NOVEL METHODS IMPROVE GENOME ANNOTATION

by

Markus Sommer

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

August, 2023

Abstract

In the era of high-throughput sequencing, comprehensive genome annotation has become critical for understanding the functional complexities of life. Here we explore the development and application of computational methods for genome annotation, focusing on human transcriptome analysis, prokaryotic gene prediction, and circular RNA annotation. The primary contributions of this work span three distinct areas, each addressing unique challenges in the field. First, we develop a structure-guided isoform identification approach that utilizes three-dimensional protein structure predictions to identify functional human gene isoforms. Our method evaluates over 230,000 isoforms of human protein-coding genes assembled from thousands of RNA sequencing experiments across various human tissues. We identify hundreds of isoforms with more confidently predicted structure and potentially superior function compared to canonical isoforms, thus demonstrating the potential of protein structure prediction as a powerful tool for genome annotation and transcriptome analysis. Second, we present a universal protein model for prokaryotic gene prediction, Balrog, which employs a temporal convolutional network to analyze amino acid sequences from a diverse set of microbial genomes. Balrog eliminates the need for genome-specific training and matches or outperforms existing state-of-the-art gene finding tools. Lastly, we introduce alignment-free methods for annotating circular RNA in humans, leveraging a simple k-mer-based data structure.

Committee: Steven Salzberg (Primary Advisor), Mihaela Pertea, Martin Steinegger

Table of Contents

List of Tables

Chapter 2

List of Figures

Chapter 1

A remark regarding collaboration: Much of the research described in this thesis has been performed in close cooperation with other researchers. Although I strive to emphasize my personal input, I have incorporated some findings and written content from my collaborators and mentors to maintain context and thoroughness. In each section, I identify and acknowledge the individuals who contributed to the work. I am deeply appreciative of the numerous exceptional scientists I have had the pleasure of collaborating with during my PhD journey.

Chapter 1: Structure-guided isoform identification for the human transcriptome

A version of chapter 1 has appeared in:

M. J. Sommer, S. Cha, A. Varabyou, N. Rincon, S. Park, I. Minkin, M. Pertea, M. Steinegger, S. L. Salzberg, Structure-guided isoform identification for the human transcriptome. *eLife*. 11 (2022), doi:10.7554/eLife.82556.

1.1 Introduction

More than twenty years after the initial publication the human genome, the scientific community is still trying to determine the complete set of human protein-coding genes. Although the number of genes is converging around 20,000, we do not yet have agreement on the precise number. The true number of different isoforms of human genes – variations due to alternative splicing, alternative transcription initiation sites,

and alternative transcription termination sites – is even less certain. Currently, the major human gene annotation databases each contain well over 100,000 protein-coding transcripts [1–5], but the sets of transcripts vary widely among them. The disagreement between human transcriptome databases was clearly demonstrated when, in 2018, GENCODE and RefSeq were shown to agree on fewer than 50,000 of the nearly 300,000 total transcripts in their human annotations [4].

Although the functions of many human genes are known, elucidating gene function remains a complex and time-consuming task. Given that at least 92% of human genes express more than one isoform [6], and that the human transcriptome contains an average of seven or more unique transcripts per protein-coding gene [7], the only feasible way to determine which isoforms are functional on a genome-wide scale is by using computational methods. Until now, the primary tools used to investigate gene function were sequence alignment and gene expression. Alignment relies on the long-established observation that if a protein is conserved in other species, then it is likely to be functional, particularly if the conservation extends to distantly-related species [8]. This rule applies to isoforms as well: if we can find evidence that a particular sequence – e.g., a protein that uses an alternative exon – is present in species that diverged tens of millions of years ago, then the conservation of the sequence argues in favor of its function.

In a similar vein, the use of RNA sequencing (RNA-seq) to detect gene expression also provides clues to function: if a transcript is consistently expressed in multiple samples, it

is more likely to be functional than one for which little expression evidence can be found. Genes may encode multiple transcripts that fold into distinct isoforms with well-defined functions [6], but recent work has shown that most assembled human transcripts are found at very low levels in the transcriptomes of individual tissues [4], and may simply reflect biological noise, products of intrinsically stochastic biochemical reactions [9,10]. The large majority of assembled isoforms are unlikely to be functional, and indeed only a small percentage are included in current human genome annotation databases [1–5]. Because transcription is noisy, the observation of transcription in RNA-seq data is insufficient evidence to conclude that a sequence is functional [11].

This study explores a fundamentally new line of evidence that can be used to investigate protein function: computational prediction of three-dimensional structure. The recently-developed AlphaFold2 system can automatically predict three-dimensional protein structure with accuracy that often matches far more time-consuming laboratory methods [12,13], allowing us to generate structure predictions for thousands of gene isoforms. In proteins where a substantial portion folds into an ordered structure, estimated to be 68% of human proteins [13,14], a well-folded structure within an isoform argues in favor of its functionality. Conversely, a poorly-folded isoform may indicate loss of function.

In a recent effort to create a single consensus annotation of all human protein-coding genes, two of the leading human genome annotation centers created the MANE (Matched Annotation from NCBI and EMBL-EBI) database [15], a high-quality collection of

protein-coding isoforms for which the annotation databases RefSeq (NCBI) and Ensembl-GENCODE (EMBL) match precisely. The goal of MANE is to identify just one isoform for each protein-coding gene that is well-supported by experimental data, and to ensure that both databases agree on all exon boundaries as well as the sequence of the associated protein. In addition to the one-isoform-per-gene collection known as MANE Select, a small number of additional transcripts with special clinical significance, known as MANE Plus Clinical, are included in the database. Upon its initial release, MANE included only around 50% of human protein-coding genes. The latest version, v1.0, includes 19,062 genes and 19,120 transcripts, with an additional 58 transcripts included in the MANE Plus Clinical set. These transcripts have been described as a "universal standard" for human gene annotation, and they provide a valuable resource to scientists and clinicians who need a consistent set of functional primary transcripts.

Here we describe our use of protein folding predictions from ColabFold [16], an open-source accelerated version of AlphaFold2, alongside experimental RNA-seq expression data from the Genotype-Tissue Expression project (GTEx) [17], to present substantial evidence for functional isoforms that can be used to improve human gene annotation, including the MANE gene set as well as the comprehensive human annotation databases RefSeq [2], GENCODE [3], and the Comprehensive Human Expressed SequenceS database (CHESS) [4]. We dive into a few exemplary predictions to explain, biologically and evolutionarily, the three-dimensional structure of our alternate isoforms. We also present an example of a novel protein isoform in mouse to demonstrate the

general applicability of this structure-guided approach to improving functional annotation of any genome.

1.2 Results

Scoring the transcriptome:

Using CHESS, a large set of transcripts assembled from nearly 10,000 human RNA-sequencing experiments, we identified all protein-coding gene isoforms that were 1000aa or less in length (see Methods). The 233,973 transcripts at 20,666 gene loci that fit this description encoded 127,398 distinct protein isoforms, and we predicted structures for all of them. Additionally, we included 3302 structure predictions for proteins with length >1000aa from the AlphaFold Protein Structure Database [18] for isoforms with an exact protein sequence match in CHESS 3. This resulted in a total of 237,295 transcripts at 20,817 gene loci encoding 130,700 distinct protein isoforms. As shown in Fig. 1, we observe no strong trend in the relationship between pLDDT and either protein length or overall transcript expression in the GTEx data. The lack of a clear linear relationship implies protein structure prediction may provide an orthogonal source of useful information for genome annotation efforts.

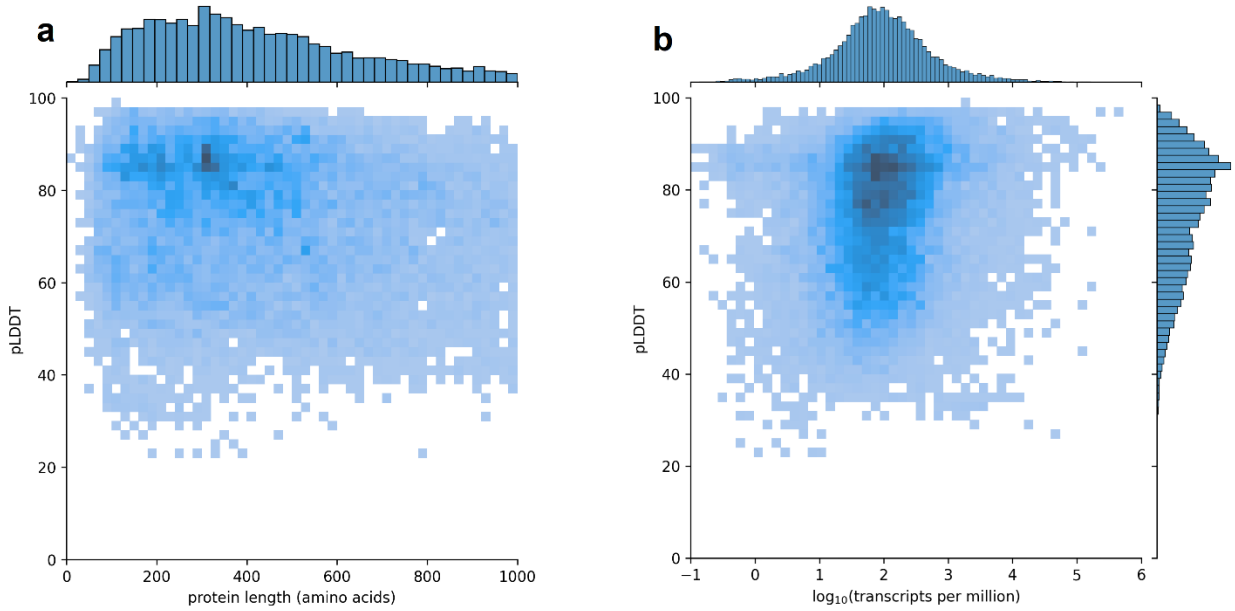

Figure 1: pLDDT distribution across the human transcriptome. 2D joint histograms comparing pLDDT to protein amino-acid length (a) and expression (b) measured in transcripts per million (TPM). For each protein coding gene, only the isoform found in the highest number of GTEx samples is plotted. No strong trend is visible in the relationship between pLDDT and either protein length (a) or transcript expression (b).

In total, 22,644 transcripts (9.7% of all examined transcripts) encoded an isoform that scored a higher pLDDT than the isoform encoded by the corresponding MANE transcript. However, many of these higher-scoring transcripts encoded relatively short, often low-expressed, likely-nonfunctional fragments of larger proteins. Therefore, we incorporated filters based on RNA-seq expression data when determining which higher-scoring isoforms appeared superior to MANE isoforms (see *Filtering MANE comparisons*). Based on the combination of RNA-seq evidence and protein foldability, we identified 940 unique alternate isoforms at 632 loci (3.4% of all MANE loci) which appeared to have a more stable structure than the annotated primary isoform. Data for these 940 alternate isoforms can be found in Supplementary File 3. Worth noting is that

over 96% of the MANE loci evaluated here contained no higher-scoring alternate isoforms that passed our filtering criteria. This is a testament to both the high degree of consistency in MANE as well as the sensitivity of protein structure prediction for finding instances where alternative isoforms may create functional products. Additionally, for 35% of all human protein coding gene loci in our analysis, the most commonly observed isoform scored a pLDDT below 70, suggesting that intrinsic disorder may be an important feature of proteins at these loci.

Gene identifiers for all predicted protein isoforms as well as predicted Local Distance Difference Test (pLDDT) scores and evolutionary conservation data from mouse can be found in Supplementary File 1. Predicted scores and GTEx expression data for all isoforms overlapping a MANE locus can be found in Supplementary File 2. All predicted protein structures as well as data for all tables in this paper are publicly available at our website, isoform.io.

Exemplary predictions:

To illustrate the improvements in human gene annotation that can be obtained using accurate structure prediction, we describe a small set of proteins, selected from Supplementary File 3, where an alternate isoform appears to be superior to the isoform chosen for inclusion in MANE. For these examples, the alternative isoform is clearly functional based on structure as well as evolutionary conservation and, in some cases, additional expression evidence from RNA-sequencing data. For some of these

examples, the MANE isoform is missing critical structural elements and may not be functional at all.

Acetylserotonin O-methyltransferase

Acetylserotonin O-methyltransferase (*ASMT*, alternatively *HIOMT*) is responsible for the final catalytic step in the production of melatonin, a critical hormone in sleep, metabolism, immune response, and neuronal development [19]. Depressed levels of circulating melatonin have been associated with autism spectrum disorder, and clinical studies have classified *ASMT* as a susceptibility gene due to the highly significant association between *ASMT* activity and autism [19,20].

The CHESS and GENCODE gene databases contain a 345aa isoform of *ASMT* (CHS.57426.4, ENST00000381229.9) while RefSeq is missing this isoform. The MANE version of this gene is 373aa long and appears in the CHESS (CHS.57426.2), GENCODE (ENST00000381241.9), and RefSeq (NM_001171038.2) gene databases. The predicted structures of both isoforms are shown in Fig. 2.

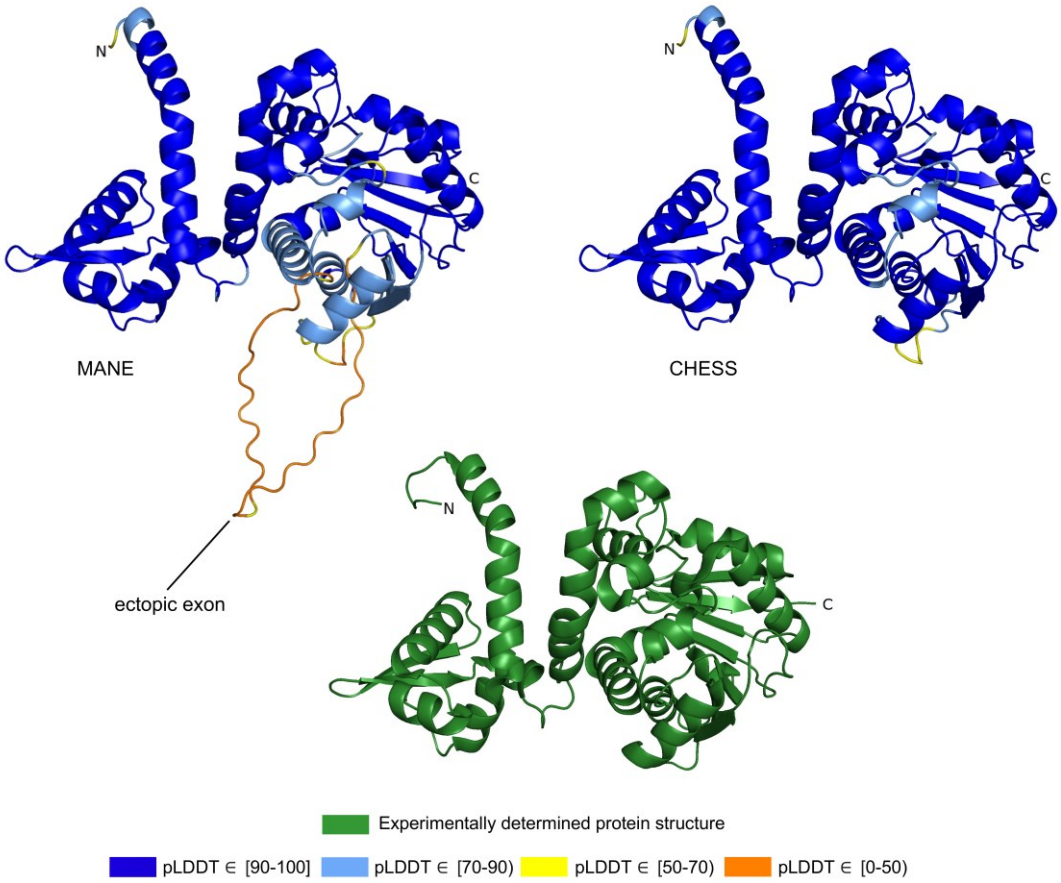

Figure 2: *ASMT* isoform comparison. Comparison of predicted structures of acetylserotonin O-methyltransferase (ASMT), showing the 373aa isoform from MANE (CHS.57426.2, RefSeq NM_001171038.2, GENCODE ENST00000381241.9) on the left, and a 345aa alternate isoform from CHESS (CHS.57426.4, GENCODE ENST00000381229.9) on the right. The CHESS 345aa isoform closely matches the experimentally determined X-ray crystal structure of the biologically active protein [21], shown at the bottom.

We hypothesized that the highest scoring isoform of *ASMT* according to AlphaFold2 corresponds to the biologically active version of the protein. The score we use for these comparisons is the pLDDT score, which has been demonstrated to be a well-calibrated, consistent measure of protein structure prediction accuracy [12,13]. A pLDDT score above 70 (the maximum is 100) indicates that a predicted structure can generally be trusted, while a score below 70 may indicate folding prediction failure or intrinsic disorder within

a protein. Scores above 90 imply structure predictions accurate enough for highly shape-sensitive tasks such as chemical binding site characterization. The structure of the 345aa isoform has a very high pLDDT score of 94.7, versus the somewhat lower score of 87.1 for the 373aa MANE isoform.

Because melatonin is primarily synthesized within the human pineal gland at night, we quantified ASMT isoform expression using RNA-sequencing data from a previously-published experiment that used tissue extracted from the pineal gland of a patient who died at midnight [21]. In this tissue sample, the 345aa isoform of *ASMT* was expressed at a level of 327 transcripts per million (TPM), while the 373aa isoform from MANE was expressed at 34 TPM, nearly 10 times lower, supporting our hypothesis that the higher scoring 345aa isoform is functional.

Further evidence for the functionality of the 343aa isoform is shown in Fig. 2. An ectopic exon in the MANE *ASMT* protein creates an unstructured loop that bulges out from the primary structure. The alternate isoform, missing this ectopic exon, closely matches the experimentally determined *ASMT* X-ray crystal structure of the biologically active protein. Furthermore, as reported in Botros et al. 2013 [21], the insertion of exon 6, corresponding to the ectopic exon in the MANE isoform, distorts the structure and destroys its ability to bind S-adenosyl-L-methionine and to synthesize melatonin. Thus, the structural comparison, the expression evidence, and a melatonin synthesis activity assay all combine to support our hypothesis that the 345aa isoform represents the primary biologically functional isoform of *ASMT*.

Gamma-N crystallin

Gamma-N crystallin (CRYGN) is a highly-conserved member of the crystallin family of proteins, responsible for the transparency of the lens and cornea in vertebrate eyes [22]. The intron-exon structure of CRYGN has been conserved across at least 400 million years of vertebrate evolution, with close orthologs present in the genomes of chimpanzees, mice, frogs, and the white-rumped snowfinch. Given this extensive evolutionary history, Wistow et al. 2005 [23] were surprised to observe that the primate CRYGN gene has lost its canonical stop codon, leading them to conclude “the human gene has clearly changed its expression and may indeed be heading for extinction.”

As shown in Fig. 3a, the MANE isoform (CHS.52273.5, RefSeq NM_144727.3, GENCODE ENST00000337323.3) that matches descriptions by Wistow et al. [23] includes sequences that do not fold well, as indicated by its pLDDT score of 67.7. However, we found an alternate CRYGN isoform, assembled from GTEx [17] data, that had a far higher pLDDT score of 92.2, shown in Fig. 3b. Small differences between pLDDT scores may not be meaningful, but large score differences, such as the 24-point gap between the two isoforms of CRYGN discussed here, represent a substantial difference in prediction confidence across a large portion of the protein. The higher-scoring CRYGN isoform is present in CHESS (CHS.52273.9) and GENCODE (ENST00000644350.1), and it was also present in RefSeq v109 (XM_005249952.4) but was removed in the next release, v110.

Both the MANE and alternate isoforms are exactly the same length despite having different C-terminal sequence content. Visual comparison of the predicted structure for the alternate isoform reveals a marked improvement in the structure in the β domain and a clear recovery of CRYGN's dimer-like characteristic, with two structurally similar domains as shown in Fig. 3b and Video 1. Ramachandran plots (Fig. 3c,d) also support the structure of the alternate CHESS isoform.

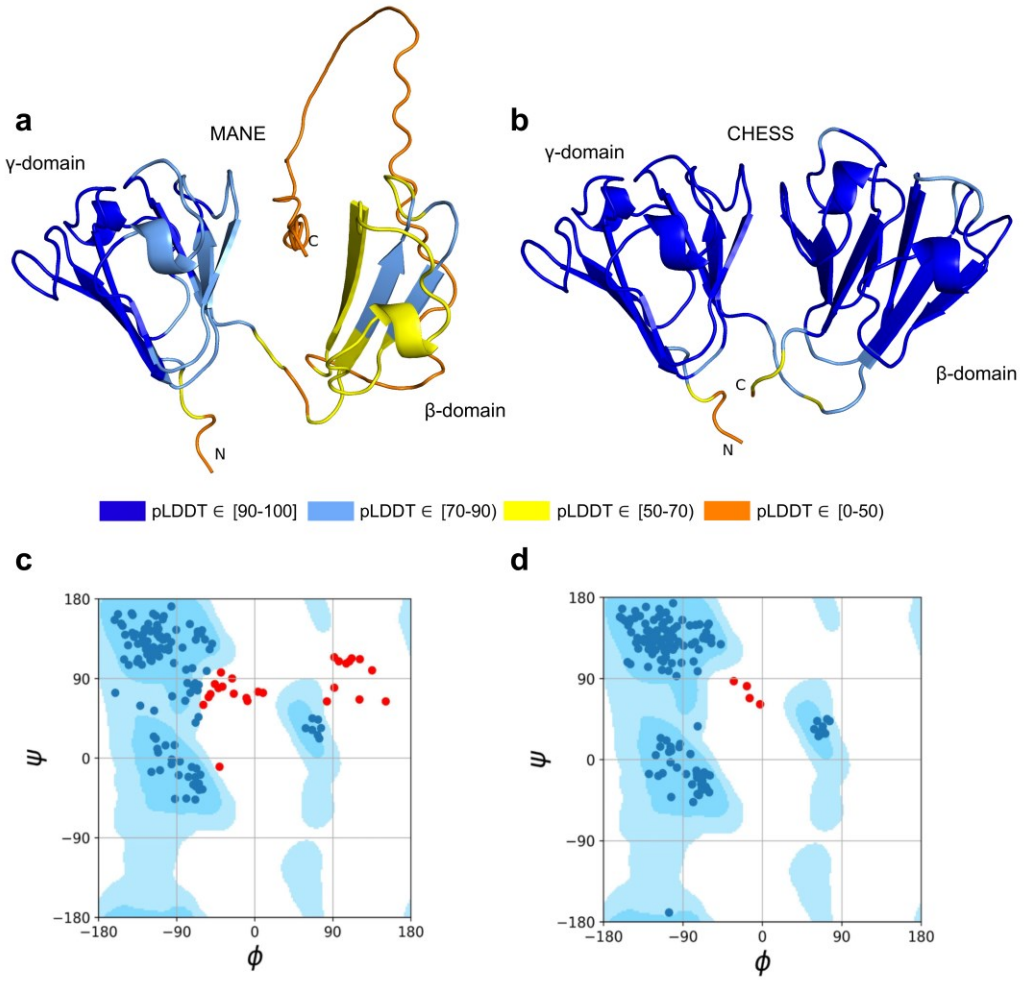

Figure 3: CRYGN isoform comparison. (a) Predicted protein structure for the MANE isoform (CHS.52273.5, RefSeq NM_144727.3, GENCODE ENST00000337323.3) of gamma-N crystallin (CRYGN), colored by pLDDT score. (b) Predicted protein structure

for a CRYGN alternate isoform (CHS.52273.9, GENCODE ENST00000644350.1). (c) Ramachandran plot for the MANE (CRYGN) isoform. Dark blue areas represent “favored” regions while light blue represent “allowed” regions [24]. The 32 red dots represent amino acid residues with secondary structures that fall outside the allowed regions. (d) Ramachandran plot for the alternate CRYGN isoform with 4 red dots falling in disallowed regions, compared to 32 disallowed in MANE. All residues associated with the 4 red dots in the alternate isoform are shared with the MANE isoform in the poorly folded N-terminal region.

Encouraged by the substantially improved folding of this alternate isoform, we examined the intron-exon structure of *CRYGN* in human to determine how it recovered its functional shape despite losing its original stop codon. We found that the common vertebrate four-exon structure of *CRYGN* has changed to a five-exon structure in humans, as shown in Fig. 4. In the well-folded alternate isoform, a novel primate-specific splice site removes the last four amino acids as compared to other vertebrates, but the new primate-specific fifth exon contains a downstream stop codon that adds four residues. The poorly folding MANE isoform (Fig. 4, bottom), in contrast, entirely skips the fourth exon, resulting in a frameshift that adds 43 C-terminal amino acids which have no similarity to any *CRYGN* sequence outside of primates.

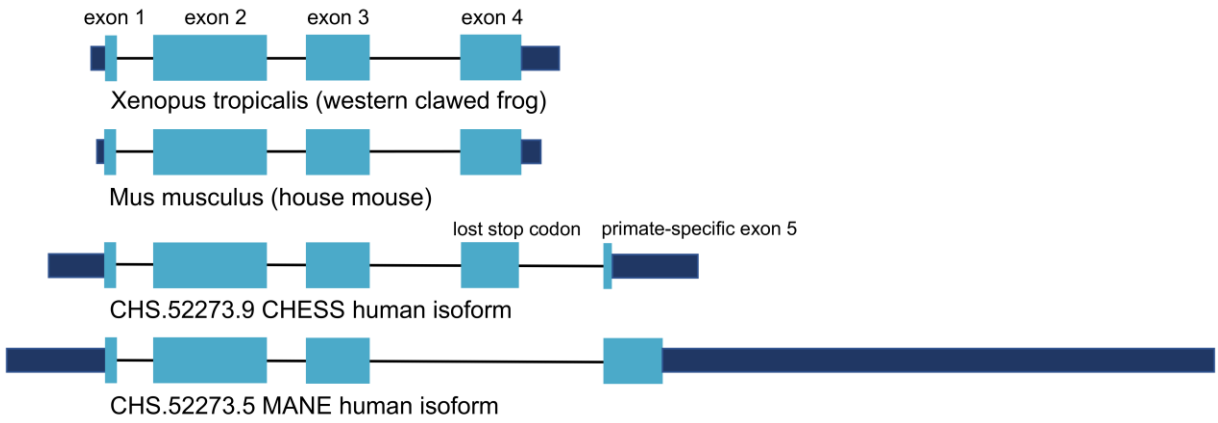

Figure 4: CRYGN intron-exon structure. Comparison of gamma-N crystallin (CRYGN) transcript structures in frog, mouse, and human. Exons 1, 2, and 3 are highly conserved across all species. Exon 4 is missing from the poorly folding MANE isoform, while exon 5 shows no homology to any species outside of primates. The loss of a stop codon in human exon 4 appears to be balanced by the inclusion of a short novel exon that adds only four amino acids to the final protein. Coding portions of exons are shown with thicker rectangles in teal. Intron lengths are reduced proportionately for the purpose of display.

Thioredoxin domain-containing protein 8

The thioredoxin protein family represents an ancient group of highly-conserved small globular proteins found in all forms of life [25]. Thioredoxin domain-containing protein 8 (TXNDC8, alternatively PTRX3) is a testis-specific enzyme responsible for catalyzing redox reactions via the oxidation of cysteine from dithiol to disulfide forms [26].

As shown in Fig. 5, several canonical protein motifs appear altered or missing in the predicted structure of the human TXNDC8 MANE transcript, as it lacks a highly-conserved sequence that should start only eight residues away from the CGPC dithiol/disulfide active site. The α2 helix is severely truncated, leading directly to the α3 helix, thereby entirely skipping the β3 sheet. The β5 sheet, normally providing an

interaction bridge between the α3 and α4 helices, is similarly missing in its entirety. Finally, the α4 helix is present but rotated 140 degrees relative to its canonical position. These large alterations to the fundamental thioredoxin protein structure result in the MANE transcript receiving a pLDDT score of 56.7.

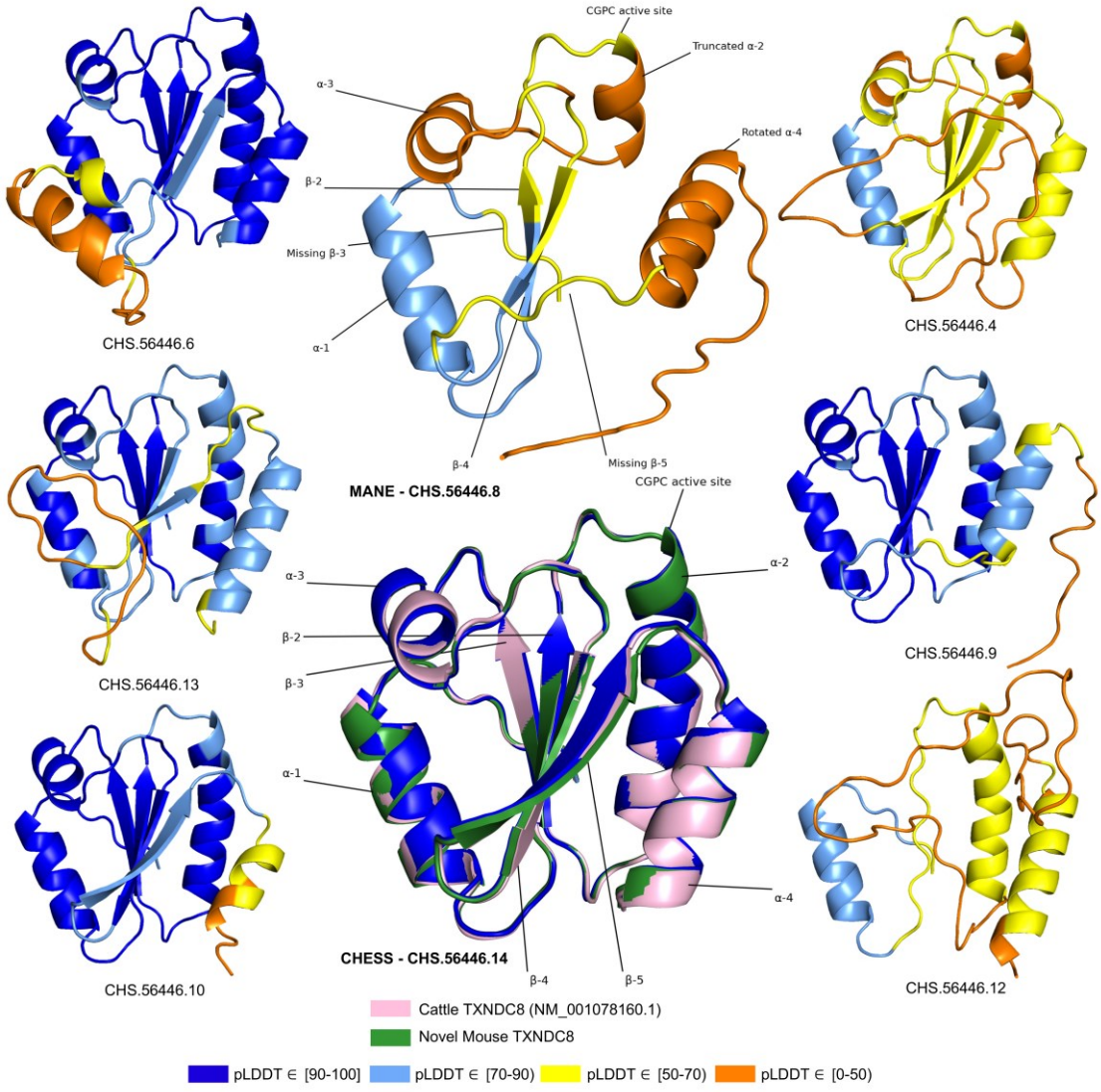

Figure 5: TXNDC8 isoform comparison. Predicted protein structures for seven distinct human isoforms of thioredoxin domain-containing protein 8 (TXNDC8), as well as the primary cattle transcript and a novel mouse transcript. Alternate human isoforms 4, 9, and 12 (right side of figure) lack multiple canonical thioredoxin structures and thus

appear non-functional. Several canonical protein motifs are missing or altered in the predicted structure of the MANE transcript (top center). In contrast, the alternate human transcript 14 matches cattle and mouse to within 0.8 Angstroms. Human transcript CHS.56446.14 is colored solid dark blue because every amino-acid residue scores a pLDDT above 90.

In stark contrast to the poorly-folded MANE isoform (CHS.56446.8, RefSeq NM_001286946.2, GENCODE ENST00000423740.7), an alternate isoform assembled as part of the CHESS project (CHS.56446.14) and RefSeq (NM_001364963.2) has a pLDDT score of 96.9, an improvement of 40 points. Inspection of the alternate transcript (Fig. 5) reveals full recovery of the canonical α2, α3, and α4 helices as well as the β3 and β5 sheets. Moreover, 3D alignment of the protein encoded by CHS.56446.14 to the protein from the primary *TXNDC8* isoform in *Bos taurus* (cow) reveals a very close structural correspondence between the two proteins, with a predicted root-mean-square deviation (RMSD) of 0.8Å. Fig. 5 shows multiple alternative isoforms of human *TXNDC8* from the CHESS annotation, as well as the three-dimensional alignment of CHS.56446.14 to its orthologs in cow and mouse. Given its substantially higher pLDDT score and near-perfect structural conservation in other species, the CHESS transcript appears to be a much better candidate for the canonical form of this protein. The MANE isoform, because it is missing multiple key structures, may represent a non-functional product of transcriptional noise.

All isoforms of *TXNDC8* shown in Fig. 5 were assembled from RNA sequencing data during the construction of the CHESS database. This figure illustrates another potential

use of structure prediction, namely the ability to distinguish among multiple functional and non-functional isoforms when annotating a genome. As discussed above, the MANE *TXNDC8* isoform appears non-functional, lacking several key structures. In addition, isoform 12 in Fig. 5 appears clearly non-functional, lacking all four of the β sheets and one of the α helices of isoform 14. Isoforms 4 and 9 also appear likely to be non-functional: both are missing one of the β sheets, and isoform 4 has a low pLDDT score of just 52. Although this example is only one of many, it illustrates how one can employ accurate 3D structure prediction in virtually any species as a powerful new tool to improve gene annotation.

Assemblies of RNA-seq experiments typically reveal thousands of un-annotated gene isoforms, as was illustrated by the use of GTEx to discover more than 100,000 new isoforms when building the original CHESS gene database [4]. The approach used here, computationally folding each distinct protein encoded by alternate isoforms, allows us to compare the structure of these predicted proteins to the “best” structure for each protein-coding gene locus. For those proteins with at least one high-confidence structure, this strategy may allow us to identify and remove potentially non-functional isoforms.

Interleukin 36 beta

Interleukin 36 beta (IL36B, alternatively IL1F8, FIL1-ETA, or IL1H2) mediates inflammation as part of a signaling system in epithelial tissue. The pro-inflammatory properties of IL36B have been implicated in the pathogenesis of psoriasis [27], a common

disease characterized by scaly rashes on the skin. In vitro and in vivo studies have found consistently increased expression of IL36B in psoriatic lesions, making the gene a potential target for future anti-psoriatic drugs [28].

The ColabFold structure of the IL36B MANE isoform (CHS.30565.1, RefSeq NM_014438.5, GENCODE ENST00000259213.9) averaged a pLDDT of only 50.2, a score indicative of near-complete folding failure, while an alternate isoform in CHESS (CHS.30565.4) and RefSeq (XM_011510962.1), shown in Fig. 6, scored 93.0, the largest relative score increase of any isoform we examined. This alternate isoform contains two C-terminal exons that are not present in the MANE isoform and that contribute nearly half of the total protein sequence. A BLASTP homology search of the 34aa coding sequence of the final C-terminal exon in the MANE transcript yielded no hits beyond primates using default search parameters. In contrast, a BLASTP search of sequence unique to the alternate isoform revealed significant similarity (e-values of 0.001 and smaller) to IL36B orthologs in 1517 organisms including mouse, rat, and Hawaiian monk seal. In addition, expression of the MANE isoform was observed in only 1 sample in the GTEx data with an expression level of just 0.01 TPM, while the CHESS isoform was found in 775 samples with a much higher expression level of 8.4 TPM (Supplementary File 3).

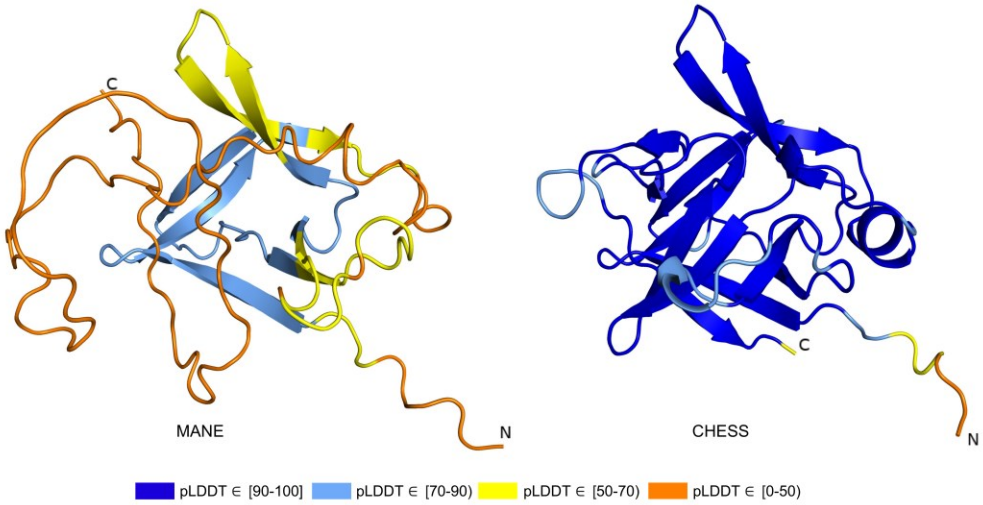

Figure 6: IL36B isoform comparison. Comparison of predicted structures for interleukin 36 beta (IL36B) for the MANE isoform (CHS.30565.1, RefSeq NM_014438.5, GENCODE ENST00000259213.9) and an alternate isoform from CHESS and RefSeq (CHS.30565.4, RefSeq XM_011510962.1). The highly-conserved protein sequence of the alternate human isoform achieves a very high pLDDT score of 93.0, versus the MANE isoform's much lower pLDDT of 50.2.

Further probing the functionality of our high-scoring isoform, we aligned the predicted three-dimensional structures for *IL36B* in human, mouse, and rat. As expected, the mouse and rat proteins aligned to each other remarkably well, with a RMSD of 0.60Å. We found the low-scoring MANE protein aligned poorly to the structures for mouse and rat, averaging a distance of 2.74Å, while the alternate isoform aligned far better with an RMSD of 0.76Å. This close similarity in both sequence and structure to conserved orthologs in distant species strongly reinforces the argument that the alternate isoform represents the functional version of the protein in human.

Post-GPI attachment to proteins 2

The protein known as post-GPI attachment to proteins 2 (PGAP2, alternatively FRAG1 or CWH43N) is required for stable expression of glycosylphosphatidylinositol (GPI)-anchored proteins [29] attached to the external cellular plasma membrane via a post-translational modification system ubiquitous in eukaryotes [30]. Mutations in GPI pathway proteins have been linked to a wide variety of rare genetic disorders [31], while mutations in PGAP2 specifically have been shown to cause to intellectual disability, hyperphosphatasia, and petit mal seizures [32].

Out of 85 GTEx-assembled transcripts for PGAP2 produced during the latest build of the CHESS database, encoding 33 distinct protein isoforms, the single highest scoring isoform according to ColabFold was CHS.7860.59 (RefSeq NM_001256240.2, GENCODE ENST00000463452.6), with a pLDDT of 87.9. The coding sequence of this isoform exactly matches the sequence of the assumed biologically active protein [32,33], and all intron boundaries are conserved in mouse. For comparison, the annotated MANE protein (CHS.7860.58, RefSeq NM_014489.4, GENCODE ENST00000278243.9) has a pLDDT of 78.0 and the intron boundaries are not conserved in mouse. Predicted structures of both proteins are shown in Fig. 7.

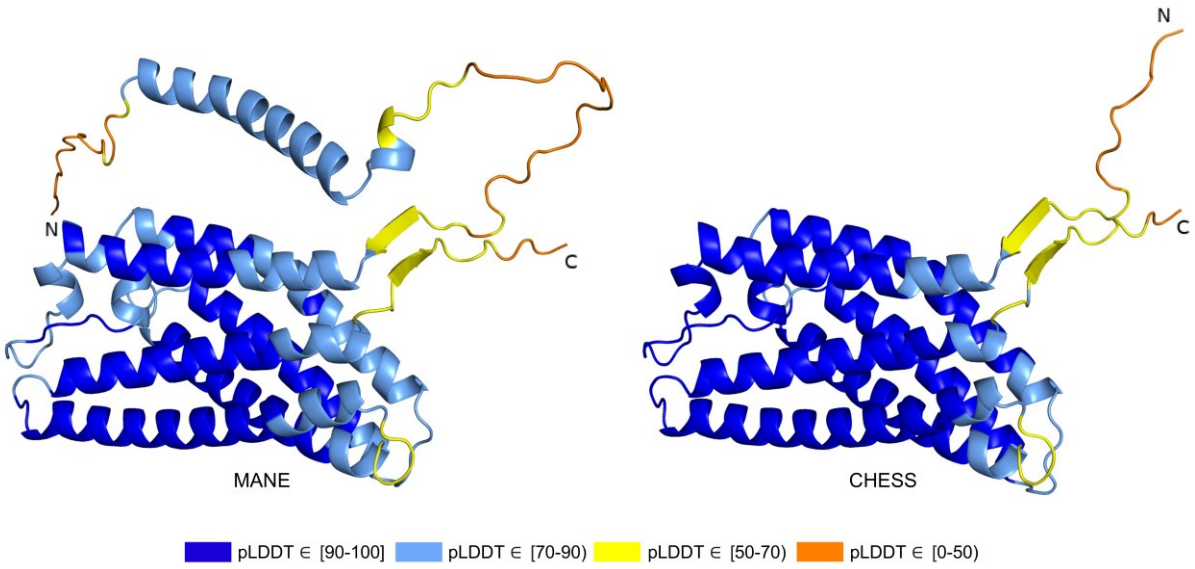

Figure 7: PGAP2 isoform comparison. Comparison of the structure of the MANE isoform (CHS.7860.58, RefSeq NM_014489.4, GENCODE ENST00000278243.9) versus the highest scoring alternate isoform (CHS.7860.59, RefSeq NM_001256240.2, GENCODE ENST00000463452.6) for PGAP2. Of 33 distinct annotated protein isoforms of PGAP2, the one with the highest pLDDT represents the biologically active version [32,33] of PGAP2 in humans.

RNA-seq data from GTEx also showed that the higher scoring isoform, CHS.7860.59, was expressed in 8776 samples with an average expression level of 2.6 TPM, compared to only 4116 samples with an 0.9 TPM average for the MANE isoform. Comparing their intron-exon structure revealed that the MANE transcript has one extra exon (the second exon out of six). On average across all 31 tissues in the GTEx data, five times more spliced reads supported skipping that exon, as in CHS.7860.59, rather than including it.

Functional splice variants may not fold well:

Alternative splicing allows genes to code for multiple functional protein products [34]. Thus, rejecting all but the top-scoring isoform based on predicted structure may eliminate lower-scoring yet functional proteins. The risk of discarding functional transcripts by relying too heavily on the pLDDT score is well-illustrated by vascular endothelial growth factor B (*VEGFB*), a growth factor implicated in cancer and diabetes-related heart disease [35]. The human *VEGFB* gene encodes two well-characterized protein isoforms: *VEGFB-167* and *VEGFB-186*. Alternative splicing that skips part of the sixth exon in *VEGFB-167* leads to sequestration of the protein to the cell surface due to a highly basic C-terminal heparin binding domain. Full inclusion of exon six in *VEGFB-186* results in a soluble protein freely transported to the blood stream [36].

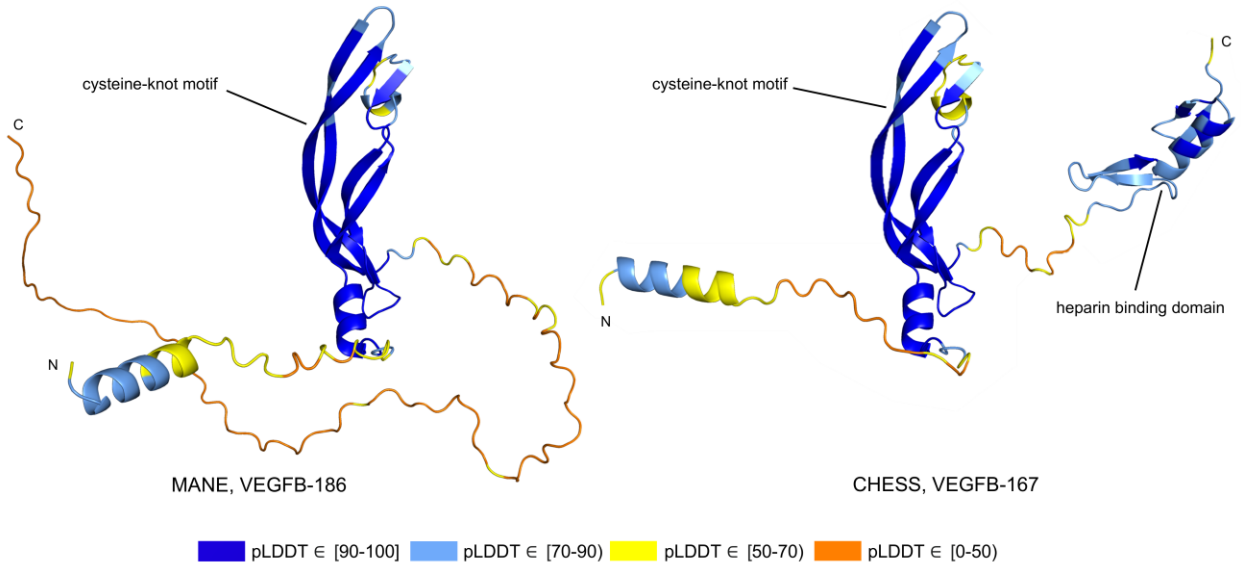

Figure 8: VEGFB isoform comparison. Vascular endothelial growth factor B (VEGFB) isoforms VEGFB-186 (a) and VEGFB-167 (b). The inclusion of a heparin binding domain in VEGFB-167 results in sequestration to the cell surface while VEGFB-186 remains freely soluble. Relying solely on pLDDT comparisons in this case would be misleading, as both isoforms represent well-understood functional protein products.

Both isoforms shown in Fig. 8 represent highly-expressed and similarly functional products, containing a well conserved cysteine-knot motif [37], yet *VEGFB-167* receives a pLDDT score of 81.7 while *VEGFB-186* receives a much lower pLDDT score of 69.5. The MANE isoform (CHS.9039.1, RefSeq NM_003377.5, GENCODE ENST00000309422.7) encodes the freely-soluble protein *VEGFB-186*, while the alternate isoform (CHS.9039.2, RefSeq NM_001243733.2, GENCODE ENST00000426086.3) encodes the sequestered protein *VEGFB-167*. Additionally, *VEGFB-186* is present as a full-length cDNA clone (MGC:10373 IMAGE:4053976) in the Mammalian Gene Collection [38], a fact which strongly supports its functionality. Due to the large pLDDT score difference between the two functional *VEGFB* isoforms, a naïve attempt to use protein folding prediction scores as the sole oracle of protein function might inadvertently discard *VEGFB-186*, a clearly functional transcript. Thus, one must be careful to incorporate multiple sources of information when making decisions about isoform functionality.

A novel protein-coding transcript in mouse:

While examining the evolutionary conservation of the three-dimensional structure of TXNDC8 in human, we noticed that the predicted structure for the same gene in Mus musculus (house mouse) seemed to contain a poorly folded region similar to CHS.56446.6, a low-scoring human protein. Further inspection of TXNDC8 in the mouse genome revealed that the primary RefSeq transcript contains a misfolding fifth exon, while the alternate RefSeq transcript skips the misfolding exon, similar to the functional human isoform, shown in Fig. 9. Interestingly, both mouse transcripts contain

a third exon homologous to the sequence missing in the human MANE isoform. A BLASTP search of the misfolding exon resulted in only a single significant hit outside the order Rodentia to an unnamed protein product. In contrast, a BLASTP search of the third exon present in all mouse transcripts, homologous to the missing exon in the human MANE transcript, revealed significant hits to at least 31 homologs outside Rodentia.

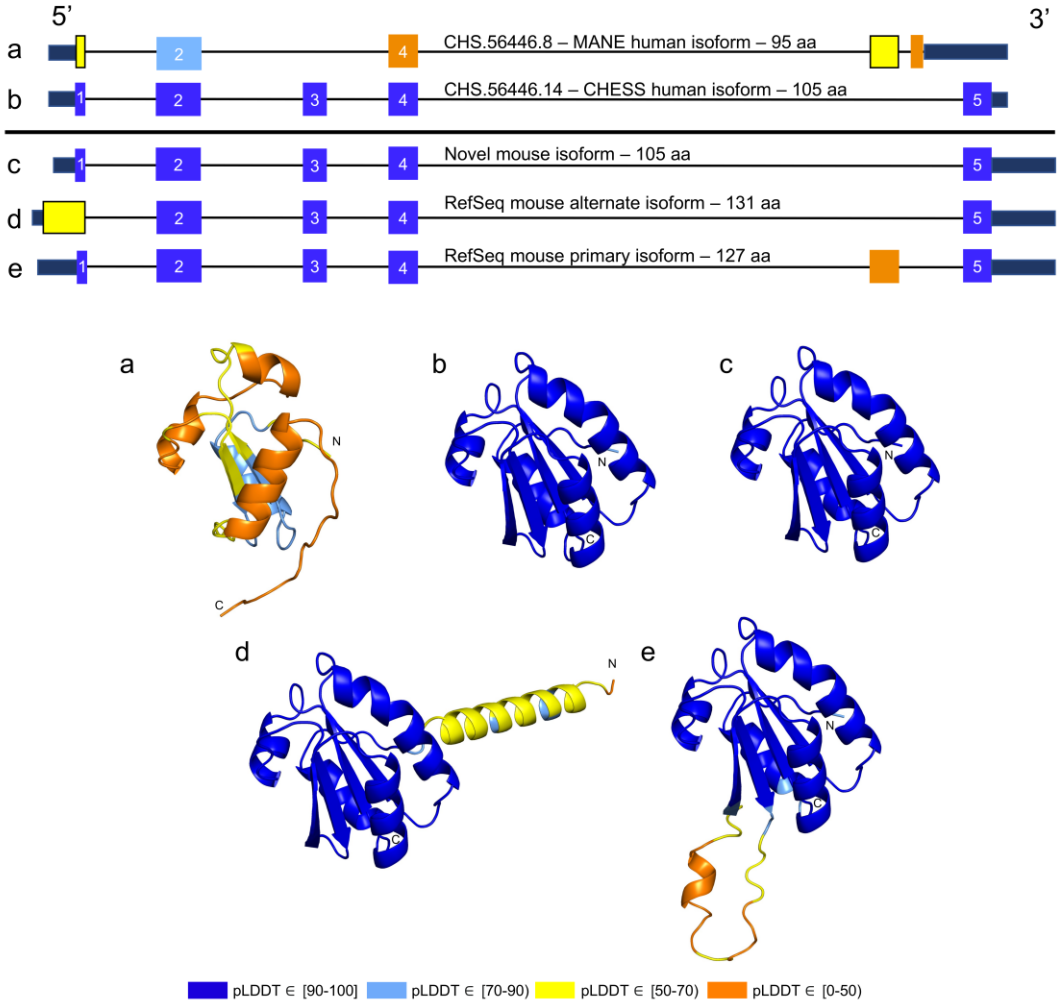

Figure 9: TXNDC8 human and mouse comparison. Intron-exon and predicted protein structure for TXNDC8 in human (a and b) and mouse (c, d, and e). Exons are colored

according to their average pLDDT score. The highest-scoring isoforms in both human (b) and mouse (c) share conserved intron-exon structure and nearly identical predicted protein structure.

Unlike the functional human isoform, however, the RefSeq alternate mouse transcript is annotated with a start site 26 codons upstream of the translation initiation site annotated in the human orthologs. A BLASTP search of the 26 amino-acid additional sequence resulted in zero significant hits outside *Rodentia*. Folding this alternate mouse transcript revealed that these additional N-terminal amino acids fail to form any confident predicted protein structure. As a result, none of the three transcripts annotated in mouse fold into the highly-conserved structure of *TXNDC8* in human. Predicted structures and exon alignments for isoforms in human and mouse are shown in Fig. 9.

We hypothesized that a truly functional isoform of *TXNDC8* in mouse should be similar to the human isoform in three-dimensional structure. Based on our observations, this similarity might be realized if the mouse alternate isoform simply started at the downstream start site that matches human, yielding a 105aa protein rather than the 131aa protein that is currently annotated. Folding the coding sequence of the alternate mouse transcript, minus the 26 N-terminal amino-acid residues, revealed a predicted structure remarkably similar to both human and cattle *TXNDC8*. Fig. 5 shows the three-dimensional alignment of the human and cow proteins to our predicted (105aa) isoform of mouse *TXNDC8*. Remarkably, the predicted average RMS deviation between aligned heavy atoms of the putative human and mouse proteins is just 0.83Å. For reference, the atomic diameter of one carbon atom is 1.4Å.

In a further investigation of mouse *TXNDC8* transcription, we aligned 8 gigabases of RNA-seq cDNA from a mouse testis sample (SRR18337982) to the GRCm38 reference genome using HISAT2 [39] then assembled transcripts using StringTie2 [40]. This resulted in two putative transcripts at the mouse *TXNDC8* locus, with neither transcript containing the upstream start site present in the RefSeq annotation. Examination of the read coverage confirmed that both putative *TXNDC8* transcripts appear to use the start site of our proposed shorter protein-coding sequence, with 1060 reads supporting the canonical start site and zero reads supporting the upstream start site. As hypothesized, one of these newly assembled transcripts contained the protein-coding sequence necessary to exactly match our predicted structure-conserved isoform.

We believe this represents the first experimentally confirmed novel isoform in any organism discovered due to a hypothesis derived from comparison of computationally predicted protein structures. All in all, the structure-guided identification and subsequent experimental confirmation of a novel functional *TXNDC8* isoform in mouse demonstrates the potential of three-dimensional protein structure prediction to enhance functional annotation in any genome.

A resource for human annotation:

The examples discussed here are only a small subset of the 130,700 unique protein structures we generated at 20,817 human gene loci. These structures and associated pLDDT scores have already been used to avoid filtering out transcripts encoding

functional, clinically relevant isoforms while building the latest version of the CHESS human gene catalog. We provide all of these structures as a searchable and downloadable database, at isoform.io, to create a public resource for improving the annotation of the human genome. The large majority of CHESS isoforms in this collection have direct support from RNA-sequencing data, as they were assembled from the large GTEx collection, a high-quality set of deep RNA-sequencing experiments across dozens of human tissues. A very small number of MANE proteins were not assembled from GTEx data, but structures of these too are included in this analysis so that no MANE genes would be omitted.

In the online resource, we provide for each transcript: (1) the nucleotide and amino-acid sequences; (2) the predicted structure, as a file that can be viewed in a standard structure viewer such as PyMOL [41]; (3) the pLDDT score of that structure; (4) the length of the isoform; (5) the number of GTEx samples in which the isoform was observed; (6) the maximum expression of the isoform in any tissue; (7) an indicator based on alignment of whether all introns are conserved in the mouse genome; (8) an interactive table with functionality to search and sort transcripts to find isoforms of interest; and (9) a Foldseek [42] interface to search any given protein structure against all 237,295 transcripts presented here. These predictions can be mined to discover, for example, cases where a known protein gets a surprisingly low pLDDT score, or where alternative isoforms have structures that get higher scores and appear more stable than previously-reported forms of the same protein.

1.3 Discussion

In this analysis, we demonstrated the ability to improve protein-coding gene annotation by predicting three-dimensional protein structures. We searched tens of thousands of predicted structures of alternate isoforms of human genes and identified a subset that appear to fold more confidently than the isoforms found in MANE, a recently-developed "universal standard" for human gene annotation [15]. We found hundreds of gene isoforms, all of which were supported by RNA sequencing data, that outscored the corresponding MANE transcript. Inclusion of truly functional protein isoforms in future releases of human gene catalogs, particularly clinically focused catalogs such as MANE, may enable more accurate downstream analyses of these genes.

In the illustratory examples described here, we provide biological and evolutionary context for cases where an alternate human isoform appears clearly superior in structure to its canonical protein. Given the many additional high-scoring transcripts that we identified (Supplementary File 3), we expect further improvements in human annotation are yet to be discovered. More generally, we followed a structure-guided annotation strategy that may prove useful in refining the annotation of many non-human species as well.

We expect computational protein structure prediction to become an indispensable tool for future transcriptome annotation efforts. Still, the functionality of many proteins may not be revealed by structure prediction alone. Cases where substantial portions of a protein fail to form a stable structure, such as intrinsically disordered proteins, were not

examined here. Additionally, non-functional isoforms may achieve a higher pLDDT simply due to the omission of small, yet functionally important intrinsically disordered regions. In these cases, it is necessary to contextualize results with both expression data, when available, and evolutionary sequence conservation analysis. If a low-scoring region is highly conserved across species or is consistently expressed, this still provides a strong indication of function regardless of foldability.

Although we restricted our analysis to whole-protein comparisons, comparing local structural portions of a protein, potentially near shape-sensitive ligand binding sites [43], may enable similar analysis in these proteins. Further advances in predicting structures for multi-chain protein complexes [44], as well as improvements in prediction efficiency in large proteins, may expand the range of genes that may be analyzed. An important caveat is that in some cases, truly functional isoforms may receive low predicted folding scores relative to well-folded functional alternate isoforms within the same gene. Thus, structure prediction alone is not always sufficient to make functional claims about any individual protein isoform.

Ideas and speculation:

Though the complete sequence of the human genome has been revealed [45], the annotation of the human genome, by far the most comprehensively studied, remains far from finished. The use of accurate predicted protein structures for gene annotation, as we have done here, represents a new paradigm, not only for human gene discovery but for all other species as well. For decades, the scientific community has relied principally

on two methods to discover and validate protein-coding genes at the genome scale: sequencing of transcripts (or cDNAs), and alignment of DNA and protein sequences to detect evolutionary conservation in other species. Protein structure holds valuable information regarding biological functionality, providing an independent and powerful tool to complement these methods. The analysis described here takes a first step toward improving genome annotation of humans using structure prediction, but specific methods deploying this powerful new tool on a broader scale will require fundamentally new computational strategies. As such, development of comprehensive genome-annotation protocols incorporating protein structure prediction will remain an area of active investigation for years to come.

1.4 Materials and Methods

Protein structure prediction:

We folded all transcripts in the Comprehensive Human Expressed SequenceS (CHESS) annotation less than 1000 amino acids in length. Similar to the initial effort to fold the human proteome [13], the length limit was chosen to make the overall computational runtime feasible. This yielded 233,973 transcripts representing 127,398 unique protein-coding sequences at 20,666 loci. Coding sequences for CHESS transcripts were determined with ORFanage [46]. For each protein sequence, we generated a multiple sequence alignment by aligning them with ColabFold's MMseqs2 [47] workflow (colabfold_search) against the UniRef100 [48] (2021/03) and ColabFoldDB (2021/08) database. Structure predictions were made with ColabFold (commit 3398d3)

using AlphaFold2 and MMseqs2 version 13.45111. To speed up the search, we set the sensitivity setting to 7 (-s 7). We predicted each structure using colabfold_batch and stopped the process early if a pLDDT of at least 85 was reached by any model (--stop-at-score 85) or if a model produced a pLDDT less than 70 (--stop-at-score-below 70). All models were ranked by pLDDT in descending order. Run-time was estimated from a sample of 500 proteins randomly selected from the 127,398 structures. Prediction of all structures on 8 x A5000 GPUs required 34 days. Multiple sequence alignment took 34 hours using MMseqs2 on an AMD EPYC 7742 CPU with 64 cores.

Filtering MANE comparisons:

To generate the 940 protein isoforms in Supplementary File 3, we used the following filtering criteria and procedures. For each isoform with a distinct coding sequence located at a MANE v1.0 locus and with coding sequence overlapping a MANE annotated protein, we compared the pLDDT score to that of the associated MANE protein. As described previously, pLDDT is a reliable measure of the confidence in a structure, where predictions with 70 ≤ pLDDT ≤ 90 are confident, those with pLDDT > 90 are highly confident, and those below 50 represent low-confidence structures and may be disordered proteins [18]. We only considered alternative isoforms that had a pLDDT score ≥ 70, indicating a generally well-folded protein, to avoid including any intrinsically disordered proteins [49]. Filtering and general analysis was performed in Colab (https://colab.research.google.com) with Python version 3.8.15. A full list of the filtered subset can be found in Table 1 below.

CHESS_ID_isoform	CHESS_ID_MANE	gene	aa_length_isoform	aa_length_MANE	length_ratio	pLDDT_isoform	pLDDT_MANE	pLDDT_ratio	GTEx_samples_observed_isoform	GTEx_samples_observed_MANE	GTEx_top_tissue_name_isoform	GTEx_top_tissue_name_MANE	GTEx_top_tissue_TPM_isoform	GTEx_top_tissue_TPM_MANE	introns_conserved_in_mouse_isoform	introns_conserved_in_mouse_MANE
CHS.10000.2	CHS.10000.4	NCAM1	726	858	0.8462	87.5	81.5	1.07	6098	5156	Brain	Uterus	35.08	26.99	TRUE	TRUE
CHS.10005.3	CHS.10005.4	DRD2	414	443	0.9345	73.4	68.8	1.07	962	1200	Pituitary	Pituitary	19.54	62.50	FALSE	FALSE
CHS.10008.7	CHS.10008.9	TMPRSS5	413	457	0.9037	83.1	79	1.05	1586	2006	Nerve	Nerve	5.70	25.09	FALSE	FALSE
CHS.10133.59	CHS.10133.10	PHLDB1	721	1377	0.5236	72.3	59	1.23	8159	6998	Colon	Brain	2.30	11.17	TRUE	TRUE
CHS.10182.2	CHS.10182.8	USP2	396	605	0.6545	85.3	65.7	1.30	7695	5407	Kidney	Testis	24.36	126.52	TRUE	TRUE
CHS.10323.13	CHS.10323.1	VSIG2	284	327	0.8685	87	80.7	1.08	3791	3735	Stomach	Stomach	24.50	302.78	TRUE	TRUE
CHS.10326.8	CHS.10326.13	MSANTD2	329	559	0.5886	76.2	63.9	1.19	9695	7130	Thyroid	Thyroid	17.04	5.59	TRUE	TRUE
CHS.10376.18	CHS.10376.15	KIRREL3	600	778	0.7712	83.8	71.1	1.18	2033	2011	Vagina	Brain	0.50	9.34	TRUE	TRUE
CHS.10627.2	CHS.10627.7	NINJ2	135	142	0.9507	76.4	68.2	1.12	5889	6455	Blood	Brain	3.94	9.60	TRUE	TRUE
CHS.10627.5	CHS.10627.7	NINJ2	106	142	0.7465	75.5	68.2	1.11	8351	6455	Lung	Brain	19.15	9.60	TRUE	TRUE
CHS.10698.14	CHS.10698.9	CRACR2A	395	731	0.5404	74.9	68.6	1.09	2335	1190	Salivary_Gland	Bladder	0.53	1.60	FALSE	FALSE
CHS.10701.7	CHS.10701.5	PARP11	257	338	0.7604	86	80.8	1.06	6881	6571	Cervix_Uteri	Ovary	2.06	4.14	FALSE	TRUE
CHS.10725.50	CHS.10725.33	DYRK4	327	634	0.5158	73.3	67.8	1.08	605	18	Prostate	Testis	17.18	5.89	FALSE	TRUE
CHS.10725.8	CHS.10725.33	DYRK4	520	634	0.8202	76.8	67.8	1.13	232	18	Skin	Testis	3.86	5.89	TRUE	TRUE
CHS.10747.36	CHS.10747.7	ANO2	599	998	0.6002	77.8	71	1.10	86	27	Lung	Salivary_Gland	4.04	8.58	FALSE	TRUE
CHS.10747.4	CHS.10747.7	ANO2	546	998	0.5471	76.9	71	1.08	96	27	Testis	Salivary_Gland	4.39	8.58	FALSE	TRUE
CHS.10813.1	CHS.10813.18	TPI1	286	249	1.1486	87.5	87.5	1.00	9791	0	Bone_Marrow	-	726.15	0.00	TRUE	TRUE
CHS.10813.13	CHS.10813.18	TPI1	167	249	0.6707	94.9	87.5	1.08	9529	0	Brain	-	11.81	0.00	TRUE	TRUE
CHS.10904.19	CHS.10904.14	KLRG1	110	189	0.582	93.3	82.1	1.14	3632	3195	Spleen	Spleen	1.91	5.86	FALSE	TRUE
CHS.10959.9	CHS.10959.11	CLEC7A	201	247	0.8138	83	75.6	1.10	6133	4401	Blood	Blood	15.85	8.90	TRUE	TRUE
CHS.11072.10	CHS.11072.1	GPRC5D	300	345	0.8696	79.1	73.8	1.07	702	264	Lung	Skin	2.88	18.47	TRUE	TRUE
CHS.11074.3	CHS.11074.7	GSG1	326	362	0.9006	70.6	67.1	1.05	422	218	Nerve	Testis	1.09	98.49	FALSE	FALSE
CHS.11361.19	CHS.11361.18	BICD1	835	975	0.8564	76.6	68.8	1.11	3998	2447	Brain	Ovary	4.96	4.36	TRUE	TRUE
CHS.11361.4	CHS.11361.18	BICD1	827	975	0.8482	77.1	68.8	1.12	2742	2447	Brain	Ovary	4.32	4.36	TRUE	TRUE
CHS.11645.3	CHS.11645.1	ANKRD33	272	452	0.6018	80.9	63.9	1.27	16	14	Liver	Nerve	0.68	1.16	FALSE	FALSE
CHS.11790.2	CHS.11790.1	HNRNPA1	320	372	0.8602	70	64.6	1.08	9794	9790	Bone_Marrow	Ovary	710.97	472.34	FALSE	FALSE
CHS.11792.16	CHS.11792.15	COPZ1	160	177	0.904	91.9	87.2	1.05	9760	9786	Bone_Marrow	Bone_Marrow	16.43	228.22	TRUE	TRUE
CHS.1184.11	CHS.1184.8	TFAP2E	271	442	0.6131	80.2	64.2	1.25	2354	1355	Fallopian_Tube	Brain	5.15	7.11	TRUE	TRUE
CHS.11943.6	CHS.11943.8	ARHGAP9	420	731	0.5746	74.9	64.5	1.16	5179	2523	Blood	Spleen	27.36	7.78	FALSE	FALSE
CHS.12119.5	CHS.12119.1	NUP107	896	925	0.9686	79.4	75.4	1.05	83	9225	Testis	Bone_Marrow	1.91	34.95	TRUE	TRUE
CHS.12155.2	CHS.12155.8	PTPRR	412	657	0.6271	78.3	68.8	1.14	1360	834	Bladder	Skin	6.05	32.01	TRUE	TRUE
CHS.12176.68	CHS.12176.5	TBC1D15	649	674	0.9629	71.2	67.6	1.05	0	9703	-	Adipose_Tissue	0.00	18.73	TRUE	TRUE
CHS.12203.25	CHS.12203.59	CAPS2	382	557	0.6858	83.9	75.3	1.11	1280	27	Thyroid	Adrenal_Gland	3.73	2.77	FALSE	FALSE

CHS.12203.55	CHS.12203.59	CAPS2	350	557	0.6284	80	75.3	1.06	339	27	Fallopian_Tube	Adrenal_Gland	3.20	2.77	FALSE	FALSE
CHS.126.10	CHS.126.4	CALML6	156	181	0.8619	83.3	75.9	1.10	2276	933	Muscle	Muscle	9.95	7.71	FALSE	FALSE
CHS.126.2	CHS.126.4	CALML6	164	181	0.9061	81.6	75.9	1.08	722	933	Muscle	Muscle	3.69	7.71	FALSE	FALSE
CHS.126.9	CHS.126.4	CALML6	138	181	0.7624	83.3	75.9	1.10	4022	933	Muscle	Muscle	3.01	7.71	FALSE	FALSE
CHS.12617.9	CHS.12617.1	TXNRD1	461	649	0.7103	96.5	89.3	1.08	4547	906	Bladder	Testis	1.58	11.72	TRUE	TRUE
CHS.12653.10	CHS.12653.2	CKAP4	550	602	0.9136	77.1	73.4	1.05	1073	9725	Skin	Skin	7.87	95.26	FALSE	TRUE
CHS.12808.20	CHS.12808.7	TPCN1	748	816	0.9167	85	80.2	1.06	1708	9355	Brain	Thyroid	1.24	33.36	FALSE	TRUE
CHS.12956.10	CHS.12956.15	BICDL1	602	621	0.9694	76.9	73	1.05	60	2866	Blood	Testis	2.49	42.38	TRUE	TRUE
CHS.12956.14	CHS.12956.15	BICDL1	561	621	0.9034	78.7	73	1.08	166	2866	Pituitary	Testis	22.33	42.38	TRUE	TRUE
CHS.12956.20	CHS.12956.15	BICDL1	588	621	0.9469	76.8	73	1.05	1254	2866	Skin	Testis	105.11	42.38	TRUE	TRUE
CHS.12956.44	CHS.12956.15	BICDL1	584	621	0.9404	77.3	73	1.06	94	2866	Pituitary	Testis	25.79	42.38	TRUE	TRUE
CHS.12956.56	CHS.12956.15	BICDL1	585	621	0.942	77.4	73	1.06	11	2866	Pituitary	Testis	28.87	42.38	TRUE	TRUE
CHS.12956.7	CHS.12956.15	BICDL1	562	621	0.905	78.7	73	1.08	1	2866	Colon	Testis	0.41	42.38	TRUE	TRUE
CHS.13007.6	CHS.13007.7	P2RX7	364	595	0.6118	87.2	82.6	1.06	5711	5324	Lung	Nerve	4.39	9.03	TRUE	TRUE
CHS.13049.1	CHS.13049.24	DIABLO	186	239	0.7782	91.4	79.7	1.15	9755	9735	Testis	Testis	17.56	88.39	FALSE	FALSE
CHS.13090.1	CHS.13090.8	KMT5A	295	352	0.8381	70.2	63.4	1.11	6333	3508	Bone_Marrow	Prostate	44.51	38.01	FALSE	FALSE
CHS.13233.4	CHS.13233.12	GLT1D1	346	266	1.3008	94.3	85.3	1.11	2857	2399	Blood	Blood	37.07	6.71	TRUE	TRUE
CHS.13233.7	CHS.13233.12	GLT1D1	324	266	1.218	89.7	85.3	1.05	1	2399	Brain	Blood	1.34	6.71	TRUE	TRUE
CHS.13233.8	CHS.13233.12	GLT1D1	362	266	1.3609	91.3	85.3	1.07	396	2399	Muscle	Blood	3.77	6.71	TRUE	TRUE
CHS.13233.9	CHS.13233.12	GLT1D1	308	266	1.1579	92.2	85.3	1.08	14	2399	Spleen	Blood	3.25	6.71	TRUE	TRUE
CHS.13365.30	CHS.13365.5	CHFR	591	652	0.9064	72.4	68	1.06	104	9466	Lung	Spleen	3.48	16.11	FALSE	FALSE
CHS.13365.40	CHS.13365.5	CHFR	592	652	0.908	71.6	68	1.05	236	9466	Spleen	Spleen	3.19	16.11	FALSE	FALSE
CHS.13480.2	CHS.13480.14	PSPC1	393	523	0.7514	80.4	68.9	1.17	9769	9762	Uterus	Testis	18.80	22.35	FALSE	TRUE
CHS.13529.17	CHS.13529.7	ZDHHC20	320	365	0.8767	91	83.4	1.09	8183	7159	Nerve	Bladder	11.25	8.50	TRUE	TRUE
CHS.13577.7	CHS.13577.8	MIPEP	651	713	0.913	94.5	90	1.05	299	9028	Prostate	Skin	3.87	20.72	TRUE	TRUE
CHS.13667.16	CHS.13667.5	FLT1	687	1338	0.5135	84.6	72.7	1.16	8286	7077	Thyroid	Thyroid	42.28	44.85	TRUE	TRUE
CHS.13726.11	CHS.13726.23	TEX26	158	289	0.5467	72.5	58.1	1.25	655	641	Testis	Colon	17.56	30.21	TRUE	TRUE
CHS.13738.2	CHS.13738.3	B3GLCT	449	498	0.9016	91.5	85.9	1.07	14	5882	Skin	Uterus	3.01	19.58	TRUE	TRUE
CHS.13802.5	CHS.13802.8	CCNA1	421	465	0.9054	71.9	66.1	1.09	1395	2269	Testis	Testis	45.12	9.96	FALSE	FALSE
CHS.13809.5	CHS.13809.4	SMAD9	430	467	0.9208	83.8	79.6	1.05	5002	5297	Thyroid	Thyroid	11.05	14.74	TRUE	FALSE
CHS.1381.27	CHS.1381.55	CCDC30	710	938	0.7569	75.6	71	1.06	878	0	Thyroid	-	4.91	0.00	FALSE	FALSE
CHS.1381.29	CHS.1381.55	CCDC30	752	938	0.8017	75.4	71	1.06	8	0	Thyroid	-	2.50	0.00	FALSE	FALSE
CHS.13823.29	CHS.13823.10	POSTN	778	836	0.9306	83.9	79.7	1.05	96	5233	Lung	Blood	2.49	15.00	TRUE	TRUE
CHS.13823.5	CHS.13823.10	POSTN	749	836	0.8959	87.4	79.7	1.10	5577	5233	Blood_Vessel	Blood	2.28	15.00	TRUE	TRUE
CHS.13823.7	CHS.13823.10	POSTN	779	836	0.9318	84.9	79.7	1.07	6681	5233	Blood_Vessel	Blood	24.90	15.00	TRUE	TRUE
CHS.13823.8	CHS.13823.10	POSTN	781	836	0.9342	84	79.7	1.05	6745	5233	Blood_Vessel	Blood	54.98	15.00	TRUE	TRUE
CHS.13825.7	CHS.13825.8	TRPC4	893	977	0.914	74.3	70	1.06	1097	2145	Uterus	Uterus	0.59	8.44	TRUE	TRUE
CHS.14097.1	CHS.14097.4	FAM124A	301	546	0.5513	80.9	61.3	1.32	6772	6088	Prostate	Spleen	1.12	9.32	TRUE	TRUE
CHS.14108.8	CHS.14108.10	DHRS12	317	268	1.1828	95.8	84.9	1.13	9592	2935	Thyroid	Pituitary	17.44	5.33	FALSE	FALSE
CHS.14285.10	CHS.14285.14	PIBF1	688	757	0.9089	81.5	77.1	1.06	13	9260	Thyroid	Thyroid	1.73	17.34	TRUE	TRUE

CHS.14534.28	CHS.14534.12	MBNL2	288	391	0.7366	72	61.6	1.17	217	50	Nerve	Brain	3.73	2.69	TRUE	TRUE
CHS.14638.1	CHS.14638.12	DAOA	82	153	0.5359	71.4	44.9	1.59	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.14638.6	CHS.14638.12	DAOA	84	153	0.549	83.1	44.9	1.85	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.14754.25	CHS.14754.59	ARHGEF 7	646	862	0.7494	70.6	66	1.07	9556	2789	Spleen	Prostate	20.47	20.61	FALSE	TRUE
CHS.14754.46	CHS.14754.59	ARHGEF 7	547	862	0.6346	70.5	66	1.07	3915	2789	Brain	Prostate	4.39	20.61	TRUE	TRUE
CHS.1479.29	CHS.1479.7	MUTYH	478	521	0.9175	84.8	80.2	1.06	10	9580	Blood_V essel	Nerve	3.70	8.56	FALSE	TRUE
CHS.14920.5	CHS.14920.3	KLHL33	533	797	0.6688	87.8	82.6	1.06	872	439	Cervix_ Uteri	Muscle	1.80	11.88	TRUE	FALSE
CHS.14971.5	CHS.14971.31	RPGRIP1	928	1286	0.7216	70.7	67.1	1.05	83	0	Testis	-	7.69	0.00	FALSE	FALSE
CHS.1498.3	CHS.1498.5	PIK3R3	417	461	0.9046	84.8	79.8	1.06	318	8417	Spleen	Adipose _Tissue	11.83	40.10	TRUE	TRUE
CHS.15049.13	CHS.15049.1	PABPN1	178	306	0.5817	70	63	1.11	9787	9787	Testis	Testis	26.41	115.77	TRUE	TRUE
CHS.1509.5	CHS.1509.4	UQCRH	82	91	0.9011	88.2	83.6	1.06	4300	9796	Bone_M arrow	Bone_M arrow	18.43	1693.53	TRUE	TRUE
CHS.15096.8	CHS.15096.7	TSSK4	328	338	0.9704	85.7	79.1	1.08	249	226	Testis	Testis	17.91	17.12	TRUE	TRUE
CHS.1525.4	CHS.1525.18	ATPAF1	240	328	0.7317	81.2	70.5	1.15	447	0	Testis	-	15.15	0.00	TRUE	TRUE
CHS.1525.5	CHS.1525.18	ATPAF1	177	328	0.5396	93	70.5	1.32	755	0	Testis	-	11.88	0.00	FALSE	TRUE
CHS.1525.7	CHS.1525.18	ATPAF1	351	328	1.0701	70.5	70.5	1.00	9781	0	Adrenal _Gland	-	46.47	0.00	TRUE	TRUE
CHS.15449.6	CHS.15449.4	L2HGDH	418	463	0.9028	92.5	85.9	1.08	4	6785	Adipose _Tissue	Blood	1.37	4.08	TRUE	TRUE
CHS.15449.7	CHS.15449.4	L2HGDH	426	463	0.9201	91.6	85.9	1.07	47	6785	Adipose _Tissue	Blood	1.90	4.08	TRUE	TRUE
CHS.15451.15	CHS.15451.8	CDKL1	276	357	0.7731	86.9	82.2	1.06	1440	637	Fallopia n_Tube	Cervix_ Uteri	5.48	7.59	FALSE	FALSE
CHS.15462.4	CHS.15462.3	ABHD12 B	255	362	0.7044	90.1	85.3	1.06	1307	1069	Skin	Skin	6.44	5.34	TRUE	TRUE
CHS.15467.42	CHS.15467.51	TRIM9	718	795	0.9031	80.8	75.6	1.07	3	629	Brain	Skin	3.29	61.31	TRUE	TRUE
CHS.15467.5	CHS.15467.51	TRIM9	710	795	0.8931	82.5	75.6	1.09	3302	629	Brain	Skin	18.73	61.31	TRUE	TRUE
CHS.15467.53	CHS.15467.51	TRIM9	732	795	0.9208	79.8	75.6	1.06	3	629	Brain	Skin	3.46	61.31	TRUE	TRUE
CHS.1547.6	CHS.1547.2	CMPK1	196	228	0.8596	95.7	89.3	1.07	9795	0	Bone_M arrow	-	158.52	0.00	TRUE	TRUE
CHS.1566.30	CHS.1566.5	AGBL4	455	503	0.9046	91.4	85.9	1.06	41	764	Fallopia n_Tube	Brain	3.02	3.53	FALSE	FALSE
CHS.1566.8	CHS.1566.5	AGBL4	464	503	0.9225	90.2	85.9	1.05	61	764	Testis	Brain	1.23	3.53	FALSE	FALSE
CHS.15759.20	CHS.15759.26	CCDC19 6	240	297	0.8081	76.1	67.6	1.13	190	189	Testis	Testis	2.70	7.01	FALSE	FALSE
CHS.15759.4	CHS.15759.26	CCDC19 6	165	297	0.5556	75.4	67.6	1.12	208	189	Testis	Testis	62.28	7.01	FALSE	FALSE
CHS.15759.5	CHS.15759.26	CCDC19 6	215	297	0.7239	77.8	67.6	1.15	199	189	Testis	Testis	29.63	7.01	FALSE	FALSE
CHS.15759.6	CHS.15759.26	CCDC19 6	162	297	0.5455	72.1	67.6	1.07	194	189	Testis	Testis	3.64	7.01	FALSE	FALSE
CHS.15804.40	CHS.15804.10	ACTN1	826	914	0.9037	82.5	77.9	1.06	17	8639	Blood_V essel	Blood_V essel	10.53	127.83	TRUE	TRUE
CHS.15804.49	CHS.15804.10	ACTN1	883	914	0.9661	81.8	77.9	1.05	1	8639	Brain	Blood_V essel	12.19	127.83	TRUE	TRUE
CHS.15938.4	CHS.15938.5	PGF	149	170	0.8765	80.1	76.2	1.05	8985	8726	Thyroid	Thyroid	40.56	37.05	FALSE	FALSE
CHS.16023.4	CHS.16023.5	VIPAS39	444	493	0.9006	82.6	77.4	1.07	5568	9604	Vagina	Testis	0.61	23.98	TRUE	TRUE
CHS.16167.16	CHS.16167.9	RPS6KA 5	549	802	0.6845	75.4	66.4	1.14	6968	6877	Testis	Esophag us	1.54	2.76	TRUE	TRUE
CHS.16181.10	CHS.16181.13	PPP4R3 A	553	833	0.6639	88.8	76.5	1.16	7844	2817	Cervix_ Uteri	Fallopia n_Tube	14.87	5.73	TRUE	TRUE
CHS.1643.3	CHS.1643.5	LRP8	700	963	0.7269	80.2	69.3	1.16	3310	3149	Thyroid	Testis	9.14	19.08	TRUE	TRUE
CHS.1643.34	CHS.1643.5	LRP8	888	963	0.9221	74.5	69.3	1.08	7	3149	Brain	Testis	1.11	19.08	TRUE	TRUE
CHS.1643.7	CHS.1643.5	LRP8	901	963	0.9356	74.1	69.3	1.07	22	3149	Heart	Testis	2.21	19.08	TRUE	TRUE
CHS.16436.27	CHS.16436.20	PPP2R5 C	449	579	0.7755	87.3	78.6	1.11	9771	3291	Adrenal _Gland	Bone_M arrow	15.37	13.24	FALSE	FALSE
CHS.16436.30	CHS.16436.20	PPP2R5 C	485	579	0.8377	85.8	78.6	1.09	9757	3291	Bone_M arrow	Bone_M arrow	13.64	13.24	FALSE	FALSE

CHS.16436.61	CHS.16436.20	PPP2R5C	504	579	0.8705	84.3	78.6	1.07	7378	3291	Bone_Marrow	Bone_Marrow	128.34	13.24	FALSE	FALSE
CHS.16496.1	CHS.16496.18	KLC1	560	639	0.8764	76.4	71.9	1.06	9754	580	Brain	Spleen	68.51	6.81	FALSE	FALSE
CHS.16496.112	CHS.16496.18	KLC1	556	639	0.8701	76.6	71.9	1.07	2778	580	Brain	Spleen	6.45	6.81	FALSE	FALSE
CHS.16496.113	CHS.16496.18	KLC1	565	639	0.8842	75.6	71.9	1.05	3006	580	Nerve	Spleen	9.51	6.81	FALSE	FALSE
CHS.16496.20	CHS.16496.18	KLC1	564	639	0.8826	76.1	71.9	1.06	7988	580	Nerve	Spleen	29.68	6.81	FALSE	FALSE
CHS.16496.60	CHS.16496.18	KLC1	551	639	0.8623	77	71.9	1.07	1632	580	Brain	Spleen	103.38	6.81	FALSE	FALSE
CHS.16496.61	CHS.16496.18	KLC1	542	639	0.8482	78	71.9	1.08	1178	580	Brain	Spleen	11.06	6.81	FALSE	FALSE
CHS.16576.1	CHS.16576.4	BTBD6	485	538	0.9015	86	80.5	1.07	97	9641	Lung	Testis	8.41	117.86	TRUE	TRUE
CHS.16586.14	CHS.16586.16	TEDC1	426	495	0.8606	74.6	67.5	1.11	8970	736	Spleen	Fallopian_Tube	7.02	3.35	FALSE	FALSE
CHS.16586.15	CHS.16586.16	TEDC1	393	495	0.7939	71.1	67.5	1.05	7880	736	Nerve	Fallopian_Tube	2.96	3.35	FALSE	FALSE
CHS.16586.8	CHS.16586.16	TEDC1	385	495	0.7778	74.1	67.5	1.10	1007	736	Blood_Vessel	Fallopian_Tube	3.54	3.35	FALSE	FALSE
CHS.16586.9	CHS.16586.16	TEDC1	302	495	0.6101	77.8	67.5	1.15	1461	736	Brain	Fallopian_Tube	6.16	3.35	FALSE	FALSE
CHS.166726.2	CHS.166726.3	ALDOA	364	418	0.8708	95	87.3	1.09	9786	9696	Bone_Marrow	Muscle	4113.77	238.92	FALSE	FALSE
CHS.16840.1	CHS.16840.2	GABRG3	253	467	0.5418	80.2	74.3	1.08	665	522	Prostate	Brain	3.08	3.30	TRUE	TRUE
CHS.17218.1	CHS.17218.3	CATSPER2	414	530	0.7811	78.2	71.4	1.10	1222	806	Testis	Testis	14.04	6.90	FALSE	FALSE
CHS.17305.11	CHS.17305.5	DUT	164	252	0.6508	87.8	71.8	1.22	9785	9729	Bone_Marrow	Thyroid	91.82	29.94	FALSE	FALSE
CHS.17484.19	CHS.17484.21	RORA	477	523	0.912	77.2	73.4	1.05	5	3411	Skin	Fallopian_Tube	13.08	9.64	TRUE	TRUE
CHS.17484.9	CHS.17484.21	RORA	468	523	0.8948	78	73.4	1.06	7808	3411	Skin	Fallopian_Tube	50.11	9.64	TRUE	TRUE
CHS.17693.2	CHS.17693.7	KIF23	856	974	0.8789	77.5	71.1	1.09	2240	1161	Bone_Marrow	Bone_Marrow	16.06	13.61	FALSE	FALSE
CHS.17711.10	CHS.17711.7	TLE3	698	769	0.9077	70.6	66.5	1.06	1	9571	Blood	Spleen	7.32	11.11	TRUE	TRUE
CHS.17711.21	CHS.17711.7	TLE3	693	769	0.9012	70.9	66.5	1.07	4	9571	Blood	Spleen	2.55	11.11	TRUE	TRUE
CHS.17744.8	CHS.17744.9	LRRC49	642	686	0.9359	80.6	75.8	1.06	344	6933	Testis	Pituitary	5.03	14.74	TRUE	TRUE
CHS.17750.15	CHS.17750.1	THSD4	658	1018	0.6464	76.8	66.7	1.15	4138	561	Cervix_Uteri	Kidney	36.37	46.90	TRUE	TRUE
CHS.18147.2	CHS.18147.7	PDE8A	757	829	0.9131	81.4	77.2	1.05	930	8833	Spleen	Adrenal_Gland	5.82	27.42	FALSE	FALSE
CHS.18267.18	CHS.18267.14	PRC1	566	620	0.9129	81.7	77.2	1.06	2915	6435	Testis	Bone_Marrow	30.48	49.35	FALSE	FALSE
CHS.18267.8	CHS.18267.14	PRC1	562	620	0.9065	82.1	77.2	1.06	600	6435	Bone_Marrow	Bone_Marrow	6.48	49.35	FALSE	FALSE
CHS.18267.9	CHS.18267.14	PRC1	563	620	0.9081	81.9	77.2	1.06	359	6435	Bone_Marrow	Bone_Marrow	4.66	49.35	FALSE	FALSE
CHS.1836.5	CHS.1836.4	UBE2U	226	317	0.7129	81.6	70.4	1.16	202	166	Testis	Testis	0.87	16.20	TRUE	TRUE
CHS.18401.1	CHS.18401.2	PGPEP1L	196	142	1.3803	89.9	85.6	1.05	788	287	Muscle	Esophagus	5.71	3.67	FALSE	FALSE
CHS.18401.6	CHS.18401.2	PGPEP1L	204	142	1.4366	93	85.6	1.09	219	287	Blood_Vessel	Esophagus	1.91	3.67	TRUE	FALSE
CHS.18477.11	CHS.18477.12	TM2D3	221	247	0.8947	83.3	79.2	1.05	9782	9648	Uterus	Ovary	49.91	23.33	TRUE	TRUE
CHS.1852.1	CHS.1852.3	LEPR	958	1165	0.8223	74.6	66	1.13	3677	3179	Bone_Marrow	Pituitary	4.52	2.14	FALSE	TRUE
CHS.1852.8	CHS.1852.3	LEPR	896	1165	0.7691	78.6	66	1.19	5774	3179	Nerve	Pituitary	36.49	2.14	FALSE	TRUE
CHS.1856.17	CHS.1856.5	PDE4B	531	736	0.7215	79.8	67.8	1.18	562	453	Prostate	Prostate	3.02	12.69	TRUE	TRUE
CHS.1856.18	CHS.1856.5	PDE4B	564	736	0.7663	77.3	67.8	1.14	8862	453	Spleen	Prostate	27.25	12.69	TRUE	TRUE
CHS.1856.21	CHS.1856.5	PDE4B	503	736	0.6834	80.5	67.8	1.19	4851	453	Pituitary	Prostate	1.10	12.69	TRUE	TRUE
CHS.18655.4	CHS.18655.3	LUC7L	325	371	0.876	73.9	68.4	1.08	9679	9677	Blood	Blood	79.01	33.96	TRUE	TRUE
CHS.18706.42	CHS.18706.30	CCDC78	319	470	0.6787	77.5	71.2	1.09	886	296	Pituitary	Fallopian_Tube	11.20	89.14	FALSE	FALSE
CHS.18896.30	CHS.18896.16	IL32	178	188	0.9468	77.8	72.9	1.07	225	9628	Lung	Lung	3.96	37.68	FALSE	FALSE

CHS.18993.4	CHS.18993.6	SMIM22	83	88	0.9432	74.7	71	1.05	3265	3226	Colon	Stomach	37.32	37.92	FALSE	FALSE
CHS.19027.28	CHS.19027.1	ABAT	455	500	0.91	97.1	92.3	1.05	396	7338	Pituitary	Liver	6.82	140.92	TRUE	TRUE
CHS.19080.1	CHS.19080.2	CLEC16A	906	1053	0.8604	75.7	70.7	1.07	9010	1546	Testis	Testis	8.83	25.94	TRUE	TRUE
CHS.19080.9	CHS.19080.2	CLEC16A	951	1053	0.9031	74.3	70.7	1.05	12	1546	Testis	Testis	4.38	25.94	FALSE	TRUE
CHS.19161.10	CHS.19161.3	PARN	585	639	0.9155	84.3	79.7	1.06	85	9568	Uterus	Testis	1.82	34.05	TRUE	TRUE
CHS.19161.6	CHS.19161.3	PARN	577	639	0.903	85.3	79.7	1.07	5	9568	Testis	Testis	1.24	34.05	TRUE	TRUE
CHS.19191.12	CHS.19191.2	BMERB1	187	204	0.9167	77.6	73.4	1.06	1757	9417	Testis	Brain	3.02	109.02	TRUE	TRUE
CHS.19260.7	CHS.19260.5	TMC5	760	1006	0.7555	76.6	63.1	1.21	1267	1225	Testis	Small_Intestine	7.67	50.17	TRUE	FALSE
CHS.19296.20	CHS.19296.14	LDAF1	158	161	0.9814	77.8	68.4	1.14	218	9432	Vagina	Muscle	6.02	14.73	TRUE	TRUE
CHS.19518.4	CHS.19518.8	SEZ6L2	840	923	0.9101	76.3	71.8	1.06	87	5007	Brain	Pituitary	2.98	54.59	TRUE	TRUE
CHS.19518.5	CHS.19518.8	SEZ6L2	853	923	0.9242	76.1	71.8	1.06	3193	5007	Pituitary	Pituitary	25.91	54.59	TRUE	TRUE
CHS.19620.2	CHS.19620.3	PYCARD	176	195	0.9026	81	76	1.07	8903	9729	Spleen	Spleen	35.22	75.08	TRUE	TRUE
CHS.19650.30	CHS.19650.2	KRBOX5	475	126	3.7698	79.6	67.6	1.18	0	9524	-	Bone_Marrow	0.00	3.65	FALSE	FALSE
CHS.1966.4	CHS.1966.3	ACADM	385	421	0.9145	97.1	91.9	1.06	9495	9755	Bone_Marrow	Muscle	6.10	97.42	FALSE	FALSE
CHS.19740.6	CHS.19740.5	MYLK3	478	819	0.5836	72.4	59.6	1.21	1015	602	Heart	Heart	15.60	22.56	FALSE	FALSE
CHS.19947.3	CHS.19947.4	MT1G	61	62	0.9839	74.8	69.7	1.07	9035	9387	Liver	Liver	833.93	1739.90	FALSE	FALSE
CHS.19992.14	CHS.19992.23	KIFC3	768	826	0.9298	78.8	73.9	1.07	2662	8438	Kidney	Nerve	18.33	36.67	TRUE	TRUE
CHS.19992.42	CHS.19992.23	KIFC3	795	826	0.9625	77.7	73.9	1.05	1072	8438	Kidney	Nerve	13.65	36.67	FALSE	TRUE
CHS.20040.9	CHS.20040.4	CDH8	532	799	0.6658	82.8	78.1	1.06	1154	1114	Brain	Brain	5.15	3.20	TRUE	TRUE
CHS.20082.9	CHS.20082.7	TK2	168	265	0.634	90.8	83.5	1.09	9706	9703	Testis	Adrenal_Gland	15.83	18.08	TRUE	TRUE
CHS.20141.4	CHS.20141.2	AGRP	100	132	0.7576	71.6	67.3	1.06	900	892	Lung	Adrenal_Gland	5.35	12.78	TRUE	TRUE
CHS.20155.10	CHS.20155.9	TSNAXIP1	366	712	0.514	90.3	79.7	1.13	5140	891	Fallopian_Tube	Fallopian_Tube	7.24	6.23	TRUE	TRUE
CHS.20155.8	CHS.20155.9	TSNAXIP1	658	712	0.9242	84.3	79.7	1.06	850	891	Testis	Fallopian_Tube	23.25	6.23	TRUE	TRUE
CHS.20296.2	CHS.20296.1	HPR	228	348	0.6552	92.3	85.3	1.08	6037	591	Heart	Liver	46.63	483.65	FALSE	FALSE
CHS.2046.1	CHS.2046.9	PRKACB	257	398	0.6457	93.3	87.2	1.07	7199	6451	Brain	Spleen	2.06	13.58	TRUE	TRUE
CHS.2046.2	CHS.2046.9	PRKACB	351	398	0.8819	93.8	87.2	1.08	8812	6451	Bladder	Spleen	30.78	13.58	TRUE	TRUE
CHS.20525.10	CHS.20525.2	CIBAR2	277	304	0.9112	85.1	80.5	1.06	3	702	Lung	Fallopian_Tube	0.90	26.18	FALSE	FALSE
CHS.2060.21	CHS.2060.5	SSX2IP	560	614	0.9121	70.4	66.9	1.05	3	6035	Testis	Bone_Marrow	8.99	18.55	FALSE	FALSE
CHS.2095.62	CHS.2095.10	ODF2L	513	644	0.7966	85.7	80.4	1.07	3324	2398	Breast	Fallopian_Tube	5.00	8.76	FALSE	FALSE
CHS.21162.6	CHS.21162.13	DLG4	664	724	0.9171	80.9	76.6	1.06	8370	9582	Brain	Brain	2.86	95.24	TRUE	TRUE
CHS.21163.3	CHS.21163.4	ACADVL	579	655	0.884	95.7	89.2	1.07	9786	9776	Adrenal_Gland	Muscle	197.06	490.50	TRUE	TRUE
CHS.21181.7	CHS.21181.8	TMEM95	184	176	1.0455	76.4	65.1	1.17	90	193	Testis	Testis	0.88	8.50	TRUE	TRUE
CHS.21217.15	CHS.21217.5	TP53	234	393	0.5954	76.6	72.4	1.06	9647	7015	Adrenal_Gland	Cervix_Uteri	0.24	5.31	TRUE	FALSE
CHS.21319.5	CHS.21319.8	GAS7	336	476	0.7059	88.7	82.4	1.08	9086	8274	Brain	Cervix_Uteri	46.21	32.36	TRUE	TRUE
CHS.21359.5	CHS.21359.2	ARHGAP44	591	818	0.7225	76.2	62.9	1.21	2123	1968	Testis	Brain	0.83	11.80	FALSE	FALSE
CHS.21410.34	CHS.21410.42	TRIM16	348	564	0.617	85.1	80.3	1.06	2452	1865	Vagina	Vagina	1.75	29.31	FALSE	FALSE
CHS.2144.5	CHS.2144.6	KYAT3	420	454	0.9251	94.6	90	1.05	9640	9588	Bone_Marrow	Nerve	24.80	25.37	FALSE	FALSE
CHS.21533.14	CHS.21533.5	GRAPL	116	118	0.9831	79.2	70.8	1.12	49	2762	Spleen	Spleen	2.96	5.10	FALSE	FALSE
CHS.21533.4	CHS.21533.5	GRAPL	144	118	1.2203	78	70.8	1.10	185	2762	Spleen	Spleen	2.47	5.10	FALSE	FALSE
CHS.21533.9	CHS.21533.5	GRAPL	100	118	0.8475	87.5	70.8	1.24	3374	2762	Spleen	Spleen	4.09	5.10	FALSE	FALSE

CHS.21574.30	CHS.21574.4	SPECC1	703	1068	0.6582	74.3	66.4	1.12	5672	4786	Ovary	Ovary	1.38	8.72	FALSE	FALSE
CHS.21574.31	CHS.21574.4	SPECC1	709	1068	0.6639	74	66.4	1.11	6964	4786	Testis	Ovary	4.43	8.72	TRUE	FALSE
CHS.21580.13	CHS.21580.2	LGALS9B	324	356	0.9101	87.1	82.3	1.06	54	437	Spleen	Stomach	3.37	22.13	FALSE	FALSE
CHS.21580.7	CHS.21580.2	LGALS9B	323	356	0.9073	87	82.3	1.06	121	437	Blood	Stomach	2.40	22.13	FALSE	FALSE
CHS.21659.7	CHS.21659.6	LGALS9	323	355	0.9099	87.5	82.8	1.06	9570	8849	Spleen	Colon	55.07	45.29	FALSE	FALSE
CHS.21834.4	CHS.21834.3	ASIC2	512	563	0.9094	84.7	77.9	1.09	827	1440	Brain	Uterus	6.76	9.17	TRUE	TRUE
CHS.21924.5	CHS.21924.3	C17orf50	173	174	0.9943	71.3	57.6	1.24	169	268	Testis	Testis	11.23	129.73	TRUE	TRUE
CHS.21930.2	CHS.21930.3	CCL5	91	154	0.5909	84.6	52.1	1.62	8848	3021	Blood	Blood	125.77	43.67	FALSE	FALSE
CHS.21934.15	CHS.21934.12	RDM1	236	284	0.831	83.3	76.3	1.09	987	591	Bone_Marrow	Testis	2.98	8.16	FALSE	FALSE
CHS.21957.2	CHS.21957.4	CCL4L2	64	103	0.6214	84.2	50.3	1.67	6869	89	Spleen	Blood	6.23	7.91	FALSE	FALSE
CHS.21957.8	CHS.21957.4	CCL4L2	87	103	0.8447	84.3	50.3	1.68	2224	89	Blood	Blood	12.18	7.91	FALSE	FALSE
CHS.22052.23	CHS.22052.2	CACNB1	478	598	0.7993	71.6	64.4	1.11	5253	5249	Muscle	Brain	2.18	24.09	TRUE	TRUE
CHS.22154.5	CHS.22154.4	KRT23	285	422	0.6754	84.4	77.2	1.09	1655	259	Testis	Adipose_Tissue	15.44	17.68	FALSE	FALSE
CHS.22356.1	CHS.22356.2	MPP3	315	585	0.5385	83.4	77.6	1.07	7559	2261	Heart	Brain	16.86	17.77	FALSE	FALSE
CHS.22358.9	CHS.22358.21	CD300LG	247	332	0.744	76.9	66.3	1.16	4523	3871	Adipose_Tissue	Adipose_Tissue	43.34	64.01	TRUE	TRUE
CHS.22367.5	CHS.22367.6	PYY	90	97	0.9278	75.1	69.2	1.09	678	616	Colon	Colon	30.95	66.63	TRUE	TRUE
CHS.2237.8	CHS.2237.7	FNBP1L	551	605	0.9107	81.1	76.6	1.06	8668	623	Thyroid	Brain	24.40	5.47	TRUE	TRUE
CHS.2237.9	CHS.2237.7	FNBP1L	547	605	0.9041	81.4	76.6	1.06	8640	623	Breast	Brain	17.20	5.47	TRUE	TRUE
CHS.22680.1	CHS.22680.2	SPATA20	742	802	0.9252	92.3	87.7	1.05	9581	9565	Thyroid	Thyroid	74.82	46.54	FALSE	FALSE
CHS.22691.2	CHS.22691.1	ANKRD40CL	119	114	1.0439	80.7	74.5	1.08	212	346	Colon	Small_Intestine	6.12	26.52	FALSE	FALSE
CHS.22734.2	CHS.22734.4	TOM1L1	346	476	0.7269	76.3	65.1	1.17	8572	8341	Thyroid	Thyroid	3.36	36.28	FALSE	FALSE
CHS.22852.1	CHS.22852.2	RPS6KB1	472	525	0.899	76.4	72.2	1.06	2567	0	Cervix_Uteri	-	5.04	0.00	TRUE	TRUE
CHS.22852.3	CHS.22852.2	RPS6KB1	451	525	0.859	77.4	72.2	1.07	1479	0	Pituitary	-	1.16	0.00	TRUE	TRUE
CHS.22937.12	CHS.22937.9	STRADA	306	431	0.71	83.1	78.5	1.06	8462	7130	Testis	Pituitary	2.68	15.37	TRUE	TRUE
CHS.22937.22	CHS.22937.9	STRADA	285	431	0.6613	86.3	78.5	1.10	8519	7130	Thyroid	Pituitary	1.37	15.37	TRUE	TRUE
CHS.22937.31	CHS.22937.9	STRADA	266	431	0.6172	85.7	78.5	1.09	8168	7130	Nerve	Pituitary	1.28	15.37	TRUE	TRUE
CHS.22937.4	CHS.22937.9	STRADA	394	431	0.9142	83	78.5	1.06	8742	7130	Testis	Pituitary	7.48	15.37	TRUE	TRUE
CHS.22937.5	CHS.22937.9	STRADA	348	431	0.8074	83.6	78.5	1.07	8743	7130	Nerve	Pituitary	4.57	15.37	TRUE	TRUE
CHS.22937.8	CHS.22937.9	STRADA	373	431	0.8654	86	78.5	1.10	8796	7130	Testis	Pituitary	31.08	15.37	TRUE	TRUE
CHS.22945.3	CHS.22945.5	SMARCD2	483	531	0.9096	78.2	74.1	1.06	7835	9743	Skin	Skin	1.94	101.80	TRUE	TRUE
CHS.2297.8	CHS.2297.6	PTBP2	480	531	0.904	79.4	71.2	1.12	2476	7981	Uterus	Testis	0.36	17.01	TRUE	TRUE
CHS.23027.1	CHS.23027.2	PITPNC1	268	332	0.8072	91	80.9	1.12	9643	8317	Brain	Brain	15.52	3.03	TRUE	TRUE
CHS.23152.1	CHS.23152.2	CPSF4L	190	179	1.0615	85.1	80.6	1.06	12	24	Testis	Breast	1.91	2.10	FALSE	FALSE
CHS.23152.3	CHS.23152.2	CPSF4L	104	179	0.581	86.1	80.6	1.07	40	24	Thyroid	Breast	1.59	2.10	FALSE	FALSE
CHS.23235.6	CHS.23235.3	SLC16A5	460	505	0.9109	84.4	79.3	1.06	4754	8323	Kidney	Kidney	34.99	11.96	FALSE	FALSE
CHS.23357.35	CHS.23357.13	SEPTIN9	335	568	0.5898	84.2	63.1	1.33	9625	9227	Spleen	Cervix_Uteri	10.61	67.63	TRUE	TRUE
CHS.2339.8	CHS.2339.12	CDC14A	383	594	0.6448	88.9	70.3	1.26	6343	6326	Breast	Fallopian_Tube	0.41	11.00	TRUE	TRUE
CHS.23394.4	CHS.23394.6	C17orf99	247	265	0.9321	78.5	73.6	1.07	143	328	Bone_Marrow	Bone_Marrow	2.19	7.17	FALSE	FALSE
CHS.23425.29	CHS.23425.19	CYTH1	372	398	0.9347	88.2	84	1.05	1550	9741	Vagina	Spleen	14.47	55.55	TRUE	TRUE
CHS.237.25	CHS.237.24	KCNAB2	300	415	0.7229	91.4	85	1.08	7083	216	Kidney	Brain	15.35	0.59	FALSE	FALSE

CHS.23961.4	CHS.23961.5	COLEC12	725	742	0.9771	70.6	65.2	1.08	534	7902	Cervix_Uteri	Cervix_Uteri	3.84	74.03	TRUE	TRUE
CHS.24209.1	CHS.24209.2	SEH1L	360	421	0.8551	85.7	78.2	1.10	9587	8314	Bone_Marrow	Muscle	64.05	20.41	TRUE	TRUE
CHS.24209.4	CHS.24209.2	SEH1L	380	421	0.9026	83.2	78.2	1.06	124	8314	Testis	Muscle	2.08	20.41	TRUE	TRUE
CHS.24244.1	CHS.24244.2	POTEC	384	542	0.7085	72.4	64.7	1.12	196	192	Testis	Testis	1.36	1.86	FALSE	FALSE
CHS.24423.12	CHS.24423.4	KLHL14	408	628	0.6497	94	87.6	1.07	768	489	Thyroid	Fallopian_Tube	13.52	20.26	TRUE	TRUE
CHS.24436.11	CHS.24436.114	DTNA	683	770	0.887	70.4	67	1.05	5728	30	Bladder	Brain	13.59	4.15	TRUE	FALSE
CHS.24436.2	CHS.24436.114	DTNA	513	770	0.6662	79.1	67	1.18	56	30	Brain	Brain	16.48	4.15	FALSE	FALSE
CHS.24436.32	CHS.24436.114	DTNA	510	770	0.6623	79.5	67	1.19	6706	30	Brain	Brain	21.26	4.15	TRUE	FALSE
CHS.24436.36	CHS.24436.114	DTNA	540	770	0.7013	77.6	67	1.16	71	30	Brain	Brain	14.14	4.15	FALSE	FALSE
CHS.24436.60	CHS.24436.114	DTNA	567	770	0.7364	75.8	67	1.13	1999	30	Pituitary	Brain	7.85	4.15	FALSE	FALSE
CHS.24550.5	CHS.24550.3	ATP5F1A	503	553	0.9096	91.9	86.6	1.06	9727	9786	Testis	Heart	69.09	287.25	FALSE	TRUE
CHS.24565.41	CHS.24565.3	PIAS2	507	621	0.8164	75.1	67.3	1.12	8914	8895	Testis	Fallopian_Tube	2.45	4.18	FALSE	FALSE
CHS.24584.3	CHS.24584.2	SMAD2	437	467	0.9358	80.3	75.5	1.06	8911	9762	Thyroid	Thyroid	1.28	4.36	TRUE	TRUE
CHS.2465.19	CHS.2465.1	AMPD2	761	825	0.9224	83.1	78.8	1.05	7514	8951	Pituitary	Brain	1.61	10.57	FALSE	FALSE
CHS.24667.22	CHS.24667.4	DCC	965	1447	0.6669	88.2	68	1.30	137	84	Testis	Brain	5.84	3.09	TRUE	TRUE
CHS.24849.3	CHS.24849.4	CDH7	630	785	0.8025	85.5	79	1.08	647	644	Nerve	Skin	0.37	6.18	TRUE	TRUE
CHS.2497.1	CHS.2497.13	LAMTOR5	90	91	0.989	87.4	66.1	1.32	6959	0	Testis	-	18.07	0.00	FALSE	TRUE
CHS.2497.12	CHS.2497.13	LAMTOR5	72	91	0.7912	93.8	66.1	1.42	1	0	Blood	-	5.23	0.00	TRUE	TRUE
CHS.25065.2	CHS.25065.8	SLC66A2	253	271	0.9336	86.5	81.4	1.06	6013	7805	Testis	Nerve	81.91	72.15	TRUE	TRUE
CHS.25249.16	CHS.25249.15	IZUMO4	214	232	0.9224	83.2	67.4	1.23	7438	8646	Testis	Testis	173.13	366.17	TRUE	FALSE
CHS.25482.9	CHS.25482.6	TRIP10	545	601	0.9068	82.1	78	1.05	9633	8883	Lung	Muscle	91.28	217.40	TRUE	TRUE
CHS.25532.8	CHS.25532.7	CD209	404	404	1	71.3	67.6	1.05	330	2505	Vagina	Fallopian_Tube	4.72	12.12	TRUE	TRUE
CHS.25533.12	CHS.25533.7	CLEC4M	376	399	0.9424	71.6	66.9	1.07	0	807	-	Liver	0.00	18.25	FALSE	FALSE
CHS.25533.2	CHS.25533.7	CLEC4M	371	399	0.9298	78.9	66.9	1.18	20	807	Breast	Liver	2.15	18.25	FALSE	FALSE
CHS.25663.17	CHS.25663.6	PDE4A	647	886	0.7302	70.1	62.2	1.13	3041	2188	Brain	Brain	9.38	10.00	TRUE	TRUE
CHS.25676.1	CHS.25676.4	ILF3	706	898	0.7862	70.4	61.8	1.14	9780	9779	Bone_Marrow	Uterus	64.09	71.96	TRUE	TRUE
CHS.25676.2	CHS.25676.4	ILF3	702	898	0.7817	70.5	61.8	1.14	9779	9779	Bone_Marrow	Uterus	65.70	71.96	TRUE	TRUE
CHS.25709.2	CHS.25709.1	PLPPR2	343	452	0.7588	78.8	66.2	1.19	9748	1447	Brain	Pituitary	40.45	21.04	FALSE	FALSE
CHS.25735.3	CHS.25735.11	ZNF439	363	504	0.7202	78.2	66.8	1.17	4822	2508	Testis	Bladder	4.09	5.49	FALSE	FALSE
CHS.25978.23	CHS.25978.6	EPS15L1	601	910	0.6604	75.1	64.7	1.16	9719	9703	Skin	Brain	8.15	17.48	FALSE	FALSE
CHS.26148.2	CHS.26148.1	YJEFN3	249	299	0.8328	92.8	81.6	1.14	5018	3358	Brain	Testis	74.26	24.97	FALSE	FALSE
CHS.26329.4	CHS.26329.8	C19orf12	77	141	0.5461	88.4	54.3	1.63	8487	7951	Adipose_Tissue	Muscle	13.57	5.59	FALSE	FALSE
CHS.26383.1	CHS.26383.2	WDR88	426	472	0.9025	77.8	73.9	1.05	192	193	Testis	Testis	6.26	9.98	FALSE	TRUE
CHS.26463.1	CHS.26463.2	MAG	582	626	0.9297	84.9	80.5	1.05	1642	1892	Nerve	Brain	47.30	104.75	FALSE	FALSE
CHS.26577.2	CHS.26577.3	ZNF585A	714	769	0.9285	77.1	70.8	1.09	4058	6541	Bladder	Brain	2.73	2.10	FALSE	FALSE
CHS.26623.2	CHS.26623.7	YIF1B	283	314	0.9013	73.6	69.4	1.06	8585	9695	Bone_Marrow	Bone_Marrow	9.88	17.98	TRUE	TRUE
CHS.26656.5	CHS.26656.4	SIRT2	234	389	0.6015	85.9	79.7	1.08	9709	9651	Testis	Testis	3.16	20.25	TRUE	TRUE
CHS.26656.6	CHS.26656.4	SIRT2	352	389	0.9049	84.4	79.7	1.06	9782	9651	Brain	Testis	115.77	20.25	TRUE	TRUE
CHS.26722.14	CHS.26722.3	ZNF546	798	836	0.9545	74.3	69.7	1.07	2098	980	Cervix_Uteri	Cervix_Uteri	3.94	2.27	FALSE	FALSE
CHS.26722.4	CHS.26722.3	ZNF546	810	836	0.9689	73.5	69.7	1.05	1663	980	Testis	Cervix_Uteri	2.50	2.27	FALSE	FALSE

CHS.26732.5	CHS.26732.2	CCNP	329	307	1.0717	78.8	71.6	1.10	806	900	Fallopian_Tube	Fallopian_Tube	0.55	2.94	FALSE	FALSE
CHS.26852.15	CHS.26852.6	LIPE	821	1076	0.763	74.6	64.1	1.16	8175	4537	Adipose_Tissue	Testis	65.25	39.41	FALSE	FALSE
CHS.26947.3	CHS.26947.2	ZNF229	479	825	0.5806	83.8	58.5	1.43	3047	2226	Thyroid	Fallopian_Tube	3.18	2.32	FALSE	FALSE
CHS.26967.6	CHS.26967.7	CBLC	431	474	0.9093	85.3	79.3	1.08	2	2990	Thyroid	Skin	2.32	44.49	FALSE	FALSE
CHS.26970.2	CHS.26970.3	NECTIN2	479	538	0.8903	79	73.6	1.07	9642	9297	Lung	Testis	67.14	66.36	TRUE	TRUE
CHS.26987.4	CHS.26987.5	NKPD1	657	832	0.7897	74	64.6	1.15	249	167	Skin	Skin	2.85	4.02	TRUE	TRUE
CHS.26991.2	CHS.26991.3	EXOC3L2	476	802	0.5935	87.6	74	1.18	4100	238	Thyroid	Uterus	38.24	13.82	TRUE	TRUE
CHS.27152.11	CHS.27152.14	LIG1	851	919	0.926	75.9	71.8	1.06	1031	9038	Bladder	Bone_Marrow	0.43	30.04	TRUE	TRUE
CHS.27157.39	CHS.27157.60	CARD8	487	537	0.9069	75.2	69.6	1.08	8996	9115	Thyroid	Bone_Marrow	2.23	2.05	FALSE	FALSE
CHS.27157.49	CHS.27157.60	CARD8	486	537	0.905	75.7	69.6	1.09	20	9115	Blood	Bone_Marrow	5.64	2.05	FALSE	FALSE
CHS.27250.14	CHS.27250.1	PIH1D1	273	290	0.9414	89.7	81	1.11	1042	9779	Spleen	Bone_Marrow	6.37	133.37	FALSE	FALSE
CHS.27273.19	CHS.27273.13	PRMT1	342	371	0.9218	92.4	87.8	1.05	4237	9741	Cervix_Uteri	Muscle	10.89	32.93	FALSE	TRUE
CHS.27353.31	CHS.27353.28	KLK4	205	254	0.8071	92.8	87.8	1.06	1233	28	Prostate	Uterus	401.88	11.46	TRUE	TRUE
CHS.27353.6	CHS.27353.28	KLK4	159	254	0.626	95.6	87.8	1.09	1233	28	Prostate	Uterus	401.88	11.46	FALSE	TRUE
CHS.27358.3	CHS.27358.5	KLK8	139	260	0.5346	94.3	87.8	1.07	1660	1327	Esophagus	Skin	13.49	25.49	TRUE	TRUE
CHS.27404.4	CHS.27404.3	SIGLEC5	500	551	0.9074	79.7	75.7	1.05	8	83	Blood	Blood	6.79	73.75	FALSE	FALSE
CHS.27443.2	CHS.27443.1	ZNF480	492	535	0.9196	71.9	67.4	1.07	5449	9666	Bladder	Bone_Marrow	0.26	7.30	FALSE	FALSE
CHS.27564.27	CHS.27564.9	LILRB1	599	652	0.9187	75.8	71.5	1.06	6	3408	Blood	Fallopian_Tube	1.33	3.22	FALSE	FALSE
CHS.27564.28	CHS.27564.9	LILRB1	598	652	0.9172	76	71.5	1.06	4	3408	Blood	Fallopian_Tube	3.85	3.22	FALSE	FALSE
CHS.27572.1	CHS.27572.2	KIR2DL4	273	342	0.7982	82	66.5	1.23	615	446	Spleen	Spleen	4.06	1.53	FALSE	FALSE
CHS.27581.1	CHS.27581.3	GP6	321	620	0.5177	76.7	49.3	1.56	226	216	Bone_Marrow	Brain	0.91	2.01	FALSE	FALSE
CHS.27581.2	CHS.27581.3	GP6	339	620	0.5468	74.5	49.3	1.51	488	216	Bone_Marrow	Brain	5.54	2.01	FALSE	FALSE
CHS.27587.2	CHS.27587.4	TNNT1	251	278	0.9029	77.4	73	1.06	6705	4424	Muscle	Muscle	1676.32	608.39	FALSE	FALSE
CHS.27662.3	CHS.27662.1	NLRP8	891	1048	0.8502	85.5	80.5	1.06	11	0	Breast	-	1.49	0.00	FALSE	FALSE
CHS.2847.4	CHS.2847.6	HJV	313	426	0.7347	85.1	80.8	1.05	1673	1299	Muscle	Muscle	89.92	54.80	TRUE	TRUE
CHS.28605.26	CHS.28605.25	RNASEH1	169	286	0.5909	91.6	82.2	1.11	6383	0	Testis	-	4.75	0.00	TRUE	TRUE
CHS.28610.12	CHS.28610.3	COLEC11	245	271	0.9041	80.3	76.3	1.05	2747	6233	Kidney	Liver	5.82	60.25	TRUE	TRUE
CHS.28738.17	CHS.28738.13	RRM2	339	389	0.8715	86.6	80.3	1.08	151	0	Spleen	-	2.87	0.00	TRUE	TRUE
CHS.28754.10	CHS.28754.9	ATP6V1C2	381	427	0.8923	91.2	84.5	1.08	1876	1490	Kidney	Skin	24.93	44.51	FALSE	FALSE
CHS.29005.1	CHS.29005.2	FAM166C	184	201	0.9154	78.6	74.5	1.06	420	1681	Testis	Testis	19.70	20.13	TRUE	TRUE
CHS.29020.25	CHS.29020.11	AGBL5	801	886	0.9041	71.9	67.6	1.06	68	8231	Pituitary	Testis	6.38	158.74	TRUE	TRUE
CHS.29020.8	CHS.29020.11	AGBL5	717	886	0.8093	76.6	67.6	1.13	9648	8231	Bone_Marrow	Testis	32.15	158.74	TRUE	TRUE
CHS.29020.9	CHS.29020.11	AGBL5	805	886	0.9086	71.7	67.6	1.06	39	8231	Testis	Testis	6.16	158.74	TRUE	TRUE
CHS.29035.9	CHS.29035.10	SLC30A3	412	388	1.0619	79.4	74.6	1.06	877	2077	Testis	Brain	11.54	33.14	TRUE	TRUE
CHS.29099.18	CHS.29099.7	TOGARAM2	546	1019	0.5358	82.5	61.4	1.34	2257	835	Muscle	Fallopian_Tube	9.98	16.29	TRUE	FALSE
CHS.29198.14	CHS.29198.4	VIT	629	693	0.9076	80.9	76.5	1.06	487	5679	Nerve	Nerve	2.61	36.05	FALSE	FALSE
CHS.29198.31	CHS.29198.4	VIT	628	693	0.9062	80.4	76.5	1.05	1	5679	Thyroid	Nerve	0.83	36.05	FALSE	FALSE
CHS.29315.6	CHS.29315.11	MTA3	515	594	0.867	82.3	77.3	1.06	9473	9093	Adrenal_Gland	Uterus	23.13	11.66	TRUE	TRUE
CHS.2934.4	CHS.2934.11	NOTCH2NLC	236	293	0.8055	85.2	77.6	1.10	1809	579	Testis	Cervix_Uteri	21.14	20.55	TRUE	TRUE
CHS.29447.5	CHS.29447.6	FBXO11	843	927	0.9094	86.9	82.4	1.05	177	9656	Liver	Uterus	4.86	51.68	FALSE	FALSE

CHS.29500.29	CHS.29500.16	ACYP2	99	172	0.5756	89.1	61.7	1.44	9775	8724	Muscle	Adrenal_Gland	61.31	10.33	FALSE	FALSE
CHS.29631.41	CHS.29631.7	WDPCP	673	746	0.9021	81.9	75.7	1.08	44	6780	Prostate	Fallopian_Tube	1.94	3.92	FALSE	FALSE
CHS.29631.42	CHS.29631.7	WDPCP	693	746	0.929	79.9	75.7	1.06	34	6780	Lung	Fallopian_Tube	2.56	3.92	FALSE	FALSE
CHS.29770.23	CHS.29770.5	NFU1	230	254	0.9055	82.2	77.9	1.06	7919	9777	Pituitary	Muscle	4.90	35.01	FALSE	TRUE
CHS.29800.4	CHS.29800.9	FAM136A	138	245	0.5633	95.2	62.4	1.53	9772	8842	Bone_Marrow	Bone_Marrow	48.12	6.96	TRUE	TRUE
CHS.29878.1	CHS.29878.4	DGUOK	180	277	0.6498	93.6	85.7	1.09	9789	9775	Pituitary	Testis	35.29	34.68	FALSE	TRUE
CHS.29904.8	CHS.29904.7	CCDC142	566	750	0.7547	82.7	71.7	1.15	1154	5	Kidney	Nerve	7.95	1.28	FALSE	TRUE
CHS.29929.2	CHS.29929.1	TACR1	311	407	0.7641	79.4	71.3	1.11	3535	3242	Cervix_Uteri	Cervix_Uteri	1.21	8.03	TRUE	TRUE
CHS.29951.11	CHS.29951.3	LRRTM4	518	590	0.878	77.9	71.6	1.09	1195	934	Nerve	Brain	2.01	3.33	TRUE	TRUE
CHS.30186.13	CHS.30186.1	KCNIP3	230	256	0.8984	72.6	66.2	1.10	7478	2498	Testis	Brain	25.01	23.10	TRUE	TRUE
CHS.30265.21	CHS.30265.13	ZAP70	312	619	0.504	86	78.6	1.09	7353	5009	Spleen	Spleen	3.80	40.91	FALSE	FALSE
CHS.3029.13	CHS.3029.7	MINDY1	327	469	0.6972	79	71.4	1.11	8742	6237	Thyroid	Esophagus	4.62	11.79	TRUE	FALSE
CHS.30350.14	CHS.30350.2	IL1RL1	328	556	0.5899	87	81	1.07	5393	1498	Lung	Lung	168.15	3.55	FALSE	FALSE
CHS.30490.17	CHS.30490.8	RGPD6	904	1765	0.5122	81.6	65.8	1.24	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.30565.4	CHS.30565.1	IL36B	157	164	0.9573	93	50.2	1.85	775	1	Skin	Skin	8.38	0.01	FALSE	FALSE
CHS.30584.13	CHS.30584.10	CBWD2	359	395	0.9089	79	75	1.05	9721	9768	Bone_Marrow	Thyroid	12.43	17.84	TRUE	TRUE
CHS.30750.10	CHS.30750.19	TEX51	170	166	1.0241	78.1	61.9	1.26	62	207	Testis	Testis	3.56	42.71	FALSE	FALSE
CHS.30750.11	CHS.30750.19	TEX51	227	166	1.3675	71.7	61.9	1.16	6	207	Testis	Testis	3.23	42.71	FALSE	FALSE
CHS.30750.22	CHS.30750.19	TEX51	198	166	1.1928	76.9	61.9	1.24	190	207	Testis	Testis	30.34	42.71	FALSE	FALSE
CHS.30750.26	CHS.30750.19	TEX51	197	166	1.1867	74.1	61.9	1.20	31	207	Testis	Testis	3.69	42.71	FALSE	FALSE
CHS.30750.27	CHS.30750.19	TEX51	175	166	1.0542	71.9	61.9	1.16	37	207	Testis	Testis	3.45	42.71	FALSE	FALSE
CHS.30750.3	CHS.30750.19	TEX51	190	166	1.1446	76.5	61.9	1.24	15	207	Testis	Testis	3.40	42.71	FALSE	FALSE
CHS.30750.4	CHS.30750.19	TEX51	162	166	0.9759	72.4	61.9	1.17	9	207	Testis	Testis	2.28	42.71	FALSE	FALSE
CHS.30750.5	CHS.30750.19	TEX51	156	166	0.9398	75.2	61.9	1.21	48	207	Testis	Testis	2.88	42.71	FALSE	FALSE
CHS.30750.8	CHS.30750.19	TEX51	204	166	1.2289	79.2	61.9	1.28	30	207	Testis	Testis	3.30	42.71	FALSE	FALSE
CHS.30750.9	CHS.30750.19	TEX51	203	166	1.2229	79	61.9	1.28	19	207	Testis	Testis	3.75	42.71	FALSE	FALSE
CHS.30752.20	CHS.30752.33	BIN1	409	593	0.6897	78.6	65.4	1.20	9664	3171	Nerve	Brain	23.46	41.98	FALSE	FALSE
CHS.30752.21	CHS.30752.33	BIN1	439	593	0.7403	75.6	65.4	1.16	9714	3171	Adrenal_Gland	Brain	50.33	41.98	FALSE	FALSE
CHS.30752.22	CHS.30752.33	BIN1	475	593	0.801	72.1	65.4	1.10	6791	3171	Nerve	Brain	13.25	41.98	FALSE	FALSE
CHS.30752.23	CHS.30752.33	BIN1	482	593	0.8128	71.4	65.4	1.09	9684	3171	Cervix_Uteri	Brain	53.51	41.98	FALSE	FALSE
CHS.30752.26	CHS.30752.33	BIN1	424	593	0.715	77.3	65.4	1.18	8437	3171	Muscle	Brain	450.85	41.98	FALSE	FALSE
CHS.30752.27	CHS.30752.33	BIN1	454	593	0.7656	73.8	65.4	1.13	8467	3171	Muscle	Brain	401.37	41.98	FALSE	FALSE
CHS.30752.28	CHS.30752.33	BIN1	497	593	0.8381	70.5	65.4	1.08	8446	3171	Muscle	Brain	134.54	41.98	FALSE	FALSE
CHS.30752.30	CHS.30752.33	BIN1	506	593	0.8533	72.5	65.4	1.11	5000	3171	Muscle	Brain	10.00	41.98	FALSE	FALSE
CHS.30758.1	CHS.30758.2	CYP27C1	416	537	0.7747	90.9	85.3	1.07	357	193	Bladder	Cervix_Uteri	4.83	7.77	FALSE	FALSE
CHS.30758.12	CHS.30758.2	CYP27C1	447	537	0.8324	90	85.3	1.06	357	193	Bladder	Cervix_Uteri	4.83	7.77	FALSE	FALSE
CHS.30758.4	CHS.30758.2	CYP27C1	372	537	0.6927	90.7	85.3	1.06	418	193	Cervix_Uteri	Cervix_Uteri	3.39	7.77	FALSE	FALSE
CHS.30833.15	CHS.30833.6	IMP4	206	291	0.7079	90.8	85	1.07	9790	9775	Brain	Bone_Marrow	14.55	30.92	TRUE	TRUE
CHS.30878.1	CHS.30878.4	ANKRD30BL	251	258	0.9729	85.7	76.5	1.12	245	284	Testis	Testis	8.23	6.97	FALSE	FALSE
CHS.31062.17	CHS.31062.3	LYPD6B	183	207	0.8841	80	73.2	1.09	2700	2389	Skin	Bladder	24.49	10.43	FALSE	TRUE

CHS.31099.11	CHS.31099.8	CACNB4	473	520	0.9096	72	68.4	1.05	618	564	Brain	Brain	10.34	4.78	TRUE	TRUE
CHS.31231.29	CHS.31231.12	GCA	198	217	0.9124	84.7	78.9	1.07	2546	9696	Thyroid	Blood	0.74	94.61	TRUE	TRUE
CHS.31359.3	CHS.31359.4	MAP3K20	455	800	0.5687	72.4	66.4	1.09	8825	7530	Muscle	Bone_Marrow	109.05	12.77	TRUE	TRUE
CHS.31617.60	CHS.31617.28	ANKAR	782	1434	0.5453	89.8	84.9	1.06	196	68	Thyroid	Testis	3.30	7.05	FALSE	FALSE
CHS.31712.25	CHS.31712.3	CCDC150	748	1101	0.6794	79.2	74.4	1.06	252	2	Liver	Testis	2.13	5.40	FALSE	FALSE
CHS.31748.1	CHS.31748.11	FTCDNL1	77	138	0.558	85.1	77.8	1.09	4475	2121	Thyroid	Thyroid	2.28	0.40	FALSE	FALSE
CHS.31748.14	CHS.31748.11	FTCDNL1	147	138	1.0652	78	77.8	1.00	5383	2121	Thyroid	Thyroid	2.87	0.40	FALSE	FALSE
CHS.31771.4	CHS.31771.3	CLK1	333	484	0.688	94.9	77.7	1.22	9789	9789	Thyroid	Nerve	153.59	175.95	TRUE	TRUE
CHS.31782.12	CHS.31782.11	CFLAR	445	480	0.9271	84	80	1.05	6457	9784	Muscle	Thyroid	2.27	7.46	FALSE	FALSE
CHS.31784.6	CHS.31784.3	CASP10	479	522	0.9176	82.8	78.5	1.05	4300	6373	Lung	Spleen	4.39	14.05	FALSE	FALSE
CHS.31868.13	CHS.31868.3	MDH1B	472	518	0.9112	92	86.4	1.06	1109	3095	Fallopian_Tube	Testis	14.76	20.03	FALSE	FALSE
CHS.31922.1	CHS.31922.2	MYL1	150	194	0.7732	91.6	86.6	1.06	5928	4101	Muscle	Muscle	2842.57	1533.06	FALSE	FALSE
CHS.3203.2	CHS.3203.23	TPM3	158	285	0.5544	95.8	90.4	1.06	8414	7018	Brain	Muscle	2.17	662.22	TRUE	TRUE
CHS.32039.20	CHS.32039.2	PNKD	361	385	0.9377	92.8	82.2	1.13	8208	5965	Liver	Brain	86.24	26.20	TRUE	TRUE
CHS.3207.21	CHS.3207.7	C1orf43	189	253	0.747	79.8	71.6	1.11	9795	9795	Pituitary	Muscle	3.53	183.97	TRUE	TRUE
CHS.32305.17	CHS.32305.3	PSMD1	924	953	0.9696	76.4	72.2	1.06	113	9735	Muscle	Bone_Marrow	2.24	104.92	TRUE	TRUE
CHS.32305.18	CHS.32305.3	PSMD1	923	953	0.9685	76.5	72.2	1.06	35	9735	Muscle	Bone_Marrow	2.94	104.92	TRUE	TRUE
CHS.32305.21	CHS.32305.3	PSMD1	893	953	0.937	76.4	72.2	1.06	15	9735	Muscle	Bone_Marrow	1.95	104.92	TRUE	TRUE
CHS.32305.4	CHS.32305.3	PSMD1	922	953	0.9675	75.9	72.2	1.05	9326	9735	Bone_Marrow	Bone_Marrow	2.48	104.92	TRUE	TRUE
CHS.32310.53	CHS.32310.12	ARMC9	665	818	0.813	79.5	69.8	1.14	6699	2603	Esophagus	Uterus	14.63	12.58	FALSE	TRUE
CHS.32363.31	CHS.32363.1	EIF4E2	229	245	0.9347	82.9	77.9	1.06	49	9791	Testis	Bone_Marrow	19.83	29.55	TRUE	TRUE
CHS.32363.7	CHS.32363.1	EIF4E2	231	245	0.9429	82.8	77.9	1.06	68	9791	Testis	Bone_Marrow	5.43	29.55	TRUE	TRUE
CHS.32369.6	CHS.32369.8	NGEF	618	710	0.8704	75.1	70.2	1.07	4574	782	Brain	Small_Intestine	99.54	24.68	TRUE	TRUE
CHS.32507.34	CHS.32507.5	LRRFIP1	428	640	0.6687	78.7	70.9	1.11	2218	1544	Blood_Vessel	Heart	18.46	23.67	TRUE	TRUE
CHS.32507.4	CHS.32507.5	LRRFIP1	404	640	0.6312	78.3	70.9	1.10	3239	1544	Blood_Vessel	Heart	23.54	23.67	TRUE	TRUE
CHS.32507.50	CHS.32507.5	LRRFIP1	394	640	0.6156	77.6	70.9	1.09	3720	1544	Adipose_Tissue	Heart	22.00	23.67	TRUE	TRUE
CHS.32507.57	CHS.32507.5	LRRFIP1	418	640	0.6531	79	70.9	1.11	1642	1544	Cervix_Uteri	Heart	17.61	23.67	TRUE	TRUE
CHS.3251.12	CHS.3251.5	ADAM15	796	863	0.9224	76.9	73.2	1.05	8200	9638	Nerve	Skin	2.49	22.11	FALSE	FALSE
CHS.32599.12	CHS.32599.15	CROCC2	897	1653	0.5426	81	72.7	1.11	16	0	Lung	-	7.16	0.00	FALSE	FALSE
CHS.32606.30	CHS.32606.29	MTERF4	193	381	0.5066	85.3	76	1.12	9739	9625	Cervix_Uteri	Thyroid	4.29	20.63	TRUE	TRUE
CHS.32668.1	CHS.32668.6	C20orf96	328	363	0.9036	79.2	75.2	1.05	1036	5824	Uterus	Fallopian_Tube	5.81	43.51	FALSE	FALSE
CHS.32668.10	CHS.32668.6	C20orf96	310	363	0.854	81.7	75.2	1.09	7331	5824	Testis	Fallopian_Tube	18.26	43.51	FALSE	FALSE
CHS.32668.12	CHS.32668.6	C20orf96	261	363	0.719	85.3	75.2	1.13	5847	5824	Testis	Fallopian_Tube	28.17	43.51	FALSE	FALSE
CHS.32712.1	CHS.32712.4	SDCBP2	207	292	0.7089	93.9	80.9	1.16	9462	7125	Brain	Small_Intestine	6.84	111.40	TRUE	TRUE
CHS.32745.2	CHS.32745.5	STK35	401	534	0.7509	79.6	65.6	1.21	8228	4684	Testis	Colon	37.35	8.16	TRUE	TRUE
CHS.3277.2	CHS.3277.1	FDPS	353	419	0.8425	94.1	84.6	1.11	9794	8401	Bone_Marrow	Bone_Marrow	403.43	15.82	FALSE	FALSE
CHS.3277.4	CHS.3277.1	FDPS	248	419	0.5919	94.4	84.6	1.12	9398	8401	Bone_Marrow	Bone_Marrow	1.36	15.82	FALSE	FALSE
CHS.328.6	CHS.328.7	TMEM201	392	666	0.5886	75	58	1.29	8542	7623	Muscle	Bone_Marrow	25.17	10.87	TRUE	TRUE
CHS.32815.2	CHS.32815.34	PANK2	279	460	0.6065	91	69	1.32	5953	0	Testis	-	3.61	0.00	TRUE	TRUE
CHS.32850.12	CHS.32850.13	TMEM230	120	183	0.6557	76.2	64.4	1.18	9783	8149	Thyroid	Uterus	77.57	12.44	FALSE	FALSE

CHS.33206.23	CHS.33206.6	ABHD12	359	398	0.902	89.2	84.1	1.06	34	9763	Esophagus	Thyroid	8.24	119.02	TRUE	TRUE
CHS.33267.4	CHS.33267.3	HM13	377	426	0.885	82.1	74	1.11	9784	9670	Cervix_Uteri	Prostate	146.08	3.18	TRUE	TRUE
CHS.33289.49	CHS.33289.3	TTLL9	355	439	0.8087	83.2	78	1.07	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.33289.52	CHS.33289.3	TTLL9	379	439	0.8633	85.2	78	1.09	178	0	Lung	-	5.00	0.00	FALSE	FALSE
CHS.33289.55	CHS.33289.3	TTLL9	378	439	0.861	83.9	78	1.08	25	0	Nerve	-	3.19	0.00	FALSE	FALSE
CHS.33289.65	CHS.33289.3	TTLL9	436	439	0.9932	84.7	78	1.09	72	0	Cervix_Uteri	-	13.23	0.00	FALSE	FALSE
CHS.33289.75	CHS.33289.3	TTLL9	347	439	0.7904	86.1	78	1.10	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.33289.76	CHS.33289.3	TTLL9	341	439	0.7768	83.6	78	1.07	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.33289.8	CHS.33289.3	TTLL9	347	439	0.7904	86.5	78	1.11	754	0	Testis	-	4.98	0.00	FALSE	FALSE
CHS.33289.9	CHS.33289.3	TTLL9	230	439	0.5239	83.7	78	1.07	158	0	Nerve	-	1.65	0.00	FALSE	FALSE
CHS.3337.8	CHS.3337.1	TTC24	489	582	0.8402	82.5	73.2	1.13	3	2	Pituitary	Pituitary	3.11	0.95	FALSE	FALSE
CHS.33402.20	CHS.33402.3	SPAG4	312	437	0.714	85.4	69.3	1.23	3186	1904	Pancreas	Testis	24.06	27.16	TRUE	TRUE
CHS.33455.3	CHS.33455.1	NNAT	54	81	0.6667	82	67.2	1.22	7381	5622	Pituitary	Pituitary	1145.45	838.76	TRUE	TRUE
CHS.33600.29	CHS.33600.40	PABPC1L	330	619	0.5331	90.9	77.4	1.17	6156	0	Thyroid	-	37.05	0.00	FALSE	FALSE
CHS.33600.35	CHS.33600.40	PABPC1L	449	619	0.7254	82.3	77.4	1.06	5	0	Thyroid	-	7.76	0.00	FALSE	FALSE
CHS.33642.23	CHS.33642.20	WFDC3	137	231	0.5931	75.3	68.2	1.10	309	201	Skin	Testis	2.15	9.30	FALSE	FALSE
CHS.33745.15	CHS.33745.13	STAU1	496	577	0.8596	70.5	66.8	1.06	9774	9334	Thyroid	Muscle	40.05	14.89	FALSE	FALSE
CHS.34057.7	CHS.34057.11	CDH26	749	832	0.9002	77.4	73.3	1.06	0	97	-	Esophagus	0.00	2.67	FALSE	FALSE
CHS.341.13	CHS.341.11	LZIC	176	190	0.9263	85.4	80.7	1.06	97	9717	Pituitary	Uterus	1.63	10.26	FALSE	FALSE
CHS.3440.12	CHS.3440.11	FCRL6	397	434	0.9147	78.5	73.9	1.06	12	830	Blood	Prostate	1.45	1.90	FALSE	FALSE
CHS.34534.37	CHS.34534.2	GRIK1	865	949	0.9115	84.8	80.2	1.06	0	744	-	Adrenal_Gland	0.00	2.40	TRUE	TRUE
CHS.34534.38	CHS.34534.2	GRIK1	880	949	0.9273	84.7	80.2	1.06	0	744	-	Adrenal_Gland	0.00	2.40	TRUE	TRUE
CHS.34636.2	CHS.34636.3	IFNAR2	331	515	0.6427	74	58.1	1.27	9630	8959	Bone_Marrow	Spleen	20.99	4.33	FALSE	FALSE
CHS.34682.21	CHS.34682.9	RUNX1	250	480	0.5208	72.7	54.2	1.34	3787	2612	Lung	Muscle	3.10	9.48	FALSE	FALSE
CHS.34849.5	CHS.34849.4	PDE9A	533	593	0.8988	86.3	81.5	1.06	7750	7622	Colon	Small_Intestine	25.70	99.66	FALSE	FALSE
CHS.34903.4	CHS.34903.1	RRP1	415	461	0.9002	78.5	73.8	1.06	242	9766	Vagina	Bone_Marrow	2.78	39.78	FALSE	FALSE
CHS.3497.12	CHS.3497.5	CD244	329	365	0.9014	72.6	68.2	1.06	8	1307	Adipose_Tissue	Spleen	2.13	13.82	FALSE	FALSE
CHS.34978.7	CHS.34978.3	SLX9	214	230	0.9304	74.7	70.3	1.06	1277	9748	Bone_Marrow	Blood_Vessel	11.07	34.21	FALSE	FALSE
CHS.35043.2	CHS.35043.3	SPATC1L	186	340	0.5471	82.1	67.8	1.21	9371	2414	Bone_Marrow	Testis	130.50	14.57	FALSE	FALSE
CHS.3507.21	CHS.3507.6	NECTIN4	460	510	0.902	81.7	77.8	1.05	4	2286	Stomach	Skin	1.22	149.20	TRUE	TRUE
CHS.35196.13	CHS.35196.1	MRPL40	162	206	0.7864	85.1	79.2	1.07	9781	9778	Thyroid	Bone_Marrow	3.27	101.49	FALSE	TRUE
CHS.35226.22	CHS.35226.2	RANBP1	150	278	0.5396	85.6	72.2	1.19	9643	9349	Bone_Marrow	Bone_Marrow	81.34	15.47	FALSE	FALSE
CHS.35226.8	CHS.35226.2	RANBP1	201	278	0.723	82.9	72.2	1.15	9783	9349	Bone_Marrow	Bone_Marrow	435.96	15.47	FALSE	FALSE
CHS.35226.9	CHS.35226.2	RANBP1	200	278	0.7194	82.2	72.2	1.14	9782	9349	Bone_Marrow	Bone_Marrow	226.55	15.47	FALSE	FALSE
CHS.35270.29	CHS.35270.4	LRRC74B	228	392	0.5816	90.9	84.5	1.08	121	3	Cervix_Uteri	Lung	6.29	2.64	FALSE	FALSE
CHS.35332.5	CHS.35332.4	IGLL5	139	214	0.6495	83.7	69	1.21	4640	1414	Spleen	Kidney	241.90	3366.61	FALSE	FALSE
CHS.35375.4	CHS.35375.3	DERL3	205	235	0.8723	89.6	81.7	1.10	7266	5781	Spleen	Spleen	2.97	7.00	TRUE	TRUE
CHS.35534.14	CHS.35534.4	AP1B1	877	949	0.9241	83.9	77.9	1.08	0	7319	-	Brain	0.00	30.93	FALSE	FALSE
CHS.35651.1	CHS.35651.5	RFPL3	288	317	0.9085	88.7	83.8	1.06	88	1478	Testis	Testis	1.42	2.62	FALSE	FALSE
CHS.35755.12	CHS.35755.9	KCTD17	213	314	0.6783	78.3	69.3	1.13	8627	7028	Brain	Testis	20.28	20.63	TRUE	TRUE

CHS.35885.2	CHS.35885.4	SYNGR1	191	233	0.8197	87.6	77.8	1.13	9734	8627	Ovary	Brain	50.08	142.35	TRUE	TRUE
CHS.3592.11	CHS.3592.5	PBX1	347	430	0.807	75.2	67	1.12	9260	9205	Uterus	Uterus	90.77	64.88	TRUE	TRUE
CHS.35937.7	CHS.35937.12	L3MBTL2	637	705	0.9035	76.3	71.9	1.06	111	9685	Skin	Pituitary	3.29	24.38	TRUE	TRUE
CHS.36016.2	CHS.36016.3	PACSIN2	445	486	0.9156	83.1	79	1.05	7668	9776	Pituitary	Blood_Vessel	6.66	116.53	FALSE	FALSE
CHS.36181.11	CHS.36181.2	TAFA5	125	132	0.947	88.9	78.7	1.13	6876	5389	Nerve	Brain	89.56	46.07	TRUE	TRUE
CHS.36460.7	CHS.36460.2	IL5RA	335	420	0.7976	89.4	83.7	1.07	413	162	Prostate	Fallopian_Tube	1.41	15.13	FALSE	FALSE
CHS.36460.9	CHS.36460.2	IL5RA	333	420	0.7929	89.6	83.7	1.07	579	162	Vagina	Fallopian_Tube	0.82	15.13	FALSE	FALSE
CHS.36550.31	CHS.36550.63	TTLL3	574	815	0.7043	79.7	67.8	1.18	3	2	Blood	Blood_Vessel	1.33	1.21	FALSE	TRUE
CHS.36550.44	CHS.36550.63	TTLL3	434	815	0.5325	75.6	67.8	1.12	1150	2	Cervix_Uteri	Blood_Vessel	22.81	1.21	TRUE	TRUE
CHS.36550.59	CHS.36550.63	TTLL3	531	815	0.6515	77.3	67.8	1.14	8	2	Adipose_Tissue	Blood_Vessel	1.30	1.21	TRUE	TRUE
CHS.36550.60	CHS.36550.63	TTLL3	607	815	0.7448	80.4	67.8	1.19	112	2	Testis	Blood_Vessel	8.83	1.21	TRUE	TRUE
CHS.36550.65	CHS.36550.63	TTLL3	613	815	0.7521	72.8	67.8	1.07	4	2	Testis	Blood_Vessel	7.75	1.21	TRUE	TRUE
CHS.36550.66	CHS.36550.63	TTLL3	621	815	0.762	78	67.8	1.15	10	2	Skin	Blood_Vessel	1.54	1.21	TRUE	TRUE
CHS.36550.67	CHS.36550.63	TTLL3	564	815	0.692	77.6	67.8	1.14	17	2	Skin	Blood_Vessel	3.07	1.21	TRUE	TRUE
CHS.36550.68	CHS.36550.63	TTLL3	622	815	0.7632	75.8	67.8	1.12	7	2	Testis	Blood_Vessel	6.32	1.21	FALSE	TRUE
CHS.36550.69	CHS.36550.63	TTLL3	578	815	0.7092	75.9	67.8	1.12	5	2	Muscle	Blood_Vessel	1.32	1.21	TRUE	TRUE
CHS.36575.3	CHS.36575.1	IRAK2	576	625	0.9216	71.2	67.1	1.06	544	6824	Spleen	Bladder	1.74	16.76	TRUE	TRUE
CHS.36606.1	CHS.36606.2	SYN2	478	582	0.8213	74.6	68.6	1.09	3800	3765	Brain	Brain	84.86	16.17	TRUE	TRUE
CHS.36712.17	CHS.36712.1	OXNAD1	285	312	0.9135	91.4	85.4	1.07	5873	9502	Testis	Bone_Marrow	1.98	5.39	FALSE	FALSE
CHS.36794.17	CHS.36794.9	RARB	336	448	0.75	84.7	77.5	1.09	6797	0	Prostate	-	13.30	0.00	TRUE	TRUE
CHS.36794.18	CHS.36794.9	RARB	399	448	0.8906	83.1	77.5	1.07	988	0	Testis	-	1.43	0.00	TRUE	TRUE
CHS.36794.8	CHS.36794.9	RARB	405	448	0.904	82.4	77.5	1.06	0	0	-	-	0.00	0.00	FALSE	TRUE
CHS.36930.11	CHS.36930.1	LRRFIP2	400	721	0.5548	81	67.3	1.20	9683	990	Testis	Uterus	27.01	5.65	FALSE	FALSE
CHS.36930.19	CHS.36930.1	LRRFIP2	458	721	0.6352	81.1	67.3	1.21	1270	990	Skin	Uterus	33.70	5.65	FALSE	FALSE
CHS.36930.59	CHS.36930.1	LRRFIP2	424	721	0.5881	80.9	67.3	1.20	9756	990	Blood_Vessel	Uterus	37.44	5.65	FALSE	FALSE
CHS.36981.14	CHS.36981.1	SLC25A38	287	304	0.9441	82.2	77.8	1.06	67	9793	Prostate	Thyroid	2.97	69.21	FALSE	FALSE
CHS.36981.18	CHS.36981.1	SLC25A38	286	304	0.9408	82.4	77.8	1.06	1049	9793	Testis	Thyroid	1.36	69.21	FALSE	FALSE
CHS.37012.14	CHS.37012.4	CTNNB1	720	781	0.9219	84.1	79.9	1.05	281	9790	Heart	Cervix_Uteri	4.00	168.94	TRUE	TRUE
CHS.37181.1	CHS.37181.2	KIF9	725	790	0.9177	78.2	73.2	1.07	706	950	Testis	Testis	6.43	14.34	TRUE	TRUE
CHS.3729.1	CHS.3729.2	DNM3	555	863	0.6431	86.7	78.1	1.11	2678	1213	Cervix_Uteri	Brain	1.16	24.93	TRUE	TRUE
CHS.37312.3	CHS.37312.7	RASSF1	270	340	0.7941	77.4	73.3	1.06	9790	8922	Testis	Spleen	54.30	31.34	TRUE	TRUE
CHS.37350.1	CHS.37350.2	IQCF2	86	164	0.5244	95.3	86.1	1.11	225	225	Testis	Testis	105.51	33.68	FALSE	FALSE
CHS.37354.3	CHS.37354.1	RRP9	431	475	0.9074	88.4	84	1.05	986	9651	Uterus	Bone_Marrow	5.45	33.62	TRUE	TRUE
CHS.37638.3	CHS.37638.4	PROK2	108	129	0.8372	82.1	73.7	1.11	2367	1632	Blood	Blood	82.69	55.62	TRUE	TRUE
CHS.37756.1	CHS.37756.2	EPHA3	539	983	0.5483	87.6	79.6	1.10	3522	3268	Adrenal_Gland	Prostate	0.40	20.47	TRUE	TRUE
CHS.37972.1	CHS.37972.3	CD96	402	569	0.7065	74.5	64.5	1.16	2545	1957	Thyroid	Bladder	1.33	3.77	TRUE	TRUE
CHS.37976.1	CHS.37976.2	PLCXD2	341	305	1.118	86.1	78.5	1.10	1017	145	Brain	Spleen	3.74	0.62	FALSE	FALSE
CHS.37976.5	CHS.37976.2	PLCXD2	289	305	0.9475	85.5	78.5	1.09	480	145	Pituitary	Spleen	1.23	0.62	FALSE	FALSE
CHS.37982.3	CHS.37982.13	TMPRSS7	717	843	0.8505	85.5	81.4	1.05	26	0	Testis	-	1.11	0.00	FALSE	FALSE
CHS.37982.4	CHS.37982.13	TMPRSS7	706	843	0.8375	86.1	81.4	1.06	3	0	Testis	-	1.59	0.00	FALSE	FALSE

CHS.38022.33	CHS.38022.6	CFAP44	982	1854	0.5297	84.2	74.9	1.12	3147	2983	Cervix_Uteri	Fallopian_Tube	0.36	8.75	FALSE	FALSE
CHS.38038.3	CHS.38038.4	DRD3	367	400	0.9175	77.4	71.8	1.08	1	246	Testis	Brain	0.06	3.68	TRUE	TRUE
CHS.38173.18	CHS.38173.10	PARP15	375	678	0.5531	85	80.3	1.06	4201	938	Spleen	Spleen	2.18	5.27	FALSE	FALSE
CHS.38190.25	CHS.38190.11	MYLK	991	1914	0.5178	70.4	64	1.10	2205	365	Colon	Esophagus	35.82	163.20	FALSE	FALSE
CHS.38190.43	CHS.38190.11	MYLK	992	1914	0.5183	70.9	64	1.11	6175	365	Colon	Esophagus	486.47	163.20	TRUE	FALSE
CHS.38194.7	CHS.38194.8	ROPN1	120	212	0.566	91	85.2	1.07	1063	228	Breast	Testis	9.73	144.12	TRUE	TRUE
CHS.38356.2	CHS.38356.1	ALG1L2	120	215	0.5581	79.7	72.2	1.10	76	10	Breast	Spleen	2.62	1.58	FALSE	FALSE
CHS.38435.1	CHS.38435.18	CEP63	495	703	0.7041	87	78	1.12	9545	7512	Bladder	Testis	6.77	28.09	FALSE	FALSE
CHS.38435.15	CHS.38435.18	CEP63	541	703	0.7696	86.7	78	1.11	9391	7512	Ovary	Testis	3.02	28.09	FALSE	FALSE
CHS.38435.36	CHS.38435.18	CEP63	475	703	0.6757	85.3	78	1.09	9297	7512	Muscle	Testis	5.91	28.09	FALSE	FALSE
CHS.38483.10	CHS.38483.9	FAIM	179	201	0.8905	92.8	87.6	1.06	9457	1449	Fallopian_Tube	Bladder	20.57	3.50	FALSE	FALSE
CHS.38508.1	CHS.38508.5	NMNAT3	215	346	0.6214	85.8	77.5	1.11	4420	2960	Skin	Bladder	1.16	1.95	FALSE	FALSE
CHS.38508.14	CHS.38508.5	NMNAT3	252	346	0.7283	88.7	77.5	1.14	8080	2960	Skin	Bladder	7.36	1.95	FALSE	FALSE
CHS.38799.20	CHS.38799.2	MLF1	194	283	0.6855	76.5	64.4	1.19	4613	902	Testis	Heart	7.95	28.87	TRUE	TRUE
CHS.38799.23	CHS.38799.2	MLF1	223	283	0.788	72	64.4	1.12	1470	902	Testis	Heart	4.24	28.87	TRUE	TRUE
CHS.38811.27	CHS.38811.1	SCHIP1	244	487	0.501	74.1	57.1	1.30	7496	1248	Brain	Brain	58.28	10.98	TRUE	TRUE
CHS.38811.32	CHS.38811.1	SCHIP1	260	487	0.5339	72.9	57.1	1.28	2949	1248	Muscle	Brain	57.52	10.98	TRUE	TRUE
CHS.38981.2	CHS.38981.4	ECT2	854	914	0.9344	73.2	66.1	1.11	9	2428	Blood	Skin	1.48	4.63	FALSE	TRUE
CHS.38981.6	CHS.38981.4	ECT2	929	914	1.0164	70.5	66.1	1.07	1	2428	Brain	Skin	1.55	4.63	TRUE	TRUE
CHS.39044.8	CHS.39044.7	KCNMB3	257	275	0.9345	75.7	71	1.07	7572	8844	Prostate	Bone_Marrow	1.58	3.17	TRUE	TRUE
CHS.39070.1	CHS.39070.2	TTC14	439	770	0.5701	86.4	64.3	1.34	9614	9447	Thyroid	Fallopian_Tube	60.46	6.78	TRUE	TRUE
CHS.39075.7	CHS.39075.6	FXR1	539	621	0.868	74.3	68.2	1.09	9788	9779	Adrenal_Gland	Testis	5.14	43.62	TRUE	TRUE
CHS.39199.2	CHS.39199.1	KNG1	427	644	0.663	78.9	62.3	1.27	351	351	Liver	Liver	394.01	196.47	TRUE	TRUE
CHS.39269.13	CHS.39269.10	TP63	393	680	0.5779	74.2	60.7	1.22	919	837	Ovary	Muscle	3.40	4.08	TRUE	TRUE
CHS.39269.35	CHS.39269.10	TP63	416	680	0.6118	71.1	60.7	1.17	919	837	Ovary	Muscle	1.65	4.08	TRUE	TRUE
CHS.39295.3	CHS.39295.2	CCDC50	306	482	0.6349	73.7	61.3	1.20	9762	6287	Adipose_Tissue	Thyroid	37.65	13.52	TRUE	TRUE
CHS.39787.5	CHS.39787.7	CRMP1	572	686	0.8338	89.6	81.1	1.10	4416	2120	Pituitary	Brain	98.04	19.13	FALSE	FALSE
CHS.39815.18	CHS.39815.6	TBC1D14	413	693	0.596	87.2	64	1.36	9477	9168	Skin	Spleen	14.91	36.59	TRUE	TRUE
CHS.3983.1	CHS.3983.3	PDC	194	246	0.7886	81.3	77.2	1.05	17	11	Breast	Skin	0.29	0.37	FALSE	FALSE
CHS.39970.7	CHS.39970.4	BST1	289	318	0.9088	91.9	87.2	1.05	4840	8033	Blood	Blood	9.75	44.93	FALSE	FALSE
CHS.39989.13	CHS.39989.3	LDB2	348	373	0.933	74.5	70.5	1.06	2	8710	Adipose_Tissue	Uterus	1.04	30.24	TRUE	TRUE
CHS.39989.32	CHS.39989.3	LDB2	349	373	0.9357	74.7	70.5	1.06	3	8710	Lung	Uterus	2.65	30.24	TRUE	TRUE
CHS.40025.11	CHS.40025.15	KCNIP4	229	250	0.916	73	68.8	1.06	1252	1465	Thyroid	Skin	4.51	39.80	TRUE	TRUE
CHS.40025.13	CHS.40025.15	KCNIP4	225	250	0.9	74.1	68.8	1.08	608	1465	Colon	Skin	3.35	39.80	TRUE	TRUE
CHS.40224.13	CHS.40224.17	KLHL5	568	709	0.8011	86	79.9	1.08	9279	8378	Bladder	Brain	16.93	2.88	TRUE	TRUE
CHS.40331.7	CHS.40331.2	CORIN	975	1042	0.9357	74.4	70.1	1.06	9	1374	Spleen	Heart	4.18	13.46	TRUE	TRUE
CHS.40462.10	CHS.40462.4	HOPX	73	91	0.8022	79.4	71.3	1.11	8715	4122	Prostate	Cervix_Uteri	18.28	91.69	FALSE	FALSE
CHS.40468.11	CHS.40468.9	SPINK2	84	134	0.6269	83.9	69.7	1.20	2944	1970	Testis	Testis	207.32	108.04	TRUE	TRUE
CHS.40558.11	CHS.40558.4	UGT2A1	483	527	0.9165	92.8	87.6	1.06	192	13	Pituitary	Pituitary	1.09	1.72	TRUE	FALSE
CHS.40558.3	CHS.40558.4	UGT2A1	527	527	1	93.8	87.6	1.07	71	13	Pituitary	Pituitary	4.87	1.72	FALSE	FALSE

CHS.40589.29	CHS.40589.26	RUFY3	567	620	0.9145	80.3	76.1	1.06	2895	9355	Thyroid	Pituitary	10.09	26.41	TRUE	TRUE
CHS.40700.6	CHS.40700.9	NUP54	459	507	0.9053	79.8	75.5	1.06	4235	9598	Bone_Marrow	Bone_Marrow	3.66	52.65	TRUE	TRUE
CHS.4081.21	CHS.4081.12	DENND1B	426	775	0.5497	89	64.2	1.39	1894	1788	Brain	Small_Intestine	0.68	4.93	TRUE	TRUE
CHS.40815.11	CHS.40815.29	COQ2	328	371	0.8841	82.3	77.7	1.06	1301	0	Cervix_Uteri	-	2.26	0.00	FALSE	TRUE
CHS.40815.15	CHS.40815.29	COQ2	421	371	1.1348	77.7	77.7	1.00	9607	0	Adrenal_Gland	-	18.56	0.00	TRUE	TRUE
CHS.40815.23	CHS.40815.29	COQ2	298	371	0.8032	91.7	77.7	1.18	54	0	Breast	-	2.05	0.00	TRUE	TRUE
CHS.40815.27	CHS.40815.29	COQ2	260	371	0.7008	88	77.7	1.13	136	0	Prostate	-	1.47	0.00	TRUE	TRUE
CHS.41089.8	CHS.41089.2	MCUB	307	336	0.9137	75.2	70.7	1.06	378	8921	Nerve	Adipose_Tissue	2.69	16.94	TRUE	TRUE
CHS.4110.13	CHS.4110.1	NR5A2	469	541	0.8669	77.7	71.5	1.09	1585	386	Pancreas	Pancreas	5.43	5.68	TRUE	TRUE
CHS.4110.2	CHS.4110.1	NR5A2	495	541	0.915	75.3	71.5	1.05	597	386	Pancreas	Pancreas	7.95	5.68	TRUE	TRUE
CHS.41135.27	CHS.41135.19	CAMK2D	478	533	0.8968	86.6	81.5	1.06	9277	1751	Adipose_Tissue	Pituitary	19.99	8.06	FALSE	FALSE
CHS.41135.30	CHS.41135.19	CAMK2D	492	533	0.9231	85.7	81.5	1.05	1287	1751	Brain	Pituitary	25.41	8.06	FALSE	FALSE
CHS.41135.37	CHS.41135.19	CAMK2D	489	533	0.9174	85.6	81.5	1.05	44	1751	Blood_Vessel	Pituitary	4.06	8.06	FALSE	FALSE
CHS.41135.54	CHS.41135.19	CAMK2D	478	533	0.8968	85.8	81.5	1.05	9266	1751	Lung	Pituitary	2.16	8.06	FALSE	FALSE
CHS.41245.1	CHS.41245.2	TRPC3	848	921	0.9207	78.5	74.1	1.06	1242	762	Pituitary	Stomach	6.29	5.76	TRUE	TRUE
CHS.41329.16	CHS.41329.2	JADE1	509	842	0.6045	74.5	60.3	1.24	9095	9011	Thyroid	Thyroid	9.86	17.12	TRUE	TRUE
CHS.41632.2	CHS.41632.3	RBM46	485	533	0.9099	78.3	72.8	1.08	217	207	Testis	Testis	11.33	2.31	TRUE	TRUE
CHS.41729.5	CHS.41729.1	TRIM61	471	209	2.2536	86.1	86	1.00	4760	173	Testis	Ovary	7.61	1.86	FALSE	FALSE
CHS.42006.5	CHS.42006.6	PDLIM3	276	364	0.7582	71.6	65.3	1.10	6461	5764	Muscle	Muscle	80.33	1016.37	TRUE	TRUE
CHS.42393.10	CHS.42393.3	FBXL7	444	491	0.9043	85.6	81.5	1.05	545	8183	Nerve	Uterus	3.79	37.15	TRUE	TRUE
CHS.42523.2	CHS.42523.1	CDH6	663	790	0.8392	84	78.6	1.07	4536	4444	Kidney	Kidney	5.94	11.59	TRUE	TRUE
CHS.4253.12	CHS.4253.7	GOLT1A	130	132	0.9848	80.3	76.4	1.05	141	4034	Liver	Liver	42.44	96.64	TRUE	TRUE
CHS.4253.15	CHS.4253.7	GOLT1A	123	132	0.9318	80.9	76.4	1.06	1004	4034	Liver	Liver	24.47	96.64	TRUE	TRUE
CHS.42565.1	CHS.42565.2	C1QTNF3	246	319	0.7712	79	68.5	1.15	6363	5846	Esophagus	Esophagus	5.84	5.93	TRUE	TRUE
CHS.42606.4	CHS.42606.7	LMBRD2	637	695	0.9165	79.8	75.9	1.05	322	6372	Nerve	Brain	2.50	8.54	FALSE	FALSE
CHS.42854.4	CHS.42854.5	IL31RA	582	764	0.7618	76	72	1.06	281	258	Blood_Vessel	Blood_Vessel	0.43	1.37	FALSE	FALSE
CHS.42930.11	CHS.42930.15	PDE4D	673	809	0.8319	72.8	64.1	1.14	2282	1482	Muscle	Brain	14.79	1.78	TRUE	TRUE
CHS.42930.4	CHS.42930.15	PDE4D	518	809	0.6403	78.7	64.1	1.23	1789	1482	Brain	Brain	1.35	1.78	TRUE	TRUE
CHS.42930.5	CHS.42930.15	PDE4D	507	809	0.6267	80	64.1	1.25	4905	1482	Skin	Brain	3.63	1.78	TRUE	TRUE
CHS.42930.6	CHS.42930.15	PDE4D	585	809	0.7231	75.8	64.1	1.18	6951	1482	Muscle	Brain	4.11	1.78	TRUE	TRUE
CHS.42930.7	CHS.42930.15	PDE4D	679	809	0.8393	72.4	64.1	1.13	4109	1482	Blood_Vessel	Brain	6.50	1.78	TRUE	TRUE
CHS.42930.8	CHS.42930.15	PDE4D	687	809	0.8492	71.9	64.1	1.12	2360	1482	Muscle	Brain	28.82	1.78	TRUE	TRUE
CHS.43067.20	CHS.43067.19	CCDC125	386	511	0.7554	73.5	66.5	1.11	4987	2285	Bone_Marrow	Cervix_Uteri	2.95	3.33	FALSE	FALSE
CHS.43067.37	CHS.43067.19	CCDC125	467	511	0.9139	70.6	66.5	1.06	198	2285	Nerve	Cervix_Uteri	3.96	3.33	FALSE	FALSE
CHS.43079.1	CHS.43079.2	SERF1B	62	110	0.5636	79.7	61	1.31	9776	9764	Testis	Brain	102.23	11.13	FALSE	FALSE
CHS.43155.2	CHS.43155.1	BTF3	162	206	0.7864	76.6	69.7	1.10	9794	9793	Ovary	Ovary	893.52	196.66	FALSE	FALSE
CHS.43203.1	CHS.43203.2	F2RL2	352	374	0.9412	78.8	74.1	1.06	556	3661	Uterus	Breast	1.27	3.79	TRUE	TRUE
CHS.4321.2	CHS.4321.10	SRGAP2	790	1071	0.7376	84.2	71.7	1.17	1904	8	Nerve	Testis	7.13	2.68	TRUE	TRUE
CHS.4321.3	CHS.4321.10	SRGAP2	841	1071	0.7852	81.3	71.7	1.13	11	8	Testis	Testis	1.32	2.68	TRUE	TRUE
CHS.4321.4	CHS.4321.10	SRGAP2	789	1071	0.7367	84.4	71.7	1.18	9466	8	Nerve	Testis	35.05	2.68	FALSE	TRUE

CHS.4321.45	CHS.4321.10	SRGAP2	836	1071	0.7806	81.3	71.7	1.13	3676	8	Nerve	Testis	5.72	2.68	TRUE	TRUE
CHS.4321.46	CHS.4321.10	SRGAP2	832	1071	0.7768	81.5	71.7	1.14	157	8	Adipose_Tissue	Testis	9.98	2.68	FALSE	TRUE
CHS.4321.48	CHS.4321.10	SRGAP2	833	1071	0.7778	81.1	71.7	1.13	162	8	Brain	Testis	32.59	2.68	TRUE	TRUE
CHS.43219.5	CHS.43219.15	PDE8B	805	885	0.9096	78.4	73.7	1.06	207	5049	Adipose_Tissue	Thyroid	9.27	56.80	TRUE	TRUE
CHS.43252.82	CHS.43252.22	TENT2	450	484	0.9298	78.9	74.4	1.06	1	9691	Blood	Bone_Marrow	0.18	18.95	TRUE	TRUE
CHS.43258.4	CHS.43258.5	MTX3	248	312	0.7949	89.4	84.3	1.06	5503	3904	Thyroid	Brain	4.79	4.77	TRUE	TRUE
CHS.4328.9	CHS.4328.1	RASSF5	265	418	0.634	80.1	68.8	1.16	8619	7130	Blood	Salivary_Gland	58.20	28.78	TRUE	TRUE
CHS.43437.6	CHS.43437.18	MCTP1	692	999	0.6927	76.7	64	1.20	4098	2033	Adipose_Tissue	Brain	2.39	3.12	FALSE	FALSE
CHS.43437.7	CHS.43437.18	MCTP1	778	999	0.7788	75.1	64	1.17	5743	2033	Adipose_Tissue	Brain	12.29	3.12	FALSE	FALSE
CHS.43443.29	CHS.43443.11	FAM81B	288	452	0.6372	85.2	74.9	1.14	362	193	Fallopian_Tube	Testis	11.10	88.93	TRUE	TRUE
CHS.43449.1	CHS.43449.3	RFESD	157	210	0.7476	89.1	77.2	1.15	7007	5514	Bone_Marrow	Bone_Marrow	4.83	9.13	FALSE	FALSE
CHS.43527.11	CHS.43527.70	PAM	866	973	0.89	74.1	68.4	1.08	9565	6737	Adrenal_Gland	Blood_Vessel	30.77	84.97	FALSE	FALSE
CHS.43527.117	CHS.43527.70	PAM	884	973	0.9085	74.4	68.4	1.09	176	6737	Blood_Vessel	Blood_Vessel	5.99	84.97	FALSE	FALSE
CHS.43527.51	CHS.43527.70	PAM	883	973	0.9075	74.7	68.4	1.09	150	6737	Blood_Vessel	Blood_Vessel	5.55	84.97	FALSE	FALSE
CHS.43625.4	CHS.43625.3	DCP2	385	420	0.9167	70.6	67	1.05	2593	7297	Bone_Marrow	Bone_Marrow	3.05	33.37	TRUE	TRUE
CHS.43700.3	CHS.43700.17	TNFAIP8	188	198	0.9495	91.9	87.5	1.05	8633	7844	Spleen	Spleen	35.22	6.10	TRUE	TRUE
CHS.43754.10	CHS.43754.8	CSNK1G3	416	456	0.9123	78	73.2	1.07	2570	905	Bone_Marrow	Nerve	4.53	3.08	TRUE	TRUE
CHS.43754.13	CHS.43754.8	CSNK1G3	415	456	0.9101	77.9	73.2	1.06	2247	905	Bone_Marrow	Nerve	5.10	3.08	TRUE	TRUE
CHS.43754.16	CHS.43754.8	CSNK1G3	311	456	0.682	77.8	73.2	1.06	3268	905	Bladder	Nerve	0.57	3.08	TRUE	TRUE
CHS.43812.13	CHS.43812.11	CCDC192	263	273	0.9634	82.2	77.8	1.06	17	350	Skin	Testis	2.27	19.24	FALSE	FALSE
CHS.43902.30	CHS.43902.4	SEPTIN8	440	483	0.911	79.5	75.5	1.05	7090	9104	Brain	Brain	15.66	105.31	TRUE	TRUE
CHS.43902.37	CHS.43902.4	SEPTIN8	442	483	0.9151	79.6	75.5	1.05	7164	9104	Brain	Brain	27.89	105.31	TRUE	TRUE
CHS.43902.38	CHS.43902.4	SEPTIN8	429	483	0.8882	81	75.5	1.07	9386	9104	Bone_Marrow	Brain	56.20	105.31	TRUE	TRUE
CHS.44126.2	CHS.44126.3	SRA1	162	224	0.7232	76.3	71.6	1.07	9794	9783	Pituitary	Bone_Marrow	6.17	25.29	TRUE	TRUE
CHS.44143.3	CHS.44143.1	PCDHAC2	884	1007	0.8779	76.8	71.9	1.07	8768	4705	Thyroid	Brain	0.11	3.93	TRUE	TRUE
CHS.44144.1	CHS.44144.2	PCDHA1	807	950	0.8495	80.5	72.8	1.11	7460	898	Testis	Skin	1.59	3.60	TRUE	TRUE
CHS.44145.1	CHS.44145.2	PCDHA9	842	950	0.8863	77.9	72.1	1.08	8244	1391	Bladder	Testis	5.68	1.69	TRUE	TRUE
CHS.44147.1	CHS.44147.2	PCDHA8	814	950	0.8568	79.8	72.7	1.10	7364	934	Liver	Spleen	3.53	1.65	TRUE	TRUE
CHS.44150.2	CHS.44150.3	PCDHA13	807	950	0.8495	80.1	72.5	1.10	7320	2331	Testis	Brain	0.19	1.15	TRUE	TRUE
CHS.44152.2	CHS.44152.3	PCDHA7	789	937	0.842	81.4	73.5	1.11	8841	957	Testis	Brain	3.08	1.51	TRUE	TRUE
CHS.44153.1	CHS.44153.2	PCDHA12	792	941	0.8417	81.4	73.3	1.11	7775	2640	Testis	Nerve	0.29	2.24	FALSE	FALSE
CHS.44155.1	CHS.44155.3	PCDHA2	824	948	0.8692	79.2	72.6	1.09	8440	1312	Pituitary	Pituitary	1.18	1.56	TRUE	TRUE
CHS.44189.1	CHS.44189.2	PCDHGB5	818	923	0.8862	80.2	74.6	1.08	9729	9112	Skin	Skin	0.37	12.76	TRUE	TRUE
CHS.44192.1	CHS.44192.2	PCDHGB7	808	929	0.8698	80.7	74.4	1.08	9841	9153	Bladder	Lung	45.42	18.10	TRUE	TRUE
CHS.44194.1	CHS.44194.2	PCDHGB2	811	931	0.8711	80.8	74.3	1.09	9654	8274	Bone_Marrow	Brain	0.43	9.25	TRUE	TRUE
CHS.44195.1	CHS.44195.2	PCDHGA10	850	936	0.9081	78.1	73.8	1.06	1	8514	Liver	Brain	2.42	6.31	TRUE	TRUE
CHS.44197.1	CHS.44197.3	PCDHGA3	829	932	0.8895	79.3	74.1	1.07	9573	8616	Bone_Marrow	Brain	0.26	10.68	TRUE	TRUE
CHS.44199.1	CHS.44199.2	PCDHGB3	814	929	0.8762	80.5	74.3	1.08	9535	7969	Ovary	Nerve	0.27	2.46	TRUE	FALSE
CHS.44200.1	CHS.44200.2	PCDHGA4	851	962	0.8846	79	73.5	1.07	9585	8650	Bone_Marrow	Brain	0.29	6.49	TRUE	TRUE
CHS.44201.1	CHS.44201.2	PCDHGA11	837	935	0.8952	79.5	74	1.07	9635	8905	Skin	Skin	0.55	4.56	TRUE	TRUE

CHS.44202.1	CHS.44202.2	PCDHGA 9	828	932	0.8884	80	74.5	1.07	9698	8681	Testis	Brain	0.30	10.20	TRUE	TRUE
CHS.44203.1	CHS.44203.2	PCDHGB 1	810	927	0.8738	80.4	74.3	1.08	9472	5549	Bladder	Brain	0.40	6.69	TRUE	TRUE
CHS.44204.1	CHS.44204.2	PCDHGA 7	817	932	0.8766	80.1	74.2	1.08	9668	8612	Ovary	Uterus	0.39	3.16	TRUE	TRUE
CHS.44209.1	CHS.44209.2	PCDHGA 12	820	932	0.8798	80.6	74.4	1.08	9669	8904	Liver	Skin	0.11	9.82	TRUE	TRUE
CHS.44311.6	CHS.44311.10	DPYSL3	570	684	0.8333	89.8	80.9	1.11	8849	8591	Blood_V essel	Uterus	147.55	83.87	TRUE	TRUE
CHS.44334.10	CHS.44334.14	HTR4	360	388	0.9278	82.6	76.8	1.08	13	602	Adrenal _Gland	Brain	0.35	3.07	FALSE	FALSE
CHS.44334.15	CHS.44334.14	HTR4	371	388	0.9562	82.5	76.8	1.07	66	602	Colon	Brain	0.60	3.07	FALSE	FALSE
CHS.44451.5	CHS.44451.3	GLRA1	307	449	0.6837	87.6	83.4	1.05	74	73	Brain	Brain	2.20	1.03	FALSE	TRUE
CHS.4463.6	CHS.4463.3	NEK2	384	445	0.8629	82	77.7	1.06	2295	2235	Bone_M arrow	Bone_M arrow	7.53	57.37	TRUE	TRUE
CHS.44692.6	CHS.44692.5	NPM1	265	294	0.9014	75.1	69.9	1.07	9786	9794	Cervix_ Uteri	Bone_M arrow	27.45	1760.73	FALSE	FALSE
CHS.44706.11	CHS.44706.8	FBXW11	508	563	0.9023	86.9	82	1.06	9200	110	Uterus	Pituitary	11.82	2.60	FALSE	FALSE
CHS.44706.4	CHS.44706.8	FBXW11	510	563	0.9059	86.9	82	1.06	793	110	Nerve	Pituitary	3.59	2.60	FALSE	FALSE
CHS.44706.5	CHS.44706.8	FBXW11	510	563	0.9059	87.4	82	1.07	8458	110	Brain	Pituitary	9.03	2.60	FALSE	FALSE
CHS.44915.7	CHS.44915.3	CLK4	445	481	0.9252	82.1	77.7	1.06	4713	9743	Nerve	Uterus	8.81	31.34	TRUE	TRUE
CHS.44938.11	CHS.44938.7	RUFY1	600	708	0.8475	80.3	75	1.07	9781	9640	Pituitary	Spleen	20.41	40.97	TRUE	TRUE
CHS.44962.6	CHS.44962.5	MAPK9	382	424	0.9009	85.9	80.9	1.06	9505	9577	Brain	Bladder	11.45	9.85	TRUE	TRUE
CHS.44962.8	CHS.44962.5	MAPK9	382	424	0.9009	85.7	80.9	1.06	3712	9577	Brain	Bladder	9.89	9.85	TRUE	TRUE
CHS.44996.1	CHS.44996.6	TRIM7	329	511	0.6438	89.9	84.9	1.06	7891	7274	Testis	Muscle	1.64	26.83	TRUE	TRUE
CHS.4517.7	CHS.4517.5	VASH2	290	355	0.8169	83.3	76.8	1.08	424	325	Blood	Blood	6.51	6.71	TRUE	TRUE
CHS.4517.8	CHS.4517.5	VASH2	251	355	0.707	83.2	76.8	1.08	467	325	Cervix_ Uteri	Blood	2.41	6.71	TRUE	TRUE
CHS.45442.6	CHS.45442.11	SIRT5	202	310	0.6516	93.4	88.6	1.05	8558	7654	Prostate	Liver	2.18	4.18	FALSE	FALSE
CHS.45850.5	CHS.45850.7	HLA-F	346	442	0.7828	89.4	81.5	1.10	9683	8652	Spleen	Spleen	102.84	9.84	FALSE	FALSE
CHS.45850.6	CHS.45850.7	HLA-F	254	442	0.5747	86	81.5	1.06	9009	8652	Spleen	Spleen	2.26	9.84	FALSE	FALSE
CHS.45902.1	CHS.45902.7	ATAT1	323	409	0.7897	72.4	61.7	1.17	9388	7293	Pituitary	Testis	16.88	5.24	TRUE	TRUE
CHS.45902.3	CHS.45902.7	ATAT1	300	409	0.7335	75	61.7	1.22	8290	7293	Pituitary	Testis	7.18	5.24	TRUE	TRUE
CHS.45902.5	CHS.45902.7	ATAT1	248	409	0.6064	81.7	61.7	1.32	7945	7293	Testis	Testis	8.68	5.24	TRUE	TRUE
CHS.45940.16	CHS.45940.2	CCHCR1	782	871	0.8978	79.9	74.8	1.07	9494	8673	Bone_M arrow	Testis	16.35	100.36	FALSE	TRUE
CHS.45972.2	CHS.45972.1	NCR3	177	201	0.8806	86.6	80.7	1.07	2100	1798	Spleen	Spleen	10.68	8.55	FALSE	FALSE
CHS.46090.4	CHS.46090.3	VPS52	656	723	0.9073	89.9	85.4	1.05	5169	9756	Testis	Bone_M arrow	3.81	57.06	TRUE	TRUE
CHS.46093.5	CHS.46093.6	WDR46	556	610	0.9115	85.8	80.8	1.06	0	9769	-	Bone_M arrow	0.00	52.01	TRUE	TRUE
CHS.46101.13	CHS.46101.12	PHF1	419	567	0.739	77.6	68.4	1.13	9781	9780	Testis	Fallopia n_Tube	224.50	44.87	TRUE	TRUE
CHS.46102.6	CHS.46102.3	CUTA	156	179	0.8715	82	77	1.06	9793	9788	Pituitary	Pituitary	100.60	182.48	FALSE	FALSE
CHS.46177.5	CHS.46177.3	PPARD	402	441	0.9116	85	80.2	1.06	4441	9743	Ovary	Thyroid	0.91	48.19	TRUE	TRUE
CHS.46183.10	CHS.46183.2	TULP1	349	542	0.6439	77.9	62.1	1.25	22	7	Pituitary	Uterus	2.58	1.69	TRUE	FALSE
CHS.46217.7	CHS.46217.2	PXT1	83	134	0.6194	81.1	66.2	1.23	440	15	Testis	Testis	3.83	4.30	TRUE	FALSE
CHS.46252.11	CHS.46252.1	PIM1	296	313	0.9457	81.3	74.4	1.09	1763	0	Fallopia n_Tube	-	18.65	0.00	TRUE	TRUE
CHS.46252.16	CHS.46252.1	PIM1	404	313	1.2907	74.4	74.4	1.00	9756	0	Bone_M arrow	-	420.49	0.00	TRUE	TRUE
CHS.46300.2	CHS.46300.7	MOCS1	385	636	0.6053	86.2	78.5	1.10	9171	8138	Testis	Breast	1.22	10.57	TRUE	TRUE
CHS.46300.4	CHS.46300.7	MOCS1	385	636	0.6053	85.7	78.5	1.09	9124	8138	Adipose _Tissue	Breast	41.74	10.57	TRUE	TRUE
CHS.46319.7	CHS.46319.8	TREML1	199	311	0.6399	78	65.7	1.19	1255	998	Blood	Muscle	18.33	15.22	TRUE	TRUE

CHS.46418.4	CHS.46418.5	YIPF3	315	350	0.9	74.6	69.4	1.07	5866	9792	Prostate	Adrenal_Gland	17.44	154.52	TRUE	TRUE
CHS.46539.1	CHS.46539.2	DEFB110	62	67	0.9254	81.9	77.7	1.05	3	2	Skin	Thyroid	0.64	0.16	TRUE	TRUE
CHS.46544.2	CHS.46544.1	TFAP2D	305	452	0.6748	74.5	60.9	1.22	20	4	Brain	Brain	1.65	1.12	TRUE	TRUE
CHS.4662.24	CHS.4662.2	SUSD4	290	490	0.5918	75.5	69.6	1.08	5346	4711	Vagina	Ovary	8.13	33.71	FALSE	TRUE
CHS.46865.3	CHS.46865.1	RIPPLY2	70	128	0.5469	72.2	59.7	1.21	4348	1763	Brain	Brain	14.57	1.58	TRUE	FALSE
CHS.4689.2	CHS.4689.40	WDR26	661	761	0.8686	77.3	70.8	1.09	9784	0	Bone_Marrow	-	36.19	0.00	TRUE	TRUE
CHS.4689.28	CHS.4689.40	WDR26	513	761	0.6741	85.4	70.8	1.21	26	0	Adipose_Tissue	-	3.15	0.00	TRUE	TRUE
CHS.4689.8	CHS.4689.40	WDR26	660	761	0.8673	76.5	70.8	1.08	339	0	Nerve	-	4.40	0.00	TRUE	TRUE
CHS.46890.2	CHS.46890.20	SYNCRIP	561	623	0.9005	73.9	70.1	1.05	4116	9782	Bone_Marrow	Skin	5.78	21.67	FALSE	FALSE
CHS.46890.6	CHS.46890.20	SYNCRIP	562	623	0.9021	74	70.1	1.06	9782	9782	Bone_Marrow	Skin	52.87	21.67	FALSE	FALSE
CHS.46923.1	CHS.46923.4	CNR1	439	472	0.9301	72.5	68.9	1.05	3174	6222	Pituitary	Adipose_Tissue	1.42	11.25	TRUE	TRUE
CHS.4739.2	CHS.4739.27	LIN9	523	542	0.9649	74.2	69.9	1.06	4104	0	Bone_Marrow	-	4.20	0.00	TRUE	TRUE
CHS.4739.22	CHS.4739.27	LIN9	449	542	0.8284	78.8	69.9	1.13	0	0	-	-	0.00	0.00	TRUE	TRUE
CHS.4739.6	CHS.4739.27	LIN9	524	542	0.9668	73.8	69.9	1.06	206	0	Kidney	-	2.04	0.00	TRUE	TRUE
CHS.4739.7	CHS.4739.27	LIN9	490	542	0.9041	75.6	69.9	1.08	692	0	Testis	-	2.25	0.00	TRUE	TRUE
CHS.47520.4	CHS.47520.7	SGK1	431	526	0.8194	79.3	71.7	1.11	9745	3538	Vagina	Ovary	171.47	117.19	TRUE	TRUE
CHS.47520.5	CHS.47520.7	SGK1	445	526	0.846	78.1	71.7	1.09	9235	3538	Adipose_Tissue	Ovary	25.51	117.19	TRUE	TRUE
CHS.47520.6	CHS.47520.7	SGK1	459	526	0.8726	76.2	71.7	1.06	6276	3538	Small_Intestine	Ovary	244.70	117.19	TRUE	TRUE
CHS.47580.6	CHS.47580.3	IL22RA2	231	263	0.8783	85.7	79.9	1.07	76	0	Breast	-	1.62	0.00	TRUE	FALSE
CHS.47683.10	CHS.47683.11	NMBR	212	390	0.5436	81.6	77.5	1.05	39	0	Testis	-	2.06	0.00	FALSE	TRUE
CHS.47799.14	CHS.47799.17	RAET1E	209	263	0.7947	88	81	1.09	3295	2759	Cervix_Uteri	Vagina	0.97	1.10	FALSE	FALSE
CHS.47800.1	CHS.47800.2	RAET1G	213	334	0.6377	85.3	68.9	1.24	2697	2514	Adrenal_Gland	Esophagus	8.31	9.60	FALSE	FALSE
CHS.47818.19	CHS.47818.3	MTHFD1L	913	978	0.9335	89.7	85.4	1.05	3041	7828	Bone_Marrow	Bone_Marrow	4.31	21.66	TRUE	TRUE
CHS.47847.6	CHS.47847.10	MTRF1L	238	380	0.6263	88.3	83.9	1.05	9658	9643	Fallopian_Tube	Bone_Marrow	5.83	5.05	FALSE	FALSE
CHS.47856.31	CHS.47856.6	OPRM1	300	400	0.75	81.4	76.5	1.06	395	314	Brain	Brain	1.09	2.83	TRUE	TRUE
CHS.47856.33	CHS.47856.6	OPRM1	292	400	0.73	81.1	76.5	1.06	912	314	Breast	Brain	0.19	2.83	TRUE	TRUE
CHS.48020.21	CHS.48020.7	QKI	317	341	0.9296	71.3	67.8	1.05	9173	9761	Nerve	Bone_Marrow	46.12	19.19	TRUE	TRUE
CHS.48020.22	CHS.48020.7	QKI	311	341	0.912	71.7	67.8	1.06	720	9761	Nerve	Bone_Marrow	11.91	19.19	TRUE	TRUE
CHS.48096.2	CHS.48096.1	TTLL2	519	592	0.8767	75.3	69.2	1.09	221	212	Colon	Testis	0.96	16.69	FALSE	FALSE
CHS.4875.4	CHS.4875.3	TRIM67	721	783	0.9208	81	77	1.05	4	318	Brain	Skin	2.63	5.50	TRUE	TRUE
CHS.4882.1	CHS.4882.2	SPRTN	250	489	0.5112	83	61.4	1.35	9307	9193	Bone_Marrow	Testis	1.93	10.82	TRUE	TRUE
CHS.4925.6	CHS.4925.1	SLC35F3	421	490	0.8592	76.2	69.8	1.09	1993	834	Brain	Brain	6.17	6.29	TRUE	TRUE
CHS.4930.9	CHS.4930.4	COA6	79	155	0.5097	86.8	68.9	1.26	9795	9791	Brain	Bone_Marrow	8.98	92.08	FALSE	FALSE
CHS.49971.17	CHS.49971.4	ANKMY2	399	441	0.9048	92.8	85.5	1.09	864	9370	Bone_Marrow	Bone_Marrow	11.73	58.92	TRUE	TRUE
CHS.50025.11	CHS.50025.3	TMEM196	171	179	0.9553	76.2	62.2	1.23	54	689	Uterus	Brain	1.81	2.64	TRUE	TRUE
CHS.50025.4	CHS.50025.3	TMEM196	178	179	0.9944	76.9	62.2	1.24	440	689	Brain	Brain	2.23	2.64	TRUE	TRUE
CHS.50025.5	CHS.50025.3	TMEM196	185	179	1.0335	76.2	62.2	1.23	129	689	Pituitary	Brain	1.66	2.64	TRUE	TRUE
CHS.50025.8	CHS.50025.3	TMEM196	188	179	1.0503	74.7	62.2	1.20	36	689	Uterus	Brain	2.79	2.64	TRUE	TRUE
CHS.50076.30	CHS.50076.10	RAPGEF5	611	883	0.692	77.9	71.7	1.09	4457	2803	Lung	Brain	29.31	21.63	FALSE	FALSE
CHS.50078.2	CHS.50078.16	STEAP1B	245	264	0.928	81.4	77.1	1.06	1514	2595	Skin	Skin	7.15	12.69	FALSE	FALSE

CHS.50168.51	CHS.50168.27	HNRNPA2B1	313	341	0.9179	70.7	67	1.06	40	9791	Esophagus	Bone_Marrow	20.53	250.70	FALSE	FALSE
CHS.50294.16	CHS.50294.4	PDE1C	694	709	0.9788	71.3	67.8	1.05	235	4304	Brain	Heart	3.82	7.40	TRUE	TRUE
CHS.50294.17	CHS.50294.4	PDE1C	769	709	1.0846	72.4	67.8	1.07	9	4304	Brain	Heart	2.75	7.40	TRUE	TRUE
CHS.50333.6	CHS.50333.5	DPY19L1	675	748	0.9024	88.3	83.1	1.06	510	6005	Small_Intestine	Bone_Marrow	4.29	27.82	TRUE	TRUE
CHS.50337.7	CHS.50337.6	TBX20	297	447	0.6644	76.7	65.6	1.17	787	784	Nerve	Bladder	0.42	18.36	TRUE	TRUE
CHS.50360.7	CHS.50360.4	KIAA0895	417	517	0.8066	82.3	71.8	1.15	3097	3029	Fallopian_Tube	Testis	4.36	6.04	TRUE	TRUE
CHS.50510.2	CHS.50510.10	CAMK2B	449	666	0.6742	88.9	70.6	1.26	2227	1736	Prostate	Muscle	1.68	48.00	TRUE	TRUE
CHS.50510.29	CHS.50510.10	CAMK2B	603	666	0.9054	75.7	70.6	1.07	83	1736	Heart	Muscle	2.58	48.00	TRUE	TRUE
CHS.50510.3	CHS.50510.10	CAMK2B	517	666	0.7763	82.7	70.6	1.17	3360	1736	Brain	Muscle	8.80	48.00	TRUE	TRUE
CHS.50510.4	CHS.50510.10	CAMK2B	479	666	0.7192	87.5	70.6	1.24	2839	1736	Pituitary	Muscle	3.93	48.00	TRUE	TRUE
CHS.50510.5	CHS.50510.10	CAMK2B	503	666	0.7553	84	70.6	1.19	3515	1736	Pituitary	Muscle	40.16	48.00	TRUE	TRUE
CHS.50510.6	CHS.50510.10	CAMK2B	518	666	0.7778	82.8	70.6	1.17	3363	1736	Pituitary	Muscle	8.25	48.00	TRUE	TRUE
CHS.50510.7	CHS.50510.10	CAMK2B	492	666	0.7387	78.8	70.6	1.12	1863	1736	Brain	Muscle	5.11	48.00	TRUE	TRUE
CHS.50510.9	CHS.50510.10	CAMK2B	542	666	0.8138	80.6	70.6	1.14	2553	1736	Brain	Muscle	109.48	48.00	TRUE	TRUE
CHS.50615.9	CHS.50615.50	GRB10	536	594	0.9024	76.2	71.9	1.06	8883	866	Brain	Pituitary	4.05	8.57	FALSE	FALSE
CHS.50755.8	CHS.50755.5	ZNF138	172	319	0.5392	76.5	61.6	1.24	8183	6688	Ovary	Bone_Marrow	3.16	1.91	FALSE	FALSE
CHS.50863.4	CHS.50863.3	NSUN5	429	466	0.9206	91.7	87	1.05	9770	9770	Spleen	Bone_Marrow	8.59	42.00	FALSE	FALSE
CHS.50999.3	CHS.50999.2	DTX2	575	622	0.9244	73.1	68.6	1.07	9730	9162	Testis	Testis	4.46	32.50	TRUE	TRUE
CHS.51117.1	CHS.51117.3	ADAM22	859	964	0.8911	74.3	68.6	1.08	949	102	Testis	Brain	0.16	1.85	TRUE	TRUE
CHS.51117.10	CHS.51117.3	ADAM22	869	964	0.9015	73	68.6	1.06	52	102	Brain	Brain	8.11	1.85	TRUE	TRUE
CHS.51117.20	CHS.51117.3	ADAM22	887	964	0.9201	72.1	68.6	1.05	34	102	Adipose_Tissue	Brain	3.80	1.85	TRUE	TRUE
CHS.51117.7	CHS.51117.3	ADAM22	870	964	0.9025	72.9	68.6	1.06	2067	102	Brain	Brain	13.50	1.85	TRUE	TRUE
CHS.51117.8	CHS.51117.3	ADAM22	890	964	0.9232	72.1	68.6	1.05	52	102	Brain	Brain	1.85	1.85	TRUE	TRUE
CHS.51145.18	CHS.51145.7	CDK14	423	469	0.9019	75.2	70.9	1.06	4683	8663	Brain	Blood_Vessel	15.43	16.49	TRUE	TRUE
CHS.51163.1	CHS.51163.9	LRRD1	528	860	0.614	90.6	80	1.13	249	14	Testis	Testis	4.82	12.58	TRUE	FALSE
CHS.51163.5	CHS.51163.9	LRRD1	431	860	0.5012	92.4	80	1.16	56	14	Brain	Testis	1.12	12.58	TRUE	FALSE
CHS.51182.22	CHS.51182.3	VPS50	905	964	0.9388	74.7	70.1	1.07	79	8817	Breast	Brain	1.23	9.40	TRUE	TRUE
CHS.51295.2	CHS.51295.4	ATP5MF	88	94	0.9362	76.1	67.3	1.13	9791	9791	Bone_Marrow	Bone_Marrow	354.98	228.99	FALSE	FALSE
CHS.51297.10	CHS.51297.9	CPSF4	244	269	0.9071	76.9	71.5	1.08	9730	9752	Bone_Marrow	Bone_Marrow	21.33	35.28	TRUE	TRUE
CHS.51297.11	CHS.51297.9	CPSF4	243	269	0.9033	76.4	71.5	1.07	2081	9752	Bone_Marrow	Bone_Marrow	5.63	35.28	TRUE	TRUE
CHS.51374.1	CHS.51374.3	TFR2	630	801	0.7865	91.6	83.9	1.09	2326	975	Brain	Liver	2.54	39.45	TRUE	TRUE
CHS.51427.3	CHS.51427.1	MYL10	167	226	0.7389	82.2	59.8	1.37	41	7	Prostate	Muscle	17.71	0.20	FALSE	FALSE
CHS.51427.4	CHS.51427.1	MYL10	147	226	0.6504	85.3	59.8	1.43	25	7	Muscle	Muscle	2.42	0.20	FALSE	FALSE
CHS.51464.16	CHS.51464.14	ARMC10	308	343	0.898	85.2	79.7	1.07	9754	9508	Bone_Marrow	Bone_Marrow	22.92	15.62	FALSE	FALSE
CHS.51630.18	CHS.51630.1	CAV1	147	178	0.8258	87.3	77.4	1.13	9731	9536	Lung	Adipose_Tissue	210.34	292.83	FALSE	TRUE
CHS.5167.8	CHS.5167.9	ZNF124	289	351	0.8234	82.4	77.2	1.07	6090	3105	Bone_Marrow	Bone_Marrow	2.12	1.75	FALSE	FALSE
CHS.51777.3	CHS.51777.8	IMPDH1	509	599	0.8497	88.2	82.7	1.07	7130	2582	Bone_Marrow	Blood	10.16	5.13	FALSE	FALSE
CHS.51777.4	CHS.51777.8	IMPDH1	514	599	0.8581	89.4	82.7	1.08	2945	2582	Bone_Marrow	Blood	21.84	5.13	FALSE	FALSE
CHS.51777.9	CHS.51777.8	IMPDH1	522	599	0.8715	88.3	82.7	1.07	6906	2582	Blood	Blood	25.90	5.13	FALSE	FALSE
CHS.51908.14	CHS.51908.1	AKR1B15	316	344	0.9186	96.3	85.1	1.13	183	100	Skin	Prostate	11.93	9.34	FALSE	FALSE

CHS.51908.3	CHS.51908.1	AKR1B15	248	344	0.7209	96.3	85.1	1.13	327	100	Vagina	Prostate	6.23	9.34	FALSE	FALSE
CHS.51973.8	CHS.51973.7	ZC3HAV1L	276	300	0.92	86.2	81.5	1.06	615	2417	Prostate	Ovary	3.80	1.74	FALSE	FALSE
CHS.52202.19	CHS.52202.16	ACTR3C	194	210	0.9238	90.9	85.6	1.06	55	2127	Thyroid	Liver	4.55	6.40	FALSE	FALSE
CHS.52202.21	CHS.52202.16	ACTR3C	189	210	0.9	91.7	85.6	1.07	279	2127	Liver	Liver	5.75	6.40	FALSE	FALSE
CHS.52273.9	CHS.52273.5	CRYGN	182	182	1	92.2	67.7	1.36	391	996	Thyroid	Thyroid	3.70	4.06	FALSE	FALSE
CHS.52277.5	CHS.52277.8	PRKAG2	328	569	0.5764	88.6	64.7	1.37	9624	8110	Bladder	Bladder	28.67	22.55	TRUE	TRUE
CHS.52315.10	CHS.52315.20	DPP6	801	865	0.926	90.3	85.5	1.06	2845	1511	Uterus	Uterus	54.05	9.66	TRUE	TRUE
CHS.52315.22	CHS.52315.20	DPP6	803	865	0.9283	90.3	85.5	1.06	2609	1511	Brain	Uterus	29.09	9.66	TRUE	TRUE
CHS.5236.5	CHS.5236.12	SH3BP5L	361	393	0.9186	75.4	71.8	1.05	4	9773	Lung	Spleen	0.65	25.81	TRUE	TRUE
CHS.52917.5	CHS.52917.6	BMP1	479	730	0.6562	90.1	81.7	1.10	9182	9155	Cervix_Uteri	Cervix_Uteri	4.23	4.01	FALSE	FALSE
CHS.52917.8	CHS.52917.6	BMP1	735	730	1.0068	86.9	81.7	1.06	9150	9155	Adrenal_Gland	Cervix_Uteri	16.81	4.01	FALSE	FALSE
CHS.52995.1	CHS.52995.14	ADAM28	540	775	0.6968	79.2	74.4	1.06	2035	1969	Stomach	Stomach	1.30	22.95	TRUE	FALSE
CHS.53035.7	CHS.53035.1	DPYSL2	572	677	0.8449	89.4	81.8	1.09	9488	6865	Brain	Brain	186.26	18.95	TRUE	TRUE
CHS.53038.10	CHS.53038.12	ADRA1A	427	466	0.9163	72.2	66	1.09	24	3221	Heart	Liver	4.97	6.62	TRUE	TRUE
CHS.53038.9	CHS.53038.12	ADRA1A	429	466	0.9206	70.9	66	1.07	2375	3221	Liver	Liver	38.03	6.62	TRUE	TRUE
CHS.53042.3	CHS.53042.4	STMN4	189	216	0.875	80.4	73	1.10	2864	2147	Brain	Brain	52.68	20.41	TRUE	TRUE
CHS.53126.18	CHS.53126.4	RBPMS	179	196	0.9133	70.4	66.6	1.06	7	9262	Blood_Vessel	Bladder	7.20	151.71	TRUE	TRUE
CHS.53126.20	CHS.53126.4	RBPMS	186	196	0.949	71.5	66.6	1.07	7	9262	Heart	Bladder	1.64	151.71	TRUE	TRUE
CHS.53126.24	CHS.53126.4	RBPMS	179	196	0.9133	70	66.6	1.05	2315	9262	Blood_Vessel	Bladder	111.86	151.71	FALSE	TRUE
CHS.53251.29	CHS.53251.5	ADAM9	752	819	0.9182	80.2	75.8	1.06	8	9566	Lung	Skin	1.38	84.55	FALSE	FALSE
CHS.53411.2	CHS.53411.1	OPRK1	291	380	0.7658	82.8	78	1.06	22	0	Pituitary	-	2.27	0.00	TRUE	TRUE
CHS.53416.13	CHS.53416.1	RGS20	241	388	0.6211	73.2	59	1.24	2601	96	Brain	Brain	12.26	0.83	FALSE	FALSE
CHS.53416.20	CHS.53416.1	RGS20	220	388	0.567	76.6	59	1.30	364	96	Thyroid	Brain	2.31	0.83	TRUE	FALSE
CHS.53416.22	CHS.53416.1	RGS20	224	388	0.5773	74.6	59	1.26	720	96	Brain	Brain	2.49	0.83	TRUE	FALSE
CHS.53416.8	CHS.53416.1	RGS20	217	388	0.5593	77.5	59	1.31	973	96	Brain	Brain	5.17	0.83	FALSE	FALSE
CHS.53548.1	CHS.53548.40	ASPH	686	758	0.905	72.5	68.5	1.06	36	9405	Adipose_Tissue	Adrenal_Gland	23.33	54.05	FALSE	FALSE
CHS.53548.24	CHS.53548.40	ASPH	711	758	0.938	73.5	68.5	1.07	28	9405	Blood	Adrenal_Gland	3.17	54.05	FALSE	FALSE
CHS.53548.25	CHS.53548.40	ASPH	729	758	0.9617	72.2	68.5	1.05	26	9405	Colon	Adrenal_Gland	7.83	54.05	FALSE	FALSE
CHS.53548.27	CHS.53548.40	ASPH	730	758	0.9631	72.1	68.5	1.05	27	9405	Pituitary	Adrenal_Gland	16.39	54.05	FALSE	FALSE
CHS.53548.7	CHS.53548.40	ASPH	724	758	0.9551	72.7	68.5	1.06	5	9405	Stomach	Adrenal_Gland	3.67	54.05	FALSE	FALSE
CHS.53558.7	CHS.53558.5	NKAIN3	197	218	0.9037	74.4	70	1.06	3	730	Nerve	Stomach	0.31	1.79	FALSE	TRUE
CHS.53611.5	CHS.53611.4	DNAJC5B	180	199	0.9045	77.3	69	1.12	11	330	Testis	Testis	3.76	55.98	FALSE	FALSE
CHS.53613.13	CHS.53613.12	TRIM55	452	548	0.8248	71.7	65.2	1.10	1695	1288	Heart	Muscle	22.22	10.65	TRUE	TRUE
CHS.53660.3	CHS.53660.2	PRDM14	316	571	0.5534	73.7	60.5	1.22	29	14	Testis	Testis	1.26	1.23	TRUE	TRUE
CHS.53700.42	CHS.53700.6	STAU2	549	570	0.9632	72.9	68.8	1.06	2627	9035	Bone_Marrow	Muscle	8.12	9.72	FALSE	FALSE
CHS.53700.43	CHS.53700.6	STAU2	517	570	0.907	72.4	68.8	1.05	5412	9035	Bone_Marrow	Muscle	8.37	9.72	FALSE	FALSE
CHS.53727.5	CHS.53727.18	HNF4G	408	455	0.8967	77.5	72.9	1.06	429	243	Small_Intestine	Small_Intestine	16.00	21.12	FALSE	FALSE
CHS.53867.1	CHS.53867.3	CNGB3	671	809	0.8294	73.3	67	1.09	93	4	Testis	Ovary	3.98	2.88	TRUE	TRUE
CHS.53905.1	CHS.53905.4	OSGIN2	505	549	0.9199	86.4	81	1.07	5584	9200	Testis	Nerve	14.70	26.29	TRUE	TRUE
CHS.5414.12	CHS.5414.13	FBH1	969	1043	0.9291	81	76.9	1.05	4484	9648	Fallopian_Tube	Bone_Marrow	9.90	26.15	FALSE	TRUE

CHS.54278.1	CHS.54278.5	SLC30A8	320	369	0.8672	84.1	75.3	1.12	478	208	Pancreas	Pancreas	4.22	3.99	FALSE	FALSE
CHS.54658.20	CHS.54658.3	TSNARE1	277	513	0.54	73.1	67.2	1.09	5494	2908	Bladder	Nerve	1.76	1.38	FALSE	FALSE
CHS.54733.2	CHS.54733.1	CCDC166	399	439	0.9089	73	69.3	1.05	43	371	Testis	Testis	1.93	5.80	FALSE	TRUE
CHS.54803.17	CHS.54803.6	ZNF34	499	539	0.9258	70.6	67	1.05	432	9211	Breast	Nerve	1.41	7.64	FALSE	FALSE
CHS.54940.4	CHS.54940.3	CBWD1	359	395	0.9089	79.1	75	1.05	9656	9696	Thyroid	Thyroid	12.34	17.87	FALSE	TRUE
CHS.54984.43	CHS.54984.12	RFX3	682	749	0.9105	70.2	66	1.06	30	4064	Vagina	Fallopian_Tube	1.52	7.80	TRUE	TRUE
CHS.55319.28	CHS.55319.27	AQP7	257	342	0.7515	91.8	82.8	1.11	5334	4961	Adipose_Tissue	Adipose_Tissue	18.02	66.98	FALSE	FALSE
CHS.55340.2	CHS.55340.15	PRSS3	139	247	0.5628	96.4	91.5	1.05	5226	4045	Pancreas	Pancreas	360.29	2856.79	FALSE	FALSE
CHS.55379.11	CHS.55379.1	DNAJB5	382	420	0.9095	74.2	69.1	1.07	2937	6703	Brain	Pituitary	2.54	6.17	TRUE	FALSE
CHS.55379.8	CHS.55379.1	DNAJB5	348	420	0.8286	78.4	69.1	1.13	9411	6703	Blood_Vessel	Pituitary	64.16	6.17	FALSE	FALSE
CHS.5572.4	CHS.5572.27	FAM107B	170	306	0.5556	76.2	60.6	1.26	5116	314	Colon	Testis	2.79	8.23	TRUE	TRUE
CHS.55721.9	CHS.55721.5	CFAP95	130	229	0.5677	70.9	67.4	1.05	1067	845	Pituitary	Testis	8.74	22.92	FALSE	FALSE
CHS.55846.5	CHS.55846.4	TLE4	704	773	0.9107	70	66.4	1.05	3302	9356	Testis	Testis	6.96	52.78	TRUE	TRUE
CHS.55860.19	CHS.55860.14	TLE1	696	770	0.9039	70.7	66.8	1.06	542	9447	Fallopian_Tube	Nerve	3.33	66.85	TRUE	TRUE
CHS.55860.30	CHS.55860.14	TLE1	695	770	0.9026	70.6	66.8	1.06	439	9447	Cervix_Uteri	Nerve	4.35	66.85	TRUE	TRUE
CHS.55898.8	CHS.55898.7	SLC28A3	622	691	0.9001	85.1	79.2	1.07	330	1382	Breast	Adipose_Tissue	4.43	6.61	FALSE	TRUE
CHS.55901.25	CHS.55901.33	NTRK2	477	838	0.5692	81.1	75.8	1.07	6121	29	Brain	Brain	39.21	5.62	TRUE	FALSE
CHS.56089.19	CHS.56089.7	CARD19	170	183	0.929	75.6	69.5	1.09	19	9775	Esophagus	Testis	4.52	20.93	TRUE	TRUE
CHS.56137.32	CHS.56137.18	AOPEP	529	819	0.6459	87.8	82.4	1.07	5670	257	Uterus	Blood_Vessel	6.52	7.44	TRUE	TRUE
CHS.56174.30	CHS.56174.25	CDC14B	462	498	0.9277	82	78	1.05	109	7008	Nerve	Testis	5.16	7.22	TRUE	TRUE
CHS.56239.1	CHS.56239.4	TBC1D2	468	928	0.5043	87.2	74.2	1.18	9540	8509	Brain	Lung	3.00	36.58	TRUE	TRUE
CHS.56430.11	CHS.56430.8	EPB41L4B	518	900	0.5756	71.2	55.5	1.28	5649	1859	Pancreas	Brain	88.56	12.00	TRUE	TRUE
CHS.56446.10	CHS.56446.8	TXNDC8	105	95	1.1053	87.5	56.7	1.54	3	186	Testis	Testis	1.71	1.28	TRUE	TRUE
CHS.56446.13	CHS.56446.8	TXNDC8	128	95	1.3474	78.7	56.7	1.39	4	186	Testis	Testis	0.90	1.28	TRUE	TRUE
CHS.56446.14	CHS.56446.8	TXNDC8	105	95	1.1053	96.9	56.7	1.71	58	186	Testis	Testis	4.37	1.28	TRUE	TRUE
CHS.56446.6	CHS.56446.8	TXNDC8	127	95	1.3368	84.4	56.7	1.49	194	186	Testis	Testis	5.51	1.28	TRUE	TRUE
CHS.56446.9	CHS.56446.8	TXNDC8	115	95	1.2105	82.3	56.7	1.45	181	186	Testis	Testis	1.42	1.28	TRUE	TRUE
CHS.56522.3	CHS.56522.5	WDR31	242	367	0.6594	94.6	86.2	1.10	3259	3186	Pituitary	Testis	1.92	4.08	FALSE	FALSE
CHS.56730.1	CHS.56730.5	NR6A1	438	480	0.9125	79.8	75.6	1.06	51	1629	Small_Intestine	Testis	4.02	19.14	TRUE	TRUE
CHS.56730.19	CHS.56730.5	NR6A1	437	480	0.9104	79.9	75.6	1.06	40	1629	Colon	Testis	3.57	19.14	TRUE	TRUE
CHS.56760.8	CHS.56760.1	PBX3	351	434	0.8088	74.9	66.2	1.13	9480	9464	Ovary	Adrenal_Gland	35.44	27.24	TRUE	TRUE
CHS.56771.1	CHS.56771.8	MVB12B	221	319	0.6928	80.4	72.7	1.11	8798	8771	Lung	Nerve	0.56	36.26	TRUE	TRUE
CHS.56815.10	CHS.56815.16	SH2D3C	703	860	0.8174	70.1	62.3	1.13	8857	5091	Lung	Spleen	49.24	30.66	TRUE	TRUE
CHS.56815.7	CHS.56815.16	SH2D3C	506	860	0.5884	72.5	62.3	1.16	6605	5091	Lung	Spleen	15.54	30.66	TRUE	TRUE
CHS.56827.9	CHS.56827.11	ST6GALNAC6	299	333	0.8979	88.1	83.9	1.05	9748	9733	Brain	Brain	16.15	13.27	TRUE	TRUE
CHS.56863.2	CHS.56863.18	ODF2	657	912	0.7204	77.8	72.2	1.08	1809	28	Testis	Liver	3.08	2.39	TRUE	TRUE
CHS.56863.36	CHS.56863.18	ODF2	767	912	0.841	78.5	72.2	1.09	77	28	Uterus	Liver	4.22	2.39	TRUE	TRUE
CHS.56863.37	CHS.56863.18	ODF2	824	912	0.9035	76.2	72.2	1.06	9578	28	Bone_Marrow	Liver	54.86	2.39	TRUE	TRUE
CHS.56863.39	CHS.56863.18	ODF2	748	912	0.8202	79.3	72.2	1.10	35	28	Adipose_Tissue	Liver	2.13	2.39	TRUE	TRUE
CHS.56863.42	CHS.56863.18	ODF2	638	912	0.6996	77.9	72.2	1.08	4686	28	Testis	Liver	12.07	2.39	TRUE	TRUE

CHS.56863.53	CHS.56863.18	ODF2	638	912	0.6996	78.2	72.2	1.08	922	28	Testis	Liver	122.57	2.39	TRUE	TRUE
CHS.56863.6	CHS.56863.18	ODF2	576	912	0.6316	80.9	72.2	1.12	6833	28	Bone_Marrow	Liver	1.02	2.39	TRUE	TRUE
CHS.56863.79	CHS.56863.18	ODF2	657	912	0.7204	77	72.2	1.07	62	28	Testis	Liver	62.02	2.39	TRUE	TRUE
CHS.56863.82	CHS.56863.18	ODF2	633	912	0.6941	78.2	72.2	1.08	678	28	Kidney	Liver	8.51	2.39	TRUE	TRUE
CHS.56863.83	CHS.56863.18	ODF2	652	912	0.7149	78.3	72.2	1.08	480	28	Uterus	Liver	8.30	2.39	TRUE	TRUE
CHS.56863.9	CHS.56863.18	ODF2	805	912	0.8827	76.5	72.2	1.06	1168	28	Bone_Marrow	Liver	13.15	2.39	FALSE	TRUE
CHS.56953.21	CHS.56953.10	FNBP1	556	617	0.9011	81.1	76.1	1.07	188	9131	Uterus	Uterus	7.49	30.93	TRUE	TRUE
CHS.56953.22	CHS.56953.10	FNBP1	560	617	0.9076	80.6	76.1	1.06	736	9131	Uterus	Uterus	5.19	30.93	TRUE	TRUE
CHS.56953.23	CHS.56953.10	FNBP1	561	617	0.9092	80.6	76.1	1.06	111	9131	Stomach	Uterus	3.27	30.93	TRUE	TRUE
CHS.56953.3	CHS.56953.10	FNBP1	556	617	0.9011	80.7	76.1	1.06	1213	9131	Bladder	Uterus	26.68	30.93	TRUE	TRUE
CHS.57019.3	CHS.57019.7	DDX31	674	746	0.9035	76.1	72	1.06	30	8668	Nerve	Thyroid	3.30	14.29	TRUE	TRUE
CHS.57048.3	CHS.57048.2	MED22	140	200	0.7	89.2	75.2	1.19	9727	8186	Uterus	Brain	33.39	7.25	FALSE	FALSE
CHS.57107.1	CHS.57107.11	OLFM1	467	485	0.9629	80.4	75.5	1.06	9029	8833	Brain	Brain	98.90	49.72	TRUE	TRUE
CHS.57107.8	CHS.57107.11	OLFM1	458	485	0.9443	83.3	75.5	1.10	731	8833	Brain	Brain	10.30	49.72	TRUE	TRUE
CHS.57164.1	CHS.57164.2	GPSM1	457	675	0.677	82.5	69.5	1.19	8493	8436	Nerve	Nerve	0.78	52.71	TRUE	TRUE
CHS.57426.3	CHS.57426.2	ASMT	298	373	0.7989	93	87.1	1.07	728	256	Blood	Blood	0.93	0.87	FALSE	FALSE
CHS.57426.4	CHS.57426.2	ASMT	345	373	0.9249	94.7	87.1	1.09	75	256	Cervix_Uteri	Blood	1.12	0.87	FALSE	FALSE
CHS.57447.9	CHS.57447.10	GYG2	430	470	0.9149	72.9	68.5	1.06	28	3989	Brain	Adipose_Tissue	1.46	46.65	FALSE	FALSE
CHS.57496.3	CHS.57496.7	TBL1X	526	577	0.9116	86.7	82.3	1.05	5839	7404	Uterus	Uterus	13.88	55.08	FALSE	FALSE
CHS.57568.3	CHS.57568.4	GPM6B	265	328	0.8079	83.1	77.5	1.07	5853	3673	Nerve	Brain	38.50	4.10	TRUE	TRUE
CHS.57568.8	CHS.57568.4	GPM6B	246	328	0.75	84.8	77.5	1.09	8091	3673	Brain	Brain	220.19	4.10	TRUE	TRUE
CHS.57640.2	CHS.57640.1	RS1	192	224	0.8571	81.9	76.7	1.07	275	0	Lung	-	2.95	0.00	TRUE	TRUE
CHS.57851.18	CHS.57851.4	DDX3X	598	662	0.9033	74.5	70.7	1.05	4	9772	Adipose_Tissue	Fallopian_Tube	5.79	102.21	FALSE	FALSE
CHS.58025.7	CHS.58025.4	CLCN5	746	816	0.9142	85.9	79.9	1.08	758	1260	Brain	Ovary	5.99	8.40	TRUE	TRUE
CHS.58077.7	CHS.58077.6	RIBC1	238	379	0.628	90.3	85.7	1.05	3979	2406	Fallopian_Tube	Testis	10.80	43.75	FALSE	FALSE
CHS.58147.12	CHS.58147.25	ARHGEF9	414	523	0.7916	84.5	77.6	1.09	7482	140	Brain	Ovary	5.26	8.36	FALSE	TRUE
CHS.58212.6	CHS.58212.4	DLG3	512	817	0.6267	76.2	72	1.06	7323	1063	Salivary_Gland	Skin	15.41	22.42	TRUE	TRUE
CHS.58315.2	CHS.58315.17	MAGT1	367	335	1.0955	83.4	83.4	1.00	9655	0	Thyroid	-	51.56	0.00	TRUE	TRUE
CHS.58424.11	CHS.58424.3	DRP2	879	957	0.9185	77.3	72.7	1.06	423	1141	Nerve	Nerve	215.04	15.44	TRUE	FALSE
CHS.58489.17	CHS.58489.16	PLP1	242	277	0.8736	84.1	76.1	1.11	7674	7207	Nerve	Brain	358.47	1626.60	TRUE	TRUE
CHS.58539.2	CHS.58539.3	MID2	685	735	0.932	83.1	78.2	1.06	1595	6816	Nerve	Uterus	2.98	3.85	TRUE	TRUE
CHS.58771.8	CHS.58771.5	SLC25A14	290	325	0.8923	81.3	68.2	1.19	9063	5532	Pituitary	Brain	15.29	10.78	TRUE	TRUE
CHS.58840.2	CHS.58840.1	CT55	242	264	0.9167	76.7	72	1.07	399	380	Bone_Marrow	Bone_Marrow	13.99	19.99	FALSE	FALSE
CHS.58861.24	CHS.58861.30	SLC9A6	512	679	0.7541	80.1	70.9	1.13	1	0	Breast	-	0.53	0.00	FALSE	FALSE
CHS.58862.11	CHS.58862.3	FHL1	280	323	0.8669	88.7	68.5	1.29	9681	8843	Heart	Cervix_Uteri	3.00	154.64	TRUE	TRUE
CHS.58862.22	CHS.58862.3	FHL1	309	323	0.9567	87.3	68.5	1.27	8746	8843	Muscle	Cervix_Uteri	2.20	154.64	TRUE	TRUE
CHS.58862.23	CHS.58862.3	FHL1	296	323	0.9164	87.6	68.5	1.28	9645	8843	Muscle	Cervix_Uteri	3330.57	154.64	TRUE	TRUE
CHS.58862.29	CHS.58862.3	FHL1	210	323	0.6502	78.4	68.5	1.14	9401	8843	Muscle	Cervix_Uteri	62.28	154.64	TRUE	TRUE
CHS.58862.5	CHS.58862.3	FHL1	194	323	0.6006	78.8	68.5	1.15	9426	8843	Heart	Cervix_Uteri	19.06	154.64	TRUE	TRUE
CHS.59047.3	CHS.59047.1	SRPK3	533	567	0.94	80.7	76.1	1.06	2010	3629	Prostate	Muscle	1.85	32.30	TRUE	TRUE

CHS.59048.14	CHS.59048.4	IDH3G	363	393	0.9237	91.5	87.1	1.05	380	9782	Small_Intestine	Blood_Vessel	7.92	67.09	TRUE	TRUE
CHS.5906.26	CHS.5906.8	NRP1	644	923	0.6977	83.8	77.8	1.08	8463	8422	Kidney	Adipose_Tissue	1.23	55.14	TRUE	TRUE
CHS.59079.10	CHS.59079.8	TAFAZZIN	262	292	0.8973	94	86.2	1.09	9753	9740	Spleen	Spleen	9.62	9.82	FALSE	FALSE
CHS.59107.19	CHS.59107.9	MPP1	420	466	0.9013	83.9	78.6	1.07	39	9754	Ovary	Bone_Marrow	4.04	473.00	FALSE	TRUE
CHS.59117.1	CHS.59117.12	VBP1	192	197	0.9746	85.6	76.6	1.12	57	0	Bone_Marrow	-	4.51	0.00	TRUE	TRUE
CHS.59117.3	CHS.59117.12	VBP1	233	197	1.1827	76.6	76.6	1.00	9774	0	Bone_Marrow	-	138.74	0.00	TRUE	TRUE
CHS.59117.7	CHS.59117.12	VBP1	160	197	0.8122	91.6	76.6	1.20	2443	0	Fallopian_Tube	-	3.38	0.00	TRUE	TRUE
CHS.6033.6	CHS.6033.4	ZNF32	248	273	0.9084	82.1	75.7	1.08	7080	9748	Thyroid	Uterus	4.30	64.19	TRUE	TRUE
CHS.6106.52	CHS.6106.6	PTPN20	228	420	0.5429	93.3	77.2	1.21	135	116	Blood_Vessel	Testis	1.21	2.22	FALSE	FALSE
CHS.6106.9	CHS.6106.6	PTPN20	339	420	0.8071	87.7	77.2	1.14	1219	116	Spleen	Testis	1.84	2.22	FALSE	FALSE
CHS.6139.3	CHS.6139.1	FAM25C	139	89	1.5618	75.9	57	1.33	480	1710	Salivary_Gland	Salivary_Gland	4.80	19.31	FALSE	FALSE
CHS.6144.16	CHS.6144.13	MAPK8	384	427	0.8993	86	81.2	1.06	5526	5192	Bone_Marrow	Brain	10.78	5.20	TRUE	TRUE
CHS.6165.5	CHS.6165.6	CHAT	666	748	0.8904	87.9	81.7	1.08	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.6165.8	CHS.6165.6	CHAT	630	748	0.8422	91.5	81.7	1.12	135	0	Vagina	-	3.07	0.00	FALSE	FALSE
CHS.6339.6	CHS.6339.5	DNAJC12	107	198	0.5404	75.8	64.7	1.17	8543	7948	Bone_Marrow	Adrenal_Gland	13.24	68.79	TRUE	TRUE
CHS.6367.13	CHS.6367.1	DDX21	715	783	0.9132	74.4	70.4	1.06	4759	9586	Fallopian_Tube	Bone_Marrow	1.89	204.87	TRUE	TRUE
CHS.6368.7	CHS.6368.1	DDX50	672	737	0.9118	78.2	73.5	1.06	5421	9716	Bone_Marrow	Bone_Marrow	6.60	70.35	TRUE	TRUE
CHS.64.16	CHS.64.6	TTLL10	404	673	0.6003	72.3	66.8	1.08	279	202	Testis	Cervix_Uteri	16.68	3.21	FALSE	FALSE
CHS.645.52	CHS.645.10	TMCO4	580	634	0.9148	75.1	71.2	1.05	2	5866	Skin	Prostate	1.10	7.91	FALSE	FALSE
CHS.6496.2	CHS.6496.4	ANXA7	426	466	0.9142	81.3	77.4	1.05	2600	9784	Bone_Marrow	Bone_Marrow	3.90	136.20	FALSE	FALSE
CHS.6517.11	CHS.6517.28	CAMK2G	542	588	0.9218	80.3	76.3	1.05	155	63	Blood_Vessel	Brain	6.17	5.79	TRUE	TRUE
CHS.6517.116	CHS.6517.28	CAMK2G	535	588	0.9099	81.2	76.3	1.06	0	63	-	Brain	0.00	5.79	TRUE	TRUE
CHS.6517.12	CHS.6517.28	CAMK2G	506	588	0.8605	83.4	76.3	1.09	666	63	Brain	Brain	10.54	5.79	TRUE	TRUE
CHS.6517.15	CHS.6517.28	CAMK2G	529	588	0.8997	81.8	76.3	1.07	4797	63	Blood_Vessel	Brain	15.80	5.79	TRUE	TRUE
CHS.6517.17	CHS.6517.28	CAMK2G	516	588	0.8776	82.9	76.3	1.09	8551	63	Blood_Vessel	Brain	10.53	5.79	TRUE	TRUE
CHS.6517.26	CHS.6517.28	CAMK2G	550	588	0.9354	80.2	76.3	1.05	56	63	Pituitary	Brain	6.30	5.79	TRUE	TRUE
CHS.6517.32	CHS.6517.28	CAMK2G	495	588	0.8418	84.5	76.3	1.11	9477	63	Blood_Vessel	Brain	9.51	5.79	TRUE	TRUE
CHS.6517.33	CHS.6517.28	CAMK2G	504	588	0.8571	83.7	76.3	1.10	672	63	Bladder	Brain	7.36	5.79	TRUE	TRUE
CHS.6517.34	CHS.6517.28	CAMK2G	518	588	0.881	82.8	76.3	1.09	5348	63	Blood_Vessel	Brain	42.49	5.79	TRUE	TRUE
CHS.6517.36	CHS.6517.28	CAMK2G	539	588	0.9167	81	76.3	1.06	8713	63	Bladder	Brain	27.64	5.79	TRUE	TRUE
CHS.6517.37	CHS.6517.28	CAMK2G	527	588	0.8963	82.1	76.3	1.08	6057	63	Brain	Brain	6.57	5.79	TRUE	TRUE
CHS.6517.76	CHS.6517.28	CAMK2G	519	588	0.8827	81.2	76.3	1.06	64	63	Vagina	Brain	10.32	5.79	TRUE	TRUE
CHS.6517.78	CHS.6517.28	CAMK2G	533	588	0.9065	81.1	76.3	1.06	217	63	Bladder	Brain	8.40	5.79	TRUE	TRUE
CHS.6517.8	CHS.6517.28	CAMK2G	510	588	0.8673	82	76.3	1.07	1104	63	Prostate	Brain	6.70	5.79	TRUE	TRUE
CHS.6534.9	CHS.6534.2	DUSP13	198	334	0.5928	89.8	66.6	1.35	1318	1308	Testis	Testis	22.76	54.51	TRUE	TRUE
CHS.657.6	CHS.657.3	PLA2G2F	168	211	0.7962	86.4	79.5	1.09	743	313	Bladder	Skin	43.87	4.65	TRUE	TRUE
CHS.6787.3	CHS.6787.2	HTR7	432	479	0.9019	75.7	71.3	1.06	1976	620	Testis	Testis	3.23	3.14	FALSE	FALSE
CHS.6802.5	CHS.6802.18	PCGF5	236	256	0.9219	84.1	79.4	1.06	2551	9357	Cervix_Uteri	Blood_Vessel	5.77	12.30	TRUE	TRUE
CHS.6927.2	CHS.6927.25	ANKRD2	360	333	1.0811	76.4	76.4	1.00	0	0	-	-	0.00	0.00	TRUE	TRUE
CHS.6943.2	CHS.6943.9	CRTAC1	610	661	0.9228	90.8	85.7	1.06	7	5914	Brain	Bladder	2.37	75.58	TRUE	TRUE

CHS.7029.1	CHS.7029.4	FGF8	204	244	0.8361	85.1	78.9	1.08	116	79	Breast	Adipose_Tissue	5.20	1.22	TRUE	TRUE
CHS.7029.2	CHS.7029.4	FGF8	215	244	0.8811	84.9	78.9	1.08	108	79	Testis	Adipose_Tissue	1.07	1.22	TRUE	TRUE
CHS.7029.5	CHS.7029.4	FGF8	140	244	0.5738	83.6	78.9	1.06	669	79	Testis	Adipose_Tissue	1.76	1.22	FALSE	TRUE
CHS.7034.13	CHS.7034.17	KCNIP2	225	270	0.8333	74.5	66.5	1.12	3802	2858	Brain	Brain	20.38	44.36	TRUE	TRUE
CHS.7034.14	CHS.7034.17	KCNIP2	220	270	0.8148	75.3	66.5	1.13	4874	2858	Heart	Brain	114.16	44.36	TRUE	TRUE
CHS.7118.16	CHS.7118.4	CFAP43	940	1665	0.5646	84.4	77.4	1.09	691	403	Fallopian_Tube	Fallopian_Tube	2.06	38.39	FALSE	FALSE
CHS.7208.8	CHS.7208.6	TECTB	182	329	0.5532	85	67.8	1.25	24	1	Thyroid	Brain	2.13	0.66	TRUE	TRUE
CHS.7287.17	CHS.7287.6	SHTN1	573	631	0.9081	72.3	68.2	1.06	158	8468	Small_Intestine	Pituitary	3.70	26.29	TRUE	TRUE
CHS.7394.34	CHS.7394.9	BTBD16	271	506	0.5356	86.7	78.1	1.11	508	340	Bladder	Testis	19.57	13.60	FALSE	FALSE
CHS.7394.35	CHS.7394.9	BTBD16	338	506	0.668	91.8	78.1	1.18	578	340	Bladder	Testis	33.31	13.60	FALSE	FALSE
CHS.7396.55	CHS.7396.3	PLEKHA1	372	404	0.9208	71.4	67.3	1.06	27	7812	Thyroid	Skin	2.29	5.32	TRUE	TRUE
CHS.7396.7	CHS.7396.3	PLEKHA1	366	404	0.9059	71.9	67.3	1.07	14	7812	Testis	Skin	6.29	5.32	TRUE	TRUE
CHS.7396.97	CHS.7396.3	PLEKHA1	380	404	0.9406	72.1	67.3	1.07	0	7812	-	Skin	0.00	5.32	TRUE	TRUE
CHS.7460.20	CHS.7460.3	CTBP2	513	985	0.5208	72.6	56.4	1.29	3917	1203	Bladder	Pituitary	0.42	0.73	TRUE	TRUE
CHS.75.18	CHS.75.17	SCNN1D	638	802	0.7955	74.2	63.8	1.16	5300	4266	Testis	Testis	7.41	8.62	FALSE	FALSE
CHS.7640.2	CHS.7640.6	SYCE1	282	351	0.8034	85.3	77.3	1.10	1300	911	Brain	Testis	2.49	9.91	FALSE	FALSE
CHS.7704.14	CHS.7704.13	RASSF7	320	373	0.8579	79	73.3	1.08	6969	6805	Kidney	Kidney	3.46	3.70	TRUE	TRUE
CHS.7812.7	CHS.7812.4	TSPAN32	288	320	0.9	83	70.1	1.18	1567	2655	Prostate	Blood	1.05	2.88	FALSE	FALSE
CHS.7860.23	CHS.7860.58	PGAP2	294	315	0.9333	84	78	1.08	463	4116	Testis	Lung	6.96	1.48	FALSE	FALSE
CHS.7860.54	CHS.7860.58	PGAP2	254	315	0.8063	87.9	78	1.13	8776	4116	Skin	Lung	7.60	1.48	FALSE	FALSE
CHS.7860.82	CHS.7860.58	PGAP2	250	315	0.7937	86.5	78	1.11	8675	4116	Skin	Lung	7.79	1.48	FALSE	FALSE
CHS.7895.6	CHS.7895.5	MMP26	191	261	0.7318	89	78.2	1.14	0	0	-	-	0.00	0.00	FALSE	FALSE
CHS.8003.4	CHS.8003.5	OVCH2	302	565	0.5345	84.8	80.6	1.05	30	0	Kidney	-	2.69	0.00	FALSE	FALSE
CHS.809.20	CHS.809.21	NIPAL3	367	406	0.9039	78.6	74	1.06	80	8664	Brain	Brain	6.57	40.81	TRUE	TRUE
CHS.809.25	CHS.809.21	NIPAL3	368	406	0.9064	79.3	74	1.07	7448	8664	Prostate	Brain	8.86	40.81	FALSE	TRUE
CHS.8123.15	CHS.8123.5	MICAL2	934	1124	0.831	74	70.2	1.05	4876	2560	Blood_Vessel	Skin	6.49	54.10	FALSE	FALSE
CHS.8123.23	CHS.8123.5	MICAL2	955	1124	0.8496	75.6	70.2	1.08	8931	2560	Blood_Vessel	Skin	39.42	54.10	FALSE	FALSE
CHS.8123.52	CHS.8123.5	MICAL2	877	1124	0.7802	76.9	70.2	1.10	2842	2560	Blood_Vessel	Skin	14.17	54.10	FALSE	FALSE
CHS.8211.15	CHS.8211.10	KCNC1	528	585	0.9026	76.8	71.4	1.08	33	1299	Brain	Skin	8.94	94.10	TRUE	TRUE
CHS.8211.9	CHS.8211.10	KCNC1	511	585	0.8735	78.7	71.4	1.10	2161	1299	Brain	Skin	10.73	94.10	TRUE	TRUE
CHS.8404.16	CHS.8404.6	FBXO3	415	471	0.8811	94.3	87.5	1.08	9701	9688	Bone_Marrow	Thyroid	6.39	29.00	TRUE	TRUE
CHS.8404.22	CHS.8404.6	FBXO3	430	471	0.913	93	87.5	1.06	46	9688	Salivary_Gland	Thyroid	5.35	29.00	TRUE	TRUE
CHS.8409.7	CHS.8409.18	LMO2	158	227	0.696	87	71.7	1.21	9580	476	Lung	Thyroid	30.19	8.19	TRUE	FALSE
CHS.8426.2	CHS.8426.6	ELF5	232	255	0.9098	75.4	69.3	1.09	5	1201	Lung	Salivary_Gland	1.90	55.84	TRUE	TRUE
CHS.8462.18	CHS.8462.1	PRR5L	205	368	0.5571	77.8	65.6	1.19	5017	1839	Spleen	Brain	5.50	3.53	TRUE	TRUE
CHS.851.22	CHS.851.4	LDLRAP1	281	308	0.9123	71	67.2	1.06	132	9690	Brain	Spleen	2.43	84.73	TRUE	TRUE
CHS.851.7	CHS.851.4	LDLRAP1	278	308	0.9026	70.8	67.2	1.05	233	9690	Lung	Spleen	3.12	84.73	TRUE	TRUE
CHS.858.2	CHS.858.1	SELENON	556	590	0.9424	81.9	76.9	1.07	9629	9231	Thyroid	Muscle	59.85	9.01	TRUE	FALSE
CHS.8600.5	CHS.8600.7	CSTPP1	274	331	0.8278	86.1	79.5	1.08	9714	9701	Pituitary	Brain	19.27	27.49	TRUE	TRUE
CHS.8627.10	CHS.8627.5	KBTBD4	541	534	1.0131	84.1	79.4	1.06	1314	9653	Pituitary	Bladder	1.90	11.02	TRUE	TRUE

CHS.8627.13	CHS.8627.5	KBTBD4	518	534	0.97	85.1	79.4	1.07	5631	9653	Pituitary	Bladder	1.85	11.02	TRUE	TRUE
CHS.8627.7	CHS.8627.5	KBTBD4	543	534	1.0169	84.2	79.4	1.06	4	9653	Brain	Bladder	2.12	11.02	FALSE	TRUE
CHS.8634.16	CHS.8634.3	AGBL2	819	902	0.908	75.8	71.2	1.06	2	253	Thyroid	Testis	1.88	2.60	FALSE	FALSE
CHS.8634.21	CHS.8634.3	AGBL2	826	902	0.9157	75	71.2	1.05	1	253	Lung	Testis	1.10	2.60	FALSE	FALSE
CHS.8634.8	CHS.8634.3	AGBL2	828	902	0.918	75	71.2	1.05	27	253	Cervix_ Uteri	Testis	2.41	2.60	FALSE	FALSE
CHS.8857.5	CHS.8857.1	ZP1	345	638	0.5408	75.1	68.4	1.10	294	81	Testis	Skin	11.32	1.92	FALSE	FALSE
CHS.888.31	CHS.888.53	UBXN11	477	520	0.9173	72.9	68.9	1.06	145	6550	Pituitary	Testis	23.86	75.51	FALSE	FALSE
CHS.888.32	CHS.888.53	UBXN11	469	520	0.9019	72.8	68.9	1.06	257	6550	Thyroid	Testis	10.47	75.51	FALSE	FALSE
CHS.8896.2	CHS.8896.18	SYT7	403	686	0.5875	77.7	59.2	1.31	4437	1618	Pituitary	Brain	18.39	19.34	TRUE	TRUE
CHS.8896.3	CHS.8896.18	SYT7	478	686	0.6968	71.5	59.2	1.21	2120	1618	Brain	Brain	10.35	19.34	TRUE	TRUE
CHS.8896.4	CHS.8896.18	SYT7	447	686	0.6516	72.7	59.2	1.23	2620	1618	Prostate	Brain	10.60	19.34	TRUE	TRUE
CHS.9039.2	CHS.9039.1	VEGFB	188	207	0.9082	81.7	69.5	1.18	9776	9776	Adipose _Tissue	Adipose _Tissue	79.04	146.33	TRUE	TRUE
CHS.9156.6	CHS.9156.5	KAT5	494	546	0.9048	76.1	71	1.07	9780	9760	Bladder	Fallopia n_Tube	11.70	10.96	TRUE	TRUE
CHS.9219.4	CHS.9219.3	PELI3	445	469	0.9488	84.8	80.4	1.05	8935	9183	Brain	Bladder	16.83	7.68	FALSE	FALSE
CHS.9412.12	CHS.9412.16	XNDC1N	207	234	0.8846	82.7	75.4	1.10	251	190	Vagina	Prostate	3.65	2.60	FALSE	FALSE
CHS.9412.20	CHS.9412.16	XNDC1N	134	234	0.5726	85.5	75.4	1.13	327	190	Lung	Prostate	1.56	2.60	FALSE	FALSE
CHS.9412.5	CHS.9412.16	XNDC1N	179	234	0.765	85.3	75.4	1.13	1564	190	Cervix_ Uteri	Prostate	3.23	2.60	FALSE	FALSE
CHS.9471.5	CHS.9471.6	MRPL48	194	212	0.9151	78.3	73.9	1.06	1993	9756	Cervix_ Uteri	Bone_M arrow	0.83	27.28	TRUE	TRUE
CHS.9526.14	CHS.9526.2	DGAT2	358	388	0.9227	90.7	85.7	1.06	429	8025	Skin	Adipose _Tissue	21.43	276.06	TRUE	TRUE
CHS.9526.5	CHS.9526.2	DGAT2	360	388	0.9278	90.3	85.7	1.05	43	8025	Breast	Adipose _Tissue	8.51	276.06	TRUE	TRUE
CHS.9636.13	CHS.9636.54	DLG2	819	975	0.84	71.8	67.2	1.07	1167	241	Skin	Uterus	14.99	5.89	FALSE	FALSE
CHS.9642.1	CHS.9642.4	TMEM1 26B	210	230	0.913	85.8	77.3	1.11	7085	9737	Bone_M arrow	Bone_M arrow	3.27	28.88	TRUE	TRUE
CHS.9642.3	CHS.9642.4	TMEM1 26B	200	230	0.8696	87.3	77.3	1.13	9747	9737	Bone_M arrow	Bone_M arrow	44.33	28.88	TRUE	TRUE
CHS.9816.4	CHS.9816.6	TRPC6	845	931	0.9076	77.9	73.7	1.06	463	3683	Lung	Lung	4.48	16.84	TRUE	TRUE
CHS.9827.4	CHS.9827.3	BIRC2	569	618	0.9207	76.3	71.3	1.07	2857	9762	Testis	Nerve	0.90	38.06	FALSE	FALSE
CHS.9866.4	CHS.9866.5	CASP5	376	434	0.8664	80.3	74.3	1.08	752	644	Prostate	Small_In testine	8.70	10.99	FALSE	FALSE
CHS.9867.8	CHS.9867.11	CASP1	383	404	0.948	83.8	79.6	1.05	7973	9008	Spleen	Spleen	22.25	50.58	TRUE	FALSE
CHS.9971.11	CHS.9971.7	DIXDC1	472	683	0.6911	73.3	68.5	1.07	8189	6051	Bladder	Brain	44.26	9.87	TRUE	TRUE

Table 1: A filtered set of CHESS transcripts compared to MANE according to the criteria detailed in the “Filtering MANE comparisons” section of the Methods. Uses the same column names as Supplementary File 2.

We selected isoforms that, when compared to the MANE isoform for the same gene, scored at least 5% higher as measured by pLDDT and were at least 90% as long. Additionally, to capture cases where the MANE transcript might be missing functional sequence elements, we selected alternate isoforms that were at least 5% longer than

the MANE isoform, that had equal or higher pLDDT scores, and that were assembled in an equal or higher number of GTEx samples. Finally, to capture cases where an alternate isoform might be functional despite being substantially shorter than the MANE protein, we selected isoforms at least 50% as long as the MANE protein where the alternate isoform scored at least 5% higher and was assembled in an equal or higher number of GTEx samples. After applying these filters, we observed that in some cases, a processed pseudogene [50] contained within an intron outscored the associated primary transcript. To eliminate such cases, we used GFFcompare [51] to ensure that isoforms overlapped their MANE transcript's coding sequence. When multiple alternate isoforms contained the same coding sequence, and thus received the same pLDDT score, we selected the isoform assembled in the highest number of GTEx samples.

Annotation sources:

Transcript annotations for MANE, CHESS, RefSeq, and GENCODE were retrieved from the following sources. The MANE v1.0 database was downloaded from NCBI at https://www.ncbi.nlm.nih.gov/refseq/MANE. Annotations from the CHESS v3.0 database were retrieved from http://ccb.jhu.edu/chess. Additional CHESS annotations came from an unpublished set of transcript assemblies created as part of the process of building CHESS v3.0; these were assembled from approximately 10,000 GTEx RNA-seq experiments across 31 tissues using StringTie2 [40]. Transcripts from this set were given a CHESS ID starting in “hypothetical” if the locus was missing from CHESS v3.0, or else given an ID ending in “altN” if the locus was present in CHESS v3.0 but the exact isoform was not. Note that many of these, particularly those with a poor protein folding

score, were not retained in the final CHESS v3.0 database. RefSeq annotations (releases 109 and 110) were downloaded from https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml. GENCODE (v38, v39, v40) annotations were collected from https://www.gencodegenes.org/human/.

Visualization and atomic alignment:

All visualizations and three-dimensional protein structure atomic alignments were performed in PyMol [41] version 2.5.2 using non-orthoscopic view, white background, and ray trace 1200,1200. Root-mean-square deviations were calculated without excluding any outliers. Ramachandran plots were created using PyRAMA version 2.0.2 with Richardson [24] standard psi and phi values for all amino acids excluding glycine and proline. Intron-exon structure plots were produced with MISO [52] (commit b714021) and TieBrush [53] (commit e986d64).

RNA-sequencing quantification of the human ASMT gene:

RNA-sequencing data were downloaded from NCBI for run SRR5756467 from BioSample SAMN07278516, a pineal gland from a patient who died at midnight. A detailed summary of the experimental protocol used to generate these data can be found in NCBI BioProject PRJNA391921. Isoform-level quantification was performed using Salmon [54] version 1.8.0.

RNA-sequencing assembly of mouse TXNDC8:

RNA-sequencing data were downloaded for run SRR18337982 from BioSample SAMN26725167, a tissue sample from the testis of a control mouse. A detailed summary of the experimental protocol used to generate these data can be found in NCBI BioProject PRJNA816862. cDNA reads were aligned to the mm39 reference genome using HISAT2 [39] version 2.1.0 then assembled into transcripts using StringTie2 [40] version 2.2.1.

Intron conservation in human and mouse:

We assessed the conservation of GT-AG intron boundaries between CHESS human transcripts and transcripts from the GRCm38 mouse reference genome. Data for the human-mouse alignment was extracted from a 30-species alignment anchored on GRCh38 that was downloaded from the UCSC genome browser [55]. We used MafIO in BioPython [56] version 1.71 to check if all intron boundaries were conserved between mouse and human transcripts. In the supplementary tables, a value of “TRUE” in the “introns in mouse” column indicates that all boundaries were conserved, while “FALSE” means that at least one splice site (either a GT at a donor site or an AG at an acceptor site) was not conserved in the alignment.

Chapter 1 Acknowledgments:

The authors would like to thank all members of the Salzberg, Pertea, and Steinegger labs, as well as David J. Lipman for helpful feedback during project conceptualization,

Benjamin Langmead for publicly hosting bulk data files, and Do-Yoon Kim for creating the isoform.io logo.

Data and materials availability:

Gene identifiers for all predicted protein isoforms as well as pLDDT scores and evolutionary conservation data from mouse can be found in Supplementary File 1. Predicted scores and GTEx expression data for all isoforms overlapping a MANE locus can be found in Supplementary File 2. Data for the 940 alternate isoforms with evidence of relatively superior structure, and possibly superior function, can be found in Supplementary File 3. Additionally, all data can be downloaded from the project website, isoform.io.

Supplementary Captions:

All supplementary files can be found as part of Sommer et al. 2022 [57]

Supplementary File 1: All isoform summary.

Folding scores from ColabFold for each transcript from a preliminary new build of the CHESS database that contained a protein-coding sequence (CDS) that was under 1000aa in length. For transcripts already contained in the released CHESS v3.0 database, the identifier from that database is provided. If the transcript maps to a known gene locus X but is a novel isoform, it is shown with the identifier CHS.X.altY. If a transcript occurs at a novel locus X, the identifier is hypothetical.X.Y, where Y identifies the isoform number. Additional columns show the gene name, the RefSeq ID (release

110), the GENCODE ID (release 40), the pLDDT (folding) score, and a flag indicating whether all intron boundaries (for multi-exon genes) are conserved in the mouse genome.

Supplementary File 2: MANE comparison summary.

Folding scores and additional data for all CHESS transcripts that match genes in the MANE v1.0 dataset, limited to protein sequences under 1000aa in length. Transcripts must overlap the annotated CDS of the MANE transcript to be included. Columns include: CHESS_ID_isoform, the CHESS identifier of the alternate isoform transcript; CHESS_ID_MANE, the CHESS identifier of the MANE transcript at the same locus; gene, the gene name; aa_length_isoform, the amino-acid length of the alternate isoform's CDS; aa_length_MANE, the amino-acid length of the MANE transcript's CDS; length_ratio, the ratio of the alternate isoform length to the MANE isoform length; pLDDT_isoform, the predicted folding score of the alternate isoform; pLDDT_MANE, the predicted folding score of the MANE isoform; pLDDT_ratio, the ratio of the alternate isoform folding score to the MANE isoform folding score; GTEx_samples_observed_isoform, the total number of GTEx samples where the alternate isoform was observed at least once; GTEx_samples_observed_MANE, the total number of GTEx samples where the MANE isoform was observed at least once; GTEx_top_tissue_name_isoform, the name of the tissue in which the alternate isoform was observed in the highest number of samples; GTEx_top_tissue_name_MANE, the name of the tissue in which the MANE isoform was observed in the highest number of samples; GTEx_top_tissue_TPM_isoform, the average TPM of the alternate isoform in

the named tissue; GTEx_top_tissue_TPM_MANE, the observed TPM of the MANE isoform in the named tissue; introns_conserved_in_mouse_isoform, an indicator of whether introns are conserved between the alternate human isoform and any annotated isoform in the GRCm38 mouse reference genome; introns_conserved_in_mouse_MANE, an indicator of whether introns are conserved between the MANE human isoform and any annotated isoform in the GRCm38 mouse reference genome.

Video 1: CRYGN comparison.

A 3D animation comparing the predicted protein structure of the MANE isoform (CHS.52273.5, RefSeq NM_144727.3, GENCODE ENST00000337323.3) of gamma-N crystallin (CRYGN) vs. the predicted protein structure for the highest-scoring CRYGN alternate isoform (CHS.52273.9, GENCODE ENST00000644350.1).

Chapter 2: Balrog, a universal protein model for prokaryotic gene prediction

A version of chapter 2 previously appeared as:

M. J. Sommer, S. L. Salzberg, Balrog: A universal protein model for prokaryotic gene prediction. PLoS Computational Biology 17, e1008727 (2021).

2.1 Introduction

One of the most important steps after sequencing and assembling a microbial genome is the annotation of its protein-coding genes. Methods for finding protein-coding genes within a prokaryotic genome are highly sensitive, and thus have seen little change over the past decade. Widely used prokaryotic gene finders include various iterations of Glimmer [58,59], GeneMark [60,61], and Prodigal [62], all of which are based on Markov models and which utilize an array of biologically-inspired heuristics. Each of these previous methods requires a bootstrapping step to train its internal gene model on each new genome. This requirement also limits their ability to detect genes in fragmented sequences commonly seen in metagenomic samples [63].

The lack of recent advances in ab initio bacterial gene finding tool development is partly due to the perception that bacterial gene finding is a solved problem. Currently available tools achieve near 99% sensitivity for known genes (i.e., genes with a functional annotation), so there appears to be little room for improvement. However, all current software tools predict hundreds or thousands of “extra" genes per genome, i.e., genes

that do not match any gene with a known function and are usually given the name “hypothetical protein." Many of these hypothetical genes likely represent genuine protein coding sequences, but many others may be false positive predictions. It is difficult if not impossible to prove that a predicted open reading frame is not a gene; thus, these hypothetical proteins have remained in genome annotation databases for many years. However, systematically annotating false positives as genes may create problems for downstream analyses of genome function [64].

In line with evaluation metrics used by other gene finders, if a program can find nearly all true positive genes while predicting fewer genes overall, it is reasonable to assume this is primarily due to a reduction in false positive predictions [59,62]. Thus, we would prefer a method that makes fewer overall predictions while retaining very high sensitivity to known genes.

Currently available gene finders were developed in the late 1990's and 2000's, when relatively few prokaryotic genomes were available. Today, tens of thousands of diverse bacterial genomes from across the prokaryotic tree of life have been sequenced and annotated. We hypothesized that it should therefore be feasible to build a data-driven gene finder by training a machine learning model on a large, diverse collection of high-quality prokaryotic genomes. The program could then be applied, without any further re-training or adjustment, to find genes in any prokaryotic species. Balrog was developed with this strategy in mind. In the experiments below, we show that Balrog, when trained on all high-quality prokaryotic genomes available today, matches the sensitivity of

current state-of-the-art gene finders while reducing the total number of hypothetical gene predictions. By integrating protein-coding gene predictions from Balrog, standard prokaryotic annotation and analysis pipelines such as NCBI PGAP (Prokaryotic Genome Annotation Pipeline) [65], MGnify [66], or Prokka [67] may improve their genome annotation quality.

2.2 Results

Gene prediction sensitivity

We compared the performance of Balrog, Prodigal, and Glimmer3 by running each tool with default settings on a test set of 30 bacteria and 5 archaea that were not included in the Balrog training set. Following the conventions established in multiple previous studies, we considered a protein-coding gene to be known if it was annotated with a name not including “hypothetical" or “putative." In standard annotation pipelines, proteins are labeled hypothetical if they have no significant match to known protein sequences and are not otherwise covered by a standard naming rule [68]. For most bacterial genomes, more than two-thirds of their annotated genes fall into the “known" category, with the rest being hypothetical. The hypothetical genes include a mixture of true genes and false positive predictions. In our experiments, we measured the total number of genes predicted in each genome and calculated the sensitivity of each program to known, non-hypothetical genes. Predictions were considered correct if the stop codon was correctly predicted, i.e., if the 3' position of the gene was correct. Results for this gene finder comparison can be found in Table 2.

Genome			Balrog			Prodigal			Glimmer3		
	GC	genes	3′ matches		extra	3′ matches		extra	3′ matches		extra
Bacteria	%	#	#	%	#	#	%	#	#	%	#
T. narugense	30	1570	**1559**	99.3	**271**	1557	99.2	302	**1559**	99.3	367
C. fetus	31	1486	**1476**	99.3	**216**	1475	99.3	248	1473	99.1	279
T. wiegelii	33	2359	2265	96.0	**505**	2255	95.6	557	**2267**	96.1	715
Nat. thermophilus	34	2419	2397	99.1	**479**	2401	99.3	554	**2403**	99.3	648
D. thermolithotrophum	34	1360	**1336**	98.2	**197**	**1336**	98.2	220	1332	97.9	257
D. thermophilum	37	1630	1607	98.6	**250**	**1609**	98.7	281	**1609**	98.7	333
P. UFO1	38	3873	**3834**	99.0	**725**	3829	98.9	970	3831	98.9	1134
T. takaii	40	1496	1484	99.2	**322**	**1486**	99.3	373	1485	99.3	422
K. pacifica	41	1608	**1597**	99.3	**405**	1596	99.3	441	1594	99.1	543
B. bacteriovorus	42	1897	1883	99.3	**840**	**1887**	99.5	921	1884	99.3	1027
P. HL-130-GSB	45	1882	1804	95.9	**515**	1809	96.1	604	**1810**	96.2	783
C. thermautotrophica	46	2137	2107	98.6	**508**	**2116**	99.0	595	2114	98.9	696
A. aeolicus	46	885	**884**	99.9	**784**	883	99.8	826	879	99.3	840
M. thermoacetica	49	2299	2227	96.9	**679**	2233	97.1	808	**2238**	97.3	1134
Nov. thermophilus	49	2850	2769	97.2	**789**	2754	96.6	929	**2771**	97.2	1103
T. oceani	49	1998	1941	97.1	**305**	1932	96.7	375	**1943**	97.2	533
D. indicum	50	2178	2152	98.8	**461**	**2154**	98.9	492	2134	98.0	679
L. boryana	50	4031	3947	97.9	**1588**	**3956**	98.1	1868	3953	98.1	2423
D. multivorans	51	3128	3061	97.9	**667**	3064	98.0	796	**3065**	98.0	1585
E. coli K-12 MG1655	52	3529	**3451**	97.8	914	3408	96.6	**911**	3368	95.4	1110
D. acetoxidans	52	2322	**2273**	97.9	**554**	2268	97.7	698	2268	97.7	1165
C. parvum	54	1780	**1753**	98.5	**301**	1752	98.4	348	1746	98.1	489
T. ammonificans	56	1382	1373	99.3	**306**	**1377**	99.6	354	1373	99.3	362
A. acidocaldarius	58	2499	2393	95.8	**617**	**2397**	95.9	724	**2397**	95.9	908
R. radiotolerans	60	2196	2155	98.1	**563**	**2166**	98.6	608	2160	98.4	742
D. desulfuricans	62	2889	2849	98.6	**578**	2853	98.8	619	**2854**	98.8	858
S. thermophilum	63	2612	2564	98.2	**652**	**2567**	98.3	730	2562	98.1	847
V. incomptus	65	2498	2451	98.1	**1131**	**2465**	98.7	1176	2447	98.0	1540
C. bipolaricaulis	65	1022	997	97.6	**237**	**1008**	98.6	260	1000	97.8	286
S. amylolyticus	73	4880	4778	97.9	**3631**	**4821**	98.8	3887	4728	96.9	4789
Averages:	49	2289	2248	98.2	**664**	**2250**	98.3	747	2245	98.1	949
Archaea											
M. ruminantium	36	1710	1678	98.1	**455**	1682	98.4	517	**1687**	98.7	570
A. GW2011 AR10	39	621	618	99.5	**607**	**621**	100.0	720	**621**	100.0	778
M. sp. WWM596	46	2757	2567	93.1	**840**	2545	92.3	1123	**2581**	93.6	1999
M. labreanum	50	1390	1372	98.7	**379**	1370	98.6	446	**1376**	99.0	581
H. lacusprofundi	61	2047	2001	97.8	**613**	**2017**	98.5	731	2015	98.4	884
Averages:	46	1705	1661	97.4	**565**	1663	97.6	691	**1670**	97.9	949

Table 2: Non-hypothetical gene prediction comparison. “genes” refers to all protein-coding genes in the NCBI annotation where the description does not contain “hypothetical” or “putative.” Genes with descriptions containing “hypoth” or “etical” are also excluded to catch the most common misspellings of hypothetical. “3’ matches” counts the number of genes with stop sites exactly matching between the annotation and prediction on the same strand. “extra” counts the number of genes predicted by each program that do not share strand and stop site with an annotated non-hypothetical gene. The lowest number of extra genes and the highest number of 3’ matches are bolded for each organism.

All three tools achieved similar sensitivity on the bacterial genomes in the test set. On average, Balrog found 2 non-hypothetical genes fewer than Prodigal (2,248 vs. 2,250) and 3 genes more than Glimmer3 (2,248 vs. 2,245). This represents a difference of less than 0.1% in sensitivity. Balrog predicted the fewest genes overall, reducing the number of “extra" gene predictions by 11% vs. Prodigal (664 vs. 747) and 30% vs Glimmer3 (664 vs. 949).

Balrog predicted more genes than Prodigal for only one bacterial genome, E. coli K-12 MG1655 (the standard laboratory strain). On that genome, Balrog predicted 3 more extra genes than Prodigal, but at the same time it found 43 more true annotated genes. It is worth noting here that all organisms in the Escherichia and Shigella genuses were excluded from the Balrog training data set.

On the five genomes in the archaea test set, we observed more pronounced differences in the number of extra gene predictions. Glimmer3 found the most known genes, averaging 1670, versus 1663 for Prodigal and 1661 for Balrog. However, Balrog predicted the fewest genes overall, 18% fewer extra genes than Prodigal and 40% fewer than Glimmer3.

Similar results were observed when the gene model was trained on a set excluding organisms sharing a family, rather than a genus, with any organism in the test set. On average, the gene model achieved sensitivity of 98.12% with family excluded vs. 98.15% with genus excluded (2247 vs. 2248 genes) in bacteria and 97.50% vs. 97.44%

(1662 vs. 1661) in archaea. The family-excluded model predicted on average 25 more extra genes than the genus-excluded model in bacteria (689 vs. 664) and 32 more in archaea (597 vs. 565).

2.3 Materials and Methods

Training and testing data

In selecting genomes on which to train our gene model, we aimed to cover as much microbial diversity as possible while limiting sequence redundancy. As a whole, currently available prokaryotic genomes are biased toward clinically relevant organisms. Many low-abundance environmental species may be absent from public databases, whereas organisms important to human disease may have full genomes for hundreds of closely related strains [69]. We cannot fully account for the missing diversity within available sequence databases (indeed, millions of bacterial species probably remain unsequenced), but to limit the impact of highly-overrepresented species, we randomly selected only one genome for each bacterial and archaeal species within the Genome Taxonomy Database (GTDB, https://gtdb.ecogenomic.org) for the training set [70]. Only high-quality genomes were selected, defined by GTDB as over 90% complete with less than 5% contamination. Because high-quality protein annotations were also necessary, we required selected genomes to be available in RefSeq or GenBank with the tag “Complete Genome" and without the tag “Anomalous assembly."

From this set of high-quality complete genomes with gene annotations, 29 bacterial and five archaeal species were randomly selected to serve as a test set. Escherichia coli was also put in the test set because it is often used as a benchmark organism to compare gene finders. All genomes sharing a GTDB genus with any species in the test set were excluded from the training set. Full organism names and accession numbers for the testing data are available in S1 Appendix, while data for all organisms used to train the gene model are available in S2 Appendix. Though many gene sequences likely overlap between training and test data, we feel this test set should allow a reasonably conservative estimate of generalization error when predicting genes on a newly sequenced prokaryotic genome, which likely shares many gene sequences with previously seen genomes. Overall, this genome selection process yielded 3290 genomes in the training set and 36 in the test set. Additionally, a separate training set was constructed excluding all organisms sharing a family with any organism in the test set. This yielded 3085 genomes in the training set while the test set remained the same.

From all genomes, we extracted amino-acid sequences from annotated non-hypothetical genes. All genes with a description containing “hypothetical" or “putative" were removed from analysis, as many of these are not true genes but instead are the predictions of other gene finding programs. Additionally, genes with descriptions containing “hypoth" or “etical" were excluded in an effort to catch the most common misspellings of hypothetical. All non-hypothetical gene sequences were translated in all five alternative reading frames, and from these translations we extracted open reading

frames (ORFs) longer than 100 amino acids to use as training examples of non-protein sequence.

We extracted amino-acid shingles (overlapping subsequences) in the 3' to 5' direction of length 100 and overlapping by 50 from all protein and non-protein sequences. These were used as positive and negative gene examples, respectively. In total, ~27 gigabases (9 billion amino acids) of translated gene and non-gene sequence was generated to train the gene model.

Training the gene model

A temporal convolutional network (TCN) was trained using the methods and open source Python framework of Bai et al. [71], slightly modified to enable binary classification of protein sequence. We use the state of the last node of the linear output layer as representative of the binary classifier, with a value close to 1 predicting a protein-coding gene sequence and 0 predicting an out-of-frame sequence. During training, binary cross-entropy loss was calculated on the state of this last node. Backpropagation from this loss minimizes gene prediction error based on the full context of our 100 amino-acid sequence shingles. This works because we set parameters such that the receptive field size of the network was sufficient to cover the whole length of a sequence shingle. Fig. 10 shows an example TCN with each parameter explained.

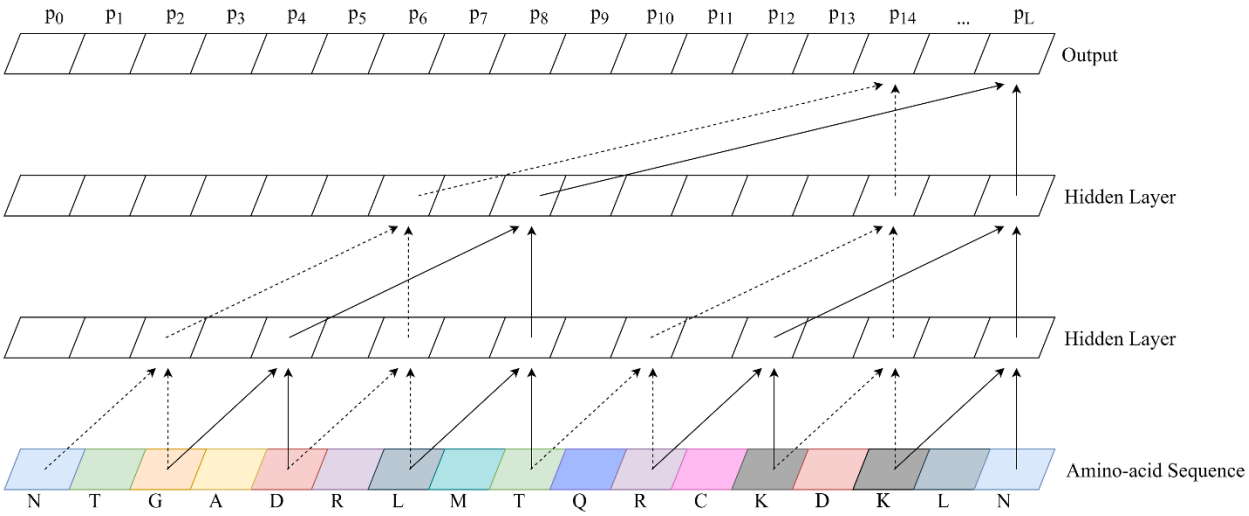

Figure 10: Example temporal convolutional network. A temporal convolutional network (TCN) with 2 hidden layers and a convolutional kernel size of 2. The number of connections exponentially increases as hidden layers are added, enabling a wide receptive field. Notice the output of a TCN is the same length as the input. Balrog's TCN used 8 hidden layers, a convolutional kernel size of 8, a dilation factor of 2, and 32*L hidden units per layer where L is the length of the amino-acid sequence.

During inference, we use the output from the pre-trained TCN to predict a single score for an ORF of any given length. To predict a single probability between 0 and 1, we combine all output scores from the TCN according to Equation 1 where L is the length of the ORF and p_i Sberro2019-qo is the predicted gene probability by the TCN model at position i. This represents taking a weighted average predicted gene probability, then applying the logistic sigmoid function to map back from (-infinity, infinity) to (0, 1). This method has the effect of more heavily weighing TCN predictions that are close to 0 or 1 [72,73]. We expect certain regions of a gene may contain recognizable protein sequence motifs, causing the TCN to predict a probability near 1. Other regions of a gene may contain little recognizable information, causing the TCN to predict near 0.5. By combining scores using this function, a single prediction near 1, caused by a

recognizable protein motif, can force the combined gene score closer to 1. Simply put, this equation allows us to improve gene scores based on the presence of conserved motifs in true proteins.

$$\text{Predicted gene probability} = \frac{1}{1+e^{-x}},\ x = \frac{1}{L}\sum_{i=1}^{L} ln\left(\frac{p_i}{1-p_i}\right)$$

Equation 1: log averaging predicted probabilities to retrieve single gene score.

Our gene model TCN used 8 hidden layers, 32*L hidden units per layer, a dilation factor of 2, and a convolutional kernel size of 8. Dropout was performed on 5% of nodes during training to mitigate overfitting. We used adaptive moment estimation with decoupled weight decay regularization (AdamW) [74] to minimize loss during initial training, while final loss minimization was performed by stochastic gradient descent with a learning rate of 10^(-4) and Nesterov momentum of 0.90 [75,76]. We performed all training on Google Colab servers with 32GB of RAM and a 16GB NVIDIA Tesla P100 GPU over the course of 48 hours.

Training the translation initiation site model

Though not the main focus of this work, a good start site model provides a boost in accuracy for a prokaryotic gene finder. In bacteria, the initiation of translation is usually marked by a ribosome binding site (RBS), which manifests as a conserved 5-6 bp sequence just upstream of the start codon of a protein-coding gene. Experimentally-validated start sites are not available for the vast majority of bacterial genes, so we

made the assumption (also used in previous methods [59]) that the annotated start sites of known genes would usually, but not always, be correct. Thus, to create a RBS model, we extracted 16 nucleotides upstream and downstream from all annotated non-hypothetical gene start sites in the training set genomes. For each start site, we also found the closest downstream start codon within the gene and extracted the same sized windows for use as examples of false start sites.

Similar to the gene model, we trained a TCN on the positive and negative examples of gene start sites. A slightly smaller model was used due to the reduced complexity and length of the start site sequence data. Our start site model used 5 layers with 25*L hidden units per layer and a convolutional kernel size of 6. The model was trained for 12 hours on the same Google Colab server type as the gene model.

Gene finding

A powerful gene sequence model is necessary for finding genes, but additional features such as open reading frame (ORF) length can also be taken into account. In particular, longer ORFs are more likely to be protein-coding genes, by the simple argument that a long stretch of DNA without stop codons is less likely, in random DNA sequence, than a short stretch. Balrog begins by identifying and translating all ORFs longer than a user-specified minimum. Its task is to determine for each of these ORFs whether it represents a protein-coding gene.

We also developed an optional k-mer-based filter, using amino-acid sequences of length 10, which runs before the gene model to positively identify genes. This filtering procedure simply identifies all amino-acid 10-mers found in annotated non-hypothetical genes from the training data set and flags any ORF containing at least two of these 10-mers as a true protein. This initial step finds many common prokaryotic genes with a very high specificity and near-zero false positive rate.

Next, ORFs are scored by the pre-trained temporal convolutional network in the 3' to 5' direction. The region surrounding each potential start site of each ORF is then scored by the start site model. A directed acyclic graph is constructed for each contig, with nodes representing all possible ORFs. Edges are added between compatible ORFs overlapping by less than a user-specified minimum. To avoid creating a graph with O(n^2) edges, we only connect a constant number of nodes to each node. Because prokaryotes are gene dense, we do not expect any large region with a significant number of non-gene ORFs. Therefore, we can keep the number of edges to O(n*C) where C=50 was empirically found to be sufficient for all tested genomes. Edge weights are calculated by a linear combination of the gene model score of the ORF, the gene start site model of the potential start site, a bonus for ATG vs. GTG vs. TTG start codon usage, and penalties for overlap depending on the 3'/5' orientation of the overlap.

The global maximum score of the directed acyclic graph is computed by finding the longest weighted path through the graph as shown in Fig. 11. Because we are searching for the maximum score and some ORFs can receive negative scores,

Dijkstra's algorithm does not work in this context [77]. Instead, we take advantage of the fact that our genome is implicitly topologically sorted to find the longest weighted path in two steps. First, we sweep forward along the genome, keeping track of the maximum attainable score at each node as well as its predecessor node. Then, we simply backtrack along the predecessors from the global maximum attainable score to find the longest weighted path. This is similar to finding the “critical path" in a task scheduling problem [78]. In practice, ORFs must only be connected locally to a relatively small set of other ORFs because no real prokaryotic genome should have a very large gap between genes. This makes the complexity of finding the maximum score scale linearly with the size of the genome. The highest scoring path through the graph represents the best predicted set of all compatible genes in the genome and is converted into an annotation file for the user. To benchmark gene finding performance, Glimmer3 and Prodigal were run with default settings and allowed to train on each genome in the test set.

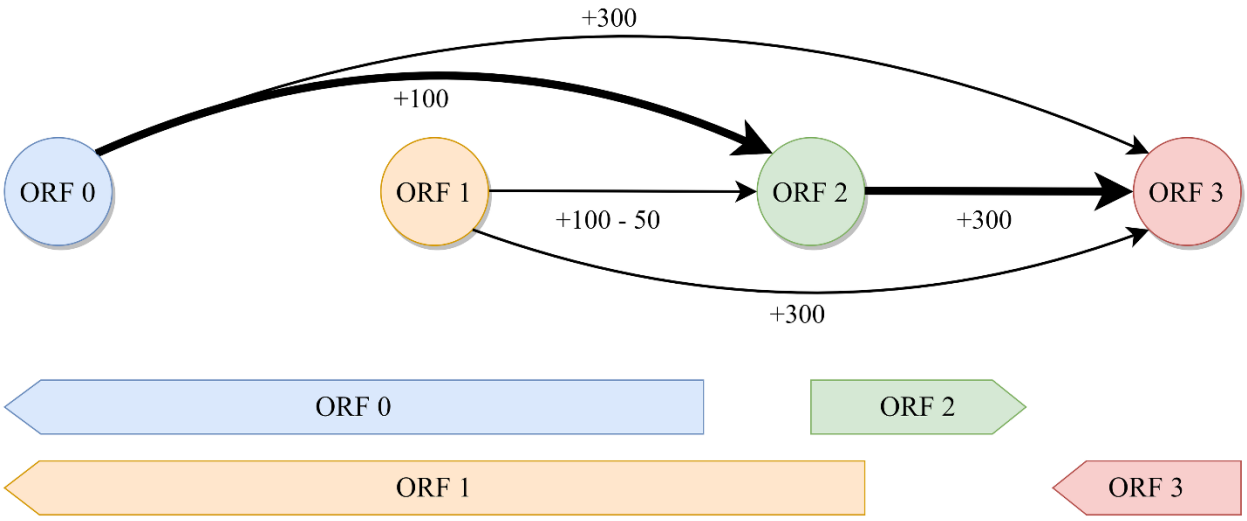

Figure 11: Example ORF connection graph. A directed acyclic graph with nodes representing open reading frames (ORFs) and edges representing possible connections. Each edge is weighted by the ORF score at the tip of the arrow minus any penalty for overlap. ORFs that overlap by too much are not connected. In this example, the maximum score is achieved by following the bolded path connecting 0-2-3. ORF 1 is not included because it is mutually exclusive with ORF 0 and results in a lower score due to overlap with ORF 2.

Parameter optimization

In the spirit of building a data-driven model, nearly all parameters were optimized with respect to the data rather than being hand-tuned. Ten genomes were randomly selected from the training data set to use for optimization of weights used in the scoring function for genome graph construction.

The score for each ORF node was calculated by a linear combination of features including the gene model score, start site model score, start site codon usage, and the length of the ORF. Additionally, final scores for edges between nodes are penalized by the length and direction of overlap, if any, between the connected ORFs. Depending on

the type of overlap, per-base penalties are multiplied by the length of the overlap and subtracted from the edge connection score. Different penalties are learned for divergent overlap (3' to 3'), convergent overlap (5' to 5'), and unidirectional overlap (3' to 5' or 5' to 3').

This scoring system was used to combine features so the linear weights can be learned with respect to the data to maximize gene finding sensitivity. Optimization of all weights with respect to gene sensitivity was accomplished using a tree-structured Parzen estimator [79] and a covariance matrix adaptation evolution strategy [80]. Because ORFs do not need to be re-scored by the TCN during parameter optimization, only the graph construction and longest path finding steps must be iterated to maximize gene sensitivity. All optimization was carried out using the Optuna framework [81] over the course of 9 hours on two 10 core Intel Xeon E5-2680 v2 processors at 2.8GHz.

Filtering with MMseqs2

Our gene model is tuned to maximize sensitivity to known genes without regard to the total number of predictions. In order to keep down the number of false positive predictions, users may optionally run a post-processing step with MMseqs2 [47]. In this step, we run all predictions against non-hypothetical protein coding gene sequence from a set of 177 diverse bacterial genomes. All reference genomes in this step do not share a genus with any of the test set organisms. Predictions are also run against the SWISS-PROT curated protein sequence database [82]. Any Balrog prediction that maps to a known gene with an E-value less than 0.001 is marked as a predicted gene. Finally, any

gene below a set cutoff ORF score is discarded unless it was found by the k-mer filter or MMseqs2. This process allows low-scoring predictions to be discarded as false positives while retaining many low-scoring genes that easily map to conserved known genes. All genomes used in this step can be found in S3 Appendix. Fig. 12 shows a flow chart with a broad overview of all steps performed by Balrog.

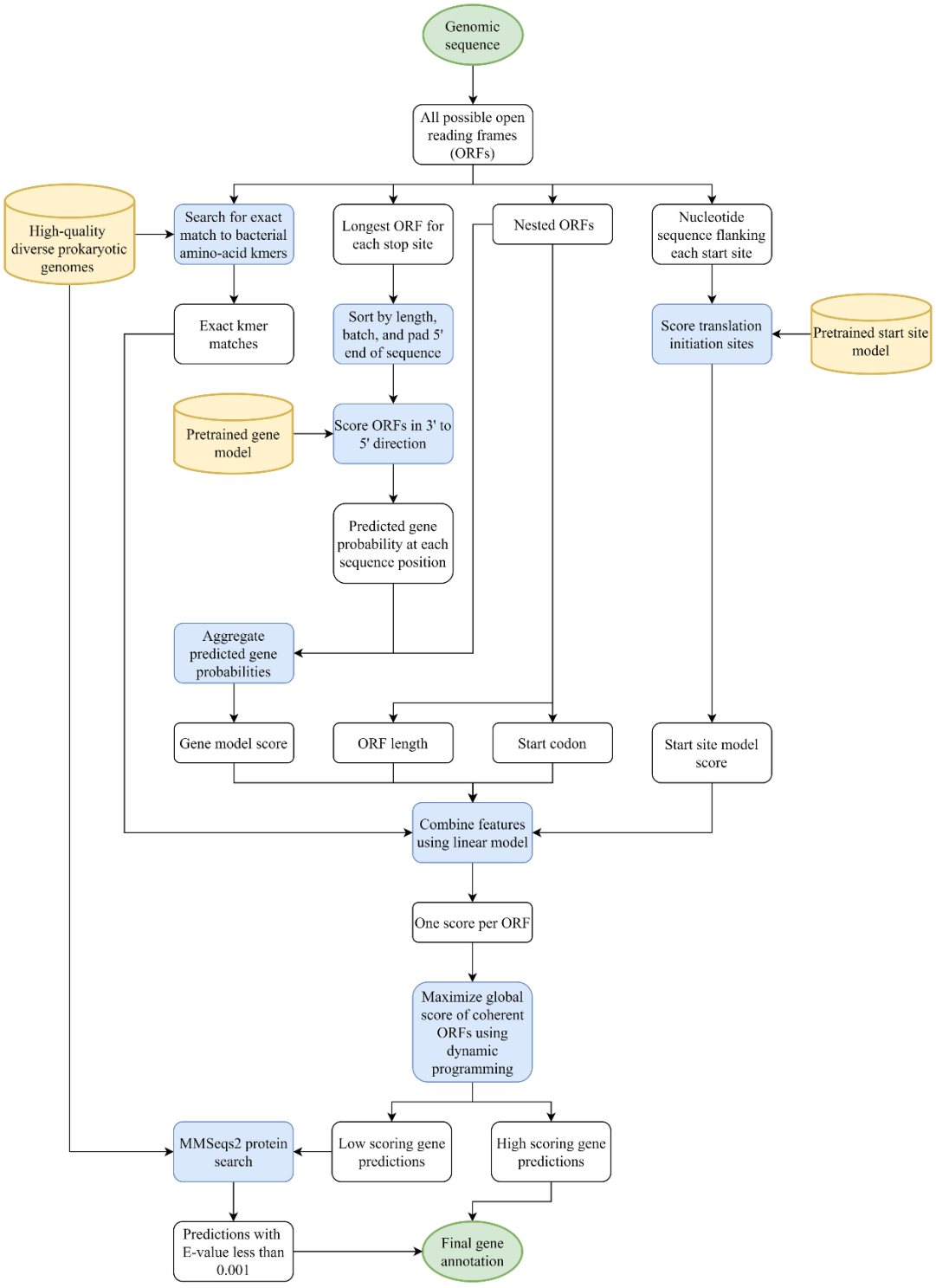

Figure 12: Balrog gene finding flow chart. A diagram showing all steps from genomic sequence in to gene predictions out. Green circles represent input and output data. White squares represent intermediate data. Blue squares represent processes. Yellow cylinders represent databases and pretrained models.

2.4 Discussion

Balrog demonstrates that a data-driven approach to gene finding with minimal hand-tuned heuristics can match or outperform current state-of-the-art gene finders. By training a single gene model on nearly all available high-quality prokaryotic gene data, Balrog matches the sensitivity of widely used gene finders while predicting fewer genes overall. Balrog also requires no retraining or fine-tuning on any new genome.

Balrog predicted consistently fewer genes than both Prodigal and Glimmer3 on both the bacterial and archaeal genome test sets. The sensitivity of all three gene finders was nearly identical and likely well within the range of noise in our sample on average, though Prodigal appears to achieve higher sensitivity than both Balrog and Glimmer3 on high-GC% genomes. A stronger bias against short ORFs, similar to Prodigal's penalty on ORFs shorter than 250bp, may help Balrog perform better in genomes with particularly high GC content. However, incorporating a bias against small genes may provide higher specificity at the cost of sensitivity to small genes. Heuristics used by current gene finders, including default minimum ORF lengths of 90 for Prodigal and 110 for Glimmer3, have led to a blind spot around functionally important small prokaryotic proteins [83]. Balrog's default minimum ORF length is 60 nucleotides. Further work on finding small genes without significantly increasing false positive predictions may help illuminate this underappreciated category of prokaryotic genome function.

Our test set deliberately represented a near-worst-case scenario for Balrog, where no organism from the same genus was used to train the model. On organisms closely related to those in the large and diverse training set, we expect Balrog may perform better as a result of overfitting. Overfitting of a gene model in this context is a complex issue. Simply memorizing and aligning to all known genes can be thought of as the ultimate overfit model, yet that strategy would likely prove effective at finding conserved bacterial genes. Finding prokaryotic genes is not a standard machine learning task where memorization inevitably leads to higher generalization error. Conserved amino-acid sequences in prokaryotic genes may represent functionally important protein motifs and memorization of short amino-acid sequences as indicators of protein coding sequence may prove useful in finding genes even in novel organisms. Still, we attempted to be as fair as possible to competing gene finders by removing all organisms with a shared genus. We felt this should provide a conservative estimate of the true generalization error of our model to relatively distant genomes.

An alternative approach to training a universal protein model could use protein clusters to capture diversity in protein sequences with less redundancy than our whole-genome approach. However, we wanted our final evaluation metric to be as fair as possible to all gene finders and reflective of a real-world situation where a newly sequenced prokaryote would likely contain many proteins from many different clusters.

Balrog in its current form is relatively slow. While tools like Prodigal and GeneMarkS-2 may analyze a genome in a matter of seconds, Balrog may take minutes per genome.

This is due to a wide range of factors including the complexity of the gene model and the optional gene filtering step with MMseqs2. Optimization of run time represents a possible future improvement for Balrog.

Balrog was designed primarily to find genes without much regard for identifying the exact location of their translation initiation site (TIS). TIS identification is a challenging problem with relatively little available ground-truth data. A reasonably accurate start site predictor helps to guide a gene finder, so Balrog does include a small TIS model, but accurate start site prediction was not a primary focus of this work. Further complicating the issue, nearly all available start site locations are based solely on predictions of previous gene finders. Demonstrating true improvement in start site prediction would require comparing Balrog to other gene finders on a large ground-truth data set which is simply not currently available. Incorporating TIS models used by Prodigal or GeneMark may enable improvement in start site identification in the future.

Supplementary Captions:

All supplementary files can be found as part of Sommer and Salzberg 2021 [84]

S1 Appendix. Gene model testing organism information. Full organism names and accession numbers of all genomes used in the gene finder comparison in Table 2.

S2 Appendix. Gene model training organism information. Full organism names and accession numbers of all genomes used to train the gene model.

S3 Appendix. MMseqs2 and k-mer filter organism information. Full organism names and accession numbers of all genomes used in the protein k-mer and MMseqs2 filtering steps.

Chapter 2 Acknowledgements:

We would like to thank Christopher Pockrandt for helping distribute the C++ version of Balrog, Jennifer Lu and Alaina Shumate for helping brainstorm cool program names, everyone on the Center for Computational Biology Slack channel for voting on said cool names, Martin Steinegger for helpful conversations and creating MMseqs2, @genexa_ch for providing via Twitter a small set of diverse GTDB genomes on which the k-mer filter and MMseqs2 are run, and everyone in the S. Salzberg and M. Pertea labs.

Chapter 3: cirkit, alignment-free circRNA detection

3.1 Introduction

Circular RNA (circRNA) is a class of RNA molecule expressed in a wide variety of eukaryotes [85] with predicted functional roles in human heart [86], brain [87], and cancer [88,89]. All circRNAs are generated by back-splicing whereby the spliceosome connects a downstream donor site for one intron to an upstream acceptor site for a previous intron, in the opposite direction of canonical linear RNA splicing [92]. Back-splicing creates a circular RNA molecule which includes a back-splice junction (BSJ), the defining sequence characteristic of circRNA. As nearly all sequence within circRNA will be identical to linear RNA from the same locus, sequencing-based methods for identifying circRNA rely on identification of the BSJ.

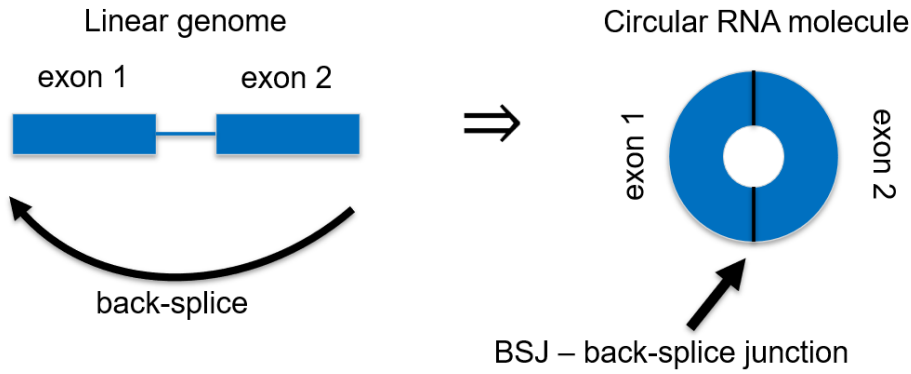

Figure 13: Creation of circRNA from back-splicing. circRNA is produced by a connection between a downstream donor and upstream acceptor, creating a back-splice junction (BSJ).

Current methods for circRNA detection from RNA sequencing (RNA-seq) have substantial room for improvement [93–95]. Specifically, current tools tend to predict a high number of circRNAs supported by small numbers of reads. In other words, when run on reasonably deep coverage RNA-seq data, the vast majority of circRNA predictions will be observed few times in the data. As of July 2023, circAtlas [96] annotated 768,986 unique circRNA molecules in human. To put this into perspective, human circRNA annotations alone amount to more than double the total number of human protein coding transcripts, pseudogenes, and non-coding RNAs combined in GENCODE v41 [97]. Whether genome annotation databases should include all observed genetic molecules, regardless of biological functionality, is up for debate, but irrationally large numbers of low-abundance circRNA molecules can stymie downstream analyses and overwhelm any computational attempt to study human circRNAs.

Given that many circRNAs are simply products of natural splicing errors [92], one might expect circRNA detection tools to suffer from poor precision on benchmark experimental data. This is not the case. Though most circRNAs may represent biological noise, circRNA annotation tool benchmarks have shown consistently high precision of circRNA prediction tools when measured experimentally [98]. When verified experimentally by qPCR, RNase R, or amplicon sequencing, circRNA prediction precision ranges from 95% to nearly 99%. This apparently high precision may be a consequence of the true existence of stochastically produced circRNA in biological samples, which can be enriched by experimental techniques and detected by current tools even at low expression levels. Still, benchmarking studies have shown the number of unique

circRNA predictions may vary by orders of magnitude between tools, and the agreement between sets of predictions is weak at best [95]. Furthermore, a high rate of low-abundance BSJ prediction can lead to sparse circRNA expression matrices in human data, violating assumptions of differential expression analysis tools such as DESeq2 and edgeR [93].

At least one circRNA, circRims2, is known to be conserved between mouse and human and in both species is 20 times more abundant than linear RNA from the same locus [90]. Recent work has shown circRims2 deficiency causes retinal degeneration [91], suggesting circRNA can play a key role in healthy biological function. As the number of circRNAs with known functional roles in human health and disease continues to increase, there is a need for methods capable of detecting functional circRNA molecules among stochastic background circRNA, akin to finding a whisper at the veritable rock concert of circRNA noise.

Here we demonstrate an alignment-free k-mer-based method, cirkit, for circRNA detection in RNA-seq data. Run in conjunction with widely used circRNA detection and quantification tools, we show improved specificity on experimentally generated benchmark data. When run on real-world RNA-seq data from post-mortem human brain samples, cirkit positively identifies circHomer1a, a known neuronally enriched circRNA [99]. Our results suggest filtering circRNA quantification results with cirkit may improve downstream differential expression analyses without removing true positive BSJs. This work represents a proof-of-concept for alignment-free circRNA detection, further work

on which may enable substantial computational improvements over existing alignment-based tools.

3.2 Results

Both cirkit and CIRI2 were run on paired-end RNA-seq data generated from untreated human lung fibroblast (HLF) cells (SRX13414572) and RNAseR treated HLF cells (SRX13414575) generated by Vromman et al. 2023 for the purpose of benchmarking circRNA detection tools. Because RNAseR degrades linear RNA molecules, we expect that predicted circRNAs will be enriched in the treated cells as compared to the untreated cells. When we treat a sample with RNAseR, an enzyme that degrades RNA from the end of the molecule, we expect the total amount of linear RNA to decrease in the sample. We do not, however, expect much impact on the amount of circRNA in the sample, as circles do not have the 3’ ends upon which the RNAseR acts. Thus, by using RNAseR to preferentially digest linear RNA in one sample while not treating a paired biological control sample, we can determine which circRNA predictions likely come from true circRNA molecules and which may be false-positive predictions. Based on these paired samples, cirkit showed a moderate increase in false-positive circRNA predictions (1.0% vs. 0.2%) as well as a 33% increase in positive predictions (17,598 vs 13,218). A proportional Venn-diagram comparing cirkit and CIRI2 predictions is shown in Fig. 14 with RNAseR predicted false positive predictions labeled in red. It may be worth noting that the intersection of cirkit and CIRI2 predictions contains zero RNAseR-determined false-positive predictions.

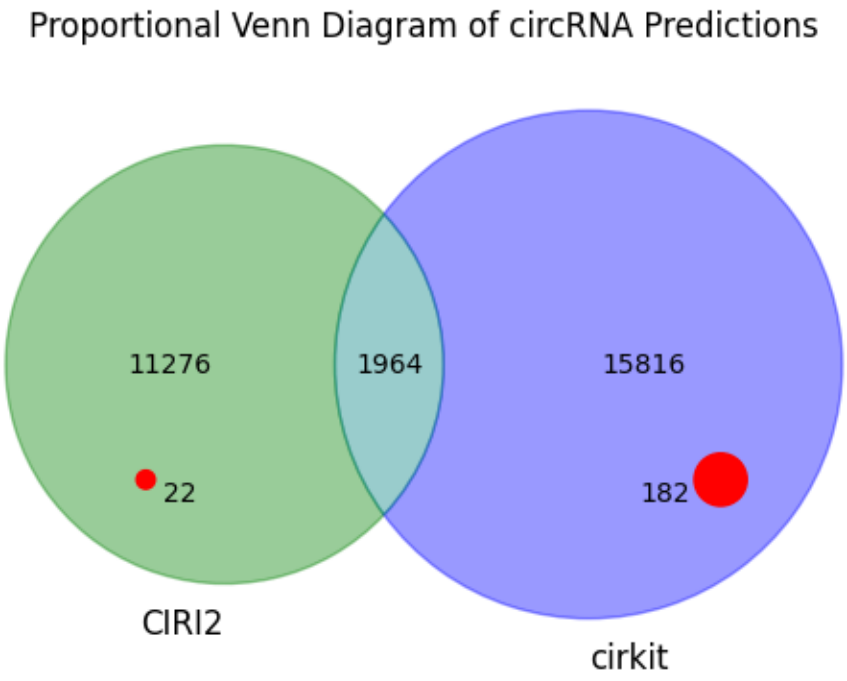

Figure 14: Proportional Venn diagram of circRNA predictions from Vromman et al. human lung fibroblast (HLF) cell line benchmarking data. Red circles show the RNAseR-determined false positive predictions, with 22 (0.2%) CIRI2 false positives and 182 (1.0%) cirkit false positives. The intersection of CIRI2 and cirkit contained zero RNAseR-determined false positive circRNA predictions.

As a preliminary demonstration of its potential real-world performance, cirkit was also run on RiboZero RNA-seq samples from the Lieber Institute for Brain Development. Biological samples were sourced from the dorsolateral prefrontal cortex (DLPFC) of a patient with schizophrenia (R2948) and a control patient (R2905). A known brain-specific circRNA, circHomer1a (BSJ T2T-CHM13v2.0 Chromosome 5: 79923195 - 79941203), was found in both samples. As would be expected based on prior work [99], circHomer1a was depleted two-fold in the sample from the schizophrenic patient (52 vs.

104 BSJ spanning read pairs). When run on the same samples, CIRI2 detected circHomer1a with a similar depletion percentage (75 vs 158 BSJ spanning read pairs).

3.3 Methods

Overview of cirkit

Relative to existing circRNA detection tools, which often rely on one or more alignment steps followed by postprocessing in Perl or Python, cirkit is relatively simple. We take advantage of the apparent inversion of paired-end short reads that occurs when the alignment spans a back-splice junction. In effect, when paired-end reads span a BSJ, their direction when aligned to the genome swaps from pointing towards one another to pointing away from one another. This inversion effect is illustrated in Fig. 15.

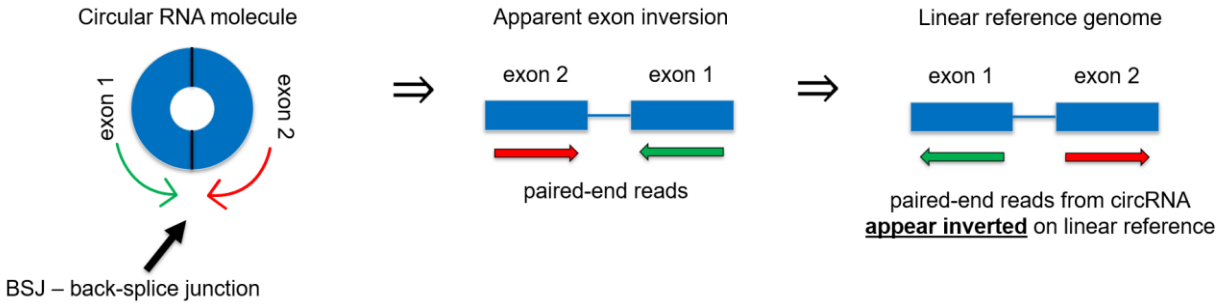

Figure 15: Illustration of apparent paired-end read inversion when spanning circRNA BSJ. This inversion remains in effect for individual k-mers at the ends of each read as long as the BSJ is at least k bases away from the end of the read. Additionally, no individual k-mer must span the BSJ, as the reads will invert even if the BSJ is contained in the paired-end insert.

Our analysis used the complete T2T human genome T2T-CHM13v2.0 [45]. As we care primarily about canonical exonic circRNA, we build the cirkit index from annotated

exons in the human genome. We create a hash table storing information on the gene ID, contig, strand, and genomic coordinate or each 25-mer in each exon of the genome. Two additional vectors are created for the gene ID and contig information such that the 64-bit integer can simply store bits encoding the position of the gene ID and contig, rather than storing the gene ID or contig as a string in the table. For simplicity in this proof-of-concept work, our analysis used the standard Python 3.8.5 dict() as our hash table implementation. More appropriate hash tables optimized for genomic information and k-mers may substantially improve memory and runtime performance in future work. A simple diagram demonstrating information retrieval for k-mers from paired-end reads is shown in Fig. 16.

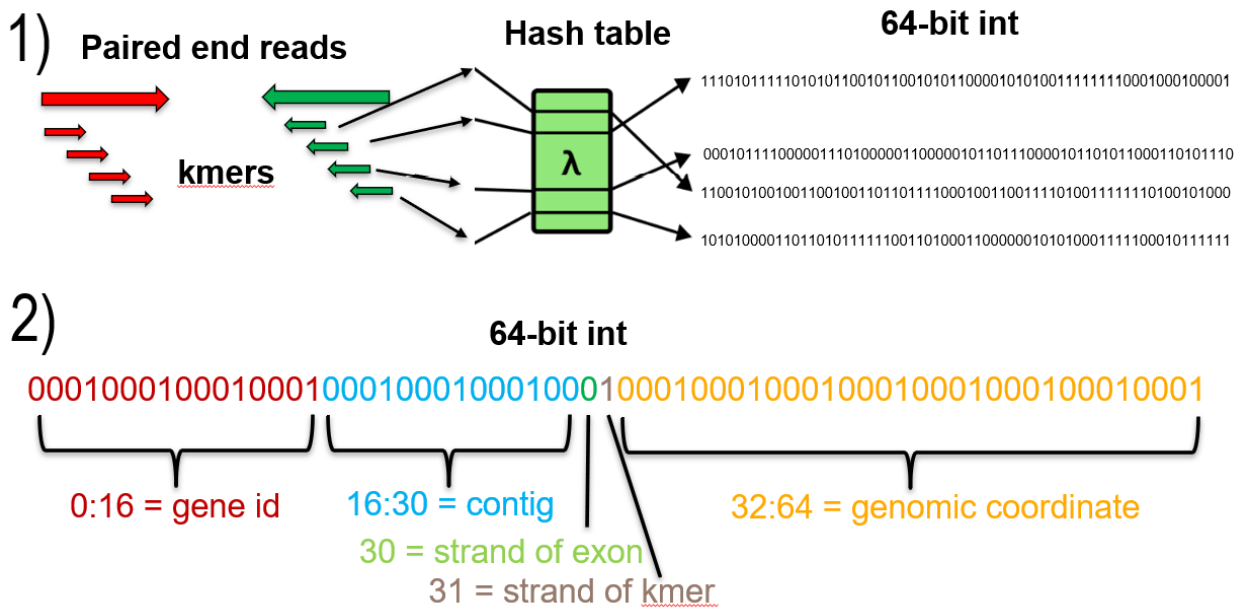

Figure 16: cirkit k-mer data structure. At runtine**,** cirkit works relatively simply by extracting k-mers from paired-end reads, using a hash table with the k-mer as the key and the 64-bit integer as the value. (illustration based on Vromman et al. 2023). The hash table has been constructed based on the T2T genome annotation and contains k-mers from all annotated exons and their associated genomic coordinates. We use the

information stored in the 64-bit integer to look for k-mer inversions in the paired-end reads.

circRNA detection benchmarking

Paired-end 150bp RNA-seq data for human lung fibroblast (HLF) cells (SRX13414572) and RNAseR treated HLF cells (SRX13414575) were originally generated by Vromman et al. 2023. cirkit was run on each sample using a single thread with default settings (k=25, m=30). CIRI2 was run on each sample with default settings. We do not know the true set of circRNAs present in these samples. Thus, it is not possible to confirm the true accuracy of cirkit from these data. Rather, we compare cirkit to existing tools and measure RNAseR-determined negative predictions. RNAseR preferentially degrades linear transcripts rather than circular transcripts, so enrichment of BSJs in the control data suggests such predictions are false positives. BSJ predictions were filtered with a minimum paired-read count of 5 for each tool. In the enrichment analysis, a fold-change of factor of 3 was required for calling enriched vs. depleted in the untreated vs. RNAseR samples. All enrichment analysis was performed in Python version 3.8.5.

circHomer1a analysis

Paired-end 100bp RiboZero RNA-seq data from the dorsolateral prefrontal cortex (DLPFC) of a patient with schizophrenia (R2948) and a control patient (R2905) were generated by the Lieber Institute for Brain Development (LIBD) and analyzed as part of the CHESS-BRAIN (Comprehensive Human Expressed Sequences in Brain) project. As

CHESS-BRAIN was not specific to circRNA, no RNAseR enrichment was performed on these data. cirkit was run with default settings (k=25, m=30) on both samples. CIRI2 was run with default settings on both samples. Reads spanning the back-splice associated with circHomer1a (BSJ T2T-CHM13v2.0 Chromosome 5: 79923195 - 79941203) were counted from the cirkit and CIRI2 output results.

3.4 Discussion

Accurate detection of circRNA is critical to solidify our understanding of the functional RNA landscape in human, yet extensive generation of circRNA by stochastic cellular processes complicates any computational analysis of these molecules. In this study, we present cirkit, an alignment-free k-mer-based method for circRNA detection in RNA-seq data which may be run in conjunction with widely used circRNA detection and quantification tools to improve specificity on experimentally generated benchmark data. cirkit positively identified circHomer1a, a known neuronally enriched circRNA, in RiboZero RNA-seq data from human brain samples. This represents neither a definitive analysis of circHomer1a in the schizophrenic brain nor a statistically significant result, as it is based on only two RNA-seq samples. Instead, our analysis demonstrates that an alignment-free circRNA detection algorithm can generate expected results with relatively low-quality non-RNAseR treated RiboZero RNA-seq data. This work represents a proof-of-concept for alignment-free circRNA detection, and further algorithmic and implementation improvements may enable substantial computational improvements over existing alignment-based tools.

cirkit does not attack the core problem of circRNA analysis, i.e. sorting out which circRNAs are products of noise versus those which may be biologically functional. Still, using multiple tools with orthogonal detection approaches may limit false positive predictions and simplify downstream analyses. Even in gold-standard experimental circRNA data, used here to benchmark cirkit, more than two thirds of all predicted circRNA molecules were not present in any existing circRNA atlas[101]. This is despite the fact that current circRNA databases already contain well over one million unique circRNAs previously observed in the human transcriptome. Moreover, only 55 out of 315,312 unique predicted circRNAs (0.02%) were agreed upon by all tools in the Vromman study. This suggests that increasingly deep sequencing experiments aimed at detecting human circRNA may be increasingly sensitive to intrinsic biological noise from which circRNA may arise.

Interestingly, most low-abundance circRNA predictions may not be false-positives in the traditional sense. Recent work elucidating the chemical structure of the spliceosomal E complex confirmed a single unified mechanism is responsible for both linear splicing and back-splicing, leading to the conclusion that “circRNA is a natural byproduct of spliceosome-mediated splicing in all eukaryotic species” [92]. Thus, low-abundance circRNA may be the result of true back-splicing events which occur naturally in all eukaryotes as a product of biological noise [9]. Eukaryotic cells are not perfect machines, and RNA molecules may be generated by inherently stochastic reactions with no useful or evolutionarily conserved function in the cell [102]. By creating a human annotation

based on all observed circRNA molecules rather than prioritizing a functional few, researchers risk sowing confusion in the nascent scientific field of human circRNA research [103].

Further work explicitly aimed at detecting functional or conserved circRNA may lead to more trustworthy and biologically relevant conclusions. As the inherently high rate of low-abundance circRNA prediction can lead to sparse expression matrices, violating assumptions of differential expression analysis tools such as DESeq2 and edgeR [93], novel methods specifically developed for differential transcript usage may prove useful in finding important circRNA in human tissues [104]. In a similar vein, analyzing conservation of circRNA is a deceptively difficult task in the context of circRNA. This is because all genetic sequence of exonic circRNA is inherently shared with linear RNA from the same locus. Off-the-shelf conservation scores such as those from phyloP must be interpreted with caution [90]. Thus, future development of circRNA-specific sequence conservation scoring tools may be useful in finding functional circRNA molecules.

In its current form, cirkit is designed to be run alongside and in addition to existing circRNA detection and quantification tools. We have demonstrated the ability of cirkit to polish circRNA annotations, potentially reducing false positive predictions, but further work would be needed to allow cirkit to be run as a standalone circRNA annotation tool without reducing specificity. Additionally, quantification of circRNA would require careful future development due to substantial shared sequence with each circRNA's corresponding linear transcripts. Still, the algorithmic simplicity of cirkit naturally lends

itself to future computational optimization and efficiency enhancements. The single-table lookup of cirkit should enable embarrassingly parallel analysis. Using an optimized hash table rather than an off-the-shelf python dictionary should also improve computational performance substantially.

Chapter 3 Acknowledgements:

I would like to thank Hayden Ji for consistently running CIRI2 as well as Kuan-Hao Chao for helpful discussions and calculating the theoretical maximum number of unique intra-locus exon-exon BSJs in the human genome, 877,973.

Chapter 4: Conclusion

Computational tools for genome annotation play a key role in expanding the frontiers of genomics, helping scientists analyze how organisms function at a molecular level. Here, we developed three novel methods for human transcriptome analysis, prokaryotic gene prediction, and circular RNA annotation. Bound by the common thread of improving genome annotation, this work spanned three distinct areas within computational biology, each with unique challenges and solutions.

The CHESS Human Protein Structure database, a publicly available protein structure resource, uses three-dimensional protein structure predictions to identify functional human gene isoforms. In our analysis, we evaluated over 230,000 isoforms of human protein-coding genes assembled from thousands of RNA sequencing experiments across the whole human body. We identified hundreds of isoforms with potentially superior function compared to canonical isoforms, thus establishing protein structure prediction as a powerful tool for transcriptome analysis.

Balrog, a universal protein model for prokaryotic gene prediction, employs a temporal convolutional network to score amino acid sequences. This work demonstrated that a data-driven approach to gene finding can match or outperform current state-of-the-art gene finders. By training a single gene model on a diverse set of high-quality prokaryotic gene data, Balrog matched the sensitivity of widely used gene finders while predicting fewer genes overall.

Finally, we presented cirkit, an alignment-free method for annotating circRNA in humans. Further work will be needed to match the circRNA detection quality of existing alignment-based tools, but considering the simplicity of the algorithm presented, cirkit performed remarkably well even on noisy real-world data. The ideas used to build cirkit provide a foundation upon which future tools may be built to enable fast, sensitive, and specific circRNA detection.

Together, these methods have expanded our understanding of the biological world through the development and application of novel genomic tools. We hope scientists building upon this work will continue to unravel the intricacies of the genome by bridging the gap between raw data and meaningful biological insights, opening new avenues for research by deciphering the intricate genetic mechanisms governing life.

References

1. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
2. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
3. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
4. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
5. Salzberg, S. L. Open questions: How many genes do we have? *BMC Biol.* **16**, 94 (2018).
6. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
7. Tung, K.-F., Pan, C.-Y., Chen, C.-H. & Lin, W.-C. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci. Rep.* **10**, 16245 (2020).
8. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
9. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **20**, 536–548 (2019).

10. Ponting, C. P. & Haerty, W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. *Annu. Rev. Genomics Hum. Genet.* (2022) doi:10.1146/annurev-genom-112921-123710.
11. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6**, 2 (2015).
12. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
13. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
14. Deiana, A., Forcelloni, S., Porrello, A. & Giansanti, A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One* **14**, e0217889 (2019).
15. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 1–6 (2022).
16. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* 1–4 (2022).
17. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
18. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
19. Melke, J. *et al.* Abnormal melatonin synthesis in autism spectrum disorders. *Mol. Psychiatry* **13**, 90–98 (2008).

20. Rossignol, D. A. & Frye, R. E. Melatonin in autism spectrum disorders: a systematic review and meta-analysis. *Dev. Med. Child Neurol.* **53**, 783–792 (2011).
21. Botros, H. G. *et al.* Crystal structure and functional mapping of human ASMT, the last enzyme of the melatonin synthesis pathway. *J. Pineal Res.* **54**, 46–57 (2013).
22. Andley, U. P. Crystallins in the eye: Function and pathology. *Prog. Retin. Eye Res.* **26**, 78–98 (2007).
23. Wistow, G. *et al.* gammaN-crystallin and the evolution of the betagamma-crystallin superfamily in vertebrates. *FEBS J.* **272**, 2276–2291 (2005).
24. Lovell, S. C. *et al.* Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437–450 (2003).
25. Modi, T., Huihui, J., Ghosh, K. & Ozkan, S. B. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20170184 (2018).
26. Jiménez, A. *et al.* Spermatocyte/Spermatid-specific Thioredoxin-3, a Novel Golgi Apparatus-associated Thioredoxin, Is a Specific Marker of Aberrant Spermatogenesis*. *J. Biol. Chem.* **279**, 34971–34982 (2004).
27. Carrier, Y. *et al.* Inter-regulation of Th17 cytokines and the IL-36 cytokines in vitro and in vivo: implications in psoriasis pathogenesis. *J. Invest. Dermatol.* **131**, 2428–2437 (2011).
28. Uppala, R. *et al.* “Autoinflammatory psoriasis”-genetics and biology of pustular psoriasis. *Cell. Mol. Immunol.* **18**, 307–317 (2021).
29. Tashima, Y. *et al.* PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. *Mol. Biol. Cell* **17**, 1410–1420 (2006).

30. Englund, P. T. The structure and biosynthesis of glycosyl phosphatidylinositol protein anchors. *Annu. Rev. Biochem.* **62**, 121–138 (1993).
31. Bellai-Dussault, K., Nguyen, T. T. M., Baratang, N. V., Jimenez-Cruz, D. A. & Campeau, P. M. Clinical variability in inherited glycosylphosphatidylinositol deficiency disorders. *Clin. Genet.* **95**, 112–121 (2019).
32. Hansen, L. *et al.* Hypomorphic mutations in PGAP2, encoding a GPI-anchor-remodeling protein, cause autosomal-recessive intellectual disability. *Am. J. Hum. Genet.* **92**, 575–583 (2013).
33. Krawitz, P. M. *et al.* PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. *Am. J. Hum. Genet.* **92**, 584–589 (2013).
34. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
35. Lal, N., Puri, K. & Rodrigues, B. Vascular Endothelial Growth Factor B and Its Signaling. *Front Cardiovasc Med* **5**, 39 (2018).
36. Li, X. VEGF-B: a thing of beauty. *Cell Res.* **20**, 741–744 (2010).
37. Iyer, S. & Acharya, K. R. Tying the knot: the cystine signature and molecular-recognition processes of the vascular endothelial growth factor family of angiogenic cytokines. *FEBS J.* **278**, 4304–4322 (2011).
38. MGC Project Team *et al.* The completion of the Mammalian Gene Collection (MGC). *Genome Res.* **19**, 2324–2333 (2009).

39. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
40. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
41. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. Preprint at (2015).
42. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
43. Greer, J., Erickson, J. W., Baldwin, J. J. & Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J. Med. Chem.* **37**, 1035–1054 (1994).
44. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
45. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
46. Varabyou, A., Erdogdu, B., Salzberg, S. L. & Pertea, M. Investigating Open Reading Frames in Known and Novel Transcripts using ORFanage. *bioRxiv* (2023) doi:10.1101/2023.03.23.533704.
47. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

48. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
49. Ruff, K. M. & Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **433**, 167208 (2021).
50. Zhang, Z., Harrison, P. & Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12**, 1466–1482 (2002).
51. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, (2020).
52. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
53. Varabyou, A., Pertea, G., Pockrandt, C. & Pertea, M. TieBrush: an efficient method for aggregating and summarizing mapped reads across large datasets. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab342.
54. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
55. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
56. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

57. Sommer, M. J. *et al.* Structure-guided isoform identification for the human transcriptome. *Elife* **11**, (2022).
58. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).
59. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
60. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
61. Lomsadze, A., Gemayel, K., Tang, S. & Borodovsky, M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.* **28**, 1079–1089 (2018).
62. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
63. McHardy, A. C. & Kloetgen, A. Finding Genes in Genome Sequence. *Methods Mol. Biol.* **1525**, 271–291 (2017).
64. Wang, Q., Lei, Y., Xu, X., Wang, G. & Chen, L.-L. Theoretical prediction and experimental verification of protein-coding genes in plant pathogen genome Agrobacterium tumefaciens strain C58. *PLoS One* **7**, e43176 (2012).
65. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
66. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).

67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

68. Haft, D. H. *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).

69. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

70. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0501-8.

71. Bai, S., Zico Kolter, J. & Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv [cs.LG]* (2018).

72. Stearns, S. C. Daniel Bernoulli (1738): evolution and economics under risk. *J. Biosci.* **25**, 221–228 (2000).

73. Satopää, V. A. *et al.* Combining multiple probability predictions using a simple logit model. *Int. J. Forecast.* **30**, 344–356 (2014).

74. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv [cs.LG]* (2017).

75. Kiefer, J. & Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *Ann. Math. Stat.* **23**, 462–466 (1952).

76. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. in *International Conference on Machine Learning* 1139–1147 (jmlr.org, 2013).

77. Dijkstra, E. W. & Others. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
78. Kelley, J. E. & Walker, M. R. Critical-path planning and scheduling. in *Papers presented at the December 1-3, 1959, eastern joint IRE-AIEE-ACM computer conference* 160–173 (Association for Computing Machinery, 1959).
79. Eggensperger, K. *et al.* Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. https://www.cs.ubc.ca/~hoos/Publ/EggEtAl13.pdf (2013).
80. Hansen, N. & Ostermeier, A. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.* **9**, 159–195 (2001).
81. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, 2019).
82. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
83. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* (2019) doi:10.1016/j.cell.2019.07.016.
84. Sommer, M. J. & Salzberg, S. L. Balrog: A universal protein model for prokaryotic gene prediction. *PLoS Comput. Biol.* **17**, e1008727 (2021).
85. Wang, P. L. *et al.* Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* **9**, e90859 (2014).

86. Tan, W. L. W. *et al.* A landscape of circular RNA expression in the human heart. *Cardiovasc. Res.* **113**, 298–309 (2017).

87. Hanan, M., Soreq, H. & Kadener, S. CircRNAs in the brain. *RNA Biol.* **14**, 1028–1034 (2017).

88. Patop, I. L. & Kadener, S. circRNAs in Cancer. *Curr. Opin. Genet. Dev.* **48**, 121–127 (2018).

89. Wei, G., Zhu, J., Hu, H.-B. & Liu, J.-Q. Circular RNAs: Promising biomarkers for cancer diagnosis and prognosis. *Gene* **771**, 145365 (2021).

90. Rybak-Wolf, A. *et al.* Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* **58**, 870–885 (2015).

91. Sun, L.-F. *et al.* Circular Rims2 Deficiency Causes Retinal Degeneration. *Adv Biol (Weinh)* **5**, e2100906 (2021).

92. Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019).

93. Buratin, A., Bortoluzzi, S. & Gaffo, E. Systematic benchmarking of statistical methods to assess differential expression of circular RNAs. *Brief. Bioinform.* **24**, (2023).

94. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).

95. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Res.* **44**, e58 (2016).

96. Wu, W., Ji, P. & Zhao, F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* **21**, 101 (2020).
97. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
98. Vromman, M. *et al.* Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. *bioRxiv* 2022.12.06.519083 (2022) doi:10.1101/2022.12.06.519083.
99. Zimmerman, A. J. *et al.* A psychiatric disease-related circular RNA controls synaptic gene expression and cognition. *Mol. Psychiatry* **25**, 2712–2727 (2020).
100. Stemmler, K. Hash Tables. https://khalilstemmler.com/blogs/data-structures-algorithms/hash-tables/ (2022).
101. Vromman, M. *et al.* Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. *Nat. Methods* (2023) doi:10.1038/s41592-023-01944-6.
102. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105 (2007).
103. Zhang, J. & Xu, C. Gene product diversity: adaptive or not? *Trends Genet.* **38**, 1112–1122 (2022).
104. Erdogdu, B., Varabyou, A., Hicks, S. C., Salzberg, S. L. & Pertea, M. Detecting differential transcript usage in complex diseases with SPIT. *bioRxiv* 2023.07.10.548289 (2023) doi:10.1101/2023.07.10.548289.