatgctgatgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaagagttcctggtcaa *HYDIN* full set
caagctgaggtctcccttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcactattatgtctccaag *HYDIN* full set
gtgattgcagggttatttttttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcacacacgttcac *HYDIN* full set
gcagcaaagataaaaccttaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgagggcctttgata *HYDIN* full set
caaccttgcttgaaggaggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttagggaatagtttaaagaa *HYDIN* full set
agatgtgtgtttataattgccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgtgatcttctgcttag *HYDIN* full set
gatctcccaggaatctgctcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttcctgggtgaatttttt *HYDIN* full set
atgctaaagaaagtaggaaagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcagccaggaattggga *HYDIN* full set
gctctgaaggtaaacaagataacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgctgtggccattcc *HYDIN* full set
caaagccaggttgagttctcagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgttggcagggctgtgtt *HYDIN* full set
atgacatgatggtctatcacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaatacaaggtggggga *HYDIN* full set
atcagctgagtgtagtggtacgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcatagtggctcaca *HYDIN* full set
gcacccccaggccagtgcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctctagggagaatggagtc *HYDIN* full set
gatacatggcacccttcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacagaatattctcatacctg *HYDIN* full set
gttttgaggatgaataagagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatggaagatggttaaggg *HYDIN* full set
cagctaagctaattaccaaaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtctcaggtttgggaaat *HYDIN* full set
ggaaattaacacaaaattgccccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctctccacatcctcca *HYDIN* full set
gagaattgcttgctccggaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacctgaggtcaggagttc *HYDIN* full set
cctgatgcaaaatccatatcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatttctgggtaggtcatc *HYDIN* full set
aacccactatttttcaggagatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaattgggtggtgattgtc *HYDIN* full set
gatgctttatttgaacaactctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttacaatccggttttgc *HYDIN* full set

caattttgatacaccctgtgatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctgctcaagactgatg *HYDIN* full set
aatgcaagtactttacagttgtaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgcttgtcaagaagcc *HYDIN* full set
ggaaactgtagatggcttagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcatggaaaaagaactgtg *HYDIN* full set
acagtgcccggccactaaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgcccatccaatttttgta *HYDIN* full set
cactaagctctgattttctttcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagccagccagagact *HYDIN* full set
atccccatgacttacaaactgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagcttagcagtttcta *HYDIN* full set
ggtcttttgtagtccacatgctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtgtgtgcaggaaggc *HYDIN* full set
agacacaaatacattgggcaaagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctaagcagctgagc *HYDIN* full set
gtctgaaggaaagataaacctaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcgctgttagtgat *HYDIN* full set
cctcaggatgaggaagctgaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaaacaacaacagcaaaa *HYDIN* full set
gtcattgatttggaaagaaacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagttgcagagggaatca *HYDIN* full set
gaccctcaatctctcctagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcttccagtacatacaat *HYDIN* full set
gacaagttggtcttcgaaaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggggagcaataccttgc *HYDIN* full set
ccaccaaaggtgctttagaaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttcctgctaaaaccttt *HYDIN* full set
acacagtcaacaaatggctgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccagatctggacactgat *HYDIN* full set
gtgtatgtcttcctaagattcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaggtcacagtgggaac *HYDIN* full set
aataggtcttgaaggaatactaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtagggaaaggggctag *HYDIN* full set
gagagccagcagtgacagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaaaatcacaaactttgcgg *HYDIN* full set
caatggcaaagggtaggatagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagaaatgggactctgag *HYDIN* full set
agtaccagctttgaatttgcagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaggggagttgaagc *HYDIN* full set
gactggtccaaagtgtcacagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcacagtgccagttctgg *HYDIN* full set
ccatgccaaacagggccaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggactagaagcagaactcc *HYDIN* full set
gaggtggactttgggcactgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatgtgacactgacaaatgt *HYDIN* full set
cctgtgaactaaccatgcatggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggattgctagctgcatg *HYDIN* full set
cagcagctatcaccttccttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatatgattggatttgcagat *HYDIN* full set
ggggtcaggggaagttgtttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagacagaactagaagcc *HYDIN* full set
gaatgaatgaacaactgtccagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcactggacagcaccat *HYDIN* full set
ccttaggaaaaggaaaaggtccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccccaattcggccaaaag *HYDIN* full set
atgtttatccttcaactcagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatcctagtgttttacgc *HYDIN* full set
catcccttgagctaaggcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtacagttcagtaggggagc *HYDIN* full set
caatggggaaaagattctctagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctccatccatgttgct *HYDIN* full set
gacatttagacctgctagtcttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaggaggagggagag *HYDIN* full set
ggcacctctggtttgaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgattagtacaaggaagaat *HYDIN* full set
catctgaattcctgtttcctggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccagtggctgagggtc *HYDIN* full set
cacttacaagttgtggcaattaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccgtgctcctttcaccat *HYDIN* full set
acttcattttatgcagaggggtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaaaggagcacggttt *HYDIN* full set
catctgtggagggaaatgatgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgactccagggccttt *HYDIN* full set
aacacatacccactcccctTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgtttctgatgggtttta *HYDIN* full set
cagcacgtgctaggcatgaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaagctgagtgatggctg *HYDIN* full set
gatgttgccaaattactctccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgctgcagtgaacagaa *HYDIN* full set
aacaaagaaagaaagagaaagaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggctgaggtgggagga *HYDIN* full set
ccatttatctatctcagtgtcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagaaagccattcgtcc *HYDIN* full set
gtatttctgtgctttttgtggttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgccagggattagtgg *HYDIN* full set
gcactgtggggtgactggttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatcatagaagtagagagta *HYDIN* full set
accccacctgagacctttttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagccatctgaagtgaccc *HYDIN* full set
atgctggtaatggggaagaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgattttgtactggcttaa *HYDIN* full set
gagaaaacgaaacagaggtagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacttattgttgcatgcaa *HYDIN* full set
ccctcccggatgcatcccaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcctcactttgatggtac *HYDIN* full set
agggttcaaggcctttctatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaattagatatacttgcct *HYDIN* full set
agggaaatagtttggaatgacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaagcagtcctaaatgc *HYDIN* full set
actagaacagaccctcaattcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgtccagggagataaaa *HYDIN* full set
gatgcccaactaacagtttatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagagggagagggagag *HYDIN* full set
aggtgggaggattatttcagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgttttggtttctttctt *HYDIN* full set
aactcctgaacaaactgcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcagcatatttgatgaacct *HYDIN* full set
acagaaacactagaggaaaaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcgaggcacaccataa *HYDIN* full set
gcaaacactaaactcagctgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctttttttcttgcccact *HYDIN* full set
gtactctactgctacttcttaagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaaaaagcagcagccc *HYDIN* full set
gctgagtaagggacacactcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcagcaggtcccagtaag *HYDIN* full set
ccttaaatttgttcatgtgacctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcagagaagaagctggc *HYDIN* full set
cctctgccccccgtccctatcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgggtggttaagggaa *HYDIN* full set
cctgctggtaccctccccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgttaaatgttaacttgaag *HYDIN* full set
cacttctgccttttgtggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagattcttgcagatttcaac *HYDIN* full set
atcaccagtggtcacggaattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaacctataaagctctggct *HYDIN* full set
gagtcgtttgagcagctcgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaagagaaacctaatacag *HYDIN* full set
gccttaagccgctatatttgtgatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaagactcccgaaag *HYDIN* full set
gatgagaaacctgagaccaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgatgacgatagctttcg *HYDIN* full set
ggtggttgacttttagggctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaaccagacagcaaaatat *HYDIN* full set
gaaaacataaatcaatgccattttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggggcagctagagaga *HYDIN* full set
gtgaaccaacaaggaaactttggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctgcagttcccaaa *HYDIN* full set
aaccaagagttaaatatgtcagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggccagtggaagtga *HYDIN* full set
ccagctggccagtaagagaagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggcccacatgaattcca *HYDIN* full set
gtctggttactcaagcatgagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagcttctgttgtataaag *HYDIN* full set
atatacttcttctagcaaaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaagaagccctccaaa *HYDIN* full set
agctccctgagtttggtaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagatgtggagagtatttag *HYDIN* full set

ccccattttgattgctttgtcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcatgctgggtttaatac *HYDIN* full set
acctcattcaaaaggatgttttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgtccatgtgttctcat *HYDIN* full set
aatctaaagaacaagagcaaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgaggtccaggtgttg *HYDIN* full set
gtcaaagccctctacacttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagttgaacatcagtttct *HYDIN* full set
catatgctgagcatcacctctactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgttcaggcttccgg *HYDIN* full set
gtcttcttgttgctgacatacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtacggttgagagtctaaa *HYDIN* full set
gactcccatagtaccgacaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttagccaggatggtctcg *HYDIN* full set
actgctaggattataggcgtgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgggactacacgcacc *HYDIN* full set
gtggggattcttgattcctcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaggttgaggctacagtg *HYDIN* full set
ggttgaattgcatgcttttagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggagatgtgactgacatt *HYDIN* full set
ccgcttctaattgcatatttttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcacactacaagggaaag *HYDIN* full set
caatcccggtggtgttctggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttttttttagcaaatagcc *HYDIN* full set
gggacattccactcagtcactaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcattagtctgagatgggt *HYDIN* full set
gcctgctggggaagacttacgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaattttctcagcacagct *HYDIN* full set
aaaacacccagtttttagccaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcatccgttgctttcgg *HYDIN* full set
gtgggtcacctgtggtcgggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtaggctttttgagtccaaag *HYDIN* full set
gaatgtgtactggcctctgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgactaaaatgcatattcag *HYDIN* full set
gcacacgaggaaactgagataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctaccactgattgattag *HYDIN* full set
gaagggctcgagttaggaaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgttcgtccatgtggtc *HYDIN* full set
cattctctgtccaaggtggcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcagtgaatctgaccatg *HYDIN* full set
acattccattcagtggccagcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactgtggggcagtgact *HYDIN* full set
gcttggtcttcccagcaggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaacattcttctgcaaacct *HYDIN* full set
ggcttggcttaatagatgagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggggcttaaaacctagatg *HYDIN* full set
gtgcactcccagggtgactttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgatatgactctgtcttc *HYDIN* full set
gaaccctctttctagggaatacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcaaatgaaaacaagg *HYDIN* full set
atttttctcattctgatgttggggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacccgcctcagcctc *HYDIN* full set
gtggcactgagtcgagatcacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctggccaacatggtgaa *HYDIN* full set
ggtaaatacagtaagcttttagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagatgttgacagcaa *HYDIN* full set
cagacagcttgccagaacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctgatttaccactcatttt *HYDIN* full set
attactcattctcaggcagttcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctgtctctcctgct *HYDIN* full set
gtcattacatattgcagaatttccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccctcagcatgtgggg *HYDIN* full set
aaaattcaacctatggaaagggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcctactcagtctgtgc *HYDIN* full set
ggaaaaatcagccaacccgaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatgacctagcgaagcac *HYDIN* full set
gcttagggtcaggttttcaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctgagagacagagaac *HYDIN* full set
gatcggagtgtaaaaggacatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgatgagtgcatcacc *HYDIN* full set
cgggcatgtgaccaggtcaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaacatcatggtagaagcgt *HYDIN* full set
aattaacgctctttcttggaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccatgatgttggcggag *HYDIN* full set
cctcttgtcccctcaacaactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgtaatcaagaaggtgac *HYDIN* full set
ccctgcgttcctctccagatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaccacagctcctactc *HYDIN* full set
cccatctccttggtgattcacatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctcctccccagccc *HYDIN* full set
gtaggcaattctcactcatctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagaagtgtaagagaggg *HYDIN* full set
ccccctcccacatctgcatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagaaagaagtaaaatggac *HYDIN* full set
ggctggcgctcatcatagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaaaatctacttttgaagc *HYDIN* full set
acttcgggagactgaggtaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacagcgcttggattcta *HYDIN* full set
ggtccatctgtggttgtctaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccctggcccccttaa *HYDIN* full set
catataagggcacctgtggttaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctgaatttccagagctt *HYDIN* full set
gaccatatcattagaagctttggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggttccagggattcaa *HYDIN* full set
gtaactcactggcactggttgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgccatattacctccatt *HYDIN* full set
catattctttgtggtctggaaagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggtgggtactcaatca *HYDIN* full set
caccatgagtactatgcagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattctaccaaaaagacaca *HYDIN* full set
cagccattgtgaaagcagtttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcaaatcaaaaccacaat *HYDIN* full set
gataatgtctcataccagtcagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggctttgcagtcagt *HYDIN* full set
gagtgttcagagaaggttcgtttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttatccacccaccttg *HYDIN* full set
catgtttgtctttgggggctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaataataacagtgtttccta *HYDIN* full set
catcaagagtttataagctgaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcagggaaaagggca *HYDIN* full set
ccctgtctcttccctccctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgaactgtggaacaaaca *HYDIN* full set
ggctggctccataccagagttatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgctagaggtgggagc *HYDIN* full set
gatgaacctgtgggcagaatcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaacttgaatttccaaacc *HYDIN* full set
gcactaattccgtgaggctggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagatcacagtgccagca *HYDIN* full set
gatcatttactgaataggccatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctcaaactcctgacctt *HYDIN* full set
cctccaactagctgggatgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaataaggaagaatttttt *HYDIN* full set
gttgaatttgagaactttctgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcaatcctcctacctt *HYDIN* full set
cacagaagaacagatatgaatggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccaacagagagcagtt *HYDIN* full set
ggatacctgaaagacaactgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgagctattatgggtaa *HYDIN* full set
gcttctcaaactatgtatctatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaggccctgggaagg *HYDIN* full set
gaagctatctccatgttagggccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtagtcgacctgcaagt *HYDIN* full set
aaaggctgccaaacaggtaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctgtcaggtgagtagtg *HYDIN* full set
gtgactttgctggagagtctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactgcctaatttaatcttca *HYDIN* full set
catctcctaagctctgagtaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggttgggattcaggctt *HYDIN* full set
gaggctgaaactcaagttactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaggggaagccaactaaag *HYDIN* full set
ccttccttcctctgactctacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatctgatgcaaaccatg *HYDIN* full set
aatgtcaatattcttagtggtgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcgtgatgaaagctaa *HYDIN* full set
cggcaacttgaaaactacctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaagcacaactgtttgtaa *HYDIN* full set
acatttgagcaaaacaagatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagatgcaaggcctgaa *HYDIN* full set

ccccatggctactatcccccttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcagggaagatgttccgg *HYDIN* full set
cattagggtgatgccccccagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattgattcattagacttggc *HYDIN* full set
aatccaccatacttctttcctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacatgcatgtcagaca *HYDIN* full set
gttccctgtagctgccagagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtatgctgacccaaacag *HYDIN* full set
gccacaggagacagtgaaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtagagggtcttcatacc *HYDIN* full set
ccacggcagctcctaaatctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgacctggagtgatttcataa *HYDIN* full set
gctgactttacctctggctgaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcaccttgtctgtcaac *HYDIN* full set
ggacacctttactatttctagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtgtctgacttcctct *HYDIN* full set
atcccactaccatcttcgtgatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccctcccttggaagct *HYDIN* full set
cctatctcaagcccttgatgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatatgtccaggttccact *HYDIN* full set
gaaacccgagcctgcacttccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtacagttgcttttcccct *HYDIN* full set
catctaacagggatggtaaaaatacTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccctgctttctggtg *HYDIN* full set
ccccttcacacttatttttaccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccctacaccctccctt *HYDIN* full set
gtccccgagatggatttaaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccctgtcccatgctttg *HYDIN* full set
atgtcttctatgtgtaggcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNattttgttaagtaagtgca *HYDIN* full set
gatggtaccctctccctatccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtgaggacataatgtttg *HYDIN* full set
ccccattttaccaatgcagggacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaatccacctgcctcagc *HYDIN* full set
catcgtcatatcctcaggatatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactcgaccatcttgcttg *HYDIN* full set
cacctactgagaccatctccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccatatctcctaccaaatt *HYDIN* full set
cgagcttactttctcactattcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaacatgcggtgtttg *HYDIN* full set
ggagctgtcacacccttatccatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcagccccacagcca *HYDIN* full set
ccaactttccccatgttcgtcagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtggggctcgtaggttt *HYDIN* full set
atagaagagagagaaatgcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgatctgtatcctatacag *HYDIN* full set
ggaaactccttttaaagacctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccagtggggaactactc *HYDIN* full set
ggctgccatagacgatttcttgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggacagctcggggaaag *HYDIN* full set
aagagacagggtcttgcctatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacatctgtagtcccagct *HYDIN* full set
ggtgggaatctctccaccgattaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaacctcccaaagtgc *HYDIN* full set
agaaaaccccaagcaagccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacttgacccaaacatatctt *HYDIN* full set
gccagtccttcactaaagcacaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagcactgggagagctcc *HYDIN* full set
gctctccttccaaccccccactccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcctcaccctctgaagca *HYDIN* full set
gaacatttcatacttttcatttgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcgtgcaggctggagct *HYDIN* full set
atccctctggctgccatgtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaatcaggaagctaggaga *HYDIN* full set
gtaaggaggggactgagacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgacagtggggacaggg *HYDIN* full set
atctcacattcggtagccacgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccccatttcctgtgattgc *HYDIN* full set
catatatgtggctagagagccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaatatatcctccctcat *HYDIN* full set
cagcctgggaacccaataagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaactctgtgcttggctcc *HYDIN* full set
atccaagccttttcattcctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaggtagaagagaaagaaa *HYDIN* full set
atctggttggctctataggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagctaacacagttggtcac *HYDIN* full set
gagaataaaactgtgcacaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttcaggtccataaggg *HYDIN* full set
caaggaacccatcagtactgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccgaacccacatcatcaag *HYDIN* full set
actgggtcagaatgactgctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatatttccgtttcacaatga *HYDIN* full set
cacacagctcctatgtgaaccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgagactgaatgaggcta *HYDIN* full set
gggtatatgcactgtgtttactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaactgacatgacaccatt *HYDIN* full set
cattttgaagattaaatgcacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgctgagggtattaggt *HYDIN* full set
actttgaccaattccaactcaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcactgatttcttggtg *HYDIN* full set
atctgaagacctcctatcttagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccatccctagaaatcctg *HYDIN* full set
gttcctagtctgatttcactggggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcagaatgccagagc *HYDIN* full set
gctcccttgctcctgaggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctgaaggttctcatttcatg *HYDIN* full set
ccttgcccaagtcatttcagaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcggagagccagaaatac *HYDIN* full set
aaggtagggcagcagggtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttttaaattagaggatggaa *HYDIN* full set
cgtaaccatttatctgtgacttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtattgaatgcccaacacc *HYDIN* full set
attgccactctgtctattactacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcatattgcctcctt *HYDIN* full set
ccccatgacaaaatttcagaactcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtagggctgagctctg *HYDIN* full set
atgaggttacagagaagcaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggggaatatttgatcaat *HYDIN* full set
aaaattagtgaaaaaagtgccgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctccaaatgcagtcacat *HYDIN* full set
agagctgcaacctctcagatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatacatttgcttagaaaatg *HYDIN* full set
gtatggctcctagcccatcctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggatgaggaaagcctgg *HYDIN* full set
cctgctaagattctggcagactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccggagaagctctgatct *HYDIN* full set
gcaggaaagattcaggccttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaaaatggggccataatg *HYDIN* full set
ggttccaacatcttgcaaaacttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagcctcacccatgatt *HYDIN* full set
ccaccttgcccggcctagctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccatgcccggctaattttt *HYDIN* full set
ccaagatatattatgtgaaaagcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggcggaggttgcagtg *HYDIN* full set
ccttttatgggaaaagtttgccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtattggaacatagcca *HYDIN* full set
ggccccttttgggctaataatttgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcatgcatgctaca *HYDIN* full set
ccaacttcctccatggagtggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacctgacaagggtgtccc *HYDIN* full set
acttcgtaaccactcagaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccaaaattacatgtgcta *HYDIN* full set
gaaaggttcttcaaaggaagggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcaggtgtgtttgtgta *HYDIN* full set
cacagaatcttcatgaagaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgacctggaaagagccac *HYDIN* full set
caccctgtccatttccttcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccggctaatttttgtatt *HYDIN* full set
ccttatgcaaggtgaaataggacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggttgcagtgagccaag *HYDIN* full set
gcttctagaaatgttgcctgcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgacttcatggcttcacat *HYDIN* full set
gtagttcctcctgtctctgattcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtctcgaactcctaac *HYDIN* full set
ggaaggaggcagagcactcagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgggaaaacccagcgagga *HYDIN* full set
aagagcgactttgggatgaatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtggggaccaggacaatg *HYDIN* full set

gaaaacttcccacatcaccctt**CTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acaatgattggcctcaac *HYDIN* full set
ctgcatggtggagaattgaagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ttgggacagtttcacacta *HYDIN* full set
gtggccatcttgaatccctcttac**CTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aacaaccccccctcccc *HYDIN* full set
gggctataaggagagggtgaatttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gctcccacgaacactc *HYDIN* full set
gagcagaagtgaaagtactggtagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gtatgggaagcccagc *HYDIN* full set
ggagtcagaattgggatgttctC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ttcacaggggacactgt *HYDIN* full set
accagttagaaatcctgatcttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aaagcatgagttttccaa *HYDIN* full set
agatgactaactctggtaagcaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ttctcaatttcatggcct *HYDIN* full set
gaggattaggtttgccagggtC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tgaatagatgcagagctga *HYDIN* full set
acttgacgttcccagaagaggaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**cttaagcctgtttctgct *HYDIN* full set
gcttgtgtcagaatgtaaatgaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccttggagcgtaatgatt *HYDIN* full set
gaccagattagtctcgaactccC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ctgcaacctctgtctccc *HYDIN* full set
atttgttctgtttatgcttttactC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**caggcagaggagggcg *HYDIN* full set
actgtacccctccatcctttctgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccctgtaaggtggccag *HYDIN* full set
gctgtcagcaaagcgttcatC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gaaatgcacagttggcttaa *HYDIN* full set
gtgtgtgtacctatctttcaatcaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ggaggtggggctctga *HYDIN* full set
gctggatgactgctgagggatgcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ggagaggagcccacagt *HYDIN* full set
acatttcaatttagatgatttaggC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**atacggccgttgtaga *HYDIN* full set
cggtgagaatgctgtgctagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tacatgatcgcttatttaat *HYDIN* full set
cagtggtcatttgtcacagcaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aggtgacactgtttctgtt *HYDIN* full set
ggcagggaaagatgttaggaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ctcatgtttgttaacatgta *HYDIN* full set
attgtgggctcttctcccctaaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ctcagcagtcatcaagc *HYDIN* full set
gatgtgatagtaggagcaggggC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccatagcagagggactct *HYDIN* full set
gtctccaaaacatggatgaggaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccccacactaagacccta *HYDIN* full set
gtagtttctgtgtggttcccagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**caccccatccctgcaatc *HYDIN* full set
gatctcgaactgaactcctgggC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acctcctgggttcaagca *HYDIN* full set
ggaagagaaggaaagaaagaaaaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gttcaggggcccccat *HYDIN* full set
ggagactggttatagcttcacC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**cctagttctgtatgcttag *HYDIN* full set
gactaaagaccttcctgcttccC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tcttgagactcttggact *HYDIN* full set
gatctggaaatggagaggctgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**agaaggaggaaagccagaa *HYDIN* full set
gtgacactccctgctgggaacC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aacgaagtgaagggcattc *HYDIN* full set
acaaaaacatacccacacataaaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acgctgaggcaggaga *HYDIN* full set
aatgaaatctgctggtgattctgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tccaccttggcctccca *HYDIN* full set
gcaagaaatagaagactcaattcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tcacattcatgctgctg *HYDIN* full set
catcttagtcactgaccactgcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gagccaaagaagaggaac *HYDIN* full set
caagatggtgtggaactcctgacC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aacctccgcctcccagg *HYDIN* full set
aatagaaacataactccctgagagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tggaaccagcacaggt *HYDIN* full set
gccatcgtccttgtttctactttgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**atccacctgccttggc *HYDIN* full set
gtgctaccctcttctaaattatatC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tagctgggactgcagg *HYDIN* full set
gaagaggcagtgtgaccgttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**attgtgctccttctgtgtta *HYDIN* full set
accccaagtaggataaaatacaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aggtttgcaaaaaggcc *HYDIN* full set
atgttccacgtgtgtttgaataC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gcttcattttgccttctc *HYDIN* full set
atgacccaaaaaacgatggagccC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**catgtctccaagcctga *HYDIN* full set
cacaaaacaagtttaaaagtgcagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acactggctgcactgc *HYDIN* full set
gcacctttccaaaatgcacaatC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gaccactacacgctgtg *HYDIN* full set
gtgtcatgccttctaggttcctcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccaggcacaacctagc *HYDIN* full set
gtcctcctcccctttcccctcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**agagtccatacctgctctc *HYDIN* full set
gtcacacatctagattagtccatcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**caggggtgcttctccc *HYDIN* full set
caggcatgagtcaacttttaaaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**catgcaccagcacgacc *HYDIN* full set
gaccacccatcccgcctcccgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aaggcagacaattaaacaa *HYDIN* full set
acaacctttgcagtccccattC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gggttattttcccttgact *HYDIN* full set
caatttgaaatgtcaccagggaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acatctgggctcaagcaa *HYDIN* full set
cagaggaaggaagcaaggaagC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**agagttatagtgagttgag *HYDIN* full set
cctgccccctctcccaatcacC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**agttgttctggagctaaaa *HYDIN* full set
atcctccgcccctggatacaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aactgcaggaataaatcaag *HYDIN* full set
ccgatatgcctgatgaacataC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gatacccatcctattgaca *HYDIN* full set
atcctagaactgacaaaagaattC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**cctcgccataacaatca *HYDIN* full set
cattcacagctacctgaatgatC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**agcacaattcacaattgc *HYDIN* full set
ggcaaatgtatgggtaagtgcagaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aggggctggggtgctg *HYDIN* full set
aaaaaagtcttcctttggcaaattC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**acccccagcccctcttg *HYDIN* full set
agcaaactaatttgccaaaggC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**aggtgcttttctgtagtgt *HYDIN* full set
gagaatccctatggaaaattggC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**caccaccagagaagacag *HYDIN* full set
gatatttctgtagttacatgaagcC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**cctgctggacacccag *HYDIN* full set
caaggagatgagaaagagagctC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ggttctcaaaggagagat *HYDIN* full set
gttattcagggcttcttttcttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccctctcctcagtattca *HYDIN* full set
caagcttctccaaagagttatcttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**cccttcccatctcctg *HYDIN* full set
attgtaaacctaccttctgcaaC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tcataccccctctatttc *HYDIN* full set
aattcgtgaggactgaactgggtC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gcaatctgcccacctca *HYDIN* full set
catggcctatctaaaaccttttttC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gtgccggatggaagtga *HYDIN* full set
gggccttctggtttgcttcgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ctgactctcatttggtctct *HYDIN* full set
gagggtttgggagaaggaggtgC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**taactgagtctccctcag *HYDIN* full set
cacatctgttttagagatgaccC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**gtttcaccatgttagcca *HYDIN* full set
cctgctgaagttcatccactC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**tctaacttctagagcaaaga *HYDIN* full set
catttcttaaccctttttcgaattC**TTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNN**ccctgacctcgtgatct *HYDIN* full set

caacaaacctgcacatgaaccccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaggatgcttgcgggtg *HYDIN* full set
ggacttctacagaggtttcctaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcctgcctcggcctc *HYDIN* full set
agccaagacagtgcgtgggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagagggattctctcaaagg *HYDIN* full set
gccctgatgcagttcatataaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaccttgattttggacatc *HYDIN* full set
ccacagaaggaaatgccttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttttgtattttcacaactg *HYDIN* full set
gggctgacaggatgtgtgtaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaactcaagttttaattcaga *HYDIN* full set
ccatcttctctttcattagatgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctctggggagtccacc *HYDIN* full set
ggggcagtgggggaagctctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaactcaatcaaatcaacaaa *HYDIN* full set
aactcttgaactcaaacaatcctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcaagccatcctccc *HYDIN* full set
atccctgcaatactaacaacacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagctctgcaaatttac *HYDIN* full set
caaaaaagcagagacggttcttaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctatggttccaactggc *HYDIN* full set
gaatgtaatgaaaaaggaaagcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctgggattacaggca *HYDIN* full set
acttttttgtattttgtagaggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgaggctggagtgcag *HYDIN* full set
cccgtgtgccacccttcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaactgacaggggaaaagtc *HYDIN* full set
cagggaagcaaggaagcacagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtggaatcttgcctgttgg *HYDIN* full set
caatggcctctcttgaccatgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctctggtttctgtgtca *HYDIN* full set
gtagggatttcaacacctaatataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggctaaacagggctgg *HYDIN* full set
gaaggatacatgacctaatccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgtccaagttttcttcc *HYDIN* full set
gtgaaataaagaacatggcttcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaagctgagtcccatca *HYDIN* full set
cacttcctctcttccacgcttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgaagaacatggagtttt *HYDIN* full set
aaaattactgcaagaagatagtaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccttctttgcccaag *HYDIN* full set
ggggaaccatgacgtccaccccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatgaggtgccccaccc *HYDIN* full set
gactatggaggaggttggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttatacagcaaaaggactt *HYDIN* full set
cctaccccatttaaatatgccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatacctgcttctctcatt *HYDIN* full set
gtggtggctggagttgcagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcttgacaagtggaaacca *HYDIN* full set
catcacaagaaactgaggtgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccttctttgcctaccatat *HYDIN* full set
gaactgaggcagcagaaagctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcctgggcactggtgatt *HYDIN* full set
gtctaccaggtactgaccactcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggtcatcaacgttgac *HYDIN* full set
gtcttgtcttcctcttttgtacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatgctgccttgattgct *HYDIN* full set
ggagcctgtatgcgtgagtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgggaaagatcacctgtga *HYDIN* full set
accctcgagtattacctgggcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtctgcatgtccctgacac *HYDIN* full set
gataaaagaaacttcaggagggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctactgtccaaggtagtg *HYDIN* full set
cctgtagtagtgtgcacgttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttgtagggtttaaaatatt *HYDIN* full set
cactgtgatagtgaggcctcccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctttcccaggctgttctc *HYDIN* full set
cagtggaataatatctcccatttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggggtgtggctgtaaa *HYDIN* full set
ggggaatggttttgggttgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaaactgtagaaaacaaagt *HYDIN* full set
caacctttgtctcccgggttctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagagcatttgccagtc *HYDIN* full set
gagctcaaggcaacaacaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggtgaaaattgttaagaaa *HYDIN* full set
ggaattgacatcaatgcaggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagaaaccaccactttct *HYDIN* full set
actcctagctcatcctgcaggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccctaatgaagcagctac *HYDIN* full set
acaatctgggataggaagactagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggatggagcttggaac *HYDIN* full set
gggataaattttgctttttggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacccccagtgtcttcact *HYDIN* full set
acaaaaaaaaagtagtatctgcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagctacttgggggct *HYDIN* full set
gtgtttgtttcttcacaggcaaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgtgggtgcacagaag *HYDIN* full set
gcagcaagacgttcaagaggtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaactggagcaaaggtgac *HYDIN* full set
gagaaaatgagagtgaaaaagtcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccaggtccctcccac *HYDIN* full set
ggcaaaaaatgcatgttaagcctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttggcacacccagggtg *HYDIN* full set
gtcagattggtgcttgatgtcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagcatgtgtagaagtttag *HYDIN* full set
atggctgagaattacacaggtagtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccactgtgcctggcct *HYDIN* full set
gatcaataagctctctgctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagaaatactcagccttatt *HYDIN* full set
gagccgagattgtgtcactgtatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatggtgaaaccctgtc *HYDIN* full set
gtgctctgcatctctgaggacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttcatttagcctgtctg *HYDIN* full set
cattggcactgatttgaagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatggaatgaaagagctttt *HYDIN* full set
ggtgcatgggcaaaaagaacagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttctctgcagggacccc *HYDIN* full set
cccctggagacttgataagggagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgagctctttgaaggc *HYDIN* full set
ccaacccctaaactgagaactcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagctatgcccattcatt *HYDIN* full set
cacagggacttccttcctttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtcttaaagcaggaaaatt *HYDIN* full set
gtgaaatagattttgtttcctctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttggcctcccaaagt *HYDIN* full set
accttgtggtgtctgcagagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaccaaggtaaattcagtg *HYDIN* full set
cattgggttagcttttctctcactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccccagcagctcactt *HYDIN* full set
cctcaatccctgtcccactctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaccctaccctgactgtgg *HYDIN* full set
ccaatagcacctactccttctgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcttcccagatggggg *HYDIN* full set
gtggttctcatctgcacatacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatggcagccattcactaa *HYDIN* full set
ggaccaaaaatggtgaaaggaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgagaagcagagagctctaa *HYDIN* full set
agtggagccatgtcccatggcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtaactggacaccaatgc *HYDIN* full set
gtcatctaccagagaagccctgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtgcactccagcctgggg *HYDIN* full set
cggaaggctgagtcaggagaatagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgccgagacgggcagat *HYDIN* full set
ggtctttaagtgacagtcatatcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccaggagctgggtcca *HYDIN* full set
ccttggaatgggtaccaactctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccactgccagctggcc *HYDIN* full set
cctgctggtaggcaaaactacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagaagtgtcttttaccac *HYDIN* full set
atcccaccactgtccaggtaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagagaagccagggagac *HYDIN* full set
gctgcctcagtctctcgagtaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccggctaattattgtatt *HYDIN* full set
atgcaccacagttagttgaacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggaatgttgtattcacaga *HYDIN* full set
gtggagattctggaacattcagagCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaacctgggaggcggg *HYDIN* full set

gcttttctgatagatgtgtaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatggggtttcaccatgt	*HYDIN* full set
gtctctttccagtttcctgctcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtatgtgtgttgtcaggcc	*HYDIN* full set
cagcaatctctttacaaaaaacaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatcctggcgaggctgc	*HYDIN* full set
gtgacattaaatgtcaaggccgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttggcctagtcaggtaa	*HYDIN* full set
gcaaggtcttcttacacctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtacattctggggatagagc	*HYDIN* full set
atactccgagaattgttgcaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttgattgccaaattgttg	*HYDIN* full set
ggtattactaattctaggtccctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttgggttggttccaagt	*HYDIN* full set
caaatcctcaggaagaaggtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtccaggattgaccagatat	*HYDIN* full set
caaaacagatgaagacaggttccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgatgcagccataaggcc	*HYDIN* full set
gcaactttgctgctgagtttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaaccagagtagaaagatctt	*HYDIN* full set
gcctctttattcaagatagaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagccttgttctccttct	*HYDIN* full set
cagccatggtccaacatttatgtCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggaaatggcaccctga	*HYDIN* full set
gtgggatgttctagtccaaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatagacttggaaactgagg	*HYDIN* full set
gcttccagaagtaattaggtaacaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccacccacctcgacct	*HYDIN* full set
gtagagtgaagaatattgggcaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatcgcttgaacctggg	*HYDIN* full set
ggaaacacattggagattccaataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcatctgcccctctggg	*HYDIN* full set
caatggggtgttaaggaaagaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNactagacacaagaaggaaa	*HYDIN* full set
gggtttaaagatagttgggtggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcacccattatgtattgg	*HYDIN* full set
gcctctactgccactgttctaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNccagtgggtgctgctca	*HYDIN* full set
actgattaaaatggtgcctggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcttctttacggagaatgg	*HYDIN* full set
gaagtcatcttctcaagtcctagcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaagccagagctctctc	*HYDIN* full set
cagggaaggctcttagaatggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgacaatgaatgacatggaa	*HYDIN* full set
aatagagaactgggctgtctacacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcacctgcctcagcctc	*HYDIN* full set
caggaaagggaaatgatacttattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgaaactgaggcaggga	*HYDIN* full set
gaggaagtgttcatggaatgcatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctgactctgggtgcta	*HYDIN* full set
gtaagggagcagtgtctcccacccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcttagcggctgctgcc	*HYDIN* full set
cattgaggacacccttacaacatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaggaagtgaccttggc	*HYDIN* full set
aggcagtcctctgagtctaaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNctcatttctccattagcc	*HYDIN* full set
cctttccattcttcagctctcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgaccccagcaagtgaa	*HYDIN* full set
cctaactgaaccatcagctgtgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgcctgctttgacatccct	*HYDIN* full set
gtaattccagctacttaagagctgCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcagcactttgggaggc	*HYDIN* full set
aagagtccaggctgcagataCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtactgatgtgggattgactg	*HYDIN* full set
gtgcttgcggttctccaagttcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccaaagtgctgggatta	*HYDIN* full set
gatgggaccggctggctgtaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtttccagagcctccattatc	*HYDIN* full set
gaaccctggagccatgcactCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtattaatacctcctacaga	*HYDIN* full set
cactgctggatgcgtcttgcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggtaacttttttagatgt	*HYDIN* full set
cacctatatttcaaagcatacattCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcactccccttttcccc	*HYDIN* full set
cgctcatgctgtcttctgaccCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatgttccccagcactgatt	*HYDIN* full set
gtaatttataattcaggtatgcacCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcccccaccttcagtaa	*HYDIN* full set
catggcatcttcaccatgaatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgctgtttgaaagcatttta	*HYDIN* full set
gttttcagtttaccttgcccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNacagttacttcctctactt	*HYDIN* full set
aatacatttcatagggcaatagttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggacacccaggagctc	*HYDIN* full set
cacctaggaaataatggaagctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNaggcatcgtttcattttc	*HYDIN* full set
ccgaagtctgcttccttgctaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNagtagagatggggtttcac	*HYDIN* full set
ggaggccaacacggacagatCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtggcatacagtaatcattaa	*HYDIN* full set
ctggaacaaagaaaacaaacaaaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtcaaactcctgacctca	*HYDIN* full set
catgcactcctaaccactgtcaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtgggctgaaggtggaaa	*HYDIN* full set
cctgtggttctactcctaggCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNttagcagtttatcagattag	*HYDIN* full set
gggagagattgtctgctacttaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcaagccatatagagtcag	*HYDIN* full set
caaaatggcaaacgcatcctttCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNgtaatcccagctactcgg	*HYDIN* full set
gtatctaaggctctctcctctCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNatcttgtcttcactgttac	*HYDIN* full set
cccggcctgtggcttctattcCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNcggctaatttttgtatttt	*HYDIN* full set
ggtggggattattgaggaccaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNtagaagccacagtctttca	*HYDIN* full set
gtcacaaacaaacaaacccgaCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNNNggtggcgcatatctgtaat	*HYDIN* full set

*Most individuals in this study were genotyped using a reduced set of 153 MIPs targeting *HYDIN* singly-unique nucleotides (SUNs), a subset of 717 SUN-targeting *HYDIN* MIPs designed for paralog-specific copy number genotyping.

Additional Data 1: Measured HYDIN isoforms (fasta)

>hydin_transcript_A
AGGTACTAGCGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAA
GGCCTTGGAAAATCGGAGCGATTAGAGCGGCCGAGAAATCCGTCTGCCCCTCCGAATCAAAGGGGAAGGCATGGGACCTAAGATTCACTTCAACTTTGAA
TTGCTGGATATTGGGAAAGTTTTCACTGGATCTGCACATTGTTATGAGGCGATACTGTACAACAAAGACAGCATCGATGCTCTCTTCAACATGACCCCTC
CAACTTCAGCTTTGGGGGCCTGCTTTGTTTTCAGTCCCAAGGAAGGCATCATTGAACCAAGTGGAGTCCAAGCTATCCAGATCTCCTTCAGCTCTACCAT
CCTGGGAAACTTTGAAGAAGAGTTCCTGGTCAATGTCAATGGGTCACCTGAGCCTGTGAAACTGACCATTAGAGGCTGTGTCATTGGACCTACCTTCCAT
TTTAATGTTCCAGCTCTGCACTTTGGTGATGTTTCCTTTGGGTTTCCTCATACCTTGATATGTTCCCTCAATAATACCTCTTTGATCCCCATGACTTACA
AACTGCGTATCCCTGGGGATGGCCTTGGCCATAAAAGCATTTCATATTGTGAGCAGCATGTGGACTACAAAAGACCGTCTTGGACCAAGGAAGAAATATC
CTCAATGAAACCAAAAGAATTCACCATCTCTCCTGACTGTGGCACCATTCGCCCCCAGGGATTTGCTGCTATCAGGGTGACATTATGCTCCAACACTGTA
CAGAAATACGAGCTGGCACTCGTGGTGGACGTGGAGGGCATCGGAGAAGAGGTGCTGGCGCTCTTAATTGCAGCAAGGTGTGTTGTACCTGCCCTCCACC
TGGTCAATACAGAGGTGGACTTTGGGCACTGCTTCCTGAAGTACCCGTATGAGAAAACACTCCAGCTTGCCGATCAAGATGACCTCCCAGGATTCTATGA
GGTCCAGCCTCAGGTGTGTGAGGAGGTGCCTACTGTGCTGTTTTCCAGCCCCACCCCCAGCGGGGTCATCTCCCCAAGCAGCACCATCCACATACCACTG
GTCCTGGAGACCCAGGTCACTGGAGAACACAGATCCACGGTTTACATCTCAATCTTTGGGAGCCAGGACCCCCCTTTGGTATGTCACTTAAAGAGCGCTG
GAGAAGGCCCAGTTATCTACGTCCATCCCAATCAAGTGGACTTCGGGAATATCTACAGTCCTAAAAGACTCTTCCAGGATTCTCAACCTATGCAACCAGTC
CTTCATTCCCGCATTTTTCCAGGCACACATGGCACACAAAAAATCCCTTTGGACGATTGAACCCAATGAAGGCATGGTTCCTCCAGAAACTGATGTTCAA
CTGGCACTGACCGCCAACCTGAATGACACACTGACATTCAAGGACTGTGTTATTTTGGACATTGAAAATAGCAGTACCTATCGGATTCCTGTTCAGGCTT
CCGGAACTGGTTCCACTATTGTTTCAGATAAGCCCTTTGCTCCAGAACTCAATTTGGGGGCACATTTTAGGTAAGGAAATCTCATGGACCAATTTTTATA
GATATCACATATGTTACATATGCTGAGCATCACCTCTACTTTAGAAAATTATTTTAAGTAATTAAAACTAGCCACCTCAGTCCCTCTCCTTTCCCATCCT
CACTCTTCCACCTCCTGATGTGTATGTCAGCAACAAGAAGACATTTTCTAGCCATAACTTGGTTGCATTCTCTAATCCGTGTCCCTCTACATGTGAGAAG
AATATGGTTTGTAGGAAATGCATTCTTTTCTCAAACTGGCCACTCTGTGATTAATTTAGACTCTCAACCGTAATTTCTCCTCTTTGATTCCAATACACTG
TTAATCCCTCTCCACAAATGT
>hydin_transcript_B
AGGTACTAGCGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAA
GGCCTTGGAAAATCGGAGCGATTAGAGCGGTTCAAACAATACAGAAGTGTATAAAGTAAGAATTGGAAGTCTCTTCGCTGCTCTAATCTCACTCCTCAGA
GGTGATCACTGGACCTGAGTGAGCAGAAGGCTTTTCAGAATTACAGCCTGAACTATTTCTTGAGCAATTTGGGAACTCATTTCTAAAGCAGTCCTCAGAT
TCACACAACGTAGTTTGTGACGGAGTATGCCCAGAGCAGTGTGGAAGATGATCAAAGTATGACATTTCTTCTACCCCAGAGGCTGGCATTTCTGCAGGCA
CAAATGAGCTTGGTTTTCAAGCACTGCTCTTTGAACACAAGGAATTCCTTTAATGATCACGCCGGACTATATAGACTTTGGAATCAGACTCTTTGAATTC
CAGCTCTGCCATTTACCAACTTTGACACTTTGGACAACGTCTCATTCTGTCACCCAGGCTGGAGTGCAATGGCACAATTATAGCTCACTGCAGCCTCAAA
CACCTGGGCTCAAGCAATCCTTCCACCTCTGCCTCCTGAGTAATTGGGACTACAGGTATTATGTTTCTATTAATGCAGTCTAAAGGGTTTTGGAAATCTG
GATTGTAGCAGGGCATTGTGGATCTGCCTGTTGGTGGAGAACCCTAGACTGCATGCAGCCATGCACATGATCTGGAGACACAGCAGAGACAAGGTCTTTG
TCTGTTGCTCAGGACGGAGTGTAGTGGTGCCATAGCACACTGTAATCTTGAATTTCTGGACTACTGGCATACACCACCATGCTTAGGTTCGGGCCACAAC
ATCCGTGACTTTACCGGTTTTGGTCTCTTCCTTGTAACTTTCGGCTCCAGGATGCTGCGCAGCTCCTCTCTGCACTATCTATCAGAACGTTGCTCTGCGT
GCTTGAAAGTTTCGCCAAACAGAAGTACACAAAAACAAACAGAGGCGAAACACAAATCCCCTCAGCCCTGGAGGAAAAGGCCTTGAAGCTCTTCCAGTGC
AGAGGCCACGCAGAGGACATGCTCCTACCTGTGCACATCGGCCACGCCTGGGGTGCCTCCTGGAGTCCAGGCTGCTGGCGATGGAATCCATCATGTTGCT
GATGTCACTGATTACGCCCATACGTTGACCTGTGTTACTGAAAGAGAAAAGTTTATTGTACCCATCAAAGCTAGAGGGGCACGAGCCATTCTCGATTTTC
CTGACAAGCTGAATTTTTCCACTTGTCCTGTCAAATACAGCACCCAGAAGATTCTGCTGGTACGAAACATTGGCAACAAAAATGCTGTATTTCACATCAA
AACTTGTAGGCCTTTCTCTATAGAACCAGCTATTGGAACTCTTAATGTGGGAGAGTCCATGCAACTGGAAGTGGAGTTTGAGCCACAGAGTGTGGGCGAT
CACAGTGGAAGACTTATCGTGTGTTATGACACAGGTGAAAAGGTGTTTGTATCTCTCTATGGAGCTGCCATAGACATGAATATAAGGCTGGATAAGAATT
CCTTGACCATCGAGAAAACCTACATATCTCTGGCCAATCAGCGAACTATAACCATTCACAATCGCAGTAATATCATTGCCCATTTCCTGTGGAAGGTATT
TGCTACCCAGCAAGAAGAGGACAGAGAAAAATATAGGGCCTGTGATGATCTGATCAAAGAGGAGAAGGATGAGACTGATGAGTTTTTTGAAGAGTGCATT
ACTGATCCTTTACTCCGAGAACATCTTTCTGTTCTGTCCCGAACCTTTGCGAATCAAAGGAGGCTGGTGCAGGGAGACAGCAAACTGTTCTTCAATAACG
TTTTCACTGTGGAGCCCCTGGAAGGTGATGTCTGGCCCAACTCATCAGCTGAAATCACCGTGTACTTTAACCCACTAGAAGCCAAGCTCTATCAACAGAC
CATTTACTGCGACATTTTAGGCCGAGAAATCCGTCTGCCCCTCCGAATCAAAGGGGAAGGCATGGGACCTAAGATTCACTTCAACTTTGAATTGCTGGAT
ATTGGGAAAGTTTTCACTGGATCTGCACATTGTTATGAGGCGATACTGTACAACAAAGACAGCATCGATGCTCTCTTCAACATGACCCCTCCAACTTCAG
CTTTGGGGGCCTGCTTTGTTTTCAGTCCCAAGGAAGGCATCATTGAACCAAGTGGAGTCCAAGCTATCCAGATCTCCTTCAGCTCTACCATCCTGGGAAA
CTTTGAAGAAGAGTTCCTGGTCAATGTCAATGGGTCACCTGAGCCTGTGAAACTGACCATTAGAGGCTGTGTCATTGGACCTACCTTCCATTTTAATGTT
CCAGCTCTGCACTTTGGTGATGTTTCCTTTGGGTTTCCTCATACCTTGATATGTTCCCTCAATAATACCTCTTTGATCCCCATGACTTACAAACTGCGTA
TCCCTGGGGATGGCCTTGGCCATAAAAGCATTTCATATTGTGAGCAGCATGTGGACTACAAAAGACCGTCTTGGACCAAGGAAGAAATATCCTCAATGAA
ACCAAAAGAATTCACCATCTCTCCTGACTGTGGCACCATTCGCCCCCAGGGATTTGCTGCTATCAGGGTGACATTATGCTCCAACACTGTACAGAAATAC
GAGCTGGCACTCGTGGTGGACGTGGAGGGCATCGGAGAAGAGGTGCTGGCGCTCTTAATTGCAGCAAGGTGTGTTGTACCTGCCCTCCACCTGGTCAATA
CAGAGGTGGACTTTGGGCACTGCTTCCTGAAGTACCCGTATGAGAAAACACTCCAGCTTGCCGATCAAGATGACCTCCCAGGATTCTATGAGGTCCAGCC
TCAGGTGTGTGAGGAGGTGCCTACTGTGCTGTTTTCCAGCCCCACCCCCAGCGGGGTCATCTCCCCAAGCAGCACCATCCACATACCACTGGTCCTGGAG
ACCCAGGTCACTGGAGAACACAGATCCACGGTTTACATCTCAATCTTTGGGAGCCAGGACCCCCCTTTGGTATGTCACTTAAAGAGCGCTGGAGAAGGCC
CAGTTATCTACGTCCATCCCAATCAAGTGGACTTCGGGAATATCTACGTCCTAAAAGACTCTTCCAGGATTCTCAACCTATGCAACCAGTCCTTCATTCC
CGCATTTTTCCAGGCACACATGGCACACAAAAAATCCCTTTGGACGATTGAACCCAATGAAGGCATGGTTCCTCCAGAAACTGATGTTCAACTGGCACTG
ACCGCCAACCTGAATGACACACTGACATTCAAGGACTGTGTTATTTTGGACATTGAAAATAGCAGTACCTATCGGATTCCTGTTCAGGCTTCCGGAACTG
GTTCCACTATTGTTTCAGATAAGCCCTTTGCTCCAGAACTCAATTTGGGGGCACATTTTAGGTAAGGAAATCTCATGGACCAATTTTTATAGATATCACA
TATGTTACATATGCTGAGCATCACCTCTACTTTAGAAAATTATTTTAAGTAATTAAAACTAGCCACCTCAGTCCCTCTCCTTTCCCATCCTCACTCTTCC
ACCTCCTGATGTGTATGTCAGCAACAAGAAGACATTTTCTAGCCATAACTTGGTTGCATTCTCTAATCCGTGTCCCTCTACATGTGAGAAGAATATGGTT
TGTAGGAAATGCATTCTTTTCTCAAACTGGCCACTCTGTGATTAATTTAGACTCTCAACCGTAATTTCTCCTCTTTGATTCCAATACACTGTTAATCCCT
CTCCACAAATGT
>hydin_transcript_F
AGGTACTAGCGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAA
GGCCTTGGAAAATCGGAGCGATTAGAGCGGTTCAAACAATACAGAAGTGTATAAAGTAAGAATTGGAAGTCTCTTCGCTGCTCTAATCTCACTCCTCAGA
GGTGATCACTGGACCTGAGTGAGCAGAAGGCTTTTCAGAATTACAGCCTGAACTATTTCTTGAGCAATTTGGGAACTCATTTCTAAAGCAGTCCTCAGAT
TCACACAACGTAGTTTGTGACGGAGTATGCCCAGAGCAGTGTGGAAGATGATCAAAGTATGACATTTCTTCTACCCCAGAGGCTGGCATTTCTGCAGGCA
CAAATGAGCTTGGTTTTCAAGCACTGCTCTTTGAACACAAGGAATTCCTTTAATGATCACGCCGGACTATATAGACTTTGGAATCAGACTCTTTGAATTC
CAGCTCTGCCATTTACCAACTTTGACACTTTGGACAACGTCTCATTCTGTCACCCAGGCTGGAGTGCAATGGCACAATTATAGCTCACTGCAGCCTCAAA
CACCTGGGCTCAAGCAATCCTTCCACCTCTGCCTCCTGAGTAATTGGGACTACAGGTATTATGTTTCTATTAATGCAGTCTAAAGGGTTTTGGAAATCTG
GATTGTAGCAGGGCATTGTGGATCTGCCTGTTGGTGGAGAACCCTAGACTGCATGCAGCCATGCACATGATCTGGAGACACAGCAGAGACAAGGTCTTTG
TCTGTTGCTCAGGACGGAGTGTAGTGGTGCCATAGCACACTGTAATCTTGAATTTCTGGACTACTGGCATACACCACCATGCTTAGGTTCGGGCCACAAC
ATCCGTGACTTTACCGGTTTTGGTCTCTTCCTTGTAACTTTCGGCTCCAGGATGCTGCGCAGCTCCTCTCTGCACTATCTATCAGAACGTTGCTCTGCGT
GCTTGAAAGTTTCGCCAAACAGAAGTACACAAAAACAAACAGAGGCGAAACACAAATCCCCTCAGCCCTGGAGGAAAAGGCCTTGAAGCTCTTCCAGTGC
AGAGGCCACGCAGAGGACATGCTCCTACCTGTGCACATCGGCCACGCCTGGGGTGCCTCCTGGAGTCCAGGCTGCTGGCGATGGAATCCATCATGTTGCT
GATGTCACTGATTACGCCCATACGTTGACCTGTGTTACTGAAAGAGAAAAGTTTATTGTACCCATCAAAGCTAGAGGGGCACGAGCCATTCTCGATTTTC
CTGACAAGCTGAATTTTTCCACTTGTCCTGTCAAATACAGCACCCAGAAGATTCTGCTGGTACGAAACATTGGCAACAAAAATGCTGTATTTCACATCAA
AACTTGTAGGCCTTTCTCTATAGAACCAGCTATTGGAACTCTTAATGTGGGAGAGTCCATGCAACTGGAAGTGGAGTTTGAGCCACAGAGTGTGGGCGAT
CACAGTGGAAGACTTATCGTGTGTTATGACACAGGTGAAAAGGTGTTTGTATCTCTCTATGGAGCTGCCATAGACATGAATATAAGGCTGGATAAGAATT
CCTTGACCATCGAGAAAACCTACATATCTCTGGCCAATCAGCGAACTATAACCATTCACAATCGCAGTAATATCATTGCCCATTTCCTGTGGAAGGTATT
TGCTACCCAGCAAGAAGAGGACAGAGAAAAATATAGGGCCTGTGATGATCTGATCAAAGAGGAGAAGGATGAGACTGATGAGTTTTTTGAAGAGTGCATT

ACTGATCCTTTACTCCGAGAACATCTTTCTGTTCTGTCCCGAACCTTTGCGAATCAAAGGAGGCTGGTGCAGGGAGACAGCAAACTGTTCTTCAATAACG
TTTTCACTGTGGAGCCCCTGGAAGGTGATGTCTGGCCCAACTCATCAGCTGAAATCACCGTGTACTTTAACCCACTAGAAGCCAAGCTCTATCAACAGAC
CATTTACTGCGACATTTTAGGCCGAGAAATCCGTCTGCCCCTCCGAATCAAAGGGGAAGGCATGGGACCTAAGATTCACTTCAACTTTGAATTGCTGGAT
ATTGGGAAAGTTTTCACTGGATCTGCACATTGTTATGAGGCGATACTGTACAACAAAGACAGCATCGATGCTCTCTTCAACATGACCCCTCCAACTTCAG
CTTTGGGGGCCTGCTTTGTTTTCAGTCCCAAGGAAGGCATCATTGAACCAAGTGGAGTCCAAGCTATCCAGATCTCCTTCAGCTCTACCATCCTGGGAAA
CTTTGAAGAAGAGTTCCTGGTCAATGTCAATGGGTCACCTGAGCCTGTGAAACTGACCATTAGAGGCTGTGTCATTGGACCTACCTTCCATTTTAATGTT
CCAGCTCTGCACTTTGGTGATGTTTCCTTTGGGTTTCCTCATACCTTGATATGTTCCCTCAATAATACCTCTTTGATCCCCATGACTTACAAACTGCGTA
TCCCTGGGGATGGCCTTGGCCATAAAAGCATTTCATATTGTGAGCAGCATGTGGACTACAAAAGACCGTCTTGGACCAAGGAAGAAATATCCTCAATGAA
ACCAAAAGAATTCACCATCTCTCCTGACTGTGGCACCATTCGCCCCCAGGGATTTGCTGCTATCAGGGTGACATTATGCTCCAACACTGTACAGAAATAC
GAGCTGGCACTCGTGGTGGACGTGGAGGGCATCGGAGAAGAGGTGCTGGCGCTCTTAATTGCAGCAAGGTGTGTTGTACCTGCCCTCCACCTGGTCAATA
CAGAGGTGGACTTTGGGCACTGCTTCCTGAAGTACCCGTATGAGAAAACACTCCAGCTTGCCGATCAAGATGACCTCCCAGGATTCTATGAGGTCCAGCC
TCAGGTGTGTGAGGAGGTGCCTACTGTGCTGTTTTCCAGCCCCACCCCCAGCGGGGTCATCTCCCCAAGCAGCACCATCCACATACCACTGGTCCTGGAG
ACCCAGGTCACTGGAGAACACAGATCCACGGTTTACATCTCAATCTTTGGGAGCCAGGACCCCCCTTTGGTATGTCACTTAAAGAGCGCTGGAGAAGGCC
CAGTTATCTACGTCCATCCCAATCAAGTGGACTTCGGGAATATCTACGTCCTAAAAGACTCTTCCAGGATTCTCAACCTATGCAACCAGTCCTTCATTCC
CGCATTTTTCCAGGCACACATGGCACACAAAAAATCCCTTTGGACGATTGAACCCAATGAAGGCATGGTTCCTCCAGAAACTGATGTTCAACTGGCACTG
ACCGCCAACCTGAATGACACACTGACATTCAAGGACTGTGTTATTTTGGACATTGAAAATAGCAGTACCTATCGGATTCCTGTTCAGGCTTCCGGAACTG
GTTCCACTATTGTTTCAGATAAGCCCTTTGCTCCAGAACTCAATTTGGGGGCACATTTTAGCCTGGATACCCACTATTACCACTTTAAGTTGATCAACAA
GGGACGTCGGATCCAACAGTTGTTCTGGATGAATGATAGCTTCCGACCCCAGGCCAAGCTGAGTAAGAAGGGCCGGGTTAAGAAGGGACATGCTCATGTC
CAACCCCAGCCCAGTGGCTCTCAGGAGCCCAGGGATCCACAGAGCCCCGTGTTTCATCTCCACCCCGCCAGCATGGAGCTGTACCCAGGCCAGGCAATTG
ATGTGATACTCGAAGGCTATTCTGCTACTCCCAGGATAGTGAAAGAGAAGCTGGTGTGCCACGCCATCATCGGGGCACAGAAGGGGAAGAGCTTGGTGAT
GGCTGTGAACATCACCTGTGAGTTCGTCGCACCTCTCATCCAGCTCTCCACCAAGCAGCTCATCTACCGACTGGAGAAGAAACCTAACAGTATCCTGAAA
CCTGATTACCAGCCCTTGGCCGTAAAGAACATTTCCACCCTGCCCGTGAACTTGTTGCTGTCAACATCTGGACCCTTCTTTATATGTGAGACTGATAAAT
CCCTGCTGCCGGCAACTCCTGAGCCTATTAAACTGGAAATTGATGAAGAAAAAAACCTGCTGATCAAGTTTGACCCTTCCTACAGAAACGATCTGAACAA
CTGGGTGGCAGAAGAAATTCTAGCAATTAAGTATGTGGAACACCCTCAGATAGACAGCCTGGACCTGCGCGGAGAAGTGCATTACCCCAACCTCAGCTTT
GAGACAAAGGAGCTGGATTTTGGCTGCATCCTGAACGATACTGAGCTCATTCGCTACGTTACCATCACCAACTGCAGTCCGTTGGTTGTGAAGTTTCGCT
GGTTCTTCTTGGTGAATGATGAGGAAAATCAGATAAGGTTTGTGACATTGCCAAAGAAGCCCTACAGTGCCCCACTGTCCCAGATGGAGTCCATCCCAGC
AACCTCAGAGGCTGCCAGCCCACCAGCAATCCTAGTTACAGTAGAGTCCCCCGAGATGGATTTAAATGATTTTGTTAAGACTGTCCTTGTGGATGAAGAT
GCCAGGCCTGAAGAAAAAGAACTAAGAAAAACAAAAGCTTCCAGTGTGATCTCAGATGAAATAAAAATTAGCTCTACTGAAATAGAAAGAATATACTCAA
GCCAGAGCCAGGTGGAGGATCAGGAATCCCTACAGACCTGTGAACAGAATGAGATGCTTTCCATTGGGATAGAAGAAGTGTTTGATATTTTGCCCCTGTT
TGGAGTGTTGCAGCCACACAGTAGCCACCAAATATCGTTCACCTTCTATGGACACGCTAACATCATTGCACAAGCTAAAGCTCTGTGTGAAGTGGAAGAA
GGACCCACCTACGAAATAACACTGAAGGGAGAGGCGTCCCTGGTCAACTATTCCTTTGACACCAAGGATATTCACTACGGATTACAGCTGTTTGACCATG
TCACAGAGAGGGAAATCACGCTGACGAACATGGGGAAAGTTGGCTTTGAGTTCAAGGTTCTGACTGACCACCAGTCTTCTCCAGACAACCTTCTCCCTGG
AGTGCCACTAATCCTGCCTGTGTCTGGCTTTATCAGTTCACATCAAGAGCAGGTATTAAAAGTTTACTACCTACCTGGAGTACCTGAGGTCTTTAAAAGG
AGTTTCCAGATACAGATCGCCCACCTGGACCCAGAAAATATCACTCTGAGCGGAGAGGGAATCTTTCCCCAAATCTGCCTCGATCTCCCCAGGAACCTCA
CAGCAAATGAAAAGTATGAAATGTTCTTGAATCAAGCCAGGAAAAACACAGACAAAGAGTATAACAAATGTGAAATGCTCGATCACTTTGACGTAATAAC
TGAGGAAGTGCCAGAAGACGAGCCTGCTGAGGTAAGTGCTCATCTCCAGATGGAGGTAGAAAGACTTATAGTCCAAAGCTATGTCCTAGAACATCAGAAA
ACAACCACCCCTGATCCTATGGATGACCCCTGCTTCAGCCATCGGAGTCGCCGCAAACTGGCCAAAATCCAGCTACCAGAGTACATCCTGGACTTTGGCT
ACATCATCCTTGGCGAAGTCCGAACCCACATCATCAAGATCATCAACACCAGTCACTTTCCAGTGTCATTCCATGCAGACAAGCGTGTCCTTCATGAGAC
AGGATTCAGTACTGAGCTAGATCGTGTAAAGAATCTGCCTCATTGTGAAACGGAAATATTTGAAGTGAGATTTGACCCACAGGGGGCCAATCTTCCTGTT
GGAAGCAAAGAAGTCATTCTGCCCATCAAGGTGGTTGGAGGGCCAACAGTTCACATCTGTCTCCAAGCCAAGGTGACCATTCCAACCATGACTCTCTCTC
GTGGAAAAGTGGACTTTGCCACAATTCAGTGTGGACAGTGCCTGGTGGAAACTATTCAGCTTTCCAATCATCTCCAAGTCCCTTGTGAATGGTTCGTCCA
GAGCCAAAAGCCTGTTGACAAGCTGGAGAAACACATGCCGAAGTACTTAAGACAGAAACTACGCGCTGAATTAAAGCCAAAGACACGGATCTTCGAAATC
CAGCCCATTTCTGGAGTCTTGGATCCTGGTGAGAAGTCCAACGTGCAAGTGAAATTCATGCCAAAAGAAGAGAAATTCTACAGCCAAACCCTGGTGTTTC
AGATTGCCCAGAGTGCTCAAAAGCTTACCCTCCTGGCACGTGGGCAAGGTCTAGAGCCACGCCTGGAATTTAGTCCTTCAGTCCTGGATCTGGGGCCACT
GCTACTTTGTGCACCTGGAGACGAGGCCGAGGTGATAGTGAAGAATCCCTGCAACTTCCCCATTGAGTTTTATTCCTTAGAATTTGATCAGCAGTATCTC
ATAGAAGAAAAGATCTTGCGGAAGCTGAAGGGCTATGATTCCTACAACACCCTGCTGCTGCCTCCCCGCAACCCTGGGGAGAAGCTGCCCCCAGAACTGT
ACGAGTACTTCAAAGAGATAAAGAAGTCAAAAGAGGAGCAGATGAGGGCGAAATATCTGGAGAATCTGGCACAGGAGAATGAAGAGGAAGATATAACCTC
ATCAGATCAGGGAACCTCCAATAGCACAAAGAGGACATCGCTGAGCCGAGGGATCTCTGTCACATCCAACCTGGAAGAATGGCACGCCCTGTTGGTCGAG
TCCAAAACCTACCTAGAGGAAGAGGAGGATGAGGAAAGCCTGGAAAAAATCATTTTCCAAACTGACAAGCTTCAGAGCATTGACAGCCACTCCATGGAGG
AAGTTGGAGAGGTGGAAAACAACCCAGTGAGCAAAGCAATCGCTCGCCACCTGGGCATTGACATTTCTGCAGAAGGCCGCCTGGCCAAGAACCGGAAAGG
CATCGCCATTATCATTCACGGGACACCCTTGTCAGGAAAGTCAGCCAATGCCGTTAGCGTGGCCAAGTACTACAACGCAGCCTGCCTGAGCATCGACTCC
ATTGTGCTGGAAGCTGTGGCCAACAGCAACAACATCCCAGGGATCCGGGCCTGTGAGCTCTGCATCAGGGCTGCCATAGAGCAGTCCATGAAGGAAGGAG
AGGAGGCTGCTGAGTGACTGCTACCGAGGAGTGGTGTTTGATGGCCTCGACACTCTCTTTGCTCGGAATGCTGCAGCCGCCCTCCTCTGCCTGCTGAAGG
CCATTGGCAGCCGGGAGCATATATACATTCTCAACATGGCCCAGGATTACGCAGCCATGAAGGCCCGGGAGAAAGCCAAAAAGGAGCAAGAAGGCAAGGC
CTCCTACCTTTCCATTGGGCTTAAGTGAAACTTTGACATTTGCTATTCCGTTGGCGCCTCACCTCTGCCCTGTAAGGTGGCCAGGAGTTGTATTTCTGCT
TTGCTAGGAGGGAACTGGGACGCGAGTGGCTGGCTGGGAAGTTCTATGATGAACGCTTTGCTGACAGCAGGTGCATTCCTGCTTCCCTCCCATCACACTG
TACCCCTCCATCCTTTCTGCAATTACAGCCATTTTGGGTCCC
>hydin_transript_G
AGGTACTAGCGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAA
GGCCTTGGAAAATCGGAGCGATTAGAGCGGTTCAAACAATACAGAAGTGTATAAAGTAAGAATTGGAAGTCTCTTCGCTGCTCTAATCTCACTCCTCAGA
GGTGATCACTGGACCTGAGTGAGCAGAAGGCTTTTCAGAATTACAGCCTGAACTATTTCTTGAGCAATTTGGGAACTCATTTCTAAAGCAGTCCTCAGAT
TCACACAACGTAGTTTGTGACGGAGTATGCCCAGAGCAGTGTGGAAGATGATCAAAGTATGACATTTCTTCTACCCCAGAGGCTGGCATTTCTGCAGGCA
CAAATGAGCTTGGTTTTCAAGCACTGCTCTTTGAACACAAGGAATTCCTTTAATGATCACGCCGGACTATATAGACTTTGGAATCAGACTCTTTGAATTC
CAGCTCTGCCATTTACCAACTTTGACACTTTGGACAACGTCTCATTCTGTCACCCAGGCTGGAGTGCAATGGCACAATTATAGCTCACTGCAGCCTCAAA
CACCTGGGCTCAAGCAATCCTTCCACCTCTGCCTCCTGAGTAATTGGGACTACAGGTATTATGTTTCTATTAATGCAGTCTAAAGGGTTTTGGAAATCTG
GATTGTAGCAGGGCATTGTGGATCTGCCTGTTGGTGGAGAACCCTAGACTGCATGCAGCCATGCACATGATCTGGAGACACAGCAGAGACAAGGTCTTTG
TCTGTTGCTCAGGACGGAGTGTAGTGGTGCCATAGCACACTGTAATCTTGAATTTCTGGACTACTGGCATACACCACCATGCTTAGGTTCGGGCCACAAC
ATCCGTGACTTTACCGGTTTTGGTCTCTTCCTTGTAACTTTCGGCTCCAGGATGCTGCGCAGCTCCTCTCTGCACTATCTATCAGAACGTTGCTCTGCGT
GCTTGAAAGTTTCGCCAAACAGAAGTACACAAAAACAAACAGAGGCGAAACACAAATCCCCTCAGCCCTGGAGGAAAAGGCCTTGAAGCTCTTCCAGTGC
AGAGGCCACGCAGAGGACATGCTCCTACCTGTGCACATCGGCCACGCCTGGGGTGCCTCCTGGAGTCCAGGCTGCTGGCGATGGAATCCATCATGTTGCT
GATGTCACTGATTACGCCCATACGTTGACCTGTGTTACTGAAAGAGAAAAGTTTATTGTACCCATCAAAGCTAGAGGGGCACGAGCCATTCTCGATTTTC
CTGACAAGCTGAATTTTTCCACTTGTCCTGTCAAATACAGCACCCAGAAGATTCTGCTGGTACGAAACATTGGCAACAAAAATGCTGTATTTCACATCAA
AACTTGTAGGCCTTTCTCTATAGAACCAGCTATTGGAACTCTTAATGTGGGAGAGTCCATGCAACTGGAAGTGGAGTTGAGCCACAGAGTGTGGGCGAT
CACAGTGGAAGACTTATCGTGTGTTATGACACAGGTGAAAAGGTGTTTGTATCTCTCTATGGAGCTGCCATAGACATGAATATAAGGCTGGATAAGAATT
CCTTGACCATCGAGAAAACCTACATATCTCTGGCCAATCAGCGAACTATAACCATTCACAATCGCAGTAATATCATTGCCCATTTCCTGTGGAAGGTATT
TGCTACCCAGCAAGAAGAGGACAGAGAAAAATATAGGGCCTGTGATGATCTGATCAAAGAGGAGAAGGATGAGACTGATGAGTTTTTTGAAGAGTGCATT
ACTGATCCTTTACTCCGAGAACATCTTTCTGTTCTGTCCCGAACCTTTGCGAATCAAAGGAGGCTGGTGCAGGGAGACAGCAAACTGTTCTTCAATAACG
TTTTCACTGTGGAGCCCCTGGAAGGTGATGTCTGGCCCAACTCATCAGCTGAAATCACCGTGTACTTTAACCCACTAGAAGCCAAGCTCTATCAACAGAC
CATTTACTGCGACATTTTAGGCCGAGAAATCCGTCTGCCCCTCCGAATCAAAGGGGAAGGCATGGGACCTAAGATTCACTTCAACTTTGAATTGCTGGAT
ATTGGGAAAGTTTTCACTGGATCTGCACATTGTTATGAGGCGATACTGTACAACAAAGACAGCATCGATGCTCTCTTCAACATGACCCCTCCAACTTCAG
CTTTGGGGGCCTGCTTTGTTTTCAGTCCCAAGGAAGGCATCATTGAACCAAGTGGAGTCCAAGCTATCCAGATCTCCTTCAGCTCTACCATCCTGGGAAA
CTTTGAAGAAGAGTTCCTGGTCAATGTCAATGGGTCACCTGAGCCTGTGAAACTGACCATTAGAGGCTGTGTCATTGGACCTACCTTCCATTTTAATGTT
CCAGCTCTGCACTTTGGTGATGTTTCCTTTGGGTTTCCTCATACCTTGATATGTTCCCTCAATAATACCTCTTTGATCCCCATGACTTACAAACTGCGTA

TCCCTGGGGATGGCCTTGGCCATAAAAGCATTTCATATTGTGAGCAGCATGTGGACTACAAAAGACCGTCTTGGACCAAGGAAGAAATATCCTCAATGAA
ACCAAAAGAATTCACCATCTCTCCTGACTGTGGCACCATTCGCCCCCAGGGATTTGCTGCTATCAGGGTGACATTATGCTCCAACACTGTACAGAAATAC
GAGCTGGCACTCGTGGTGGACGTGGAGGGCATCGGAGAAGAGGTGCTGGCGCTCTTAATTGCAGCAAGGTGTGTTGTACCTGCCCTCCACCTGGTCAATA
CAGAGGTGGACTTTGGGCACTGCTTCCTGAAGTACCCGTATGAGAAAACACTCCAGCTTGCCGATCAAGATGACCTCCCAGGATTCTATGAGGTCCAGCC
TCAGGTGTGTGAGGAGGTGCCTACTGTGCTGTTTTCCAGCCCCACCCCCAGCGGGGTCATCTCCCCAAGCAGCACCATCCACATACCACTGGTCCTGGAG
ACCCAGGTCACTGGAGAACACAGATCCACGGTTTACATCTCAATCTTTGGGAGCCAGGACCCCCCTTTGGTATGTCACTTAAAGAGCGCTGGAGAAGGCC
CAGTTATCTACGTCCATCCCAATCAAGTGGACTTCGGGAATATCTACGTCCTAAAAGACTCTTCCAGGATTCTCAACCTATGCAACCAGTCCTTCATTCC
CGCATTTTTCCAGGCACACATGGCACACAAAAAATCCCTTTGGACGATTGAACCCAATGAAGGCATGGTTCCTCCAGAAACTGATGTTCAACTGGCACTG
ACCGCCAACCTGAATGACACACTGACATTCAAGGACTGTGTTATTTTGGACATTGAAAATAGCAGTACCTATCGGATTCCTGTTCAGGCTTCCGGAACTG
GTTCCACTATTGTTTCAGATAAGCCCTTTGCTCCAGAACTCAATTTGGGGGCACATTTTAGCCTGGATACCCACTATTACCACTTTAAGTTGATCAACAA
GGGACGTCGGATCCAACAGTTGTTCTGGATGAATGATAGCTTCCGACCCCAGGCCAAGCTGAGTAAGAAGGGCCGGGTTAAGAAGGGACATGCTCATGTC
CAACCCCAGCCCAGTGGCTCTCAGGAGCCCAGGGATCCACAGAGCCCCGTGTTTCATCTCCACCCCGCCAGCATGGAGCTGTACCCAGGCCAGGCAATTG
ATGTGATACTCGAAGGCTATTCTGCTACTCCCAGGATAGTGAAAGAGAAGCTGGTGTGCCACGCCATCATCGGGGCACAGAAGGGGAAGAGCTTGGTGAT
GGCTGTGAACATCACCTGTGAGTTCGTCGCACCTCTCATCCAGCTCTCCACCAAGCAGCTCATCTACCGACTGGAGAAGAAACCTAACAGTATCCTGAAA
CCTGATTACCAGCCCTTGGCCGTAAAGAACATTTCCACCCTGCCCGTGAACTTGTTGCTGTCAACATCTGGACCCTTCTTTATATGTGAGACTGATAAAT
CCCTGCTGCCGGCAACTCCTGAGCCTATTAAACTGGAAATTGATGAAGAAAAAAACCTGCTGATCAAGTTTGACCCTTCCTACAGAAACGATCTGAACAA
CTGGGTGGCAGAAGAAATTCTAGCAATTAAGTATGTGGAACACCCTCAGATAGACAGCCTGGACCTGCGCGGAGAAGTGCATTACCCCAACCTCAGCTTT
GAGACAAAGGAGCTGGATTTTGGCTGCATCCTGAACGATACTGAGCTCATTCGCTACGTTACCATCACCAACTGCAGTCCGTTGGTTGTGAAGTTTCGCT
GGTTCTTCTTGGTGAATGATGAGGAAAATCAGATAAGGTTTGTGACATTGCCAAAGAAGCCCTACAGTGCCCCACTGTCCCAGATGGAGTCCATCCCAGC
AACCTCAGAGGCTGCCAGCCCACCAGCAATCCTAGTTACAGTAGAGTCCCCCGAGATGGATTTAAATGATTTTGTTAAAGACTGTCCTTGTGGATGAAGAT
GCCAGGCCTGAAGAAAAAGAACTAAGAAAAACAAAAGCTTCCAGTGTGATCTCAGATGAAATAAAAATTAGCTCTACTGAAATAGAAAGAATATACTCAA
GCCAGAGCCAGGTGGAGGATCAGGAATCCCTACAGACCTGTGAACAGAATGAGATGCTTTCCATTGGGATAGAAGAAGTGTTTGATATTTTGCCCCTGTT
TGGAGTGTTGCAGCCACACAGTAGCCACCAAATATCGTTCACCTTCTATGGACACGCTAACATCATTGCACAAGCTAAAGCTCTGTGTGAAGTGGAAGAA
GGACCCACCTACGAAATAACACTGAAGGGAGAGGCGTCCCTGGTCAACTATTCCTTTGACACCAAGGATATTCACTACGGATTACAGCTGTTTGACCATG
TCACAGAGAGGGAAATCACGCTGACGAACATGGGGAAAGTTGGCTTTGAGTTCAAGGTTCTGACTGACCACCAGTCTTCTCCAGACAACCTTCTCCCTGG
AGTGCCACTAATCCTGCCTGTGTCTGGCTTTATCAGTTCACATCAAGAGCAGGTATTAAAAGTTTACTACCTACCTGGAGTACCTGAGGTCTTTAAAAGG
AGTTTCCAGATACAGATCGCCCACCTGGACCCAGAAAATATCACTCTGAGCGGAGAGGGAATCTTTCCCCAAATCTGCCTCGATCTCCCCAGGAACCTCA
CAGCAAATGAAAAGTATGAAATGTTCTTGAATCAAGCCAGGAAAAACACAGACAAAGAGTATAACAAATGTGAAATGCTCGATCACTTTGACGTAATAAC
TGAGGAAGTGCCAGAAGACGAGCCTGCTGAGGTAAGTGCTCATCTCCAGATGGAGGTAGAAAGACTTATAGTCCAAAGCTATGTCCTAGAACATCAGAAA
ACAACCACCCCTGATCCTATGGATGACCCCTGCTTCAGCCATCGGAGTCGCCGCAAACTGGCCAAAATCCAGCTACCAGAGTACATCCTGGACTTTGGCT
ACATCATCCTTGGCGAAGTCCGAACCCACATCATCAAGATCATCAACACCAGTCACTTTCCAGTGTCATTCCATGCAGACAAGCGTGTCCTTCATGAGAC
AGGATTCAGTACTGAGCTAGATCGTGTAAAGAATCTGCCTCATTGTGAAACGGAAATATTTGAAGTGAGATTTGACCCACAGGGGGCCAATCTTCCTGTT
GGAAGCAAAGAAGTCATTCTGCCCATCAAGGTGGTTGGAGGGCCAACAGTTCACATCTGTCTCCAAGCCAAGGTGACCATTCCAACCATGACTCTCTCTC
GTGGAAAAGTGGACTTTGCCACAATTCAGTGTGGACAGTGCCTGGTGGAAACTATTCAGCTTTCCAATCATCTCCAAGTCCCTTGTGAATGGTTCGTCCA
GAGCCAAAAGCCTGTTGACAAGCTGGAGAAACACATGCCGAAGTACTTAAGACAGAAACTACGCGCTGAATTAAAGCCAAAGACACGGATCTTCGAAATC
CAGCCCATTTCTGGAGTCTTGGATCCTGGTGAGAAGTCCAACGTGCAAGTGAAATTCATGCCAAAAGAAGAGAAATTCTACAGCCAAACCCTGGTGTTTC
AGATTGCCCAGAGTGCTCAAAAGCTTACCCTCCTGGCACGTGGGCAAGGTCTAGAGCCACGCCTGGAATTTAGTCCTTCAGTCCTGGATCTGGGGCCACT
GCTACTTTGTGCACCTGGAGACGAGGCCGAGGTGATAGTGAAGAATCCCTGCAACTTCCCCATTGAGTTTTATTCCTTAGAATTTGATCAGCAGTATCTC
ATAGAAGAAAAGATCTTGCGGAAGCTGAAGGGCTATGATTCCTACAACAACCCTGCTGCTGCCTCCCCGCAACCCTGGGGAGAAGCTGCCCCCAGAACTGT
ACGAGTACTTCAAAGAGATAAAGAAGTCAAAAGAGGAGCAGATGAGGGCGAAATATCTGGAGAATCTGGCACAGGAGAATGAAGAGGAAGATATAACCTC
ATCAGATCAGGGAACCTCCAATAGCACAAAGAGGACATCGCTGAGCCGAGGGATCTCTGTCACATCCAACCTGGAAGAATGGCACGCCCTGTTGGTCGAG
TCCAAAACCTACCTAGAGGAAGAGGAGGATGAGGAAAGCCTGGAAAAAATCATTTTCCAAACTGACAAGCTTCAGAGCATTGACAGCCACTCCATGGAGG
AAGTTGGAGAGGTGGAAAACAACCCAGTGAGCAAAGCAATCGCTCGCCACCTGGGCATTGACATTTCTGCAGAAGGCCGCCTGGCCAAGAACCGGAAAGG
CATCGCCATTATCATTCACGGGACACCCTTGTCAGGAAAGTCAGCCAATGCCGTTAGCGTGGCCAAGTACTACAACGCAGCCTGCCTGAGCATCGACTCC
ATTGTGCTGGAAGCTGTGGCCAACAGCAACAACATCCCAGGGATCCGGGCCTGTGAGCTCTGCATCAGGGCTGCCATAGAGCAGTCCATGAAGGAAGGAG
AGGAGGCTGCTGAGTGACTGCTACCGAGGAGTGGTGTTTGATGGCCTCGACACTCTCTTTGCTCGGAATGCTGCAGCCGCCCTCCTCTGCCTGCTGAAGG
CCATTGGCAGCCGGGAGCATATATACATTCTCAACATGGCCCAGGATTACGCAGCCATGAAGGCCCGGGAGAAAGCCAAAAAGGAGCAAGAAGAACGCAA
GCACAAGGGAGCTCTTGAGAAAGAGAAGGAGGGTCTCCAAAACATGGATGAGGAAGAATATGATGCCCTGACTGAGGAGGAGAAACTCACATTCGATCGG
GGGATTCAGCAGGCGCTCCGCGAGCGGAAGAAGAGGGAGCAGGAGAGGCTGGCAAAGGAAATGCAAGAAAAGAAGCTACAGCAGGAGCTGGAGCGACAAA
AGGAAGAGGATGAGCTGAAACGGAGGGTCAAAAAAGGAAAGCAGGGACCCATTAAGGAGGAGCCCCCCATGAAGAAATCTCAAGCAGCAAACAAGCAGGT
TCCTCCGCTCACCAAAGTGGATGTCAAGATGGAGACAATCGAAAGGAAAATATCTGTTAGGGAACAAACAATGTCTGAGAAGGAAGAGCTAAATAAGAAG
AAAAGGAACATGGGCGATGTCAGCATGCATGGGCTTCCTCTTGTCCAGGACCAAGAGGACAGTGAAGGGGACATCTCAAAGGACCCCGACAAGCAACTGG
CCCAGAAGTTTAAGACCTATGAATTGACACTGAAGGATGTCCAGAACATCCTCATGTACTGGGACCGGAAGCAAGGAGTCCAGCTGCCTCCTGCAGGGAT
GGAGGAAGCTCCCCATGAGCCCGACGACCAGCGCCAGGTCCCCTCGGGTGGGCGCAGGGGCCGCAAGGACCGGGAGAGAGAGCGCCTGGAGAAGGAGCGC
CTGGAGAGGGAGAAGGCGGAGCGGGAGCGCCTGGAGAAGCTGCGAGCCCTGGAGGAGCGGAGCGACTGGGAGGGGAAGGGGAGGAGGACCACGAAGAGA
AGAAGGAGAAGGACCTGGGCGTACCCTTCCTTAACATCCAGACACCAGACTTTGAAGGCTTGAGCTGGAAGCAGGCCCTAGAGAGCGACAGGCTTCCCAA
AGGAGAGCAGATCCTAGACATCTTGGGTCTGGGTGCCTCCGGACCACCCATCCCGCCTCCCGCCTTATTCTCAATCGTCTCCTACCCGGTGAAGCGGCCA
CCTTTGACCATGACAGACGACCTGGAGCATTTTGTATTTGTGATCCCACCATCCGAAGATATATCTCTGGATGAAAAAAAGAAATGGAAATAGAATCAG
ACTTTTTGGCCACCACAAACACTACAAAGGCTCAAGAGGAGCAGACCAGCTCATCTAAGGGGGGCAAACAGAAAATGAAAGAAAAGATAGACCAAGTCTT
CGAGAGTCAGAAAGACAAGCGTCACATGGCCTTAAACAGGAAGGTCCTTTCTGGGGAACCTGCTGGAACCATTTCCCAGCTGTCAGATACAGACCTGGAC
AACTTCAACGGGCAGCACTCCCAGGAGAAATTCACCAGACTGAATCACTTCCGGTGGATCGTGCCAGCCAATGGCGAGGTAACGTTGCAGGTGCACTTCT
CTTCTGATGAGTTCGGGAACTTTGACCAAACCTTTAACTTTGAGATCCTAGGAACTTGCTGCCAGTACCAGCTCTACTGCCGAGGCATCTGCACTTACCC
ATACATTTGCCAAGACCCAAAAGTGGTATTTCCTCAGCGGAAGATGGACATGAAGACAAATGAGGTCATCTTTAAGAAGTATGTTATGAGCACGGAGACG
TACTACTTTGGGCCACTACTTTGTGGAAAATCAAGAGATAAGTACAAGTCATCCTTATTCCCAGGCAACATGGAGACGCTAACAATCCTGAACACTTCCT
TAATGGTGGTGGAGGCATCCTTCTATTTTCAGAATGATGTCAAAGCAAACACGTACTTCCTGGAACCCAACACCATGGTCCTGAAACCCAATGAGAAGCA
GATATTAAACGTATGGGCCTACCCTACTTCAGTTGGTGTCTTTGAAGACAGCATTGTCTGCTGCATCAATGACAACCCAGAGCCAGCCATCTTCCAATTA
AGCTGCCAGGGGATCCGCCCGGAACTAGAGCTGGAACCCAGGCAATTACATTTTGACCGGCTCTTGCTGCACAGACAGGAATCCAGGGTAGTTCTTCTCC
GCAATGTCACGCTCCTGCCTGTGGCCTGGCGGATCACCAGCCTGGAGCACTTGGGTGATGATTTCACTGTATCCCTGATGCAGGGGACCATCCCCCCTGA
GGCTGAGTACGGCCTGCACCTGTACTTTCAGCCCACCAAGCCTGTCAACATCAAGAAGGCTATTCGGTTGGAGGTTTTAGATGCAGAAAATCTTCTTGGT
GTTGTTCAGATTGAAAATATCATGGTCTTTGCAGAGGCATACGACATCGCCTTGGACATCACCTTCCCCAAAGGAGCTGAAGGGGGACTGGATTTTGGGA
TTGTCAGGGTCACAGAGGAGGCGAAGCAGCCCCTGCAATTGAAGAACCGTGGGAAATATGAGATCGCGTTCAGCTTTTCCGTGGACTCTGTAGGGATTTC
AACACCTAATATAAATTCCATGATCTCAGTCCAACCCAAAAAGGGTTCACTGACCCCAACAGAAAAACCCACAAATGTCCAAGTTTTCTTCCATGCAAAA
AAGGAAGTGAAGATTGAGCACCAGCCTGTTCTGCGCTGTCAGATTATTGAGCCCAATATTTCAGAAGGAGGTGAGATCATTGCCAGCATCCCAATTAAGT
TTTCCGCGAATGCAGTATATTCCAAATACAACATCACCCCCTCCTCTGTCATCAACTTTGGAGCTTTGATCTGTGGCACTCGTAAAAGCACCACCTTCAC
CATAGAAAATCAAGGTGTTACTGACTTCAAGTTCGCCCTTTATAAGCTGACAGGGGAGAGCCCCATTCATCAAAAGAAAGCAGCCAGCCACGTCAGACAT
GCAAGATCCCGAGAAAGTGAGAGCTTCTACAAAACTGGCTCTTCCAGAGCAGCCAAGTTCTCTGACACGATTCAGAAAGAAGTAACCACCACAGGCCAGG
CCCGCTTCGCCCATGGCATGTTCACCGTGTACCCTGGGTTTGGCTCCATTCCTTCCGGAGGACAGCAGGTCATCAACGTTGACTGTGTGGCTGACCCCAT
GGGAAAGTGTGAGGAGTTTATAGCCATCGATATCTCCGGCCGAGACCCTGCAGTCCACCCTGCCGGCATTCCTTACACTTTGCTAGCTGAAGCCTGTCTA
CCAGCCTTCGTGACCGAAAACAATGCCTTGATATTTGAAGAGCACCAGATATGTACCAGTGCCAACCTGCACCACATCCTGCAGACCATAGAGAGCGGGG
GGCTGTTCGTCGAGGATGAGAACAAGTTCATCTTCTGCAATGTCCTGGTGGGCCGCCAAGCCAAGGCTCGTTTCAAGATCAGCAACGTGGGAAAGATCAC
CTGTGATGTCAACATTGTAGTCAGGCCTATCTCCAATAAGCCCTTTGCCCGCATCGTCGACATTTTTGAAGTGGAACCCAGCAAGATGTGCATTGCCAGT
CATTCCCATGCCTTTGCCACGGTGTCCTTCACCCCGCAGATCATGCAGAACTACCAGTGCATCTTTGAGGCTACCTTGGATGGCTTGCCCAGCACTCTGG
CCAAGAGCCGAGGCCTCGTGTTTGACATCGCTGGTGAGGGGAACCTCCCTCGAGTGACGGTTGTGCGGCCAGTTCTTCATAACCAATATGGAAACCCCTT

GCTCCTCTTTAAGAGGCTTCTCCTTGGTCATTCAGAGAAGCTGCCTCTCATCCTCAAGAACAATGGTGTCCTCCCTGCCCAGCTGCATGTTGACCTGCAG
GATGAGCTAGGAGTCTTCTCCCTGAAAGGGAGGCCCACCACCGCGTATATCTACATCACAGAGGAAAACAAACCACATGTAAAAGCAAAGAAAGCTCACA
CAGCCTCCTTGGTTGTTTCTCCTGGAGATACAGCTGAATTTGATGTCGTTTTCCACTCCCAGAAGGTTGGGAGGATGAGAGGTATCATCCACTTGTCAGT
GATCAACAACCAATATGAGGAGACCTCCATCCACATGGTGGGAGAGGGCTATGAGGATGACATCACCTTGGACAACATCCATGGACTGGTGGCCCCCACC
AGCCAGGAAGACATAAGTATCTCTGAGTTCACAGAGATCATCGAGGACAATGATATGGAAGACTTGGTGGCAGCTGCTCTGGTGGACCACATCCAATTTG
GGGACTGCCACATTGGACACAGCTATAATGCGAGCTTCACAGTCACAAATCACAGCCAAGTGAACTTGATACGGTTTGAATGGCCTGTTTCAGCTACAAT
TGCTTTCTCCCCACAGATGGGCCATCTCCACCCTGGGTGTGCCAAGGACATAGTGGTGACCATGAAGTCAGATGTACCCATCAACCTAAAGAATATGCGG
ATCAGGTGCAAGCTCTCCAGGATTATGTTTCAGCTCCCTACAGACCAGATCCCCGACTGGGATGACCGCATGCACACAGTCAAGTGGGTGGACGTACCCA
GAAACATGCCTGGGACTTTCACTACAAAACGAAAAGTGATAGAGACGGATCCGGAACCTGCTCACTCAGTACTAGAAGAAAACTACCAAGAACTGCAGCT
TCAAATCAGTGCCAATGTGGATTTCGCTTCATACCATTGCCAAGCAAGAGATGTGCGCTTTAAGGAAACCTTGGTTTACCAGACCCGAGTGTTTGAGTTC
GATGTGATTAATTCAGGACGTGTCCAGCTGGAATTCAGCTGGCTCTCAGAAGATACCTCAAAGGCAGTCAGCTTTGCAAAACCAGATCACCAAGGTTCAG
CTCAAAAAGATCAGCTTAGTCAGGGCACGAGGCACACAGGCAGCACCCTGGACAGCACCATGGACCACTGGGCCGAGGGTTCCCCACAGCCCTTCTGT
GGAGCCCTCTTCGGGAATCGTGCCGGTGGGGAAGATCAGAAGTTCAAAGTAAAATTCTCCCCGTTGGACATTGGAGACTTCGAGAGCAACCTTTTCTGCC
AGATTCCCAACCTGCCACCTGGAGAGCAAGGTCCGGTCCTGGTAGCAAAAGGGCGGAGCACCTTGCCCATCTGCCATTTTGATCTGAAAGACTCGGACTA
CATAAGTGGCCATCAGCGCAACCCAGAGCTCCGAGGGTCCAGTGGGGGAGCTCTGGATCCAAACACCCGGGTGATTGAGTTCACCACTGTGGGCATAGGA
GGGAAGAATCTCCGGACCTTTACCATCCTAAACCCAGCCAATAGCACCTACTCCTTCTGCTGGATCTCTGAAGAAATTGAAAGTCTCCAGAACCCTGCAG
CATTCACATGCCTTACAGAAAAGGGCTTCATCCACCCTGAAAAGAAAGCCGAGATCGTGTTCCAGTTCACACCTTTCCATCTGGGCATCACTGAGTCATC
ATGGACCTTCCTAATTCCCGAGCACAACATCACAGTCCCTTTCCTGCTGGTAGGCAAAACTACCGAACCTCTCATCTCTCTGAACAAGTCACACCTCAAC
TTCAGCTCTCTCCTCATTGGCAGAGAAGCCAGGGAGACTGTGCAGATCATCAACAAGGAGGAGCAGGGGTTCGATTTTTCCTTCCAGGACAACTCCCGCT
ATTCTGAAGGTTTCAGCAACAGCCTGCTTGTATGTCCCATGGAAGGCTGGATCCCACCACTGTCCAGGTTCCCAATTGATATTTTCTTCACTCCAAAGCA
GCAAGGAGATGTGAACTTTAATTTGATCTGCAATGTGGAAAAGAAAGTCCACCCTGTGACATTAAATGTCAAGGCCGAGGGCTACACTATGAATGTGGAG
ATCAAGTGCAAGGACAGGACAGGCTCCATCACTCTGTTGACTCCCAACCAGACTAACATCATCAACTTCTATGAGGTGGAGTTAAATGAATGTGTCCAGT
GTGAATTCAACTTTATCAACACTGGAAAGTTCACCTTCAGCTTCCAGGCACAGCTGTGTGGCTCCAAAACCTTGCTGCAGTACTTGGAATTTTCACCCAT
CGACAGCACTGTGGATGTAGGACAGAGTGTACATGCCACCCTGTCTTTTCAACCATTAAAGAAGTGTGTCTTGACAGACCTGGAACTCATAATCAAGATC
AGCCATGGTCCAACATTTATGTGCAACATCTCAGGCTGTGCTGTGAGCCCGGCTATCCATTTCTCCTTCACCAGCTACAACTTTGGGACCTGCTTTATCT
ATCAAGCTGGGATGCCCCCATACAAACAAACCCTGGTAATTACCAACAAGGAAGAAACACCTATGAGCATAGATTGTCTGTACACCAACACCACTCACCT
CGAGGTGAACTGCCGTGTTGATGTGGTAAAGCCAGGAAACACATTGGAGATTCCAATAACTTTTTATCCTCGAGAAAGTATCAACTATCAAGAACTCATT
CCCTTTGAAATCAATGGGCTCTCACAACAAACAGTCGAAATCAAAGGGAAGGGTACCGAAATGAAGATTTTAGTCCTAGATCCAGCCAACAGGATTGTGA
AGTTGGGAGCTGTCCTACCAGGGCAGGTTGTGAAAAGAACAGTTTCCATCATGAACAACAGCCTGGCCCAGCTCACATTTAATCAGTCCATTCTGTTCAC
AATTCCAGAACTCCAGGAACCCAAGGTCCTCACCCTGGCGCCCTTCCACAACATCACACTGAAGCCCAAAGAAGTCTGTAAACTGGAAGTCATCTTTGCC
CCGAAGAAGCGTGTCCCTCCCTTCTCTGAGGAAGTGTTCATGGAATGCATGGGGCTCCTGCGCCCCCTCTTCCTCCTTAGCGGCTGCTGCCAGGCCCTGG
AGATCTCACTGGACCAGGAACATATTCCCTTTGGACCCGTGGTGTATCAGACGCAAGCCACGCGTCGCATCCTCATGTTGAACACAGGCGATGTGGGTGC
AAGGTTTAAATGGGACATCAAAAAATTTGAGCCTCATTTCTCCATTAGCCCAGAAGAAGGCTATATTACCTCAGGCATGGAGGTTTCTTTTGAAGTGACC
TACCATCCCACCGAGGTGGGAAAGGAGAGCCTTTGTAAAAAAATTCTCTGCTACATCCAGGGAGGCAGTCCTCTGAGTCTAACCCTGTCTGGAGTCTGCG
TGGGACCACCTGCAGTAAAAGAGGTAGTGAATTTCATGTGCCAGGTGCGCTCCAAGCACACGCAGACCATCCTGCTGTCAAACTGCACCAACCAGACCTG
GAATCTGCACCCCATCTTTGAGGGCGAGCACTGGGAGGGGCCTGAGTTCATCACCCTGGAGGCCCACCAGCAAAACAAGCCCTATGAGATCACCTACAGG
CCCCGCACCATGAACTTGGAGAACCGCAAGCACCAGGGCACCCTCTTCTTCCCCCTCCCAGATGGGACCGGCTGGCTGTATGCTCTGCATGGGACTTCTG
AGCTCCCCAAAGCTGTAGCCAATATCTATCGTGAAGTGCCATGTAAGACCCCCTACACTGAGCTTCTGCCAATCACCAACTGGCTGAACAAGCCCCAGAG
ATTCCGGGTCATCGTGGAAATACTGAAACCAGAGAAGCCGGACCTAAGCATCACTATGAAGGGCCTTGATTACATTGATGTACTGTCTGGCTCTAAGAAA
GACTACAAGCTGAACTTCTTTTCCCACAAGGAGGGAACGTACGCTGCAAAGGTGATCTTCCGAAACGAGGTGACAAATGAGTTCTTGTACTACAATGTGA
GTTTCAGGGTCATCCCTTCAGGCATCATCAAAACCATCGAGATGGTGACCCCAGTCCGGCAAGTTGCGTCAGCCTCCATCAAGTTGGAGAACCCTCTGCC
CTACTCGGTGACCTTCTCCACAGAATGCCAGATGCCCGACATCGCCCTGCCCTCCCAGTTTGTGGTGCCTGCCAACTCCGAGGCGGGTGATGTGAAGATA
AATGTGATAACGTCTACAAAGTCCTCCAAATTCTCGGGATGAAATGTGCTATGTAAATGGAACACAGCCCCGTTATGATTGAAGTTTAATGCAGCAGCTC
TCAGGCCTATACAGGGAATGCAGTCTGGCTGGAAGgcattctgaaggcaaatcctgttTAGACCCAGGCAAAGGTTCCTGGTGACCCAGGCTCTCATGAG
CAGATTGTCCCTTATCATTCTCCTGAGGATGTCTGGAGCTTCAGTGCTGTGTGCTCTTGGCCTCCACACTGGGGGTGCCACCAACTCCCACTGTCCAGGG
CTTCTGGTGGACTCTCCGAGGCACCGATGTAGAAACTTACCCATTCAGTGCACCAAGAGCAGCCTCACATGGTGTGGGCCGACATGAAGAGCTGCCAGAT
CCAACAGCCCTTGAAGATAAGGATGGTCAAACAAAATAATATCATACCTGGAGAAACTCAGATCTTGCTAAGATTTACTGGTTGGGAATCCAAAGTTAAT
GCCAAGAAGCAGCTGCCAGTTGGGATCAAATGTGAGCCTATGGATCAAGAAAATGAGCAAACAGGTGGCCATGAAACAGATGGTCATAGAATTGTTTCAG
TGCTTGTGAGTGCAGCAACCCAAGAGTGTCTTATCTGAAATACCACCAGGAATGTCTGGACACAGTAGACAAAGGTTTTTCAACTGGATGCCTTAGGATA
CATGCTTCCAAAAACAAAGTAGCCAAAAAGAAACCagagtcacagaatatcagagccagaggaacatttggagACGGTTACCTGGCACGCTGGCCACAAT
CTACCTCACTCTTATCAGAGTCTGAGCAGTGCTTTCAGCTCTGAGTTGAGGCACCTCGAACCTTGTTTTTGTGGTGAAGGATCCTAAAGTGCTGTGGGGA
GTGATCACATTTTTCACAACATCCCTGACTCCACCTCTTCTGCCACAAACGTCAGCATGGTGGTATCAGCCGGCCCTTGGTCCAGCGAGAAGGCAGAGAT
GAACATTCTAGAAATCAACGAGAAATTGCGCCCCCAGTTGGCAGAGAACAAACAGCAGTTCAGAAACCTCAAAGAGAGATGTTTTCTAACTCAACTGGCC
GGCTTCCTGGCCAACCGACAGAAGAAATACAAGTATGAAGAGTGTAAAGACCTCATAAAATTTATGCTGAGGAATGAGCGACAGTTCAAGGAGGAGAAGC
TTGCAGAGCAGCTGAAACAAGCTGAGGAGCTCAGGCAATATAAAGTCCTGGTTCACTCTCAGGAACGAGAGCTGACCCAGTTAAGGGAGAAGTTACGGGA
AGGGAGAGATGCCTCCCGCTCATTGAATGAGCATCTCCAGGCCCTCCTCACTCCGGATGAGCCGGACAAGTCCCAGGGGCAGGACCTCCAAGAACAGCTG
GCTGAGGGGTGTAGACTGGCACAGCAACTTGTCCAAAAGCTCAGCCCAGAAAATgatgaagatgaggatgaagatgttcaagttgaggaggatgagAAAG
TACTGGAATCATCTGCCCCCAGGGAGGTGCAGAAGGCTGAAGAGAGCAAAGTCCCTGAGGACTCACTGGAGGAATGTGCCATCACTTGTTCAAATAGCCA
CGGCCCTTGTGACTCCATCCAGCCTCACAAGAACATCAAAATCACATTTGAGGAAGACAAAGTCAACTCAACTGTGGTTGTAGACAGAAAATCCTCTCAT
GATGAATGTCAGGATGCTCTAAACATTCTCCCAGTCCCTGGCCCCACCTCTTCTGCCACAAACGTCAGCATGGTGGTATCAGCCGGCCCTTTGTCCAGCG
AGAAGGCAGAGATGAACATTCTAGAAATCAATGAGAAGTTGCGCCCCCAGCTGGCAGAGAAGAAACAGCAGTTCAGAAGCCTCAAAGAGAAATGTTTTGT
AACTCAACTGGCCGGCTTCCTGGCCAAGCAGCAGAACAAATACAAATATGAAGAGTGCAAAGACCTCATAAAATCTATGCTGAGGAATGAGCTACAGTTC
AAGGAGGAGAAGCTTGCAGAGCAGCTGAAGCAAGCTGAGGAGCTCAGGCAATATAAAGTCCTGGTTCACTCTCAGGAACGAGAGCTGACCCAGTTAAGGG
AGAAGTTACGGGAAGGGAGAGATGCCTCCCGCTCATTGAATGAGCATCTCCAGGCCCTCCTCACTCCGGATGAGCCGGACAAGTCCCAGGGGCAGGACCT
CCAAGAACAGCTGGCTGAGGGGTGTAGACTGGCACAGCACCTTGTCCAAAAGCTCAGCCCAGAAAATgatgaagatgaggatgaagatgttcaagttgag
gaggatgagAAAGTGCTGGAATCATCTTCCCCCAGGGAGATGCAGAAGGCTGAAGAAAGCAAAGTCCCTGAGGACTCACTGGAGGAATGTGCCATCACTT
GTTCAAATAGCCACGGCCCTTGTGACTCCAACCAGCCTCACAAGAACATCAAAATCACATTTGAGGAAGACAAAGTCAACTCATCTCTGGTTGTAGACAG
AGAATCCTCTCATGATGAATGTCAGGATGCTCTAAACATTCTCCCAGTCCCTGGCCCCACCTCTTCTGCCACAAACGTCAGCATGGTGGTATCAGCCGGC
CCTTTGTCCAGCGAGAAGGCAGAGATGAACATTCTAGAAATCAATGAGAAGTTGCGCCCCCAGCTGGCAGAGAAGAAACAGCAGTTCAGAAGCCTCAAAG
AGAAATGTTTTGTAACTCAAGTGGCCTGCTTCCTGGCCAAGCAGCAGAACAAATACAAATATGAAGAGTGCAAAGACCTCCTAAAATCTATGCTGAGGAA
TGAGCTACAGTTCAAGGAGGAGAAGCTTGCAGAGCAGCTGAAGCAAGCTGAGGAGCTCAGGCAATATAAAGTCCTGGTTCACTCTCAGGAACGAGAGCTG
ACCCAGTTAAGGGAGAAGTTACGGGAAGGGAGAGATGCCTCCCGCTCATTGAATGAGCATCTCCAGGCCCTCCTCACTCCGGATGAGCCGGACAAGTCCC
AGGGGCAGGACCTCCAAGAACAGCTGGCTGAGGGGTGTAGACTGGCACAACACCTTGTCCAAAAGCTCAGCCCAGAAAATGATAACGATGACGATGAAGA
TGTTCAAGTTGAGGTGGCTGAGAAAGTGCAGAAATCGTCTTCCCCCAGGGAGATGCAGAAGGCTGAAGAAAAGGAAGTCCCTGAGGACTCACTGGAGGAG
TGTGCCATCACTTGTTCAAATAGCCATGGCCCTTATGACTCCAACCAGCCACATAGGAAAACCAAAATCACATTTGAGGAAGACAAAGTCGACTCAACTC
TCATTGGCTCATCCTCTCATGTTGAATGGGAGGATGCTGTACACATTATCCCAGAAAATGAAAGTGATGATGAGGAAGAGGAAGAAAAAGGGCCAGTGTC
TCCCAGGAATCTGCAGGAGTCTGAAGAGGAGGAAGTCCCCCAGGAGTCCTGGGATGAAGGTTATTCGACTCTCTCAATTCCTCCTGAAAGGTTGGCCTCA
TACCAGTCTTACAGCAGCACATTTCACTCATTAGAGGAACAGCAAGTCTGCATGGCTGTTGACATAGGCAGACATCGGTGGGATCAAGTGAAAAAGGAGG
ACCAAGAGGCAACAGGTCCCAGGCTCAGCAGGGAGCTGCTGGCTGAGAAAGAGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTATTCAACTCCTTC
AGTTTATCTTGGACTGACTGACTCATGCCAGCCCTACAGAAGTGCCTTTTACGTATTGGAGCAACAGCGTGTTGGCTTGGCTGTTGACATGGATGAAATT
GAAAAGTACCAAGAAGTGGAAGAAGACCAAGACCCATCATGCCCCAGGCTCAGCAGGGAGCTGCTGGCTGAGAAAGAGCCTGAAGTCTTGCAGGACTCAC
TGGATAGATGTTATTCGACTCCTTCAGGTTATCTTGAACTGCCTGACTTAGGCCAGCCCTACAGAAGTGCTGTTTACTCATTGGAGGAACAGTACCTTGG
CTTGGCTCTTGACGTGGACAGAATTAAAAAGGACCAAGAAGAGGAAGAAGACCAAGGCCCACCATGCCCCAGGCTCAGCAGGGAGCTGCTGGAGGTAGTA

GAGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTATTCAACTCCTTCCAGTTGTCTTGAACAGCCTGACTCCTGCCAGCCCTACAGAAGTTCCTTTT
ATGCATTGGAGGAAAAACATGTTGGCTTTTCTCTTGACGTGGGAGAAATTGAAAAGAAGGGGAAGGGGAAGAAAAGAAGGGGAAGAAGATCAAAGAAGAA
AAGAAGAAGGGGAAGAAAAGAAGGGGAAGAAGATCAAAACCCACCATGCCCCAGGCTCAGCAGGGAGCTGCTGGCTGAGAAAGAGCCTGAAGTCTTGCAG
GACTCACTGGATAGATGGTATTCGACTCCTTCAGTTTATCTTGGACTGACTGACCCATGCCAGCCCTACAGAAGTGCCTTTTACGTATTGGAGCAACAGC
GTGTTGGCTTGGCTGTTGACATGGATGAAATTGAAAAGTACCAAGAAGTGGAAGAAGACCAAGACCCATCATGCCCCAGGCTCAGCAGGGAGCTGCTGGC
TGAGAAAGAGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTATTCGACTCCTTCAGGTTATCTTGAACTGCCTGACTTAGGCCAGCCCTACAGAAGT
GCTGTTTACTCATTGGAGGAACAGTACCTTGGCTTGGCTCTTGACGTGGACAGAATTAAAAAGGACCAAGAAGAGGAAGAAGACCAAGGCCCACCATGCC
CCAGGCTCAGCAGGGAGCTGCTGGAGGTAGTAGAGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTATTCAACTCCTTCCAGTTGTCTTGAACAGCC
TGACTCCTGCCAGCCCTACAGAAGTTCCTTTTATGCATTGGAGGAAAAACATGTTGGCTTTTCTCTTGACGTGGGAGAAATTGAAAAGAAGGGGAAGGGG
AAGAAAAGAAGGGGAAGAAGATCAAAGAAGAAAAGAAGAAGGGGAAGAAAAGAAGGGGAAGAAGATCAAAACCCACCATGCCCCAGGCTCAACAGCGTGC
TGATGGAAGTGGAAGAGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTATTCGACTCCATCAATGTACTTTGAACTACCTGACTCATTCCAGCACTA
CAGAAGTGTGTTTTACTCATTTGAGGAACAGCACATCACCTTTGCCCTTGACATGGACAATAGCTTTTTTACTTTGACGGTGACAAGTCTCCACCTGGTC
TTCCAGATGGGAGTCATATTCCCACAATAAGCAGCCCTTACTAAGCCGAGAGGTTTCATTCCTGCAGGCAGGACCTATAGGCACCTGAAGATTTGAATGA
AACTATAGTTCCATTTGGAAGCCCAGACATAGGATGGGTCAGTGGGCATGGCTCTATTCCTATTCTCAGAGCATGCCAGTGGCAACCTGTGCTCAGTCTG
AAGACAATGGACCCACGTTAGGTGTGACACGTTCACATAACTGTGCAGCACATGCCGGGAGTGATCAGCCGGACATTTTAATTTGAACCATGTATCTCTG
GGTAGCTACAAAATTCCTCAGGGATTTCATTTTGCAGGCATGTCTCTGAGCTTCTATACCTACTCAAGGTCAGTGTCATCTTTGTGTTTAGTTCATCCAA
AGGTGTTACCCTGGTTTCAATGAACCTAACCTCATTATTTGTGTCTTCAGTGTTGGCTTGTTTTAGCTGATCCATCTGTAACACAGGAGGGATCCTTGGC
TGAGGATTGTATTTCAGAACCACCAACTGCTCTTGACAATTGTTAACCCACTAGGCTCCTTTGGTTAGAGAAGCCACAGTCCTTCAGCCTCCAATTGGTG
TCAGTACTTAGGAAGACCACAGCTAGATGGACAAACAGCATTGGGAGGCCTTAGCCCTGCTCCTCTCAATTCCATCCTGTAGAGAACAGGAGTCAGGAGC
CGCTGGCAGGAGACAGCATGTCACCCAGGACTCTGCCGGTGCAGAATATGAGCAATGCCAtgttcttgcagaaaacgcttagcctgagtttcataggagg
taatcaccagacaactgcagaatgtagaacactgagcaggacaactgacctgtctccttcacatagtccatatcaccacaaatcacacaacaaaaaggag
aagagatattttgggttgaaaaaaagtaaaaagataATGTAGCTGCAtttctttagttattttgaaccccaaatatttcctcatctttttgttgttgtca
ttgatggtggtgacatggacttgtttatagaggacaggtcagctctctggctcaatgatctacattctgaagttgtctgaaaatgtcttcatgattaaat
tcagcctaaacttttgctgggaacactgcagagacaatgctgtgagtttccaacctcagcccatctgcgggcagagaaggtctagtttgtccatcacca
ttatgatatcaggactggttacttggttaaggaggggtctaggagatctgtccctttagagacaccttacttataatgaagtacttgggaaagcggttt
tcaagagtataaatatcctgtattctaatgatcatcctctaaacattttatcatttattaatcctccctgcctgtgtctattattaTATACATATCTCTA
CGCTGCAAATTTTGGGTCTCAATTTTTACTGTGCCTTTGTTTTTACTAGTGTCTGCTGTTGCAAAAAGAAGAAAACATTCTCTGCCTGAGTTTTAATTTT
TGTCCAAAGTTAATTTTAATCTATACAATTAAAACCTTTTGCCTATCACTCTGGACTTTTGGATTGTTTTTTACATTCAGTGTTATAATATTTGATTATG
CTGATTGGTTTTGGTGGGTACTGATGTGAATTAATAAAAACATTTCATTTCCATGTT

APPENDIX B: ADDITIONAL MATERIAL FOR CHAPTER 3

Supplementary Figures

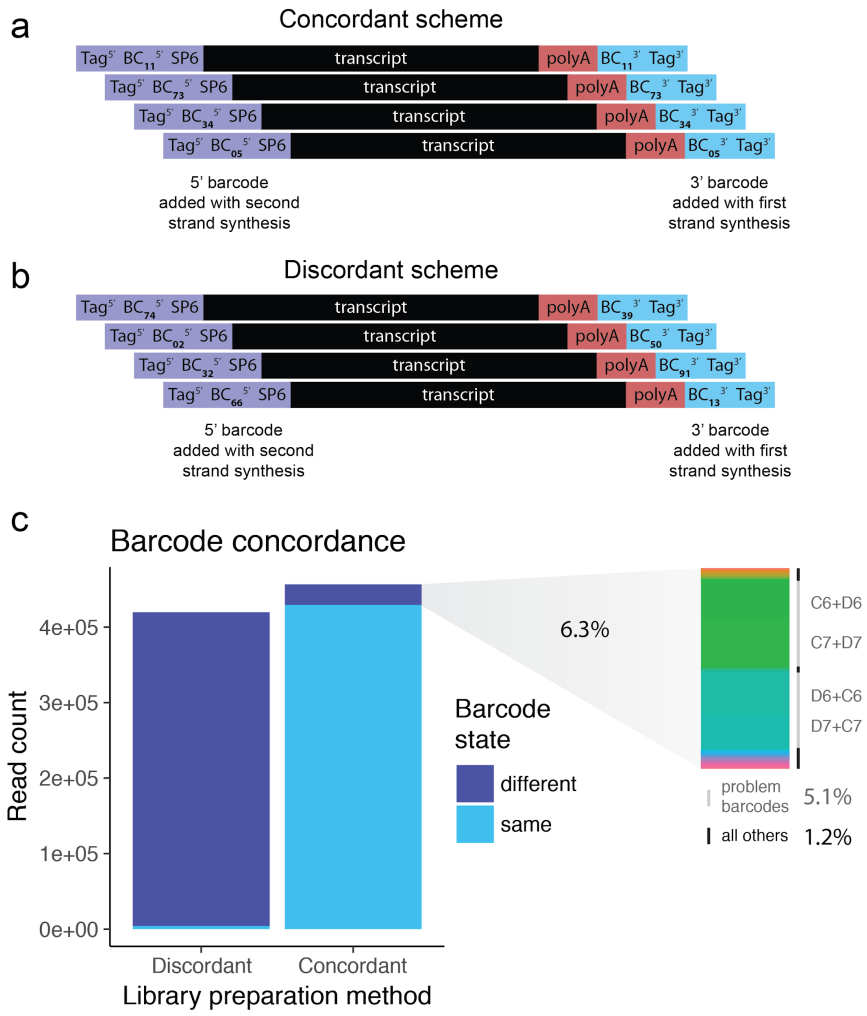


Figure S1. Barcoding strategy. A. In the concordant barcode scheme, 1 of 96 identical barcodes are appended to the 3' and 5' ends of transcript sequences during first strand and second strand synthesis, respectively. This allows for detection of false chimeric sequences generated during amplification. **B.** In the discordant barcode scheme, 1 of 96 barcodes are appended at random to 3' and 5' ends of transcript sequences. In this method, the combination (96x96) can be used as a pseudo-unique molecular identifier, to measure the frequency of PCR duplicates. **C.** The concordance of reads from libraries prepared under both schemes is shown, measured as number of putative full-length reads with barcodes that are the same or different. Among discordant-scheme libraries, we obtain 1.0% reads with matching barcodes, consistent with a 1 in 96 chance. Among concordant-scheme libraries, we obtain 93.7% reads with matching barcodes, indicating that 6.3% are not as expected. When we examine which barcode pairs comprise these unexpectedly discordant reads, we observe that the majority of these (80%) are derived from two particular pairings of barcodes (in both orientations), here labelled by their position in a 96-well plate. Contamination specifically between the barcodes of wells C6 and D6, and wells C7 and D7, are mostly responsible. Because pattern holds true for all concordant-scheme libraries over multiple separate

library preparations, we believe this is most likely due to a manufacturing error that led to contamination. Eliminating those “problem pairings”, we observe a rate of 1.2% discordancy, indicating that chimeric formation during PCR is rare.

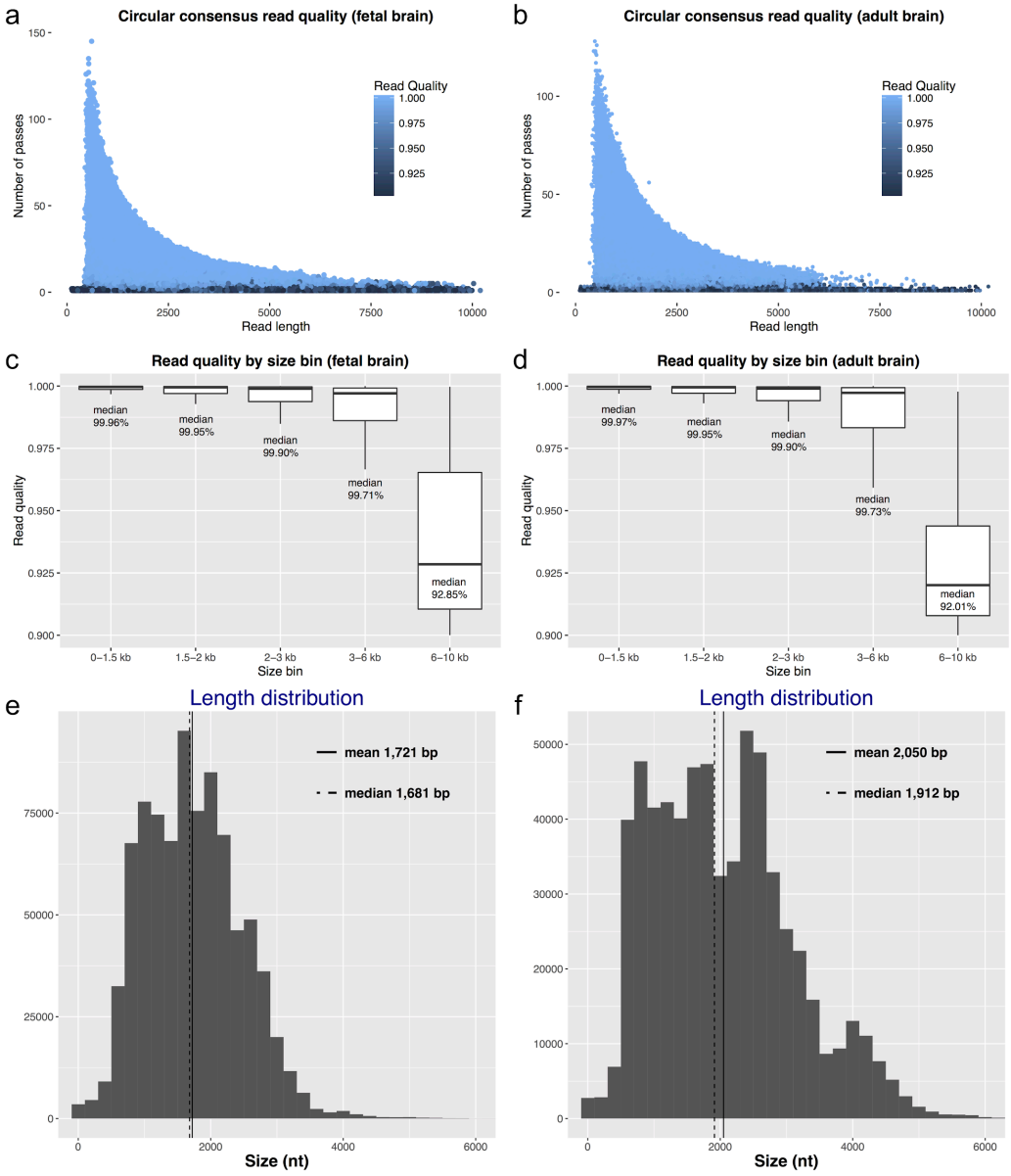


Figure S2. CCS read quality is highly dependent on read length. In both fetal brain (**A**) and adult brain (**B**), there is an inverse relationship between read length and the number of passes by the sequencing polymerase. This determines the number of subreads from which a given read is generated, which directly affects read quality. Each point represents a CCS read. Reads with a high number of passes (i.e., subreads) have very high average read quality. When reads are binned by size (**C**, **D**) we see that reads smaller than 3 kbp have on average >99.9% read quality, but that for reads over 6 kpb, that falls to ~92%. The ICE algorithm clusters these low quality reads and uses the resulting multiple sequence alignment to arrive at cluster-based isoforms with >99% read quality. Data is shown for HSD1-enriched cDNA, for which we generated the most sequencing reads. Quality values are taken from the "rq" field of bost-CCS bam files. **E** and **F**. Size distributions of putative full-length reads designated for (**A**) adult brain and (**B**) fetal brain.

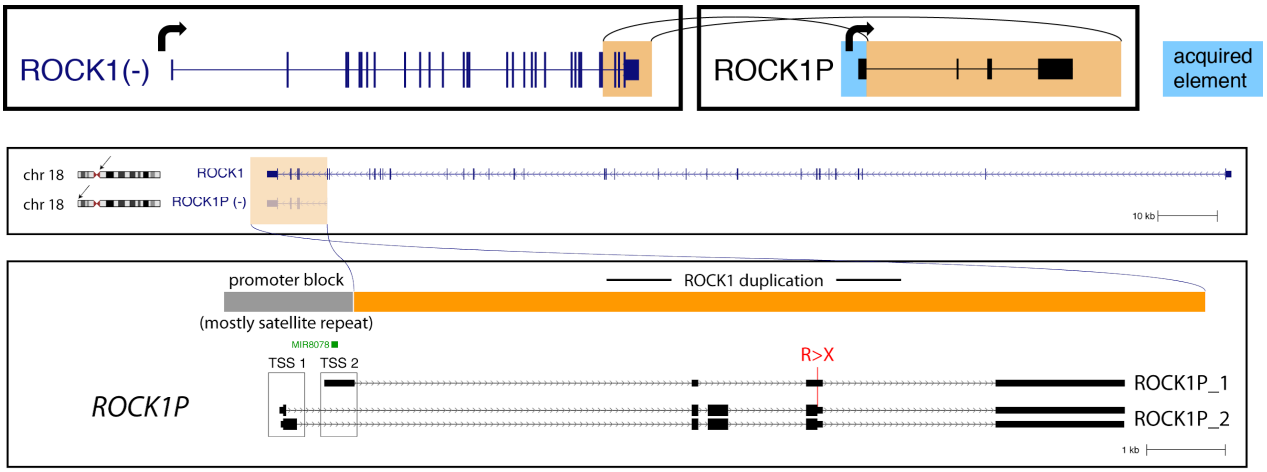


Figure S3. Promoter exaptation in the 5'-truncated duplicate gene *ROCK1P*. The final 4 exons and preceding splice donor of the large gene *ROCK1* underwent an inverted intrachromosomal duplication on chromosome 18, generating *ROCK1P*. "Rescue" of 5'-truncated gene is facilitated by promoter borrowing from the adjacent duplication block, that creates two nearby transcription start sites (TSS 1 and TSS 2) that overlap a microRNA identified in testis. *ROCK1P* has highest expression in the testis, in contrast to the ancestral widely expressed *ROCK1* gene, likely indicating a role for this acquired promoter in *ROCK1P*'s expression pattern. A stop gain on the penultimate exon of *ROCK1P* makes the open-reading frame even shorter than by duplication alone, possibly indicating neutrality of the coding sequence.

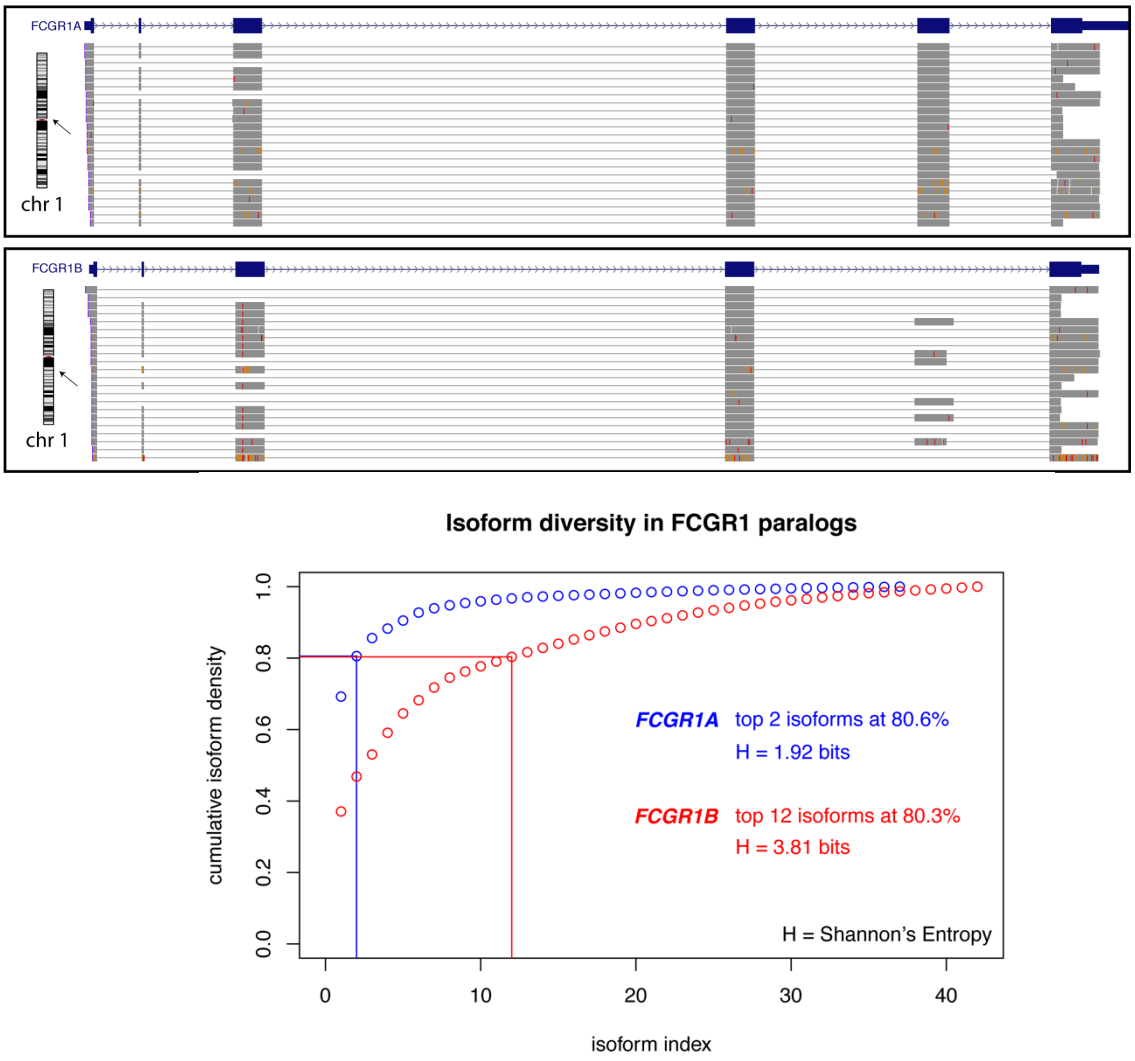


Figure S4. Relaxed selection on splicing results in increased disorder of isoforms. A. A random subset of ~25 reads each for *FCGR1A* and *FCGR1B* is shown. A visually striking difference in the orderliness of isoforms is apparent, with *FCGR1B* appearing substantially more disorderly. When compared to the ancestral paralog *FCGR1A*, *FCGR1B* displays a greater diversity of splice isoforms, particularly due to variable choice of splice donor and acceptor sites at exon 5, as well as exclusion of exon 5. In *FCGR1A*, this lack of incorporation of this exon is not observed (<0.4%). **B.** Cumulative distributions of isoform abundances in *FCGR1A* (blue) and *FCGR1B* (red) are plotted. The 2 most abundant *FCGR1A* isoforms comprise 80.6% of the total *FCGR1A* reads, while for *FCGR1B*, it is not until the 12th most abundant isoform that 80.3% of total reads is reached, indicating a significant "flattening out" of relative isoform abundances (p = 1.3e-7), also manifesting as an increase in the entropy of this distribution.

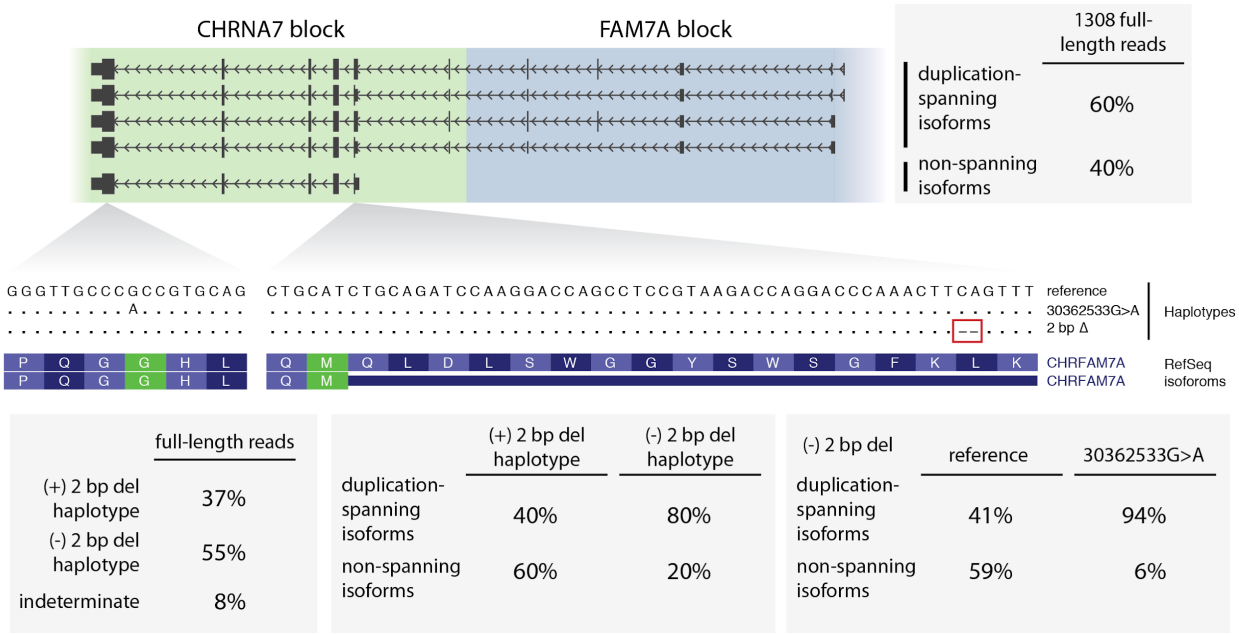


Figure S5. Identification of 5'-truncated *CHRFAM7A* isoform. Single-molecule long-read sequencing of probe-enriched cDNA generated from pooled fetal and adult brain yielded 1308 putative full-length *CHRFAM7A* reads (MAPQ > 40). 60% of these reads spanned the *CHRNA7* duplication boundary and included upstream exons from the *FAM7A* block, while 40% of reads were initiated from a transcription start site internal to the *CHRNA7* block, in the vicinity of the splice acceptor of the exon paralogous to exon 6 of canonical *CHRNA7* (referred to as *CHRFAM7A* exon 6). While we do not have access to the genotypes of the individuals from whom these reads were generated, three alleles can be discerned from variants within the transcribed sequence: the reference allele, the 30363533G>A allele, and the 2 bp Δ allele. The 2 bp Δ allele causes a frameshift in the open-reading frame (ORF) found in some of the duplication-spanning isoforms, but is outside the ORF that begins in *CHRFAM7A* exon 6, which is contained in the other duplication spanning isoforms as well as the non-spanning isoform. 2 bp Δ transcripts have a smaller proportion of duplication spanning isoforms when compared to transcripts without the deletion. Reads with the 30363533G>A allele are almost exclusively duplication-spanning, while the other two alleles share a similar ratio. Naively, the non-spanning isoform should have a greater probability of having the relevant, translated ORF, given it is unaffected by the 2 bp Δ and has a 5' UTR that is shorter and composed of fewer exons.

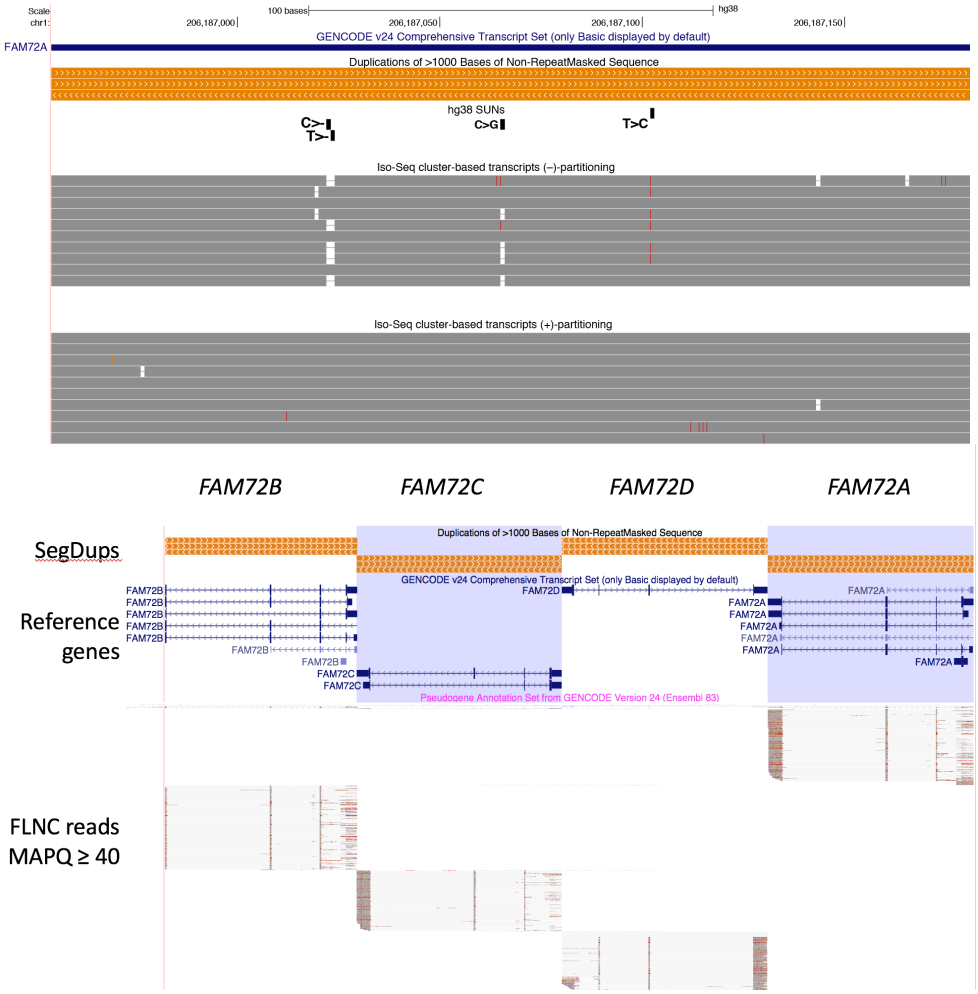


Figure S6. Pre-partitioning of reads mitigates generation of *in silico* isoform chimeras. Sown is a UCSC Genome Browser screen shot of the 3'-UTR of *FAM72A*, a gene with 4 reference paralogs. A cluster of PSVs which distinguish this paralog from all others are shown below the segmental duplication track (orange bars, indicating >99% sequence identity). Below the PSVs are two tracks showing a random selection of 10 isoforms from the output of ICE mapped back to the genome using GMAP with a mapping quality cutoff of 40. The above track uses the default application of the ICE pipeline, and *in silico* chimeric reads are apparent, as a consequence of the clustering and alignment of reads from not only *FAM72A* but also other *FAM72* paralogs, resulting in apparent SNVs corresponding with known PSVs. The bottom track uses ICE after an additional "partitioning" step, wherein putative full-length reads are first mapped with a mapping quality cutoff of 40, then partitioned into separate regions to ensure that reads belonging to different paralogs are not clustered together during ICE. Applying this strategy, the spurious variants disappear, indicating the output of ICE is no longer confounded by mixing paralogs.

Supplementary Tables

Table S1. Probe design, including targeted gene families and number of probes.

Design	Category	Reigion	Gene Family	Duplication type‡	Copies‡	No human CN reversions****	Length (anc)	No. probes	Notes
HSD Panel I	HSD	chr1	*NOTCH2*	partial	5	?	11474	8	
515 total probes	HSD	1q12	*GPR89*	complete	2	x	2150	12	
	HSD	chr1	*FAM72*	complete	4	x	2400	4	
	HSD	chr1	*SRGAP2*	partial	4	x	4917	29	includes 21 designed to SRGAP2A and 8 designed to SRGAP2C
	HSD	2p11.2,2q13	*CD8B*	complete	2	x	934	7	complete for canonical gene model (other isoforms exist that are truncated)
	HSD	2q21.1	*TISP43*	complete	2		2036	12	
	HSD	2q21.1	*CFC1*	complete	2	x	1447	11	
	HSD	7q11.23	*GTF2I*	partial	3	x	4445	18	
	HSD	7q11.23	*NCF1*	complete	3	x	1459	10	
	HSD	7q11.23	*GTF2IRD2*	partial	3	x	625	10	
	HSD	10q11.22	*GPRIN2*	complete	2	x	1823	12	
	HSD	10q11.22	*PTPN20*	partial	2	x	1885	11	
	HSD	10q11.22	*NPY4R*	complete	2	x	1800	10	
	HSD	10q11.22	*FRMPD2*	partial	2	x	1892	10	
	HSD	15q13.3	*ARHGAP11*	partial	2	x	5898	33	includes 2 designed only to ARHGAP11B
	HSD	16p11.2	*BOLA2*	complete	3	x	1012	6	
	HSD	16p11.2	*CORO1A*	partial	3		1815	2	
	HSD	16p11.2	*SLX1A*	complete	3		828	5	
	HSD	16q22.2,1q21.1	*HYDIN*	partial	2	x	3122	80	includes 3 designed only to HYDIN2
	HSD	18q11.1,18p11.2	*ROCK1*	partial	2		6648	5	
	ASD	1q21.3	*POGZ*	n/a	1		6153	22	
	ASD	14q11.2	*CHD8*	n/a	1		7475	49	
	ASD	20q13.13	*ADNP*	n/a	1		5966	19	
	ASD	21q22.13	*DYRK1A*	n/a	1		5088	17	
	splicing	2p16.3	*NRXN1*	n/a	1		5575	43	
	splicing	11q13.1	*NRXN2*	n/a	1		3535	32	
	splicing	14q24.3-31.1	*NRXN3*	n/a	1		8578	38	
HSD Panel II	HSD	1q21.1,14q23.1	*GNRHR2*	partial	2		1035	4	
525 total probes	HSD	1q21.1,14q23.1	*RBM8*	partial	2		4974	5	
	HSD	chr1	*FCGR1A*	complete	4	x	2268	7	
	HSD	chr1	*HIST2H2BF*	complete	4	x	999	8	includes 4 designed to HIST2H3D
	HSD	2p11.2,2q13	*LINC00152*	partial	2		518	5	
	HSD	2p11.2,2q13	*ANAPC1*	partial	2		7753	48	complex duplication
	HSD	5q13.2	*OCLN*	partial	2		6451	13	
	HSD	5q13.2	*GTF2H2*	complete	2		1951	8	
	HSD	5q13.2	*SERF1*	complete	2		1935	2	
	HSD	5q13.2	*SMN1*	complete	2		1641	6	
	HSD	5q13.2	*NAIP*	partial	3		5880	30	
	HSD	16p11.2/6p	*DUSP22*	complete	2		3518	6	
	HSD	6p22.2,9q22.3	*ZNF322*	partial	2		4892	11	
	HSD	7q35	*TCAF2*	partial	2		2997	23	
	HSD	7q35	*OR2A1***	complete	2		933	11	
	HSD	7q35	*ARHGEF5*	partial	3		5505	40	
	HSD	10q11.22	*LOC643650*	complete	2		2872	8	
	HSD	15q13.3	*CHRNA7*	partial	4		3443	10	
	HSD	7q35	*TCAF1*	partial	2		5682	18	
	ASD	4q25-26	*ANK2*	n/a	1		8082	47	
	ASD	6q25.3	*ARID1B*	n/a	1		9609	30	
	ASD	12p13.1	*GRIN2B*	n/a	1		5941	38	
	ASD	17p13.1	*KDM6B*	n/a	1		6704	20	
	ASD	21q22.2	*DSCAM*	n/a	1		7055	50	
	SV gene	1p21.1	*amylase*	n/a	6‡‡		1862	60	includes AMY1A,AMY1B,AMY1C,AMY2A,AMY2B
	SV gene	17q21.31	*KANSL1*	n/a	1		5059	17	

‡For HSDs, rounded haploid copy number in HGDP. Derived from Dennis et al. 2017 Table S1. Exceptions are NOTCH2,GNRHR2,RBM8,LINC00152,ANAPC1,ZNF322,LOC643650 (taken from GRCh38), SLX1A (inferred, adjacent to BOLA2).

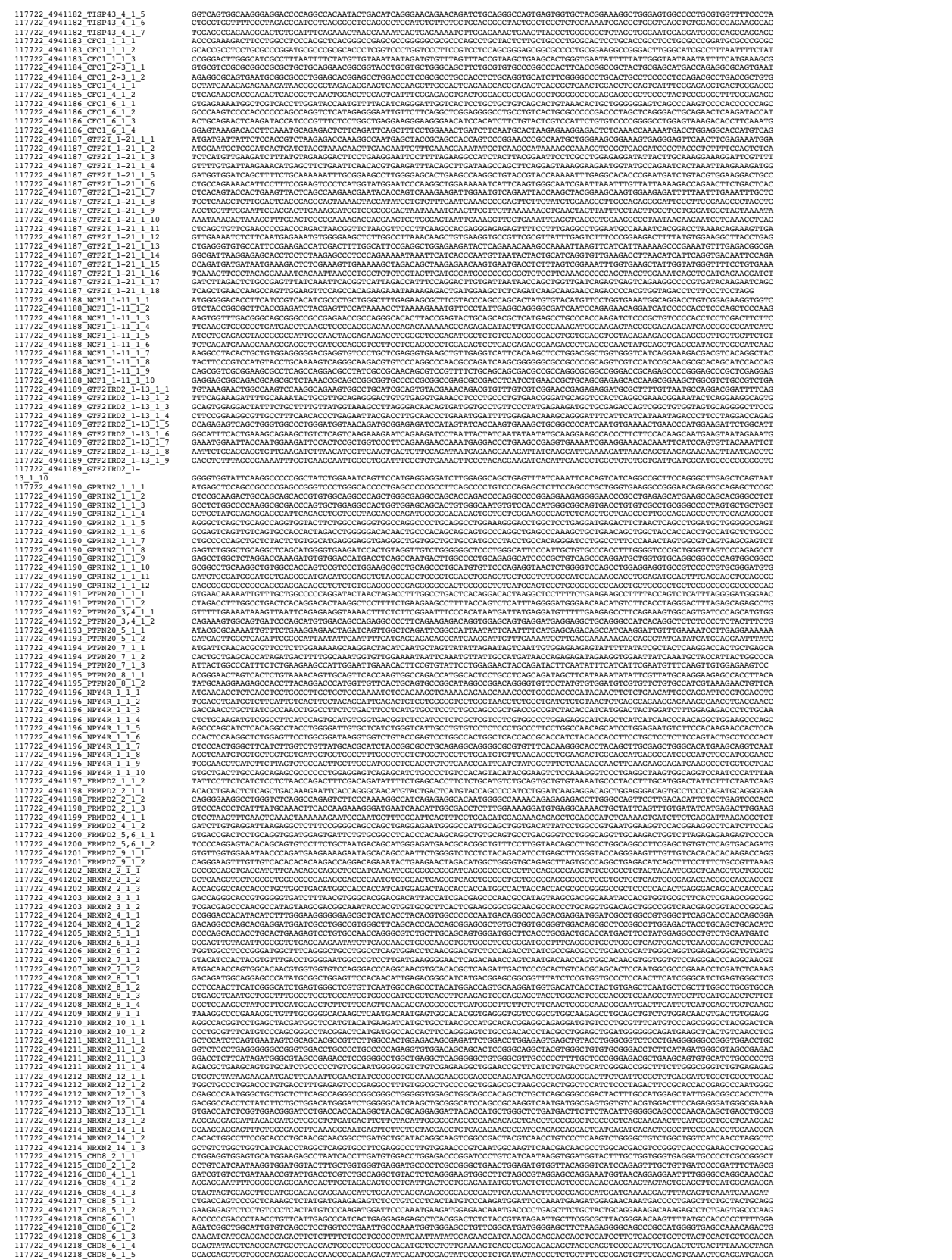
***aka "OR3A"

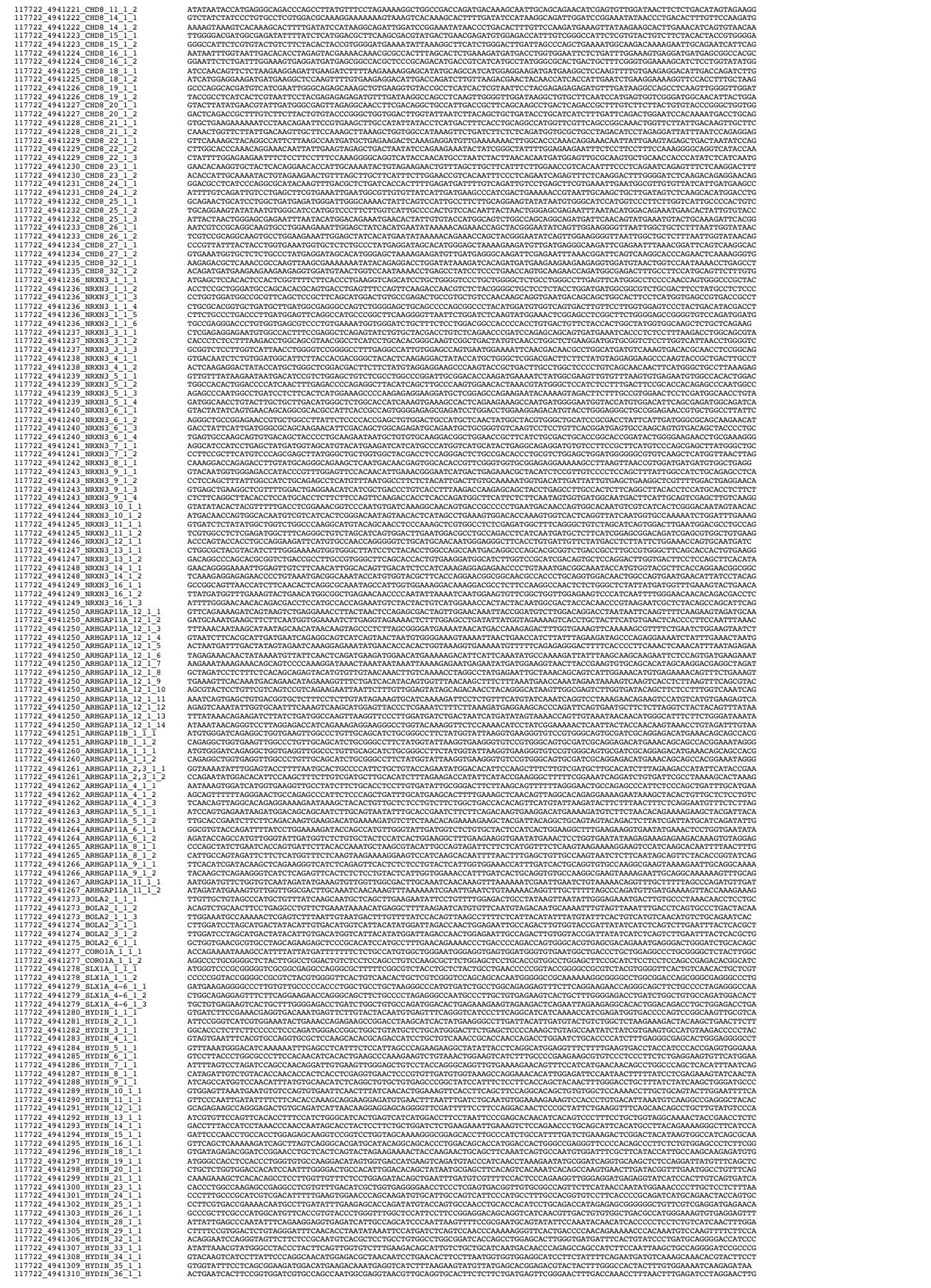
‡‡Highly polymorphic, 6 in reference counting 1 pseudogene

****Derived from Dennis et al. 2017 Table S8

Table S2a. HSD1 probe design sequences

SequenceName	Sequence
117722_4941114_NOTCH2_1_1_1	GTAAGGAGTGCCAATGGACGGATGCCTGCCTGTCTCATCCCTGTGCAAATGGAAGTACCTGTACCACTGTGGCCAACCAGTTCTCCTGCAAATGCCTCACAGGCTTCACAGGGCAGAAAT
117722_4941114_NOTCH2_1_1_2	CAGGGCAGAAATGTGAGACTGATGTCAATGAGTGTGACATTCCAGGACACTGCCAGCATGGTGGCACCTGCCTCAACCTGCCTGGTTCCTACCAGTGCCAGTGCCCTCAGGGCTTCACAG
117722_4941114_NOTCH2_1_1_3	AGGGCTTCACAGGCCAGTACTGTGACAGCCTGTATGTGCCCTGTGCACCCTCACCTTGTGTCAATGGAGGCACCTGTCGGCAGACTGGTGACTTCACTTTTGAGTGCAACTGCCTTCCAG
117722_4941115_NOTCH2_2_1_1	ATGTCCAGAAGGCTTCTTGGGGGAATATTGTCAACATCGAGACCCCTGTGAGAAGAACCGCTGCCAGAATGGTGGGACTTGTGTGGCCCAGGCCATGCTGGGGAAAGCCACGTGCCGATG
117722_4941115_NOTCH2_2_1_2	GGTGGGACTTGTGTGGCCCAGGCCATGCTGGGGAAAGCCACGTGCCGATGTGCCTCAGGGTTTACAGGAGAGGACTGCCAGTACTCAACATCTCATCCATGCTTTGTGTCTCGACCCTGC
117722_4941115_NOTCH2_2_1_3	AGGACTGCCAGTACTCAACATCTCATCCATGCTTTGTGTCTCGACCCTGCCTGAATGGCGGCACATGCCATATGCTCAGCCGGGATACCTATGAGTGCACCTGTCAAGTCGGGTTTACAG
117722_4941116_NOTCH2_3,4_1_1	ACCGAGAAGATGCCCGCCCTGCGCCCCGCTCTGCTGTGGGCGCTGCTGGCGCTCTGGCTGTGCTGCGCGGCCCCCGCGCATGCATTGCAGTGTCGAGATGGCTATGAACCCTGTGTAAAT
117722_4941116_NOTCH2_3,4_1_2	GCTGGCGCTCTGGCTGTGCTGCGCGGCCCCCGCGCATGCATTGCAGTGTCGAGATGGCTATGAACCCTGTGTAAATGAAGGAATGTGTGTTACCTACCACAATGGCACAGGATACTGCAA
117722_4941117_SRGAP2C_1_1_1	GCAAGTGAACCGGAGCAAACGACTTCCGATCCAGTCTGCGCTGTTGCGGCTCCCGTTTGGGATTTGATTTGCAGCATCTTTGAGCCTCTACGACAAAAAACCGCGAAGCACGCCCAGCCC
117722_4941117_SRGAP2C_1_1_2	AACCGCGAAGCACGCCCAGCCCTCCCCCGGCACCCCGAAAAGCACCCACTCCCTCCCGGGGACACAGCTGGGCGCGTCCACACCCCCGCAGCCCCACACCATGTTGTGCGGAAGGACTTC
117722_4941117_SRGAP2C_1_1_3	CCATGTTGTGCGGAAGGACTTCCACTCCCCGCCTGTGTCGTTGATGTCAGACCCCAGGCCAGCCTCCGGGCGCTGCAGTTCTCCCGGCTAATGCTGAGGCTGCGGCTCCGGCTCTAGCAC
117722_4941117_SRGAP2C_1_1_4	GCTGCGGCTCCGGCTCTAGCACAGGCACCAGCCGCCGCCGCACCCGGCCTCAGCGCCCACCGTCTGCATGTGCCCGCCGTAGCCGTCTGCCCAGCCCGCAGCCCGCGCTCCACGGAGCGC
117722_4941117_SRGAP2C_1_1_5	CAGCCCGCGCTCCACGGAGCGCTGGAGACCACCGTGGGGGGCCCCTTCTGCCCTCGAGAGAAGCGGTCTTGGAGGTATTGATTTAGGTGGTTGGATTTTTTCCGTGGATCTATCAATTCA
117722_4941117_SRGAP2C_1_1_6	TTTCCGTGGATCTATCAATTCACAATTCGAATTTGGAAGAAAGAAGGAAAACATGACGTCTCCAGCCAAATTCAAAAAGGATAAGGAGATCATAGCAGAGTACGATACTCAGGTCAAAG
117722_4941123_SRGAP2C_8_1_1	GCTTCCCAGCTCTGTGCCCAGCAGCCTGTCCAGAGTGAGCTGGTACAGAGATGCCAACAACTGCAGTCTCGCTTATCCACTCTAAAGATTGAAAACGAAGAGGTGAGTTTCTTCCTTAGG
117722_4941123_SRGAP2C_8_1_2	CTTCCCAGCTCTGTGCCCAGCAGCCTGTCCAGAGTGAGCTGGTACAGAGATGCCAACAACTGCAGTCTCGCTTATCCACTCTAAAGATTGAAAACGAAGAGGTGAGTTTCTTCCTTAGGA
117722_4941130_GPR89_1-14_1_1	ATGAGTTTCCTCATCGACTCCAGCATCATGATTACCTCCCAGATACTATTTTTTGGATTTGGGTGGCTTTTCTTCATGCGCCAATTGTTTAAAGACTATGAGATACGTCAGTATGTTGTA
117722_4941130_GPR89_1-14_1_2	TGTTGTACAGGTGATCTTCTCCGTGACGTTTGCATTTTCTTGCACCATGTTTGAGCTCATCATCTTTGAAATCTTAGGAGTATTGAATAGCAGCTCCCGTTATTTTCACTGGAAAATGAA
117722_4941130_GPR89_1-14_1_3	AAATGAACCTGTGTGTAATTCTGCTGATCCTGGTTTTCATGGTGCCTTTTTACATTGGCTATTTTATTGTGAGCAATATCCGACTACTGCATAAACAACGACTGCTTTTTTCCTGTCTCT
117722_4941130_GPR89_1-14_1_4	TGTCTCTTATGGCTGACCTTTATGTATTTCTTCTGGAAACTAGGAGATCCCTTTCCCATTCTCAGCCCAAAACATGGGATCTTATCCATAGAACAGCTCATCAGCCGGGTTGGTGTGATT
117722_4941130_GPR89_1-14_1_5	TGTGATTGGAGTGACTCTCATGGCTCTTCTTTCTGGATTTGGTGCTGTCAACTGCCCATACACTTACATGTCTTACTTCCTCAGGAATGTGACTGACACGGATATTCTAGCCCTGGAACG
117722_4941130_GPR89_1-14_1_6	TGGAACGGCGACTGCTGCAAACCATGGATATGATCATAAGCAAAAAGAAAAGGATGGCAATGGCACGGAGAACAATGTTCCAGAAGGGGGAAGTGCATAACAAACCATCAGGTTTCTGGG
117722_4941130_GPR89_1-14_1_7	TTCTGGGGAATGATAAAAAGTGTTACCACTTCAGCATCAGGAAGTGAAAATCTTACTCTTATTCAACAGGAAGTGGATGCTTTGGAAGAATTAAGCAGGCAGCTTTTTCTGGAAACAGCT
117722_4941130_GPR89_1-14_1_8	AACAGCTGATCTATATGCTACCAAGGAGAGAATAGAATACTCCAAAACCTTCAAGGGGAAATATTTTAATTTTCTTGGTTACTTTTTCTCTATTTACTGTGTTTGGAAAATTTTCATGGC
117722_4941130_GPR89_1-14_1_9	TCATGGCTACCATCAATATTGTTTTTGATCGAGTTGGGAAAACGGATCCTGTCACAAGAGGCATTGAGATCACTGTGAATTATCTGGGAATCCAATTTGATGTGAAGTTTTGGTCCCAAC
117722_4941130_GPR89_1-14_1_10	TCCCAACACATTTCCTTCATTCTTGTTGGAATAATCATCGTCACATCCATCAGAGGATTGCTGATCACTCTTACCAAGTTCTTTTATGCCATCTCTAGCAGTAAGTCCTCCAATGTCATT
117722_4941130_GPR89_1-14_1_11	TGTCATTGTCCTGCTATTAGCACAGATAATGGGCATGTACTTTGTCTCCTCTGTGCTGCTGATCCGAATGAGTATGCCTTTAGAATACCGCACCATAATCACTGAAGTCCTTGGAGAACT
117722_4941130_GPR89_1-14_1_12	GAGAACTGCAGTTCAACTTCTATCACCGTTGGTTTGATGTGATCTTCCTGGTCAGCGCTCTCTCTAGCATACTCTTCCTCTATTTGGCTCACAAACAGGCACCAGAGAAGCAAATGGCAC
117722_4941131_HYDIN2_1_1_2	TCCTGAAATGGGCATATCCCTGGCTGGCCTGGAACCAACCTTTCGGACTTAGGACTATGGGAAGAATAGGGGGAGATTCTCTAGTGATGATAATACATGAATGCTTTAAAAGCTAGAACA
117722_4941132_HYDIN2_2_1_1	GGTACTAGCGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAAGGCCTTGGAAAATCGGAGCGA
117722_4941132_HYDIN2_2_1_2	CGATTCCACTGGGCGCTGTAGGCAGGGCGCACTCACCGAACTCCCCCATCCCTTCCAGCCTCAGGGCCCGCCTGGCAGGACTACTGAGCAAGGCCTTGGAAAATCGGAGCGATTAGAGCG
117722_4941139_POGZ_4_1_1	GAGTGTCATGGGCCGGCAGACATGCCTGGAGTGCAGCTTCGAGATCCCAGACTTCCCTAATCATTTCCCTACTTACGTACACTGCTCTCTGTGTCGCTATAGCACCTGCTGTTCTCGAGC
117722_4941139_POGZ_4_1_2	ATGCCTGGAGTGCAGCTTCGAGATCCCAGACTTCCCTAATCATTTCCCTACTTACGTACACTGCTCTCTGTGTCGCTATAGCACCTGCTGTTCTCGAGCTTATGCCAACCACATGATCAA
117722_4941140_POGZ_5_1_1	GTGACAATCCGGGCTTCCCGAGGGCAGCCACGAACTGTTCCTGTATCCTCTAATGATACACCTCCCAGCGCCTTGCAGGAGGCAGCACCGCTGACCTCCTCAATGGACCCTCTGCCTGTC
117722_4941140_POGZ_5_1_2	TGATACACCTCCCAGCGCCTTGCAGGAGGCAGCACCGCTGACCTCCTCAATGGACCCTCTGCCTGTCTTCCTTTATCCCCCTGTCCAGCGCAGCATCCAGAAGAGAGCTGTTAGGAAAAT
117722_4941141_POGZ_6_1_1	AAGAGAAATGTTTATCACTGCAACAAATGCCGGCTGCAGTTTCTCTTTGCCAAGGACAAAATTGAACACAAGCTTCAACACCATAAAACCTTCCGTAAACCCAAGCAGCTGGAGGGCTTG
117722_4941141_POGZ_6_1_2	CACTGCAACAAATGCCGGCTGCAGTTTCTCTTTGCCAAGGACAAAATTGAACACAAGCTTCAACACCATAAAACCTTCCGTAAACCCAAGCAGCTGGAGGGCTTGAAACCAGGCACCAAG
117722_4941142_POGZ_7_1_1	GTGTGTCAATATCGCTCCTCACTCTACTCTGAGGTAGATGTCCATTTTCGGATGATCCATGAGGATACCCGGCATCTGCTCTGCCCTTATTGCCTGAAGGTCTTCAAAAATGGCAATGCA
117722_4941142_POGZ_7_1_2	TCTGAGGTAGATGTCCATTTTCGGATGATCCATGAGGATACCCGGCATCTGCTCTGCCCTTATTGCCTGAAGGTCTTCAAAAATGGCAATGCATTCCAACAGCATTACATGAGGCACCAG
117722_4941143_POGZ_9_1_1	ATTCATGAACCATATGAAACACCACGTAGAACTCGATCAGCAGAACGGTGAGGTAGATGGTCACACTATCTGCCAGCACTGTTACCGCCAGTTTTCCACTCCCTTCCAGCTTCAGTGCCA
117722_4941143_POGZ_9_1_2	ATCAGCAGAACGGTGAGGTAGATGGTCACACTATCTGCCAGCACTGTTACCGCCAGTTTTCCACTCCCTTCCAGCTTCAGTGCCACTTGGAAAATGTTCATAGTCCCTATGAATCTACTA
117722_4941144_POGZ_10_1_1	TACTGTTGCCCAGAAATGGTTGAATACCAGAAGAAAGGAAAGTCCCTGGATTCAGAACCCAGTGTCCCATCAGCAGCAAAGCCCCCATCCCCTGAGAAAACAGCTCCTGTTGCTTCCACA
117722_4941144_POGZ_10_1_2	TTGCTTCCACACCCTCTTCTACACCTATTCCTGCTCTGTCACCGCCTACCAAAGTACCAGAACCAAATGAGAACGTGGGCGATGCCGTCCAGACCAAACTCATTATGCTTGTAGATGACT
117722_4941144_POGZ_10_1_3	TGTAGATGACTTCTACTATGGACGGGATGGTGGCAAAGTAGCCCAGCTCACAAATTTCCCTAAGGTCGCCACATCTTTCCGATGCCCACATTGTACCAAAAGGCTAAAAAACAATATTCG
117722_4941145_POGZ_12_1_1	TGAGCATTGCCAGCTTTGTCACTGTGAAGCGACCTGGTGTTACAGGCGAAAATAGCAATGAAGTGGCCAAATTGGTGAATACCCTTAACACCATCCCTTCCCTGGGCCAGAGTCCTGGGC
117722_4941145_POGZ_12_1_2	TGGTGAATACCCTTAACACCATCCCTTCCCTGGGCCAGAGTCCTGGGCCAGTGGTGGTGTCCAACAACAGCTCTGCTCATGGCTCTCAAAGAACCAGCGGACCTGAGTCTTCAATGAAAG
117722_4941146_POGZ_13_1_1	CCCCAGGTACCCAGTTTGTTAAGCCGACAGTTGGAGTTCCACAAGTGTTCTCCCAGATGACCCCTGTGAGGCCAGGCTCCACAATGCCTGTGAGGCCCACCACCAACACCTTCACCACCG
117722_4941146_POGZ_13_1_2	CCTGTGAGGCCCACCACCAACACCTTCACCACCGTCATCCCGGCCACTCTTACCATTCGAAGCACCGTCCCACAGTCCCAGTCCCAGCAGACCAAGTCCACTCCCAGCACTTCTACCACT
117722_4941146_POGZ_13_1_3	GCAGACCAAGTCCACTCCCAGCACTTCTACCACTCCCACTGCCACACAGCCAACCTCACTGGGGCAACTAGCTGTTCAGTCTCCAGGCCAGTCAAACCAGACCACGAATCCCAAGCTAG
117722_4941147_POGZ_2_1_1	CTGGCAATCCTTTGGTCCAGCAAGGTGGACAGCCACTCATCCTGACCCAGAATCCAGCCCCAGGTCTGGGCACAATGGTTACTCAACCAGTATTGAGGCCTGTTCAGGTCATGCAGAATG
117722_4941147_POGZ_2_1_2	GCCCCAGGTCTGGGCACAATGGTTACTCAACCAGTATTGAGGCCTGTTCAGGTCATGCAGAATGCCAATCATGTGACTAGTTCCCCTGTGGCCTCACAACCAATATTTATCACTACGCAG
117722_4941148_POGZ_3_1_1	TTTCTGTGAGCCAGCAGCCAGTCTCGGCTCCAGTGCCCATCGCTGCCCATGCTTCTGTTGCTGGGCACCTCTCTACATCCACCACCGTTAGTAGCAGCGGGGCACAGAACAGCGACAGTA
117722_4941148_POGZ_3_1_2	TCGCTGCCCATGCTTCTGTTGCTGGGCACCTCTCTACATCCACCACCGTTAGTAGCAGCGGGGCACAGAACAGCGACAGTACAAAGAAGACTCTTGTCACACTAATTGCCAACAACAATG
117722_4941149_FAM72A_1-4_1_1	ATGTCTACCAACATTTGTAGTTTCAAGGACAGGTGCGTGTCCATCCTGTGTTGCAAATTCTGTAAACAAGTGCTCAGCTCTAGGGGAATGAAGGCTGTTTTGCTGGCTGATACTGAAATA
117722_4941149_FAM72A_1-4_1_2	TACTGAAATAGACCTTTTCTCTACAGACATCCCTCCTACCAACGCAGTGGACTTCACTGGAAGATGCTATTTCACCAAAATCTGCAAATGTAAACTGAAGGACATCGCATGTTTAAAATG
117722_4941149_FAM72A_1-4_1_3	GTTTAAAATGTGGGAACATTGTAGGTTATCATGTGATTGTTCCATGTAGTTCCTGTCTTCTTTCCTGCAACAACGGACACTTCTGGATGTTTCACAGCCAGGCAGTTTATGATATTAACA
117722_4941149_FAM72A_1-4_1_4	GATATTAACAGACTAGACTCCACAGGTGTAAACGTCCTACTTTGGGGCAACTTGCCAGAGATAGAAGAGAGTACAGATGAAGATGTGTTAAATATCTCAGCAGAGGAGTGTATTAGATAA
117722_4941150_SRGAP2A_2_1_1	AGATCCGTGCTCAGCTCACAGAGCAGATGAAATGCCTGGACCAGCAGTGTGAGCTTCGGGTGCAACTGTTGCAGGACCTCCAGGACTTCTTCCGAAAGAAGGCAGAGATTGAGATGGACT
117722_4941150_SRGAP2A_2_1_2	GGACCTCCAGGACTTCTTCCGAAAGAAGGCAGAGATTGAGATGGACTACTCCCGCAACCTGGAGAAGCTGGCAGAACGCTTCCTGGCCAAGACACGCAGCACCAAGGACCAGCAATTCAA
117722_4941151_SRGAP2A_3_1_1	GAAGGATCAGAATGTTCTCTCTCCAGTCAACTGCTGGAATCTCCTCTTAAACCAGGTGAAGCGGGAAAGCAGGGACCATACCACCCTGAGTGACATCTACCTGAATAATATCATTCCTCG
117722_4941151_SRGAP2A_3_1_2	CTCTTAAACCAGGTGAAGCGGGAAAGCAGGGACCATACCACCCTGAGTGACATCTACCTGAATAATATCATTCCTCGATTTGTACAAGTCAGCGAGGACTCAGGAAGACTCTTTAAAAAG
117722_4941152_SRGAP2A_5_1_1	GTGATGAAGACATATCACATGTACAATGCCGACAGCATCAGTGCTCAGAGCAAACTAAAGGAGGCGGAGAAGCAGGAGGAGAAGCAAATTGGTAAATCGGTAAAGCAGGAGGACCGGCAG
117722_4941152_SRGAP2A_5_1_2	TCGGTAAAGCAGGAGGACCGGCAGACCCCACGCTCCCCTGACTCCACGGCCAACGTTCGCATTGAGGAGAAACATGTCCGGAGGAGCTCAGTGAAGAAGATTGAGAAGATGAAGGAGAAG
117722_4941153_SRGAP2A_6_1_1	CGTCAAGCCAAGTACACGGAGAATAAGCTGAAGGCCATCAAAGCCCGGAATGAGTACTTGCTGGCTTTGGAGGCAACCAATGCATCTGTCTTCAAGTACTACATCCATGACCTATCTGAC
117722_4941153_SRGAP2A_6_1_2	AAGTACACGGAGAATAAGCTGAAGGCCATCAAAGCCCGGAATGAGTACTTGCTGGCTTTGGAGGCAACCAATGCATCTGTCTTCAAGTACTACATCCATGACCTATCTGACCTTATTGAT
117722_4941154_SRGAP2A_7_1_1	TGTTGTGACTTAGGCTACCATGCAAGTCTGAACCGGGCTCTACGCACCTTCCTCTCTGCTGAGTTAAACCTGGAACAGTCGAAGCATGAGGGTCTGGATGCCATCGAGAATGCAGTAGAA
117722_4941154_SRGAP2A_7_1_2	ATCGAGAATGCAGTAGAAAACCTGGATGCCACCAGTGACAAGCAGCGCCTCATGGAGATGTACAACAACGTCTTCTGCCCCCCTATGAAGTTTGAGTTTCAGCCCCACATGGGGGATATG
117722_4941155_SRGAP2A_9_1_1	GTAAAGAAGACAATGGAGGCCACCCTGCAAACCATCCAGGACATTGTGACTGTCGAGGACTTTGATGTGTCTGACTGCTTCCAGTACAGCAACTCCATGGAGTCCGTCAAGTCCACGGTC
117722_4941155_SRGAP2A_9_1_2	TTCCAGTACAGCAACTCCATGGAGTCCGTCAAGTCCACGGTCTCTGAAACCTTCATGAGCAAGCCCAGCATTGCTAAGAGGAGAGCCAACCAGCAAGAGACAGAGCAGTTTTATTTCACA
117722_4941156_SRGAP2A_10-14_1_1	AAAATGAAAGAGTACCTGGAGGGCAGGAACCTCATCACCAAGTTACAAGCCAAGCATGACCTTCTGCAGAAAACCCTGGGAGAAAGTCAGCGGACAGATTGCAGTCTAGCCAGGCGCAGC
117722_4941156_SRGAP2A_10-14_1_2	GGAGAAAGTCAGCGGACAGATTGCAGTCTAGCCAGGCGCAGCTCAACTGTGAGGAAACAGGACTCCAGCCAGGCAATTCCTCTGGTGGTGGAAAGCTGTATCCGGTTTATCAGCAGACAC
117722_4941156_SRGAP2A_10-14_1_3	CCTCTGGTGGTGGAAAGCTGTATCCGGTTTATCAGCAGACACGGACTACAGCATGAAGGAATTTTCCGGGTGTCAGGATCCCAGGTGGAAGTGAATGACATCAAAAATGCCTTTGAGAGA
117722_4941157_SRGAP2A_15_1_1	GAGAGGACCCCCTGGCTGGGGACCAGAACGACCATGACATGGATTCCATAGCTGGTGTCCTGAAGCTTTACTTCCGGGGGCTGGAACACCCTCTCTTCCCCAAGGACATCTTTCATGACC
117722_4941157_SRGAP2A_15_1_2	CTGGGGACCAGAACGACCATGACATGGATTCCATAGCTGGTGTCCTGAAGCTTTACTTCCGGGGGCTGGAACACCCTCTCTTCCCCAAGGACATCTTTCATGACCTGATGGCCTGCGTCA
117722_4941158_SRGAP2A_17_1_1	TTTATCACAGTTCAGTGAAGAGAACATGATGGACCCCTACAACCTCGCCATCTGCTTCGGGCCCTCGCTAATGTCAGTGCCAGAGGGCCACGACCAGGTGTCCTGCCAAGCCCACGTGAA
117722_4941158_SRGAP2A_17_1_2	CCAAGCCCACGTGAATGAGCTGATCAAAACCATCATCATCCAGCATGAGAACATCTTCCCAAGCCCCAGGGAGCTGGAGGGCCCTGTCTACAGCAGAGGAGGAAGCATGGAGGATTACTG
117722_4941159_SRGAP2A_19_1_1	AATGTGAGCCCATCGAGGCCATTGCCAAGTTTGACTACGTGGGCCGGACAGCCCGAGAGCTATCCTTTAAGAAGGGAGCATCCCTGCTGCTTTACCAGCGGGCTTCCGACGACTGGTGGG
117722_4941159_SRGAP2A_19_1_2	ATCCTTTAAGAAGGGAGCATCCCTGCTGCTTTACCAGCGGGCTTCCGACGACTGGTGGGAAGGCCGGCACAATGGCATCGACGGACTCATCCCCCATCAGTACATCGTGGTCCAAGACAC
117722_4941160_NRXN1_1_1_1	CCAACCCAACCCGAGCAGGCGGCAGAGAGCCGTATCCAGGCTCAGCAGAAGTGATCCGGGAGTCCAGCAGCACCACGGGTATGGTCGTTGGGATAGTAGCCGCTGCCGCCCTGTGCATCC
117722_4941160_NRXN1_1_1_2	AGTAGCCGCTGCCGCCCTGTGCATCCTTATCCTCCTCTATGCCATGTACAAGTACAGAAACCGGGATGAAGGCTCATACCATGTGGACGAGAGTCGAAACTACATCAGTAACTCAGCACA
117722_4941160_NRXN1_1_1_3	CGAAACTACATCAGTAACTCAGCACAGTCCAATGGGGCTGTTGTAAAGGAGAAACAACCCAGCAGTGCGAAAAGCTCCAACAAAAATAAGAAAAACAAGGATAAAGAGTATTATGTCTG
117722_4941161_NRXN1_3_1_1	GGCGTCAGCTCACAATCTTCAATAGCCAAGCAACCATAATAATTGGCGGGAAAGAGCAGGGCCAGCCCTTCCAGGGCCAGCTCTCTGGGCTGTACTACAATGGCTTGAAAGTTCTGAATA
117722_4941161_NRXN1_3_1_2	TGGCTTGAAAGTTCTGAATATGGCAGCCGAAAACGATGCCAACATCGCCATAGTGGGAAATGTGAGACTGGTTGGTGAAGTGCCTTCCTCTATGACAACTGAGTCAACAGCCACTGCCAT
117722_4941161_NRXN1_3_1_3	GAGTCAACAGCCACTGCCATGCAATCAGAGATGTCCACATCAATTATGGAGACTACCACGACCCTGGCTACTAGCACAGCCAGAAGAGGAAAGCCCCCGACAAAAGAACCCATTAGCCAG
117722_4941162_NRXN1_4_1_1	CACCAGGGAAAAATTGGAGTTAAGTTTAATGTTGGGACAGATGACATCGCCATTGAAGAATCCAATGCAATCATTAATGATGGGAAATACCATGTAGTTCGTTTCACGAGGAGTGGTGGC
117722_4941162_NRXN1_4_1_2	TTGAAGAATCCAATGCAATCATTAATGATGGGAAATACCATGTAGTTCGTTTCACGAGGAGTGGTGGCAATGCCACGTTGCAGGTGGACAGCTGGCCAGTGATCGAGCGCTACCCTGCAG
117722_4941163_NRXN1_5_1_1	CTGGGACGACATATATCTTTAGCAAAGGTGGTGGACAAATCACGTATAAGTGGCCTCCTAATGACCGACCCAGTACACGAGCAGACAGACTGGCCATAGGTTTTAGCACTGTTCAGAAAG
117722_4941163_NRXN1_5_1_2	GACCGACCCAGTACACGAGCAGACAGACTGGCCATAGGTTTTAGCACTGTTCAGAAAGAAGCCGTATTGGTGCGAGTGGACAGTTCTTCAGGCTTGGGTGACTACCTAGAACTGCATATA
117722_4941164_NRXN1_6_1_1	GCCCAGCACAACCTGCCAAGAGGACTCATGTTCCAATCAAGGTGTGTGCTTGCAACAATGGGATGGCTTCAGCTGTGACTGTAGTATGACTTCCTTCAGTGGACCACTCTGCAATGACC
117722_4941165_NRXN1_7_1_1	GTGACTTATATATAGGAGGAGTAGCTAAAGAAACATACAAATCCTTACCAAAACTTGTACATGCCAAAGAAGGCTTTCAAGGCTGCCTGGCATCAGTTGATTTAAATGGACGGCTTCCGG
117722_4941165_NRXN1_7_1_2	TTGTACATGCCAAAGAAGGCTTTCAAGGCTGCCTGGCATCAGTTGATTTAAATGGACGGCTTCCGGACCTCATCTCCGATGCTCTTTTCTGCAACGGACAGATCGAGAGAGGATGTGAAG
117722_4941166_NRXN1_8_1_1	GTACTTACATTACGTGTTTGATTTGGGAAATGGTGCTAACCTCATCAAAGGAAGCTCAAATAAACCTCTCAATGACAATCAGTGGCACAACGTGATGATATCAAGGGACACCAGCAACCT
117722_4941166_NRXN1_8_1_2	ATGACAATCAGTGGCACAACGTGATGATATCAAGGGACACCAGCAACCTCCACACTGTAAAGATTGACACAAAAATCACACAACGCAAATCACCGCCGGAGCCAGGAACTTAGACCTCAAGA
117722_4941167_NRXN1_9_1_1	GTCAAATGGCAGGTGATCATACTAGGCTGGAGTTCCATAACATAGAGACTGGCATCATCACAGAACGACGGTATCTTTCTTCTGTCCCCTCCAACTTCATTGGACACCTGCAGAGCTTGA
117722_4941167_NRXN1_9_1_2	CCTCCAACTTCATTGGACACCTGCAGAGCTTGACATTTAATGGAATGGCATACATTGACCTGTGTAAAAATGGCGACATAGATTACTGTGAGCTTAATGCCAGATTTGGCTTCAGGAACA
117722_4941167_NRXN1_9_1_3	GTGAGCTTAATGCCAGATTTGGCTTCAGGAACATCATAGCAGATCCTGTCACCTTCAAGACCAAATCGAGCTATGTTGCCTTAGCTACCTTGCAAGCCTACACTTCTATGCATCTTTTTT
117722_4941167_NRXN1_9_1_4	CCTTGCAAGCCTACACTTCTATGCATCTTTTTTTCCAGTTCAAGACAACATCCCTAGATGGATTAATTCTATATAACAGTGGGGATGGAAATGACTTTATTGTGGTTGAATTAGTTAAAG
117722_4941168_NRXN1_10_1_1	GCAAAGGTCCCGAGACTCTTTTTGCTGGCTATAACCTCAATGATAACGAGTGGCACACAGTGCGTGTAGTTCGGCGTGGAAAAAGTTTAAAGTTAACAGTGGATGACCAACAGGCCATGA
117722_4941168_NRXN1_10_1_2	AAGGTCCCGAGACTCTTTTTGCTGGCTATAACCTCAATGATAACGAGTGGCACACAGTGCGTGTAGTTCGGCGTGGAAAAAGTTTAAAGTTAACAGTGGATGACCAACAGGCCATGACAG
117722_4941169_NRXN1_11_1_1	AGGCAACGGTTTTGAGCTATGATGGGAGCATGTTTATGAAAATTCAGCTCCCCGTAGTCATGCATACGGAGGCTGAGGATGTTTCCTTACGGTTCCGATCCCAGCGTGCATATGGCATTC
117722_4941169_NRXN1_11_1_2	CCTTACGGTTCCGATCCCAGCGTGCATATGGCATTCTGATGGCAACCACTTCTAGAGACTCTGCTGACACCCTCCGCCTGGAGCTAGACGCAGGACGTGTGAAACTGACGGTCAATCTAG
117722_4941170_NRXN1_12_1_1	GTACCATTTCTGTCAACACGTTGCGTACTCCCTACACTGCTCCTGGTGAGAGTGAGATTCTGGACCTGGATGATGAGTTGTACCTGGGGGGGCTGCCAGAAAATAAAGCTGGCCTTGTCT
117722_4941170_NRXN1_12_1_2	GGGGCTGCCAGAAAATAAAGCTGGCCTTGTCTTCCCCACCGAGGTGTGGACTGCTCTGCTCAACTATGGCTACGTGGGCTGCATCAGGGATTTGTTCATCGATGGCCAAAGCAAAGATAT
117722_4941170_NRXN1_12_1_3	GATTTGTTCATCGATGGCCAAAGCAAAGATATCCGGCAAATGGCTGAAGTTCAAAGTACTGCTGGAGTGAAGCCTTCCTGCTCAAAGGAAACAGCAAAACCGTGCCTTAGCAACCCTTGC
117722_4941170_NRXN1_12_1_4	AAACAGCAAAACCGTGCCTTAGCAACCCTTGCAAAAACAATGGCATGTGCAGGGATGGGTGGAACAGATATGTCTGTGATTGTTCCGGAACAGGCTATCTTGGCAGGTCCTGTGAGAGAG
117722_4941171_NRXN1_13_1_1	GTTGTATATAAAAATAATGATGTGAGGCTGGAATTATCTCGACTTGCCAAGCAAGGAGATCCTAAGATGAAGATCCATGGAGTGGTGGCATTTAAATGTGAGAATGTTGCAACTTTAGAC
117722_4941171_NRXN1_13_1_2	TTGCAACTTTAGACCCAATCACCTTTGAAACCCCAGAGTCTTTCATCTCTTTGCCTAAATGGAATGCAAAGAAAACTGGCTCCATATCATTTGATTTCCGTACAACAGAGCCAAATGGCC
117722_4941171_NRXN1_13_1_3	AGAGCCAAATGGCCTCATCTTATTTAGCCATGGCAAGCCAAGACATCAGAAAGATGCCAAGCACCCACAGATGATAAAGGTGGACTTCTTTGCTATTGAGATGCTAGATGGCCACCTCTA
117722_4941171_NRXN1_13_1_4	GATGGCCACCTCTACCTCCTCCTGGACATGGGGTCAGGTACTATAAAAATAAAAGCCCTGTTGAAGAAAGTGAATGATGGAGAATGGTATCATGTGGACTTCCAGAGAGACGGACGGTCA
117722_4941172_NRXN1_14_1_1	GTGACAATATCAGTGGATGGGATTCTTACCACAACGGGCTACACGCAAGAAGATTATACCATGCTGGGGTCTGATGACTTTTTCTATGTTGGAGGCAGTCCCAGCACAGCCGACCTTCCA
117722_4941172_NRXN1_14_1_2	ACGCAAGAAGATTATACCATGCTGGGGTCTGATGACTTTTTCTATGTTGGAGGCAGTCCCAGCACAGCCGACCTTCCAGGGTCACCAGTCAGTAACAACTTTATGGGCTGTCTCAAAGAG
117722_4941173_NRXN1_15_1_1	AGGAAAAGAAGAATATATTGCCACGTTCAAAGGATCTGAATACTTCTGCTACGACTTGTCTCAAAACCCCATTCAAAGCAGCAGTGATGAAATAACTCTGTCATTTAAAACCCTTCAGAG
117722_4941173_NRXN1_15_1_2	TAACTCTGTCATTTAAAACCCTTCAGAGGAATGGACTGATGCTTCACACTGGGAAATCGGCTGATTATGTCAATCTTGCCCTGAAAAATGGAGCTGTCTCTCTGGTCATTAATTTGGGAT
117722_4941173_NRXN1_15_1_3	GCTGTCTCTCTGGTCATTAATTTGGGATCAGGGGCCTTTGAAGCACTAGTGGAGCCTGTGAATGGAAAGTTTAATGATAATGCCTGGCATGATGTGAAAGTCACCAGGAATCTGCGTCAG
117722_4941174_NRXN1_17_1_1	ATGGGGACGGCGCTGCTCCAGCGCGGGGGCTGTTTTCTTCTGTGCCTCTCGCTGCTGCTCCTGGGCTGCTGGGCGGAGCTGGGCAGCGGGCTGGAGTTTCCGGGCGCCGAGGGCCAATGG
117722_4941174_NRXN1_17_1_2	AGGGCCAATGGACGCGCTTCCCCAAGTGGAACGCCTGCTGCGAGAGCGAGATGAGCTTCCAGCTCAAGACTCGCAGCGCCCGCGGCCTCGTGCTCTACTTCGACGACGAGGGCTTCTGCG
117722_4941174_NRXN1_17_1_3	GGGCTTCTGCGACTTCCTGGAGCTGATTCTGACGCGCGGCGGCCGCCTGCAGCTCAGCTTCTCCATCTTCTGCGCTGAGCCTGCGACGCTCCTGGCCGACACGCCGGTTAACGACGGCGC
117722_4941174_NRXN1_17_1_4	AACGACGGCGCCTGGCACAGCGTGCGCATCCGCCGCCAGTTCCGCAACACCACGCTCTTCATCGACCAGGTGGAGGCCAAGTGGGTGGAGGTCAAGTCCAAGCGCAGGGACATGACGGTG
117722_4941174_NRXN1_17_1_5	ACATGACGGTGTTCAGCGGCCTTTTCGTCGGGGGGCTGCCCCCGGAACTGCGCGCCGCGGCGCTCAAGCTCACCCTGGCCTCGGTGAGGGAGCGGGAGCCCTTCAAGGGGTGGATTCGTG
117722_4941174_NRXN1_17_1_6	GTGGATTCGTGACGTGAGGGTCAACTCCTCGCAGGTCCTGCCCGTGGACAGCGGCGAGGTGAAGCTGGACGATGAGCCGCCCAACAGCGGCGGGGGAAGCCCGTGCGAGGCGGGCGAGGA
117722_4941174_NRXN1_17_1_7	GCGGGCGAGGAGGGCGAGGGCGGGGTGTGCCTCAACGGAGGTGTGTGCTCCGTGGTGGACGACCAGGCCGTGTGCGACTGCTCGCGAACCGGCTTCCGCGGCAAGGACTGCAGCCAAG
117722_4941175_CD8B_1_1_1	GCCTCAAGGGGAAGGTATATCAGGAACCTTTGTCCCCCAATGCCTGCATGGATACTACAGCAATACTACAACCTCACAGAAGCTGCTTAACCCATGGATCCTGAAAACATAGGCAAGAAG
117722_4941176_CD8B_2-4_1_1	TTGATTTCCTTCCCACCACTGCCCAGCCCACCAAGAAGTCCACCCTCAAGAAGAGAGTGTGCCGGTTACCCAGGCCAGAGACCCAGAAGGGCCCACTTTGTAGCCCCATCACCCTTGGCC
117722_4941176_CD8B_2-4_1_2	TTGTAGCCCCATCACCCTTGGCCTGCTGGTGGCTGGCGTCCTGGTTCTGCTGGTTTCCCTGGGAGTGGCCATCCACCTGTGCTGCCGGCGGAGGAGAGCCCGGCTTCGTTTCATGAAACA
117722_4941177_CD8B_5_1_1	TCTCCATGGCAACTCAGTCCTCCAGCAGACCCCTGCATACATAAAGGTGCAAACCAACAAGATGGTGATGCTGTCCTGCGAGGCTAAAATCTCCCTCAGTAACATGCGCATCTACTGGCT
117722_4941177_CD8B_5_1_2	GAGACAGCGCCAGGCACCGAGCAGTGACAGTCACCACGAGTTCCTGGCCCTCTGGGATTCCGCAAAAGGGACTATCCACGGTGAAGAGGTGGAACAGGAGAAGATAGCTGTGTTTCGGGA
117722_4941177_CD8B_5_1_3	TGCAAGCCGGTTCATTCTCAATCTCACAAGCGTGAAGCCGGAAGACAGTGGCATCTACTTCTGCATGATCGTCGGGAGCCCCGAGCTGACCTTCGGGAAGGGAACTCAGCTGAGTGTGG
117722_4941178_CD8B_6_1_1	CCCGGCCAGCTCCTCACCCACCCCAGCCGCGACTGTCTCCGCCGAGCCCCCGGGGCCAGGTGTCCCGGGCGCGCCACGATGCGGCCGCGGCTGTGGCTCCTCTTGGCCGCGCAGCTGACA
117722_4941179_TISP43_1_1_2	GGGAAGCAGGGGCAGAGCGCTCAGGCCAGGGCCGGCGGGGGCACTGCGCACTGGCCGCCTCCCTCCTCTGGCCAATCCTCCTGCCGCTGCGCGTCTAGTCCACCTCGTCCCTCTCTGCAG
117722_4941180_TISP43_2_1_1	CCCACAGAGCACAGCAAGCAGATGACAGTGAATAAGATCATGGTGCACGCTGACTATAACGAGTTGCACCGCATGGGGAGTGACATCACCCTGCTGCAGCTGCACCGTCATGTGGAATTC
117722_4941180_TISP43_2_1_2	TGCACCGTCATGTGGAATTCAGCTCCCACATCCTCCCCGCCTGCCTTCCGGAACCAACCACGTGGCTGGCCCCTGACAGCTCCTGCTGGATATCTGGTTGGGGAATGGTCACCGAGGATG
117722_4941181_TISP43_3_1_1	TCTTCCTGCCTGAGCCCTTCCAACTTCAGGAGGCAGAGGTCGGTGTCATGGACAACACTGTCTGCGGATCCTTTTTCCAGCCCCAGTACCCCGGCCAGCCAAGCAGCAGTGACTACACCA
117722_4941181_TISP43_3_1_2	ACTGTCTGCGGATCCTTTTTCCAGCCCCAGTACCCCGGCCAGCCAAGCAGCAGTGACTACACCATCCACGAGGACATGCTGTGCGCTGGGGACCTCATAACAGGAAAGGCCATTTGCCGA
117722_4941182_TISP43_4_1_1	GCATGTTCCTACGTGAAGCATGGAGTTACGCAATGGCCTGCCTGTCCATGAAGTTCCTCTGGCCTCCTCCGATTCTGTGGCTGCAAATCCACACCTAGGAGCTCCGCAGGCCTGCTGGAG
117722_4941182_TISP43_4_1_2	CCGCAGGCCTGCTGGAGCACAGCGCCTGTAGGTTTCTGCGCCTGGAACCCACCTCAGGACCACAAGCCGCATTCAGTCTCTTTTTATAGATCTGGAACCCGAAGCCCAGCGTGGTTAGGA
117722_4941182_TISP43_4_1_3	GCCCAGCGTGGTTAGGAACTTGTCTAAGGTCACTTGGTAGTAAATACTGGAACAGGGATTCAAACCCACATCTCCAGAGAGAGCTCAGAAGCCTTGTGCTTTCCATGAGTGCAGCTGTTT
117722_4941182_TISP43_4_1_4	CATGAGTGCAGCTGTTTCCAGAGGCCTCCGGCCCTGTCTCACCTTCCCAGACTGCCTTGAGAGGCTCTAGCTGGCCTGTCTCTAGTGTGGCCAGATTCTATCTGGTCAGTGGCAAGGGAG





```
117722_4941311_HYDIN_37_1_1     GCTCAAGAGGAGCAGACCAGCTCATCTAAGGGGGCAAACAGAAAATGAAAGAAAAGATAGACCAAGTCTTTCGAGATTCAGAAAGACAAGCGTCACATGGCCTTAAACAGGAAGGTCCTT
117722_4941312_HYDIN_38_1_1     ATCCTAGACATCTTGGGTCTCGGGTGCCTCCGGACCACCCATCCCGCCTCCCGCCTTATTCTCAATCGTCTCCTACCCGGTGAAGCGGCCACCTTTGACCATGACAGACGACCTGGAGCAT
117722_4941313_HYDIN_39_1_1     GTTCCTCCGCTCACCAAAGTGGATGTCAAGATGGAGACAATCGAAAGGAAAATATCTGTTAGGGAACAAACAATGTCTGAGAAGGAAGAGCTAAATAAGAAGAAAAGGAACATGGGCGAT
117722_4941314_HYDIN_40_1_1     GGAGCAGGAGAGGCTGGCAAAGGAAATGCAAGAAAAGAAGCTACAGCAGGAGCTGGAGCGACAAAAGGAAGAGGAGTGAGCTGAAACGGAGGGTCAAAAAAGGAAAGCAGGGACCCATTAA
117722_4941315_HYDIN_41_1_1     AACGCAAGCACAAGGGAGCTCTTGAGAAAGAGAAGGAGCGTCTCTCCAAAACATGGATGAGGAAGAATATGATGCCCTGACTGAGGAGGAGAAACTCACATTCGATCGGGGGATTCAGCAGG
117722_4941316_HYDIN_42_1_1     CTGAGTGACTGCTACCGAGGAGTGGTGTTTGATGGCCTCGACACTCTCTTTGCTCAGAATGCTGCAGCCGCCCTCCTCTGCCTGCTGAAGGCCATTGGCAGCCGGGAGCATATATACATT
117722_4941317_HYDIN_43_1_1     ATTTCCTCCAGCCCTCTCCCCCCGGGGCCCATCCACCGCTGGCTCAGTGTTAGTCCCAGTGTCGGAGGCGAGACCGGGCTGATGAGCTGTGTGCTCCCGGATGAACTTCTCGTGCAGATC
117722_4941318_HYDIN_44_1_1     CCCAGGAGGCAGCTGTGGGTCAAAACGTCATAGGGCAAGGACGACTGAGCACTGACACCTTGGGCAAGTTAGCCTCCGAGATGACTCTGGTGGCCCCAGAAATTAAACCTGGAAAGAGTG
117722_4941319_HYDIN_45_1_1     GAAAGTCAGCCAATGCCTTAGCGTGGCCAAGTACTACTCAACGCAGCCTGCCTGAGCATCGACTCCATGTGCTGGAAGCTGTGGCCAACAGCAACAACATCCCAGGGATCCGGGCCGTG
117722_4941320_HYDIN_46_1_1     CTGACAAGCTTCAGAGCATTGACAGCCACTCCATGGAGGAAGTTGGAGAGGTGGAAAACAACCCAGTGAGCAAAGCAATCGCTCGCCACCTGGGCATTGACATTTCTGCAGAAGGCCGCC
117722_4941321_HYDIN_47_1_1     AAGAGGAAGATATAACCTCATCAGATCAGGGAACCTCCAATAGCACAAAGAGGACATCGCTGAGCCGAGGGATCTCTGTCACATCCAACCTGGAAGAATGGCACGCCCTGTTGGTCGAGT
117722_4941322_HYDIN_48_1_1     ATCTTGCGGAAGCTGAAGGGCTATGATTCCTACAACACCCTGCTGCTGCCTCCCCGCAACCCTGGGGAGAAGCTGCCCCCAGAACTGTACGAGTACTTCAAAGAGATAAAGAAGTCAAAA
117722_4941323_HYDIN_49_1_1     AATTCTACAGCCAAACCCTGGTGTTTCAGATTGCCCAGAGTGCTCAAAAGCTTACCCTCCTGGCACGTGGGCAAGGTCTAGAGCCACGCCTGGAATTTAGTCCTTCAGTCCTGGATCTGG
117722_4941324_HYDIN_50_1_1     CTGGAGAAACACATGCCGAAGTACTTAAGACAGAAACTACGCGCTGAATTAAAGCCAAAGACACGGATCTTCGAAATCCAGCCCATTTCTGGAGTCTTGGATCCTGGTGAGAAGTCCAAC
117722_4941325_HYDIN_51_1_1     GTGGTTGGAGGGCCAACAGTTCACATCTGTCTCCAAGCCAAGGTGACCATTCCAACCATGACTCTCTCTCGTGGAAAAGTGGACTTTGCCACAATTCAGTCTGGACAGTGCCTGGTGGAA
117722_4941326_HYDIN_52_1_1     GATTCAGTACTGAGCTAGATCGTGTAAAGAATCTACCTCATTGTGAAACGGAAATATTTGAAGTGAGATTTGACCCACAGGGGGCCAATCTTCCTGTTGGAAGCAAAGAAGTCATTCTGC
117722_4941327_HYDIN_53_1_1     AATCCAGCTACCAGAGTACATCCTGGACTTTGGCTACATCATCCTTGGCGAAGTCCGAACCCACATCATCAAGATCATCAACACCAGTCACTTTCCAGTGTCATTCCATGCAGACAAGCG
117722_4941328_HYDIN_54_1_1     GTAAGTGCTCATCTCCAGATGGAGGTAGAAAGACTTATAGTCCAAAGCTATGTCCTAGAACATCAGAAAACAACCACCCCTGATCCTATGGATGACCCCTGCTTCAGCCATCGGAGTCGC
117722_4941329_HYDIN_55_1_1     CAAATGAAAAGTATGAAATGTTCTTGAATCAAGCCAGGAAAAACACAGACAAAGAGTATAACAAATGTGAAATGCTCGATCACTTTGACATAATAACTGAGGAAGTGCCAGAAGACGAGC
117722_4941330_HYDIN_56_1_1     GGCTTTATCAGTTCACATCAAGAGCAGGTATTAAAAGTTTACTACCTACCTGGAGTACCTGAGGTCTTTAAAAGGAGTTTCCAGATACAGATCGCCCACCTGGACCCAGAAAATATCACT
117722_4941331_HYDIN_57_1_1     CTCTTTGCACATGTCACAGAGAGGGAAATCACGCTGACGAACATGGGGAAAGTTGGCTTTGAGTTCAAGGTTCTGACTGACCACCAGTCTTCTCCAGACAACCTTCTCCCTGGAGTGCCA
117722_4941332_HYDIN_58_1_1     GTGTTTGATATTTTGCCCCTGTTTGGAGTGTTGCAGCCACACAGTAGCCACCAAATATCGTTCCACCTTCTATGGACACGCTAACATCATTGCACAAGCTAAAGCTCTGTGTGAAGTGGAA
117722_4941333_HYDIN_59_1_1     AAATAAAAATTAGCTCTACTGAAATAGAAAGAATATACTCAAGCCAGAGCCAGGTGGAGGATCAGGAATCCCTACAGACCTGTGAACAGAATGAGATGCTTTCCATTGGGATAGAAGAA
117722_4941334_HYDIN_61_1_1     GTTTGTGACATTGCCGAAGAAGCCCTACAGTGCCCCAGTGTCCCAGATGGAGTCCATCCCAGCAACCTCAGAGGCTGCCAGCCCACCAGCAATCCTAGTTACAGTAGAGTCCCCCGAGAT
117722_4941335_HYDIN_62_1_1     CCTATTAAACTGGAAATTGATGAAGAAAAAAACCTGCTGATCAAGTTTGACCCTTCCTACAGAAACGATCTGAACAACTGGGTGGCAGAAGAAATTCTAGCAATTAAGTATGTGGAACAC
117722_4941336_HYDIN_63_1_1     AAACCTAACAGTATCCTGAAACCTGATTACCAGCCCTTGGCCATAAAGAACATTTCCACCCTGCCCGTGAACTTGTTGCTGTCAACATCTGGACCCTTCTTTATATGTGAGACTGATAAA
117722_4941337_HYDIN_64_1_1     ATAGTGAAAGAGAAGCTGGTGTGCCACGCCATCATCGGGGCACAGAAGGGGAAGAGCTTGGTGATGGCTGTGAACATCACCTGTGAGTTCGTCGCACCTCTCATCCAGCTCTCCACCAAG
117722_4941338_HYDIN_65_1_1     CCTGGATACCCACTATTACCACTTTAAGTTGATCAACAAGGGACGTCGGATCCACAACAGTTGTTCTGGATGAATGATAGCTTCCGACCCCAGGCCAAGCTGACTAAGAAGGCCGGGTTAA
117722_4941339_HYDIN_66_1_1     GCACACAAAAAATCCCTTTTGGACGATTGAACCCAATGAAGGCATGGTTCCTCCAGAAACTGATGTTCAACTGGCACTGACCGCCAACCTGAATGACACACTGACATTCAAGGACTGTGTT
117722_4941340_HYDIN_67_1_1     GTATGTCACTTAAAGAGCGCTGGAGAAGGCCAGTTATCTACGTCCATCCCAATCAAGTGGACTTCGGGAATATCTACGTCCTAAAAGACTCTTCCAGGATTCTCAACCTATGCAACCAG
117722_4941341_HYDIN_68_1_1     GTGTGTGAGGAGGTGCCTACTGTGCTGTTTTCCAGCCCCACCCCCAGCGGGGTCATCTCCCCAAGCAGCACCATCCACATACCACTGGTCCTGGAGACCCAGGTCACTGGAGAACACAGA
117722_4941342_HYDIN_69_1_1     GTGTGTTGTACCTGCCCTCCACCTGGTCAATACAGAGGTGGACTTTGGGCACTGCTTCCTGAAGTACCCGTATGAGAAAACACTCCAGCTTGCCAATCAAGATGACCTCCCAGGATTCTA
117722_4941343_HYDIN_71_1_1     GGTTTCCTCATACCTTGATATGTTCCCTCAATAATACCTCTTTGATCCCCATGACTTACAAACTGCGTATCCCTGGGGATGGCCTTGGCCATAAAAGCATTTCATATTGTGAGCAGCATG
117722_4941344_HYDIN_73_1_1     GCGATACTGTACAACAAAGGCAGCATCGATGCTCTCTTCCAACATGACCCCTCCACTTCAGCTTTGGGGCCTCGCTTTTGTTTTCAGTCCCAAGGAAGGCATCATTGAACCAAGTGGAGTC
117722_4941345_HYDIN_76_1_1     GGCCTGTGATGATCTGATCAAAGAGGAGAAGGATGAGACTGATGAGTTTTTGAAGAGTGCATTACTGATCCTTTACTCCGAGAACATCTTTCTGTTCTGTCCCGAACCTTTGCGAATCA
117722_4941346_HYDIN_77_1_1     GTGAAAAGGTGTTTGTATCTCTCTATGGAGCTGCCATAGACATGAATATAAGGCTGGATAAGAATTCCTTGACCATCGAGAAAACCTACATATCTCTGGCCAATCAGCGAACTATAACCA
117722_4941347_HYDIN_78_1_1     GCCTTTCTCTATAGAACCAGCTATTGGAACTCTTAATGTGGGAGAGTCCATGCAACTGGAAGTGGAGTTTGAGCCACGAGGTGTGGGCGATCACAGTGGAAGACTTATCGTGTGTTATGA
117722_4941348_HYDIN_79_1_1     GATTACGCCCATACGTTGACCTGTGTTACTGAAAGAGAAAAGTTTATTGTACCCATCAAAGCTAGAGGGGCACGAGCCATTCTCGATTTTCCTGACAAGCTGAATTTTTCCACTTGTCCT
117722_4941285_HYDIN_6_1_3      CTGCTGCCAGGCCCTGGAGATCTCACTGGACCAGGAACATATTCCCTTTGGACCCGTGGTGTATCAGCGCAAGCCACACGTCGCATCCTCATGATGAACACAGGCGATGTGGGTGCAAG
117722_4941338_HYDIN_65_1_3     ACGATGACACCTTGACAACACATCCATGGACTGGTCGCCCCACCAGCCAGGAAGACATAAGTATCTGAGTTCACAGAATATCATCGAGGACAATGATATGGAAGACTGGGTGGCAG
117722_4941313_HYDIN_39_1_3     CCCCGAAGTTTAAGACCTATGAATTGACACTGAAGGATGTCCAGAACATCCTCATGTACTGGGACCGGAAGCAAGGAGTCCAGCTGCCTCCTGCAGGGATGGAGGAAGCGCCCCATGAGC
117722_4941335_HYDIN_62_1_3     TTTTGGCTGCATCCTGAACGATACTGAGCTCATTCGCTACGTTACTCATCACCAACTGCAGTCCGTTGGTTGTGAAGTTTCCTCTTGGTGAATGATGAGGAAAATCAGATAAG
117722_4941338_HYDIN_65_1_3     GGCTCTCAGGAGCCCAGGGATCCACAGAGCCCGTGTTTCATCTCCACCCCGCCAGCATGGAGCTGTACCCAGGCCAGGCAATTGATGTGATACTCGAAGGCTATTCTGCTACTCCCAGG
117722_4941313_HYDIN_39_1_4     GAGGAAGCGCCCCATGAGCCCGACGACCAGCGCCAGGTCCCCTTGGGTGGGCGCAGGGGCCGAAGGACCGGGAGAGAGAGCGCCTGGAGAAGGAGCGCCTGGAG
117722_4941313_HYDIN_39_1_5     GGAGAAGGAGCGCCTGGAGAGGGAGAAGGCGGAGCGGAGCGCCTGGAGAAGCTGCGAGCCCTGGAGGAGCGGAGCGACTGGGAGGGGGAAGGGGAGGAGGACCACGAAGGGAAGAAGGA
117722_4941313_HYDIN_39_1_6     ACCACGAAGGGAAGAAGGAGAAGGACTGGGCGTACCTTCCTAGACATCCAGACACCAGCATTTGAGCTTGAGCTGGAAGCAGGCGCTAGAGAGCGCACAAGCTTCCCAAAGGAGAGC
117722_4941349_ROCK1_1_1_1      TAAGTTATGAGTTACATCAGCAAGAGATATCGTGCTGTTAGCATGCTTCTCAGGATGACAAAAAAATGGGTAACTCATTTAGTAAAGAAATCCCCACATCTGGTTTTG
117722_4941349_ROCK1_1_1_2      GAAAATCCCTAAGAATCCACCATCTGGTTTTGTTCGTGCTTCCCCTCGAACGCTTTCTACAAGATCCACTGCAAATCAGTCTTTCCGGAAAGTGGTCAAAAATACATCTGGAAAAACTAG
117722_4941350_ROCK1_2_1_1      ATACTATATGCAAATGAAGGTGAATGTAGAAAAGATGTAGAGATGGAACCAGTACAACAAGCTGAAAAAACTAATTTCCAAAATCACAAAGGCCATGAGTTTATTCCTACACTCTACCAC
117722_4941350_ROCK1_2_1_2      TAATTTCCAAAATCACAAAGGCCATGAGTTTATTCCTACACTCTACCACTTTCCTGCCAATTGTGATGCCTGTGCCAAACCTCTCTGGCATGTTTTTAAGCCACCCCCTGCCCTAGAGTG
117722_4941350_ROCK1_2_1_3      GTGCCAAACCTCTCTGGCATGTTTTTAAGCCACCCCCTGCCCTAGAGTGTCGAAGATGCCATGTTAAGTGCCACAGAGATCACTTAGATAAGAAAGAGGACTTAATTTGTCCATGTAAAG
117722_4941351_ADNP_1_1_1       GACTATCGGACAAAACCTTTCTGCTGCAGCGCTTGTCCATTTTCCTCAAAATTCTTCTCTGCCTACAAAAGTCATTTCCGCAATGTCCATAGTGAAGACTTTGAAAATAGGATTCTCCTT
117722_4941351_ADNP_1_1_2       CCTTAATTGCCCCTACTCTACCTCTAATGCAGACAAAAGACTTTGGAAACACACATTAAATATTTCATGCTCCGAACGCCAGCTGCACCAAGTAGCAGCCTCAGCACTTTCAAAGATAA
117722_4941351_ADNP_1_1_3       ATAAAAACAAAAATGATGGCCTTAAACCTAAGCAGGCTGACAGTGTAGAGCAAGCTGTTTATTACTGTAAGAAGTGCACTTACCGAGATCCTCTTTATGAAATAGTTAGGAAGCACATTT
117722_4941351_ADNP_1_1_4       ATTTACAGGGAACATTTTCAGCATGTGGCAGCACCTTACATAGCAAAGGCAGGAGAAAAATCACTCAATGGGGCAGTCCCCTTAGGCTCGAATGCCCGAGAAGAGAGTAGTATTCACTGC
117722_4941351_ADNP_1_1_5       CTGCAAGGCGATGCCTTTTCATGCCAAAGTCCTATGAAGCTTTGGTACAGCATGTCATCGAAGACCATGAACGTATAGGCTATCAGGTCACTGCCATGATTGGGCACACAAATGTAGTGGT
117722_4941351_ADNP_1_1_6       TGGTTCCCCGATCCAAACCCTTGATGCTAATTGCTCCCAAACCTCAAGACAAGAAGAGCATGGGACTCCCACCAAGGATCGGTTCCCTTGCTTCTGGAAATGTCCGGTCTTTACCATCAC
117722_4941351_ADNP_1_1_7       TCACAGCAGATGGTGAATCGACTCTCAATACCAAGCCTAACTTAAATTCTACAGGAGTCAACATGATGTCCAGTGTTCATCTGCAGCAGAACAACTATGGAGTCAAATCTGTAGGCCAG
117722_4941351_ADNP_1_1_8       CCAGGGTTACAGTGTTCGTCAGTCAATGAGACTGGGTCTAGGTGGCAACGCACCAGTTTCCATTCCTCAACAATCTCAGTCTGTAAAGCAGTTACTTCCAAGTGGAAACGGAAGGTCTTA
117722_4941351_ADNP_1_1_9       CTTATGGGCTTGGGTCAGAGCAGAGGTCCCAGGCACCAGCAAGATACTCCCTGCAGTCTGCTAATGCCTCTTCTCTCTCATCGGGCCAGTTAAAGTCTCCTTCCCTCCTCTCAGTCACAGG
117722_4941351_ADNP_1_1_10      CAGGCATCCAGAGTGTTAGGTCAGTCCAGTTCCAAACCTGCTGCAGCTGCCACAGGCCCTCCCCCAGGTAACACTTCCTCAACTCAAAAGTGGAAAATATGTACAATCTGTAATGAGCTT
117722_4941351_ADNP_1_1_11      GCTTTTTCCTGAAAATGTCTATAGTGTGCACTTCGAAAAAGAACATAAAGCTGAGAAAGTCCCAGCAGTAGCCAACTACATTATGAAAATACACAATTTTACTAGCAAATGCCTCTACTG
117722_4941351_ADNP_1_1_12      ACTGTAATCGCTATTTACCCACAGATACTCTGCTCAACCATATGTTAATTCATGGTCTGTCTTGTCCATATTGCCGTTCAACTTTCAATGATGTGGAAAAGATGGCCGCACACATGCGGA
117722_4941351_ADNP_1_1_13      CGGATGGTTCACATTGATGAAGAGATGGGACCTAAAACAGATTCTACTTTGAGTTTTGATTTGACATTGCAGCAGGGTAGTCACACTAACATCCATCTCCTGGTAACTACATACAATCTG
117722_4941351_ADNP_1_1_14      TCTGAGGGATGCCCCAGCTGAATCTGTTGCTTACCATGCCCAAAATAATCCTCCAGTTCCTCCAAGCCACAGCCAAAGGTTCAGGAAAAGGCAGATATCCCTGTAAAAGTTCACCTCA
117722_4941351_ADNP_1_1_15      CTCAAGTCGCAGTCCCTATAAAAAAGATGTTGGGAAACCCTTTGTCCTCTTTGCTTTTCAATCCTAAAAGGACCCATACTGATGCACTTGCACATCACTTACGAGAGGCGCACCAAG
117722_4941351_ADNP_1_1_16      CAAGTTATTCAGACGGTTCATCCAGTTGAGAAAAAGCTCACCTACAAATGTATCCATTGCCTTGGTGTGTATACCAGCAACATGACCGCCTCAACTATCACTCTGCATCTAGTTCACTGC
117722_4941351_ADNP_1_1_17      CTGCAGGGGCGTTGGAAAGACCCAAAATGGCCAGGATAAGACAAATGCACCCTCTCGGCTTAATCAGTCTCCAAGTCTGGCACCTGTGAAGCGCACTTACAGACAAATGGAATTTCCCTT
117722_4941351_ADNP_1_1_18      CCTTACTGAAAAAACGAAAGTTAGATGATGATAGTGATTCACCCAGCTTCTTTGAAGAGAAGCCTGAAGAGCCTGTTGTTTTAGCTTTAGACCCCAAGGGTCATGAAGATGATTCCTATG
117722_4941351_ADNP_1_1_19      TATGAAGCCAGGAAAAGCTTTCTAACAAAGTATTTCAACAAACAGCCCTATCCCACCAGGAGAGAAATTGAGAAGCTAGCAGCCAGTTTATGGTTATGGAAGAGTGACATCGCTTCCCAT
117722_4941352_DYRK1A_4_1_1     GTTGTAAAGGCATATGATCGTGTGGAGCAAGAATGGGTGCCATTAAAATAATAAAGAACAAGAAGCTTTTCTGAATCAAGCACAGATAGAAGTGCGACTTCTTGAGCTCATGAACAAA
117722_4941352_DYRK1A_4_1_2     AAGAATGGGTTGCCATTAAAATAATAAAGAACAAGAAGCTTTTCTGAATCAAGCACAGATAGAAGTGCGACTTCTTGAGCTCATGAACAAACATGACACTGAAATGAAATACTACATAG
117722_4941353_DYRK1A_1_1_1     GAGGAGAGACTTCAGCATGCAAACCTTCATCTGTTCGGCTTGCCACCGTCATTTTCATTCCATGCTGCTGGCCTTCAGATGGCTGGACAGATGCCCCATTCACATCAGTACAGTGACCGTC
117722_4941353_DYRK1A_1_1_2     ATGGCTGGACAGATGCCCCATTCACATCAGTACAGTGACCGTCGCCAGCCAAACATAAGTGACCAACAGGTTTCTGCCTTATCATATTCTGACCAGATTCAGCAACCTCTAACTAACCAG
117722_4941354_DYRK1A_2_1_1     TGATGCCTGATATTGTCATGTTACAGAGGCGGATGCCCCAAACCTTCCGTGACCCAGCAACTGCTCCCCTGAGAAAACTTTCTGTTGACTTGATCAAAACATACAAGCATATTAATGAG
117722_4941355_DYRK1A_3_1_1     GTTTACTATGCAAAAAGAAGCGAAGACACCAACAGGGCCAGGGAGACGATTCTAGTCATAAGAAGGAACGGAAGGTTTACAATGATGGTTATGATGATGATAACTATGATTATATTTGTA
117722_4941355_DYRK1A_3_1_2     CGGAAGGTTTACAATGATGGTTATGATGATGATAACTATGATTATATTGTAAAAAACGGAGAAAAGTGGATCGTTACGAAATTGACTCCTTGATAGGCAAAGGTCCCTTTGGACAG
117722_4941356_DYRK1A_5_1_1     TGCATTGAAATCGCCACTTTATGTTTCGAAACCATCTCTGTTTAGTTTTTGAAATGCTGTCCTACCAACCTCTATGACTTGCTGAGAACACCATTTCCGAGGGGTCTCTTCTGAACTCTAA
117722_4941356_DYRK1A_5_1_2     GAAACACCAATTTCCGAGGGGTCTCTTTGAACCTAACACGAAAGTTTGCGCAACAGATGTGCACTGCACTGCTTTTCCTTGCGACTCCAGAACTTAGTATCATTCACTGTGATCTAAAAC
117722_4941356_DYRK1A_5_1_3     CTCCAGAACTTAGTATCATTCACTGTGATCTAAAACCTGAAAATATCCTTCTTTGTAACCCCAAACGCAGTGCAATCAAGATAGTTGACTTTGGCAGTTCTTGTCAGTTGGGGCAGAGG
117722_4941357_DYRK1A_6_1_1     ATATACCAGTATATTCAGAGTCGCTTTATCGGTCTCCAGAGGTGCTACTGGGAATGCCTTATGACCTTGCCATTGATATGTGGTCCCTCGGGTGTATTTTGGTTGAAATGCACACTGGA
117722_4941357_DYRK1A_6_1_2     TATCGGTTCCTCAGAGGTGCTACTGGGAATGCCTTATGACCTTGCCATTGATATGTGGTCCCTCGGGTGTATTTTGGTTGAAATGCACACTGGAGAACCTCTGTTCAGTGGTCCAATGAG
117722_4941358_DYRK1A_7_1_1     GTAGATCAGATGAATAAAATAGTGGAAGTTCTGGGTATTCCACCTGCTCATATTCTTGACCAAGCACCAAAAGCAAGAAAGTTCTTTGAGAAGTTGCCAGATGGCACTTGGAACTTAAAG
117722_4941358_DYRK1A_7_1_2     GTGGAGTTCTGGGTATTCCACCTGCTCATATTCTTGACCAAGCACCAAAAGCAAGAAAGTTCTTTGAGAAGTTGCCAGATGGCACTTGGAACTTAAAGAAGACCAAGATGGAAAACGG
117722_4941359_DYRK1A_8_1_1     GAGTACAAACCACCAGGAACCCGTAAACTTCATAACATTCTTGGAGTGGAAACAGGAGGACCTGGTGGGCGACGTGCTGGGGAGTCAGGTCATACGGTCGCTGACTACTTGAAGTTCAAA
117722_4941359_DYRK1A_8_1_2     CGGTCGCTGACTACTTGAAGTTCAAAGACCTCATTTTAAGGATGCTTGATTATGACCCCAAAACTCGAATTCAACCTTATTATGCTCTGCAGCACAGTTTCTTCAAGAAAACAGCTGATG
117722_4941359_DYRK1A_8_1_3     CAGTTTCTTCAAGAAAACAGCTGATGAAGGTACAAATACAAGTAATAGTGTATCTACAAGCCCCGCCATGGAGCAGTCTCAGTCTTCGGGCACCACCTCCAGTACATCGTCAAGCTCAG
```

Table S2b. HSD2 probe design sequences

SequenceName	Sequence
176360_8716475_GNRHR2_2_1_1	CTGTTCCTGTTCCACACGGTCCACTGAGCTGGCCCAGTCCCTTTCACTCAGTGTGTCACCAAAGGCAGCTTCAAGGCTCAATGGCAAGAGACCACCTATAACCTCTTCACCTTCTGCTGC
176360_8716475_GNRHR2_2_1_2	CCACCTATAACCTCTTCACCTTCTGCTGCCTCTTTCTGCTGCCACTGACTGCCATGGCCATCTGCTATAGCCGCATTGTCCTCAGTGTGTCCAGGCCCCAGACAAGGAAGGGGAGCCATG
176360_8716476_RBM8_1_1_1	ATGGCGGACGTGCTAGATCTTCACGAGGCTGGGGGCGAAGATTTCGCCATGGATGAGGATGGGGACGAGAGCATTCACAAACTGAAAGAAAAAGCGAAGAAACGGAAGGGTCGCGGCTTT
176360_8716476_RBM8_1_1_2	ACGGAAGGGTCGCGGCTTTGGCTCCGAAGAGGGGTCCCGAGCGCGGATGCGTGAGGATTATGACAGCGTGGAGCAGGATGGCGATGAACCCGGACCACAACGCTCTGTTGAAGGCTGGAT
176360_8716476_RBM8_1_1_3	GCTCTGTTGAAGGCTGGATTCTCTTTGTAACTGGAGTCCATGAGGAAGCCACCGAAGAAGACATACACGACAAATTCGCAGAATATGGGGAAATTAAAAACATTCATCTCAACCTCGACA
176360_8716476_RBM8_1_1_4	ATTCATCTCAACCTCGACAGGCGAACAGGATATCTGAAGGGGTATACTCTAGTTGAATATGAAACATACAAGGAAGCCCAGGCTGCTATGGAGGGACTCAATGGCCAGGATTTGATGGGA
176360_8716476_RBM8_1_1_5	TGGCCAGGATTTGATGGGACAGCCCATCAGCGTTGACTGGTGTTTTGTTCGGGGTCCACCAAAAGGCAAGAGGAGAGGTGGCCGAAGACGCAGCAGAAGTCCAGACCGGAGACGTCGCTG
176360_8716477_FCGR1A_1_1_1	TTCTAATTTGGCTCTGGAGCCACCAGCAGAACCTCTTCAATATCTTGCATGTTACAGATTTCACTGCTCCCACCAGCTTGGAGACAACATGTGGTTCTTGACAACTCTGCTCCTTTGGG
176360_8716478_FCGR1A_4_1_1	GCTGGCTACTACTGCAGGTCTCCAGCAGAGTCTTCACGGAAGGAGAACCTCTGGCCTTGAGGTGTCATGCGTGGAAGGATAAGCTGGTGTACAATGTGCTTTACTATCGAAATGGCAAAG
176360_8716478_FCGR1A_4_1_2	GGCAAAGCCTTTAAGTTTTTCCACTGGAATTCTAACCTCACCATTCTGAAAACCAACATAAGTCACAATGGCACCTACCATTGCTCAGGCATGGGAAAGCATCGCTACACATCAGCAGGA
176360_8716478_FCGR1A_4_1_3	AGCAGGAATATCTGTCACTGTGAAAGAGCTATTTCCAGCTCCAGTGCTGAATGCATCTGTGACATCCCCACTCCTGGAGGGGAATCTGGTCACCCTGAGCTGTGAAACAAAGTTGCTCTT
176360_8716478_FCGR1A_4_1_4	TGCTCTTGCAGAGGCCTGGTTTGCAGCTTTACTTCTCCTTCTACATGGGCAGCAAGACCCTGCGAGGCAGGAACACATCCTCTGAATACCAAATACTAACTGCTAGAAGAGAAGACTCTG
176360_8716478_FCGR1A_4_1_5	GACTCTGGGTTATACTGGTGCGAGGCTGCCACAGAGGATGGAAATGTCCTTAAGCGCAGCCCTGAGTTGGAGCTTCAAGTGCTTGGCCTCCAGTTACCAACTCCTGTCTGGTTTCATGTC
176360_8716478_FCGR1A_4_1_6	TCATGTCCTTTTCTATCTGGCAGTGGGAATAATGTTTTTAGTGAACACTGTTCTCTGGGTGACAATACGTAAAGAACTGAAAAGAAAGAAAAAGTGGGATTTAGAAATCTCTTTGGATTC
176360_8716479_HIST2H2BF_1_1_1	GGTAGTTCGGATTACTTCTTTAAGTCTCTTTTCTCTTTTTTCGCGCAAAAATGCCGGATCCAGCGAAATCCGCTCCTGCTCCCAAGAAGGGCTCCAAAAAGGCTGTTACGAAAGTGCAGA
176360_8716479_HIST2H2BF_1_1_2	CTGTTACGAAAGTGCAGAAGAAGGACGGCAAGAAGCGCAAGCGCAGCCGCAAGGAGAGCTACTCCGTTTACGTGTACAAGGTGCTGAAGCAGGTCCACCCCGACACCGGCATCTCGTCCA
176360_8716479_HIST2H2BF_1_1_3	ACACCGGCATCTCGTCCAAGGCCATGGGCATCATGAACTCCTTCGTCAACGACATCTTCGAGCGCATCGCGGGAGAGGCGTCCCGCCTGGCGCACTACAACAAGCGCTCCACCATCACAT
176360_8716479_HIST2H2BF_1_1_4	AGCGCTCCACCATCACATCCCGCGAGATCCAGACGGCCGTGCGCCTGCTGCTGCCCGGCGAGCTGGCCAAGCACGCCGTGTCCGAGGGCACCAAGGCGGTCACCAAGTACACCAGCTCGA
176360_8716480_HIST2H3D_1_1_1	ATGGCCCGTACTAAGCAGACTGCCCGCAAGTCGACCGGCGGCAAGGCCCCGAGGAAGCAGCTGGCTACCAAAGCGGCCCGCAAGAGCGCGCCGGCCACGGGCGGGGTGAAGAAGCCGCAC
176360_8716480_HIST2H3D_1_1_2	CGGGCGGGGTGAAGAAGCCGCACCGCTACCGGCCCGGCACCGTGGCTCTGCGGGAGATCCGGCGCTACCAGAAGTCTACGGAGCTGCTGATCCGCAAGCTGCCCTTCCAGCGGCTGGTAC
176360_8716480_HIST2H3D_1_1_3	GCTGCCCTTCCAGCGGCTGGTACGCGAGATCGCGCAGGACTTTAAGACGGACCTGCGCTTCCAGAGCTCGGCCGTGATGGCGCTGCAGGAGGCCAGCGAGGCCTACCTGGTGGGGCTGTT
176360_8716480_HIST2H3D_1_1_4	GAGGCCTACCTGGTGGGGCTGTTCGAAGACACGAACCTGTGCGCCATCCATGCCAAGCGCGTGACCATCATGCCCAAGGACATCCAGTTGGCCCGCCGCATCCGCGGGGAGCGGGCCTA
176360_8716481_LINC00152_1_1_1	GTGCGCCTTTTTTTTTTTTTCCTTCTTAGTCGTGTGTACATCATTGGGAATGGAGGGAAATAAATGACTGGATGGTCGCTGCTTTTTAAGTTTCAAATTGACATTCCAGACAAGCGGTGC
176360_8716481_LINC00152_1_1_2	ATGGTCGCTGCTTTTTAAGTTTCAAATTGACATTCCAGACAAGCGGTGCCTGAGCCCGTGCCTGTCTTCAGATCTTCACAGCACAGTTCCTGGGAAGGTGGAGCCACCAGCCTCTCCTTG
176360_8716482_LINC00152_2_1_1	AATAACTGGGAGATGAAACAGGAAGCTCTATGACACACTTGATCGAATATGACAGACACCGAAAATCACGACTCAGCCCCCTCCAGCACCTCTACCTGTTGCCCGCCGATCACAGCCGGA
176360_8716482_LINC00152_2_1_2	TACCTGTTGCCCGCCGATCACAGCCGGAATGCAGCTGAAAGATTCCCTGGGGCCTGGTTCCAACCGCCCACTGTGGACTCTGAGGCCTCTGCATTTGCGGGTGGTCTGCCTGTGATATTT
176360_8716482_LINC00152_2_1_3	ATTTGCGGGTGGTCTGCCTGTGATATTTTGGTCATGGGCTGGTCTGGTCGGTTTCCCATTTGTCTGGCCAGTCTCTATGTGTCTTAATCCCTTGTCCTTCATTAAAAGCAAAACTAAAGA
176360_8716483_ANAPC1_1_1_1	ATGTCGAACTTCTATGAAGAAAGGACAACGATGATTGCAGCAAGGGATTTGCAGGAATTTGTTCCTTTTGGTCGAGACCACTGCAAGCACCACCCTAATGCTTTGAACCTTCAACTTCGC
176360_8716483_ANAPC1_1_1_3	TAAGTGAAATTGGAGAAGATGTGGACTATGATGAGGAACTCTATGTTGCTGGAAATATGGTGATATGGAGCAAAGGAAGTAAAAGCCAGGCATTGGCAGTTTATAAAGCATTTACAGTTG
176360_8716483_ANAPC1_1_1_5	TAACATGCATAGCATAGAAGGAAAGGATTACATAGCTTCATTACCATTTCAGGTTGCAAATGTTTGGCCCACTAAATATGGATTGCTGTTTGAACGAAGCGCTTCTTCACATGAAGTACC
176360_8716483_ANAPC1_1_1_7	GTTGTAGATCATGCAATGAAAATTGTTTTCCTCAATACTGACCCCTCTATTGTAATGACTTATGATGCTGTTCAAAATGTGCATTCTGTGTGGACTCTCCGGAGAGTCAAATCAGAGGAA
176360_8716483_ANAPC1_1_1_9	CTTTCCAGAATTACTCCTCCATTCACAGCCAGAGTCGCTCAACCTCATCACCCAGTCTACATTCTCGCTCACCTTCTATTTCCAACATGGCAGCTCTAAGTCGTGCTCATTCTCCTGCGT
176360_8716483_ANAPC1_1_1_11	ACCAGAAACGGAGCCAATTGTTCCTGAACTGTGTATTGACCATTTGTGGACAGAAACGATTACTAATATAAGAGAGAAAAATTCACAAGCCTCAAAAGTGTTTATTACATCTGACCTATG
176360_8716483_ANAPC1_1_1_13	AAGGATGCAGCACCAGTGGAGAAAATAGACACCATGCTGGTCTTGGAAGGCAGTGGAAACCTGGTGCTATACACAGGAGTGGTTCGGGTGGAAAGGTTTTTATTCCTGGACTGCCAGCT
176360_8716483_ANAPC1_1_1_15	CAGTTCCAGAACTGAGGGATTCTTCAAAACTTCATGATTCTCTCTATAATGAGGATTGTACTTTCCAACAGCTTGGAACTTACATTCATTCTATCAGAGATCCTGTCCATAACAGAGTCA
176360_8716483_ANAPC1_1_1_17	TCAGATGCTTGTCAAGTGGTACAATGTCCACAGTGCTCCAGGAGGACCCAGTTATCACTCAGAGTGGAATTTATTTGTGACTTGTCTCATGAACATGATGGGTTATAACACAGACCGCTT
176360_8716483_ANAPC1_1_1_19	TACCACCAGAATGTTGAGTCTCATCTTTTGAACAGATCTTTATGTCTGAGTCCTTCAGAAGCTTCACAGATGAAGGATGAGGATTTTTCACAGAATCTCAGTCTGGATTCTTCTACACTT
176360_8716483_ANAPC1_1_1_21	CAAGGGACTTAAAATTGGGGCCTTATGTAGATCATTACTATAGAGACTACCCAACGCTTGTCAGAACTACTGGACAAGTGTGCACAATTGATCCAGGTCAAACAGGATTTATGCATCATC
176360_8716483_ANAPC1_1_1_23	AGTCTTGAGTATTGCACTGTACATACTTGGTGATGAGAGCTTGGTTTCTGATGAATCCTCACAGTATTTAACCAGAATAACTATAGCCCCCCAGAAGTTGCAAGTAGAACAAGAGGAAAA
176360_8716483_ANAPC1_1_1_25	CTTCCCATCAGAGATGCAATTTATCACTGTCGTGAGCAGCCTGCCTCAGACTGGCCAGAAGCTGTCTGTCTCTTGATTGGACGTCAGGATCTTTCCAAGCAGGCCTGCGAAGGAAACTTA
176360_8716483_ANAPC1_1_1_27	TAAGGGTGCAGGATGTGCGAAGGCTTCTTCAGAGTGCGCATCCTGTCCGTGTCAACGTAGTGCAGTACCCAGAGCTCAGTGACCACGAGTTCATCGAGGAAAAGGAAAACAGATTGCTCC
176360_8716483_ANAPC1_1_1_29	CCCTCCTCGGAACACAACAGTAGACCTTAATAGTGGAAACATCGATGTGCCTCCCAACATGACAAGCTGGGCCAGCTTTCATAATGGTGTGGCTGCTGGCCTGAAGATAGCTCCTGCCTC
176360_8716483_ANAPC1_1_1_31	CTCAATATCCATGACTACTTGACCAAGGGCCATGAAATGACAAGCATTGGACTGCTACTTGGTGTTTCTGCTGCAAAACTAGGCACCATGGATATGTCTATTACTCGGCTTCTTAGCATT
176360_8716483_ANAPC1_1_1_33	CAGAAGTCCTGTTGGCTGAGATAGGACGGCCTCCTGGTCCTGAAATGGAATACTGCACTGACAGAGAGTCATACTCCTTAGCTGCTGGCTTGGCCCTGGGCATGGTCTGCTTGGGGCATG
176360_8716483_ANAPC1_1_1_35	TTATCAAATCAAAGAAGGAGATACCATAAATGTGGATGTGACTTGTCCAGGTGCTACTCTAGCTTTGGCTATGATCTACTTAAAAACCAATAACAGATCTATTGCAGATTGGCTCCGAGC
176360_8716483_ANAPC1_1_1_38	CTGCTTGTCTCTGGGTTTTCGATTTGCTGGCTCAGAAAACTTATCAGCATTTAACTGTTTGCATAAATTTGCCAAAGATTTTATGACTTATTTGTCCGCACCTAATGCTTCTGTTACAGG
176360_8716483_ANAPC1_1_1_41	CCTTTATCCGCACTTCCCAGCTCACAGCACTGACAACCGGTATCATCTCCAGGCTCTCCGGCACCTCTATGTGCTGGCCGCGGAGCCCAGGCTTCTAGTGCCTGTGGATGTGGACACAAA
176360_8716483_ANAPC1_1_1_44	CTCCTACAAAGAAGATCCAATGGGATGGCAAAGTTTGTTGGCTCAGACTGTTGCTAACAGGAACTCTGAAGCCCGGGCTTTCAAGCCAGAAACAATCTCAGCATTCACTTCTGATCCAGC
176360_8716483_ANAPC1_1_1_47	CCATCAGGAGCGGCTGCAGAACCACCCTAAGCGGGGGCTCTTTATGAACTCGGAATTCCTCCCTGTTGTGAAGTGCACCATTGATAATACCCTGGACCAGTGGCTACAAGTCGGGGGTGA
176360_8716484_OCLN_1_1_1	ATGTCATCCAGGCCTCTTGAAAGTCCACCTCCTTACAGGCCTGATGAATTCAAACCGAATCATTATGCACCAAGCAATGACATATATGGTGGAGAGATGCATGTTCGACCAATGCTCTCT
176360_8716485_OCLN_2_1_3	TCTGGATCTCTATATGGTTCACAAATATATGCCCTCTGCAACCAATTTTATACACCTGCAGCTACTGGACTCTACGTGGATCAGTATTTGTATCACTACTGTGTTGTGGATCCCCAGGAG
176360_8716486_OCLN_3_1_2	AACACATTTATGATGAGCAGCCCCCCAATGTCGAGGAGTGGGTTAAAAATGTGTCTGCAGGCACACAGGACGTGCCTTCACCCCCATCTGACTATGTGGAAAGAGTTGACAGTCCCATGG
176360_8716486_OCLN_3_1_3	CATACTCTTCCAATGGCAAAGTGAATGACAAGCGGTTTTATCCAGAGTCTTCCTATAAATCCACGCCGGTTCCTGAAGTGGTTCAGGAGCTTCCATTAACTTCGCCTGTGGATGACTTCA
176360_8716486_OCLN_3_1_4	GGCAGCCTCGTTACAGCAGCGGTGGTAACTTTGAGACACCTTCAAAAAGAGCACCTGCAAAGGGAAGAGCAGGAAGGTCAAAGAGAACAGAGCAAGATCACTATGAGACAGACTACACAA
176360_8716486_OCLN_3_1_5	CTGGCGGCGAGTCCTGTGATGAGCTGGAGGAGGACTGGATCAGGGAATATCCACCTATCACTTCAGATCAACAAAGACAACTGTACAAGAGGAATTTTGACACTGGCCTACAGGAATACA
176360_8716486_OCLN_3_1_6	AGAGCTTACAATCAGAACTTGATGAGATCAATAAAGAACTCTCCCGTTTGGATAAAGAATTGGATGACTATAGAGAAGAAAGTGAAGAGTACATGGCTGCTGCTGATGAATACAATAGAC
176360_8716486_OCLN_3_1_7	TGAAGCAAGTGAAGGGATCTGCAGATTACAAAAGTAAGAAGAATCATTGCAAGCAGTTAAAGAGCAAATTGTCACACATCAAGAAGATGGTTGGAGACTATGATAGACAGAAAACATAG
176360_8716487_GTF2H2_1_1_1	TTGTTGGAATACTTTGTAGAGGAATATTTTGATCAAAATCCTATTAGTCAGATTGGAATAATTGTAACTAAGAGTAAAAGAGCTGAAAAATTGACTGAACTTTCAGGAAACCCAAGAAAA
176360_8716487_GTF2H2_1_1_2	AAAACATATAACGTCTTTGAAGAAAGCTGTGGATATGACCTGCCATGGAGAGCCATCTCTTTATAATTCCCTAAGCATGGCTATGCAGACTCTAAAACACATGCCTGGACATACAAGTCG
176360_8716487_GTF2H2_1_1_3	GTCGAGAAGTACTAATCATCTTTAGCAGCCTTACAACTTGCGATCCATCTAATATTTATGATCTAATCAAGACCCTAAAGGCAGCTAAAATTAGAGTATCTGTTACTGGATTGTCTGCAG
176360_8716487_GTF2H2_1_1_5	TAGCTCAAGTTCTGAATGCTCACTTATTCGTATGGGATTTCCTCAGCACACCATTGCTTCTTTATCTGACCAGGATGCAAAACCCTCTTTCAGCATGGCGCATTTGGATGGCAATACTGA
176360_8716487_GTF2H2_1_1_7	GCACGGTCTTACCATCATTTGTTTCCTTTGGATGCTTTTCAAGAAATTCCCCTAGAAGAATATAATGGAGAAAGATTTTGTTATGGATGTCAGGGGGAATTGAAAGACCAACATGTTTAT
176360_8716487_GTF2H2_1_1_8	TTATGTTTGTGCTGTGTGCCAAAATGTTTTCTGTGTGGACTGTGATGTTTTTGTTCATGATTCTCTACACTGTTGCCCTGGCTGTATTCATAAGATTCCAGCTCCTTCAGGTGTTTGA
176360_8716488_SERF1_1_1_1	GTGGAAATCAACGAGAACTTGCCCGCCAGAAAAACATGAAGAAAACCCAGGAAATTAGCAAGGGAAAGAGGAAAGAGGATAGCTTGACTGCCTCTCAGAGAAAGCAGAGGGACTCTGAGA
176360_8716488_SERF1_1_1_2	GGAAAGAGGAAAGAGGATAGCTTGACTGCCTCTCAGAGAAAGCAGAGGGACTCTGAGATCATGCAAGAAAAGCAGAAGGCAGCTAATGAGAAGAAGTCTATGCAGACAAGAGAAAAGTGA
176360_8716489_SMN1_1_1_1	ATGGCGATGAGCAGCGGCGGCAGTGGTGGCGGCGTCCCGGAGCAGGAGGATTCCGTGCTGTTCCGGCGCGGCACAGGCCAGAGCGATGATTCTGACATTTGGGATGATACAGCACTGATA
176360_8716489_SMN1_1_1_2	GGATGATACAGCACTGATAAAAGCATATGATAAAGCTGTGGCTTCATTTAAGCATGCTCTAAAGAATGGTGACATTTGTGAAACTTCGGGTAAACCAAAAACCACACCTAAAAGAAAACC
176360_8716489_SMN1_1_1_3	CCACACCTAAAAGAAAACCTGCTAAGAAGAATAAAAGCCAAAAGAAGAATACTGCAGCTTCCTTACAACAGTGGAAAGTTGGGGACAAATGTTCTGCCATTTGGTCAGAAGACGGTTGCA
176360_8716489_SMN1_1_1_4	TGGTCAGAAGACGGTTGCATTTACCCAGCTACCATTGCTTCAATTGATTTTAAGAGAGAAACCTGTGTTGTGGTTTACACTGGATATGGAAATAGAGAGGAGCAAAATCTGTCCGATCTA
176360_8716489_SMN1_1_1_5	GCAAAATCTGTCCGATCTACTTTCCCCAATCTGTGAAGTAGCTAATAATATAGAACAAAATGCTCAAGAGAATGAAAATGAAAGCCAAGTTTCAACAGATGAAAGTGAGAACTCCAGGTC
176360_8716489_SMN1_1_1_6	AAAGTGAGAACTCCAGGTCTCCTGGAAATAAATCAGATAACATCAAGCCCAAATCTGCTCCATGGAACTCTTTTCTCCCTCCACCACCCCCCATGCCAGGGCCAAGACTGGGACCAGGAA
176360_8716490_NAIP_1_1_1	CCAAAGTAGCAATCAGTGGAGGTTTCCAGAAACTTGAGAACCTAAAGCTTTCAATCAATCACAAGATTACAGAGGAAGGATACAGAAATTTCTTTCAAGCACTGGACAACATGCCAAACT
176360_8716490_NAIP_1_1_2	CTTGCAGGAGTTGGACATCTCCAGGCATTTCACAGAGTGTATCAAAGCTCAGGCCACAACAGTCAAGTCTTTGAGTCAATGTGTGTTACGACTACCAAGGCTCATTAGACTGAACATGTT
176360_8716490_NAIP_1_1_3	TTAAGTTGGCTCTTGGATGCAGATGATATTGCATTGCTTAATGTCATGAAAGAAAGACATCCTCAATCTAAGTACTTAACTATTCTCCAGAAATGGATACTGCCGTTCTCTCCAATCATT
176360_8716490_NAIP_1_1_4	TTCAGAAATAAATGTGAATTTCTTCGGAGTAAGAAATCCTCAGAGGAAATTACCCAGTATATTCAAAGCTACAAGGGATTTGTTGACATAACGGGAGAACATTTTGTGAATTCCTGGGTC
176360_8716490_NAIP_1_1_5	TCCAGAGAGAATTACCTATGGCATCAGCTTATTGCAATGACAGCATCTTTGCTTACGAAGAACTACGGCTGGACTCTTTTAAGGACTGGCCCCGGGAATCAGCTGTGGGAGTTGCAGCAC
176360_8716490_NAIP_1_1_6	ACTGGCCAAAGCAGGTCTTTTCTACACAGGTATAAAGGACATCGTCCAGTGCTTTTCCTGTGGAGGGTGTTTAGAGAAATGGCAGGAAGGTGATGACCCATTAGACGATCACACCAGATG
176360_8716490_NAIP_1_1_7	TGTTTTCCCAATTGTCCATTTCTCCAAAATATGAAGTCCTCTGCGGAAGTGACTCCAGACCTTCAGAGCCGTGGTGAACTTTGTGAATTACTGGAAACCACAAGTGAAAGCAATCTTGAA
176360_8716490_NAIP_1_1_8	AAGATTCAATAGCAGTTGGTCCTATAGTGCCAGAAATGGCACAGGGTGAAGCCCAGTGGTTTCAAGAGGCAAAGAATCTGAATGAGCAGCTGAGAGCAGCTTATACCAGCGCCAGTTTCC
176360_8716490_NAIP_1_1_9	CCGCCACATGTCTTTGCTTGATATCTCTTCCGATCTGGCCACGGACCACTTGCTGGGCTGTGATCTGTCTATTGCTTCAAAACACATCAGCAAACCTGTGCAAGAACCTCTGGTGCTGCC
176360_8716490_NAIP_1_1_13	CGGACCTGCCTATTGATTGCTGTCCGTACAAACAGGGCCAGGGACATCCGCCGATACCTAGAGACCATTCTAGAGATCAAAGCATTTCCCTTTTATAATACTGTCTGTATATTACGGAAG
176360_8716490_NAIP_1_1_17	TTACAGCCCAGAGACTAAGACCATTCTACCGGTTTTTAAGTCCTGCCTTCCAAGAATTTCTTGCGGGGATGAGGCTGATTGAACTCCTGGATTCAGATAGGCAGGAACATCAAGATTTGG
176360_8716490_NAIP_1_1_21	TCAAGGGAGAACACTGACTTTGGGTGCGCTTAACTTACAGTACTTTTTCGACCACCCAGAAAGCTTGTCATTGTTGAGGAGCATCCACTTCCCAATACGAGGAAATAAGACATCACCCAG
176360_8716490_NAIP_1_1_25	CCAGCTCTTGAGCTGTCTAAGGCCTCTGTCACCAAGTGCTCCATAAGCAAGTTGGAACTCAGCGCAGCCGAACAGGAACTGCTTCTCACCCTGCCTTCCCTGGAATCTCTTGAAGTCTCA
176360_8716490_NAIP_1_1_26	CAGGGACAATCCAGTCACAAGACCAAATCTTTCCTAATCTGGATAAGTTCCTGTGCCTGAAAGAACTGTCTGTGGATCTGGAGGGCAATATAAATGTTTTTCAGTCATTCCTGAAGAAT
176360_8716490_NAIP_1_1_27	ATTTCCAAACTTCCACCATATGGAGAAATTATTGATCCAAATTTCAGCTGAGTATGATCCTTCCAAACTAGTAAAATTAATTCAAAATTCTCCAAACCTTCATGTTTTCCATCTGAAGTG
176360_8716490_NAIP_1_1_28	TGTAACTTCTTTTCGGATTTTGGGTCTCTCATGACTATGCTTGTTTCCTGTAAGAAACTCACAGAAATTAAGTTTTCGGATTCATTTTTTCAAGCCGTCCCATTTGTTGCCAGTTTGCCA
176360_8716490_NAIP_1_1_29	CAAATTTTATTTCTCTGAAGATATTAAATCTTGAAGGCCAGCAATTTCCTGATGAGGAAACATCAGAAAAAATTTGCCTACATTTTAGGTTCTCTTAGTAACCTGGAAGAATTGATCCTTC
176360_8716490_NAIP_1_1_30	TCCTACTGGGGATGGAATTTATCGAGTGGCCAAACTGATCATCCAGCAGTGTCAGCAGCTTCATTGTCTCCGAGTCCTCTCATTTTTCAAGACTTTGAATGATGACAGCGTGGTGGAAAT
176360_8716491_DUSP22_1_1_1	ATGGGGAATGGGATGAACAAGATCCTGCCCGGCCTGTACATCGGCAACTTCAAAGATGCCAGAGACGCGGAACAATTGAGCAAGAACAAGGTGACACATATTCTGTCTGTCCACGATAGT
176360_8716491_DUSP22_1_1_2	TTCTGTCTGTCCACGATAGTGCCAGGCCTATGTTGGAGGGAGTTAAATACCTGTGCATCCCAGCAGCGGATTCACCATCTCAAAACCTGACAAGACATTTCAAAGAAAGTATTAAATTCA
176360_8716491_DUSP22_1_1_3	CAAAGAAAGTATTAAATTCATTCACGAGTGCCGGCTCCGCGGTGAGAGCTGCCTTGTACACTGCCTGGCCGGGGTCTCCAGGAGCGTGACACTGGTGATCGCATACATCATGACCGTCAC
176360_8716491_DUSP22_1_1_4	GCATACATCATGACCGTCACTGACTTTGGCTGGGAGGATGCCCTGCACACCGTGCGTGCTGGGAGATCCTGTGCCAACCCCAACGTGGGCTTCCAGAGACAGCTCCAGGAGTTTGAGAAG
176360_8716491_DUSP22_1_1_5	AGCTCCAGGAGTTTGAGAAGCATGAGGTCCATCAGTATCGGCAGTGGCTGAAGGAAGAATATGGAGAGAGCCCTTTGCAGGATGCAGAAGAAGCCAAAAACATTCTGGGTAAATATAAGG
176360_8716491_DUSP22_1_1_6	CATTCTGGGTAAATATAAGGAGCAAGGGCGCACAGAGCCCCAGCCCGGCGCCAGGCGGTGGAGCAGTTTTCCGGCACTGGCTCCGCTGACCTACGATAATTATACGACGGAGACCTAA
176360_8716492_ZNF322_1_1_1	ATGTACACTTCAGAAGAGAAATGTAATCAGAGAACTCAAAAAAGGAAAATATATAATGTATGCCCTCGGAAGGGTAAAAAGATTTTTATTCATATGCATGAGATTATTCAGATAGATGGT
176360_8716492_ZNF322_1_1_3	TGATATGTGTGAGAAAACCTTTGTCCAAAGCTCAGATCTTACTTCACACCAGAGGATCCACAATTACGAGAAACCTTATAAATGTAGCAAATGTGAGAAGAGCTTTTGGCATCACTTAGC
176360_8716492_ZNF322_1_1_5	AGCGAAGCCATACTGGCGAGAAACCATATCTATGTAGTGAGTGTGACAAATGCTTCAGTAGAAGTACAAACCTCATAAGGCATCGAAGAACTCACACAGGTGAGAAACCATTTAAGTGTC
176360_8716492_ZNF322_1_1_7	AGTTACCGACACCGTTCAGCCTTCATTGTACATAAAAGAGTTCATACTGGGGAGAAGCCCTATAAGTGTGGTGCCTGTGAAAAATGCTTTGGCCAGAAATCAGACCTTATCGTGCACCAG
176360_8716492_ZNF322_1_1_9	CACTTTTAAATGCCTTGAATATGAAAAAAGCTTTAACTGTAGCTCAGATCTTATTGTACATCAGAGAATTCACATGGAAGAGAAACCACATCAGTGGTCTGCGTGTGAGAGTGGCTTCCT
176360_8716492_ZNF322_1_1_11	CAGCCCTTCTTCAACATCAGACAGTACACATTGGTGAAAAACCGTTTGTCTGTAATGTGAGTGAAAAAGGTCTTGAGCTTAGCCCTCCCCATGCGTCAGAAGCCTCACAGATGTCTTGA
176360_8716493_TCAF2_1_1_1	ATGGCGACCATTGCTGCTGCTGCGTTTGAGGCCCTCATGGATGGAGTGACATGCTGGGATGTCCCCAGAGGCCCCATCCCCAGTGAACTCCTTCTTATTGGAGAAGCCGCCTTCCCCGTG
176360_8716493_TCAF2_1_1_2	AGAAGCCGCCTTCCCCGTGATGGTGAATGACAAGGGCCAGGTGCTCATTGCTGCCTCCTCCTACGGCCGAGGCCGCCTCGTGGTTGTGTCCCATGAGGGCTACCTGTCGCATACTGGCTT
176360_8716493_TCAF2_1_1_3	ACCTGTCGCATACTGGCTTGGCTCCATTTCTCCTCAATGCAGTGAGCTGGCTCTGTCCCTGTCCTGGGGCTCCCGTGGGAGTGCATCCATCCCTGGCACCTCTAGTAAACATCCTACAGG
176360_8716493_TCAF2_1_1_4	CTAGTAAACATCCTACAGGATGCTGGGCTTGAGGCACAGGTCAAGCCAGAACCAGGAGAGCCCCTAGGGGTTTACTGTATCAATGCCTACAATGACACCTTGACTGCAACGCTGATCCAG
176360_8716493_TCAF2_1_1_5	GACTGCAACGCTGATCCAGTTTGTGAAACATGGAGGGGGCTTGTTAATCGGGGGCCAGGCCTGGTACTGGGCCAGCCAGCACGGCCCTGACAAGGTGCTCTCCAGGTTCCCTGGGAACAA
176360_8716493_TCAF2_1_1_6	CCAGGTTCCCTGGGAACAAGGTGACAAGTGTAGCTGGAGTGTACTTCACTGACACCTATGGGGACAGAGACCGGTTCAAGGTCTCTAAGAAGGTGCCCAAGATCCCACTCCATGTCAG
176360_8716494_TCAF2_2_1_1	GTATGGGGAGGATGTCAGGCAGGACCAGCAGCAGCTCCTGGAGGGGATCTCAGAGCTGGACATCAGGACAGGGGGAGTCCCCTCACAGCTGCTTGTACATGGAGCCCTGGCCTTCCCTCT
176360_8716494_TCAF2_2_1_2	GCCTTCCCTCTGGGGCTGGATGCCTCACTCAACTGCTTCCTGGCGGCTGCTCACTATGGCCGGGGCCGGGTGGTCCTGGCTGCCCACGAGTGCCTGCTCTGTGCTCCCAAGATGGGGCCC
176360_8716494_TCAF2_2_1_3	AGATGGGGCCCTTCTTGCTCAATGCGGTGCGCTGGCTGGCCAGAGGCCAGACAGGCAAAGTTGGGGTGAACACAAATCTAAAAGATCTGTGTCCTCTCCTATCGGAGCATGGCCTGCAAT
176360_8716494_TCAF2_2_1_4	TGGCCTGCAATGCAGCCTGGAGCCCCATCTGAACAGCGACTTGTGTGTCTACTGCTGCAAGGCGTACAGTGACAAGGAGGCTAAGCAGCTGCAGGAGTTTGTGGCTGAGGGTGGGGGCT
176360_8716494_TCAF2_2_1_5	GGTGGGGGCTGCTGATTGGGGGCCAGGCCTGGTGGTGGGCCTCCCAGAACCCTGGCCACTGCCCCTTGGCTGGCTTCCCTGGTAACATCATCCTCAACTGCTTTGGCCTCAGCATCCTG
176360_8716494_TCAF2_2_1_6	TCAGCATCCTGCCTCAGACTCTCAAAGCAGGCTGCTTCCCCGTTCCCACCCCTGAGATGAGAAGCTACCACTTCCGCAAGGCGCTCTCTCAATTCCAGGCTATACTGAACCACGAGAATG
176360_8716494_TCAF2_2_1_7	CCACGAGAATGGGAACTTGGAAAAGAGCTGTCTGGCAAAGTTGAGAGTTGATGGTGCAGCCTTCCTACAGATTCCTGCGGAGGGGGTCCCTGCTTACATATCCCTGCACAGGCTCCTGAG
176360_8716494_TCAF2_2_1_8	AGGCTCCTGAGGAAGATGCTACGAGGGTCTGGCCTCCCAGCTGTGAGCCGGGAAAATCCAGTTGCCAGTGACTCCTATGAGGCTGCGGTGCTCTCCCTGGCCACTGGGCTGGCTCACTCT
176360_8716494_TCAF2_2_1_9	TGGCTCACTCTGGAACTGACTGCTCCCAGCTGGCCCAGGGGCTTGGCACCTGGACCTGCTCCTCCAGTTTGTACCCCTCAAAACACCCCATCACCGTGGAGATCAATGGAATCAACCCAG
176360_8716495_TCAF2_3_1_1	GCAACAATGATTGCTGGGTGAGTACCGGGCTCTACCTCCTGGAAGGACAAAATGCAGAAGTCTCACTGTCTGAAGCTGCTGCCTCTGCTGGCCTGAGGGTACAGATAGGCTGCCACACCG
176360_8716495_TCAF2_3_1_2	GGGTACAGATAGGCTGCCACACCGATGACCTTACCAAGGCCAGGAAGCTATCTCGAGCCCCCGTGGTGACTCACCAATGCTGGATGGACAGGACTGAGCGGTCAGTCTCCTGCCTCTGGG
176360_8716495_TCAF2_3_1_3	AGCGGTCAGTCTCCTGCCTCTGGGGCGGCCTCCTCTACGTCATCGTGCCCAAGGGCAGCCAACTAGGCCCTGTGCCTGTCACTATCAGGGGAGCTGTGCCTGCCCCATACTACAAGCTGG
176360_8716496_TCAF2_5_1_1	TAAGACATCGCTGGAGGAGTGGAAGAGGCAGATGCAGGAGAACCTGGCTCCCTGGGGAGAGCTGGCCACGGACAACATCATCCTGACAGTGCCAACCACAAACCTTCAGGCCCTGAAGGA
176360_8716496_TCAF2_5_1_2	CCCCGAGCCTGTGCTCCGCCTCTGGGATGAGATGATGCAGGCTGTGGCCAGGCTGGCGGCTGAGCCCTTCCCTTTCCGCCGTCCTGAGAGGATTGTGGCTGATGTGCAGATCTCAGCTG
176360_8716497_TCAF2_6_1_1	GCTGGATGCATTCAGGATACCCCATCATGTGCCACCTGGAGTCTGTGAAGGAGATCATCAATGAGATGGACATGAGGAGCAGGGGTGTGTGGGGCCCCATCCATGAGCTGGGCCACAACC
176360_8716497_TCAF2_6_1_2	GGGCCACAACCAACAGCGGCATGGATGGGAGTTCCCCCACACACTACTGAGGCCACCTGTAACCTTTGGTCAGTCTACGTGCATGAAACAGTCCTGGGGATCCCCAGGGCTCAGGCCCA
176360_8716497_TCAF2_6_1_3	GCTCAGGCCCACGAGGCTCTGAGCCCTCCAGAGCGAGAGAGGAGAATCAAGGCCCACCTGGGAAAGGGAGCCCCCCTGTGTGACTGGAATGTATGGACAGCCCTGGAAACATATCTACAG
176360_8716498_OR2A1_1_1_1	TGCAGGAGGTTCACCAGCATCTGGGGCACTGTGTTGCAGGCATAGGCGATGTTGACGACGGCCAGGTGTGAGAGGAAGAAGTACATGGGGGTGTGGAGTCTGGAGTCCAGTGAGATGAGC
176360_8716498_OR2A1_1_1_2	CAGGTGTGAGAGGAAGAAGTACATGGGGGTGTGGAGTCTGGAGTCCAGTGAGATGAGCCCCAGGATGGTCCCATTCCCCAGCAGGGTGAAGACATAGAACAGGGAGAAGAGCCCAAAGAG
176360_8716498_OR2A1_1_1_3	GGATGGTCCCATTCCCCAGCAGGGTGAAGACATAGAACAGGGAGAAGAGCCCAAAGAGGAGCATCTGAATCCTTGGGCCCAGGAGAAATCCCAGTAGGAGGAACTCTGTGACCATTGTCT
176360_8716499_OR2A1_2_1_1	TGGGGGAAAATCAGACAATGGTCACAGAGTTCCTCCTACTGGGATTTCTCCTGGGCCCAAGGATTCAGATGCTCCTCTTTGGGCTCTTCTCCCTGTTCTATATCTTCACCCTGCTGGGGA
176360_8716499_OR2A1_2_1_2	GGGAACGGGGCCATCCTGGGGCTCATCTCACTGGACTCCAGACTCCACACCCCCATGTACTTCTTCCTCTCACACCTGGCTGTCGTCGACATCGCCTACACCCGCAACACGGTGCCCCAG
176360_8716499_OR2A1_2_1_3	CCAGATGCTGGCGAACCTCCTGCATCCAGCCAAGCCCATCTCCTTTGCTGGCTGCATGACGCAGACCTTTCTCTGTTTGAGTTTTGGACACAGCGAATGTCTCCTGCTGGTGCTGATGTC
176360_8716499_OR2A1_2_1_4	TGTCCTACGATCGTTACGTGGCCATCTGCCACCCTCTCCGATACTCCGTCATCATGACCTGGAGAGTCTGCATCACCCTGGCCGTCACTTCCTGGACGTGTGGCTCCCTCCTGGCTCTGG
176360_8716499_OR2A1_2_1_5	CTGGCCCATGTGGTTCTCATCCTAAGACTGCCCTTCTCTGGGCCTCATGAAATCAACCACTTCTTCTGTGAAATCCTGTCTGTCCTCAGGCTGGCCTGTGCTGACACCTGGCTCAACCAG
176360_8716499_OR2A1_2_1_6	CCAGGTGGTCATCTTTGCAGCCTGCGTGTTCTTCCTGGTGGGGCCACCCAGCCTGGTGCTTGTCTCCTACTCGCACATCCTGGCGGCCATCCTGAGGATCCAGTCTGGGGAGGGCCGCAG
176360_8716499_OR2A1_2_1_7	GCAGAAAGGCCTTCTCCACCTGCTCCTCCCACCTCTGCGTGGTGGGACTCTTCTTTGGCAGTGCCATCATCATGTACATGGCCCCCAAGTCCCGCCATCCTGAGGAGCAGCAAAAGGTCT
176360_8716499_OR2A1_2_1_8	GTCTTTTTTCTATTTTACAGTTTTTTCAACCCAACACTTAACCCCCTGATTTACAGCCTGAGGAACGGAGAGGTCAAGGGTGCCCTGAGGAGAGCACTGGGCAAGGAAAGTCATTCCTAA
176360_8716500_ARHGEF5_1_1_1	ATGGAGGCTGAGGAGGCCCAGCGTGGAGCCTCTCCTCCCATCTCTGCCATAGAGGAATTCAGCATTATCCCTGAGGCTCCCATGAGGAGCAGCCAGGTCTCTGCCTTGGGGCTTGAAGCT
176360_8716500_ARHGEF5_1_1_5	TTGGGATCTGGACAGGCAGAAGAAGAAGAGGAAACCTCTTCAGATAACTCTGGTCAGACCAGATATTATTCTCCCTGCGAAGAGCATCCTGCAGAGACCAACCAGAATGAAGGCTCTGAA
176360_8716501_ARHGEF5_1_2_2	GTCCAGAACAGGGAGAAGAGAGGAAGAGGGAGCTGCAGGTGCCAGAAGAGAATAGGGCGGACTCTCAGGACGAAAAGAGTCAAACCTTTTTGGGAAAATCAGAGGAAGTAACTGGAAAGC

176360_8716501_ARHGEF5_1_2_6	TAGATCAAAACAGCCAGCAAGAGGAATCCAGGCTGAGGAAGGGAACAGTGTCCAGCCAAGGGACTGAGGTGGTCTTTGCCAGTGCATCTGTGACTCCTCCAAGGACACCAGATTCAGCTC
176360_8716501_ARHGEF5_1_2_10	CCATCTACAGCTCTGTGACCCCAAGAAGGGACAGAAGGAGTGGTAGGGACTACAGCACCGTTTCAGCATCCCCTACTGCCTTATCCACGCTGAAGCAGGACTCTCAAGAATCCATCTCAA
176360_8716501_ARHGEF5_1_2_14	ACCATGCTCAGGCCACCCCCAGGTGGAGATACAACAAGCCGCTACCCCCTACCCCTGATTTGCCGCAGCCCCACCTTCCTCCCATTTCTGCTCCTGGTAGCTCAAGGATCTACAGGCCTC
176360_8716501_ARHGEF5_1_2_19	TGGAGAAGGCGTCCAGCTGGCCCCACAGGCGGGACTCAGGGAGGCCACCAGGGGACAGCAGTGGACAGGCTGTGGCTCCTAGTGAGGGGGCCAACAAGCACAAGGGCTGGAGCCGGCAGG
176360_8716501_ARHGEF5_1_2_20	GCCTGCGCAGACCTTCCATCTTGCCTGAGGGCTCTTCAGATTCAAGAGGTCCAGCCGTGGAGAAACATCCGGGACCCTCAGACACTGTTGTTTTTCGGGAGAAAAAACCAAAGGAGGTGA
176360_8716501_ARHGEF5_1_2_21	TGGGAGGCTTTTCAAGACGCTGCTCCAAACTCATCAACTCCTCCCAGCTGCTTTACCAGGAGTATAGTGATGTTGTCCTGAATAAGGAGATCCAGAGCCAGCAGCGGCTGGAGAGCCTGT
176360_8716501_ARHGEF5_1_2_22	CCGAGACACCCGGGCCTAGCTCTCCGCGGCAGCCTCGGAAGGCCCTGGTCTCCTCCGAGTCGTACCTGCAGCGGCTCTCCATGGCCTCCAGCGGCTCCCTCTGGCAGGAAATCCCCGTGG
176360_8716501_ARHGEF5_1_2_23	TGCGCAACAGCACCGTGCTGCTCTCCATGACCCATGAAGACCAAAAGCTGCAAGAGGTCAAATTTGAGCTGATTGTGTCAGAGGCCTCCTACCTGCGCAGTCTAAACATAGCTGTGGATC
176360_8716501_ARHGEF5_1_2_24	ATTTCCAACTTTCAACTTCACTCCGGGCCACACTTTCCAACCAGGAGCACCAATGGCTCTTCTCTCGTTTACAGGATGTGCGAGACGTCAGCGCCACGTTCCTTTCAGACCTGGAAGAGA
176360_8716501_ARHGEF5_1_2_25	ACTTTGAGAACAATATCTTCTCCTTCCAAGTATGTGACGTAGTCCTGAACCACGCCCCAGACTTCCGCCGGGTCTACCTGCCTTATGTCACCAACCAGACCTATCAGGAACGCACCTTCC
176360_8716501_ARHGEF5_1_2_26	AGAGCCTGATGAATAGCAACAGCAATTTCCGGGAGGTCTTGGAGAAGCTGGAGAGCGACCCCGTCTGCCAGCGCCTTTCCCTCAAGTCCTTTCTGATTCTGCCCTTCCAACGCATCACCC
176360_8716501_ARHGEF5_1_2_27	GCCTCAAACTGCTGCTCCAGAACATTCTGAAGAGAACACAGCCTGGCTCCTCGGAGGAGGCAGAGGCCACGAAGGCACACCACGCCCTGGAGCAGCTGATCCGGGACTGCAATAACAATG
176360_8716501_ARHGEF5_1_2_28	TCCAGAGTATGCGACGGACAGAGGAACTAATCTACCTGAGCCAGAAGATTGAGTTTGAGTGCAAAATATTCCCGCTCATTTCTCAGTCACGCTGGCTGGTGAAAAGTGGGGAGCTGACAG
176360_8716501_ARHGEF5_1_2_29	CCTTGGAGTTCAGTGCTTCCCCAGGGCTACGAAGGAAGCTGAACACGCGTCCAGTCCACCTGCACCTCTTCAATGACTGTCTGCTGCTGTCTCGGCCCCGAGAGGGTAGCCGATTCCTGG
176360_8716501_ARHGEF5_1_2_30	TATTTGACCATGCTCCCTTCTCCTCCATTCGGGGGAAAAGTGTGAAATGAAGCTACATGGACCTCACAAAAACCTGTTCCGACTCTTTCTGCGGCAGAACACTCAGGGCGCCCAGGCCG
176360_8716501_ARHGEF5_1_2_31	AGTTCCTCTTCCGCACGGAGACTCAAAGTGAAAAGCTTCGGTGGATCTCAGCCTTGGCCATGCCAAGAGAGGAGTTGGACCTTCTGGAGTGTTACAACTCCCCCCAGGTACAGTGCCTTC
176360_8716501_ARHGEF5_1_2_32	GAGCCTACAAGCCCCGAGAGAATGATGAATTGGCACTGGAGAAAGCCGACGTGGTGATGGTGACTCAGCAGAGCAGTGACGGCTGGCTGGAGGGCGTGAGGCTCTCAGACGGGGAGCGAG
176360_8716501_ARHGEF5_1_2_33	GCTGGTTTCCTGTGCAGCAGGTGGAGTTCATTTCCAACCCAGAGGTCCGTGCACAGAACCTGAAGGAAGCTCATCGAGTCAAGACTGCCAAACTACAGCTGGTGGAACAGCAAGCCTAA
176360_8716502_LOC643650_1_1_1	AATCAGGTTTCAAAGCCACTTCAACTGCACTGGCCCCTGTGGGTCACTGCTGCACCGCCCTGACCCATGTGGGTCCCTGAGGAGCGACCTGCCGGGGCCACCTGGCTGGACGAAAAAGAC
176360_8716502_LOC643650_1_1_2	TGGCCCCTGTGGGTCACTGCTGCACCGCCCTGACCCATGTGGGTCCCTGAGGAGCGACCTGCCGGGGCCACCTGGCTGGACGAAAAAGACACACCTTGGACTTAAGCCGTGAGAAAAAAA
176360_8716503_LOC643650_2_1_1	AGATGAAAAGATGGGCCAGGAGTCCAGTTTCTGGAAGGCCAAGAATCGAAGTAGCAAGCTGCAGCCGTTTTCCAGACAAGCAGGATGTGGGGATGCAGAAGAATTCAGGACTGGAGGGGC
176360_8716503_LOC643650_2_1_2	CAGGACTGGAGGGGCAAACTCCGATGTGACTGAGGCCCCACTGCCAAATGGCAGCATGCTCAGATAGCACCCAGGAATTTGGGGAAAAAAACTGGTGCTCACAGCTGCCCAGTTAAGGCA
176360_8716503_LOC643650_2_1_3	TGCCCAGTTAAGGCACAAGTCTCCCGCCTGCTGCAGGGGTTGTGAGGTGTATCAGAACCTTGAAGAGATACTCTGGGAAGAAAGGCCTTCCCAGACTGCCTCTCCGCAGCGCTGTAAGAA
176360_8716503_LOC643650_2_1_4	GCAGCGCTGTAAGAATCAGATTTCTTCTATGGGGGACTTCCAGGTCTCTTCAGCATTTTCTCAGTCATCACTGTATCCCTGCCACTCAGTGAACCACAGCAACAACTCCCAGCTTCCAGC
176360_8716503_LOC643650_2_1_5	CTCCCAGCTTCCAGCCGAGAAGCTTGACCCTGAAGCTGGCCTCGATTCTCTACAGAAGTGAGGCCACAGAGCACAGGTGAAGGGCCACTTACCTGGTATCATCTCTGGTCTCAGCTCTGG
176360_8716503_LOC643650_2_1_6	TGGTCTCAGCTCTGGAGAACTAGCTCCTCTGCCACTGCCACATCCTGTCCACCAAGACCTCCTTGCTGAACCCCACCCCCACCCGGTGCCTCAGGACCCTAACTTGTTTGCTGACCTCCA
176360_8716504_CHRNA7_1_1_1	TGCTGATGAGCGCTTTGACGCCACATTCCACACTAACGTGTTGGTGAATTCTTCTGGGCATTGCCAGTACCTGCCTCCAGGCATATTCAAGAGTTCCTGCTACATCGATGTACGCTGGTT
176360_8716504_CHRNA7_1_1_2	TGGTTTCCCTTTGATGTGCAGCACTGCAAACTGAAGTTTGGGTCCTGGTCTTACGGAGGCTGGTCCTTGGATCTGCAGATGCAGGAGGCAGATATCAGTGGCTATATCCCCAATGGAGAA
176360_8716504_CHRNA7_1_1_3	GAGAATGGGACCTAGTGGGAATCCCCGGCAAGAGGAGTGAAAGGTTCTATGAGTGCTGCAAAGAGCCCTACCCCGATGTCACCTTCACAGTGACCATGCGCCGCAGGACGCTCTACTATG
176360_8716504_CHRNA7_1_1_4	CTATGGCCTCAACCTGCTGATCCCCTGTGTGCTCATCTCCGCCCTCGCCCTGCTGGTGTTCCTGCTTCCTGCAGATTCCGGGGAGAAGATTTCCCTGGGGATAACAGTCTTACTCTCTCT
176360_8716504_CHRNA7_1_1_5	TCTCTTACCGTCTTCATGCTGCTCGTGGCTGAGATCATGCCCGCAACATCCGATTCGGTACCATTGATAGCCCAGTACTTCGCCAGCACCATGATCATCGTGGGCCTCTCGGTGGTGGTG
176360_8716504_CHRNA7_1_1_6	TGGTGACAGTGATCGTGCTGCAGTACCACCACCACGACCCCGACGGGGGCAAGATGCCCAAGTGGACCAGAGTCATCCTTCTGAACTGGTGCGCGTGGTTCCTGCGAATGAAGAGGCCCG
176360_8716504_CHRNA7_1_1_7	GCCCGGGGAGGACAAGGTGCGCCCGGCCTGCCAGCACAAGCAGCGGCGCTGCAGCCTGGCCAGTGTGGAGATGAGCGCCGTGGCGCCGCCGCCCGCCAGCAACGGGAACCTGCTGTACAT
176360_8716504_CHRNA7_1_1_8	TACATCGGCTTCCGCGGCCTGGACGGCGTGCACTGTGTCCCGACCCCCGACTCTGGGGTAGTGTGTGGCCGCATGGCCTGCTCCCCCACGCACGATGAGCACCTCCTGCACGGCGGGCAA
176360_8716504_CHRNA7_1_1_9	GGCAACCCCCCGAGGGGGACCCGGACTTGGCCAAGATCCTGGAGGAGGTCCGCTACATTGCCAACCGCTTCCGCTGCCAGGACGAAAGCGAGGCGGTCTGCAGCGAGTGGAAGTTCGCCG
176360_8716504_CHRNA7_1_1_10	CGCCGCCTGTGTGGTGGACCGCCTGTGCCTCATGGCCTTCTCGGTCTTCACCATCATCTGCACCATCGGCATCCTGATGTCGGCTCCCAACTTCGTGGAGGCCGTGTCCAAAGACTTTGC
176360_8716505_TCAF1_2_1_1	TTGTGAAGATGACCTCTCCGACGACAGAGAAGAGCTTCTGCATGGGATTTCAGAGCTGGACATCAGCAACTCGGATTGTTTCCCATCCCAGCTGCTAGTGCATGGGGCTTTAGCCTTTCC
176360_8716505_TCAF1_2_1_2	CTCTAGGGTTAGATTCCTACCATGGCTGTGTTATAGCGGCTGCCCGCTATGGCCGGGGCCGGGTGGTTGTGACTGGCCATAAGGTATTATTCACTGTTGGTAAACTGGGCCCCTTTCTGC
176360_8716505_TCAF1_2_1_3	CTCAATGCTGTCCGCTGGCTGGATGGGGGCCGCAGAGGCAAGGTTGTGGTGCAGACAGAGCTGAGAACCCTGAGTGGCCTCCTCGCAGTGGGGGCATAGACACCAGCATCGAGCCCAAT
176360_8716505_TCAF1_2_1_4	TCTGACCAGTGATGCAAGTGTCTACTGCTTTGAACCCGTGAGTGAAGTGGGGGTCAAGGAACTGCAGGAGTTTGTAGCAGAGGGTGGCGGGCTATTTGTTGGAGCCCAAGCCTGGTGGTG
176360_8716505_TCAF1_2_1_5	GGGCCTTCAAGAACCCCGGAGTGTCCCCTTTGGCTCGATTCCCAGGAAACCTCCTCCTCAACCCCTTTGGCATCAGCATTACAAGCCAAAGCCTCAATCCAGGGCCCTTTCGTACTCCTA
176360_8716505_TCAF1_2_1_6	AAAGCAGGGATAAGGACCTATCACTTCCGCTCCACCTTGGCCGAGTTCCAGGTTATAATGGGCAGGAAGAGAGGAAATGTGGAAAAGGGCTGGTTGGCAAAGCTGGGACCAGATGGTGCA
176360_8716505_TCAF1_2_1_7	AGCTTTCCTGCAGATTCCCGCAGAAGAGATCCCTGCCTACATGTCTGTGCATCGACTCCTGAGGAAGCTGCTAAGTCGATATCGGCTTCCAGTAGCAACCCGAGAGAACCCTGTTATCAA
176360_8716505_TCAF1_2_1_8	ATGACTGCTGCAGAGGTGCTATGCTTTCCCTGGCCACAGGGCTGGCCCACTCTGGAAGCGACCTCTCTCTGTTAGTCCCAGAAATTGAAGATATGTACAGCAGCCCCTATCTGCGCCCCT
176360_8716505_TCAF1_2_1_9	TCAGAATCTCCTATCACCGTCGAGGTCAACTGCACCAATCCAGGCACCAGATATTGCTGGATGAGTACTGGGCTCTACATACCTGGAAGGCAAATTATAGAAGTCTCACTGCCTGAAGCT
176360_8716505_TCAF1_2_1_10	TGCTGCCTCTGCCGACCTGAAGATACAGATTGGCTGCCACACAGATGACCTGACCAGGGCCAGCAAGCTTTTCCGAGGCCCACTCGTAATTAACCGGTGCTGCTTGGACAAACCCACAAA
176360_8716505_TCAF1_2_1_11	AATCGATCACGTGCCTCTGGGGTGGACTCCTCTATATAATTGTGCCTCAGAACAGCAAACTGGGTTCTGTGCCTGTCACCGTGAAGGGGGCTGTGCATGCTCCATACTACAAGCTGGGGG
176360_8716505_TCAF1_2_1_12	GAGACCACCCTGGAGGAGTGGAAGAGGCGTATCCAGGAGAATCCAGGGCCCTGGGGAGAGCTGGCCACGGACAACATCATTCTGACCGTGCCGACCGCAAATCTTCGTACTCTGGAGAAC
176360_8716505_TCAF1_2_1_13	CCCTGAGCCGCTGCTCCGCCTCTGGGATGAGGTGATGCAGGCTGTGGCGCGACTGGGAGCTGAGCCCTTCCCTTTGCGCCTGCCTCAGAGGATTGTTGCCGACGTGCAGATCTCAGTGGG
176360_8716505_TCAF1_2_1_14	GCTGGATGCATGCAGGGTACCCCATCATGTGCCATCTGGAGTCAGTGCAGGAGCTCATCAACGAGAAGCTCATCAGAACCAAGGGGCTGTGGGGCCCCGTCCATGAGCTGGGCCGCAACC
176360_8716505_TCAF1_2_1_15	CAGCAGCGGCAGGAGTGGGAGTTCCCACCACACACCACCGAGGCCACCTGCAACCTGTGGTGTGTGTATGTGCATGAGACGGTCTTGGGCATTCCTCGAAGCCGTGCCAATATTGCTCTG
176360_8716505_TCAF1_2_1_16	GTGGCCCCCAGTTCGGGAGAAGAGAGTCAGAATCTACCTGAGCAAGGGTCCCAATGTGAAAAACTGGAATGCATGGACCGCACTGGAAACGTATTTACAGCTCCAGGAAGCCTTTGGTTG
176360_8716505_TCAF1_2_1_17	GGGAGCCATTCATCCGTCTCTTCACCGAGTACAGGAACCAGACCAACTTGCCCACAGAAAATGTTGACAAAATGAATCTGTGGGTCAAGATGTTCTCCCACCAAGTGCAGAAGAACCTGG
176360_8716505_TCAF1_2_1_18	GCTCCGTTCTTTGAGGCCTGGGCCTGGCCCATCCAGAAGGAAGTGGCTACCAGCCTGGCCTATCTGCCTGAATGGAAGGAAAATATTATGAAATTGTACCTCCTCACACAGAT
176360_8716507_AMY1A_1_1_1	ATGAAGTTCTTTCTGTTGCTTTTCACCATTGGGTTCTGCTGGGCTCAGTATTCCCCAAATACACAACAAGGACGGACATCTATTGTTCATCTGTTTGAATGGCGATGGGTTGATATTGCT
176360_8716507_AMY1A_1_1_3	AGTTAGCTATAAATTATGCACAAGATCTGGAAATGAAGATGAATTTAGAAACATGGTGACTAGATGTAACAATGTTGGGGTTCGTATTTATGTGGATGCTGTAATTAATCATATGTGTGG
176360_8716507_AMY1A_1_1_5	CTGGAAGTGGAGATATCGAGAACTACAATGATGCTACTCAGGTCAGAGATTGTCGTCTGACTGGTCTTCTTGATCTTGCACTGGAGAAGGATTACGTGCGTTCTAAGATTGCCGAATATA
176360_8716507_AMY1A_1_1_7	TTCCCTGCAGGAAGTAAACCTTTCATTTACCAGGAGGTAATTGATCTGGGTGGTGAGCCAATTAAAAGCAGTGACTACTTTGGTAATGGCCGGGTGACAGAATTCAAGTATGGTGCAAAA
176360_8716507_AMY1A_1_1_9	CAATCAACGAGGACATGGGGCTGGAGGAGCCTCTATTCTTACCTTCTGGGATGCTAGGCTGTACAAAATGGCAGTTGGATTTATGCTTGCTCATCCTTACGGATTTACACGAGTAATGTC
176360_8716507_AMY1A_1_1_11	GCAATGACTGGGTCTGTGAACATCGATGGCGCCAAATAAGGAACATGGTTATTTTCCGCAATGTAGTGGATGGCCAGCCTTTTACAAATTGGTATGATAATGGGAGCAACCAAGTGGCTT
176360_8716507_AMY1A_1_1_13	ATTAATGGCAATTGCACAGGCATTAAAATTTACGTTTCTGATGATGGCAAAGCTCATTTTTCTATTAGTAACTCTGCTGAAGATCCATTTATTGCAATTCATGCTGAATCTAAATTGTAA
176360_8716508_AMY1B_1_1_1	ATGAAGCTCTTTTGGTTGCTTTTCACCATTGGGTTCTGCTGGGCTCAGTATTCCTCAAATACACAACAAGGACGAACATCTATTGTTCATCTGTTTGAATGGCGATGGGTTGATATTGCT
176360_8716508_AMY1B_1_1_2	CTCTTGAATGTGAGCGATATTTAGCTCCCAAGGGATTTGGAGGGGTTCAGGTCTCTCCACCAAATGAAAATGTTGCCATTCACAACCCTTTCAGACCTTGGTGGGAAAGATACCAACCAG
176360_8716508_AMY1B_1_1_4	GGTAATGCTGTGAGTGCAGGAACAAGCAGTACCTGTGGAAGTTACTTCAACCCTGGAAGTAGGGACTTTCCAGCAGTCCCATATTCTGGATGGGATTTTAATGATGGTAAATGTAAAACT
176360_8716508_AMY1B_1_1_6	TATGAACCATCTCATTGACATTGGTGTTGCAGGGTTCAGAATTGATGCTTCCAAGCACATGTGGCCTGGAGACATAAAGGCAATTTTGGACAAACTGCATAATCTAAACAGTAACTGGTT
176360_8716508_AMY1B_1_1_8	AACTCGGCACAGTTATTCGCAAGTGGAATGGAGAGAAGATGTCTTACTTAAAGAACTGGGGAGAAGGTTGGGGTTTCATGCCTTCTGACAGAGCGCTTGTCTTTGTGGATAACCATGACA
176360_8716508_AMY1B_1_1_10	TCAAGCTACCGTTGGCCAAGATATTTTGAAAATGGAAAAGATGTTAATGATTGGGTTGGGCCACCAAATGATAATGGAGTAACTAAAGAAGTTACTATTAATCCAGACACTACTTGTGGC
176360_8716508_AMY1B_1_1_12	TTTTGGGAGAGGAAACAGAGGATTCATTGTTTTCAACAATGATGACTGGACATTTTCTTTAACTTTGCAAACTGGTCTTCCTGCTGGCACATACTGTGATGTCATTTCTGGAGATAAAAT
176360_8716511_KANSL1_1_1_1	GAGACGCAGGTCAGAATGGAAATGGGCTGCAGACCGGGCAGCTATTGTCAGCCGCTGGAACTGGCTTCAGGCTCATGTTTCTGACTTGGAATATCGAATTCGTCAGCAAACAGACATTTA
176360_8716511_KANSL1_1_1_2	ACAAACAGATACGTGCTAATAAGGGGTTGATAGTTCTTGGGGAGGTACCTCCCCCAGAGCATACAACAGACTTATTTCTTCCACTTAGTTCTGAGGTGAAGACAGATCATGGGACTGATA
176360_8716511_KANSL1_1_1_4	TAACACTCTTCAGCCTGTCTTGGCAGACCACATTCCAGGTGACAGCTCTGATGCTGAGGAACAATTACATAAGAAGCAACGACTGAATCTCGTCTCTTCATCATCTGATGGCACCTGTGT
176360_8716511_KANSL1_1_1_5	TGGCAGCCCGGACACGTCCTGTACTGAGCTGTAAGAAGCGGAGGCTTGTTCGACCCAACAGCATCGTTCCTCTTTCCAAGAAGGTTCACCGGAACAGCACAATCCGCCCTGGCTGTGATG
176360_8716511_KANSL1_1_1_7	TCTAGCATTTCCAGATGATGTTCCCACAAGCCTGCATTTCCAGAGCATGCTGAAATCTCAGTGGCAGAACAAGCCTTTTGACAAAATCAAACCTCCCAAAAAGTTATCGCTTAAGCACAG
176360_8716511_KANSL1_1_1_8	GAGCACCCATGCCGGGCAGTCTGCCAGATTCAGCTCGTAAGGACAGGCACAAATTGGTCAGCTCCTTCCTAACAACAGCCAAGCTGTCCCATCACCAAACCCGGCCTGACAGGACCCACA
176360_8716511_KANSL1_1_1_10	ATTGCGAGACCATTCATCTGAGAGAAGTGAAGTGTTGAAGCATCACACAGACATGAGCAGTTCGAGCTACTTGGCAGCCACCCACCATCCTCCACACAGTCCCTTGGTGCGACAGCTCTC
176360_8716511_KANSL1_1_1_11	CCACCTCCTCAGATTCCCCTGCACCCGCCAGCTCTAGCTCACAGGTTACAGCCAGCACATCGCAGCAGCCAGTAAGGAGGAGAAGGGGAGAGAGCTCATTTGATATTAACAACATTGTCA
176360_8716511_KANSL1_1_1_13	TGAAGAGATTGAGGACCTATCCGACGCAGCCTTCGCCGCCCTGCATGCCAAATGTGAGGAGATGGAGAGGGCACGGTGGCTGTGGACCACGAGTGTGCCACCCCAGCGGCGGGGCAGCAG
176360_8716511_KANSL1_1_1_14	GGTCCTACAGGTCATCAGACGGCCGGACAACCCCCCAGCTGGGCAGTGCCAACCCCTCCACCCCCCAGCCTGCCTCCCCTGATGTCAGCAGTAGCCACTCTTTGTCAGAATACTCCCATG
176360_8716511_KANSL1_1_1_16	GGGGCTGGATGAACAGTCTGTCCAGCCCTGGGAGCGGCGGACCTTCCCCCTGGCGCACAGTCCCCAGGCGGAGTGTGAGGACCAGCTGGATGCACAGGAGCGAGCAGCCCGCTGCACTCG
176360_8716511_KANSL1_1_1_17	GACGCACCTCAGGCAGCAAGACTGGCCGGGAGACAGAGGCAGCGCCCACCTCGCCTCCCATTGTCCCCCTCAAGAGTCGGCATCTGGTGGCAGCAGCCACAGCTCAGCGCCCGACTCACA
176360_8716512_ANK2_1_1_1	TCTGACAGCAATGCAAGCTTCCTCCGTGCTGCCAGAGCAGGCAACCTGGACAAAGTTGTGGAATATCTGAAGGGGGGCATAGACATCAATACCTGCAATCAGAATGGACTCAACGCTCTC
176360_8716512_ANK2_1_1_5	AATGACACCAAAGGGAAAGTGAGGCTGCCAGCTCTGCATATTGCCGCTAGGAAAGACGACACCAAATCTGCCGCACTTCTGCTTCAGAATGACCACAATGCTGACGTACAATCCAAGATG
176360_8716512_ANK2_1_1_9	CAGGGAGACCACGTGGAATGTGTGAAGCACCTGTTACAGCACAAGGCACCTGTTGATGATGTCACCCTAGACTACCTGACAGCCCTCCACGTTGCTGCGCACTGTGGCCACTACCGTGTA
176360_8716512_ANK2_1_1_13	GAACAGACACCTTTACATATTGCCTCCCGCCTGGGTAAGACAGAAATTGTCCAGCTGCTTCTACAACATATGGCTCATCCAGATGCGGCCACTACAAATGGGTACACACCACTGCACATC
176360_8716512_ANK2_1_1_17	GCAGAGACAAACATTGTGACAAAGCAAGGAGTAACTCCACTCCATCTGGCCTCGCAGGAGGGGCACACAGATATGGTTACCTTGCTTCTGGATAAGGGAGCCAATATCCACATGTCAACT
176360_8716512_ANK2_1_1_21	GTGGTCGACACCCTGAAGGTTGTGACTGAGGAGGTCACCACCACCACCACAACTATTACAGAAAAACACAAACTAAATGTACCTGAGACGATGACTGAGGTTCTTGATGTTTCTGATGAA
176360_8716512_ANK2_1_1_24	GTTGTGATTCCCAGTCACCAGGTGTCAACTCTAGCCAAGGAGGCAGAAAGGAATTCTTATCGCCTAAGCTGGGGCACTGAGAACTTAGACAACGTGGCTCTTTCTTCTAGTCCTATTCAT
176360_8716513_ANK2_26_1_1	TTTCCTGGTTAGTTTTATGGTGGATGCCCGAGGTGGTGCTATGCGAGGATGCAGACACAATGGGCTCCGAATCATTATTCCACCTCGGAAATGTACTGCTCCAACGCGAGTCACCTGCCG
176360_8716513_ANK2_26_1_5	CATGGATGAAGTACTGGATAGCCCAGAAGACCTAGAAAAGAAACGAATCTGCCGCATCATCACCCGAGACTTCCCACAGTACTTTGCAGTGGTGTCTCGTATCAAACAGGACAGCAATCT
176360_8716513_ANK2_26_1_9	TACGCCATTAACATTTGTCAATGAATGTGTTTCCTTTACAACAAACGTGTCTGCCAGGTTCTGGCTGATAGATTGTCGACAGATCCAGGAATCCGTTACTTTTGCATCACAAGTATACAG
176360_8716513_ANK2_26_1_13	ATCCACGAGAGGCCTGGTGCATCAAGCTATTTGCAACTTAAACATCACTTTGCCGATTTATACAAAGGAATCAGAGTCAGATCAAGAACAGGAGGAAGAGATCGATATGACATCAGAAAA
176360_8716514_ANK2_37_1_1	ATCCACAGGATGAGCAGGAACGGATCGAGGAAAGGCTGGCTTATATTGCTGATCACCTTGGCTTCAGCTGGACAGAATTAGCAAGAGAACTGGATTTCACTGAGGAGCAAATTCATCAAA
176360_8716514_ANK2_37_1_5	GGTTATTCCACTTTTCAGGATGGCGTCCCCAAAACTGAGGGGGACAGCTCAGCAACAGCACTCTTTCCCCAAACTCACAAGGAGCAAGTTCAACAGGATTTCTCAGGGAAAATGCAAGAC
176360_8716514_ANK2_37_1_10	TTATGGTGCAGGGAATGCCACAGGAACCTGTCAACATCGAGGAAGGGGATGGCTATTCCAAAGTTATAAAGCGTGTTGTATTGAAGAGTGACACCGAGCAGTCAGAGGACAACAATGAGT
176360_8716515_ARID1B_2_1_1	GGCAGCCCAATGGATCCAATGGTGATGAAGAGACCTCAGTTGTATGGCATGGGCAGTAACCCTCATTCTCAGCCTCAGCAGAGCAGTCCGTACCCAGGAGGTTCCTATGGCCCTCCAGGC
176360_8716516_ARID1B_3_1_1	ATGCCACCTCAGTATGGACAGCAAGGTGTGAGTGGTTACTGCCAGCAGGGCCAACAGCCATATTACAGCCAGCAGCCGCAGCCCCCGCACCTCCCACCCCAGGCGCAGTATCTGCCGTCC
176360_8716516_ARID1B_3_1_3	TACAGCAAGAAAGACCATCAAGTTTACCAGATCTGTCTGGCTCCATTGATGACCTCCCCACGGGAACGGAAGCAACTTTGAGCTCAGCAGTCAGTGCATCCGGGTCCACGAGCAGCCAAG
176360_8716516_ARID1B_3_1_7	CATGCTGGAATCAGTAGCTTTCAGCAGAGTAACTCAAGTGGGACTTACGGTCCACAGATGAGCCAGTATGGACCACAAGGTAACTACTCCAGACCCCCAGCGTATAGTGGGGTGCCCAGT
176360_8716516_ARID1B_3_1_11	CAGCCAGCCCATGAACAACAGCTCTAGCCTGATGAACACGCAGGCGCCGCCCTACAGCATGGCGCCCGCCATGGTGAACAGCTCGGCAGCATCTGTGGGTCTTGCAGATATGATGTCTCC
176360_8716516_ARID1B_3_1_15	CAAGCAGTGCAGCGAGCTCCCTGAAAAAGCAGTATATTCAGTACCTGTTTGCCTTTGAGTGCAAGATCGAACGTGGGGAGGAGCCCCCGCCGGAAGTCTTCAGCACCGGGGACACCAAAA
176360_8716516_ARID1B_3_1_19	CCTGGAAGCAGCGAGCCCTTTATGACGCAAGGACAGATGCCCAACAGCAGCATGCAGGACATGTACAACCAAAGTCCCTCCGGAGCAATGTCTAACCTGGGCATGGGGCAGCGCCAGCAG
176360_8716516_ARID1B_3_1_23	CCCGCCTCAGATGATGGGCGGCCCGCTGCAGTCGTCCTCCAGTGAGGGGCCTCAGCAGAATATGTGGGCAGCACGCAATGATATGCCTTATCCCTACCAGAACAGGCAGGGCCCTGGCGG
176360_8716516_ARID1B_3_1_28	TGGCGTGTGATGATGTCCCTTAAATCAGGTCTTTTGGCTGAGAGTACGTGGGCTTTGGACACTATTAATATTCTTCTGTATGATGACAGCACTGTTGCTACTTTCAATCTCTCCCAG
176360_8716517_GRIN2B_1_1_1	GGTATCTACAGCTGCATCCATGGGGTGGCGATCGAGGAGCGCCAGTCTGTAATGAACTCCCCCACCGCAACCATGAACAACACACACTCCAACATCCTGCGCCTGCTGCGCACGGCCAAG
176360_8716517_GRIN2B_1_1_5	CTGGACATCGGCCTCCCCTCCTCCAAGCACAGCCAGCTCAGTGACCTGTACGGCAAATTCTCCTTCAAGAGCGACCGCTACAGTGGCCACGACGACTTGATCCGCTCCGATGTCTCTGAC
176360_8716517_GRIN2B_1_1_9	ATCAAGCACGGGACGGGCGACAAACACGGCGTGGTCAGCGGGGTACCTGCACCTTGGGAGAAGAACCTGACCAACGTGGAGTGGGAGGACCGGTCCGGGGGCAACTTCTGCCGCAGCTGT
176360_8716517_GRIN2B_1_1_13	CCCTACGCCCACATGTTTGAGATGTCAGCTGGCGAGAGCACCTTTGCCAACAACAAGTCCTCAGTGCCCACTGCCGGACATCACCACCACAACAACCCCGGCGGCGGGTACATGCTCAGC
176360_8716517_GRIN2B_1_1_17	TGGAAGCTCTCTGGCTCACTGGCATTTGTCACAATGAGAAGAATGAGGTCATGAGCAGCCAGCTGGACATTGACAACATGGCAGGGGTCTTCTACATGTTGGGGGCGGCCATGGCTCTCA
176360_8716517_GRIN2B_1_1_21	CAGCACAGAGAGAAATATTCGCAATAACTATGCAGAAATGCATGCCTACATGGGAAAGTTCAACCAGAGGGGTGTAGATGATGCATTGCTCTCCCTGAAAACAGGAGCCTGGTGGACCCT
176360_8716517_GRIN2B_1_1_25	TGAAGAGGGCCTACATGGCAGTGGGCTCACTCACCATCAATGAGGAACGATCGGAGGTGGTCGACTTCTCTGTGCCCTTCATAGAGACAGGCATCAGTGTCATGGTGTCACGCAGCAATG
176360_8716517_GRIN2B_1_1_29	GAGTGGAACCTGCATGAGGAACACAGTCCCCTGCCAAAAACGCATAGTCACTGAGTATCTGATCAATGTCACTTTTGAGGGGAGGAATTTGTCCTTCAGTGAAGATGGCTACCAGATGCA
176360_8716517_GRIN2B_1_1_33	CACCTACATCTTTGAAGTGGCCAACTCAGTAGGGCTGACTGGCTATGGCTACACGTGGATCGTGCCCAGTCTGGTGGCAGGGGATACAGACACAGTGCCTGCGGAGTTCCCCACTGGGCT
176360_8716517_GRIN2B_1_1_38	GGGGTGGTGTTTGCTGATGACACAGACCAGGAAGCCATCGCCCAGATCCTCGATTTCATTTCAGCACAGACTCTCACCCCCATCCTGGGCATCCACGGGGGCTCCTCTATGATAATGGCA
176360_8716518_KDM6B_1_1_1	ATGCATCGGGCAGTGGACCCTCCAGGGGCCCGCGCTGCACGGGAAGCCTTTGCCCTTGGGGGCCTGAGCTGTGCTGGGGCCTGGAGCTCCTGCCCGCCTCATCCCCCTCCTCGTAGCGCA
176360_8716518_KDM6B_1_1_4	CTTGGGCAACTGTACGAGTCAGAGCACGATAGTGAGGAGGCCACACGCTGCTACCACAGCGCCCTTCGATACGGAGGAAGCTTCGCTGAGCTGGGGCCCCGCATTGGCCGACTGCAGCAG
176360_8716520_KDM6B_8_1_1	GAATGCCAAGGTGAAAGGGAAGTTTCGAGAGTCCTACCTTTCCCCTGCCCAGTCTGTGAAACCGAAGATCAACACTGAGGAGAAGCTGCCCCGGGAAAAACTCAACCCCCCTACACCCAG
176360_8716520_KDM6B_8_1_4	TCGGCAGATCTGGCCTTGTGAGAGCTCCCGTTCCCACACCACCATTGCCAAGTACGCACAGTACCAGGCCTCATCCTTCCAGGAGTCTCTGCAGGAGGAGAAGGAGAGTGAGGATGAGGA
176360_8716520_KDM6B_8_1_7	GCTGTACATGAAGGTGCCCGGCAGCCGAACGCCAGGCCACCAGGAGAATAACAACTTCTGCTCCGTCAACATCAACATTGGCCCAGGCGACTGCGAGTGGTTCGCGGTGCACGAGCACTA
176360_8716520_KDM6B_8_1_10	GTACCAGCTGGCCCTGGAACGATACGAGTGGAATGAGGTGAAGAACGTCAAATCCATCGTGCCCATGATTCACGTGTCATGGAACGTGGCTCGCACGGTCAAAATCAGCGACCCCGACTT
176360_8716520_KDM6B_8_1_14	CGGAGGAGGGCACTGGGGCAGGCTGCAGGGACGGGCTTCGGGAGCCCGGCCGCGCCTTTCCCTGAGCCCCCGCCGGCTTTCTCCCCCCAGGCCCCAGCCAGCACGTCGCGATGA
176360_8716521_DSCAM_1_1_1	GTGGTCCAGGCACCAGCAGGGACCTGAGCTTAGGACAAGCATGCTTGGAACCTCAGAAAAGCCGGACCCTGAAGCGCCCCACGGTCCTGGAGCCCATCCCGATGGAAGCCGCCTCCTCCG
176360_8716521_DSCAM_1_1_5	GGACAGTCTGACCTCCAGCACCCCTTCCGAATCGGGAATCTGCAGGTTCACTGCATCTCCCCCCAAACCTCAGGATGGAGGAAGAGTAATGAATATGGCAGTTCCAAAGGCACATCGGCC
176360_8716521_DSCAM_1_1_9	CGGACTTCAGATACGTTAAGCAAGCAACAGCAGACCCTGCGAATGCACATCGACATACCCAGGGCTCAGCTTTTGATTGAAGAGAGAGACACGATGGAGACCATTGATGCAAAGAGTTTA
176360_8716521_DSCAM_1_1_13	CTCCAAGTCCTACATCCTGTATGACCTGCAGGAAGCCACCTGGTATGAGCTGCAGATGCGGGTGTGCAACAGTGCGGGCTGCGCGGAGAAGCAGGCCAACTTCGCTACGCTGAACTACGA
176360_8716521_DSCAM_1_1_17	ATTCGCACGGTGAAAGCAGAAGACTCCGGCTATTACAGCTGCATTGCCAATAACAACTGGGGATCTGATGAAATTATTTTAAACTTACAAGTACAAGCTCCTGCACGAATCCTGACCTTC
176360_8716521_DSCAM_1_1_21	TCTGCTCCCACCCCTATCCCACAAGCTGGGTGAGATTAAAAACATCACCACCACACAGCCTTCACTGGAGCTGGACGGGCTGGAAAAGTACACCAACTACAGCATCCAGGTGCTGGCCTT
176360_8716521_DSCAM_1_1_25	CTCAGTACGGCCTGGTGGTGCAGGCCTGTAACCGGGCCGGCACGGGGCCTTCTTCTCAGGAAATCATCACCACCACTCTCGAGGATGCTCCTGATGGTCCACCTCAGGAAGTTCACCTGG
176360_8716521_DSCAM_1_1_29	CTGGTTTCTTTTCCTGCCATGCTATTAATTCTTATGGGGAGGACCGTGGAATAATTCAGCTCACAGTGCAAGTTCCTGCGATGATAACATCCTATCCAAATACTACCCTGGCCACGCAGG
176360_8716521_DSCAM_1_1_33	CGGGATTTATGGCAAAGCAGTCATCCTCAATTGTTCTGCTGAGGGTTACCCTGTACCTACCATCGTGTGGAAATTCTCTAAAGTTCCGCCTTTCATACAACCCTTTGAGTTTCCAAGATT
176360_8716521_DSCAM_1_1_37	CCTGCTTCCTTTCAACCACCGCCAAGTGGCATTTGAGAACAATGGAACTCTTAAACTTTCAGATGTGCAAAAGGAAGTGGACGAGGGGGAGTACACGTGCAACGTGTTGGTTCAACCACA
176360_8716521_DSCAM_1_1_41	CAGGAAGGTTAAAAGCAGCGTGGGTAGCCAAGTTTCCTTGTCCTGCAGCGTGACAGGAACTGAGGACCAGGAACTCTCCTGGTACCGCAATGGTGAAATCCTCAACCCTGGAAAAAATGT
176360_8716521_DSCAM_1_1_45	GTCCAACAGATACGGAACTGCTAAGGTGATAGGCCGCCTGTACGTGAAACGATCTAGATTTCTCATCACATCCACGGGAGCCTTGTATATTAAAGATGTACAGAATGAAGATGGATTGTA
176360_8716521_DSCAM_1_1_50	TTATTGCACAGCTGAAAATCCTTCAGGGAAAATTAGAAGTCAGGATGTCCACATCAAGGCTGATGTGGATACTGGCTCTCTCCTTGTTCCAGAGCTTCGCGAATG

Table S3. Barcoded end sequences

f_A1	AAGCAGTGGTATCAACGCAGAGTTACTAGAGTAGCACTCCATACGATTTAGGTGACACTATAGGG	r_A1	GAGTGCTACTCTAGTAACTCTGCGTTGATACCACTGCTT
f_B1	AAGCAGTGGTATCAACGCAGAGTTGTGTATCAGTACATGCATACGATTTAGGTGACACTATAGGG	r_B1	CATGTACTGATACACAACTCTGCGTTGATACCACTGCTT
f_C1	AAGCAGTGGTATCAACGCAGAGTACACGCATGACACACTCATACGATTTAGGTGACACTATAGGG	r_C1	AGTGTGTCATGCGTGTACTCTGCGTTGATACCACTGCTT
f_D1	AAGCAGTGGTATCAACGCAGAGTGATCTCTACTATATGCCATACGATTTAGGTGACACTATAGGG	r_D1	GCATATAGTAGAGATCACTCTGCGTTGATACCACTGCTT
f_E1	AAGCAGTGGTATCAACGCAGAGTATGATGTGCTACATCTCATACGATTTAGGTGACACTATAGGG	r_E1	CAGCAGTATAGACTGTACTCTGCGTTGATACCACTGCTT
f_F1	AAGCAGTGGTATCAACGCAGAGTCGTCTATATACGTATACATACGATTTAGGTGACACTATAGGG	r_F1	AGATGTAGCACATCATACTCTGCGTTGATACCACTGCTT
f_G1	AAGCAGTGGTATCAACGCAGAGTATAGAGACTCAGAGCTCATACGATTTAGGTGACACTATAGGG	r_G1	CGATCAGCTGAGCGCGACTCTGCGTTGATACCACTGCTT
f_H1	AAGCAGTGGTATCAACGCAGAGTTAGATGCGAGAGTAGACATACGATTTAGGTGACACTATAGGG	r_H1	GTCGCGACGTCAGTGTACTCTGCGTTGATACCACTGCTT
f_A2	AAGCAGTGGTATCAACGCAGAGTCATCACTACGCTAGATCATACGATTTAGGTGACACTATAGGG	r_A2	TATACGTATATAGACGACTCTGCGTTGATACCACTGCTT
f_B2	AAGCAGTGGTATCAACGCAGAGTTATGTGATCGTCTCTCCATACGATTTAGGTGACACTATAGGG	r_B2	AGCTCTGAGTCTCTATACTCTGCGTTGATACCACTGCTT
f_C2	AAGCAGTGGTATCAACGCAGAGTCGTGTCGCGCATATCTCATACGATTTAGGTGACACTATAGGG	r_C2	TCTACTCTCGCATCTAACTCTGCGTTGATACCACTGCTT
f_D2	AAGCAGTGGTATCAACGCAGAGTATATCAGTCATGCATACATACGATTTAGGTGACACTATAGGG	r_D2	CACGATAGTCGCTATGACTCTGCGTTGATACCACTGCTT
f_E2	AAGCAGTGGTATCAACGCAGAGTGAGATCGACAGTCTCGCATACGATTTAGGTGACACTATAGGG	r_E2	ATCTAGCGTAGTGATGACTCTGCGTTGATACCACTGCTT
f_F2	AAGCAGTGGTATCAACGCAGAGTCACGCACACACGCGCGCATACGATTTAGGTGACACTATAGGG	r_F2	TGCATGCACAGATGCGACTCTGCGTTGATACCACTGCTT
f_G2	AAGCAGTGGTATCAACGCAGAGTGTAGTCTCGCACAGATCATACGATTTAGGTGACACTATAGGG	r_G2	GAGAGACGATCACATAACTCTGCGTTGATACCACTGCTT
f_H2	AAGCAGTGGTATCAACGCAGAGTGCTCGACTGTGAGAGACATACGATTTAGGTGACACTATAGGG	r_H2	AGATATGCGCGACACGACTCTGCGTTGATACCACTGCTT
f_A3	AAGCAGTGGTATCAACGCAGAGTAGAGATGTGTGATGACCATACGATTTAGGTGACACTATAGGG	r_A3	TATGCATGACTGATATACTCTGCGTTGATACCACTGCTT
f_B3	AAGCAGTGGTATCAACGCAGAGTTACGACTACATATCAGCATACGATTTAGGTGACACTATAGGG	r_B3	CGAGACTGTCGATCTCACTCTGCGTTGATACCACTGCTT
f_C3	AAGCAGTGGTATCAACGCAGAGTTATCTCTGTAGAGTCTCATACGATTTAGGTGACACTATAGGG	r_C3	CGCGCGTGTGTGCGTGACTCTGCGTTGATACCACTGCTT
f_D3	AAGCAGTGGTATCAACGCAGAGTAGAGAGAGACATGCGCCATACGATTTAGGTGACACTATAGGG	r_D3	CACACGCGCGTGCTCGACTCTGCGTTGATACCACTGCTT
f_E3	AAGCAGTGGTATCAACGCAGAGTACTCTCGCTCTGTAGACATACGATTTAGGTGACACTATAGGG	r_E3	ATCTGTGCGAGACTACACTCTGCGTTGATACCACTGCTT
f_F3	AAGCAGTGGTATCAACGCAGAGTTCTATGTCTCAGTAGTCATACGATTTAGGTGACACTATAGGG	r_F3	TCTCTCACAGTCGAGCACTCTGCGTTGATACCACTGCTT
f_G3	AAGCAGTGGTATCAACGCAGAGTGAGACTAGAGATAGTGCATACGATTTAGGTGACACTATAGGG	r_G3	GTCATCACACATCTCTACTCTGCGTTGATACCACTGCTT
f_H3	AAGCAGTGGTATCAACGCAGAGTTACGCGTGTACGCAGACATACGATTTAGGTGACACTATAGGG	r_H3	CTGATATGTAGTCGTAACTCTGCGTTGATACCACTGCTT
f_A4	AAGCAGTGGTATCAACGCAGAGTGCACATACACGCTCACCATACGATTTAGGTGACACTATAGGG	r_A4	AGACTCTACAGAGATAACTCTGCGTTGATACCACTGCTT
f_B4	AAGCAGTGGTATCAACGCAGAGTACAGTGCGCTGTCTATCATACGATTTAGGTGACACTATAGGG	r_B4	GCGCATGTCTCTCTCTACTCTGCGTTGATACCACTGCTT
f_C4	AAGCAGTGGTATCAACGCAGAGTTCACATATGTATACACATCATACGATTTAGGTGACACTATAGGG	r_C4	TCTACAGAGCGAGAGTACTCTGCGTTGATACCACTGCTT
f_D4	AAGCAGTGGTATCAACGCAGAGTACACACAGACTGTGAGCATACGATTTAGGTGACACTATAGGG	r_D4	ACTACTGAGACATAGAACTCTGCGTTGATACCACTGCTT
f_E4	AAGCAGTGGTATCAACGCAGAGTTGCTCTCGTGTACTGTCATACGATTTAGGTGACACTATAGGG	r_E4	TGTGATACTGCGAGCAACTCTGCGTTGATACCACTGCTT
f_F4	AAGCAGTGGTATCAACGCAGAGTGTGTGAGATATATATCCATACGATTTAGGTGACACTATAGGG	r_F4	ACTCTATCACACACAGACTCTGCGTTGATACCACTGCTT
f_G4	AAGCAGTGGTATCAACGCAGAGTCTCAGTGTGACACATGCATACGATTTAGGTGACACTATAGGG	r_G4	TATCGCGCTCTCACTGACTCTGCGTTGATACCACTGCTT
f_H4	AAGCAGTGGTATCAACGCAGAGTAGATATCATCAGCGAGCATACGATTTAGGTGACACTATAGGG	r_H4	CACTATCTCTAGTCTCACTCTGCGTTGATACCACTGCTT
f_A5	AAGCAGTGGTATCAACGCAGAGTGTGCAGTGATCGATGACATACGATTTAGGTGACACTATAGGG	r_A5	TCTGCGTACACGCGTAACTCTGCGTTGATACCACTGCTT
f_B5	AAGCAGTGGTATCAACGCAGAGTTGACTCGCTCATAGTCCATACGATTTAGGTGACACTATAGGG	r_B5	GTGAGCGTGTATGTGCACTCTGCGTTGATACCACTGCTT
f_C5	AAGCAGTGGTATCAACGCAGAGTATGCTGATGACGCGCTCATACGATTTAGGTGACACTATAGGG	r_C5	ATAGACAGCGCACTGTACTCTGCGTTGATACCACTGCTT
f_D5	AAGCAGTGGTATCAACGCAGAGTTCGATATACGACGTGCCATACGATTTAGGTGACACTATAGGG	r_D5	ATGTATACATATGTGAACTCTGCGTTGATACCACTGCTT
f_E5	AAGCAGTGGTATCAACGCAGAGTTCGTCATACGCTCTAGCATACGATTTAGGTGACACTATAGGG	r_E5	CTCACAGTCTGTGTGTACTCTGCGTTGATACCACTGCTT
f_F5	AAGCAGTGGTATCAACGCAGAGTCGACTACGTACAGTAGCATACGATTTAGGTGACACTATAGGG	r_F5	ACAGTACACGAGAGCAACTCTGCGTTGATACCACTGCTT
f_G5	AAGCAGTGGTATCAACGCAGAGTGTCTGATAGATACAGACATACGATTTAGGTGACACTATAGGG	r_G5	GATATATATCTCACACACTCTGCGTTGATACCACTGCTT
f_H5	AAGCAGTGGTATCAACGCAGAGTCTGATGCGCGCTGTACCATACGATTTAGGTGACACTATAGGG	r_H5	CATGTGTCACACTGAGACTCTGCGTTGATACCACTGCTT
f_A6	AAGCAGTGGTATCAACGCAGAGTACGTGAGCTCACTCGCCATACGATTTAGGTGACACTATAGGG	r_A6	GCTGACACTAGCTGACACTCTGCGTTGATACCACTGCTT
f_B6	AAGCAGTGGTATCAACGCAGAGTATAGAGAGTGTCTCAGCATACGATTTAGGTGACACTATAGGG	r_B6	CTCGCTGATGATATCTACTCTGCGTTGATACCACTGCTT
f_C6	AAGCAGTGGTATCAACGCAGAGTCATAGAGAGATAGTATCATACGATTTAGGTGACACTATAGGG	r_C6	TCATCGATCACTGCACACTCTGCGTTGATACCACTGCTT
f_D6	AAGCAGTGGTATCAACGCAGAGTATCTCGAGATGTAGCGCATACGATTTAGGTGACACTATAGGG	r_D6	GACTATGAGCGAGTCAACTCTGCGTTGATACCACTGCTT
f_E6	AAGCAGTGGTATCAACGCAGAGTGATCGACTCGAGCATCCATACGATTTAGGTGACACTATAGGG	r_E6	GCACGTCGTATATCGAACTCTGCGTTGATACCACTGCTT
f_F6	AAGCAGTGGTATCAACGCAGAGTATGCTCACTACTACATCATACGATTTAGGTGACACTATAGGG	r_F6	CTAGAGCGTATGACGAACTCTGCGTTGATACCACTGCTT
f_G6	AAGCAGTGGTATCAACGCAGAGTCGTGCACATCTATAGCCATACGATTTAGGTGACACTATAGGG	r_G6	CTACTGTACGTAGTCGACTCTGCGTTGATACCACTGCTT
f_H6	AAGCAGTGGTATCAACGCAGAGTTATGACTAGTGTACTACATACGATTTAGGTGACACTATAGGG	r_H6	TCTGTATCTATCAGACACTCTGCGTTGATACCACTGCTT
f_A7	AAGCAGTGGTATCAACGCAGAGTGACGTGTCGTAGATATCATACGATTTAGGTGACACTATAGGG	r_A7	GTGAGTCAGAGATCTAACTCTGCGTTGATACCACTGCTT
f_B7	AAGCAGTGGTATCAACGCAGAGTATAGCGACGCGATATACATACGATTTAGGTGACACTATAGGG	r_B7	GTACAGCGCGCATCAGACTCTGCGTTGATACCACTGCTT
f_C7	AAGCAGTGGTATCAACGCAGAGTATCGCTGTGTCTATAGCATACGATTTAGGTGACACTATAGGG	r_C7	GCATCGTGCACGAGTGACTCTGCGTTGATACCACTGCTT
f_D7	AAGCAGTGGTATCAACGCAGAGTTGTCGTCTATCATGTACATACGATTTAGGTGACACTATAGGG	r_D7	GCGAGTGAGCTCACGTACTCTGCGTTGATACCACTGCTT
f_E7	AAGCAGTGGTATCAACGCAGAGTCACACGAGATCTCATCCATACGATTTAGGTGACACTATAGGG	r_E7	CTGAGACACTCTCTATACTCTGCGTTGATACCACTGCTT
f_F7	AAGCAGTGGTATCAACGCAGAGTCGTGAGTAGTCAGACGCATACGATTTAGGTGACACTATAGGG	r_F7	ATACTATCTCTCTATGACTCTGCGTTGATACCACTGCTT
f_G7	AAGCAGTGGTATCAACGCAGAGTTCTCGACTGCACATATCATACGATTTAGGTGACACTATAGGG	r_G7	CGCTACATCTCGAGATACTCTGCGTTGATACCACTGCTT
f_H7	AAGCAGTGGTATCAACGCAGAGTTATACACACTCGCTCGCATACGATTTAGGTGACACTATAGGG	r_H7	TGACACGAGTGATCGTACTCTGCGTTGATACCACTGCTT
f_A8	AAGCAGTGGTATCAACGCAGAGTGTCGTAGCTGCTGTATCATACGATTTAGGTGACACTATAGGG	r_A8	GATGCTCGAGTCGATCACTCTGCGTTGATACCACTGCTT
f_B8	AAGCAGTGGTATCAACGCAGAGTCTGTACTAGAGCGTCTCATACGATTTAGGTGACACTATAGGG	r_B8	ATGTAGTAGTGAGCATACTCTGCGTTGATACCACTGCTT
f_C8	AAGCAGTGGTATCAACGCAGAGTTCGAGTGTATAGCTCACATACGATTTAGGTGACACTATAGGG	r_C8	GCTATAGATGTGCACGACTCTGCGTTGATACCACTGCTT
f_D8	AAGCAGTGGTATCAACGCAGAGTACTGTGACAGTATGATCATACGATTTAGGTGACACTATAGGG	r_D8	TCGTGCATGTGCAGTCACTCTGCGTTGATACCACTGCTT
f_E8	AAGCAGTGGTATCAACGCAGAGTTGTCTGAGACGCATACCATACGATTTAGGTGACACTATAGGG	r_E8	TAGTACACTAGTCATAACTCTGCGTTGATACCACTGCTT
f_F8	AAGCAGTGGTATCAACGCAGAGTATCGCATCGCAGAGACCATACGATTTAGGTGACACTATAGGG	r_F8	ATATCTACGACACGTCACTCTGCGTTGATACCACTGCTT
f_G8	AAGCAGTGGTATCAACGCAGAGTTACTCATATATGCTACCATACGATTTAGGTGACACTATAGGG	r_G8	TATATCGCGTCGCTATACTCTGCGTTGATACCACTGCTT
f_H8	AAGCAGTGGTATCAACGCAGAGTGTAGAGTGATCGCGTCCATACGATTTAGGTGACACTATAGGG	r_H8	CTATAGACACAGCGATACTCTGCGTTGATACCACTGCTT
f_A9	AAGCAGTGGTATCAACGCAGAGTACGACAGTCAGAGTATCATACGATTTAGGTGACACTATAGGG	r_A9	CACGCTATCAGTGAGAACTCTGCGTTGATACCACTGCTT
f_B9	AAGCAGTGGTATCAACGCAGAGTATATATAGCTGATGCGCATACGATTTAGGTGACACTATAGGG	r_B9	TACATGATAGACGACAACTCTGCGTTGATACCACTGCTT
f_C9	AAGCAGTGGTATCAACGCAGAGTTGCTATCTGAGATACTCATACGATTTAGGTGACACTATAGGG	r_C9	GATGAGATCTCGTGTGACTCTGCGTTGATACCACTGCTT
f_D9	AAGCAGTGGTATCAACGCAGAGTCAGCAGATCATGTCGACATACGATTTAGGTGACACTATAGGG	r_D9	AGTATCATGTGTATCTACTCTGCGTTGATACCACTGCTT
f_E9	AAGCAGTGGTATCAACGCAGAGTACTATCGCAGCTCAGTCATACGATTTAGGTGACACTATAGGG	r_E9	CGTCTGACTACTCACGACTCTGCGTTGATACCACTGCTT
f_F9	AAGCAGTGGTATCAACGCAGAGTCGTCTCTCGTCTGTGCCATACGATTTAGGTGACACTATAGGG	r_F9	ATATGTGCAGTCGAGAACTCTGCGTTGATACCACTGCTT
f_G9	AAGCAGTGGTATCAACGCAGAGTGAGTCTCGATATACTACATACGATTTAGGTGACACTATAGGG	r_G9	AGCAGCTACGATCGTAACTCTGCGTTGATACCACTGCTT
f_H9	AAGCAGTGGTATCAACGCAGAGTTCTGTCGATATACACTCATACGATTTAGGTGACACTATAGGG	r_H9	CGAGCGAGTGTGTATAACTCTGCGTTGATACCACTGCTT
f_A10	AAGCAGTGGTATCAACGCAGAGTACGTGCTCTATAGAGACATACGATTTAGGTGACACTATAGGG	r_A10	ATACAGCAGCTACGACACTCTGCGTTGATACCACTGCTT
f_B10	AAGCAGTGGTATCAACGCAGAGTTATCAGCACGACATGCCATACGATTTAGGTGACACTATAGGG	r_B10	AGACGCTCTAGTACAGACTCTGCGTTGATACCACTGCTT
f_C10	AAGCAGTGGTATCAACGCAGAGTGCTCTCACGATATCAGCATACGATTTAGGTGACACTATAGGG	r_C10	TGAGCTATACACTCGAACTCTGCGTTGATACCACTGCTT
f_D10	AAGCAGTGGTATCAACGCAGAGTTCTCATGTGTGAGCTACATACGATTTAGGTGACACTATAGGG	r_D10	ATCATACTGTCACAGTACTCTGCGTTGATACCACTGCTT
f_E10	AAGCAGTGGTATCAACGCAGAGTCGTAGCTCAGACACTCCATACGATTTAGGTGACACTATAGGG	r_E10	GTATGCGTCTCAGACAACTCTGCGTTGATACCACTGCTT
f_F10	AAGCAGTGGTATCAACGCAGAGTTCAGAGACACTACGAGCATACGATTTAGGTGACACTATAGGG	r_F10	ATATCACACGTGAGTGACTCTGCGTTGATACCACTGCTT
f_G10	AAGCAGTGGTATCAACGCAGAGTATCGAGCAGCAGTCGTCATACGATTTAGGTGACACTATAGGG	r_G10	GTCTCTGCGATGCGATACTCTGCGTTGATACCACTGCTT
f_H10	AAGCAGTGGTATCAACGCAGAGTGCTAGTCGATGACAGCCATACGATTTAGGTGACACTATAGGG	r_H10	GTAGCATATATGAGTAACTCTGCGTTGATACCACTGCTT
f_A11	AAGCAGTGGTATCAACGCAGAGTGTCTCTCTCTCACGCACATACGATTTAGGTGACACTATAGGG	r_A11	ACATCAGAGAGATCTGACTCTGCGTTGATACCACTGCTT
f_B11	AAGCAGTGGTATCAACGCAGAGTACACACTCTATCAGATCATACGATTTAGGTGACACTATAGGG	r_B11	GACGCGATCACTCTACACTCTGCGTTGATACCACTGCTT
f_C11	AAGCAGTGGTATCAACGCAGAGTAGCGTAGCATCTGAGCCATACGATTTAGGTGACACTATAGGG	r_C11	ATACTCTGACTGTCGTACTCTGCGTTGATACCACTGCTT
f_D11	AAGCAGTGGTATCAACGCAGAGTCTACGATGCTATGTATCATACGATTTAGGTGACACTATAGGG	r_D11	AGTATCTCAGATAGCAACTCTGCGTTGATACCACTGCTT
f_E11	AAGCAGTGGTATCAACGCAGAGTCGACTAGATCTATCATCATACGATTTAGGTGACACTATAGGG	r_E11	TCGACATGATCTGCTGACTCTGCGTTGATACCACTGCTT
f_F11	AAGCAGTGGTATCAACGCAGAGTAGTCGTGACTATGCTCCATACGATTTAGGTGACACTATAGGG	r_F11	ACTGAGCTGCGATAGTACTCTGCGTTGATACCACTGCTT
f_G11	AAGCAGTGGTATCAACGCAGAGTGTATAGACAGATGTGCCATACGATTTAGGTGACACTATAGGG	r_G11	TAGTATATCGAGACTCACTCTGCGTTGATACCACTGCTT
f_H11	AAGCAGTGGTATCAACGCAGAGTGCACTCAGAGACGCGACATACGATTTAGGTGACACTATAGGG	r_H11	GTGTGTACACATGACAACTCTGCGTTGATACCACTGCTT
f_A12	AAGCAGTGGTATCAACGCAGAGTCACGACTATATGCTCTCATACGATTTAGGTGACACTATAGGG	r_A12	AGTGTATATCGACAGAACTCTGCGTTGATACCACTGCTT
f_B12	AAGCAGTGGTATCAACGCAGAGTAGTCACACGCACGCTGCATACGATTTAGGTGACACTATAGGG	r_B12	TCTCTATAGAGCACGTACTCTGCGTTGATACCACTGCTT
f_C12	AAGCAGTGGTATCAACGCAGAGTCATACATCGCGCAGTACATACGATTTAGGTGACACTATAGGG	r_C12	GCATGTCGTGCTGATAACTCTGCGTTGATACCACTGCTT
f_D12	AAGCAGTGGTATCAACGCAGAGTCTATCTAGCACTCACACATACGATTTAGGTGACACTATAGGG	r_D12	CTGATATCGTGAGAGCACTCTGCGTTGATACCACTGCTT
f_E12	AAGCAGTGGTATCAACGCAGAGTACACGTGATAGCTACGCATACGATTTAGGTGACACTATAGGG	r_E12	TAGCTCACACATGAGAACTCTGCGTTGATACCACTGCTT
f_F12	AAGCAGTGGTATCAACGCAGAGTCGCTAGAGATCTGCTACATACGATTTAGGTGACACTATAGGG	r_F12	CTCGCGACACATCTGAACTCTGCGTTGATACCACTGCTT
f_G12	AAGCAGTGGTATCAACGCAGAGTGATACTGACACACTATCATACGATTTAGGTGACACTATAGGG	r_G12	GAGTGTCTGAGCTACGACTCTGCGTTGATACCACTGCTT
f_H12	AAGCAGTGGTATCAACGCAGAGTGAGCTGATGTACATGTCATACGATTTAGGTGACACTATAGGG	r_H12	CTCGTAGTGTCTCTGAACTCTGCGTTGATACCACTGCTT

Table S4. SMRT cells sequenced

Probe Panel	RNA source	SMRT cell	Movie name
unenriched	fetal brain	m160213_030055_42134_c100957952550000001823213806221632_s1_p0	IsoSeq_FS_JU-movie1
unenriched	fetal brain	m160213_030055_42134_c100957952550000001823213806221632_s1_p0	IsoSeq_FS_JU-movie2
unenriched	fetal brain	m160213_030055_42134_c100957952550000001823213806221632_s1_p0	IsoSeq_FS_JU-movie3
unenriched	fetal brain	m160213_153305_42134_c100957952550000001823213806221634_s1_p0	IsoSeq_FC_JU-movie1
unenriched	fetal brain	m160213_153305_42134_c100957952550000001823213806221634_s1_p0	IsoSeq_FC_JU-movie2
unenriched	fetal brain	m160213_153305_42134_c100957952550000001823213806221634_s1_p0	IsoSeq_FC_JU-movie3
unenriched	adult brain	m160213_091647_42134_c100957952550000001823213806221633_s1_p0	IsoSeq_AS_JU-movie1
unenriched	adult brain	m160213_091647_42134_c100957952550000001823213806221633_s1_p0	IsoSeq_AS_JU-movie2
unenriched	adult brain	m160213_091647_42134_c100957952550000001823213806221633_s1_p0	IsoSeq_AS_JU-movie3
unenriched	adult brain	m160213_214921_42134_c100957952550000001823213806221635_s1_p0	IsoSeq_AC_JU-movie1
unenriched	adult brain	m160213_214921_42134_c100957952550000001823213806221635_s1_p0	IsoSeq_AC_JU-movie2
unenriched	adult brain	m160213_214921_42134_c100957952550000001823213806221635_s1_p0	IsoSeq_AC_JU-movie3
HSD1	adult brain	m160614_013904_42139_c100966090630000001823217407061600_s1_p0	adult_brain_elf_6-A-movie1
HSD1	adult brain	m160614_013904_42139_c100966090630000001823217407061600_s1_p0	adult_brain_elf_6-A-movie2
HSD1	adult brain	m160614_013904_42139_c100966090630000001823217407061600_s1_p0	adult_brain_elf_6-A-movie3
HSD1	adult brain	m160614_075731_42139_c100966090630000001823217407061601_s1_p0	adult_brain_elf_7-A-movie1
HSD1	adult brain	m160614_075731_42139_c100966090630000001823217407061601_s1_p0	adult_brain_elf_7-A-movie2
HSD1	adult brain	m160614_075731_42139_c100966090630000001823217407061601_s1_p0	adult_brain_elf_7-A-movie3
HSD1	adult brain	m160614_141925_42139_c100966090630000001823217407061602_s1_p0	adult_brain_elf_8-A-movie1
HSD1	adult brain	m160614_141925_42139_c100966090630000001823217407061602_s1_p0	adult_brain_elf_8-A-movie2
HSD1	adult brain	m160614_141925_42139_c100966090630000001823217407061602_s1_p0	adult_brain_elf_8-A-movie3
HSD1	adult brain	m160614_203953_42139_c100966090630000001823217407061603_s1_p0	adult_brain_elf_9-A-movie1
HSD1	adult brain	m160614_203953_42139_c100966090630000001823217407061603_s1_p0	adult_brain_elf_9-A-movie2
HSD1	adult brain	m160614_203953_42139_c100966090630000001823217407061603_s1_p0	adult_brain_elf_9-A-movie3
HSD1	adult brain	m160226_223848_42134_c100969612550000001823215707061691_s1_p0	adult_brain_amp_0.4x-A-movie1
HSD1	adult brain	m160226_223848_42134_c100969612550000001823215707061691_s1_p0	adult_brain_amp_0.4x-A-movie2
HSD1	adult brain	m160226_223848_42134_c100969612550000001823215707061691_s1_p0	adult_brain_amp_0.4x-A-movie3
HSD1	adult brain	m160410_030309_42134_c100991602550000001823221607191632_s1_p0	adult_brain_amp_0.4x-B-movie1
HSD1	adult brain	m160410_030309_42134_c100991602550000001823221607191632_s1_p0	adult_brain_amp_0.4x-B-movie2
HSD1	adult brain	m160410_030309_42134_c100991602550000001823221607191632_s1_p0	adult_brain_amp_0.4x-B-movie3
HSD1	adult brain	m160410_091947_42134_c100991602550000001823221607191633_s1_p0	adult_brain_amp_0.4x-C-movie1
HSD1	adult brain	m160410_091947_42134_c100991602550000001823221607191633_s1_p0	adult_brain_amp_0.4x-C-movie2
HSD1	adult brain	m160410_091947_42134_c100991602550000001823221607191633_s1_p0	adult_brain_amp_0.4x-C-movie3
HSD1	adult brain	m160410_153941_42134_c100991602550000001823221607191634_s1_p0	adult_brain_amp_0.4x-D-movie1
HSD1	adult brain	m160410_153941_42134_c100991602550000001823221607191634_s1_p0	adult_brain_amp_0.4x-D-movie2
HSD1	adult brain	m160410_153941_42134_c100991602550000001823221607191634_s1_p0	adult_brain_amp_0.4x-D-movie3
HSD1	adult brain	m160511_012116_42134_c100977902550000001823223308031671_s1_p0	adult_brain_elf_78-B-movie1
HSD1	adult brain	m160511_012116_42134_c100977902550000001823223308031671_s1_p0	adult_brain_elf_78-B-movie2
HSD1	adult brain	m160511_012116_42134_c100977902550000001823223308031671_s1_p0	adult_brain_elf_78-B-movie3
HSD1	adult brain	m160416_210737_42134_c100991652550000001823221607191683_s1_p0	adult_brain_elf_78-A-movie1
HSD1	adult brain	m160416_210737_42134_c100991652550000001823221607191683_s1_p0	adult_brain_elf_78-A-movie2
HSD1	adult brain	m160416_210737_42134_c100991652550000001823221607191683_s1_p0	adult_brain_elf_78-A-movie3
HSD1	adult brain	m160226_100605_42134_c100946382550000001823214306251677_s1_p0	adult_brain_amp_0.6x-movie1
HSD1	adult brain	m160226_100605_42134_c100946382550000001823214306251677_s1_p0	adult_brain_amp_0.6x-movie2
HSD1	adult brain	m160226_100605_42134_c100946382550000001823214306251677_s1_p0	adult_brain_amp_0.6x-movie3
HSD1	adult brain	m160619_122419_42134_c101009602550000001823230710211661_s1_p0	adult_brain_elf_6-B-movie1
HSD1	adult brain	m160619_122419_42134_c101009602550000001823230710211661_s1_p0	adult_brain_elf_6-B-movie2
HSD1	adult brain	m160619_122419_42134_c101009602550000001823230710211661_s1_p0	adult_brain_elf_6-B-movie3
HSD1	adult brain	m160621_233748_42134_c101009382550000001823230410211661_s1_p0	adult_brain_elf_7-B-movie1
HSD1	adult brain	m160621_233748_42134_c101009382550000001823230410211661_s1_p0	adult_brain_elf_7-B-movie2
HSD1	adult brain	m160621_233748_42134_c101009382550000001823230410211661_s1_p0	adult_brain_elf_7-B-movie3
HSD1	adult brain	m160622_143919_42134_c101009382550000001823230410211664_s1_p0	adult_brain_elf_8-B-movie1
HSD1	adult brain	m160622_143919_42134_c101009382550000001823230410211664_s1_p0	adult_brain_elf_8-B-movie2
HSD1	adult brain	m160622_143919_42134_c101009382550000001823230410211664_s1_p0	adult_brain_elf_8-B-movie3
HSD1	adult brain	m160701_124807_42134_c101025152550000001823232111041602_s1_p0	adult_brain_elf_9-B-movie1
HSD1	adult brain	m160701_124807_42134_c101025152550000001823232111041602_s1_p0	adult_brain_elf_9-B-movie2
HSD1	adult brain	m160701_124807_42134_c101025152550000001823232111041602_s1_p0	adult_brain_elf_9-B-movie3
HSD1	fetal brain	m160615_030030_42139_c100966090630000001823217407061604_s1_p0	fetal_brain_amp_0.4x-B-movie1
HSD1	fetal brain	m160615_030030_42139_c100966090630000001823217407061604_s1_p0	fetal_brain_amp_0.4x-A-movie1
HSD1	fetal brain	m160615_030030_42139_c100966090630000001823217407061604_s1_p0	fetal_brain_amp_0.4x-C-movie1
HSD1	fetal brain	m160615_092105_42139_c100966090630000001823217407061605_s1_p0	fetal_brain_amp_0.4x-D-movie1
HSD1	fetal brain	m160615_092105_42139_c100966090630000001823217407061605_s1_p0	fetal_brain_elf_78-B-movie1
HSD1	fetal brain	m160615_092105_42139_c100966090630000001823217407061605_s1_p0	fetal_brain_elf_78-A-movie1
HSD1	fetal brain	m160615_154414_42139_c100961261270000001823212707011620_s1_p0	fetal_brain_amp_0.6x-movie1
HSD1	fetal brain	m160615_154414_42139_c100961261270000001823212707011620_s1_p0	fetal_brain_elf_6-A-movie1
HSD1	fetal brain	m160615_154414_42139_c100961261270000001823212707011620_s1_p0	fetal_brain_elf_6-B-movie1
HSD1	fetal brain	m160615_220452_42139_c100961261270000001823212707011621_s1_p0	fetal_brain_elf_7-A-movie1

HSD1	fetal brain	m160615_220452_42139_c100961261270000001823212707011621_s1_p0	fetal_brain_elf_7-B-movie1
HSD1	fetal brain	m160615_220452_42139_c100961261270000001823212707011621_s1_p0	fetal_brain_elf_8-A-movie1
HSD1	fetal brain	m160226_034933_42134_c100946382550000001823214306251676_s1_p0	fetal_brain_elf_8-B-movie1
HSD1	fetal brain	m160226_034933_42134_c100946382550000001823214306251676_s1_p0	fetal_brain_elf_9-A-movie1
HSD1	fetal brain	m160226_034933_42134_c100946382550000001823214306251676_s1_p0	fetal_brain_elf_9-B-movie1
HSD1	fetal brain	m160226_162210_42134_c100969612550000001823215707061690_s1_p0	fetal_brain_amp_0.4x-B-movie2
HSD1	fetal brain	m160226_162210_42134_c100969612550000001823215707061690_s1_p0	fetal_brain_amp_0.4x-A-movie2
HSD1	fetal brain	m160226_162210_42134_c100969612550000001823215707061690_s1_p0	fetal_brain_amp_0.4x-C-movie2
HSD1	fetal brain	m160409_081409_42134_c100991582550000001823221607191687_s1_p0	fetal_brain_amp_0.4x-D-movie2
HSD1	fetal brain	m160409_081409_42134_c100991582550000001823221607191687_s1_p0	fetal_brain_elf_78-B-movie2
HSD1	fetal brain	m160409_081409_42134_c100991582550000001823221607191687_s1_p0	fetal_brain_elf_78-A-movie2
HSD1	fetal brain	m160409_143034_42134_c100991602550000001823221607191630_s1_p0	fetal_brain_amp_0.6x-movie2
HSD1	fetal brain	m160409_143034_42134_c100991602550000001823221607191630_s1_p0	fetal_brain_elf_6-A-movie2
HSD1	fetal brain	m160409_143034_42134_c100991602550000001823221607191630_s1_p0	fetal_brain_elf_6-B-movie2
HSD1	fetal brain	m160409_204649_42134_c100991602550000001823221607191631_s1_p0	fetal_brain_elf_7-A-movie2
HSD1	fetal brain	m160409_204649_42134_c100991602550000001823221607191631_s1_p0	fetal_brain_elf_7-B-movie2
HSD1	fetal brain	m160409_204649_42134_c100991602550000001823221607191631_s1_p0	fetal_brain_elf_8-A-movie2
HSD1	fetal brain	m160416_145117_42134_c100991652550000001823221607191682_s1_p0	fetal_brain_elf_8-B-movie2
HSD1	fetal brain	m160416_145117_42134_c100991652550000001823221607191682_s1_p0	fetal_brain_elf_9-A-movie2
HSD1	fetal brain	m160416_145117_42134_c100991652550000001823221607191682_s1_p0	fetal_brain_elf_9-B-movie2
HSD1	fetal brain	m160510_190442_42134_c100977902550000001823223308031670_s1_p0	fetal_brain_amp_0.4x-B-movie3
HSD1	fetal brain	m160510_190442_42134_c100977902550000001823223308031670_s1_p0	fetal_brain_amp_0.4x-A-movie3
HSD1	fetal brain	m160510_190442_42134_c100977902550000001823223308031670_s1_p0	fetal_brain_amp_0.4x-C-movie3
HSD1	fetal brain	m160619_061122_42134_c101009602550000001823230710211660_s1_p0	fetal_brain_amp_0.4x-D-movie3
HSD1	fetal brain	m160619_061122_42134_c101009602550000001823230710211660_s1_p0	fetal_brain_elf_78-B-movie3
HSD1	fetal brain	m160619_061122_42134_c101009602550000001823230710211660_s1_p0	fetal_brain_elf_78-A-movie3
HSD1	fetal brain	m160618_213601_42134_c101009522550000001823230710211677_s1_p0	fetal_brain_amp_0.6x-movie3
HSD1	fetal brain	m160618_213601_42134_c101009522550000001823230710211677_s1_p0	fetal_brain_elf_6-A-movie3
HSD1	fetal brain	m160618_213601_42134_c101009522550000001823230710211677_s1_p0	fetal_brain_elf_6-B-movie3
HSD1	fetal brain	m160701_001557_42134_c101025152550000001823232111041600_s1_p0	fetal_brain_elf_7-A-movie3
HSD1	fetal brain	m160701_001557_42134_c101025152550000001823232111041600_s1_p0	fetal_brain_elf_7-B-movie3
HSD1	fetal brain	m160701_001557_42134_c101025152550000001823232111041600_s1_p0	fetal_brain_elf_8-A-movie3
HSD1	fetal brain	m160701_063435_42134_c101025152550000001823232111041601_s1_p0	fetal_brain_elf_8-B-movie3
HSD1	fetal brain	m160701_063435_42134_c101025152550000001823232111041601_s1_p0	fetal_brain_elf_9-A-movie3
HSD1	fetal brain	m160701_063435_42134_c101025152550000001823232111041601_s1_p0	fetal_brain_elf_9-B-movie3
HSD2	fetal brain	m170527_015245_42134_c101190302550000001823271509291792_s1_p0	Fetal_HSD2_.2-movie1
HSD2	fetal brain	m170527_015245_42134_c101190302550000001823271509291792_s1_p0	Fetal_HSD2_.2-movie2
HSD2	fetal brain	m170527_015245_42134_c101190302550000001823271509291792_s1_p0	Fetal_HSD2_.2-movie3
HSD2	adult brain	m170527_081100_42134_c101190302550000001823271509291793_s1_p0	Adult_HSD2_.2-movie1
HSD2	adult brain	m170527_081100_42134_c101190302550000001823271509291793_s1_p0	Adult_HSD2_.2-movie2
HSD2	adult brain	m170527_081100_42134_c101190302550000001823271509291793_s1_p0	Adult_HSD2_.2-movie3
HSD2	fetal brain	m170507_073233_42134_c101190152550000001823271509291706_s1_p0	fetal_hsd2_minus-movie1
HSD2	fetal brain	m170507_073233_42134_c101190152550000001823271509291706_s1_p0	fetal_hsd2_minus-movie2
HSD2	fetal brain	m170507_073233_42134_c101190152550000001823271509291706_s1_p0	fetal_hsd2_minus-movie3
HSD2	fetal brain	m170507_135004_42134_c101190152550000001823271509291707_s1_p0	fetal_hsd2_plus-movie1
HSD2	fetal brain	m170507_135004_42134_c101190152550000001823271509291707_s1_p0	fetal_hsd2_plus-movie2
HSD2	fetal brain	m170507_135004_42134_c101190152550000001823271509291707_s1_p0	fetal_hsd2_plus-movie3
HSD2	adult brain	m170507_200337_42134_c101190172550000001823271509291780_s1_p0	adult_hsd2_minus-movie1
HSD2	adult brain	m170507_200337_42134_c101190172550000001823271509291780_s1_p0	adult_hsd2_minus-movie2
HSD2	adult brain	m170507_200337_42134_c101190172550000001823271509291780_s1_p0	adult_hsd2_minus-movie3
HSD2	adult brain	m170508_043407_42134_c101190172550000001823271509291781_s1_p0	adult_hsd2_plus-movie1
HSD2	adult brain	m170508_043407_42134_c101190172550000001823271509291781_s1_p0	adult_hsd2_plus-movie2
HSD2	adult brain	m170508_043407_42134_c101190172550000001823271509291781_s1_p0	adult_hsd2_plus-movie3

Table S5. Summary sequence statistics

Probe panel	RNA source	SMRT cells sequenced	CCSs	putative full-length reads	on target rate*	enrichment
Unenriched	adult brain	2	92906	84348	0.122%, 0.058%	n/a
Unenriched	fetal brain	2	91908	85597	0.080%, 0.052%	n/a
HSD1	adult brain	15	572266	468425	70.8%	580x
HSD1	fetal brain	15	488252	376375	64.9%	811x
HSD2	adult brain	3	107364	102944	88.2%	1521x
HSD2	fetal brain	3	81394	77787	90.2%	1735x

*mapped putative full-length reads; for unenriched samples, percentages refer to HSD1 and HSD2, respectively

CCS: circular consensus sequence

Table S6. Analyzed gene set

Gene Family	Paralog	Status				Features							Counts	
Family	Gene	Duplication type	Segment of gene duplicated	Other names	State [Note a]	Pseudo (currently annotated)	New ORF	5' extension/alt	3' extension/alt	Segdup Fusion	Gene Fusion	Fused with	pFL reads	pFL reads (no mapq filt)
SRGAP2	SRGAP2A	partial	-	SRGAP2	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	9642	26433
SRGAP2	SRGAP2B	partial	5'		dup	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	.	5921	19686
SRGAP2	SRGAP2C	partial	5'		dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	8146	19407
SRGAP2	SRGAP2D	partial	5'		dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	2561	15112
FAM72	FAM72A	complete	-	FAM72A	anc	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	.	508	1746
FAM72	FAM72B	complete	whole	FAM72B	dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	565	1795
FAM72	FAM72C	complete	whole	FAM72C	dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	357	1695
FAM72	FAM72D	complete	whole	FAM72D	dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	373	1797
HYDIN	HYDIN	partial	-	HYDIN	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	2573	10123
HYDIN	HYDIN2	partial	mid	HYDIN2	dup	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	NBPF12	9697	11722
ROCK1	ROCK1	partial	-	ROCK1	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	2253	2833
ROCK1	ROCK1P	partial	3'	ROCK1P1	dup	TRUE	true (?)	TRUE	FALSE	TRUE	FALSE	iRNA MIR807	652	1955
ARHGAP11	ARHGAP11A	partial	-	ARHGAP11A	anc	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	SCG5	1541	1736
ARHGAP11	ARHGAP11B	partial	5'	ARHGAP11B	dup	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	.	276	731
PTPN20	PTPN20A	partial	-	PTPN20	t appears an	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	2651	2820
PTPN20	PTPN20B	partial	5'	PTPN20CP	t appears du	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	BMS1P7	467	657
FRMPD2	FRMPD2A	partial	-	FRMPD2	anc	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	.	6753	18382
FRMPD2	FRMPD2B	partial	3'	FRMPD2B	dup	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	.	10964	17700
GTF2I	GTF2IBc	partial	-	GTF2I	anc	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	.	31444	103278
GTF2I	GTF2IBm	partial	3'	GTF2IP1;	dup	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	.	8197	125488
GTF2I	GTF2IBt	partial	3'	GTF2IP4;	dup	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	.	49266	116083
GTF2IRD2	GTF2IRD2Bc	mixed	-	GTF2IRD2	anc	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	STAG3L2	1105	2581
GTF2IRD2	GTF2IRD2Bm	mixed	whole	GTF2IRD2I	dup	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	.	1246	2505
GTF2IRD2	GTF2IRD2Bt	mixed	3'	GTF2IRD2I	dup	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	.	20	1091
NCF	NCFBc	complete	-	NCF1	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	261	578
NCF	NCFBm	complete	whole	NCF1C	dup	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	.	172	615
NCF	NCFBt	complete	whole	NCF1B	dup	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	.	188	630
GPR89	GPR89C_cent	complete	-	GPR89A	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	15641	27991
GPR89	GPR89_telo	complete	whole	GPR89B	dup	FALSE	FALSE	FALSE	false (?)	false (?)	FALSE	na	12195	27871
CD8B	CD8B_p	complete	-	CD8B	anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	104	633
CD8B	CD8B_q	complete	5' for minor i	CD8BP	dup	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	.	550	644
NOTCH2	NOTCH2	partial	-	NOTCH2NL	ily not from	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	.	3281	5292
NOTCH2	NOTCH2NLA	partial	5'		dup	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	NBPF10	580	5028
NOTCH2	NOTCH2NLB	partial	5'		dup	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	NBPF14	654	4794
NOTCH2	NOTCH2NLC	partial	5'		dup	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	NBPF18	291	4104
NOTCH2	NOTCH2NLD	partial	5'		dup	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	NBPF26	330	4494
BOLA2	BOLA2_cent	complete	whole	BOLA2_cer	unk	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	309	1473
BOLA2	BOLA2_telo	complete	whole	BOLA2_tel	unk	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	SMG1P6	16556	17989
CORO1A	CORO1A_cent	partial	-	CORO1A	ily not from	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	17653	17745
CORO1A	CORO1A_telo	partial	3'	CORO1A_ur	dup	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	na	25	26
SLX1A	SLX1A_cent	complete	whole	SLX1A	t from Denni	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	SULT1A3	412	1889
SLX1A	SLX1A_telo	complete	whole	SLX1B	t from Denni	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	SULT1A4	1390	3062
FCGR1A	FCGR1A	complete	-		anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	4857	6953
FCGR1A	FCGR1B	complete	whole		dup	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	1580	7066
FCGR1A	FCGR1C	complete	whole	FCGR1CP	dup	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	na	588	3140
ARHGEF5	ARHGEF34P	partial	5'		dup	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	OR2A7	69	128
ARHGEF5	ARHGEF35	partial	5' (shorter)		dup	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	na	76	98
ARHGEF5	ARHGEF5	partial	-		anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	195	273
CHRNA7	CHRFAM7A	partial	3'		dup	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	na	1368	3654
CHRNA7	CHRNA7	partial	-		anc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	na	2666	4287

Note a: Taken from Dennis 2017 Table S9 unless otherwise specified

Pseudo: Previously annotated only as pseudogene
5' extension/alt: 5' extension of transcript or alternate promoter
3' extension/alt: 3' extension of transcript or alternate polyadenylation
Segdup Fusion: Transcript extends into another segdup
Gene Fusion: Transcript fuses with another annotated gene

Table S7. Fusions vs. truncations in 3' truncated genes

Gene Family	Gene	Type of truncation	Type of extension	Proportion of duplication-spanning reads	Designation
NOTCH2	NOTCH2NLC	Loss of PAS	3'	0.25	intermediate
NOTCH2	NOTCH2NLD	Loss of PAS	3'	0.38	intermediate
ARHGEF5	ARHGEF34P	Loss of PAS	3'	0.45	intermediate
NOTCH2	NOTCH2NLB	Loss of PAS	3'	0.49	intermediate
HYDIN	HYDIN2	Loss of PAS	3'	0.69	intermediate
ARHGAP11	ARHGAP11B	Loss of PAS	3'	0.86	primarily fusion
ANAPC1	ANAPC1 (3)	Loss of PAS	3'	0.87	primarily fusion
PTPN20	PTPN20B	Loss of PAS	3'	0.88	primarily fusion
ANAPC1	ANAPC1 (1)	Loss of PAS	3'	0.89	primarily fusion
ARHGEF5	AHRGEF35	Loss of PAS	3'	0.93	primarily fusion
SRGAP2	SRGAP2C	Loss of PAS	3'	0.05	primarily truncation
SRGAP2	SRGAP2D	Loss of PAS	3'	0.10	primarily truncation
SRGAP2	SRGAP2B	Loss of PAS	3'	0.14	primarily truncation
NOTCH2	NOTCH2NLA	Loss of PAS	3'	0.16	primarily truncation
ANAPC1	ANAPC1 (1)	Loss of TSS	5'	0.34	intermediate
ROCK1	ROCK1P	Loss of TSS	5'	0.46	intermediate
CHRNA7	CHRFAM7A	Loss of TSS	5'	0.55	intermediate
GTF2I	GTF2IP4	Loss of TSS	5'	0.67	intermediate
GTF2I	GTF2IP1	Loss of TSS	5'	0.68	intermediate
HYDIN	HYDIN2	Loss of TSS	5'	0.90	primarily fusion
FRMPD2	FRMPD2B	Loss of TSS	5'	0.04	primarily truncation
ANAPC1	ANAPC1 (3)	Loss of TSS	5'	0.10	primarily truncation

Table S8. Quantification of isoform abundance for Tigger-derived exon

	Sequencing		*GTF2I*			*GTF2IRD2**			*GTF2IRD2*			*GTF2IRD2B*			
Data source	**technology**	**RNA source**		alt exon	exon 1		alt exon	exon 1		alt exon	exon 1		alt exon	exon 1	
Transcript Capture (a)	PB	human brain	16%	48	261	92%	403	34	92%	154	14	92%	251	21	Counts of long-read transcripts
Data release (b)	ONT	GM12878 cells	3%	6	214	0%	0	102		—			—		
(c)	PB	human iPS cells	80%	4	1		nd			nd			nd		
(c)	PB	chimp iPS cells	83%	5	1		nd			nd			nd		
(c)	PB	gorilla iPS cells	0%	0	2		nd			nd			nd		
(c)	I	human iPS cells		10 (1)	188 (14)		—			0 (1)	0 (12)		0 (1)	0 (12)	Counts of splice junction-spanning reads (d)
(c)	I	chimp iPS cells		0 (0)	108 (12)		—			0 (2)	0 (5)		0 (2)	0 (5)	
(c)	I	gorilla iPS cells		0 (0)	122 (12)		—			0 (0)	0 (0)		0 (0)	0 (0)	
(c)	I	adult brain		1 (0)	58 (2)		—			0 (4)	1 (33)		1 (4)	0 (33)	
(c)	I	fetal brain		0 (0)	58 (1)		—			0 (6)	0 (20)		0 (6)	1 (20)	
GTEx	I	brain - cortex		6 (0)	472 (10)		—			1 (97)	4 (199)		0 (97)	16 (199)	
GTEx	I	brain - cortex	2%	0.43	25.69		—		27%	0.65	1.80	9%	0.49	4.89	In silico expression estimates by Kallisto (tpm)
GTEx	I	brain - cerbellum	1%	0.47	41.92		—		41%	1.61	2.35	20%	2.14	8.66	

* Without discrimination between GTF2IRD2 and GTF2IRD2B

PB = PacBio; ONT = Oxford Nanopore; I = Illumina; nd = no data

a) Sum of counts of putative full-length reads from adult and developing brain

b) Direct RNA sequencing; https://github.com/nanopore-wgs-consortium/NA12878

c) Kronenberg et al. 2018 *in press*

d) uniquely mapping reads (multimap reads)

VITA

Max L. Dougherty (born 1988, New York, NY) grew up in Staten Island, NY, and attended Poly Prep High School in Brooklyn, NY. He graduated with an A.B. in Chemistry from Harvard University in Cambridge, MA, in 2010 where he studied inorganic chemistry in the lab of Ted Betley, and during which time he also studied the evolution of recently transmitted HIV in Gaborone, Botswana, under Max Essex. He then pursued research on the genetics of cleft lip and palate in the lab of Eric Liao at Massachusetts General Hospital in Boston, MA, before entering the University of Washington Medical Scientist Training Program in 2012. He joined the lab of Evan Eichler in the autumn of 2014 as a member of the Department of Genome Sciences. Upon completion of his PhD, Max plans to return to medical school.