Analysis of Coordinated Skipped Exon Pairs using Single Molecule Sequencing Technology

by

Asa Adadey

B.S. Bioinformatics & Computational Biology
University of Maryland Baltimore County, 2011

Submitted to the Computational and Systems Biology Program
In Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computational and Systems Biology
at the
Massachusetts Institute of Technology

[September 2014]
August 2014

Signature redacted

Signature of Author..... .....
Computational and Systems Biology Graduate Program
August 31, 2014

Signature redacted

Certified by..... .....
Christopher Burge
Professor of Biology and Biological Engineering

Signature redacted

Accepted by..... .....
Christopher Burge
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Graduate Program

Analysis of Coordinated Skipped Exon Pairs using Single Molecule Sequencing Technology

by

Asa Adadey

Submitted to the Computational and Systems Biology Program
In Partial Fulfillment of the Requirements for the Degree of
Masters of Science in Computational and Systems Biology

Abstract

Alternative splicing of mRNA transcripts is a significant step in the production of functioning protein. This process is a major source of molecular diversity, as numerous mRNA and protein products can arise from a single gene locus, and incorrect regulation has been implicated in numerous diseases. While many robust methods exist to study genome-wide single exon splicing patterns, no methodology has been established to accurately examine multiple events over a single isoform. Read sequencing technology has been the limiting factor; however, the recent development of real time, single molecule read sequencing provides an opportunity to characterize alternative splicing on the whole transcript level. We propose a computational approach to detect the splicing patterns of pairs of alternative exons in the same gene. Using a sequenced full-length cDNA library of human MCF-7 transcripts, we are able to evaluate 761 genes and identify three with evidence of non-random splicing of distinct non-adjacent alternative exons, all of which are frame-preserving and biased toward mutual

inclusion. Characterizing their protein products reveals that the domain, secondary, and tertiary structures of the isoforms are not significantly affected. Low read coverage proves to be the greatest hindrance to a larger result set, but overall we provide a computational proof of concept for studying coordinated alternative splicing events on a transcriptomic scale.

Thesis Supervisor: Christopher Burge

Title: Professor of Biology and Biological Engineering

Contents

Introduction

Production of messenger RNA, the template for cellular protein generation, begins with the assembly of a single-stranded, uninterrupted copy of a DNA template. As this pre-mRNA is processed, intragenic regions called introns are excised via a spliceosome complex, and the remaining exons are ligated together. However, depending on the presence of cis-acting RNA silencer and enhancer motifs and associated trans-acting factors, certain exons may be variably included or excluded in the mature transcript. These alternatively spliced exons greatly expand the complexity of protein products derived from the base genetic material, and up to 62% of human deleterious mutations disrupt the normal splicing process (López-Bigasa et al. 2005).

Current research on intragenic splicing events

Alternative splicing events are categorized into several types, with skipped exons being most abundant in mammals, accounting for approximately 46% of annotated events in human databases (Sammeth et al. 2008). While the functional purposes and processes behind single exon alternative splicing have been extensively studied, understanding multi-exon events is a more challenging task. These complex occurrences are estimated to account for 20-35% of all alternative splicing events (Sammeth et al. 2008). Additionally, considerable research has been conducted on multi-exon events occurring adjacent to each other in the gene, namely mutually exclusive exons (MXEs), in which all observed isoforms contain exactly one of a given pair of adjacent exons. Many mechanisms have been described to explain this phenomenon, including

incompatible, overlapping splice sites, steric hindrance preventing spliceosome formation, and NMD pathways (Pohl et al. 2013).

With advancements in high-throughput, second generation sequencing technologies, alternative splicing events can be discovered en masse by mapping of RNA-Seq reads to splice junctions and observing how many junction-spanning reads contain or exclude the exon of interest. Unfortunately, multiple intragenic events cannot be mapped and identified using these methods, as the limited length of the mRNA fragments means one cannot determine which events occurred within the same transcript.

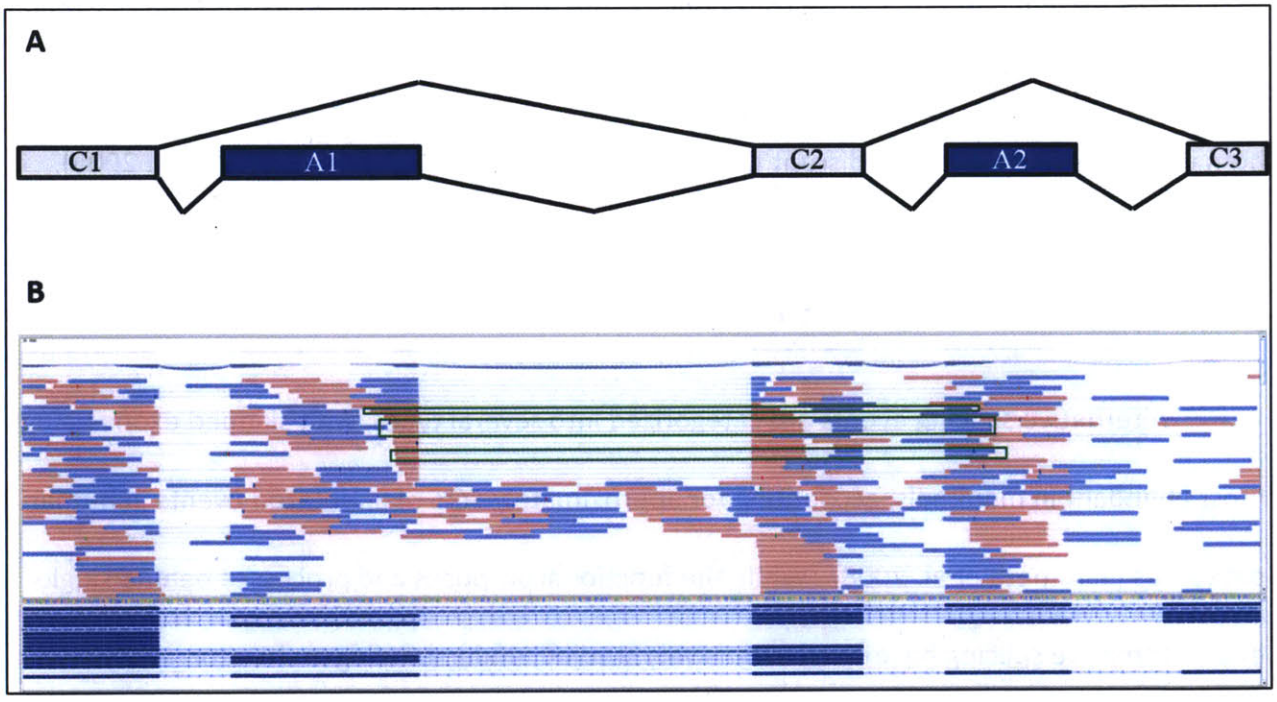


Figure 1: Initial transcript model for skipped exon pair analysis. (**A**) Representation of original skipped exon pair event for statistical analysis of paired-end reads. Inserts mapping across both alternative exons provide isoform-specific information, which can be inferred from insert length distribution. Reads that do not span exons are not informative. (**B**) Most events did not have sufficient insert length to provide relevant information; of 140 reads mapped above, only six (highlighted in green) provided isoform specific information (all for the double inclusion isoform).

Thus the goal of this work was to develop a robust method of identifying complex alternative splicing events and characterize significance behind these events. More advanced statistical models, such as MISO, make use of paired-end reads to extend the quality of exon mapping. MISO in particular creates a distribution of read insert lengths and uses this model to more robustly estimate the "Percent Spliced In" (PSI or Ψ) value of exons based on confidence intervals (Katz et al. 2010). Previous work was done to assess the feasibility of using MISO to predict multiple-exon splicing events occurring near each other to maximize insert coverage. The particular focus was to identify alternative exons separated by a single constitutive exon (Figure 1A), and to observe whether these events were maintained over several tissues and phylogenies. However, MISO was still unable to extract enough single-isoform information from these reads, and the methodology proved inconclusive (Figure 1B).

Full-length cDNA sequencing

Another approach to address the issue of simultaneously identifying multiple intra-transcript splicing events is to use third generation sequencing (TGS) technologies such as that offered by Pacific Biosciences (PacBio). PacBio has been developing long read sequencing technology capable of generating reads over 7 kbp long. By linking PCR products with single-strand hairpin adapters, full-length cDNAs can be used as templates for DNA polymerase (Travers et al., 2010). These polymerase/template complexes are then attached to 50 nm wells and saturated with γ-phosphate fluorescence-labeled nucleotides, which are excited as they are incorporated to the growing DNA polymer (Quail et al., 2012). The result is reads of theoretically unlimited length that are sequenced in real time. This method also produces an

error rate between 12% and 20% in sequenced reads, primarily through evenly distributed indels (Koren et al., 2012).

In this study, we show that full-length, single molecule sequencing has the power to inform on co-occurring splicing events on a single transcript. Querying a MCF-7 cDNA library with our computational approach, we identify three genes whose transcript profiles feature significantly correlated skipped exon events. These isoforms are later characterized on a sequence and structural level.

Methods

Data and Resources

Full-length MCF-7 cDNA sequencing data was obtained from a publicly available Pacific Biosciences dataset. This dataset was generated from reverse transcribed RNA transcripts, created SMRTBell libraries, size selected to reduce polymerization time bias, and sequenced using their TGS platform as described earlier. A full description can be found at http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html.

Exon annotation

PacBio MCF-7 cDNAs were mapped via GMAP (Wu T and Watanabe, 2005), producing a set of detected skipped regions for each read, which were used for exon calling. Low quality reads were filtered out, and reads were compiled by gene to represent the distribution of transcript isoforms. For each gene containing over 50 mapped reads, a set of consensus exons was developed by identifying every base with over 5% read coverage. Consecutive bases were grouped together and their coordinates were established as preliminary consensus exon boundaries. Then every mapped exon overlapping these boundaries (within 10 bp of each end) was aligned, and the most common start and stop coordinates were used as the final consensus exon boundaries.

Each consensus exon was then given a Ψ value where:

$$\Psi_{exon} = \frac{\#\ of\ reads\ with\ overlapping\ exons}{\#\ of\ reads\ spanning\ exon}$$

In the above calculation, overlapping was defined as being within 10 bp of the consensus exon beginning or end coordinates. Spanning reads were defined as those that completely encapsulated the exon, so that reads with alternative first/last exons do not count towards this score.

Identification of skipped exon pairs

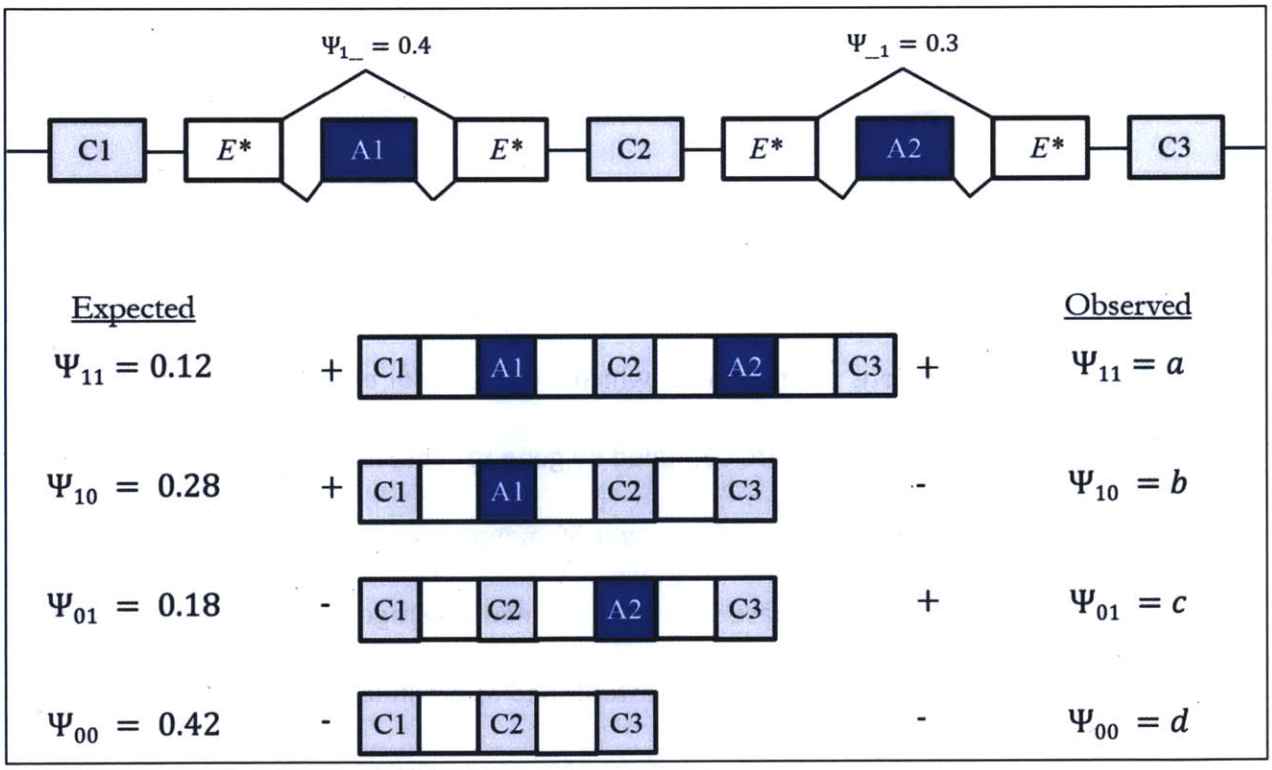


Figure 2: Schematic of skipped exon pair designation. 'A' represents an alternatively expressed exon, 'C' a constitutively expressed exon, and 'E*' represents 0 or more exons of any expression level. Also shown is an example set of expected isoform PSI values with given exon PSI values.

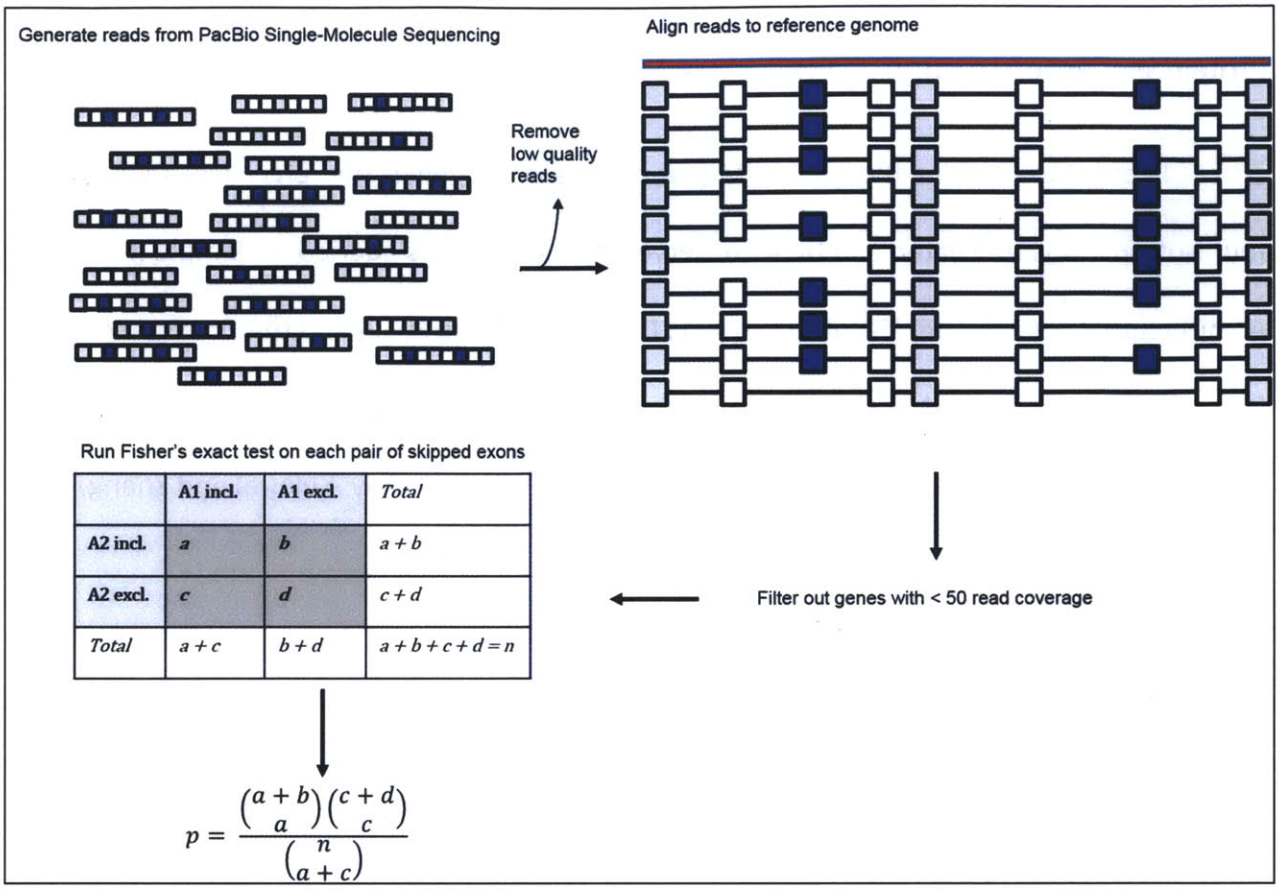


Figure 3: Overview of computational process.

Each gene was then evaluated for the presence of skipped exon pairs (SEPs), based on the set of consensus exons. A skipped exon pair was defined as two alternatively expressed exons e_1, e_2: $0.05 < \Psi_{e_1}, \Psi_{e_2} < 0.9$, that contained at least one constitutively expressed exon e_c: $\Psi_{e_c} \geq 0.95$ located prior to the pair, between the pair and after the pair (Figure 2). This definition excluded consecutive exons, which are mostly understood, as well as alternative first and last exons (outside of the scope of this study). Every read mapped to the gene was then assigned to one of the four potential isoforms regarding the SEP, or thrown out if it did not span both exons. Finally, the distribution of isoform classes was analyzed using Fisher's exact test to determine the association between the two skipped exons (Figure 3).

Protein sequence/structure analysis

In order to examine the protein sequence of relevant genes, the reference cDNA of all four isoforms was translated and their conserved domains/motifs queried via Simple Modular Architecture Research Tool (SMART) (Letunic, Doerks and Bork, 2012) and the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2011). Protein structure and functions were predicted *ab initio* by querying amino acid sequence on I-TASSER (Roy, Kukural and Zhang, 2010) and PredictProtein (Rost, Yachdav and Liu, 2004).

Results

Identification and characterization of three genes expressing correlated skipped exon pairs in human MCF-7 cells

Chrom	Gene Name	Gene Description	Gene Function	Exon Pair	Exon Pair Coordinates	Exon Pair Lengths	Exon Pair Psis	No. of reads	Expected Isoform Distribution	Observed Isoform Distribution	Delta	P-Value
chr10	RAB18	member RAS oncogene family	GTPase	3 5	(27802984, 27803349) (27820458, 27820544)	366 87	0.05 0.06	138	0.51 7.38 8.36 121.75	7 0 1 130	0.047	4.90E-11
chr17	ZNF207	zing finger protein 207	transcription factor	6 9	(30688487, 30688534) (30693684, 30693776)	48 93	0.71 0.29	185	38.12 93.17 15.59 38.12	52 80 2 51	0.075	2.10E-07
chr20	PIGT	phosphatidylinositol glycan anchor biosynthesis, class T	biosynthesis	3 9	(44047492, 44047619) (44050023, 44050223)	128 201	0.86 0.74	58	37.07 12.93 5.93 2.07	41 9 2 6	0.068	2.50E-03

Table 1: Correlated skipped exon pairs (significant p-value on Fisher's exact test). Isoform distributions are ordered as double inclusion, inclusion of only first in pair, inclusion of only second in pair, and double exclusion. Delta is calculated as observed – expected over PSI of double inclusion isoform.

PacBio sequencing of MCF-7 mRNA transcripts resulted in a mapping of 264,878 high quality reads, of which 210,663 (79.5%) were successfully mapped to 4,989 Ensembl annotated genes. 761 of these genes contained enough reads (≥ 50) to undergo analysis. Of this set, 35 genes, or 4.6% were found to express a total of 64 SEP events. A total of three genes were found to express SEPs with a correlated, non-independent distribution across isoforms: ENSG00000099246 (RAB18, $p = 4.94 \times 10^{-11}$); ENSG00000010244 (ZNF207, $p = 2.11 \times 10^{-7}$); and ENSG00000124155 (PIGT, $p = 2.51 \times 10^{-3}$) (Table 1). However, the latter result (PIGT) is questionable, as it fails under the Bonferroni-Holm correction for multiple comparisons ($p > 0.05$).

Visually comparing the mapped reads for ZNF207 and PIGT (Figures 7A, 8A), reads supporting the double inclusion isoform appeared varied enough to suggest that they represent distinct mRNA molecules and thus accurate representations of the real transcript distribution.

On the other hand, the RAB18 double inclusion reads appeared strikingly congruent (Figure 4), with only one read supporting any single inclusion isoform. This uniformity raised the possibility that these double inclusion reads were actually a single molecule that underwent excessive PCR amplification during the sequencing process. While substantial variation in sequence was observed between these reads (Figure S1), this may be explained by sequencing errors, leaving some question as to whether these are indeed uniquely sequenced transcripts.

All three SEPs showed an increased expression of the double inclusion isoform, with RAB18 in particular featuring a more than 14-fold increase of $\Psi_{e_3e_5}$ over the expected distribution. A key feature of these events is that the involved exons each maintain reading frame, suggesting that their presence or absence will not introduce a downstream premature termination codon. The exons and genes are of average length (with one or two exceptions) and feature canonical splice site motifs (Table 2). Examining the functionality, localization, and pathways of the protein products of these genes yielded no shared properties.

Gene/exon	3'SS (-20 to +5)	5'SS (-5 to +10)
RAB18 – exon 3	AATGCTAAACTGTCTCTTAG|GTAAA	GCAAG|GTAAGCTGAT
RAB18 – exon 5	TTACTTTTACTTTTCTTTAG|GTTAC	AATAC|GTAAGCAAAT
ZNF207 – exon 6	TTTGTCATACATTTTCACAG|ATTGC	AACAT|GTAAGCATCT
ZNF207 – exon 9	TCCTTTCTTTTATGCTACAG|ATGGG	CACAA|GTACGCAGGA
PIGT – exon 3	GGATTTGTGTCTCTATCCAG|TGTGG	CAATG|GTGAGATAAC
PIGT – exon 9	GTTGATATTTCTTTACACAG|AGGCC	ACCAA|GTGAGGACCT
Consensus	...YYYYYYYNYAG|GNNNN	NNNAG|GTRAGTNNNN

Table 2: Comparison of splice site motifs of skipped exons involved in coordinated exon skipping events

RAB18

RAB18 is one of over 60 Rab proteins, small GTPases belonging to the Ras superfamily that regulate membrane trafficking in intracellular vesicles. Unlike most Rabs, RAB18 is expressed ubiquitously in all studied tissue types, in both mouse and human (Lütcke et al. 1994, Schäfer et al. 2000), with particularly high expression in human endothelial cells. It also possesses unique features in primary and secondary structure that suggest different ligand interactions from the other Rabs (Lütcke et al. 1994). Point and deletion mutations in the RAB18 gene have been characterized in families afflicted with Warburg Micro syndrome, implicating a role in eye and brain development (Bem et al., 2011).

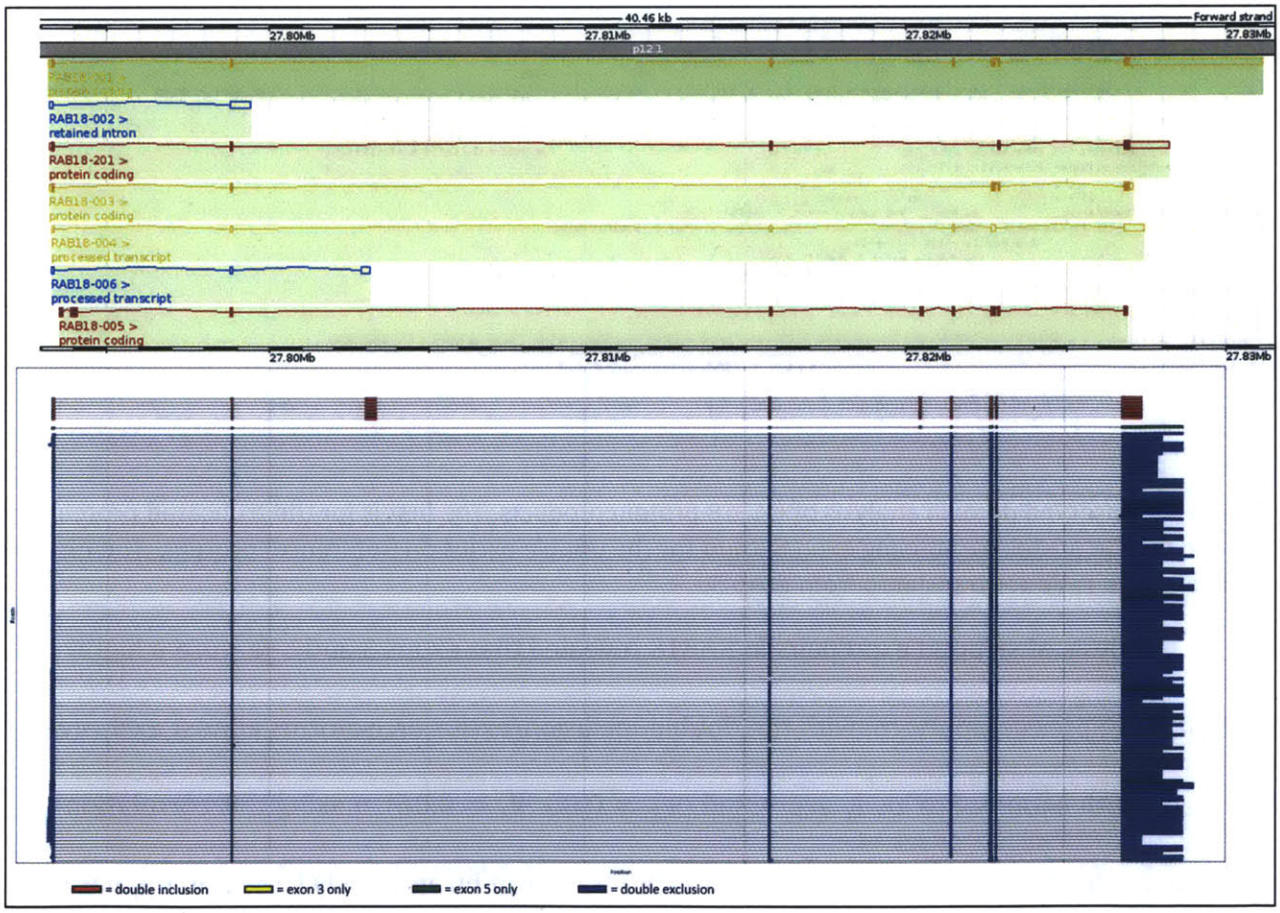


Figure 4: Distribution of read isoforms for RAB18 (below), sorted around inclusion/exclusion of exons 3 and 5. Ensembl transcript annotations (above) are shown for comparison.

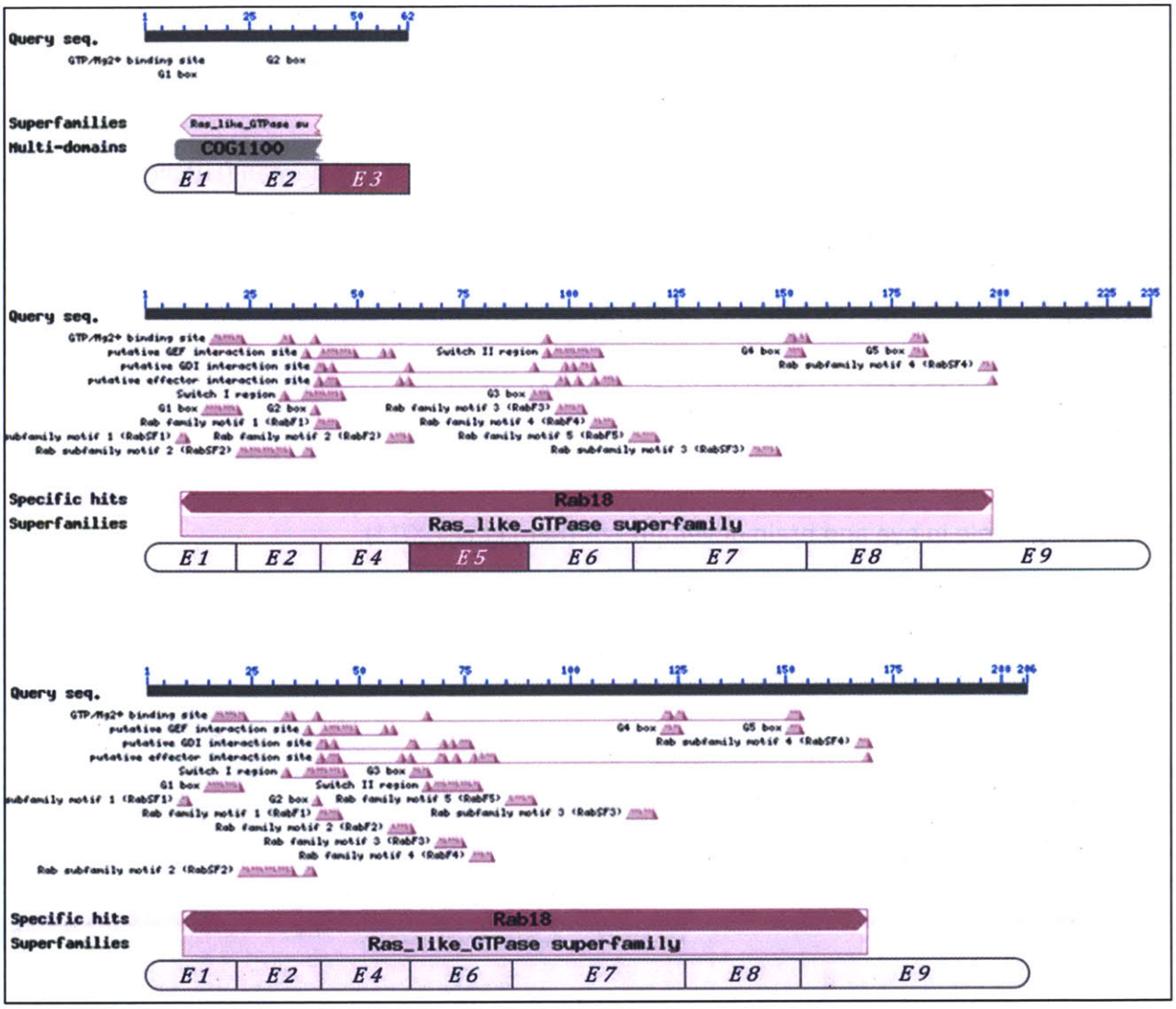


Figure 5: Conserved domain analysis of RAB18 protein products. The first is for the truncated protein, second for the protein containing exon 5, third for the double exclusion (reference) protein. The red box highlights the residues translated from exon 5.

Analysis of skipped exon pairs in RAB18 revealed that exons 3 and 5 feature a highly correlated expression pattern, with the double-inclusion/exclusion isoforms almost exclusively favored. Both exons were lowly expressed ($\Psi_3 = 0.05, \Psi_3 = 0.06$), making this correlation even more significant. Exon 3, which extends the transcript by an usually large 366 bp, has not been annotated by Ensembl, and no gene/transcript databases feature this exon. Exon 5

contains an 87 bp region and transcripts with this isoform have been well documented. The double exclusion isoform is most commonly expressed, as supported by the sequencing data.

Translating the four different isoforms revealed that exon 3 contains a stop codon, resulting in a truncated, 62-residue protein product. A transcript search on Ensembl displayed a processed transcript with its final exon overlapping exon 3, but neither the 3' nor 5' splice sites align with the observed data (Figure 4). In order to determine the nature of this abridged product, the cDNA and peptide sequences were closely analyzed. Performing a blastn and blastx alignment (Altschul et. al, 1990) of exon 3 against the human genome brought up no matches. Next, the translated amino acid sequences of the four different isoforms were analyzed for conserved domains and features using the NCBI Batch CD search tool (Figure 5). As expected, the two variants containing the full protein product matched the RAB18 domain as well as GTPase associated interaction sites such as for GTP, guanine nucleotide exchange factor (GEF), and GDP dissociation inhibitor (GDI) (Wu YW et al., 2007). Conserved domain analysis on the putative, 62-residue variant revealed a relatively weak correlation to the Ras-like GTPase superfamily and COG1100, a general match to small G proteins, suggesting some possible GTPase functionality. There was, however, no significant match to any known interaction sites. Further, looking at predicted structure, PredictProtein and I-TASSER predict that both the secondary (Figure 6A) and tertiary (Figure 6B) peptide structures of the truncated protein match the first 62 residues of full Rab18, despite differences in the last 20 residues.

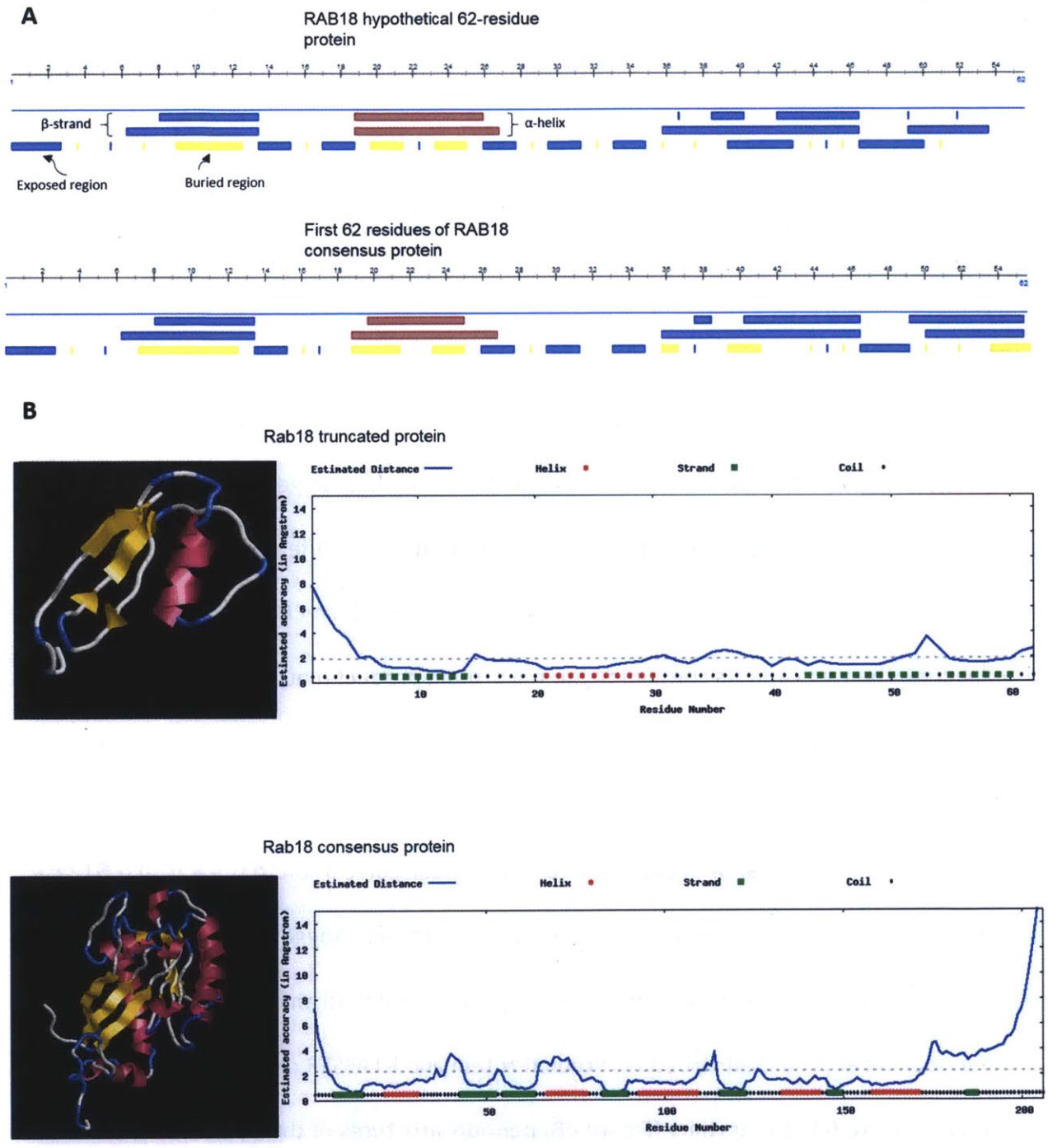


Figure 6: RAB18 peptide structure analysis. (**A**) Secondary structure comparison of truncated and reference RAB18 proteins. The first two rows show structure prediction (blue for strand, red for helix) while the third predicts exposed (blue) or buried (yellow) regions. (**B**) Tertiary structure prediction for RAB18 truncated and reference proteins. Note similarities with the triple strand component.

The introduction of an early termination codon in exon 3 may be key to its positive correlation with exon 5. However, conserved domain analysis of the exon 5-containing variant revealed a distinct gap in known features over the relevant peptides (Figure 5), as expected of symmetrical alternative exons (Magen and Ast, 2005). PredictProtein also showed a lack of stable secondary structure over the peptide sequence (Figure 6) and annotated it as a predicted high-disordered region. Therefore, no specific feature of exon 5 could be ascertained.

ZNF207

ZNF207 is one of hundreds of zinc finger proteins found in the human genome. Zinc finger proteins are defined by the presence of one or more zinc finger motifs, small 20-30 residue domains stabilized by metal ions that recognize and bind DNA motifs (Laity, Lee and Wright, 2001). ZNF207 is ubiquitously expressed in human tissue and is predicted to function as a transcriptional regulator (Pahl et al., 1998), although specific interactions have not been studied in depth.

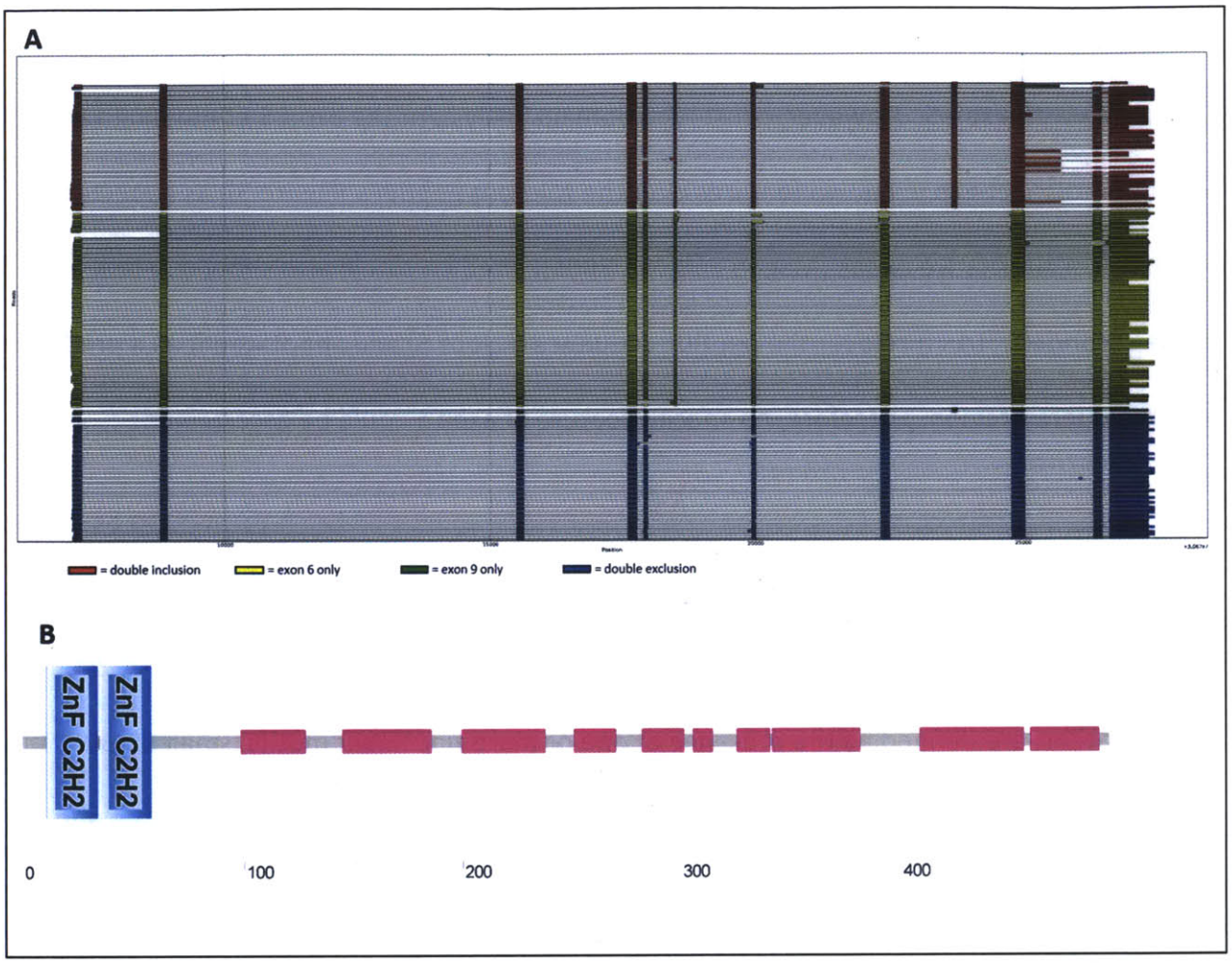


Figure 7: Investigation of the ZNF207 gene. (**A**) Read isoform distribution of ZNF207 gene, sorted over inclusion/exclusion of exons 6 and 9. (**B**) Conserved domain analysis reveals two zinc finger domains at the N-terminus, followed by a series of low-complexity regions. Shown is the double inclusion isoforms; other isoforms retain the same feature set.

Skipped exon pair analysis revealed that exons 6 and 9 in ZNF207 show a correlated isoform distribution in the MCF-7 cells (Figure 7A). The double inclusion isoform was again favored, by 1.4-fold above expectation. Both exons maintain the reading frame, with exon 6 containing 48 nucleotides and exon 9 containing 93.

Conserved domain analysis using SMART matched the expected zinc finger domain (Figure 7B), specifically two occurrences of the Cys_2His_2 motif, commonly found in mammalian

transcription factors to be involved in sequence-specific DNA binding (Pabo, Peisach and Grant, 2001). These domains are found within the first 58 residues of the ZNF207 protein, and are unaffected by the alternative exons, which are instead found in the low complexity regions. Both PredictProtein and I-TASSER showed highly disordered secondary and tertiary structure after the zinc finger domains, suggesting that these regions, including exons 6 and 9, do not perform a core function of the protein.

PIGT

PIGT encodes a protein essential for biosynthesis of glycosylphosphatidylinositol (GPI), a glycolipid that anchors proteins to the cell membrane, typically in blood cells. PIGT, along with at least four other components, forms GPI transamidase, the multiunit enzyme complex responsible for tethering GPI to nascent peptides in the endoplasmic reticulum (Ohishi, Inoue and Kinoshita, 2001). Immunoprecipitation assays and knockout studies reveal that PIGT is necessary for structural stability of the transamidase complex via a disulfide bridge formed with another component, specifically Cys^{182} (Ohishi et al., 2003). In terms of the clinical relevance of this gene, a pedigree-based sequencing analysis implicated a specific PIGT mutation, p.Thr183Pro, in development of a novel intellectual disorder (Kvarnung et al, 2013), while a dual somatic deletion and germ-line splice acceptor site mutation (affecting intron 10) was found to produce a rare variant of paroxysmal nocturnal hemoglobinuria (Krawitz et al, 2013).

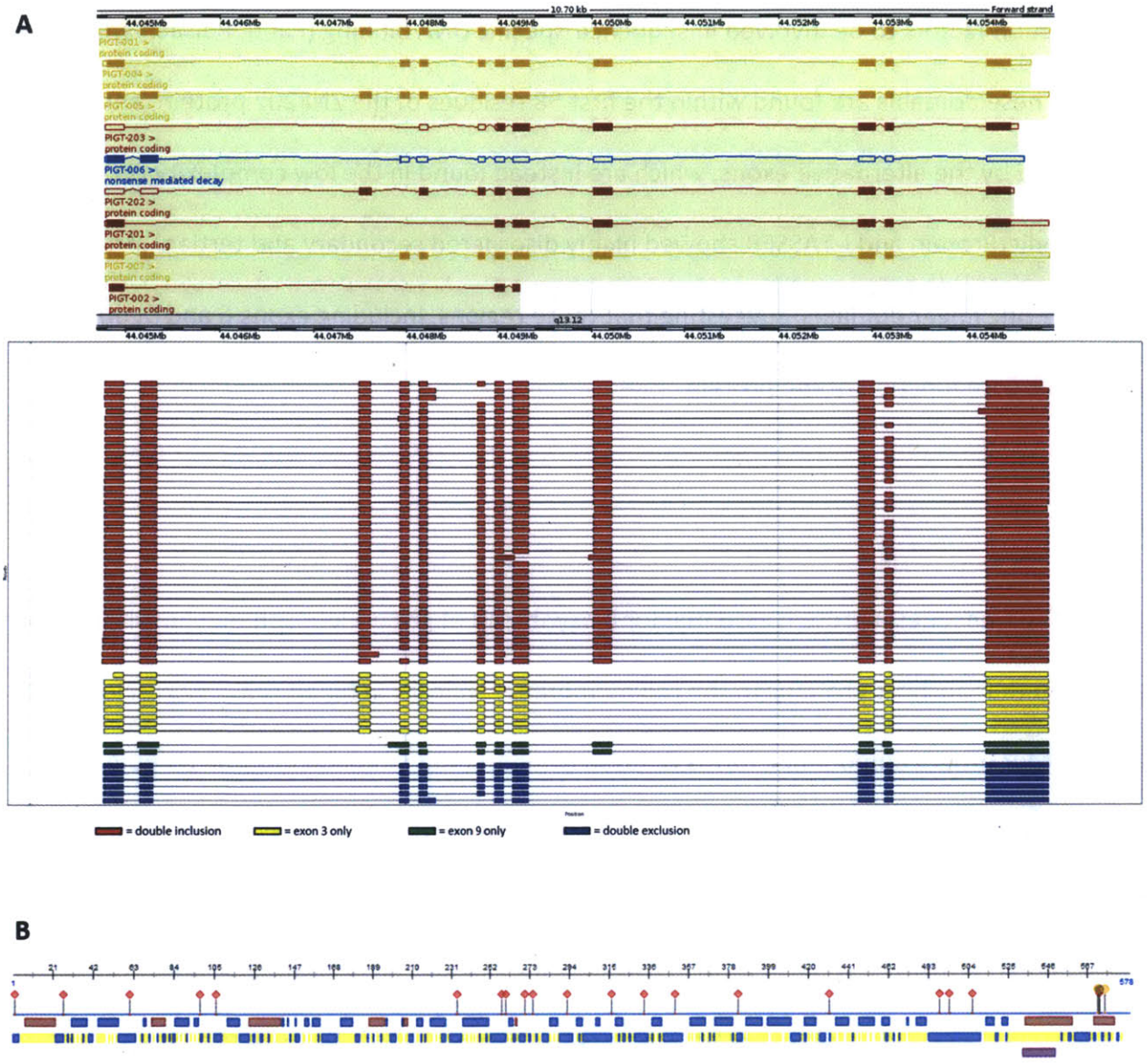


Figure 8: Investigation of the PIGT gene. (**A**) Read isoform distribution of PIGT gene, sorted over inclusion/exclusion of exons 3 and 9. Ensembl transcript annotations are shown above for comparison. (**B**) Secondary structure and feature prediction for PIGT double inclusion gene.

Compared to RAB18 and ZNF207, PIGT featured a lower correlation between alternative exons 3 and 9 (Figure 8A). Exon 3 is particularly interesting as, although its length (128 bp) is not divisible by three, all transcripts lacking this exon feature a shorter exon 2 with a complementary frame shift, maintaining the reading frame. Neither of the known deleterious

mutations overlap exon 3 or 9 specifically. Conserved domain analysis with Batch-CD predicted the expected Gpi16 superfamily for all four variants, with a higher affinity for the double inclusion variant. PredictProtein annotated a putative secondary structure along the exon 3 region of the translated protein, consisting of a large α-helix and smaller β-strands, and additional protein binding sites along the exon 9 region (Figure 8B).

Discussion

Overall, we were able to use PacBio full-length cDNA sequencing with statistical hypothesis testing on isoform distribution to present three cases of correlated exon skipping in human MCF-7 cells. This dataset is able to overcome many of the shortcomings of short-read mapping and be robustly analyzed for the occurrence of complex splicing events. Each case points to favored production of the double inclusion isoform, and visualizing the distribution of reads clearly confirms these results. These results indicate the existence of some underlying mechanism promoting inclusion or exclusion of both exons, and suggest some biological purpose behind this phenomenon.

Previous studies of coordinated exon splicing have shown occurrence of this phenomenon in a tissue-specific manner that can be disrupted by sequence manipulation within the gene. These studies also suggest numerous mechanisms may be at play, depending on the particular context. A microarray profiling of exon skipping events in mouse genes (3,707 events over 27 tissue/cell types) revealed 38 instances where exons within the same gene featured correlated inclusion profiles (Spearman correlation ≥ 0.60) (Fagnani et al., 2007). This correlation appeared in a tissue-specific manner, especially favoring central nervous system (CNS) tissue. Notably, positively correlated exon pairs were typically found in genes with specific CNS functionality, and within four exons of each other. This observed proximity-based preference implies coordinated interactions between nearby splicing factors as potential mechanism.

On the other hand, two additional studies observed coordinated alternative splicing between distal exons in different contexts. An RT-PCR analysis of splicing in *TMEM16A*, a biomarker overexpressed in many tumors, found that splicing coordination between exons 6b and 15 is more prominent in breast tumors than in normal breast tissue (Ubby et al., 2013). The fibronectin gene also features a similar alternative splicing profile, with two distal exons exhibiting coordinated inclusion in constructed minigenes and mouse embryo fibroblasts (Fededa et al., 2005). Disruption of an exonic splicing enhancer (ESE) in the upstream exon led to a subsequent 8-fold decrease in inclusion of the downstream exon (~3.4 kbp apart), but the reverse did not hold when an ESE in the downstream exon was disrupted. Assaying constructs with swapped promoters revealed that this phenomenon is dependent on slow pol II elongation or high pol II pausing. Both of these studies point to specific contextual pressures influencing the effectiveness of the coordination mechanism.

Finally, an analysis of the *slo-1* gene in *C. elegans* found three complex, coordinated alternative splicing events throughout the gene locus (Glauser et al., 2011). A known single-nucleotide, loss-of-function mutation in the intron upstream of one alternative exon alters the splicing profile, favoring exclusion of the most downstream exon, but still maintaining significant correlation. This mutation was found to disrupt a UAAAUC intronic *cis*-regulatory element, and a brief query of the *C. elegans* genome found this motif to be significantly more present in genes with multiple alternative splice site events (Glauser et al., 2011). These examples implicate widespread regulatory elements in control of coordinated alternative exon events.

Need for development of robust datasets and analytical methods

While the single molecule read mapping data is essential to the identification of full transcript isoforms, several issues with the particular Pacbio MCF-7 dataset hindered more wide-scale characterization of these events. Primarily, the read quality and quantity filtering limited the searchable dataset from over 20,000 Ensembl-annotated genes to just 761, a 96% reduction. Many genes that had shown potential for skipped exon pair events did not have the necessary read coverage when examined manually. Similarly, an attempt to corroborate these events in a single-molecule hESC dataset yielded a subset of only 185 genes with sufficient coverage to undergo analysis. While this is a side effect of the real-time sequencing methodology, this issue can be tempered by employing replicate datasets. At this point however, cost becomes a legitimate concern; at $695K for 100 Mb worth of reads, Pacbio's current platform is by far the priciest choice for high-throughput sequencing (Quail et al., 2012).

The high error rate of the sequencing dataset also proved to be an impediment to this approach. Besides the low availability of high quality read mappings, the sequences that did map still possessed enough errors to cause whole exon misplacement in several cases. This meant relaxing the constraints on identifying consensus exons, and a reduced upper bound on calling alternative exons, lowering the overall accuracy of this approach. Although the three correlated exon pairs were manually confirmed, a larger dataset (and thus result set) would suffer in reliability. A possible response to this issue is combining single molecule sequencing with short-read mappings, using the later to correct errors in the long reads. One such

algorithm creates a hybrid consensus sequence with such an approach, producing corrected reads with over 99.9% accuracy (Koren et al., 2012).

Finally, generalized inferences from identified skipped exon pair events are impossible to make without corroborating evidence in multiple RNA expression environments. Although RAB18, ZNF207, and PIGT showed positive correlation between alternative exons, these genes likely have significantly different isoform distributions in various tissues and organisms. Transcriptome studies indicate that individual exon splicing profiles are conserved by species more so than by tissue (Merkin et al., 2012, Barbosa-Morais et al., 2012); the expectation is that whole transcript isoform profiles would maintain this relationship. Additionally, these results must be validated in targeted experimental conditions to provide confidence that they are not artifacts of the sequencing technology. RT-qPCR of RAB18, ZNF207, and PIGT (or any genes that show significant exon correlation in other datasets) could provide single molecule-specific readouts of transcripts, using probes specific to each isoform. Not only can these events be validated in multiple species and tissue contexts, but approaches similar to previous studies, such as ESE disruption, exon knockout, and induction of known deleterious mutations can be replicated specifically to elucidate models of action for coordinated exon skipping.

The results of this work show promising applications of long-read sequencing technology, providing comprehensive data that reduces the complexity of transcript characterization to a simpler computation model, with verifiable results. This study, along with a growing body of transcriptome sequencing experiments, will open the window to a richer understanding of multi-exon alternative splicing events. Skipped exon triplets, alternative splice site combinations, and other events can potentially be analyzed using similar methods,

and once identified throughout the genome, these events can be followed up with targeted experimental validation.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science.* 338(6114):1587-1593.

Bem D, Yoshimura S, Nunes-Bastos R, Bond FC, Kurian MA, Rahman F, Handley MT, Hadzhiev Y, Masood I, Straatman-Iwanowska AA, Cullinane AR, McNeill A, Pasha SS, Kirby GA, Foster K, Ahmed Z, Morton JE, Williams D, Graham JM, Dobyns WB, Burglen L, Ainsworth JR, Gissen P, Müller F, Maher ER, Barr FA, Aligianis IA (2011) Loss-of-function mutations in RAB18 cause Warburg micro syndrome. *Am J Hum Genet.* 88(4):499-507.

Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, Saltzman AL, Shai O, Lee L, Rozenhek A, Mohammad N, Willaime-Morawek S, Babak T, Zhang W, Hughes TR, van der Kooy D, Frey BJ, Blencowe BJ (2007) Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol.* 8(6):R108.

Fededa JP, Petrillo E, Gelfand MS, Neverov AD, Kadener S, Nogués G, Pelisch F, Baralle FE, Muro AF, Kornblihtt AR (2005) A polar mechanism coordinates different regions of alternative splicing within a single gene. *Mol Cell.* 19(3):393-404.

Glauser DA, Johnson BE, Aldrich RW, Goodman MB (2011) Intragenic alternative splicing coordination is essential for Caenorhabditis elegans slo-1 gene function. *Proc Natl Acad Sci U S A.* 108(51):20790-20795.

Katz Y, Wang E, Airodli E, Burge C (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth.* 7(12):1009-1015.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam M Phillippy (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol.* 30(7):693-700.

Krawitz PM, Höchsmann B, Murakami Y, Teubner B, Krüger U, Klopocki E, Neitzel H, Hoellein A, Schneider C, Parkhomchuk D, Hecht J, Robinson PN, Mundlos S, Kinoshita T, Schrezenmeier H (2013) A case of paroxysmal nocturnal hemoglobinuria caused by a germline mutation and a somatic mutation in PIGT. *Blood.* 122(7):1312-1315.

Kvarnung M, Nilsson D, Lindstrand A, Korenke GC, Chiang SC, Blennow E, Bergmann M, Stödberg T, Mäkitie O, Anderlid BM, Bryceson YT, Nordenskjöld M, Nordgren A (2013) A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. *J Med Genet.* 50(8):521-528.

Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol.* 11(1):39-46.

Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40:D302-D305

López-Bigasa N, Audita B, Ouzounisa C, Parrab G, Guigób R (2005) Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* 579:1900-1903

Lütcke A, Parton RG, Murphy C, Olkkonen VM, Dupree P, Valencia A, Simons K, Zerial M (1994) Cloning and subcellular localization of novel rab proteins reveals polarized and cell type-specific expression. *J Cell Sci.* 107(12):3437-3448.

Magen A, Ast G (2005) The importance of being divisible by three in alternative splicing. *Nucleic Acids Res.* 33(17):5574-5582.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39:D225-D229

Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science.* 338(6114):1593-1599.

Ohishi K, Inoue N, Kinoshita T (2001) PIG-S and PIG-T, essential for GPI anchor attachment to proteins, form a complex with GAA1 and GPI8. *EMBO J.* 20(15):4088-4098.

Ohishi K, Nagamune K, Maeda Y, Kinoshita T (2003) Two subunits of glycosylphosphatidylinositol transamidase, GPI8 and PIG-T, form a functionally important intermolecular disulfide bridge. *J Biol Chem.* 278(16):13959-13967.

Pabo CO, Peisach E, Grant RA (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem.* 70:313-40.

Pahl PM, Hodges YK, Meltesen L, Perryman MB, Horwitz KB, Horwitz LD (1998) ZNF207, a ubiquitously expressed zinc finger gene on chromosome 6p21.3. *Genomics.* 53(3):410-412.

Pohl M, Bortfeldt R, Grützmann K, Schuster S (2013) Alternative splicing of mutually exclusive exons—A review. *BioSystems.* 114:31-38

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 13:341.

Rost B, Yachdav G, Liu J (2004) The PredictProtein server. *Nucleic Acid Res.* 32:321-326

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protocols.* 5:725-738

Sammeth M, Foissac S, Guigó R (2008) A General Definition and Nomenclature for Alternative Splicing Events. *PLoS Comput Biol.* 4(8): e1000147

Schäfer U, Seibold S, Schneider A, Neugebauer E (2000) Isolation and characterisation of the human rab18 gene after stimulation of endothelial cells with histamine. *FEBS Lett.* 466(1):148-154.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38(15):e159.

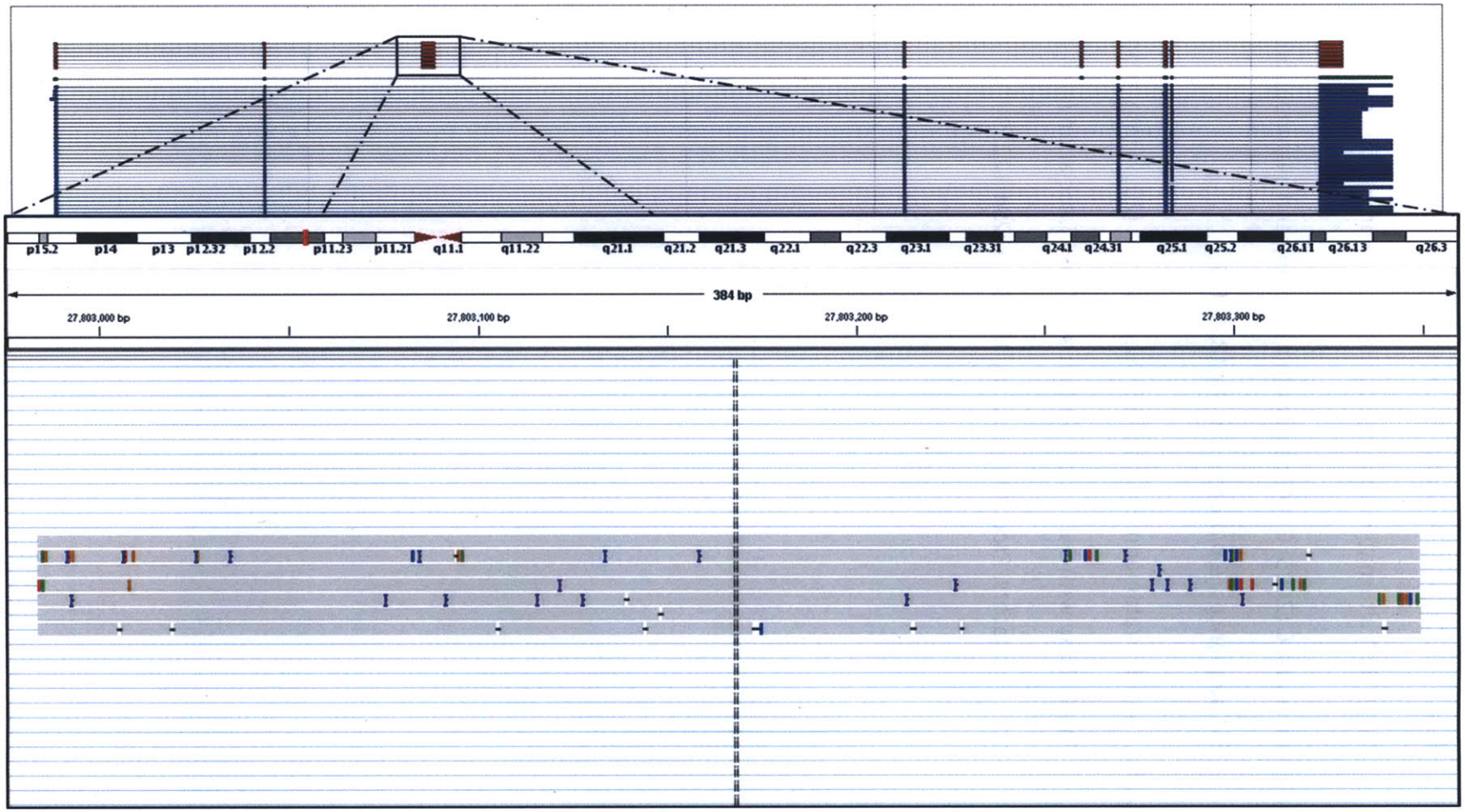
Ubby I, Bussani E, Colonna A, Stacul G, Locatelli M, Scudieri P, Galietta L, Pagani F (2013) TMEM16A alternative splicing coordination in breast cancer. *Mol Cancer.* 12:75.

Wu T, Watanabe C (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* 21:1859-1875

Wu YW, Tan KT, Waldmann H, Goody RS, Alexandrov K (2007) Interaction analysis of prenylated Rab GTPase with Rab escort protein and GDP dissociation inhibitor explains the need for both regulators. *Proc Natl Acad Sci.* 104(30):12294-12299

Appendix

A


p15.2
p14
p13
p12.32
p12.2
p11.23
p11.21
q11.1
q11.22
q21.1
q21.2
q21.3
q22.1
q22.3
q23.1
q23.31
q24.1
q24.31
q25.1
q25.2
q26.11
q26.13
q26.3
384 bp
27,803,000 bp
27,803,100 bp
27,803,200 bp
27,803,300 bp

B

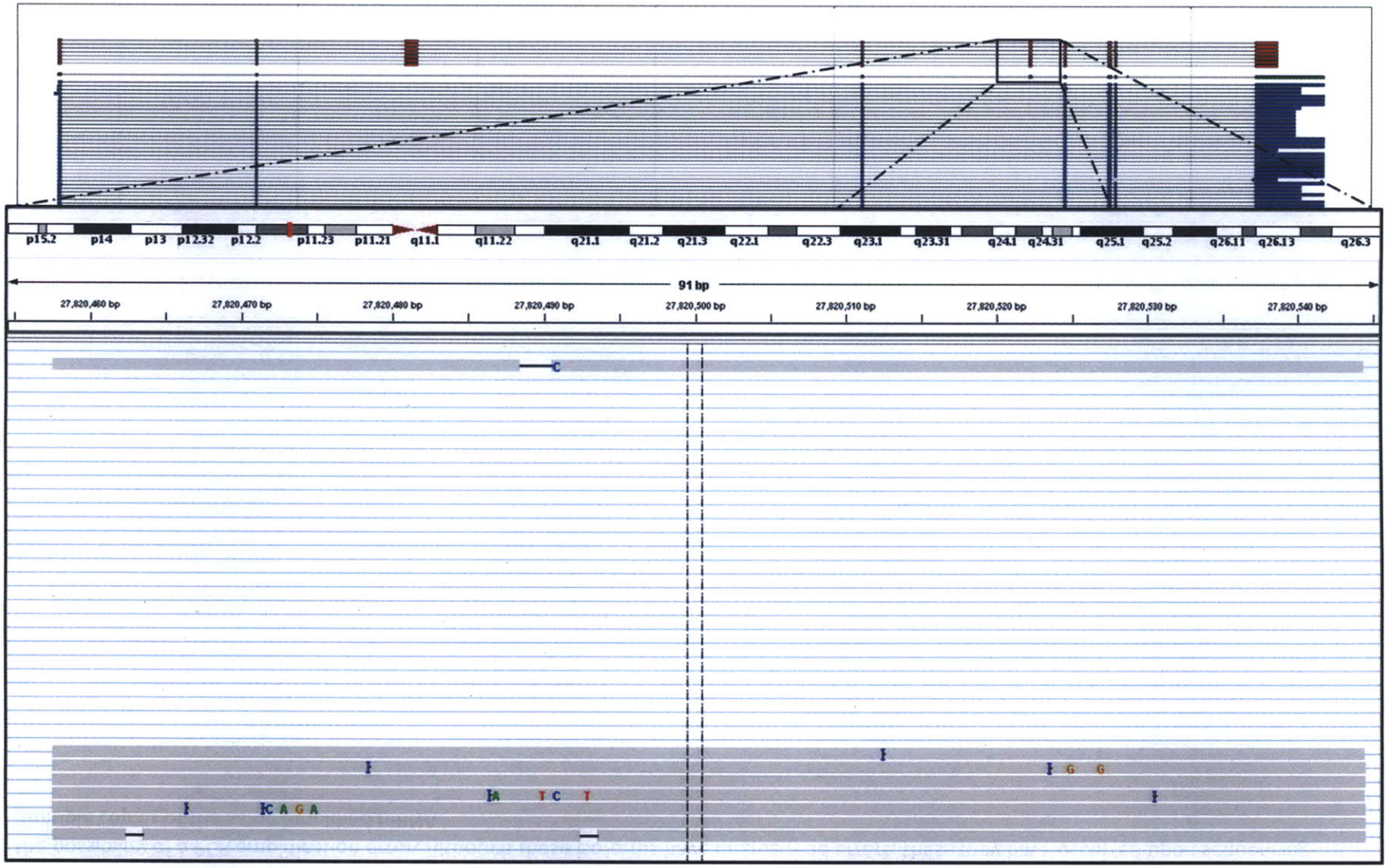
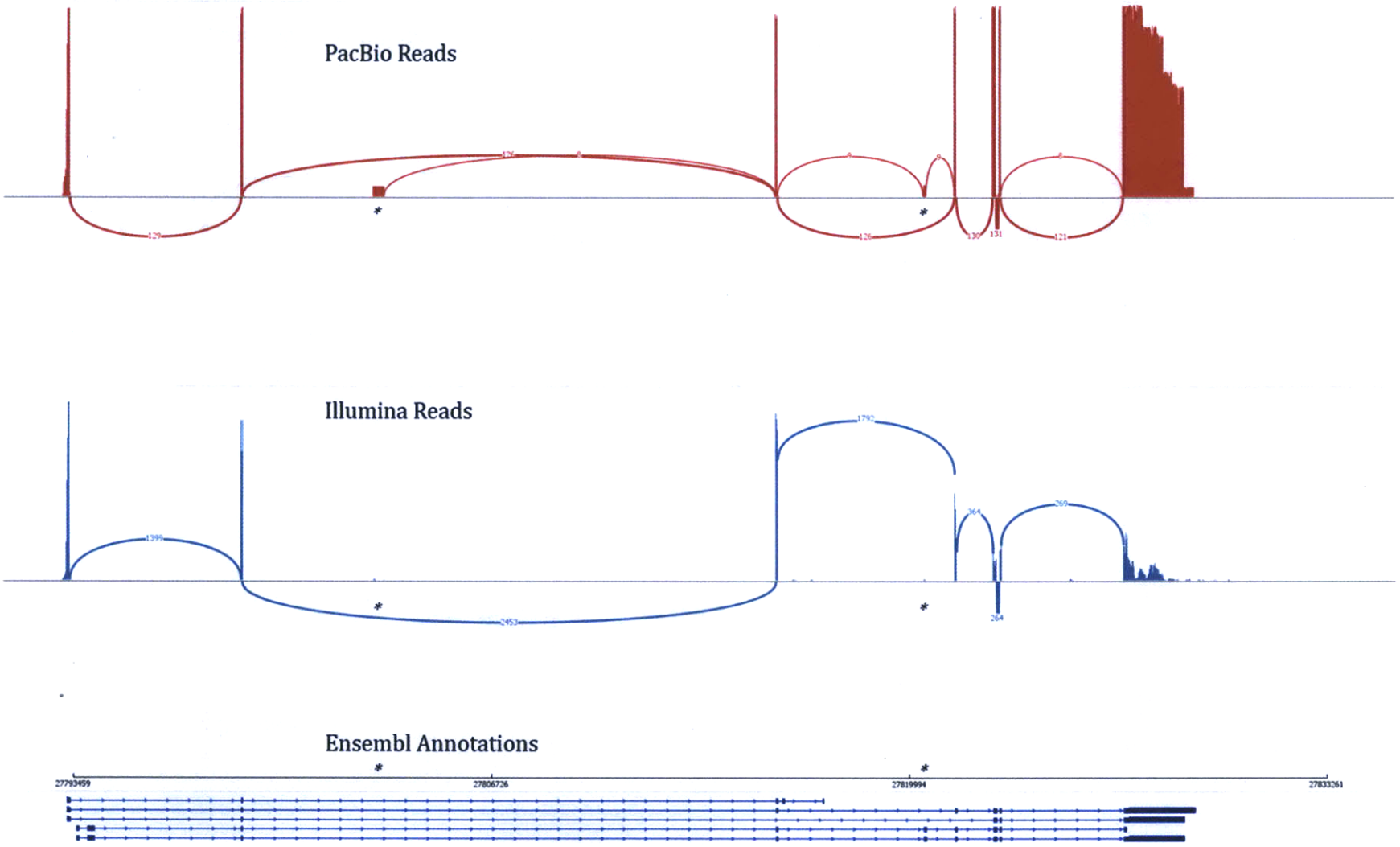
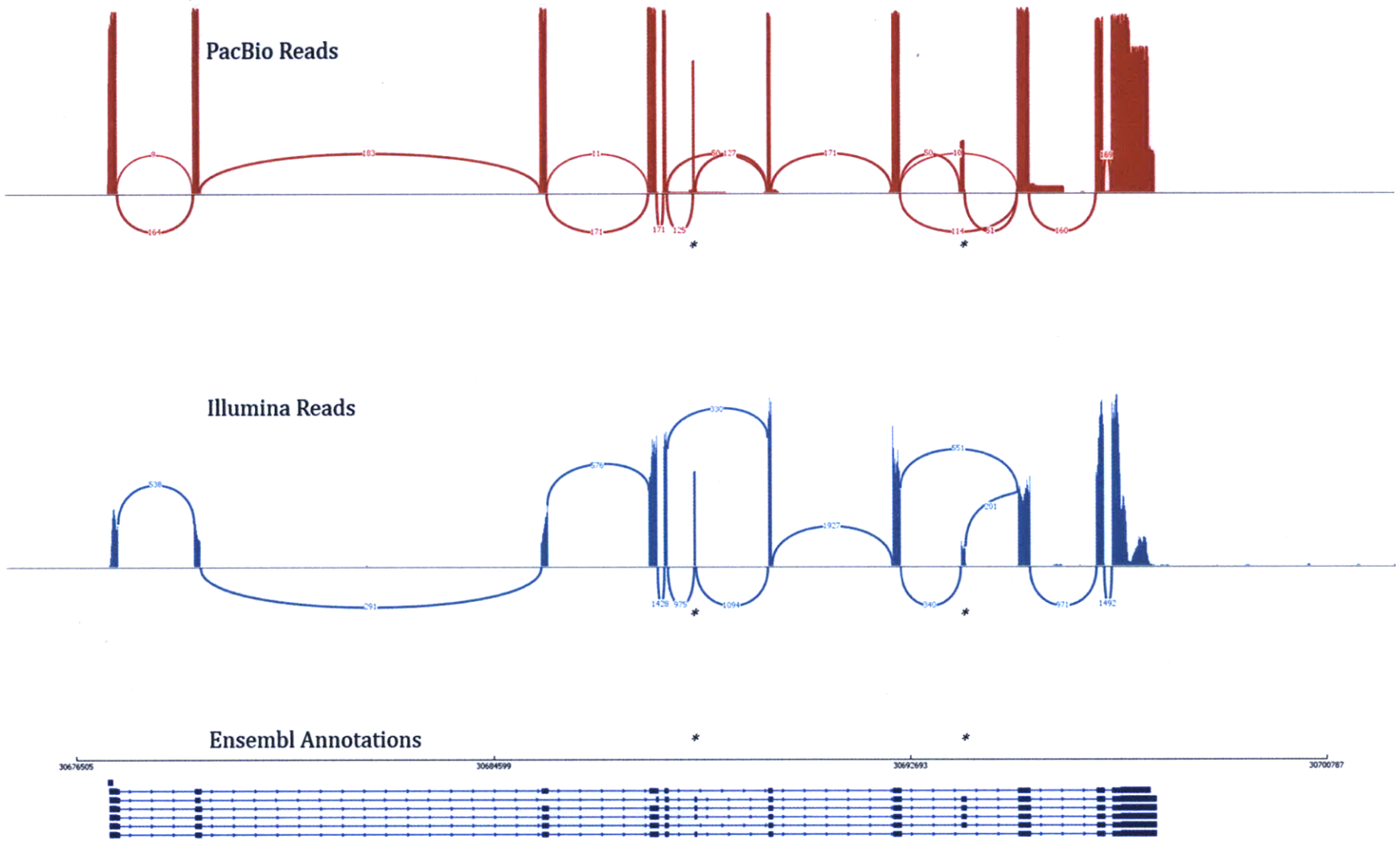

p15.2
p14
p13
p12.32
p12.2
p11.23
p11.21
q11.1
q11.22
q21.1
q21.2
q21.3
q22.1
q22.3
q23.1
q23.31
q24.1
q24.31
q25.1
q25.2
q26.11
q26.13
q26.3
91 bp
27,820,460 bp
27,820,470 bp
27,820,480 bp
27,820,490 bp
27,820,500 bp
27,820,510 bp
27,820,520 bp
27,820,530 bp
27,820,540 bp
C
C A G A
A T C T
G G

Figure S1: Error profiles for (**A**) exon 3 and (**B**) exon 5 of RAB18. The sequence length for all of the double inclusion isoforms is identical, raising the possibility of a PCR amplification error. Although these isoforms have unique read errors, these may have occurred upon sequencing multiple copies of the same amplified mRNA.

A


PacBio Reads
Illumina Reads
Ensembl Annotations
27793459
27806726
27819994
27833261

B


PacBio Reads
Illumina Reads
Ensembl Annotations
30676505
30684599
30692693
30700787

C

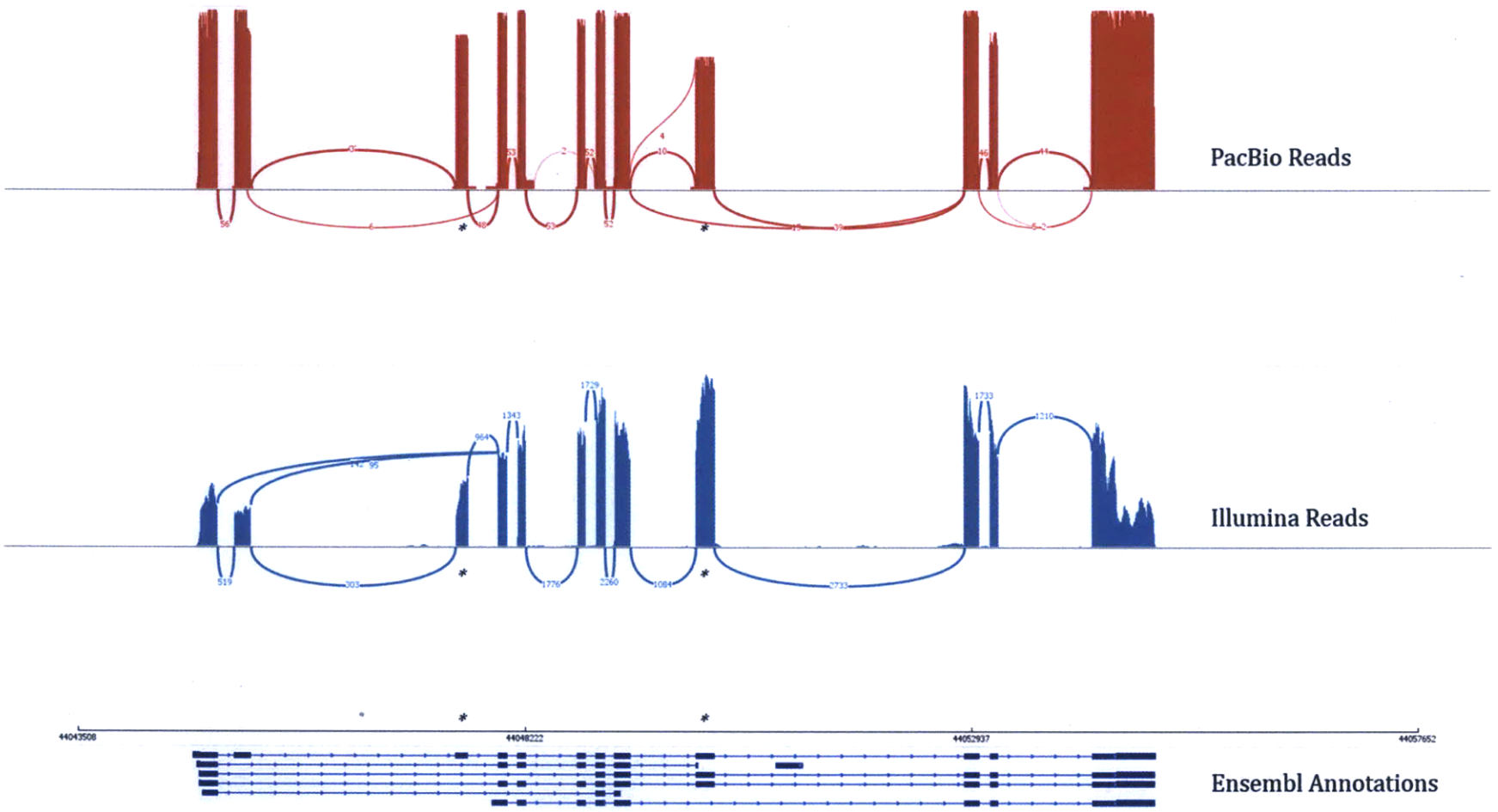

PacBio Reads
Illumina Reads
Ensembl Annotations
44043508
44048222
44052937
44057652

Figure S2: Sashimi plot comparison of PacBio and Illumina MCF7 reads over (**A**) RAB18, (**B**) ZNF207, and (**C**) PIGT. Columns represent relative read depths; connecting lines signify number of reads matching splicing profile. Asterisks denote skipped exon pairs. Ensembl annotations included for reference.