SpliceCenter: a suite of web-based bioinformatics applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Michael C. Ryan
Master of Science
George Mason University, 2006
Bachelor of Science
The College of William and Mary, 1988

Director: Dr. John Weinstein, Graduate Faculty
Department of Bioinformatics and Computational Biology

Spring Semester 2009
George Mason University
Fairfax, VA

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

SPLICECENTER: A SUITE OF WEB-BASED BIOINFORMATICS APPLICATIONS FOR EVALUATING THE IMPACT OF ALTERNATIVE SPLICING ON RT-PCR, RNAI, MICROARRAY, AND PEPTIDE-BASED STUDIES.

Michael Ryan, PhD

George Mason University, 2009

Dissertation Director: Dr. John Weinstein

Alternative splicing of gene transcripts presents biologists with challenges that have largely been ignored to-date. Assays commonly used by researchers to explore biological processes including qRT-PCR, RNAi, expression microarrays, and peptide-based technologies target sequence fragments to quantify or silence gene expression. Alternate splicing may cause failures for these assays and complicates interpretation of resulting data particularly when correlation between platforms is attempted. For example, qRT-PCR may fail to validate expression microarray results if the PCR primers and microarray probes target different splice variants. The purpose of this project is to implement a suite of tools that assist biologists in quickly and easily identifying the impact of alternative splicing on commonly used laboratory techniques. It is hoped that these tools will lead to more effective use of these technologies and to more insightful interpretation of resulting data.

1. Introduction and Previous Work

Technologies commonly used by biologists to investigate gene function include quantitative RT-PCR (qRT-PCR)[1], RNA interference (RNAi)[2], gene expression microarrays[3], and antibody-based protein assays[4]. In the context of gene expression analysis, RT-PCR involves the use of RNA primers and polymerase to amplify and quantify a specific mRNA sequence. RNA interference is used to inhibit expression of a target gene through the introduction of small interfering RNAs (siRNAs) or short hairpin RNAs (shRNAs) that invoke natural mechanisms in a cell to degrade a gene's mRNAs. Expression microarrays contain single stranded DNA sequences that hybridize to complementary, labeled mRNA sequences allowing concurrent measurement of expression levels for thousands of genes. Finally, sequential antibodies bind target regions of a protein in order to detect and quantify the protein. Each of these techniques relies on targeting a small fragment of nucleic or amino acid sequence that is uniquely related to a specific gene. Alternative splicing may remove the target sequence from some transcripts, complicating the selection of sequence targets and interpretation resulting assay data.

In eukaryotic cells, messenger RNA is modified in the nucleus by a macromolecular complex called the spliceosome. The spliceosome consists of several small nuclear RNA proteins (snRNPs) named U1, U2, U4, U5, and U6 along with other

associated proteins. After a gene is transcribed, the spliceosome removes intron regions of the transcript leaving only exon segments in the mature mRNA that is eventually translated into a protein. This editing of the transcript sequence prior to translation is a biologically regulated process that may be varied to produce different mRNA transcript sequences from a single gene. The variation in splicing of mRNA is called alternative splicing and the resulting transcripts are called alternate splice forms or splice variants. **Figure 1** provides an example of tissue specific alternative splicing of the calcitonin-related polypeptide alpha gene.

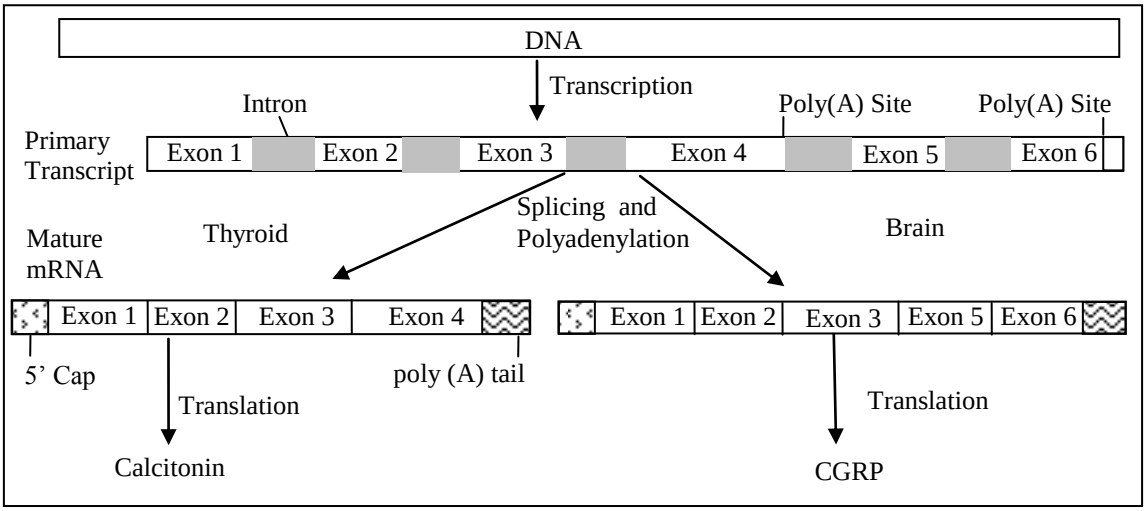


Figure 1: Example of alternate splicing of the calcitonin-related polypeptide alpha gene. In thyroid tissue, exon 4 is included in the mature transcript causing polyadenylation to occur at the end of exon 4. In brain tissue, exon 4 is splice out so exon 5 and 6 are included in the mature mRNA. The thyroid version of the transcript is translated into the calcitonin hormone while the brain version is translated into the signaling peptide, CGRP[5].

The removal of introns by the spliceosome is directed by splice site sequence motifs and branch point signals. The canonical splice signals include a 5' intron GU

sequence, an adenosine branch point near a poly pyrimidine sequence, and a 3' intron AG sequence (**Figure 2**). The U1 snRNP binds the 5' end of introns and the U2 snRNP binds the branch point near the 3' end of introns. The rest of the splicosome assembles around these snRNPs, cleaves the intron, and binds the exons. In addition to the canonical splice signals, there are a variety of non-canonical splice signals one of which is recognized by alternate U11/U12 snRNPs.

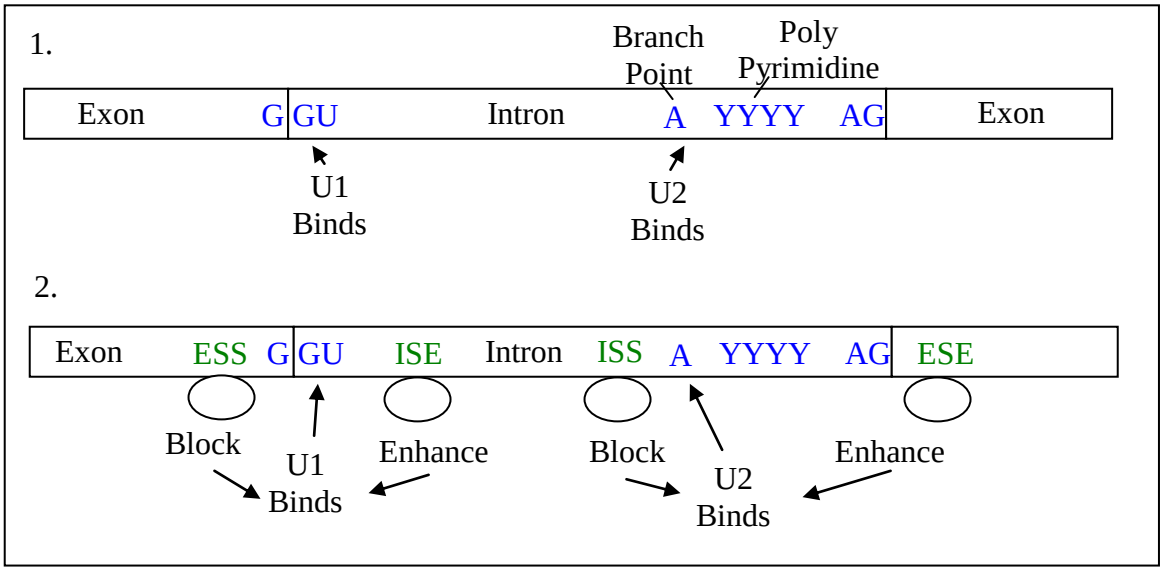


Figure 2: 1. Canonical splice signals and 1. Exon splice silencer (ESS), exon splice enhancer (ESE), intron splice silencer (ISS) and intron splice enhancer (ISE)

Regulation of alternative splicing is not completely understood but splicing enhancer and splicing silencer elements have been discovered that operate in a manner similar to transcription factors to direct alternative splicing. Splicing enhancers and silencers can occur in either introns or exons. Splicing factor proteins bind enhancer/silencer elements to facilitate or inhibit binding of U1 / U2. Many different

types of alternative splicing patterns exist including alternate promoters, cassette exons, retained introns, alternate donor site, alternate acceptor site, and alternate polyadenylation sites (**Figure 3**).

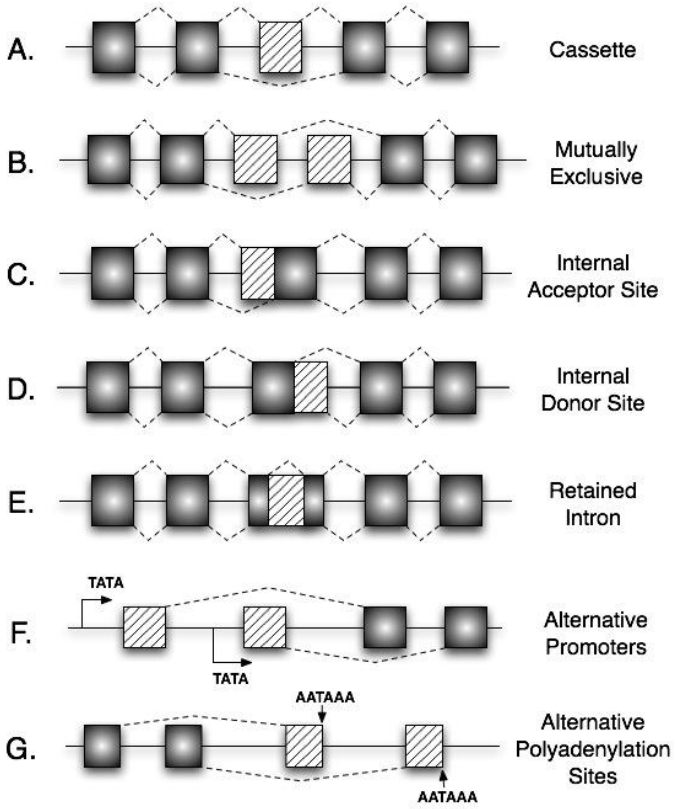


Figure 3: Types of alternative splicing.

More than 60% of protein-coding genes in vertebrates exhibit splice variation [6, 7]. Alternate splice forms have been associated with tissue-specific gene functions, developmental processes, and disease states (notably cancer)[8, 9]. Genes with splice variation in the coding region produce different proteins that may vary in form and function. Transcripts with variation in the untranslated regions (UTRs) may be

differentially regulated and therefore exhibit differences in spatial or temporal expression patterns. Splice variation is more prevalent in loop regions of proteins than core regions suggesting a role for alternative splicing in regulation of protein-protein interactions [10]. The prevalence and biological significance of alternative splice forms suggest that an accurate study of gene function must include specification of not just gene expression but rather the splice variant expression. (See Blencowe Review for additional background on the role of alternative splicing.)

Of particular interest to many biological researchers is the prevalence of alternative splicing in disease. More than 15% of heritable human diseases have now been associated with mutations in splice sites or splice regulatory elements[11]. For many intensively studies genes, more than 50% of disease causing mutations alter mRNA splicing[12]. Diseases including: cystic fibrosis, Fabry disease, Frasier syndrome, retinitis pigmentosa, spinal muscular atrophy, and myotonic dystrophy are caused by splicing mutations. Taupathies are a family of neurodegenerative disorders caused by mutations in the tau gene (MAPT). Mutations that affect the inclusion rate of a cassette exon in tau transcripts alter the balance of microtubule binding sites leading to the formation of filamentous aggregates leading to disease. Interestingly, even synonymous mutations like L284L cause the disease because a splicing repressor element is weakened[13].

Cancer has also been associated with alternative splicing in recent research. Many studies have reported tumor-specific splice variants and others have reported changes in expression of splicing regulatory factors that are associated with progression

to malignancy [14-16]. For example, the transmembrane protein encoded by the SVH gene is upregulated in liver cancer but only the SVH-B splice variants of this gene is upregulated. This was the only variant to cause tumor when introduced into mice and antisense inhibition of SVH-B cause apoptosis in hepatoma cells[14]. Large-scale bioinformatic studies of EST sequences with RT-PCR confirmation have identified hundreds of genes that demonstrate alternate splice variant expression in cancer cells when compared to normal tissue[16].

1.1 Problem Statement

Alternative splicing poses technical challenges to molecular biologists using sequence-based technologies including RT-PCR, RNAi, expression microarrays, and peptide based assays. Researchers, even those not directly interested in alternative splicing, need easy-to-use bioinformatics tools that identify the impact of alternative splicing on these technologies. The prevalence of alternative spliceforms suggests many questions that may confront biologists who employ oligo or peptide-based assays, for example:

- Which specific splice variants are targeted by my assay? What other splice variants exist?
- Do the RT-PCR primers/probe that I plan to use to validate microarray expression results target the same splice variants that were targeted by the microarray platform?

- Did an siRNA fail to mediate RNAi silencing of a gene because it did not target the dominant splice variant in my sample?
- Are there known splice forms of my gene of interest with variation occurring in the protein coding portion of the transcript? Does the sequential antibody that I plan to use target all potential protein products?
- Where could I place RT PCR primers to target all splice variants? Where could I place RT PCR primers to amplify one specific splice variant?
- Do expression values from one microarray fail to correlate with values from another microarray because the probesets target different splice variants?

In addition to the technical reasons for evaluating the impact of splice variants, the biological role of alternate splicing in tissue specific functions, developmental processes, and disease states suggests that sequence-based assay results should be interpreted at the splice variant level. For example, most microarray or RT-PCR studies of gene expression are reported in terms of up or down regulated genes but a more accurate description of the results would include identification of the specific splice variants that were differentially expressed. Again, a set of bioinformatics tools are required to support molecular biologists in understanding the breadth of splice variants and to identify the specific variants targeted by an assay.

1.2 Specific Goals

The goals of this project are as follows:

- Create a unique set of bioinformatics applications and splice variant data that assists molecular biologists with identification and analysis of the impacts of alternative splicing on their research. Include human, mouse, and rat genomes.
- Develop accessible, intuitive, and user friendly applications to encourage adoption. The daunting biological complexity and voluminous data that researchers are currently facing creates a need for tools that deliver critical information but also have minimal learning curves.
- Collaborate with research teams using RT-PCR, expression microarrays, siRNAs, and peptide-based assays to vet the applications and asses their ability to deliver technical or biological insights.

1.3 Previous Work

1.3.1 Other Splicing Databases

Several publicly-accessible websites provide data and utilities for investigating alternative splicing: AceView [17], Alternative Splicing Annotation Project database (ASAP II) [18], Alternative Splicing and Transcript Diversity Database (ATSD) [19], Friendly Alternative Splicing and Transcript Database (fastdb2)[20], Hollywood [21], Eukaryotic Splice Database (EUSplice) [22], and Genome Annotation for Alternative Splicing (ECgene) [23], and Splicy [24]. Each of these databases provides information on transcript variants and in some cases additional annotation including splicing regulatory elements, exon characteristics (*e.g.* alternative, constitutive, retained, internal),

and EST evidence/tissue of origin data. Although these sites are excellent resources for alternative splicing data, it is difficult to query the data in a way that would assist investigators to understand the impact of alternative splicing on sequence-based assays. For example, current databases do not provide a method for submitting probe sequence queries to identify splice variants that will / will not be targeted by the probe. A comparison of SpliceCenter and other alternative splicing resources is summarized in **Table 1**.

Table 1: Comparison of SpliceCenter capabilities to other splicing databases.
(✓ indicates comparable feature, * indicates partially analogous feature)

SpliceCenter Capability	AceView	ASAP	ATSD	fastdb2	Hollywood	EUSplice	ECgene	Splicy
General								
Simple, Intuitive Interface						✓		
Help and Sample Queries	✓	✓	*	*	*	✓	✓	*
Splice Variants								
Graphical display of gene's splice variants	✓	✓	✓	✓	✓	✓	✓	✓
Identifies coding regions	✓		✓			✓		
Identifies NMD targets				✓				
Human, Mouse, Rat, Worm, Fly, Cow, Rice, Zebra Fish, Arabidopsis	*	*+	*	*	*	✓	*+	*
PCR								
Primer sequence query showing position in variants				✓			✓	
Batch high-throughput primer query								
siRNA								
siRNA sequence query showing position in variants								
Batch high-throughput hit/ miss report for siRNAs								
Microarray								
Display pre-computed target positions of Affymetrix, Agilent, Illumina, and ExonHit probesets in gene's variants								*
Display integrated graphic of microarray probe targets and primer target positions.								
Batch high-throughput query of pre-computed probe / probeset target positions								*
Peptide								
Peptide sequence query displaying source coding region in splice variants								
Batch high-throughput peptide query								

1.3.2 SpliceMiner

SpliceCenter was developed using the foundation of previous work completed for my MS thesis and Ari Kahn's PhD. dissertation. The previous research lead to the development of the SpliceMiner application and the Evidence Viewer Database (EVDB) [25]. The splice variant database structure and microarray probe assignment features from the previous work were reused in SpliceCenter. However, SpliceCenter provides a substantial advance over our previous work in terms of the database content, website utilities, and potential user base. The database was rebuilt and extended to include the latest genomic and transcript data, mouse and rat genomes, position of coding regions, identification of NMD targets, protein sequence, and pre-computed microarray probe targets. The focus of SpliceCenter utilities on contemporary biological technologies expands the target user base to include biologists using RT-PCR, RNAi, microarrays, or peptide-based assays. The "niche" nature of the SpliceCenter utilities provide significant ease-of-use and time savings advantages when compared to our previous general purpose SpliceMiner utilities. To illustrate this point, the following use case comparison of the time and steps required to perform a verification of RT-PCR primers in both tools is provided (**Tables 2 & 3**). The purpose of the use case is to ensure that RT-PCR primers are targeting the same splice variants as a microarray probeset.

Table 2: Task Steps in SpliceMiner (previous work)

Step	Description	Time (Minutes)
1	Obtain probe sequences each probe in the probeset from the Affymetrix site. Either download the full sequence file and extract or use NetAFFX queries. (Time depends on familiarity with their site. Most other vendors have a way to get probe sequence but it is not always easy to find)	15 - 60
2	Obtain RT-PCR primer sequences. This is easy if the primers are custom. If not, use vendor materials to get primer sequence or reference sequence.	5
3	Format a batch FASTA file with microarray probe and RT-PCR sequences. Label each primer/probe sequence uniquely.	10
4	Submit the FASTA query file to SpliceMiner's batch sequence query page (interactive page only searches on one sequence at a time so it would be difficult to integrate all of the results)	2
5	Wait for batch results	1
6	Become familiar with SpliceMiner tabular output results. Manually build a list of the variants targeted by the probes (there will be one result line per probe per variant targeted).	15
7	Determine the variants targeted by **BOTH** the PCR primers from the tabular output. Note: SpliceMiner will only be able to match primers > 21 nts long – a new sequence matching engine was added to SpliceCenter for short sequence queries.	10
8	Using the results of steps 6 and 7, determine if the primers and probes are targeting the same variants. Use the interactive SpliceMiner gene symbol query to get a feel for the splice variant topology to understand the results	10
	Total Time:	68-113 minutes

Table 3: Task Steps in SpliceCenter

Step	Description	Time (Minutes)
1	Obtain RT-PCR primer sequences. This is easy if the primers are custom. If not, use vendor materials to get primer sequence or reference sequence.	5
2	Use the interactive Primer-Check application. Enter primer sequences and select the check box for the Affymetrix U133A microarray	1
3	Look at graphical results to see which variants are targeted by the primers and which are targeted by the probeset. Use the dropdown filter to select individual probesets if necessary.	1
	Total Time:	6 minutes

1.3.3 Microarray Analysis Software

New expression microarray platforms have been developed to investigate alternate splicing of mRNA transcripts. Typical gene expression microarrays contain 1 to 16 oligonucleotide probes per gene and often are biased in targeting toward the 3' portion of the transcript. Microarrays designed for investigation of alternate splicing contain 10s to 100s of probes per gene evenly spread along the transcript. Common commercial microarrays used to detect alternative splicing include Affymetrix's Exon 1.0 ST arrays and ExonHit's Whole Genome-Wide SpliceArray.

Many software packages are available that analyze of data from splicing microarrays to identify differences in transcript splicing between samples (e.g. the drug treated cell line is splicing out exon 3 of HYPK transcripts but untreated cells include exon 3). It would be optimal if this analysis was able to identify changes in expression from one transcript isoform to another. The DECONV program[26] attempts to do this by quantifying expression of known isoforms using maximum likelihood methods. There are, however, two issues with identifying splicing changes at the full transcript level. First, our knowledge of splice forms is not complete and novel splice forms may be the key to a biological question of interest. Second, cells often contain a complex mixture of several different transcript variants of a gene and the response to treatment may be a subtle shift in the ratio of alternative splice forms.

Because of these difficulties, the majority of splice microarray software focuses on splice event detection rather than splice isoform changes (e.g. there is an increase in skipping exon 3 or use of an alternate promoter or use of a different poly(A) site). The

Affymetrix Exon array includes small probesets (~4 probes) for each exon or alternatively spliced portion of an exon. The common technique used to identify alternate splicing of exons is to look for changes in exon expression that differ general gene expression (e.g. the treated sample showed slightly increase expression of ACP1 but exon 2 showed significantly decrease expression indicating an exon skip event). This is done by computing a Splicing Index (SI) for each exon[27]. The SI is the ratio of exon expression to that of the whole gene.

Software packages that seek to identify splice events differ in the statistical methods and models used to normalize probe values and identify significant splice events. The Affymetrix Expression Console uses Probe Logarithmic Intensity Error (PLIER) to determine gene and exon expression and pattern-based correlation (PAC) or microarray detection of alternative splicing (MIDAS) to identify alternatively spliced exons[28]. Other products use Robust Multichip Analysis (RMA) to calculate normalized gene and exon (probeset) expression and use Analysis of Variance (ANOVA) (e.g. Partek) or Iteratively Reweighted Least Squares (IRLS) (e.g. FIRMA) to fit models with terms for change in exon expression [29, 30]. None of these packages provide the visualization and dynamic data exploration capabilities of the SpliceCenter and ArrayDataViewer applications discussed in future chapters.

2. SpliceCenter Databases

SpliceCenter data is stored in two MySQL relational databases: Gene, and Microarray. The Gene database contains the splice variants for the known genes of several organisms. The splice variant data in the Gene database includes the exon structure of each variant mapped to genomic coordinates. The Microarray database contains probe target locations for common commercial microarray platforms in genomic coordinates.

2.1 Gene Database

The Gene database is a hierarchical, normalized representation of gene structures (**Figure 4**). Genes contain Variants, Variants contain Exons, and Exons have ExonPositions. Genes are identified by the unique GenBank gene ID and separate tables identify symbols and aliases related to a gene. Nucleic and protein sequence for each variant is also stored in the database. Finally, a consensus splice model of each gene is stored in the SubExon table.

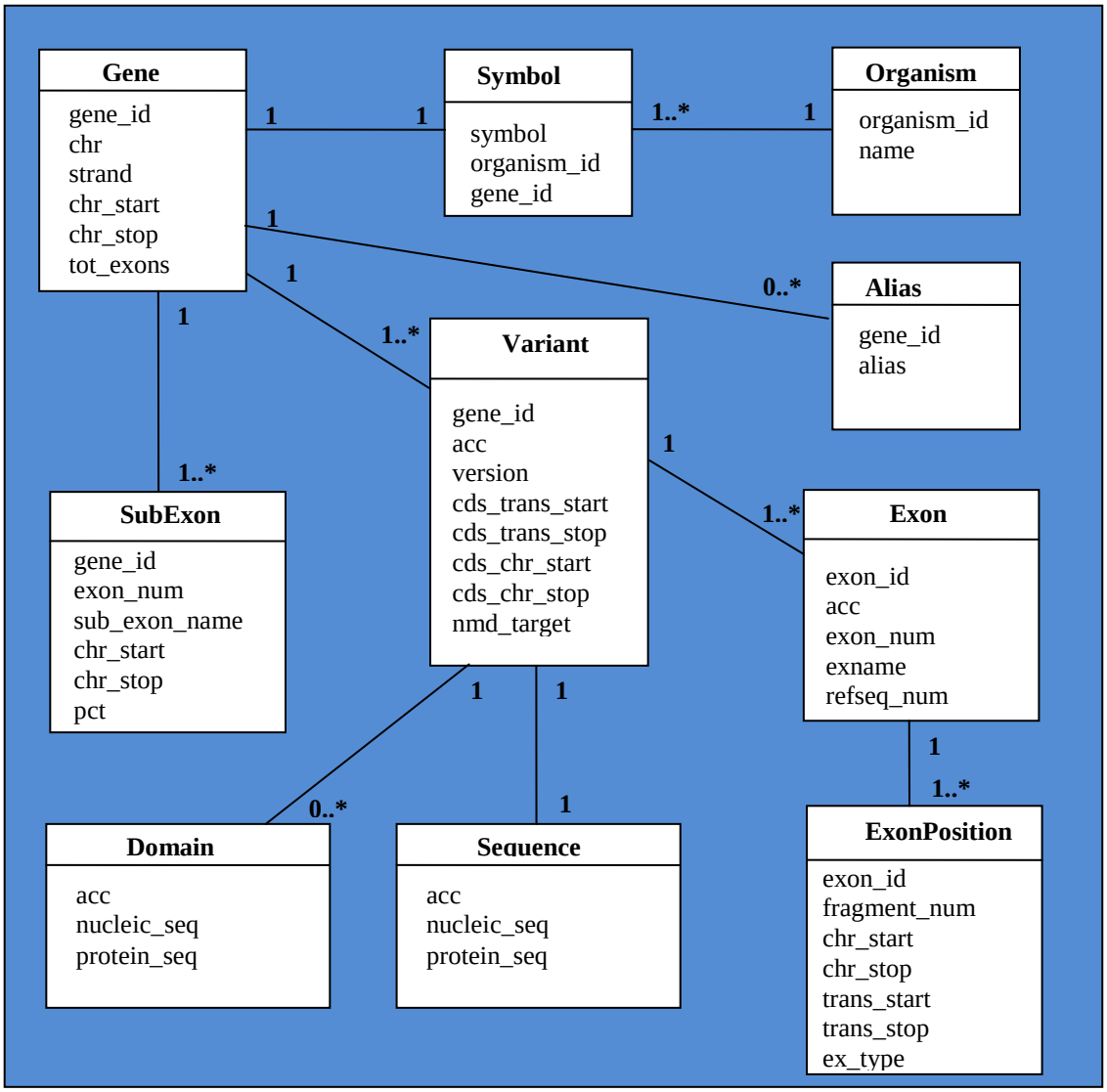


Figure 4: UML Schema diagram of the Gene database.

2.2 Precursor databases

The work of Ari Kahn, Barry Zeeberg, and Hongfang Liu laid the foundation for the current splice variant database build process. Dr. Kahn and Zeeberg developed a splice variant database, EVDB [25], which was based on NCBI Evidence Viewer [31] data. EVDB demonstrated the merits of constructing a repository of distinct splice

variants that included each variant's exon structure in chromosomal coordinates. It also identified gene families originating from alternate promoters of a single gene. The EVDB build process ultimately became very difficult to maintain due to its dependence unsupported NCBI files and spidering of NCBI websites. Subsequently, Dr. Liu and Zeeberg constructed a unique transcript database for AffyProbeMiner [32]. The techniques for obtaining and aligning transcripts developed for the AffyProbeMiner build and its methods of assuring transcript quality have been incorporated into the current SpliceCenter database build process.

2.3 Gene database build process

The Gene database is constructed by an automated build process written in Java. All Full-length mRNA transcripts from RefSeq[33] and GenBank[34] are aligned to the source organism's chromosomal sequence with BLAT[35] to determine the exon structure of expressed genes. Large gaps in the alignment of transcript to chromosome indicate the exon boundary positions in the transcript. This technique is applied to all of the transcripts of a gene in order to identify the unique splice variants of the gene.

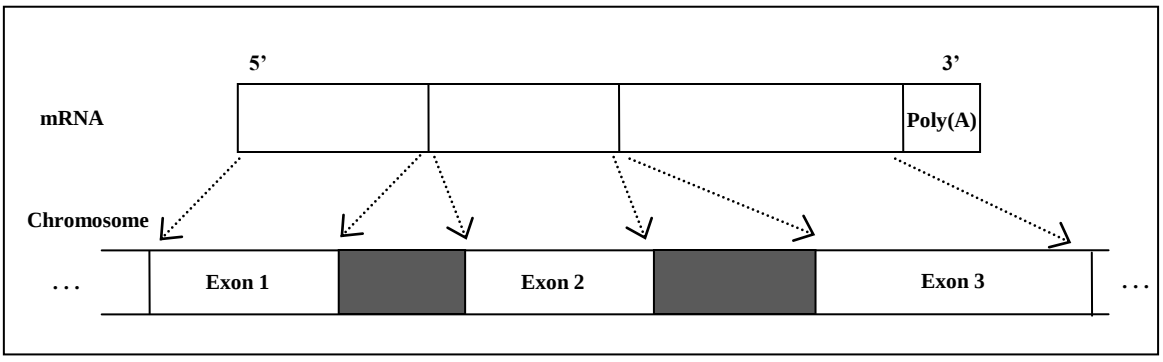


Figure 5: Exon Structure from Transcript to Chromosome Alignment

The GeneBuild program collects data from several external sources and performs a series of data processing steps to produce the Gene database. The following sections describe the processing steps performed by GeneBuild.

GeneBuild Step 1 – Setup and Gather External Data

A runtime parameter to the GeneBuild process indicates the organism for which the build should be performed. Much of the configuration used by the build process (ftp sites, directories, QA thresholds, gap sizes, etc.) is read from an external properties file to enable easy changes to build options. Next, the program clears and creates working storage directories and then retrieves external data needed for the build. The location of external data is configurable. Currently, all external data is obtained from ftp.ncbi.nih.gov. Chromosomal sequence is obtained from /genomes/<organism>/Assembled_chromosomes. RefSeq mRNA transcripts are retrieved from /refseq/release/complete (note: this directory contains all mRNA transcripts and is not organism specific). GenBank transcripts are retrieved from /genbank with an organism specific file prefix (e.g. pri for H_sapiens). This data is maintained in gzip format in local storage. The ftp download time is a significant portion of build processing time so timestamps / sizes of files are preserved and checked prior to download. Only update files will be downloaded.

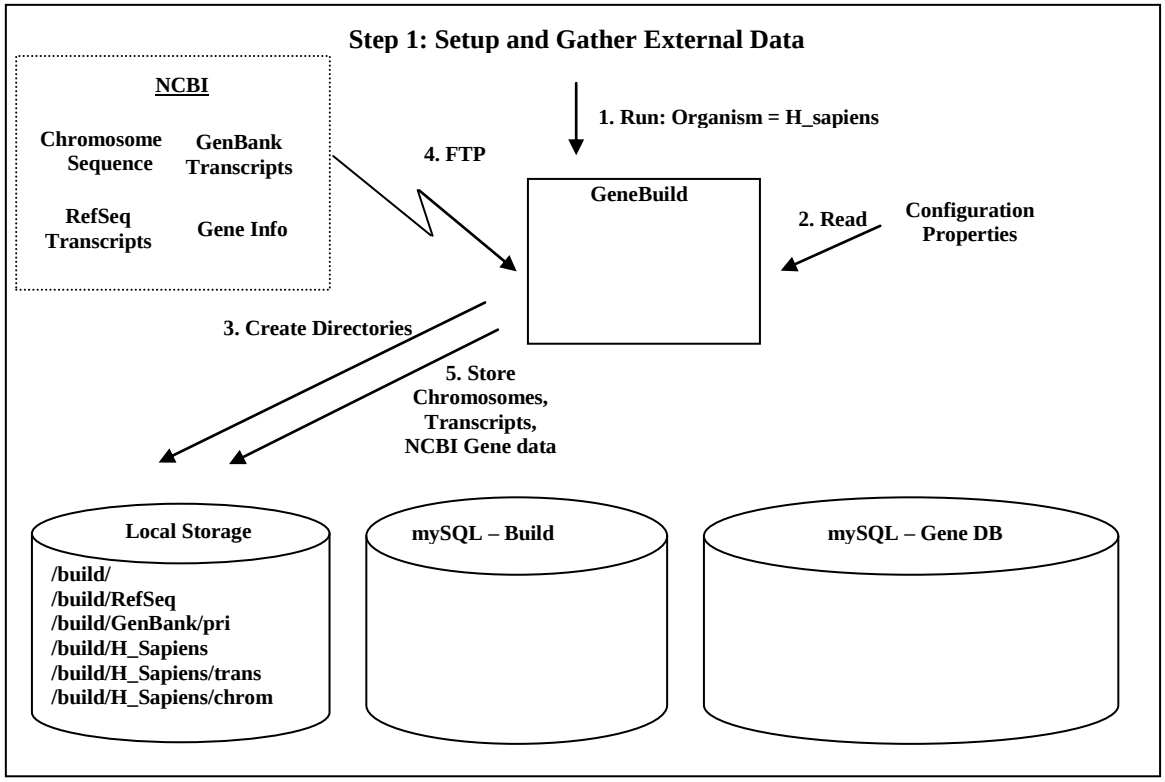


Figure 6: GeneBuild Step 1

In addition to chromosome and transcript data, the GeneBuild program obtains several NCBI Gene data files to assist with transcript assignment to genes. The gene_info.gz and gene2accession.gz are retrieved from /gene/DATA/ on the NCBI ftp server. The gene2accession file can be used to map transcript accession numbers to NCBI gene ids. The gene_info file is used to get the symbol, aliases, and chromosomal location of the genes. Finally, the mapview file seq_gene.md.gz is retrieved from /genomes/MapView/Homo_sapiens/sequence/current/initial_release. The mapview file is used in to locate a preferred location for transcripts that align equally well to multiple chromosomal locations. This is uncommon but important in constructing a coherent model for some genes.

GeneBuild Step 2 – Load Gene Information

The build process creates a set of temporary tables that are used for the build process but are not part of the production Gene database. These tables are loaded from the data obtained from NCBI Gene. Data for the target build organism is extracted from the gene_info file to construct a gene to chromosome mapping in the gene2chr table. The gene2accession file is also parsed to find transcript accession to gene id mapping for the target build organism and are stored in gene2acc. Finally, the mapview file for the target organism is parsed to obtain "GENE" records that indicate chromosomal locations for genes. Note: some organisms have multiple GENE records per gene so the appropriate source must be selected to match the NCBI genome build (e.g. "reference" for human rather than Celera entries).

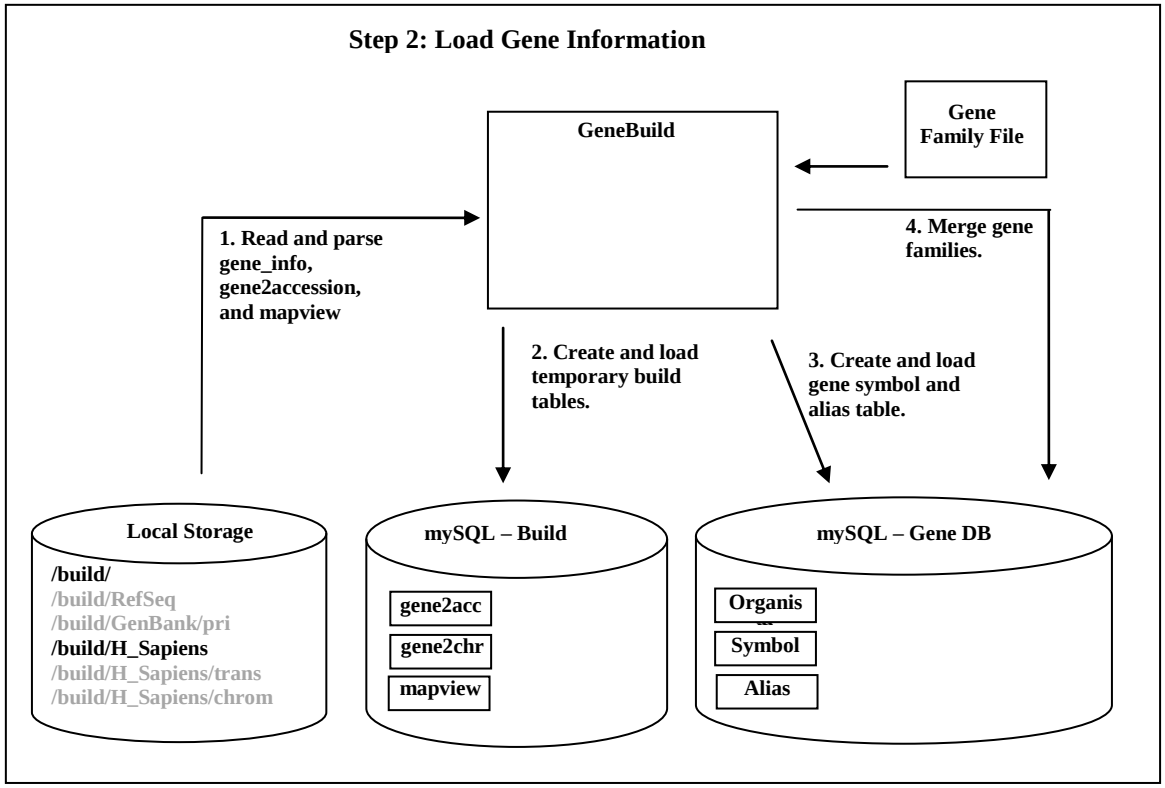


Figure 7: GeneBuild Step 2

After the temporary tables are created, some of the production data tables are populated using the gene_info file. A descriptive entry for this organism is added to the Organism table. The Symbol table is loaded with the mapping of gene symbol (e.g. HGNC symbol for human genes) to NCBI gene id (Entrez gene id) mapping. The Alias table is loaded with all alias symbols in the gene_info file and the Ensemble symbol.

Finally, a reorganization of a few genes is performed using a manually created Gene Family file. NCBI Gene treats a few gene families as separate genes that we feel are actually splice variants of a single gene. For example, PCDHA1 through PCDHA13 are the product of alternate splicing of a single gene. These are identified via a report produced by the build process that looks for genes that overlap. A manual curation is

then performed to investigate and identify gene families that should be merged into a single gene model. When these are identified, they are added to the Gene Family file and this step of the build process alters the gene2acc mapping to merge the associated transcripts into a single gene. All previous symbols associated with splice variants are added as aliases to the merged gene family.

GeneBuild Step 3 – Extract, Parse, Filter Transcripts

The RefSeq and GenBank sequence data gathered by the build process contains a variety of records. The goal of the build process is to identify alternative splice forms by aligning transcripts to the genome. To achieve this goal, it is necessary to select only mature mRNA transcripts with good sequence quality.

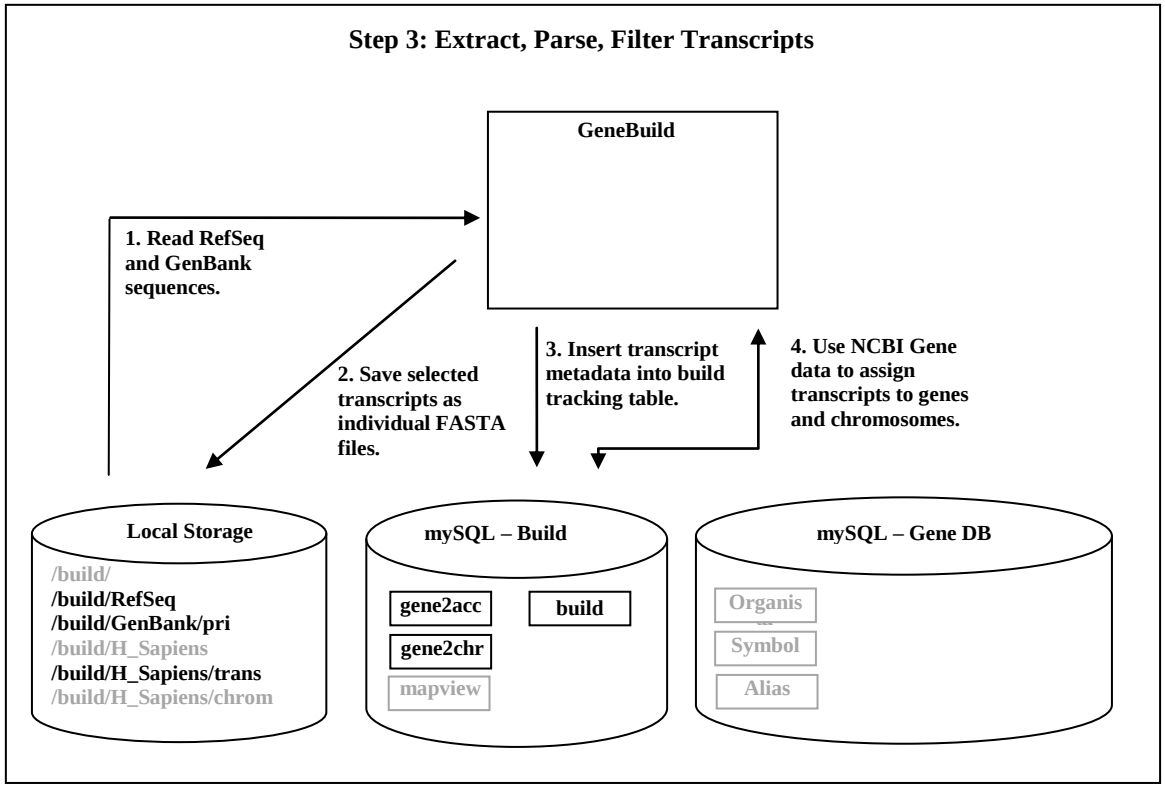


Figure 8: GeneBuild Step 3

RefSeq RNA sequences in GenBank format are stored as gzip files in the local /build/RefSeq directory. The build process uses the Java GZIPInputStream class to read directly from the compressed files. The Open Source BioJava classes are used to parse GenBank format sequence files. In order for a sequence to be accepted for further processing by GeneBuild, the sequence must meet the following criterion:

- The ORGANISM tag must identify the sequence as belonging to the target organism.
- The TYPE tag must identify the sequence as mRNA.

- For RefSeq, the sequence accession must start with ‘NM_’. We do not use predicted or model sequences.

- For GenBank, the DESCRIPTION tag must identify the sequence as ‘complete sequence’, ‘complete CDS’, or ‘complete coding sequence’. Transcripts without one of these descriptions are accepted if they contain a CDS section that identifies a coding start/stop position. In this case, the sequence must also contain canonical start/stop codons at the indicated positions.

Accepted sequences are written to the /build/<Organism>/trans director as individual FASTA files. In addition, an entry is made in the temporary build table for each accepted transcript. This table contains metadata extracted from the sequence file (Accession, Version, Start/Stop Coding Position, Protein Translation, Poly(A) sites, etc.) and is used to track that status of each transcript as build processing progresses. Finally, the gene information from NCBI Gene sources gathered in Step 2 is used to associate each transcript with a gene and to identify the chromosomal location of the gene.

GeneBuild Step 4 – Align Transcripts to the Genome

The exon structure and splicing variation of genes is identified by using BLAT to align transcripts to chromosomal sequence. GeneBuild queries the build table to create a list of all of the transcripts that originate on a given chromosome, grouped by gene. These transcripts are placed in a single query file and aligned to the source chromosome with the BLAT program. Transcript quality standards are imposed at this step by

requiring a 99% match of transcript to chromosome. The alignment must also include 95% of the transcript.

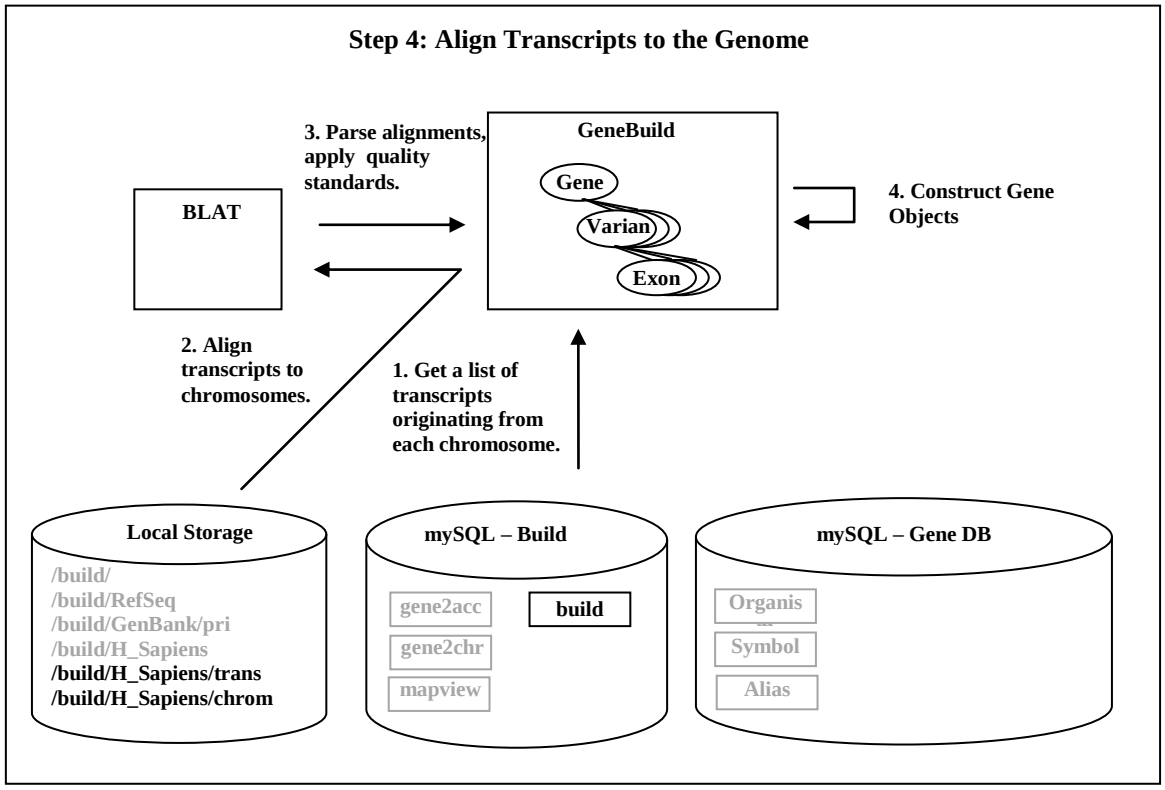


Figure 9: GeneBuild Step 4

BLAT results are parsed and the best hit for each transcript is selected. The algorithm that selects the best BLAT alignment for each transcript considers many factors. In general the longest alignment with the least gaps / indels is the best alignment. In some cases, however, the algorithm must choose between alignments with roughly equal lengths / mismatch / gap characteristics. In these cases, the algorithm considers the nominal gene location specified by NCBI Gene MapView and the location of other

transcripts for the same gene. The algorithm attempts to construct a gene model where all splice variants are mapped to the same general genomic location.

The GeneBuild process uses the transcript alignments to construct an internal object-oriented representation of a gene. Each gene object contains a collection of variant objects which in turn contain collections of exon objects. BLAT alignment blocks are parsed to find the start / stop position of exons in both transcript and genomic coordinates. Generally each block represents an exon but occasionally single base mismatches or small indels will create blocks so a minimum chromosomal distance between blocks is required for exons. The start position for each exon is specified in the qStarts and tStarts columns of the BLAT output. The end positions are found by adding the related block length value to the start position. Here is an example of BLAT block length, qStarts, and tStarts values for three transcripts aligned to human chromosome 2:

Transcript		Block_length	qStarts	tStarts
NM_004300	+	139,74,114,62..	0,139,213,327,..	254868,261865,262036,..
NM_007099	+	139,74,114,62..	0,139,213,327,. .	254868,261865,262191,265139, . .
NM_001040649	+	139,74,445,	0,139,213,	254868,261865,262036,
NM_004322	-	510,191,523,	16,526,717,	63793875,63795660,63808229,

BLAT provides 0 based alignments so chromosomal and transcript coordinates must be incremented by 1. Also negative strand alignments are tricky. The block length and tStarts lists must be reversed because the last element in the list is the first exon. Also, the "starts" will actually be the genomic end position of the exon and the start position will be the tStart + block length. For NM_004322 above, the first exon starts at 63808752 (last tStart + last block length) and ends at 63808230 (last qStart +1 for 0 based adjustment).

GeneBuild Step 5 – Process and Save Variants

After all splice variants of a gene are added to gene object in the previous step, additional processing steps are performed on the gene object and then it is saved to the database. The additional processing steps include: exon numbering, sub-exon identification, conversion of coding position coordinates, selection of potential Nonsense Mediated Decay (NMD) targets, correction of missing UTR sequence, and removal of duplicate splice forms.

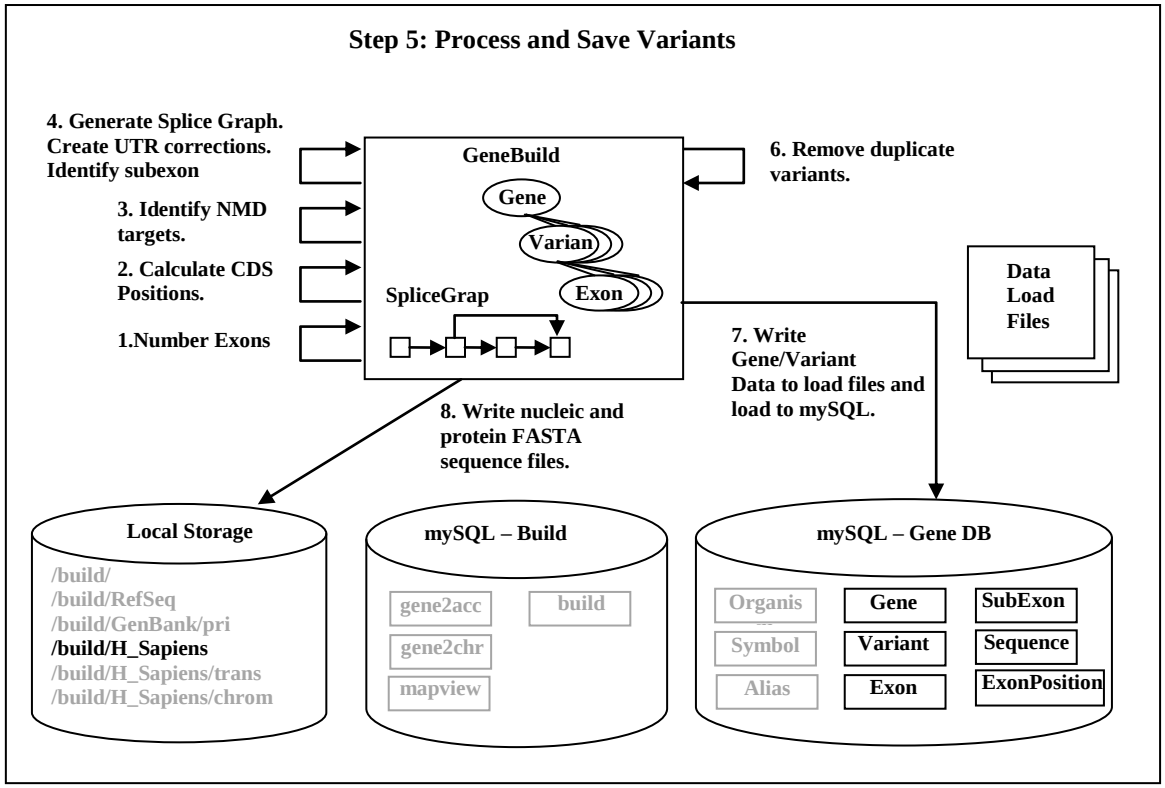


Figure 10: GeneBuild Step 5

The transcripts in GenBank often contain exons not found in RefSeq transcripts. For this reason, the Gene database cannot use RefSeq or exon numbering. After all

transcripts are added to a gene object, an exon number process is performed. Alternate promoter, poly(A), acceptor, or donor sites can create different exon isoforms so the exon numbering process must allow for variation in exon start position in length. The exon numbering algorithm sorts all exons by genomic location and numbers overlapping exons with the same exon number.

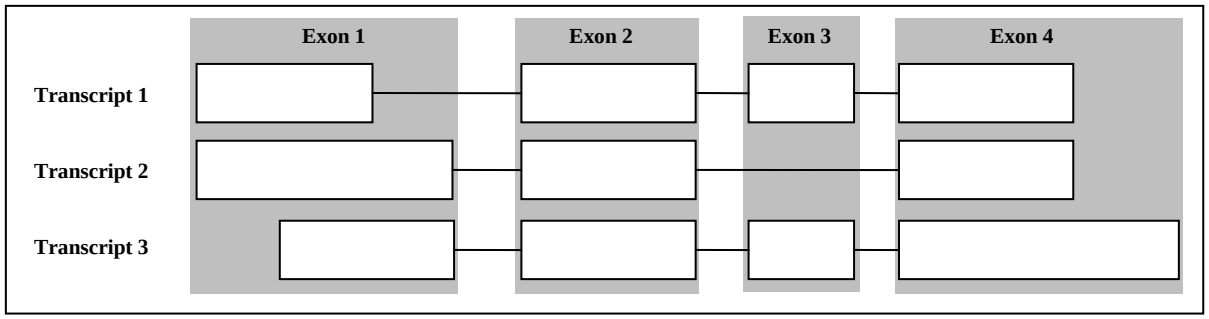


Figure 8: Numbering Exons – Transcript exons arranged by genomic location.

An important element of the Gene database is the start / stop position of the coding portion of transcripts. Alternate splicing in the coding portion of a variant will alter its protein product and changes in the UTR regions may alter transcript regulation. All position information in the Gene database must be converted into chromosomal coordinates to facilitate integration of various elements into a common reference frame for display and analysis. Coding start/stop positions are converted to genomic coordinates by locating the exon containing the start or stop position any applying the following formula:

Plus Strand:

genomic position = exon genomic start + (transcript coding start – exon transcript start)

Minus Strand:

genomic position = exon genomic start - (transcript coding start – exon transcript start)

Nonsense Mediated Decay(NMD) is a cellular process by which transcripts with early stop codons are targeted for quick degradation. Splice variants targeted for NMD will not produce a significant amount of protein product and so may be less important to some investigators splice variants that are not targeted for NMD. For this reason, SpliceCenter identifies and visually distinguishes predicted NMD targets. NMD targets are identified by reviewing each variant in the gene model. If the stop codon for the variant is more than 50 bases upstream from the last exon junction, the variant is labeled an NMD target. [36]

After NMD target identification, the build process constructs a splice graph from the gene. Splice graphs summarize the splicing patterns of a gene in a consolidated graph structure where each node is a unique exon and the weighted edges represent observed splices [37, 38].

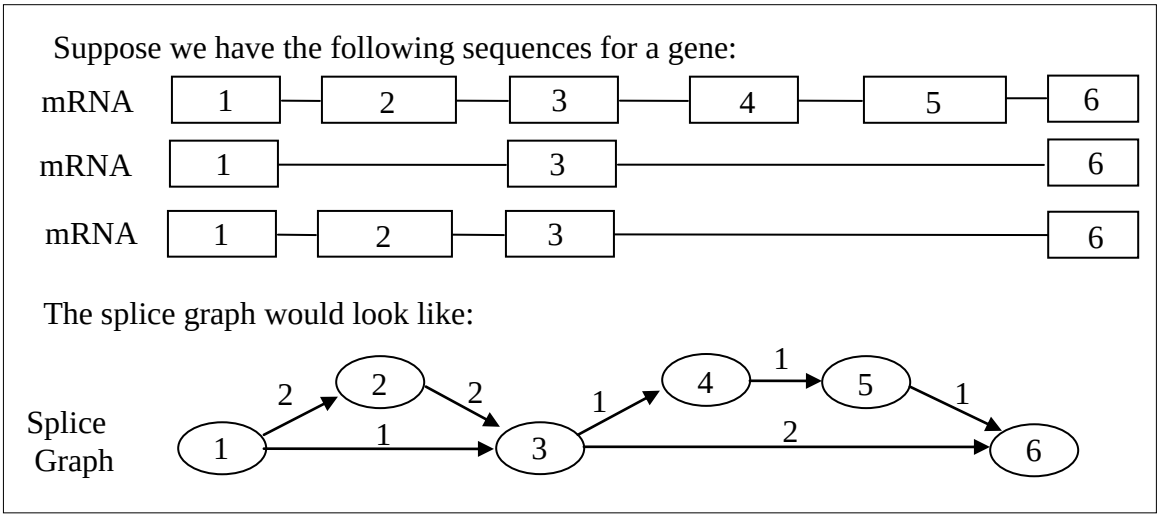


Figure 11: Gene Splice Graph

As the splice graph is constructed, exons are merged into existing nodes when possible. If a 3' splice, 5' splice, or poly(A) site of an exon conflicts with its related splice graph node, then a new node is created. In this way, the final splice graph contains the minimum set of distinct exon isoforms observed in the transcript evidence. This property of splice graphs is useful to the build process because it can be exploited to derive the sub exon structure of the gene and to correct UTR missing sequence issues.

Exons often have shorter or longer variants due to alternate promoters, alternate donor/acceptor sites or alternate poly(A) sites. A single consensus view of the exon composition of a gene including exon variation is useful for many tasks including accurate analysis of exon expression microarrays. The sub-exon map produced by the Gene build process meets this need.

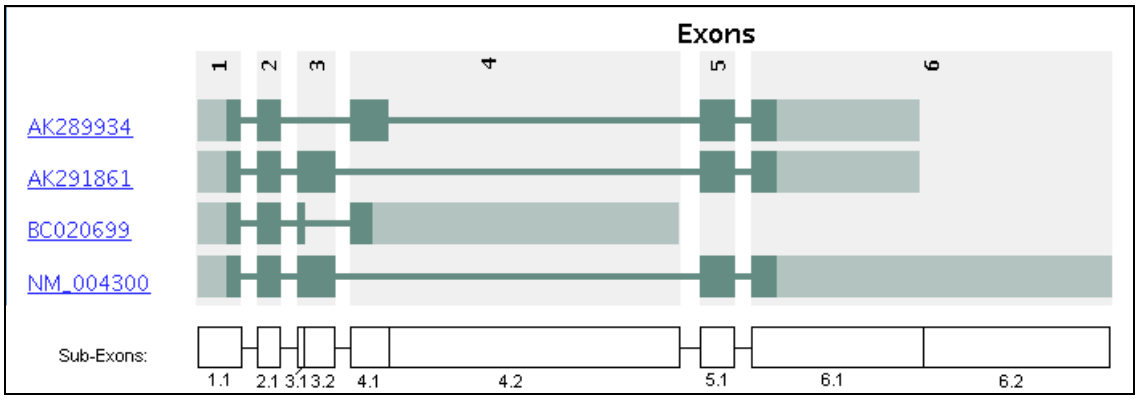


Figure 12: Sub-Exon Map

By scanning **Figure 12** vertically, one can see differences in exon structure. In this example, exons 1, 2, and 5 have no variation across the alternative splice forms. Exon 3 has a long version and a short version caused by an alternate donor site. Exon 4

has a normal size and an extended version including a stop codon and UTR region. Finally, exon 6 has a short and long version caused by alternate poly(A) sites. The sub-exon map shown below the variants presents a composite of observed exon isoforms and numbers them as <exon number>.<sub-exon number> (e.g. exon 6 has a sub-exon 6.1 and 6.2).

The build process generates a sub-exon map for each gene using the gene's splice graph. An iteration of the graph is performed to identify distinct start/end coordinates for each exon. This list of start/end sites is then used to create and number the composite sub-exon structure. The sub-exon data is written to a temporary data file for subsequent load into the MySQL SubExon table.

In addition to sub-exon map generation, the build uses the splice graph to generate corrections for missing UTR sequence. Transcripts included in the Gene database build contain the "complete coding sequence" but are not guaranteed to contain all of the 5' or 3' UTR sequence. For example, in the ACP1 gene shown in Figure 10, the variation seen in the start position of exon 1 is likely due to missing sequence. Similarly, there is a lot of variation in the 3' UTR in exon 6. Much of this is also due to missing sequence although there are some alternate poly(A) sites in this exon.

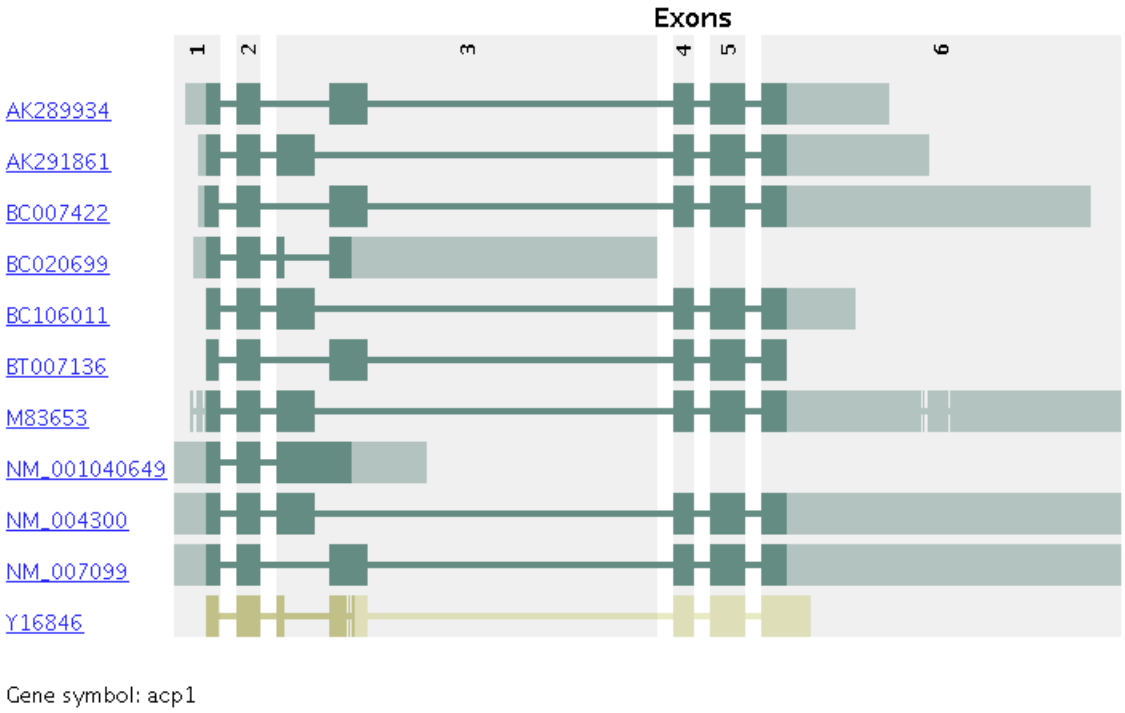


Figure 13: Missing UTR Sequence in ACP1 Variants

The splice graph contains a node for each exon isoform. The isoforms were constructed using observed splice sites and poly(A) sites. These exon isoforms are used by the build process to add "predicted UTR" regions. For example, if a transcripts 3' exon does not end at a poly(A) site, a predicted exon is created using the best fit isoform from the splice graph. The net effect of these predicted UTR segments is to smooth rough edges on the 5' end of transcripts and to extend the 3' end of transcripts to the nearest poly(A) site. These corrections are important because they improve the effectiveness of subsequent de-duplication processing and provide a better basis for determining probe targeting of variants.

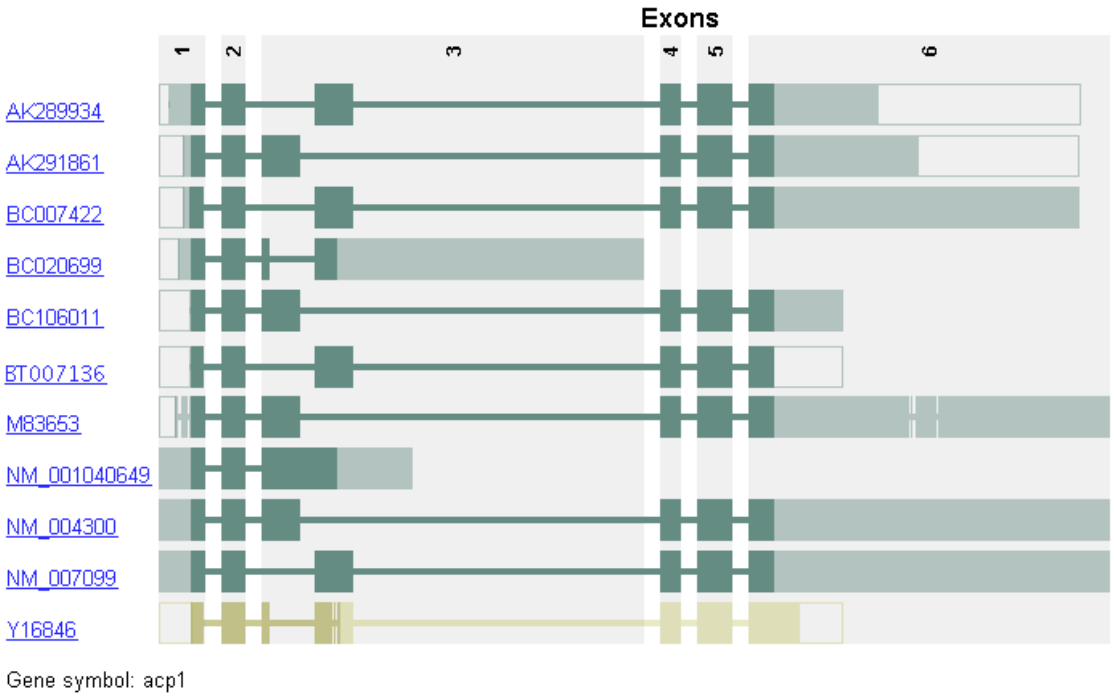


Figure 14: ACP1 with UTR corrections (drawn as hollow rectangles)

The final data processing step performed on each gene object is variant de-duplication. If two transcripts have an identical splicing pattern, one will be removed so that the Gene database will contain distinct splice variants. An exon by exon comparison of Variant objects is performed to find and eliminate duplicates. RefSeq sequences and sequences without mismatches are preferentially retained when deciding which variant to remove.

After all gene object processing steps have been completed, they are written to data load files for the Gene, Variant, Exon, and ExonPosition tables. When all BLAT results have been processed, the load files are imported into the MySQL database. Bulk file load into MySQL is much faster than iterative SQL inserts. In addition to loading splice variant data into MySQL, the build writes the nucleic sequence and protein

translation for each variant to FASTA formatted flat files. The FASTA files are used by SpliceCenter utilities to for sequence searches to identify probe target locations.

To complete the build, the BLAT faToTwoBit utility is run on the nucleic FASTA file to compress the sequence data for use by BLAT. Also, the Primer Match compress_seq utility is also run on the nucleic FASTA file to prepare a Primer Match database to support short oligo searches by SpliceCenter utilities.

GeneBuild Step 6 – Protein Domain Mapping

A key new feature of SpliceCenter is its ability to provide some insight into the impact of alternative splicing on a gene's protein products. One way this is done is by showing the portion of transcripts that are associated with Pfam domains of known function [39]. The GeneBuild application identifies Pfam domains in transcripts using the Position Specific Scoring Matrices provided by the Conserved Domain Database at NCBI [40]. The PSSMs represent the patterns identified by Hidden Markov Models developed at Pfam to identify protein domains. The BLAST utility formatrpsdb is used to create a database for RPS-BLAST from the PSSMs.

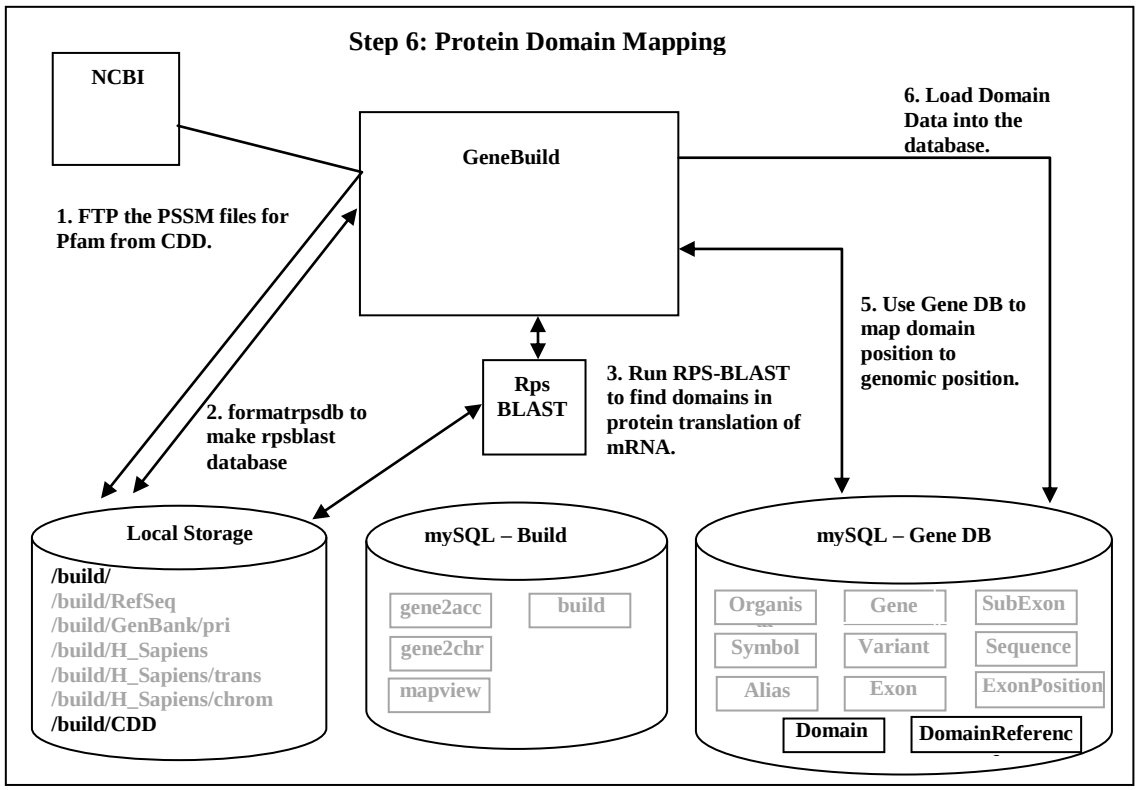


Figure 12: GeneBuild Step 6

The FASTA format file of transcript protein translations produced in the previous build step is searched using RPS-BLAST and the Pfam PSSM database to find protein domains. Currently, an e-value cutoff of .01 is used when looking for Pfam domain matches. The RPS-BLAST results are parsed to identify the highest scoring domain match for each region of the protein sequences (overlapping domains of lesser scores are discarded). RPS-BLAST alignments indicate the position of domains in the protein sequence of the splice variants. The Gene database is used to convert protein coordinates into transcript coordinates and then into genomic coordinates. Finally, the Domain data and DomainReference data with descriptions of the domains is loaded into the Gene database.

2.4 Microarray database

The Microarray database contains probe target locations for common expression microarrays. Identifying the genomic coordinates for probe targets is a computationally intensive process so SpliceCenter pre-computes these target locations. The data in the Microarray database is used by interactive and batch SpliceCenter utilities to rapidly identify probes that target a gene of interest. Currently, this database contains target positions for Affymetrix, Agilent, Illumina, and ExonHit microarrays.

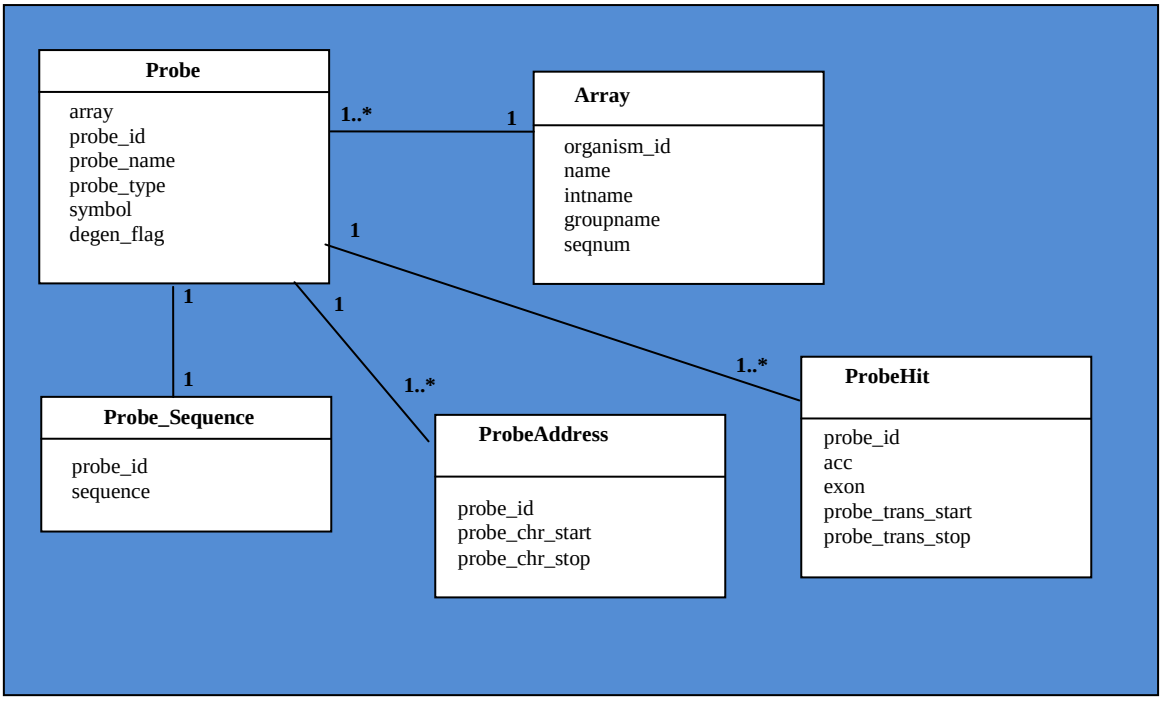


Figure 15: Microarray DB Schema

The Array table describes each microarray included in the database. The Probe table indicates the gene targeted by each probe, whether or not the probe is a junction probe (crosses an exon boundary), and whether or not the probe is degenerate (cross

hybridizes). The ProbeAddress table contains the genomic target location of the probe (two rows if the probe is a junction probe). The ProbeHit table identifies each variant targeted by the probe.

2.5 Microarray build process

The Microarray database is constructed by an automated process that is rerun when new genome builds or microarray platforms are release. The build is based on the probe assignment pipeline developed for SpliceMiner[25] and requires the nucleotide sequence of the microarray probes in a FASTA formatted file.

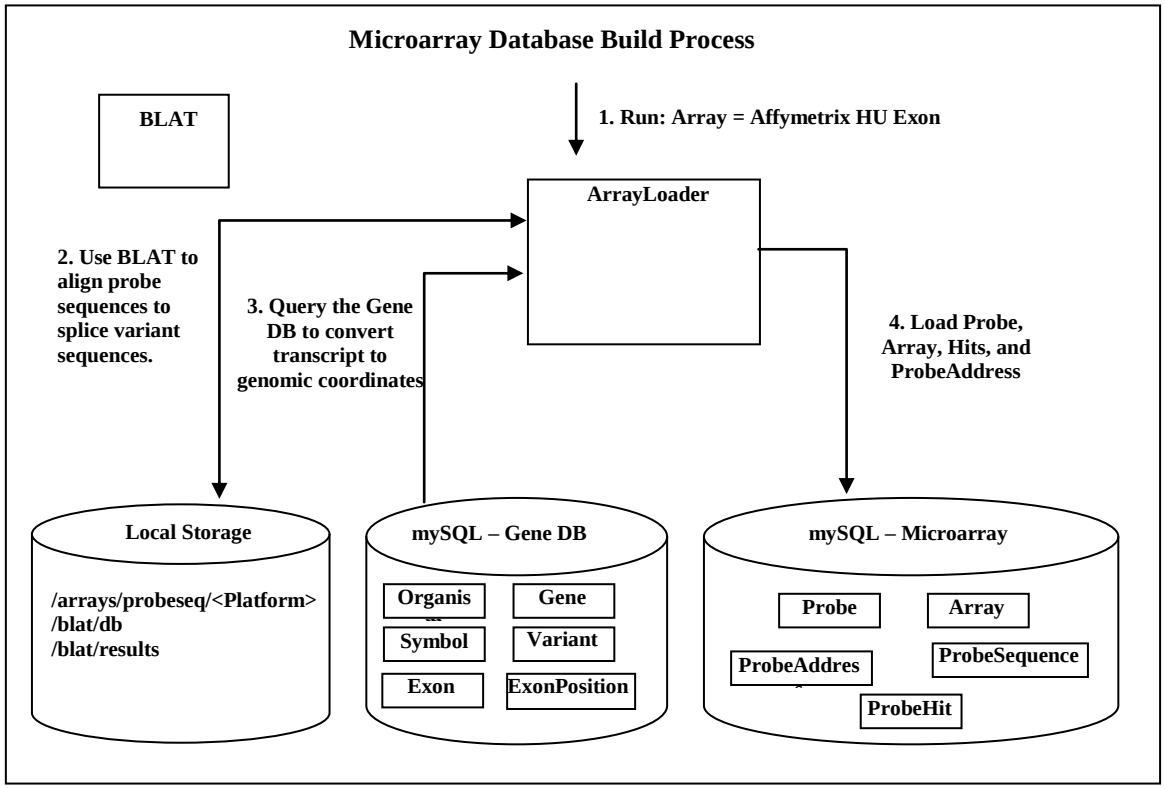


Figure 16: Microarray DB Build Process

The ArrayLoader process is a Java program that performs microarray probe target identification. Run time parameters indicate the microarray probe sequence file, array target organism, and array name. ArrayLoader uses BLAT to align the expression microarray probe sequences to the splice variant sequence data files created by the Gene build process. As can be seen in **Figure 17**, short probes may cross exon boundaries and then cover just a fraction the BLAT database 11-mers making direct alignment to the genome very difficult. For this reason, ArrayLoader aligns probe sequences to mature transcript sequences (splice variants) rather than genomic sequence.

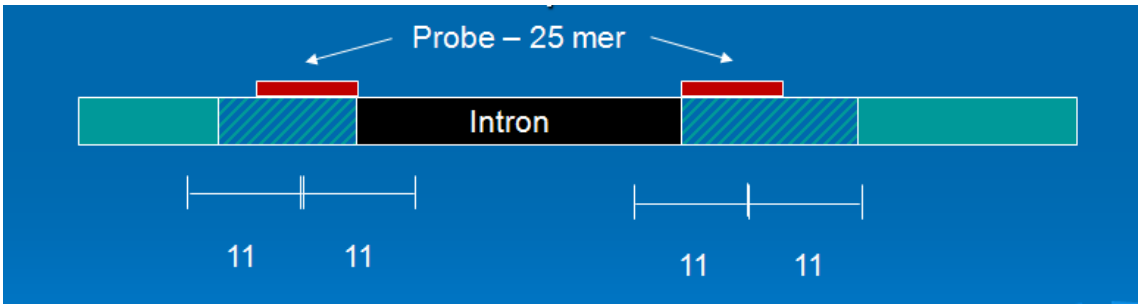


Figure 17: Alignment of probe to genome

The full set of microarray probe sequences is provided to BLAT and takes many hours to run for large arrays. The BLAT results are processed by ArrayLoader to select the best hit per splice variant and SQL queries are then used to convert transcript coordinates of each hit into genomic coordinate. The Gene database is used to find the exon(s) containing the transcript coordinates of the hit and the exon records are then used to determine the genomic coordinates of the hit. Normally a probe sequence will align with multiple transcripts and have a variety of transcript coordinates but the hits will

translate to the same genomic coordinates. In the event that a probe maps to multiple exons, it is marked as a junction probe in the probe_type column of the probe table. If a probe is complementary to more than 1 gene, it is marked as degenerate in the degen_flag column of the probe table. Currently only full sequence matches are considered for cross hybridization. In the future, 1 or 2 base mismatches will also be flagged for potential cross hybridization.

After the BLAST hits are processed and converted to genomic coordinates, the data is loaded into the Microarray database. Probe data is loaded into the probe table as described above. A single genomic address for each probe (two addresses if the probe crosses a junction boundary) is loaded into the ProbeAddress table. This address is the one used by SpliceCenter utilities to display probe target location along with other genomic data. A custom developed MySQL stored procedure selects the best address for each probe. Each transcript targeted by each probe is recorded in the ProbeHit table. The ArrayLoad process may be run in parallel for different microarrays.

3. Software Architecture

SpliceCenter is a suite of web-based tools that are implemented in Java using several standard components of Java 2 Platform Enterprise Edition (J2EE) including: Servlets, Java Server Pages (JSPs), Java Database Connectivity (JDBC), the 2D graphics library, and Web Application Archive (WAR) files. Java was selected for the development of SpliceCenter because it is highly portable and provides strong frameworks for object-oriented, model-view-controller architectures. The applications were developed on Windows machines and are deployed on UNIX servers.

Where possible, open source software was used to implement and deploy the SpliceCenter utilities. This was done to reduce deployment costs and leverage existing software packages. The production SpliceCenter servers run Fedora Linux and MySQL is used as the relational database platform. The SpliceCenter application is deployed on a Tomcat 5.5 web server and executes genetic sequence searches using BLAT and PrimerMatch [41]. The database build process and microarray target location process described in the previous chapter perform alignment of transcripts to the genome and probe sequence to variants with BLAT. PCR primer sequences and siRNA sequences are too short for accurate BLAT alignments so user provided siRNA/ PCR primer sequences are aligned to distinct splice variants with the PrimerMatch application.

Figure 18: SpliceCenter Software Component Architecture - Primer-Check provides a high-level overview of the software architecture of the Primer-Check application which is representative of the architecture of the interactive applications. The architecture follows a traditional model-view-controller pattern in which presentation logic is separate from reusable data objects.

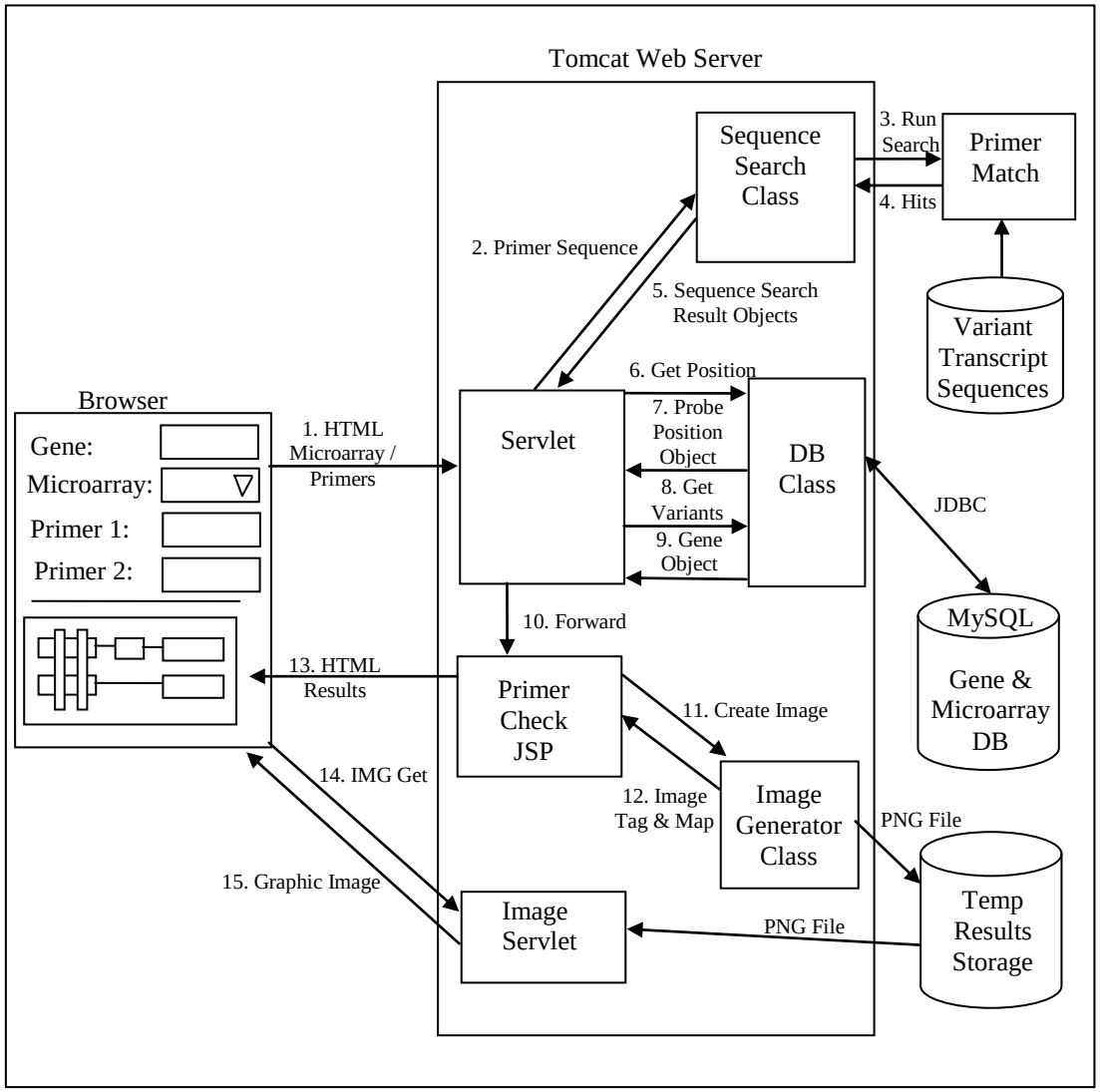


Figure 18: SpliceCenter Software Component Architecture - Primer-Check

The following is a description of the Primer-Check software components and component interactions in **Figure 18**:

1. A browser submits an HTTPRequest from an HTML form with the user query indicating a gene, microarray of interest, display options, and PCR primer sequences.
2. The HTML request is received by the Primer-Check servlet. The servlet calls the Sequence Search class with the primer sequences provided by the user.
3. The Sequence Search class runs the open source sequence search tool, PrimerMatch, to locate complementary sequences within a database of unique splice variant transcripts.
4. PrimerMatch returns hits for the PCR primer sequences. The Sequence Search class parses the PrimerMatch results and identifies the best matches.
5. Sequence Search Result objects are returned to the servlet with the start / stop position of primers within splice variant transcripts.
6. Transcript coordinates for the target position of each primer are converted into genomic coordinates by querying the Gene database. Also, the user specified microarray and gene symbol is used to query the microarray probe positions for the specified gene.
7. Genomic coordinates for each primer are packaged into primer position objects. Microarray probe coordinates are returned in probe position objects.
8. The user provided gene symbol is used to query all known splice variants from the Gene database.

9. Gene, Variant, and Exon data objects are returned. These objects contain annotations and the genomic coordinates of the exons in each unique splice variant.
10. Gene objects and probe/primer position objects are forwarded in the HTML Request to the PrimerCheck Java Server Page. The JSP is responsible for presentation of search results.
11. The JSP calls the ImageGenerator class which uses the gene and position objects to construct a graphical image of the splice variants and probe / primer positions. The Java AWT classes including Graphics2D are used by the ImageGenerator class to create graphical output. Graphical images are temporarily stored on the disk drive on the Tomcat Server
12. An image tag, complete with a tooltip map, is returned to the JSP for inclusion in the HTML results page.
13. An HTML page with the search results is returned to the user's browser.
14. The image tag on the results page refers to the Image Servlet with a unique ID of the user's result image. The browser automatically requests this image from the Image Servlet.
15. The Image Servlet delivers the PNG format image file and cleans up the image file from temporary storage.

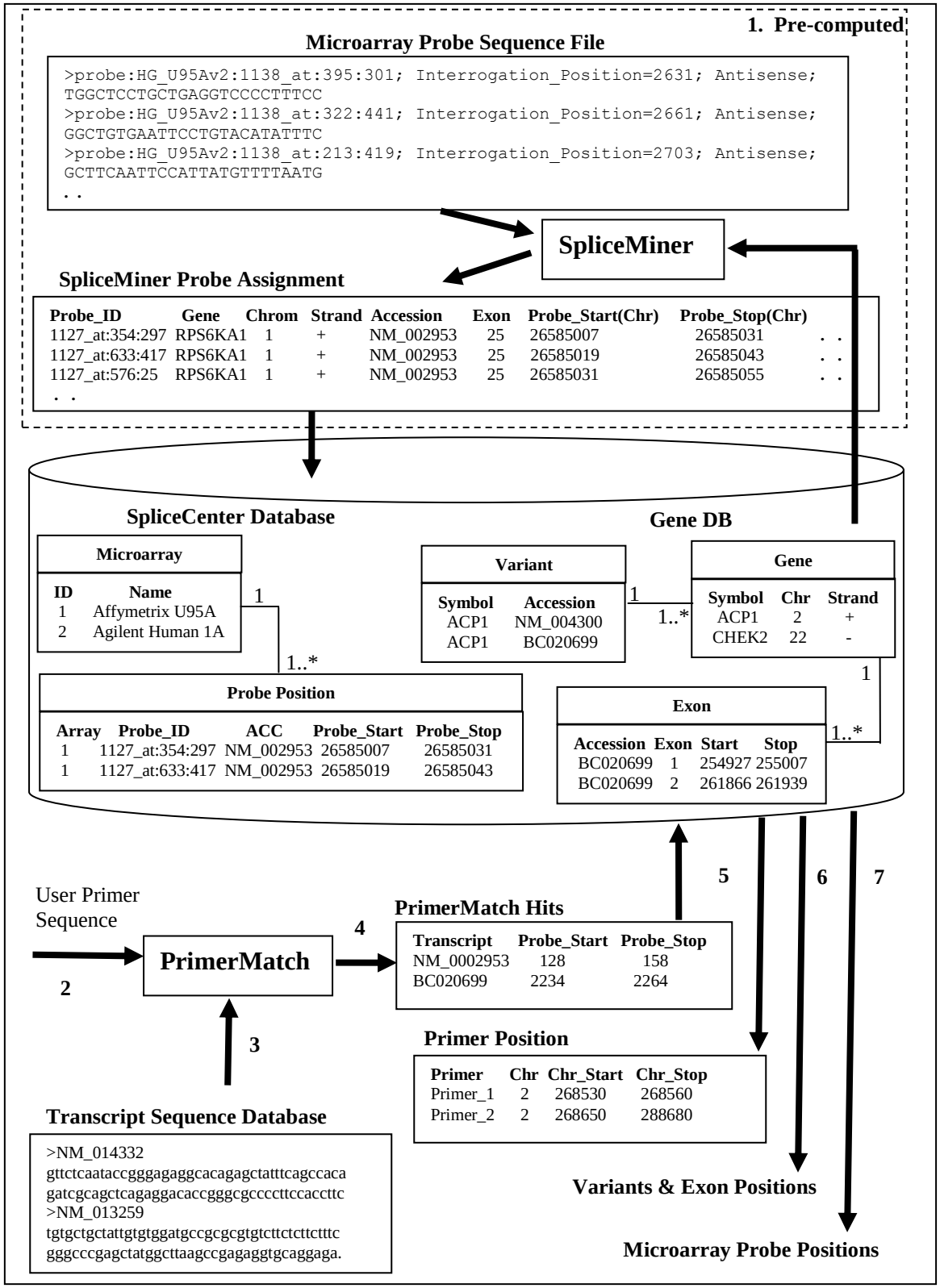


Figure 19: Primer-Check Data Flow

Figure 19 provides an overview of the data processing that is performed in order to determine the positions targeted by microarray probes and/or PCR primers. The key factor in the processing performed by SpliceCenter utilities is the conversion of various elements into a common frame of reference, genomic coordinates (chromosomal position). Identifying the genomic coordinate position of splice variants, exons, PCR primers, and microarray probes allows the creation of a graphical image that accurate displays the relative position each of these elements. The data processing flow of SpliceCenter is as follows:

1. Prior to servicing user requests, the target position of common commercial microarray probes is pre-computed. The previously developed SpliceMiner [25] probe alignment engine is used to perform this task. Microarray probe sequence files are aligned to the distinct splice variant transcripts and the Gene database is used to convert transcript coordinates to genomic coordinates. (e.g. The probe target location starts at position 110 of transcript NM_004300. This position falls 10 bases into exon 2 of this splice variant. Exon 2 starts at chromosome position 11134577 so the target start position is 11134587.) The target location of each microarray probe is loaded into the MySQL Microarray database. This process is computationally intensive so all common arrays are pre-processed and stored in the Microarray database in order to achieve fast SpliceCenter response time.
2. Users provide the primer sequences as either RNA or target DNA sequence.
3. PrimerMatch is used to search a transcript sequence database to locate the primer target position. The set of distinct splice variant transcript sequences are searched

to find primer target positions. Optionally, hits with 1-2 base mismatches can be located.

4. PrimerMatch results indicate the target location of PCR primers in transcript coordinates.
5. The Gene database is queried using transcript target locations to identify the exon(s) containing the primer target and to convert the target location into genomic coordinates.
6. A complete list of splice variant exon positions are retrieved from the database using the gene symbol provided by the user.
7. Optionally, the genomic positions of microarray probes are retrieved for the specified gene / microarray platform. All of the genomic position information for splice variant exons, microarray probes, and PCR primers are passed to the ImageGenerator to construct the graphical image results.

Each interactive application in SpliceCenter has a batch equivalent designed for high throughput processing of query files. These applications process large batch files as a separately running process. Users are notified via email when the batch processing has completed. The components of the batch processing are implemented in Java and many of the data objects and services of the interactive applications are reused for batch processing. Batch processing has been tuned for high throughput performance. Primarily this involves the use of temporary files, temporary SQL tables, and DB interactions that reuse DB connections. MySQL is much faster at bulk loading a text file with thousands of rows than performing individual inserts via JDBC. **Figure 20** presents

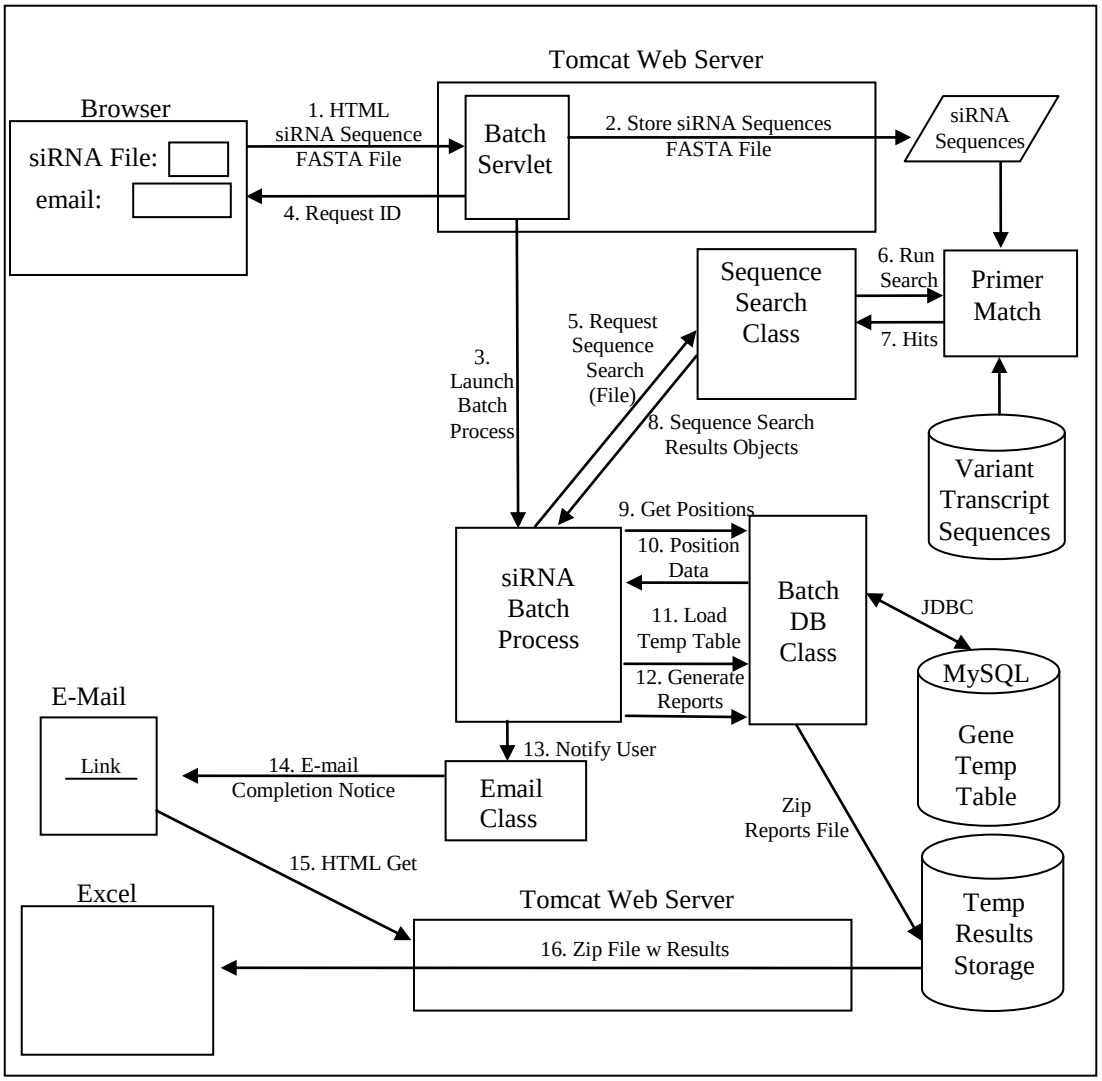


Figure 20: Batch Software Architecture

the architecture of the Batch siRNA-Check application which is representative of all the batch applications. The components shown in the figure interact as follows:

1. The user uploads a file of siRNA sequences via the web interface. The file may be plain text or a zip file with an extension of .zip. The contents of the file must be in FASTA format and should include a unique identifier for each sequence as

the first item in the header of each sequence. The file & parameters are received by the Batch servlet.

2. If the user provided file is in the proper format, it is written out to temporary storage for latter use by the siRNA Batch Process. If the file is not in the correct format or if an e-mail address is not provided, an error message is returned to the user.
3. The stand-alone siRNA Batch Process is launched to perform processing of the siRNA sequences asynchronously. A request ID is passed to the batch process that identifies the proper query file to process.
4. The system assigned request ID is returned to the user.
5. The siRNA Batch Process works with the SequenceSearch class to submit the user query file to PrimerMatch.
6. PrimerMatch is run as a stand-alone application. All user supplied siRNA sequences are processed in a single run of PrimerMatch.
7. The results from PrimerMatch are written to a file and then parsed by the SearchResult class. Matching transcript information including position within the transcript for each siRNA are returned by PrimerMatch.
8. PrimerMatch hits are delivered to the siRNA Batch Process as a SearchResult object.
9. The siRNA Batch Process uses the BatchDBUtil class to query additional information from the Gene Database.

10. Chromosomal coordinate data along with gene and exon information are retrieved for each siRNA hit. The BatchDBUtil class is optimized for high-through put processing.
11. The sequence search results augmented with Gene Database information are loaded into a temporary MySQL table for additional processing. Hit/Miss processing is performed for each siRNA to determine which splice variants are targeted and which are missed by the siRNA.
12. The Hit/Miss Report and siRNA Position Report are generated from the temporary table. The reports are written in .csv format, package into a zip file, and placed in temporary storage on the web server. Results are purged by a cron process after 3 days.
13. E-mail notification is sent to users upon completion of processing by the Email utility class.
14. The e-mail message contains a link for downloading results.
15. Clicking on the e-mail link will generate a request to the tomcat server to download the results file from temporary storage.
16. The zip file with both reports is provided to the user via their browser. The report files will normally open directly in excel.

Note: The siRNA Position report contains an HTML link to the siRNA Servlet that will return the graphical display with siRNA position relative to all known splice variants. This link contains chromosomal coordinate information so it is not necessary to perform the sequence search again.

4. SpliceCenter Utilities

The SpliceCenter web-based application suite is designed to assist biologists and bioinformaticists in analyzing the impact of alternative splicing on studies of transcripts and proteins. Each application identifies the target locations of oligonucleotides or peptides within the unique splice variants of the targeted gene or genes. Interactive investigation is supported through web-based applications that return graphical results with clickable hyperlinks. High-throughput applications for processing large query files are implemented as batch applications that return text-based result files. Currently support species include: H. sapiens, M. musculus, R. norvegicus, A. thaliana, B. taurus, C. elegans, D. melanogaster, D. rerio, and O. sativa.

4.1 General

A critical design goal of SpliceCenter is that the interface be very intuitive and user friendly. Modern biologists are already burdened with the complex analysis pipelines required to interpret voluminous genome-level assay results. As such, they may be reticent to include additional dimensions such as alternative splicing into their already complex analysis. Easy to use bioinformatics tools that provide new insights with minimal learning curve time have the best chance at general adoption. SpliceCenter attempts to do the heavy lifting required to integrate alternative splicing and specific

assays. Each utility includes extensive help and FAQ pages. The interactive utilities have single click sample queries and the batch utilities downloadable sample query files.

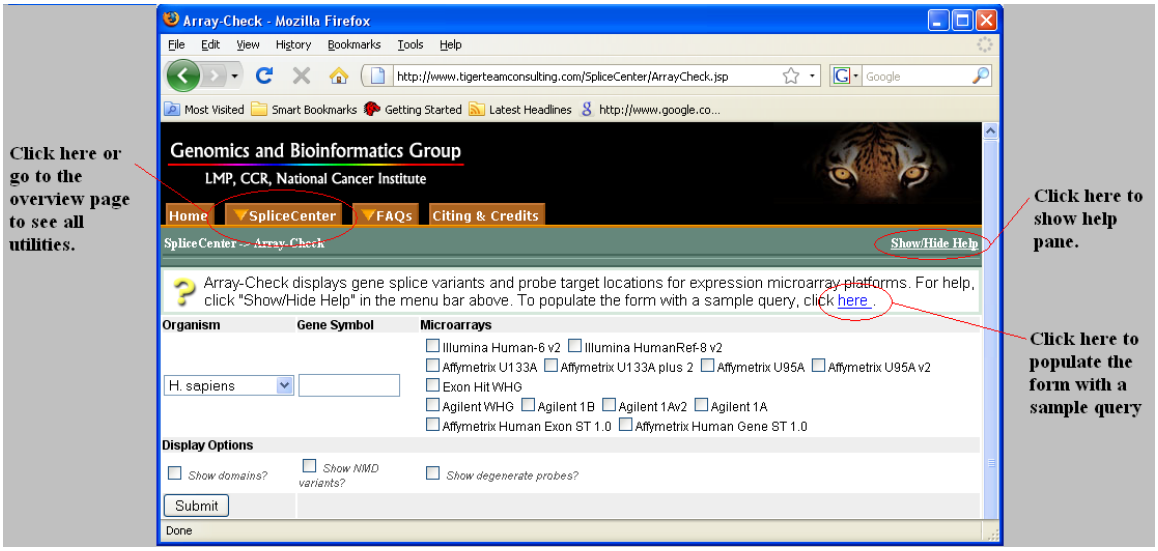


Figure 21: SpliceCenter Interface

The SpliceCenter utilities require users to enter a minimal amount of information to see the impact of alternative splicing on their assay. Generally a user identifies the target organism, gene, and an assay (microarray, primer, siRNA or peptide). If users are exploring the system and don't have a sequence to query, they can click the link to populate the form with a sample query and hit submit to see results. The 'Hide/Show Help' link opens a detailed help pane that provides detailed instructions for using the utilities and interpreting the results. The help pane may be resized and viewed while filling in queries or reviewing results. An Overview page provides a brief description and links to all of the utilities or dropdown menus may be used to navigate the site.

Finally, FAQ pages are provided to answer user questions about each utility and to describe the source of alternative splicing data including data build statistics.

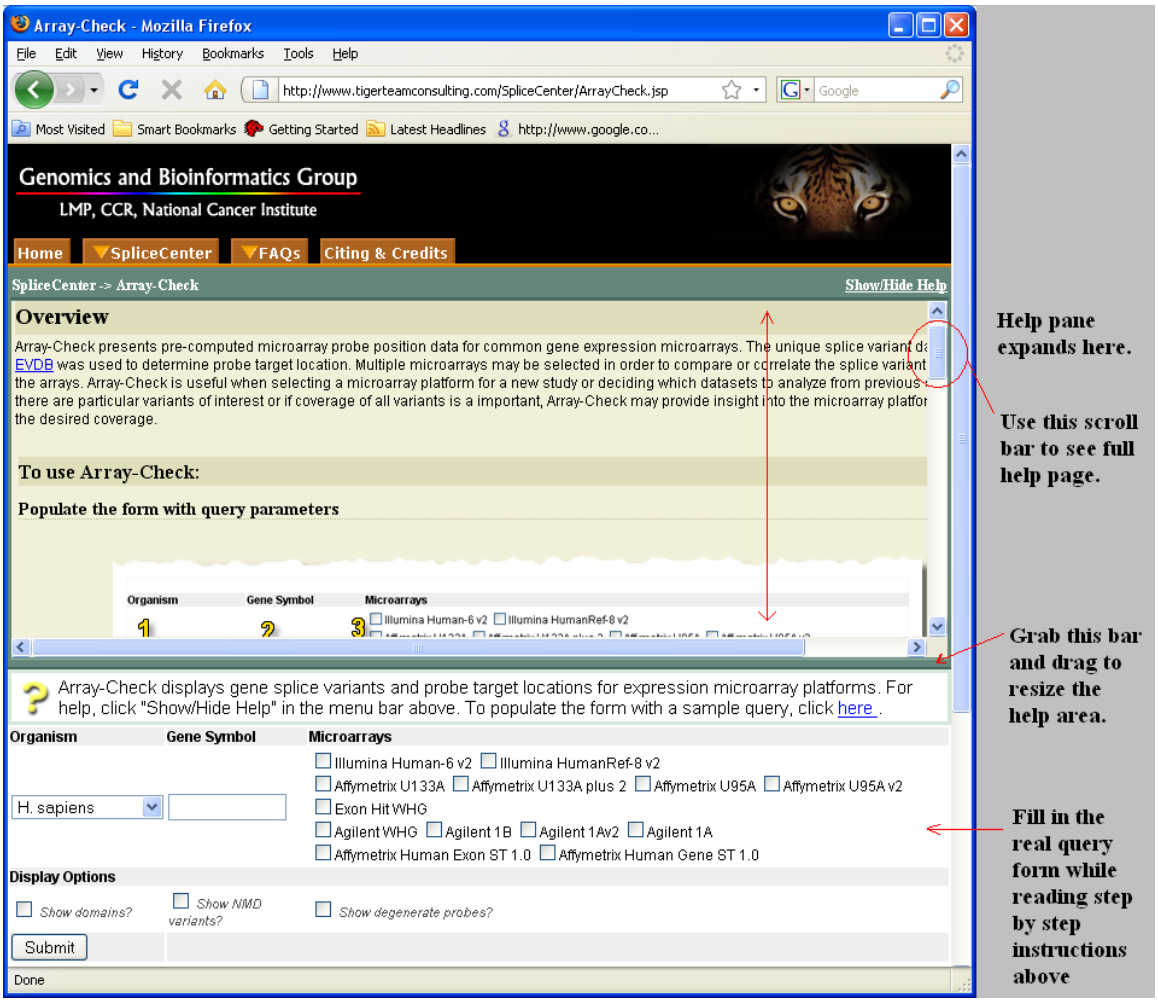


Figure 22: SpliceCenter Help

The graphical results presented by each of the interactive utilities provide an easy to understand view of the alternative splice forms of a gene, shows the impact of the splice variation on the gene's protein product, and identifies the variants that will / will not be targeted by a specific sequence based assay. When users submit queries in

SpliceCenter, the results are displayed on the same page below the query pane. **Figure 23** shows the gene and splice variant results data for the human BIRC5 gene. Each row of the results represents a unique splice form of the gene. The thick boxes indicate

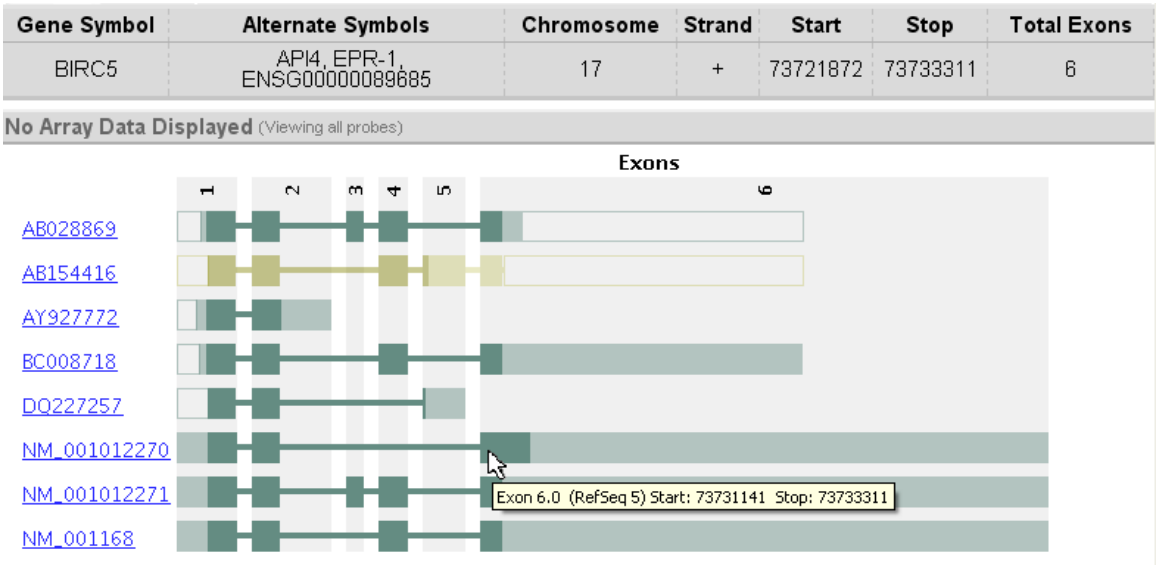


Figure 23: SpliceCenter Results

exon regions and are drawn to scale based on nucleotide length. The thinner connecting lines represent splices or intron space and are not drawn to scale. The darker regions of exons show the coding portion of the transcript. The lighter regions are the 5' and 3' untranslated regions (UTRs). Each splice variant is identified by an accession number which may be clicked to go to the source NCBI record for the transcript. Optionally, users may request to see splice variants which are predicted to be targets of nonsense mediated decay (NMD). These will be drawn in a yellow color. The hollow portions of exons at the 5' or 3' end of a splice variant are predicted to be regions of sequence that are present in the splice form but were missing in the original transcript sequence. See

Gene Build Step 5 in Chapter 2 for a description of the missing UTR prediction process. Placing the mouse pointer on any exon in the display will produce a pop-up box that show the exon number and genomic coordinates of the exon. RefSeq exon numbers sometime differ from exon numbering assigned by SpliceCenter. In these cases, the pop-up box will also show RefSeq exon numbering.

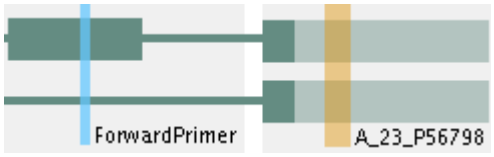


Figure 24: Primers and Probes

When users select a microarray and/or provide primer, siRNA, or peptide sequences, the results will indicate the target location of the probes/primers. Microarray probe target locations are show as vertical orange lines. Primer, siRNA, or peptide locations are drawn as vertical blue lines. When a primer/probe line intersects an exon, the splice variant containing the exon will be targeted by the primer/probe. When the primer/probe line crosses the thin intron line of a variant, then the variant WILL NOT be targeted by the probe/primer. For PCR primer pairs, both forward and reverse primers must hit an exon in the transcript in order to get an amplicon for a given variant. Optionally, users may request to see microarray probes which have been identified by SpliceCenter as degenerate or cross-hybridization probes (a probe that matches more than 1 gene). If the user asks to see degenerate probes, they will be drawn in gray.

One of the newer features in SpliceCenter is the identification of Pfam domains. Pfam [39] is a database of protein domain families generally having known functions. See Gene Build Step 6 in Chapter 2 for a description of the process used to map PFam domains to splice variants. The display of Pfam domains in SpliceCenter provides insights into the affect of alternate splicing on the function of the resultant protein. For example, if a particular splice variant is found to be upregulated in diseased organisms and that variant is missing a transmembrane domain, it may be that the variant is producing a soluble version of the protein which contributes to the disease state.

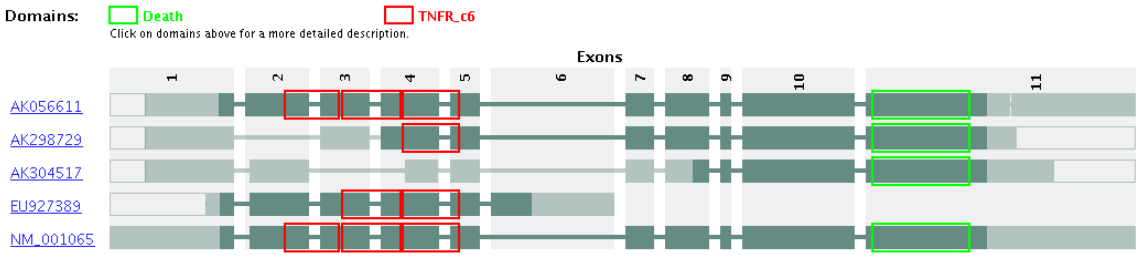


Figure 25: Splice Variants of TNFR1 with Pfam Domains

Figure 25 shows the SpliceCenter display of variants for the TNF-Alpha receptor gene, TNFR1, with the display option to show Pfam domains selected. The legend at the top of the screen identifies the domains, in this case red boxes are drawn around the portion of the transcript that codes for TNFR_c6 domains and light green lines are drawn around Death domains. Clicking on the domain in the legend will open the Pfam page describing the domain. The TNFR_c6 domain is the receptor for TNF-Alpha. The middle portion of the coding region is the transmembrane portion of the protein and the Death domain in exon 11 is a signaling domain involved in the regulation of apoptosis

and inflammation. Alternate splicing of TNFR1 is clearly producing proteins with different properties. The AK298729 has a reduced number of receptor domains and the AK304517 variant does not have the receptor domain at all. The EU927389 variant has the receptor domains but no transmembrane or signaling domains leading to a soluble version of the receptor.

The final general feature of SpliceCenter results is a pop-up display with the nucleic and protein sequence of each variant. This display is a handy way to investigate the sequence of the variants and to identify the exon boundaries in the sequence. This feature is often used as a quick easy way to get sequence within specific exons to use in designing PCR primers that target specific variants. Also, the protein sequence viewer can be used to identify exons that induce a frame shift that alters subsequent protein sequence.

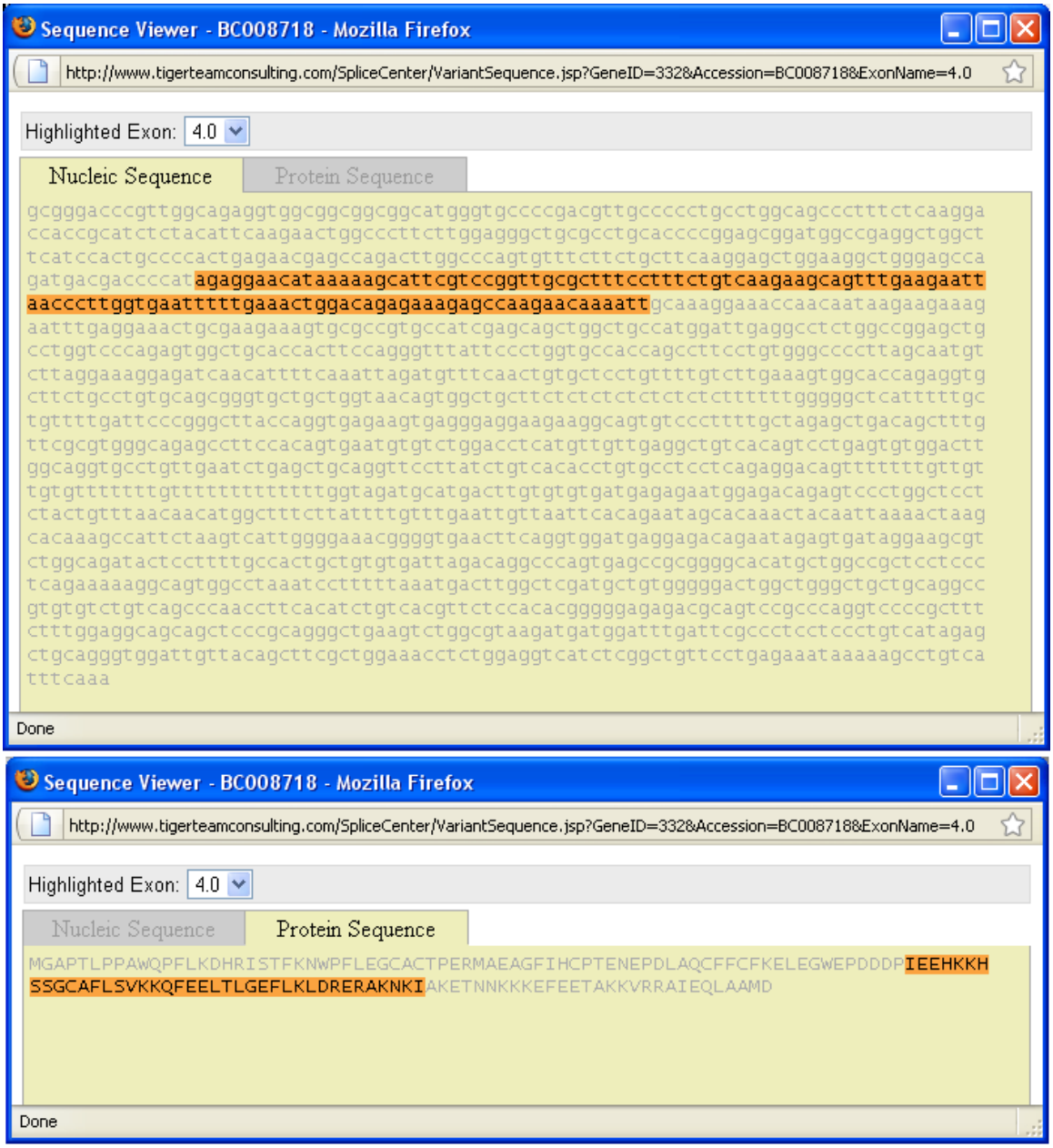


Figure 26: Sequence View of Exon 4 of ACP1 variant BC00871

Figure 26 shows the nucleic and protein sequence of a variant of the ACP1 gene. The selected exon is highlighted in dark orange. Multiple sequence viewers can be opened concurrently in order to compare the sequence of one variant to that of another.

4.2 Primer-Check

Quantitative RT-PCR (qRT-PCR) is widely accepted as the gold standard for validation of microarray expression studies and is also widely used in its own right for accurate measurement of mRNA levels. However, the process of designing oligonucleotide primers for qRT-PCR requires that all existing transcript variants be considered. For example, if qRT-PCR primers and microarray probes target different splice variants, validation of expression results may be compromised. Primer-Check shows the splice variants of a gene and indicates the target locations of PCR primers specified by the user, thereby allowing rapid determination of which variants are and are not targeted. That information can help in the design or selection of primers or in diagnosing unexpected results. The Primer-Check application also includes the option to construct graphical results that indicate the positions of both PCR primers and microarray probes to evaluate whether the two target the same variants. The user enters the target gene and sequence of the oligonucleotides selected for use as primers and/or detection probes. Primer-Check returns a graphical display of the splice variants of the gene and indicates the target locations of the primers (**Figure 27**). Common usage scenarios include the following:

- Primer design and selection: Whether designing custom primers or selecting commercial ones, it is important to identify the splice variants that will be targeted. Primer-Check can be used to ensure that selected primer pairs hybridize to all variants (or specifically targeted variants) and to screen for possible cross-hybridizations.

- Investigation of anomalous results: One potential reason for failure of RT-PCR primers is that they are not targeting the splice variant(s) present in the particular sample being analyzed. Primer-Check is useful for trouble-shooting RT-PCR primers that fail to provide the expected amplification product.

- Validation of microarray data: As already noted, qRT-PCR is considered to be the gold standard for validation of microarray results. Primer-Check can display the target locations of PCR primers and probes and the target locations of microarray probes in a single graphical display that shows directly whether PCR primers and microarray probes do, in fact, target the same variants.

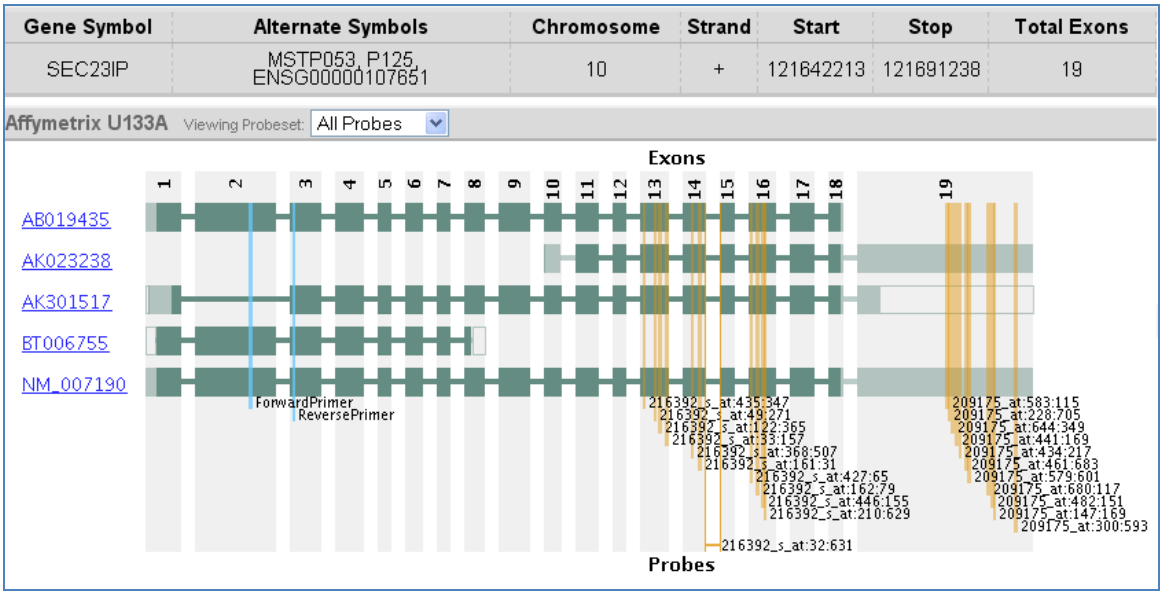


Figure 27: Primer-Check for PCR primer pair and Affymetrix U133A

The effect of splice variation on validation of microarray expression data by qRT-PCR data is by no means hypothetical. A study by Dallas and colleagues [42] found that

such correlations were negatively impacted by splice variation if PCR primers and microarray probes targeted differently spliced transcripts. **Figure 27** shows Primer-Check results for *SEC23IP* (*P125*), a gene that showed discordance between microarray and qRT-PCR results in the Dallas study [42]. Primer-Check identifies five known splice forms of the gene. Two Affymetrix probe sets *216392_s_at*, and *209175_at* on the U133A GeneChip target the RefSeq-annotated transcript (NM_007190) that corresponds to SEC23IP. Bothe probe sets miss the BT006755 transcript. In contrast, the PCR primers and probes from Applied Biosystems target BT006755 and two other transcripts but miss the AK301517 and AK023238 transcripts that are targeted by the microarray. The discrepancy in targeting of variants between the PCR primers and microarray probe sets leads to discordance between the qRT-PCR and microarray results. Primer-Check can help diagnose or even avoid such problems.

4.3 siRNA-Check

RNAi technologies based on exogenously administered siRNAs or shRNAs are used extensively to investigate gene function. For compactness in the following descriptions, we will often use the term "siRNA" to include all of the standard RNAi effector molecules. siRNAs mediate sequence-specific gene silencing through targeted cleavage of a transcript *via* the RNA interference pathway. Selecting an siRNA sequence that effectively targets a gene is a complex task that requires *in silico* prediction of the ability of the siRNA to mediate cleavage of the targeted transcript(s) while avoiding partially homologous sequences of other genes. Databases of experimentally-validated

siRNAs and several tools to aid in design are available [43, 44]. To achieve the goal of maximally silencing protein expression, it is safest to ensure that all protein-encoding transcript variants are targeted by the siRNA. Hence, in most cases siRNAs have been designed to target all splice variants of a gene that are found in the RefSeq database. But because RefSeq was not designed or intended to include all known transcripts, non-RefSeq splice variants may not be targeted. If, for example, an siRNA has been successful in one cell type and then fails to silence expression in another, the two cell types may be expressing different splice variants. siRNA-Check can be used to confirm targeting of all known variants or selective targeting of a particular variant (and, therefore, silencing of a particular protein isoform). The following are typical uses of siRNA-Check:

- Selection or design of siRNA (or shRNA) sequences: Whether designing custom siRNAs or selecting commercial ones, it is important to understand which variants will be targeted. The siRNA-Check application can be used to confirm targeting of all variants or selective targeting of a particular variant (and, therefore, silencing of a particular protein isoform). In interactive mode, the application identifies siRNA target sequences within a gene *via* an intuitive graphical display. If an siRNA targets a sequence that occurs in more than one gene, multiple graphics panels, one for each gene, are displayed.

- Clarification of anomalous results: siRNA-Check provides a quick, easy way to investigate the possibility that failure to silence a gene is due to splice variation. To cite one example from our own work, when we were trying to knock down

expression of two apoptosis-associated genes, *BAD* and *YWHAZ*, we observed differential expression of some of the untargeted transcript variants [45]. For example, as shown in **Figure 28**, two siRNAs that target *BAD* (siBAD.1 and siBAD.3) mediated a significant decrease in mRNA levels when all variants of the gene were assayed (using the Branched DNA-RNA Quantigene assay, Panomics, Fremont, CA). But siBAD.2 produced no knockdown. Transcript-specific qRT-PCR showed that NM_004322, the transcript variant targeted by siBAD.2, represents only 1% of *BAD* mRNA levels in the cell line studied. We saw analogous results for the gene *YWHAZ* where the YWHAZ_1 failed to mediate a significant decrease in mRNA levels but YWHAZ_2 was successful.

Figure 28: siRNA-Check results for 3 BAD siRNAs and 2 YWHAZ siRNAs

4.4 Array-Check

Transcript expression microarrays are being used as tools throughout basic and clinical research. The results of microarray expression studies are usually reported as lists of over- or under-expressed genes. However, failure of the oligonucleotide probes to detect all splice variants of the target gene may confound interpretation of the results. For example, a recent analysis of gene expression data obtained for different microarray platforms showed that genes exhibiting transcript variation had a lower signal agreement between platforms than did genes with less alternative splicing [46]. Array-Check enables the user to see at a glance which variants are targeted by a given microarray platform and, if desired, to compare the variants targeted by different microarray platforms (**Figure 29**). Array-Check includes a database of pre-computed microarray probe target data for the most widely-used commercial expression microarrays. The splice variant coverage of a microarray should be taken into account in a variety of research situations. The following are examples:

- Microarray platform evaluation: **Figure 29** shows an Array-Check comparison of Affymetrix 95A, U133A, and U133 Plus 2 in their coverage of the *ACP1* gene. Array-Check thus provides a quick means of performing a side-by-side comparison of the coverage of splice variants by microarray platforms. It can be used in the mining of historical microarray datasets to ensure that an older platform provided good coverage of all variants of the gene of interest. It may also be useful in selecting a platform for a new study. As shown in the figure, the

Gene Symbol	Alternate Symbols	Chromosome	Strand	Start	Stop	Total Exons
ACP1	HAAP, MGC111030, MGC3499, ENSG00000143727	2	+	254869	268283	6

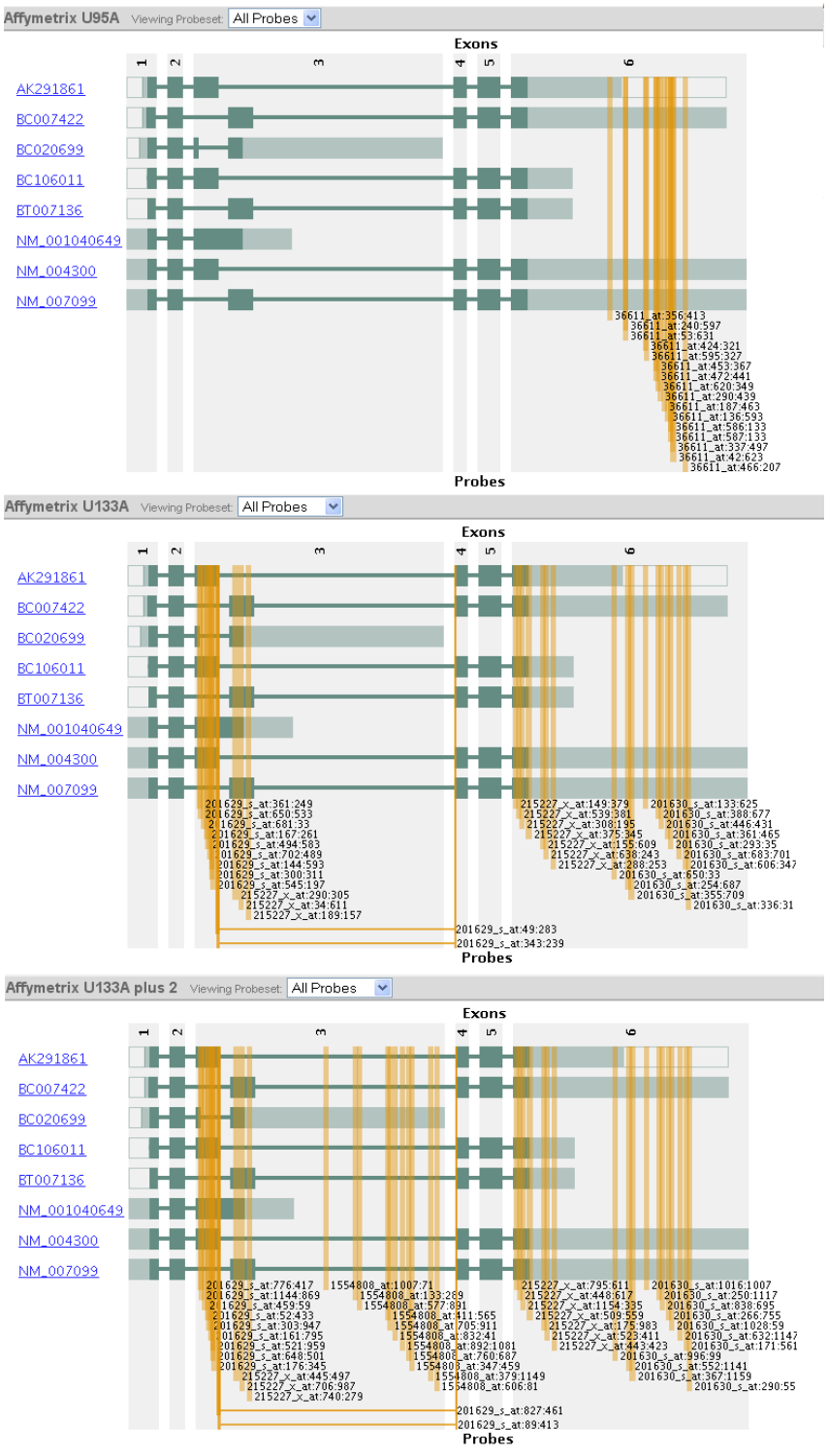


Figure 29: Array-Check results for ACP1 and Affymetrix 95A, U133A, and U133A plus 2

U133A and U133 Plus 2 Affymetrix arrays provide better coverage of the *ACP1* variants than does the U95A array.

- Microarray platform correlation: Comparison of expression values from different microarray platforms is prone to misinterpretation if the platforms target different splice variants. Array-Check provides a mechanism for comparison of probe target locations to identify potential splicing-related differences. For example, **Figure 29** shows that correlation between probe set 36611_at on the U95A array and probe set 1554808_at on the U133 Plus 2 array is unlikely because those probe sets measure non-overlapping subsets of the splice variants of *ACP1*.

- Trouble-shooting anomalous results: Alternative splicing is a potential source of inconsistent expression measurements among the probes in a nominal probe set. Array-Check provides a rapid means for ascertaining the known variants that are targeted or missed by probes on a given microarray platform**.** Older microarrays were designed before the availability of detailed annotation of many of the recently-identified transcript variants. Array-Check indicates splice variant coverage in the context of up-to-date information on transcript variation.

4.5 Peptide-Check

Alternative splicing plays a critical role in higher organisms by increasing the functional diversity of proteins. Isoforms that differ minimally in structure may perform very different functions or may perform the same function in different cell types or at

different stages of development. Failure to take splice variation into account can lead to inaccurate or incorrect interpretation of experimental results. Mass spectrometry and antibody-binding assays, the most common technologies in proteomic research, are susceptible to such problems. For those technologies, Peptide-Check provides a simple interface that accepts one or more short peptide sequences, generates a visualization of the known splice variants of the source gene, and shows the location, within the mRNA transcript, of the nucleotide sequence that codes for the peptide. Common use-cases include the following:

- Design and analysis of peptide immunogens and antigens: Peptide-Check has many applications in the context of technologies that use antibodies or other ligands that target peptide sequences. To cite one common example, animals are often immunized with a peptide to generate antibodies against a specific protein. Peptide-Check can assist in selecting an immunizing peptide that occurs in all splice forms of the protein or, conversely, in only one particular form. The latter type of specificity may be particularly useful for identification of the biological or pathological roles of individual protein isoforms. For example, antibodies raised against peptides that represent unique splice variants of p53 have helped to elucidate details of the molecule's tumor suppressor function[47, 48]. Peptide-Check provides a rapid method for identifying the target variants of those p53 antibodies(DQTSFQKENC – p53β, MLLDLRWCYFLINSS - p53γ **Figure 30**). It should be noted that Peptide-Check is capable of processing only sequential

peptide epitopes; it cannot help with conformational epitopes that are composed of multiple sequences within a protein.

- Analysis of Mass spectrometry results: Mass spectrometry is increasingly being used to identify and/or quantify proteins in a biological sample after peptidolysis. The first step is to identify peptides on the basis of mass/charge ratios, partial sequences, and/or chromatographic elution times. The identity of the original protein is then inferred from the peptides by any of a number of available software packages (reviewed in [49]). Peptide-Check can then be queried to explain the presence or absence of peptides that correspond to a given protein isoform and perhaps to give information on which isoforms are expressed in the sample. In principle, knowledge of splice variation could also be included in calculation of the protein identification probabilities provided by peptide fingerprinting programs.

Gene Symbol	Alternate Symbols	Chromosome	Strand	Start	Stop	Total Exons
TP53	FLJ92943, LFS1, TRP53, p53, ENSG00000141510	17	-	7531588	7505822	14

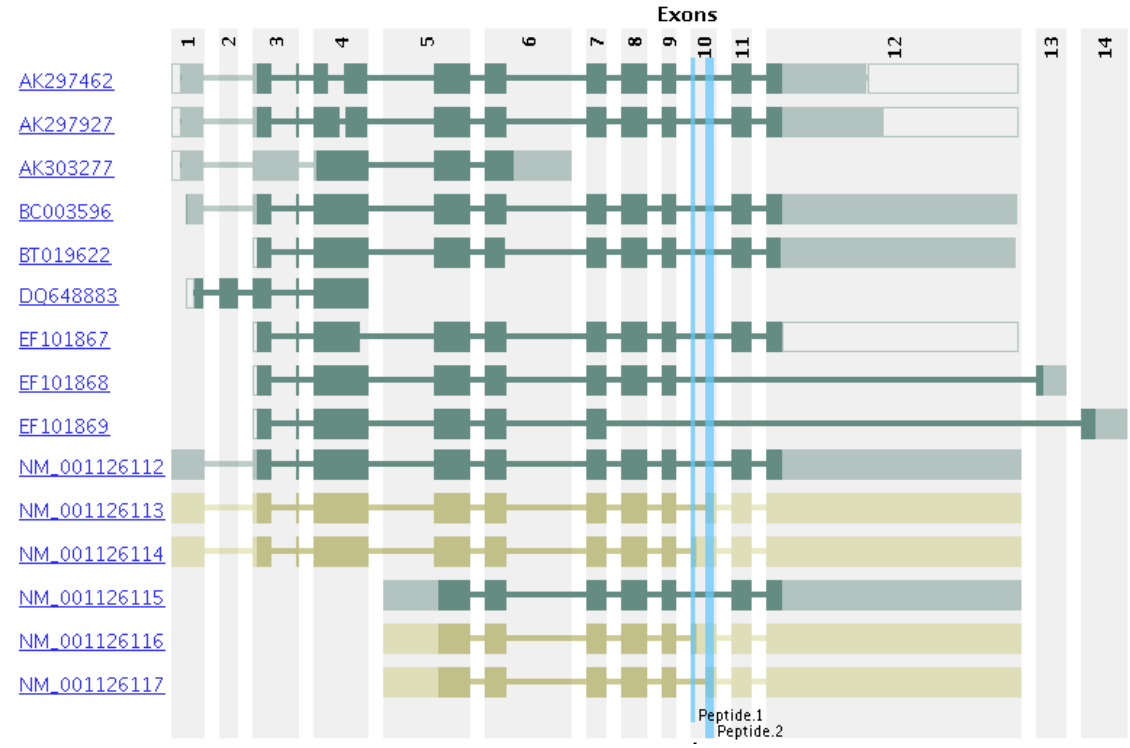


Figure 30: Peptide-Check results for targeting of sequential antibodies specific to isoforms of p53.

4.6 Batch Utilities

The interactive SpliceCenter utilities are intended for manual investigation of specific genes and their splice variants but are not well suited to analysis of high volume data. For this reason, each of the interactive SpliceCenter tools has a corresponding batch application for high-throughput processing of query files. The batch applications support submission of text or zip query files through the web interface and run asynchronously. Help pages for each provide a sample query file. Users are notified of batch processing completion via e-mail and are able to download results from the

webserver. Result files are delimited text files that are well suited to automated processing or import into spreadsheet tools. For example, the batch form of siRNA-Check accepts a FASTA-format query file containing multiple RNAi effector sequences (*e.g.* siRNA or shRNA). Two result files are produced: 1) the Hit/Miss Report and 2) the siRNA Position Report (**Figures 31**). For each query sequence, the Hit/Miss Report indicates the gene(s) targeted by the siRNA sequence as well as the splice variants that are, or are not, targeted. The siRNA Position Report provides the genes, exons, and chromosomal coordinates targeted by each query sequence. Batch siRNA-Check can be used to evaluate large libraries of siRNAs or shRNAs. The Hit/Miss report is useful for identifying those that fail to silence all known variants of a gene or those that target just a single transcript of interest. For further detailed evaluation of a specific entry, the Position Report provides a hyperlink to an interactive-mode graphical representation of the splice variants and positions of the siRNA sequences.

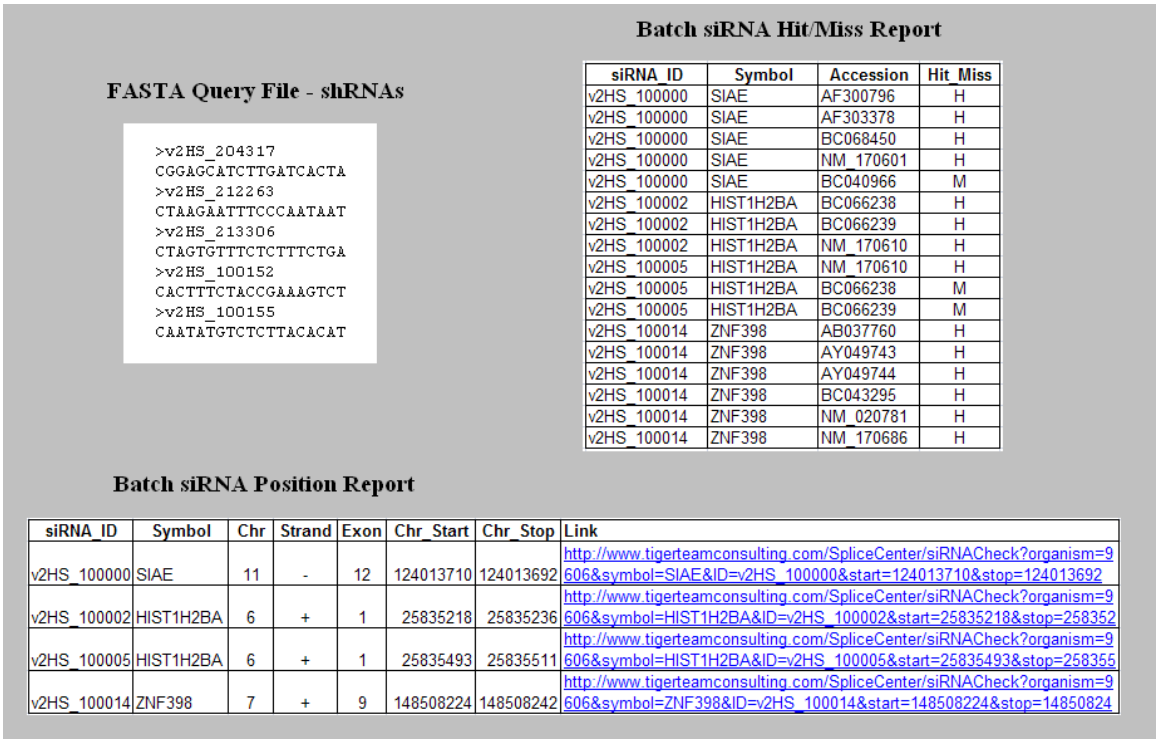


Figure 31: Batch siRNA-Check Query and Results

5. Biological Studies Conducted with SpliceCenter

SpliceCenter has been developed in collaboration with National Cancer Institute researchers in the Laboratory of Molecular Pharmacology and the Gene Silencing Section of the Genetics Branch. Suggestions and requests from these groups lead to new tools and additional features in the SpliceCenter suite. This chapter describes a few as yet unpublished studies in which SpliceCenter played a key role in the analysis of splice variation.

5.1 GSS Study of RNAi and Splice Variation

The Gene Silencing Section (GSS) at NCI develops techniques for efficient, high-throughput application of RNAi technologies including multiplexed small interfering RNA (siRNA) assays and short hairpin RNA (shRNA) libraries. GSS was interested in assessing the possible impact of transcript variation on RNAi efficacy. We used the siRNA-Check batch analysis to query the sequences of 254 synthetic siRNAs (supplied by Qiagen Inc. Germantown, MD) that target 127 human genes. The ability of those siRNAs to mediate RNAi was assessed by measuring steady state levels of the corresponding mRNA species with a branched DNA-based assay[45]. Some of the siRNAs were only partially effective. Although the siRNAs were designed to mediate

silencing against all RefSeq sequences, the batch siRNA-Check analysis indicated that some of the splice variants represented by non-RefSeq coding sequences were untargeted. A summary of the siRNA silencing efficacy (**Figure 32**) indicates that the more effective siRNAs tend to target all splice variants, whereas the least effective siRNAs include a higher portion of siRNAs that miss one or more non-RefSeq splice variants.

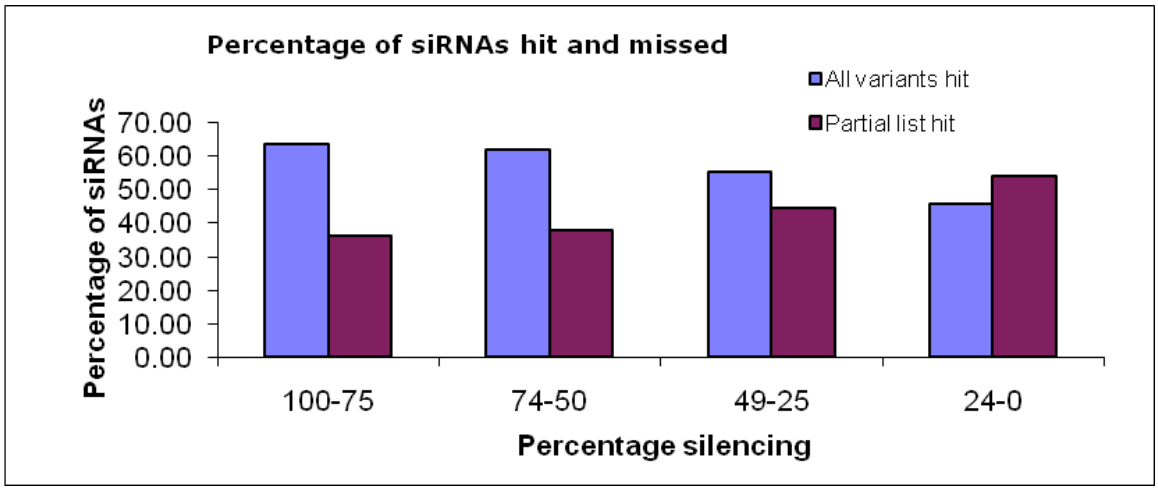


Figure 32: siRNAs that hit all vs. miss some variants binned by gene silencing efficacy

A Kolmogorov-Smirnov (K-S) test [50] was then performed on this data to determine the statistical significance of the difference in efficacy between siRNAs that target all known variants and those that miss 1 or more variant. Batch siRNA-Check results were used to classify siRNAs into two groups: (1) those that target all known variants, and (2) those that fail to target one or more variants. Each group was ordered by siRNA efficiency values (i.e., percentage decreases in target gene mRNA levels) for the K-S test. **Figure 33** shows the cumulative fraction plot of the RNAi efficiencies of the two groups. The K-S test p-value of 0.018 indicates a statistically significant difference

in the RNAi efficiencies of the groups. Splice variation is clearly not the only factor that affected RNAi mediation by the siRNAs studied. Some members of the group that targeted all variants had poor RNAi efficacy; conversely, some members of the group that did not target all variants were highly effective. As a group, however, the siRNAs that targeted all known variants were more effective at mediating RNAi. That observation suggests that splice variation plays a role in siRNA efficacy, and that tools like siRNA-Check can be of use in predicting or analyzing the relative efficiencies of individual siRNAs against particular sets of transcripts in particular cell types.

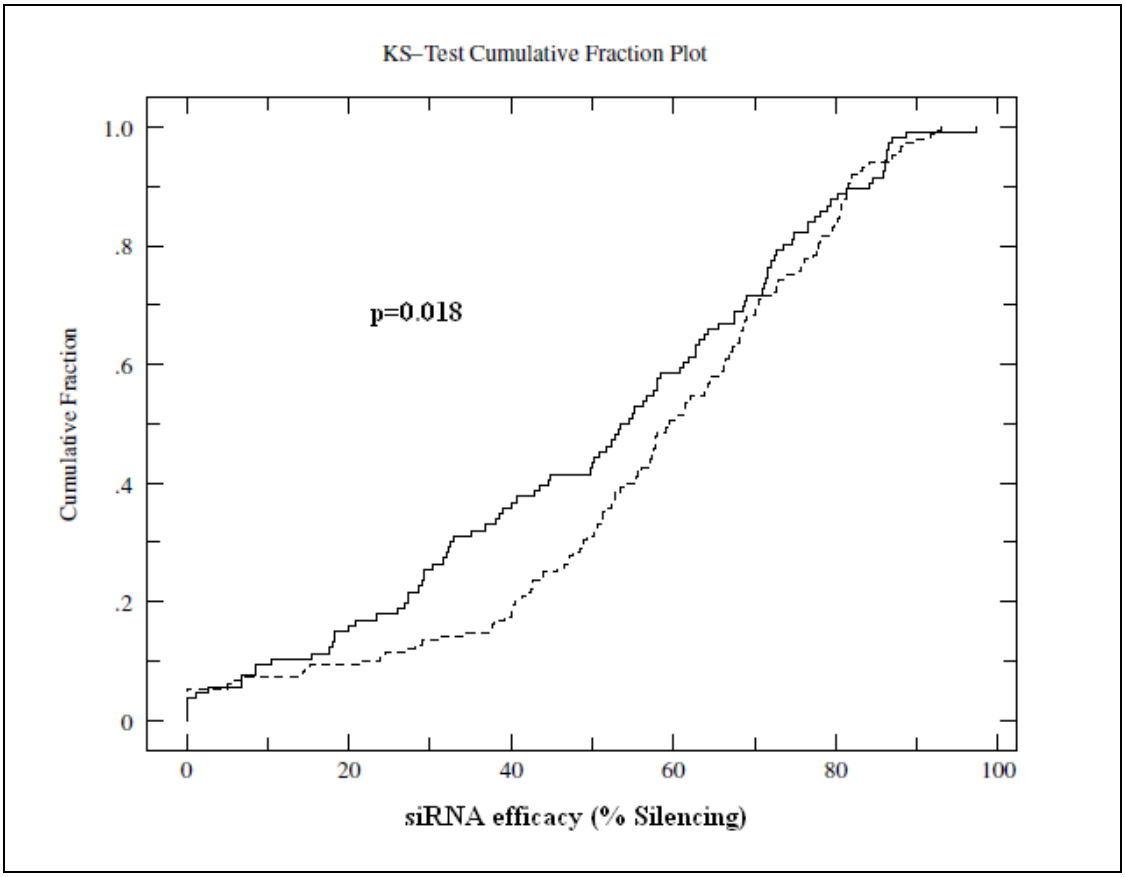


Figure 33: Cumulative Distribution Plot of siRNA efficacy comparing siRNAs that targeted all known variants (solid line) vs. siRNAs that failed to target one or more variants (dashed line). K-S test p-value = .018

We next turned our attention on short hairpin RNA Clone library initiative at NCI's Cancer Genome Anatomy Project (CGAP). CGAP investigates the gene expression profiles of normal, precancerous, and cancerous cells. As part of this project, CGAP provides a large-scale, publicly accessible library of shRNA clones for gene function analysis of cancer related and non-cancer related genes. Our previous results indicated the potentially detrimental effect of selecting RNAi targets that are not common to all known splice variants. To investigate the scope of this issue, we used batch siRNA-Check to analyze the shRNA sequences of 50,766 shRNAs in the CGAP shRNA library. The Hit / Miss Report indicated that 59% of the shRNAs target all known variants, 32% target some but not all transcripts of the target gene, and 9% do not target any currently annotated transcript.

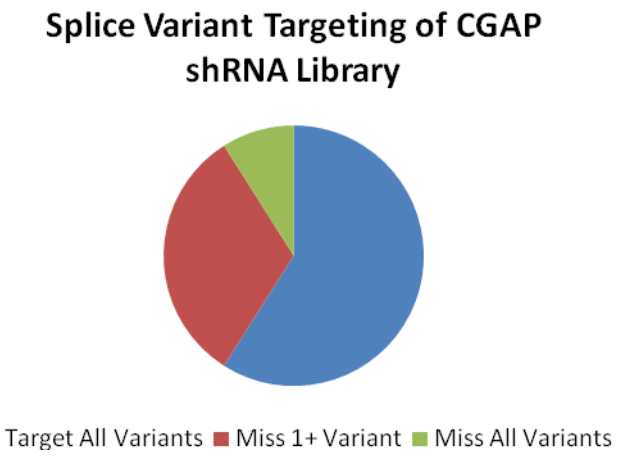


Figure 34: Batch siRNA-Check analysis of CGAP shRNA library

Clearly investigators would not want to waste time and effort on shRNAs that do not target their intended gene's transcripts. This situation arises over time with many

sequence based assays due to corrections and updates of the transcript sequence databases. Investigators may also wish to preferentially select shRNAs that target all variants when possible. The Batch siRNA-Check utility provides a simple, rapid method for screening large libraries of RNAi effectors to find those that target all variants of a given target gene.

5.2 ExonHit Camptothecin Study

Camptothecin(CPT) is an anticancer agent that was discovered during a screen of plant extracts for agents with therapeutic potential in the treatment of cancer. Camptothecin's method of action is inhibition of DNA topoisomerase I. Although topoisomerase I primarily functions as a helicase, recent studies have demonstrated that toposomerase I may also regulate mRNA splicing patterns by phosphorylating Serine/Arginine-rich proteins (SR proteins) [51, 52]. SR proteins bind specific RNA sequences and recruit spliceosome proteins. Kinase activity of topoisomerase alters mRNA splicing by phosphorolating serine residues in the RNA recognition domain of the SR proteins.

Previous studies have described alternative splicing that occurs in a few genes (Bcl-X, CD44, SC35, Clk/Sty, and CASP-2) when cell lines are treated with topoisomerase I inhibitors [51, 52]. PCR of carefully selected mRNA targets in these studies showed changes in amplicon length with treatment indicating alternate splicing. The Laboratory of Molecular Pharmacology at NCI was interested in conducting a broader exploration of the role of DNA topoisomerase I in the regulation mRNA splicing.

This study was conducted using the ExonHit Human Genome-Wide microarray to evaluate mRNA splicing in CPT treated and control samples of HCT-116 cells (colon carcinoma cell line in the NCI60). Samples were taken at 1, 2, 4, 15, and 20 hours after CPT treatment with control samples at 4 and 20 hours.

Typical gene expression microarrays contain 1 to 16 oligonucleotide probes per gene. The probes on these microarrays are 25 to 60 bases long and are complementary to the well known transcript isoforms of the gene (often with a 3' bias). These microarrays generally provide insufficient coverage of transcript variants to predict alternative splicing events. In contrast, the ExonHit microarray platform contains up to several hundred probes per gene and includes exon junction probes specifically designed to detect alternate mRNA splicing.

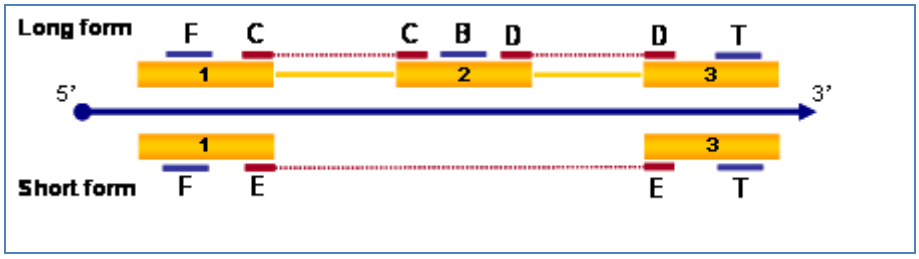


Figure 35: ExonHit probe set design

ExonHit microarrays have groups of small probe sets designed to detect splice events. **Figure 35** shows a group of probes designed to detect an exon skip event in which exon 2 is spliced out of the mature mRNA transcript. Each letter in the diagram represents 1-3 25 base oligonucleotide probes. The F and T probes are five prime and three prime probes that are outside of the region involved in the splice event. The B

probes measure expression of the alternatively splice exon. The C, D, and E probes are junction probes which cross exon boundaries of the splice event. The C and D probes should show expression of the long form of the transcript while the E probes will report expression of the short form of the transcript. If for example, a treatment of a cell line caused an increase in exon 2 skipping, it would be expected that F and T probes would remain constant, C, D, and B probes would show decreased expression, and the E probes would show increased expression.

The traditional method of analyzing oligonucleotide microarray data for alternative splicing is to generate a "splice index" for each exon by dividing exon level expression by gene level expression [27]. This technique does not incorporate indications of splicing provided by junction probes, a key feature of the ExonHit platform. For this reason, we explored several analytical methods for identifying transcript splice variation between CPT treated and control samples. Each approach started with RMA normalized, $\log_2$ expression values and used SpliceCenter probe mappings to identify the gene and exon(s) targeted by each probe.

The first analytical approach was a correlation model designed to find patterns of expression in the T, F, C, D, E, and B probes indicative of specific types of splice events. For example, an exon skip event would be characterized by an increase in expression of E probes and a decrease in C, D, and B probe expression. The study contained time course data so this pattern can be found using simple correlation coefficients in which C, D, and B are positively correlated, T and F are positively correlated, and E is negatively correlated with C, D, and B. If an exon skip is caused by treatment, then over the time

course, it is expected that E probe expression measuring the short form would increase while expression for the C, D, and B that measure the long form would decrease. **Figure 36** shows an example of the data for one splice event found with this analytical technique.

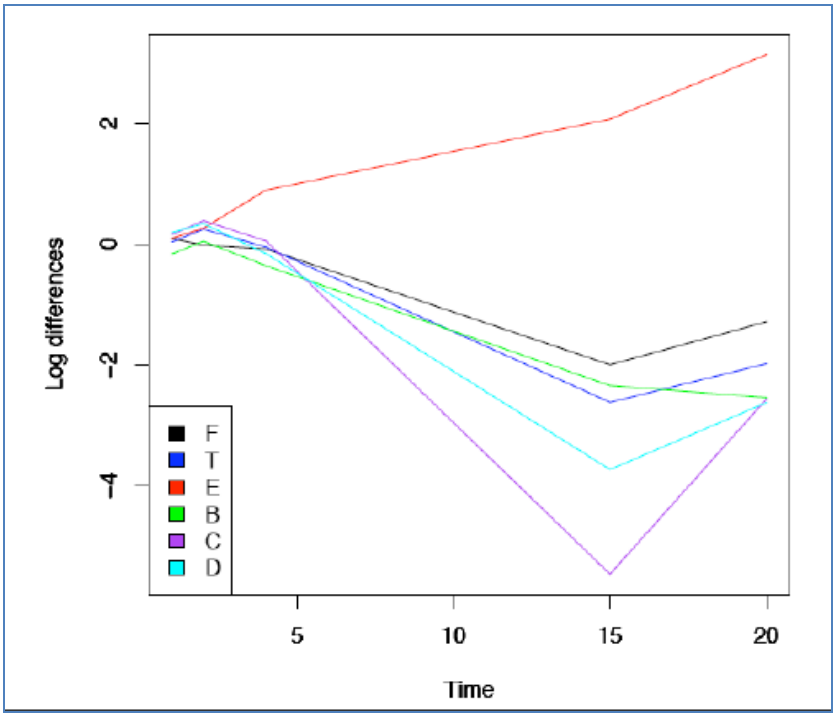


Figure 36: ExonHit 83732.011.1 log2 difference of CPT - control probe expression at 1, 2, 4, 15, and 20 hours.

The correlation model approach was able to identify splice events in the CPT treated cells that were latter confirmed via PCR. However, it was not able to find many events that fit the desired models and 4 of 10 events failed PCR confirmation. Further, many of the detected events were based on $\log_2$ differences of low expression, noisy probes. This approach was unable to find splice events in the linear range of the instrument possibly because the sample contains mixtures of transcript variants and the

probes have a range of hybridization characteristics. It may be naïve to expect that the control sample contains a single transcript isoform and that the treated sample switches to a single alternate splice form that is detectable by tightly coordinated changes in 21 probes.

The next approach at analyzing the data was designed and performed by Peter Munson and Jennifer Barb of the Analytical Biostatistics Section of the Center for Information Technology at NIH. Their technique constructs a statistical model that assumes a null hypothesis that each gene is not alternatively spliced. Deviations from the model indicate alternate splicing. The details of the statistical model will be published in the near future. The result of the analysis is a rank ordered list of genes, the top portion of which contains strong indications of splicing. These results have not yet been validated with PCR assays but a High Throughput GoMiner[53] analysis of the splicing genes showed strong coherence of functional categories.

The final analytical approach emphasized the power of junction probes to detect splice events. If we imagine a situation in which control samples contain 99% of the long form / 1% of the short form and the test sample contains 90% of the long form and 10% of the short form. The change in exon probes and the C/D junction probes would be so small as to be undetectable. The change in the E probe, however, would show a 10 fold increase and identify a potential biologically important shift in splicing. This method used standard probe cleansing techniques and used differences between test and control junction probes to identify specific splice events (See Appendix for details and results). **Figure 37** shows an example of a splice event selected by the simple junction probe

analysis. The F & T expression indicates that there is no general change in gene expression vs. the control but the E probe values indicate a dramatic increase in the short form of the transcript.

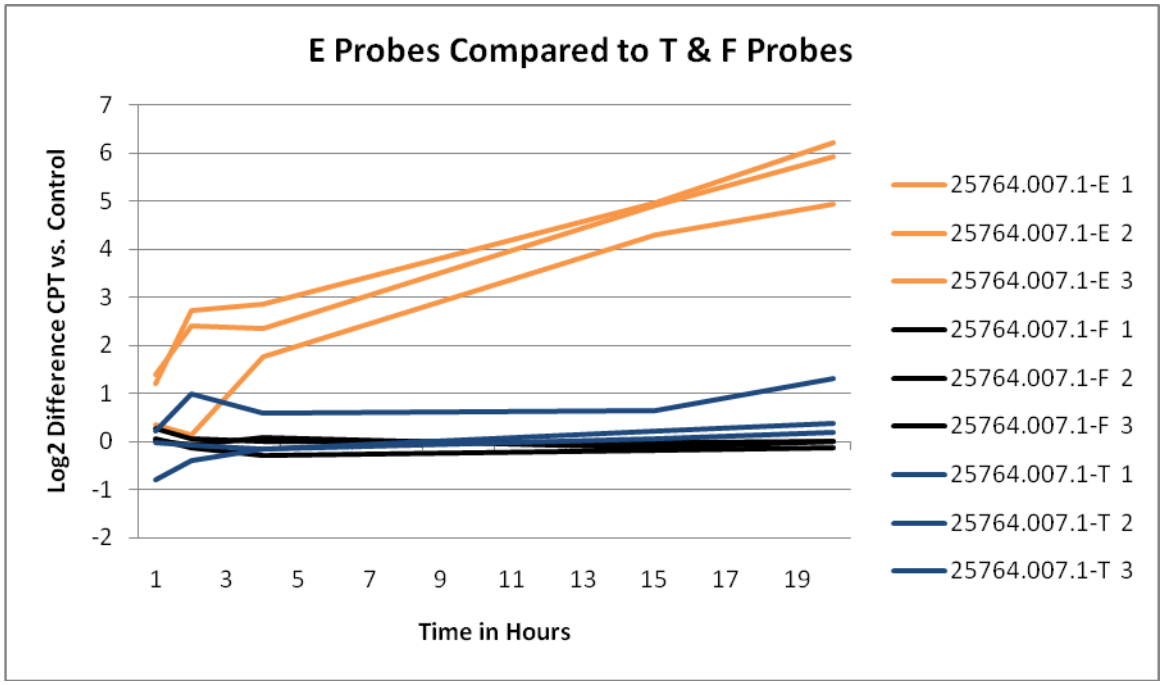
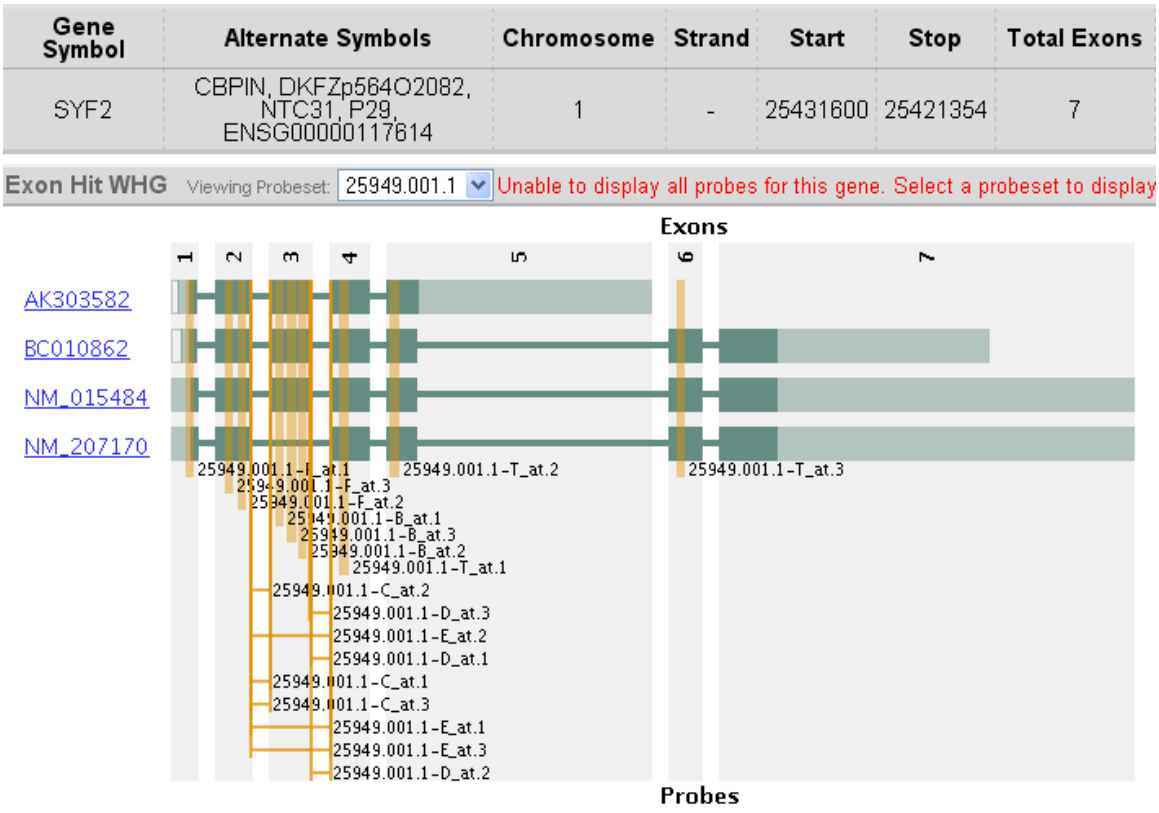


Figure 37: Log2 Difference of CPT - control probes for E, T, and F probes of 25764.007.1 on RBM8A

The list of splice events produced by each of these analytical techniques was given to biologists for further analysis. The interactive SpliceCenter Array-Check utility was used by the biologists to see the location of the splice event in the context of known splice variants. **Figure 38** shows the position of probes for a splice event, 25949.001.1, that was identified in the simple junction analysis as showing an increase in exon skip events with HCT treatment. The probe configuration shows that the exons skip is occurring at exon 3 in the coding portion of the transcript resulting in a known splice

variant of the gene SYF2 represented by a RefSeq transcript. A single click on the identifier NM_207170 displayed the NCBI record for this sequence providing the biologist with the following detail from the transcript record: "This variant (2) lacks an



Note: Click on exons to see exon sequence or use mouse over to get exon coordinates.

Figure 38: Array-Check results showing the position of ExonHit probes for the 25949.011.1 event.

alternate exon, compared to variant 1, but maintains the reading frame. The resulting protein (isoform 2) is shorter than isoform 1 and lacks both bipartite nuclear localization signals found in isoform 1." By using the 'show domains' display option in Array-Check (**Figure 39**), the biologist could see that exon 3 is part of the SYF2 domain so the exon skip may alter the function of the domain. By selecting SYF2 in the domain

legend, the Pfam site displays with a description of the domain function as "cell cycle progression and pre-mRNA splicing". With just a few clicks in SpliceCenter, the team biologists were able to visualize the splice event and gather information on the biological impact of the splicing.

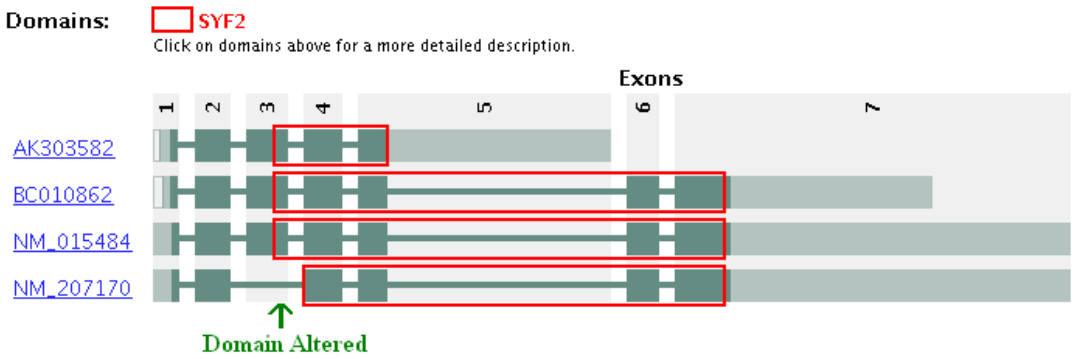


Figure 39: SpliceCenter Pfam domain display of SYF2

The final step in this study was to perform PCR validation of splice events detected by the ExonHit microarray results. For exon skip events, PCR primers need to target mRNA sequence on either side of the cassette exon so that longer amplicons will be produced for the long form and shorter amplicons will be produced when the exon is skipped. Again, SpliceCenter utilities were leverage to assist the biologist in performing PCR validation. It is often difficult to determine the exon boundaries to select the appropriate position for PCR primers. But, by selecting exons in the Array-Check viewer, sequence of the desired exon region are easily obtained and entered into a primer design tool (e.g. Primer3). **Figure 41** shows PCR confirmation of an exon splice event of exon 8 on RIOK1.

Figure 40: SpliceCenter provides a quick way to get sequences needed to design PCR primers to detect alternative splicing.

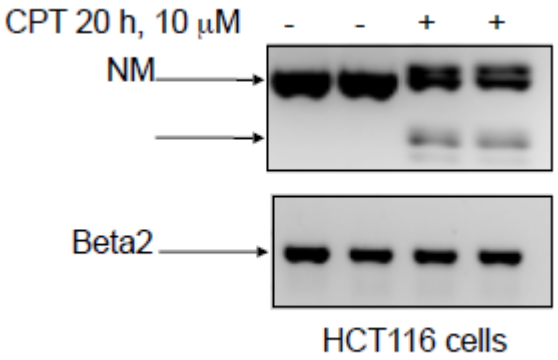


Figure 41: PCR results of exon 8 skip event in RIOK1 gene with CPT treatment. Notice single band for long form (NM) in control and two bands - one for short form and one for long form in the treated samples (+)

5.3 Usage by other research groups

A variety of research groups have contacted us with questions about SpliceCenter or requests for enhancements. The following is a brief overview of a few of these groups and how they are using SpliceCenter:

- The Finnish DNA Microarray Centre is using Array-Check to identify the target location of Illumina Rat-12 v1 microarray probes within target R. norvegicus genes[54].
- A bioinformatics group at Translational Genomics Research Institute (TGen) integrated SpliceCenter data into an internal system that they have for managing libraries of siRNAs. Batch siRNA-Check data was loaded into their database and the tools link to the interactive SpliceCenter screens for detailed views of siRNA splice variant targeting[55].
- The Bioinformatics Research Division of Katholieke Universiteit Leuven in Belgium identifies probes on the Affymetrix U133 Plus 2 array that distinguish splice variants using SpliceCenter[56].
- A PhD student at UCSF constructed a shRNA library focused on alternatively splice transcripts (specific targeting of alternatively spliced sequence) using batch SpliceCenter data and a custom database extract[57].
- A researcher at Children's Hospital in Boston is using SpliceCenter to explore the known transcript splice forms of key genes of interest to their study[58].
- Members of the Biostatistics group at Virginia Commonwealth University used Batch Array-Check to get exon level mappings of probes on the Affymetrix U133A, U133 Plus 2, and the Mouse 140A arrays[59].
- A researcher at the University of Wisconsin was working with a gene that has a complex array of splice forms. He used Primer-Check to select PCR primers that

would target specific sets of splice forms. He also used the visual output of Primer-Check to gain approval for the purchase of the reagents from the PI[60].

- A bioinformatics developer in the Oncogenomics group of NCI's Pediatric Oncology group wrote a new software system to manage their expression data from Affymetrix Exon 1.0 arrays. The software links to SpliceCenter to display probesets of interest and the splice variants targeted/not targeted by the probeset[61].
- Dr. Jameson at Cincinnati Children's Hospital Center is using the SpliceCenter splice variant transcripts to identify reads from a next generation sequencing project to explore the transcriptome of a disease[62].
- Researchers at the University of Edinburgh indicated that SpliceCenter is "really useful" for seeing how differentially expressed genes were probed by the Illumina platform so that comparable RT-PCR can be done[63].

In addition to the anecdotal evidence of adoption of SpliceCenter collected through conversations with users, web statistics have been tracked to assess usage. **Figure 42** presents web hits for the ImageGenerator servlet which is a good measure of the total use of interactive SpliceCenter utilities. Note the SpliceCenter article in BMC Bioinformatics was published in July [64].

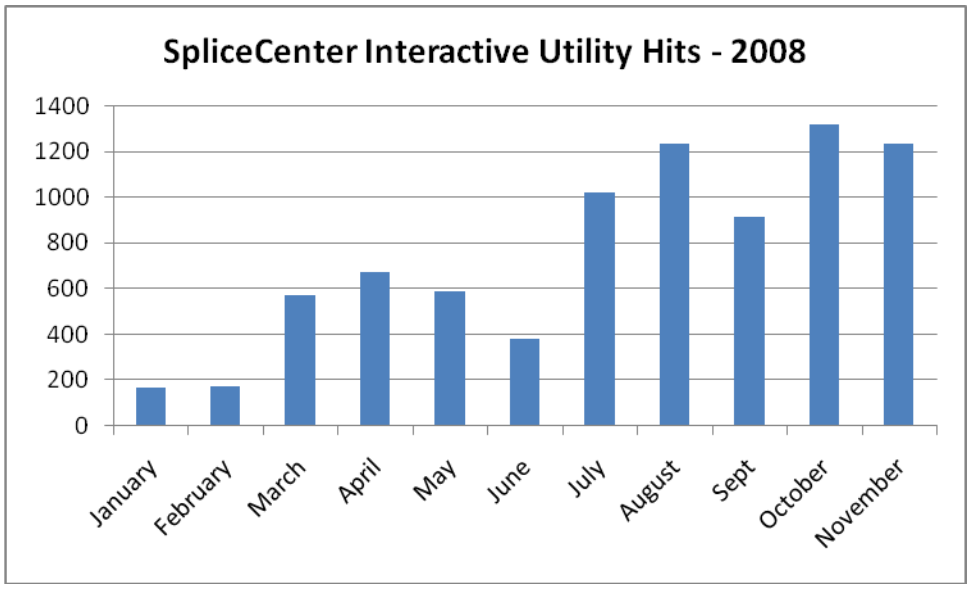


Figure 42: Web server hits for SpliceCenter

6. Spin-off Tools

Work with various labs at NIH would sometimes lead to ideas for additional utilities that could make use of the splice variant and microarray databases. Two of these "spin-off" utilities are described in this chapter.

6.1 ArrayDataViewer

The previous chapter described the benefits that the Array-Check utility provides to biologists evaluating microarray results. While it is very easy to use Array-Check as a reference for probe targeting and splice variant information, it does not support direct visualization of microarray expression data. The ArrayDataViewer is a spin-off utility that combines SpliceCenter's microarray probe target and splice variant data with user loaded microarray expression data. ArrayDataViewer is able to display probe level expression values positioned according to the target location of the probe. (Note: ArrayDataViewer currently does not support the display of junction probe expression data so it is best suited to analysis of Affymetrix Exon microarrays.)

The focus of ArrayDataViewer is the exploration of expression values in the context of gene splice variants. The tool provides a unique ability to identify regional variation in expression values in conjunction with known splice variation to identify study related alternative splicing. Microarray expression analysis is often performed by bioinformatics or statistics professionals who make decisions about how the data should

be analyzed and produce a list of splice events based on their analytical assumptions. The ArrayDataViewer takes a different approach by allowing biologists to experiment in real-time with thresholds, transforms, and comparisons of expression values.

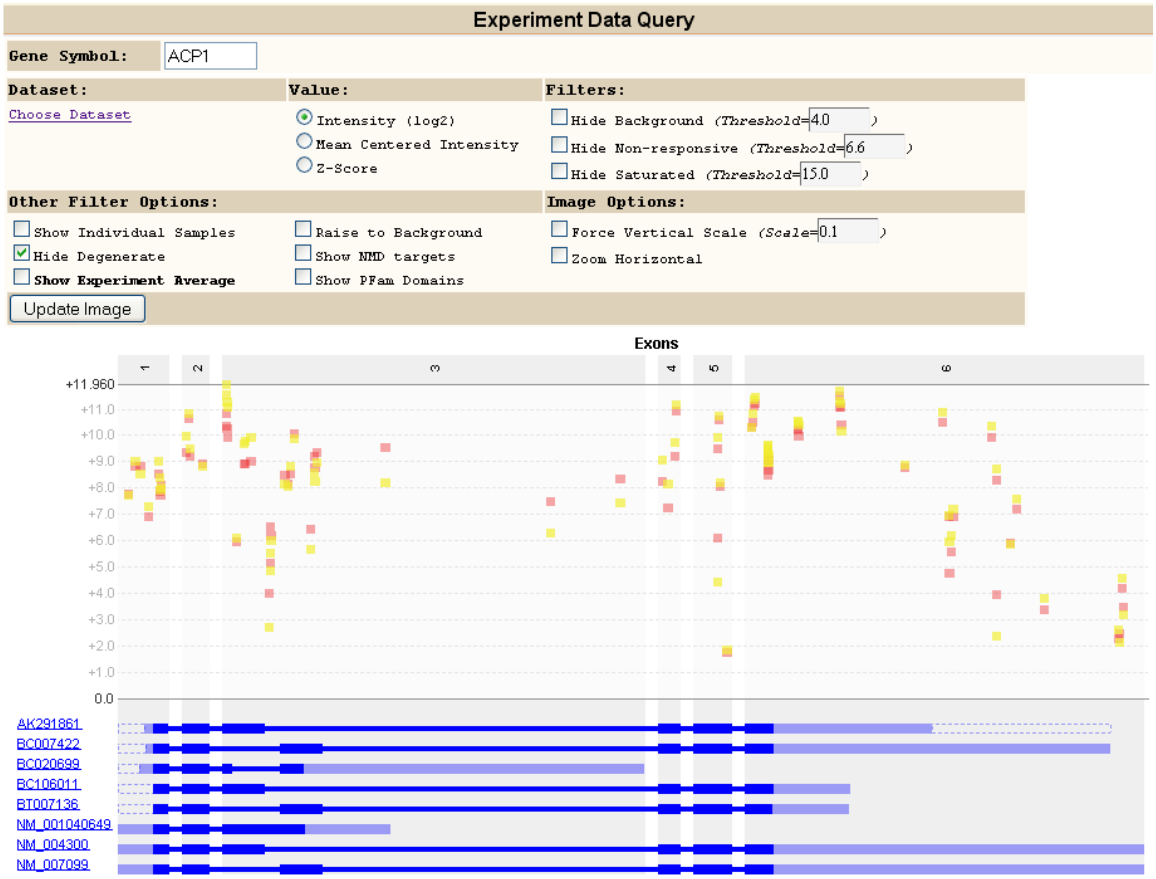


Figure 43: ArrayDataViewer display of ACP1 intensity data for ovarian cancer cell lines OVCAR-3 (yellow) and OVCAR-5 (red) from Affymetrix Human Exon microarray.

Figure 43 shows the display in ArrayDataViewer for NCI60 OVCAR-3 and OVCAR-5 expression measured on the Affemetrix Human Exon microarray in triplicate. The RMA normalized expression data is $\log_2$ intensity and is plotted above the position in the transcripts targeted by the probe. The Choose Dataset link allows user to select the

samples that they wish to display and the color to use for each. If samples are grouped, for example by tissue type, aggregated group data may be displayed instead of individual samples. A variety of thresholds and filters are available to adjust or hide values based on background levels, saturation levels, or cross hybridization potential. The data shown in **Figure 43** is difficult to interpret as there is a good bit of variation in expression level among the probes.

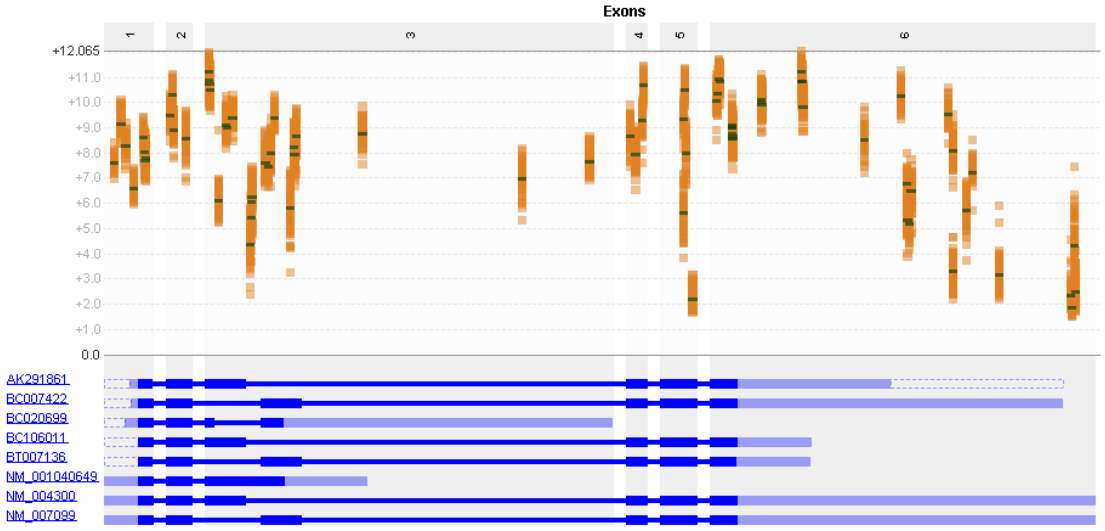


Figure 44: ArrayDataViewer plot of expression of ACP1 for all NCI60 cell lines (orange) with mean as dark line.

Figure 44 shows a plot for ACP1 of the expression of all NCI60 cell lines in triplicate (180 microarrays) with mean as a dark line. This very large set of Affymetrix Human Exon array data covering a variety of tissue types provides an opportunity to explore the behavior of individual probes. Notice that the range of expression values varies widely even for probes on the same exon (e.g. Exon 5). This pattern suggests that differences in probe binding affinity may have a larger effect on the expression values

reported by each probe than changes in gene / exon expression. Instead of averaging probe values for each exon as is done in traditional Splice Index Value (SI) analysis[27], it may be advisable to correct for binding affinity prior to analysis for alternate splicing.

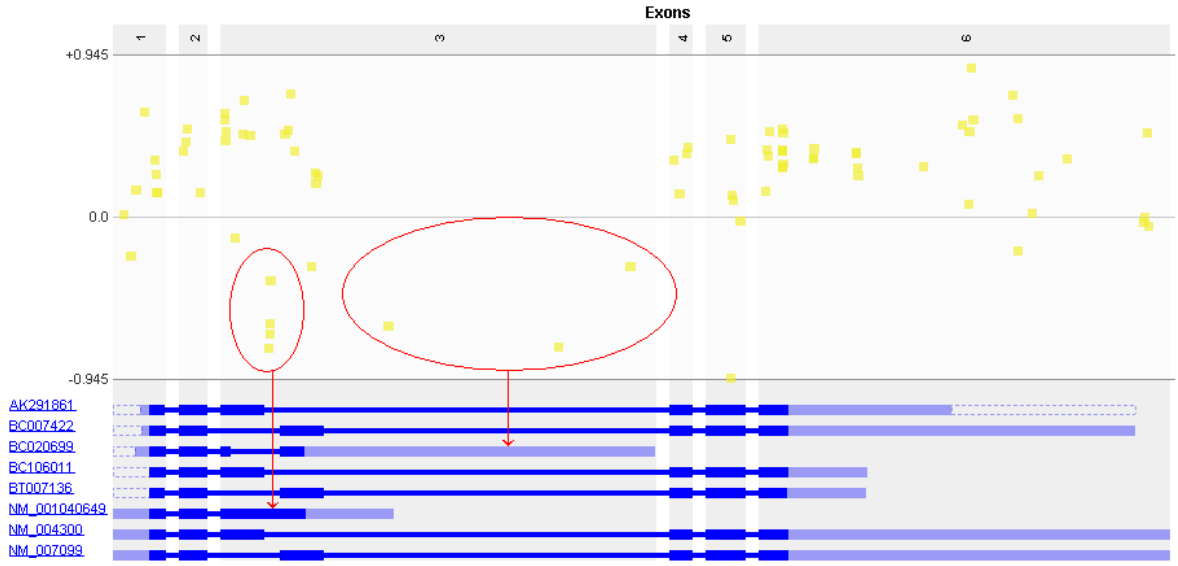


Figure 45: Mean-Centered expression values for ACP1 gene of OVCAR-3 samples.

Figure 45 shows the same expression data for OVCAR-3 that was displayed in **Figure 43** but this time the probe values are mean centered using the NCI60 mean for each probe. In general, the probe values show slightly increased expression of ACP1 in OVCAR-3 compared to average expression of this gene in all of the NCI60 samples. However, two regional groups of probes show lower than average expression that differs from the global gene expression and suggests alternative splicing. A comparison of the target position of these groups of probes to the known splice variants, suggests that OVCAR-3 may produce reduce levels of the NM_00104649 and BC020699 transcript

isoforms. ArrayDataViewer can also perform a z-score transform of expression values to adjust for both probe affinity and probe sensitivity.

Finally, ArrayDataViewer can be a very powerful tool for identifying splicing differences when an experimental design includes treatment & control samples. For example, an LMP study was investigating mitochondrial topoisomerase I (TOP1mt) using knockout and wild type mice. The Affymetrix Mouse Gene 1.0 ST microarray was used to compare gene expression in brain and testis tissue of knockout and wild type mice. **Figure 46** shows the ratio (log_2 difference) of knockout to wild type expression for the Tfam gene in testis tissue. In this scale a value of 1 indicates a 2 fold change, 2 indicates a 4 fold change, etc. The knockout mice show reduced Tfam expression (~ 4 fold reduction) but increased expression of the long 3' isoform suggesting a shift in polyadenylation of this gene in mice lacking TOP1mt. Longer 3' UTRs are potentially interesting as they often present additional targets for miRNA regulation.

Figure 46: Ratio (log2 difference) of testis Tfam expresion in TOP1mt knockout mice vs. wild type mice. Affymetrix Mouse 1.0 ST microarray was used to measure gene expression.

The current version of ArrayDataViewer does not support study wide searches for alternate splicing. It is hoped that the data exploration tools that it does provide will allow biologists to define expression patters of interest and that searches for these patters will then be incorporated into the tool.

6.2 Off-Target

Many functional genomics studies use artificially synthesized short interfering RNAs (siRNAs) to silence expression of a target gene. Introduction of an siRNA causes the endogenous RNAi pathway to knock down expression of the target gene. siRNAs are often used in conjunction with gene expression microarrays to perform a genome-wide assessment of the impact of suppression of the target gene. This approach often yields insight into gene interactions and pathways that involve the target gene.

siRNAs are short double stranded RNAs (usually 21 bases) with a sequence that is complementary to the mRNA of the target gene. siRNA sequences are selected to be specific to the target gene and follow additional guidelines that enhance siRNA efficiency. Despite the careful selection of siRNA sequence, siRNAs may cause off-target effects in which partial matches to non-target transcripts cause unintended silencing of expression[65]. When reviewing microarray results from an siRNA study, a gene that shows a strong reduction in expression may be biologically related to the target gene or may be the victim of an off-target effect. A method of distinguishing biological effects from off-target effects is critical to successful application of siRNA/microarray

techniques. No software analysis tools are currently available to assist with this specific problem.

Current Approaches

Detection of potential off-target effects is not as simple as searching non-target genes for complementary transcript regions with a few mismatches. Recent studies of off-target siRNA effects have shown that siRNAs may act as micro RNAs[66]. The mechanism for micro RNA regulation of expression is not fully understood, however, some elements of micro RNA targeting have been identified. Micro RNAs target the 3' UTR of transcripts and seem to tolerate quite a bit of mismatch between the miRNA and the target transcript. In general, there needs to be a matching 6 base "seed region" at the 5' end of the miRNA (positions 2-7). Other factors enhance miRNA targeting including a match at position 8, an adenosine in target position 1, and additional matches in the 3' end of the miRNA.

Several tools have been developed to assist with identifying potential siRNA off-target effects and other tools have been developed to find miRNA targets. A web-based tool by Chalk and Sonnhammer uses WU-BLAST and a position specific specificity scoring scheme based on experimental evidence of off-target effects[67]. Burge's lab produced the TargetRank application which uses the miRNA targeting rules to produce a rank ordered list of potential miRNA targets. Several other sites including Sanger's miRBase (http://microrna.sanger.ac.uk/index.shtml) provide precomputed miRNA target positions identified using the miRanda[68] program for target prediction.

The fuzzy set of guidelines governing miRNA targeting makes identification of miRNA interactions a stochastic rather than deterministic process. Even in the case of matches that have all of the elements that would suggest a strong miRNA match, only 45% of the matches are experimentally observed to show down regulation[66]. Conversely, poor matches containing just the seed region alignment have been shown to occasionally cause down regulation. If we were to search all known gene transcripts for potential siRNA off-target effects using miRNA targeting rules, a very long list with many false positives would be returned.

A practical, non-analytical solution to the off-target issues is to replicate each siRNA experiment with a different siRNA that has a different sequence that the initial siRNA. If a gene is down regulated by only one of the siRNAs then it might be caused by an off-target effect but if it is down regulated by both siRNAs then it is likely to be participating in a pathway with the target gene.

The Project

Prediction of potential siRNA off-target effects using miRNA targeting algorithms yields many false positives which limits the practical utility of these algorithms. However, if we apply these algorithms at a different step in the analysis, the results become much more useful. The goal of this project is to develop a web-based utility that identifies potential off-target effects from a list of genes that were shown through microarray analysis to be significantly reduced in expression. Instead of

searching the transcripts of every known gene for potential off-target interactions, we examine the genes that were shown to be down regulated and answer the question: "Which of these genes could have been down regulated due to off-target effects?"

The original motivation for this tool comes from a large-scale siRNA / microarray study of genes though to play a role in colon cancer. The hope is that the tool could be applied to lists of genes shown to be down regulated by an siRNA. By removing those gene with potential off-target effects, the resulting list of down regulated genes are more likely to be biologically related to the target gene. This study will use two siRNAs with differing sequence for each target gene. The use of two siRNAs presents an opportunity to evaluate the effectiveness of the proposed tool. (e.g. how many of the genes that show up as down regulated by one siRNA but not the other siRNA are predicted to have off-target effects). If the tool is successful in identifying off-target effects given an siRNA sequence and a list of down regulated genes, it will be generally useful for studies that are not able to use two siRNAs.

Methods

A web-based java application, OffTarget, was developed to perform prediction of potential siRNA off-target effects given an siRNA sequence and a set of genes that were shown to be down regulated in microarray results. A database that was previously developed for the SpliceCenter application[5] was used as the source of 3' transcript sequence data. Unique transcripts from both RefSeq and GeneBank are included in this

database. New code was developed in the VariantSequence object to extract 3' UTR sequence using the coding region start/stop coordinates.

The open source miRanda application was used to perform miRNA target prediction alignments. As is done in miRBase, the miRanda scale = 2 and –strict options were used to force seed matches and to properly weight 5' matches in scoring. The energy and scoring thresholds are changeable through the GUI but have good default values (-5.0 for energy and 50.0 for score) established via comparison to previous off-target analysis of a pilot study.

The OffTarget application is a Java J2EE application using servlets, data objects, and Java Server Pages (JSPs) to implement a model-view-controller architecture. MySQL server is used for relational database services and Tomcat is used as the webserver. The following diagram presents an overview of the system components and their interactions:

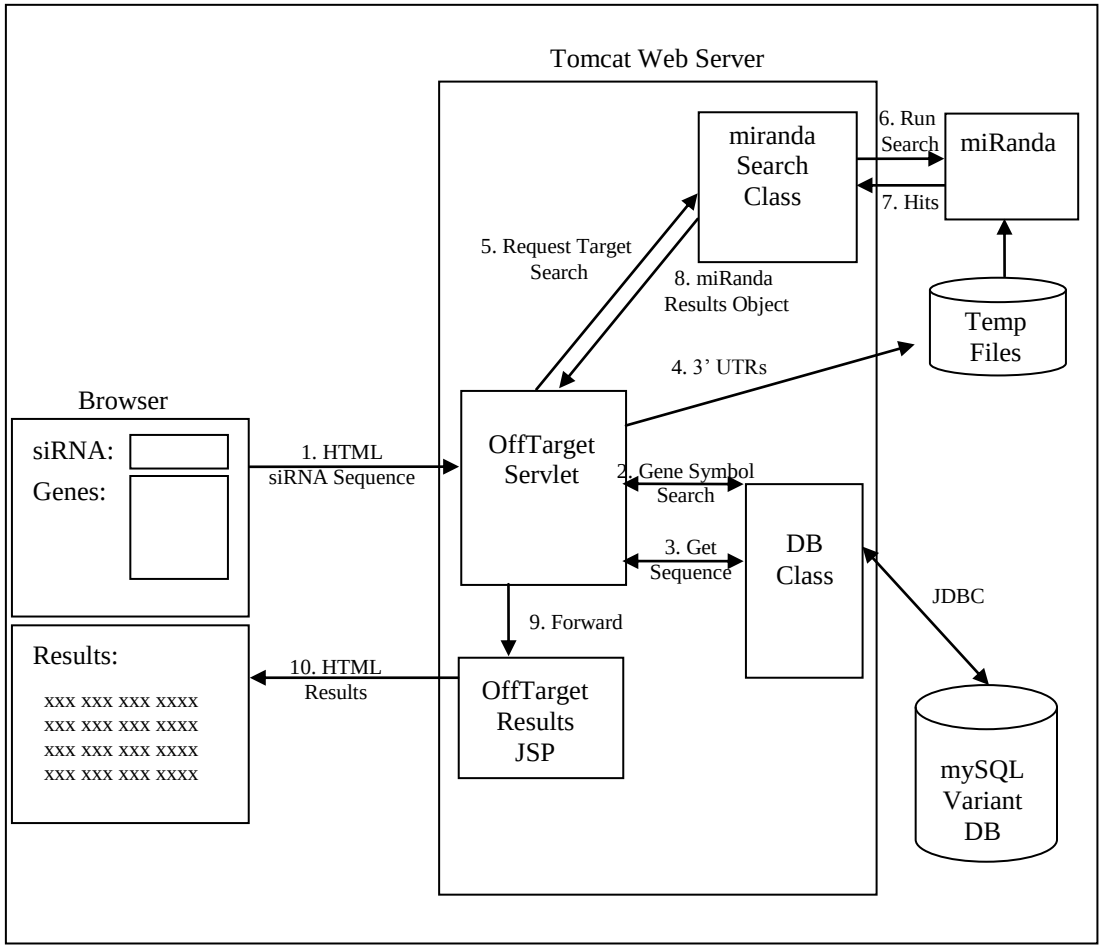


Figure 47: Off-Target Software Architecture

The processing steps depicted in **Figure 47** are as follows:

1. The OffTarget JSP page is rendered as HTML and presented to users. The page has a form allowing query submission.
2. The OffTarget servlet validates user input and presents error messages to users as appropriate. Gene symbols can come in either in the text area or as a file upload. The database class is used to lookup each gene symbol provided by the users. If the symbol is an alias for the correct gene symbol, a lookup provides the correct gene symbol.

3. For each gene, the database is queried to retrieve the sequence of all transcripts related to the gene. SequenceVariant objects are returned.
4. The collection of SequenceVariant objects are used to extract the 3' UTR of all transcripts. A FASTA file is written to temporary disk storage. This file will serve as the "database" for the miRanda search.
5. The mirandaSearch object is provided the siRNA sequence, file name of the 3' UTR file, and threshold settings.
6. The mirandaSearch object launches the 3rd party miRanda program to perform miRNA target searching.
7. The mirandaSearch object waits for miRanda to complete and collects the results.
8. A miRandaSearchResult object capable of parsing miRanda results is constructed and returned to the servlet.
9. The servlet passes the miRanda results to the OffTargetResults JSP. The JSP formats results for presentation to the user.
10. An html page of results is presented to the user.

Results and Discussion

Off-target searches are supported for human, mouse, and rat genomes. Users enter the siRNA target sequence and a list of genes that showed significantly reduced expression in microarray data from siRNA treated samples. The gene list may be entered in the text area or uploaded as a file.

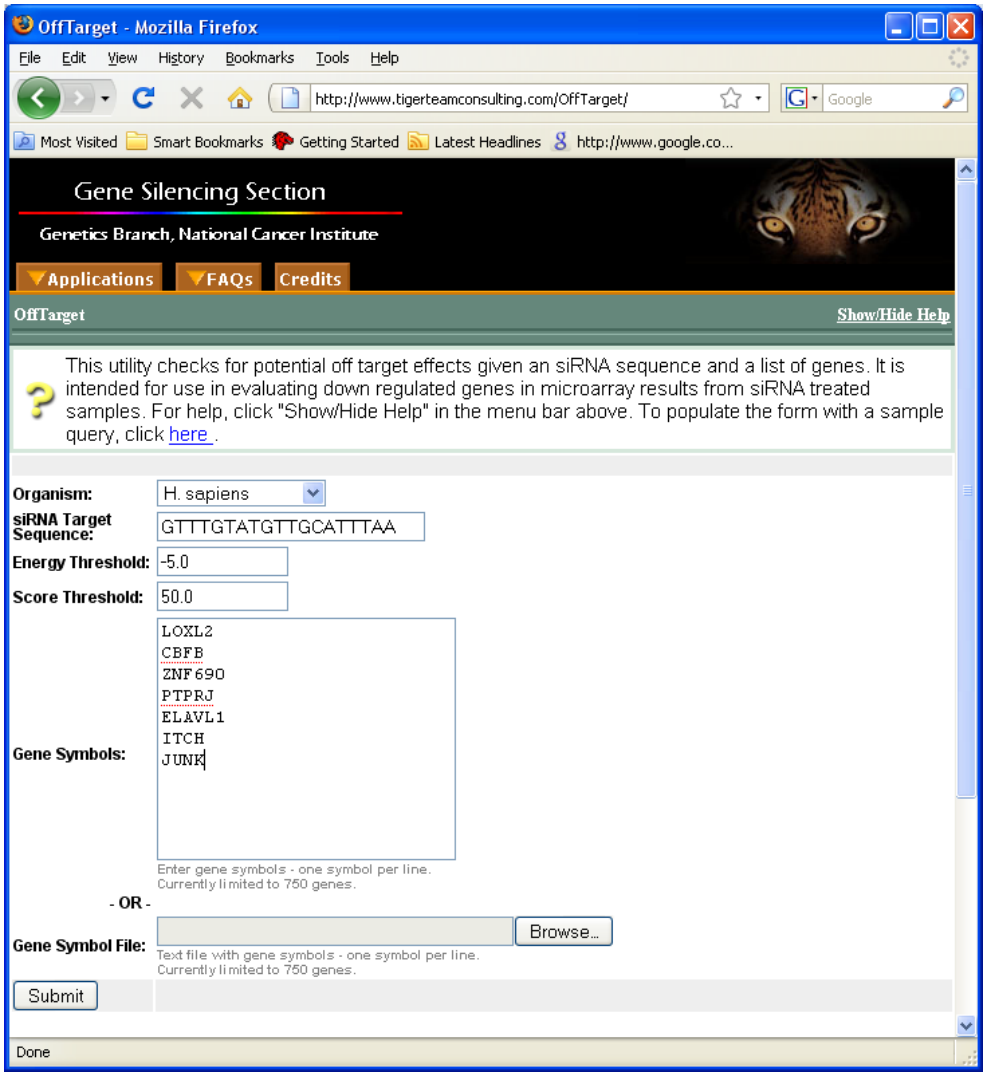


Figure 48: OffTarget Utility

Upon completion of the search, results are presented to the user. Each gene symbol in the user's request is analyzed for potential off-target effects using miRNA targeting rules. The symbols are listed in the results and will have either 'No Hits', 'Bad Symbol', or an alignment of potential off-target interactions. Genes with reduced expression and 'No Hits' are assumed to be down regulated as a result of participation in a biological relationship with the target gene. Genes with alignments need to be

reviewed by the user to determine whether the alignment(s) are strong enough to pose a reasonable threat of off-target interactions. The energy and score thresholds may be adjusted if users would like to be more aggressive in finding matches or would prefer to see only very strong matches.

Search Results

Gene	List of Hits							
CBFB	**Accession**	**Align Len**	**Alignment**	**Identity**	**Energy**	**Score**	**3' Start**	**3' Stop**
	NM_001755	25.0	CAAACATA---C---AACGTAAATT || || || | |||||||||| GTCTGAATAAGGTCATTGCATTTAA	68.0	-10.24	90.0	1522	1546
ELAVL1	**Accession**	**Align Len**	**Alignment**	**Identity**	**Energy**	**Score**	**3' Start**	**3' Stop**
	NM_001419	20.0	CAAACAT-ACAACGTAAATT |||| || | |||||||||| GTTTTTACTTTTGCATTTAA	85.0	-9.88	106.0	1161	1180
	NM_001419	19.0	CAAACATACAACGTAAATT |||| | ||||||| GTTT-TCCAGAGCATTTAA	68.421053	-7.69	77.0	837	854
ITCH	No Hits							
JUNK	Gene Symbol Not Found							
LOXL2	**Accession**	**Align Len**	**Alignment**	**Identity**	**Energy**	**Score**	**3' Start**	**3' Stop**
	NM_002318	16.0	caaACATACAACGTAAATT | ||| | |||||||| aaaTTTAT-TGGCATTTAA	81.25	-7.86	86.0	634	651
PTPRJ	No Hits							
ZNF690	No Hits							

Figure 49: OffTarget Results

An ongoing study of colon cancer related genes provided an excellent opportunity to test the OffTarget utility. Two siRNAs with different sequences were used to silence

expression of a non-histone DNA binding protein in the SW480 colon cancer cell line. Gene expression for each treated sample was measured with the Agilent Human Genome array. OffTarget was used to review all down regulated genes for potential off target effects.

"Confirmed" genes were those that showed reduced expression (> -1.5 fold) for both siRNAs. "Unconfirmed" genes were those that showed reduced expression with just one siRNA. Confirmed genes are believed to be biologically related to the target gene while unconfirmed genes are potentially down regulated by off target effects. If the OffTarget application could identify a large portion of the "Unconfirmed" genes by analyzing the gene list from a single array, then it might be possible to use one siRNA rather than two for these studies.

There were a large number "unconfirmed" down regulated genes. For siRNA1, there were 716 and for siRNA2, there were 789. The subset of these that mapped to NCBI gene symbols (599/639 respectively) was analyzed. With a relatively loose "match" criterion the OffTarget program found potential off target interactions for 29% of siRNA1 unconfirmed genes and 42% of siRNA2 unconfirmed genes.

This is clearly not good enough to drop to a single siRNA and indicates that unconfirmed genes are caused by more than just off target effects. The really unexpected finding was that of the 62 "confirmed" down regulated genes, 19 are showing potential off target interactions and 7 have off target matches for both siRNAs. It is possible that even with 2 siRNAs, some of the reduced expression genes could be off target effects rather than biological effects. A new potential use of the OffTarget tool is be to

strengthen the list of reduced expression genes by weeding out off target interactions from the list of genes down regulated by both siRNAs.

7. Conclusion and Future Work

Although alternative splicing is a ubiquitous and functionally critical phenomenon in eukaryotic gene expression, fluent software tools have not been available to assist researchers, particularly bench biologists, in determining which splice variants are targeted by particular qRT-PCR primer sets, RNAi effectors, microarray platforms or peptide-targeting reagents. Increasingly, all of those methods are being used, separately or in combination, to analyze gene function. SpliceCenter's integrated suite of applications for correlating experimental findings with the transcriptional structure of a gene should significantly aid in elucidating the roles played by splice variation in a wide range of biological processes and diseases. The SpliceCenter applications, currently including Primer-Check, Array-Check, siRNA-Check, and Peptide-Check, provide user-friendly, web-based tools for the biologist and bioinformaticist.

One aspect of this project that worked particularly well was the iterative interaction with NIH researchers who field tested the applications and provided valuable feedback. These rounds of interaction allowed us to tune the application explicitly to the needs of contemporary researchers. These interactions also led to several unique, powerful features in the SpliceCenter suite that are not provided by other bioinformatics applications. First, the novel ability of SpliceCenter to display cross-platform information like RT-PCR primers and microarray probes addresses the needs of

researchers to target primers and microarray probes to the same splice variants. Also, the peptide functions of SpliceCenter open up new dimensions of investigation for proteomics researchers by enabling them to correlate detected peptides to splice variants. Finally, the unique Pfam to splice variant feature of SpliceCenter provides a much needed initial attempt at providing a direct annotation of the functional impact of alternative splicing.

Several enhancements to the SpliceCenter suite are currently under consideration. The current splice variant data in SpliceCenter is constructed from RefSeq and GenBank transcript records. Expanding the splice variant data to include EST sequences would provide a significant expansion to the number of splice variants cataloged in SpliceCenter. Often rare or disease related splice variants are only available as EST sequences. The difficulty in using EST data is that it is very high volume, low quality, and does not include full transcript data. The splice graph algorithms currently used in the build for corrections of missing UTR sequence could be used to compile EST data into composite splice graphs that contain all observed splicing patterns in a very condensed form. Thresholds in the construction of the graph could filter out most erroneous sequence data. Display of splice graphs in the interactive SpliceCenter utilities would present EST splice variation to users of the applications.

Another planned enhancement to SpliceCenter concerns the identification of microarray probe target locations. The current pipeline is not able to identify junction probe targets unless an observed splice variant contains the targeted exon junction. For example, if a probe is designed to find transcripts where exon 1 is spliced to exon 3, we

will be able to identify the target of this probe if a RefSeq or GenBank transcript has exon 1 spliced to exon 3. If not, we can't find a target location for this probe. The target of probes designed based on ESTs or designed to find novel splicing often can't be identified by the current process. The Microarray probe alignment process could be augmented to use a different alignment process capable of finding fragmented alignments or a database of combinatorial splice sequences could be created to solve this issue.

Finally, a major adaptation of the SpliceCenter suite could be implemented to apply SpliceCenter functions to next generation sequencing of transcripts. The splice variant sequences in SpliceCenter may provide a better basis for identification sequence reads than the RefSeq transcripts used currently in some alignment processes. Our current splice graph algorithms could be used to assemble sequence reads into a simple representation of exon expression and splicing frequency. These techniques could improve the match rate for transcript reads and provide insight into alternative splicing patterns in next generation sequencing data.

Appendix

Junction Probe Analysis

A basic evaluation of junction probes was one method used to analyze the ExonHit Human Genome microarray data from the study of Camptothecin treated HCT-116 cells. A precursor analysis method that attempted to find more complex patters of expression change (e.g. E probe increase, C & D decrease, B decrease, T&F steady) was not able to locate significant numbers of splice events in the data. Further review of data from predicted splice events that passed and failed validation indicated the importance of junction probes in detecting splice events amidst a mixture of transcript variants. The simple junction analysis approach focuses on expression differences of junction probes between treated and control samples. The primary goal of this analysis was to identify with high confidence specific splice events represented in the ExonHit data. A secondary goal was to identify the timing of splicing changes after treatment and to determine whether topoisomarse I inhibition preferentially altered splicing on the 3' or 5' end of transcripts.

The junction probe analysis was conducted as follows:

1. RMA normalized $\log_2$ expression values for each probe were loaded into a MySQL table. RMA normalization was conducted with the Partek Genomics Suite with summarization set to 'None' in order to retain probe level values.
2. Any expression value below 4.4 was increased to 4.4. Downstream processing works with the ratio of test to control values so this correction is done to prevent

large ratios caused by fluctuations of values that are in the noise range of the instrument.

3. The ratio for each probe of test (CPT) vs control was calculated. The 1 hour, 2 hour, and 4 hour were match with the 4 hour control. The 15 and 20 hour samples were matched with the 20 hour control. Ratio = the log2 difference of test - control.
4. The standard deviation of the probes that make up the following groups was calculated: "E" probes, "C & D" probes, and "T & F" probes. There are actually 3 physical probes for each of these probe types in a splice event detection group.
5. Probe groups that did not show consistent measurements (standard deviation > 1.5) were filtered out. This is done to avoid misleading average values / results that can be cause by one probe with anomalous values.
6. Probes that showed no expression above background for both test and control samples were eliminated. "E" probe groups with less than 2 remaining probes and "C & D" and "T & F" groups with less than 4 remaining probes were filtered out. This step prevents non-responsive probes from influencing probe group averages and filters out probe groups with an insufficient number of responsive probes.
7. Average values for each probe group identified in step 4 was calculated (e.g average E probes, put C&D probes in a single group and calculate the average, etc.)

8. The max difference between control and test for "early" arrays (1H, 2H, 4H) and max difference for "late" arrays (15H, 20H) was identified. For probes with negative change (decrease vs. control) the min value was identified. Note: for max log2 differences value of 1 is a 2 fold change 2 is a 4 fold change 3 is an 8 fold change.
9. E probes were selected where max difference vs control was > 2 (4 fold change). These were separated into early changes that hit a max delta > 2 in the first 4 hours and those that hit > 2 in the late group.
10. The max change in E from step 9 was then compared to the general gene expression change (represented by T&F) to find E changes that differed from overall gene expression changes. (E probe could be measuring the normal variant so this must be done to ensure that the probe is detecting a splicing change rather than a general expression change.)
11. A list was constructed of probesets where the E probe had >= 4 fold change vs. control and where the general gene expression was not changed by the same magnitude. Positive fold change = exon skipping, Negative fold change = transition to exon inclusion.

Steps 9 through 11 above identify changes in long form / short form expression as identified by change in E probes. These steps were repeated with the C&D probe group to find changes that would again indicate a shift to / from the long form of the transcript. The C&D probe group behaves in a manner opposite of the “E” probe group. An

increase in the C&D probe expression in the treated samples indicates a shift to the long form of the transcript while a decrease in C&D expression shows a shift to the short form. The results of the above analysis were split into exon skip events (shift to short form) and exon inclusion events (shift to long form) and into early splice events (those showing > 4 fold difference in the 1, 2, or 4 hours sample) and late splice events (those showing > 4 fold difference in the 15 or 20 hour sample). The exon number of the exon affected by the splice event and the total number of exons in the gene are provided from the SpliceCenter database to aid in searching for a 3' or 5' bias in the splice events. Note that many of the thresholds used in this analysis could be relaxed but the first cut was focused on identification of high confidence splice events.

The E probe splice events are as follows:

Late Exon Skip Events

Probe Set	Symbol	Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Skip Exon	Gene Exons
25764.007.1	HYPK	2.51	6.88	3	4
10528.037.1	NOL5A	3.92	6.06	11	12
9939.018.1	RBM8A	2.32	5.70	3	4
84817.006.1	TXNDC17	0.17	5.01	2	4
6303.025.1	SAT1	1.79	4.95	4	6
967.028.1	CD63	2.15	4.92	6	8
54663.003.2	WDR74	2.18	4.75	11	12
25949.001.1	SYF2	1.63	4.66	3	7
80279.027.1	CDK5RAP3	2.21	4.65	10	12
114789.010.1	SLC25A25	1.67	4.60	7	12
4904.003.1	YBX1	0.83	4.51	4	8
54663.004.1	WDR74	1.23	4.49	11	12
11047.002.1	ADRM1	-2.23	4.43	3	10
6176.002.1	RPLP1	0.92	4.23	2	4
467.011.1	ATF3	1.58	4.21	3.5	4
79085.017.1	SLC25A23	1.09	4.15	6	15
5718.009.1	PSMD12	1.12	4.13	5	14
51596.009.1	CUTA	0.00	4.10	4	5
11331.001.1	PHB2	0.72	3.94	4	9
79174.004.1	CRELD2	0.66	3.90	9	11
60496.007.1	AASDHPPT	0.23	3.85	5	6
5310.032.1	PKD1	3.09	3.76	25	47
6303.022.2	SAT1	1.80	3.67	3	6

6303.020.1	SAT1	0.97	3.62	3	6
374897.003.1	SBSN	0.28	3.59	1.5	5
5713.011.1	PSMD7	0.00	3.59	4	8
328.021.1	APEX1	-0.75	3.55	3	5
6950.007.1	TCP1	1.22	3.54	7	13
26986.048.1	PABPC1	0.58	3.45	9	16
57142.037.1	RTN4	0.00	3.45	14	15
1778.030.1	DYNC1H1	0.21	3.43	17	78
5757.015.1	PTMA	0.98	3.43	5	6
5052.020.1	PRDX1	0.18	3.43	5	6
51596.001.1	CUTA	-1.20	3.27	5	5
4637.004.1	MYL6	1.01	3.26	4	8
10484.020.1	SEC23A	0.20	3.19	19	22
10521.013.1	DDX17	0.12	3.14	7	15
23015.028.1	GOLGA8A	0.12	3.14	15	23
1329.005.1	COX5B	-1.26	3.08	3	4
10329.008.1	TMEM5	0.05	3.08	5	6
7919.005.1	BAT1	1.08	3.08	3	12
10038.012.1	PARP2	-1.46	3.07	6	14
113246.005.1	C12orf57	0.92	3.05	2	3
201134.023.1	CCDC46	-0.59	3.04	15	29
283248.014.1	RCOR2	0.00	3.03	11	12
5902.011.1	RANBP1	0.00	3.00	7	7
51693.010.1	TRAPPC2L	1.13	2.99	4	5
54107.014.1	POLE3	0.56	2.98	2.5	5
4924.038.1	NUCB1	0.69	2.96	5	13
1973.016.1	EIF4A1	0.00	2.92	8	9
60481.011.1	ELOVL5	0.00	2.90	8	10
1195.019.1	CLK1	0.00	2.89	5	12
79168.011.1	LILRA6	-0.25	2.88	6	8
51491.003.1	NOP16	0.20	2.88	4	6
28988.029.1	DBNL	2.34	2.86	12	13
6923.003.1	TCEB2	-0.20	2.85	3	5
9790.021.1	BMS1	0.00	2.82	15	22
10483.013.1	SEC23B	0.92	2.82	18	21
4831.012.1	NME2	0.39	2.82	8	7
440270.008.1	GOLGA8B	0.00	2.81	15	24
23379.019.1	KIAA0947	0.46	2.80	7	19
550.014.1	AUP1	1.70	2.79	8	10
2673.010.1	GFPT1	0.00	2.78	8	19
7169.026.1	TPM2	1.42	2.77	4	7
984.005.1	CDC2L1	-0.76	2.76	8	30
968.003.1	CD68	3.20	2.76	2	6
4077.033.1	NBR1	2.15	2.73	19	23
823.037.1	CAPN1	0.12	2.72	14	23
29015.026.1	SLC43A3	1.33	2.72	8	17
11267.015.1	SNF8	-0.43	2.71	7	8
1196.014.1	CLK2	1.95	2.68	2	11
57721.008.1	KIAA1627	2.31	2.65	2	11
1349.007.1	COX7B	0.00	2.64	2	3
10061.013.1	ABCF2	2.20	2.63	6	16
4831.018.1	NME2	0.98	2.61	3	7
60509.021.1	AGBL5	0.69	2.60	4	16

23396.006.1	PIP5K1C	1.02	2.60	3	18
89797.024.1	NAV2	0.62	2.59	13	44
84191.003.1	FAM96A	-0.06	2.57	1	5
94039.007.1	ZNF101	0.98	2.52	4	6
85359.004.1	DGCR6L	1.05	2.51	3	5
833.033.1	CARS	1.06	2.51	23	24
23204.011.1	ARL6IP1	0.00	2.51	3	6
55863.002.1	TMEM126B	0.01	2.51	2	6
10969.014.1	EBNA1BP2	0.00	2.50	6	8
2193.021.1	FARSA	1.30	2.49	5	13
26227.027.1	PHGDH	0.84	2.49	7	12
718.016.1	C3	0.96	2.45	2	41
94104.011.1	C21orf66	-1.31	2.45	10	19
5426.026.1	POLE	1.00	2.45	18	50
23397.014.1	NCAPH	1.28	2.44	7	18
9261.007.1	MAPKAPK2	1.31	2.43	11	11
301.023.1	ANXA1	0.31	2.42	10	13
10311.008.1	DSCR3	0.96	2.41	2	8
25890.044.1	ABI3BP	1.43	2.41	55	65
5451.011.1	POU2F1	-0.92	2.39	11	21
4629.038.1	MYH11	2.01	2.37	19	43
9406.013.1	ZRANB2	0.29	2.35	3	11
6628.026.1	SNRPB	0.87	2.32	7	7
27309.011.1	ZNF330	0.23	2.32	7	10
220.009.1	ALDH1A3	0.69	2.32	6	13
57666.008.1	KIAA1545	-1.30	2.31	1.5	10
48.017.1	ACO1	0.00	2.30	10	22
10915.013.1	TCERG1	0.60	2.29	4	22
56647.003.1	BCCIP	0.02	2.27	4	9
3837.041.1	KPNB1	0.76	2.27	19	22
375790.031.1	AGRN	0.94	2.26	20	36
25764.005.1	HYPK	0.76	2.25	3	4
23371.024.1	TENC1	-0.15	2.24	17	30
9883.019.1	POM121	0.00	2.24	5	15
11056.004.2	DDX52	1.44	2.24	3.5	16
11176.005.1	BAZ2A	1.02	2.23	6	29
949.005.1	SCARB1	-0.20	2.23	16	16
6433.009.1	SFRS8	1.50	2.22	2	17
9871.018.1	SEC24D	1.82	2.22	9	30
9344.003.3	TAOK2	-1.61	2.20	12.5	19
375056.003.1	MIA3	-0.09	2.19	13	28
29081.015.1	METTL5	0.98	2.19	2	7
8737.011.1	RIPK1	0.54	2.19	3	10
3550.020.1	IK	0.00	2.16	10	20
3710.032.1	ITPR3	0.71	2.15	18	58
145482.010.1	PTGR2	0.50	2.14	5	8
9538.014.1	EI24	0.00	2.13	3	12
26098.013.1	C10orf137	0.00	2.13	4	26
666.004.1	BOK	1.61	2.12	5	5
84447.023.1	SYVN1	0.49	2.12	14	16
56339.005.2	METTL3	0.11	2.12	8	10
64118.002.2	DUS1L	1.30	2.11	10.5	14
1209.017.1	CLPTM1	0.68	2.11	4	16

1936.062.1	EEF1D	-0.22	2.11	12	13
5705.033.1	PSMC5	0.54	2.11	10	13
1277.039.1	COL1A1	0.73	2.10	28	51
5476.012.1	CTSA	-0.34	2.08	5	15
79050.016.1	NOC4L	0.82	2.08	13	15
2194.033.1	FASN	0.52	2.07	16	43
2013.003.1	EMP2	1.09	2.07	2	5
114990.002.1	VASN	0.45	2.05	2	2
5436.005.1	POLR2G	0.00	2.05	3	8
6464.044.1	SHC1	0.27	2.04	11	13
79065.024.1	ATG9A	0.75	2.03	9	16
30827.015.1	CXXC1	-1.07	2.03	2.5	15
55299.005.1	BXDC2	0.47	2.02	2	5
30817.023.1	EMR2	0.11	2.02	14	20
4967.041.1	OGDH	1.07	2.01	19	25
6950.003.1	TCP1	0.00	2.01	2	13
6810.019.1	STX4	0.20	2.00	12	13

Early Exon Skip Events

Probe Set	Symbol	Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Skip Exon	Gene Exons
10528.037.1	NOL5A	3.92	6.06	11	12
968.003.1	CD68	3.20	2.76	2	6
5310.032.1	PKD1	3.09	3.76	25	47
467.014.1	ATF3	2.95	3.56	4	4
5872.002.1	RAB13	2.61	-1.26	1.5	8
55870.026.1	ASH1L	2.51	1.02	22	28
25764.007.1	HYPK	2.51	6.88	3	4
1643.004.2	DDB2	2.34	-1.91	7	10
28988.029.1	DBNL	2.34	2.86	12	13
9939.018.1	RBM8A	2.32	5.70	3	4
57721.008.1	KIAA1627	2.31	2.65	2	11
137682.005.1	C8orf38	2.30	-0.55	4	15
80279.027.1	CDK5RAP3	2.21	4.65	10	12
10061.013.1	ABCF2	2.20	2.63	6	16
54663.003.2	WDR74	2.18	4.75	11	12
967.028.1	CD63	2.15	4.92	6	8
4077.033.1	NBR1	2.15	2.73	19	23
27250.001.1	PDCD4	2.08	-2.73	3	13
88455.020.1	ANKRD13A	2.06	0.73	15	16
51586.007.1	MED15	2.03	-1.98	11	19
9780.003.1	FAM38A	2.03	-0.43	34	41
4629.038.1	MYH11	2.01	2.37	19	43

Late Exon Include Events

Probe Set	Symbol	Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Include Exon	Gene Exons
8087.034.1	FXR1	0.27	-4.08	16	18
10431.008.1	TIMM23	0.75	-3.82	6.5	7
8394.012.1	PIP5K1A	-0.24	-3.52	7.5	16

55192.003.1	DNAJC17	0.42	-3.48	5.5	11
10436.003.1	EMG1	0.26	-3.37	2.5	8
2804.008.1	GOLGB1	-0.17	-3.17	1.5	22
79876.002.1	UBA5	-0.52	-3.08	2.5	13
9101.002.1	USP8	-0.14	-2.95	1	22
80011.022.1	NIP30	-0.55	-2.94	1.5	7
10016.008.1	PDCD6	0.43	-2.93	6.5	7
4149.001.1	MAX	-0.22	-2.90	2	6
23499.020.1	MACF1	-0.59	-2.88	102.5	105
221937.004.1	FOXK1	-1.10	-2.86	3.5	9
81555.006.2	YIPF5	0.65	-2.81	1	6
405.005.1	ARNT	0.67	-2.80	4.5	22
27250.001.1	PDCD4	2.08	-2.73	3	13
6760.015.1	SS18	-0.67	-2.71	3.5	13
57154.008.1	SMURF1	-0.24	-2.64	1.5	19
2289.008.1	FKBP5	-1.24	-2.63	3.5	14
23514.016.1	KIAA0146	0.87	-2.62	5.5	23
5110.007.1	PCMT1	0.46	-2.58	2.5	8
54903.002.3	MKS1	-0.86	-2.50	11.5	18
27102.011.1	EIF2AK1	-0.51	-2.42	1.5	15
7247.007.2	TSN	0.75	-2.35	1.5	6
84717.005.1	HDGF2	0.49	-2.35	6	17
90410.001.1	IFT20	0.48	-2.32	5	7
123811.008.1	C16orf63	0.30	-2.23	10	5
64963.003.1	MRPS11	0.41	-2.21	2.5	6
10527.001.1	IPO7	-0.35	-2.19	11	25
55197.010.1	RPRD1A	0.31	-2.17	2	9
4363.023.1	ABCC1	0.78	-2.16	10	31
11130.001.1	ZWINT	1.08	-2.15	2	6
23190.003.1	UBXN4	0.38	-2.14	16	12
84312.001.1	BRMS1L	-0.41	-2.13	2	11
387032.008.1	ZKSCAN4	0.66	-2.05	4	5
989.028.1	SEPT7	0.58	-2.03	5.5	14
54904.010.1	WHSC1L1	-1.59	-2.01	1.5	25

Early Exon Include Events

Probe Set	Symbol	Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Include Exon	Gene Exons
10308.002.1	ZNF267	-3.07	1.08	3.5	4
4716.001.1	NDUFB10	-2.75	0.00	2	3
54464.024.1	XRN1	-2.25	0.99	17	42
162967.003.1	ZNF320	-2.23	0.78	3.5	4
11047.002.1	ADRM1	-2.23	4.43	3	10
57570.002.1	TRMT5	-2.13	1.12	1.5	5
84527.014.1	ZNF559	-2.01	0.94	2.5	7

The C&D probe splice events are as follows:

Late Exon Include Events

Probe Set	Symbol		Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Include Exon	Gene Exons
50848.034.1	F11R	CL	0.12	2.46	9	10
84908.001.1	FAM136A	CL	0.61	2.22	2	4
2194.049.1	FASN	CL	-1.19	2.02	32	43

Late Exon Skip Events

Probe Set	Symbol		Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Skip Exon	Gene Exons
56339.005.2	METTL3	CLN	0.31	-3.57	8	10
145553.003.3	MDP-1	CLN	0.09	-3.23	4	4
55165.007.1	CEP55	CLN	0.10	-3.16	4	9
6760.009.1	SS18	CLN	0.47	-3.15	3	13
2804.009.1	GOLGB1	CLN	-0.09	-2.85	2	22
23112.018.1	TNRC6B	CLN	-0.20	-2.61	7	28
26146.009.1	TRAF3IP1	CLN	0.22	-2.57	6	17
2804.019.1	GOLGB1	CLN	0.42	-2.53	13	22
55706.013.1	TMEM48	CLN	0.54	-2.52	3	18
23512.003.1	SUZ12	CLN	0.32	-2.45	2	16
63967.007.1	CLSPN	CLN	0.41	-2.41	2	27
80005.049.1	DOCK5	CLN	-0.30	-2.39	50	52
56252.012.1	YLPM1	CLN	-0.36	-2.24	2	21
4605.021.1	MYBL2	CLN	0.95	-2.15	9	14
91869.008.1	RFT1	CLN	0.47	-2.09	5	13

Early Exon Skip Events

Probe Set	Symbol		Max Early Log2 Diff vs Control	Max Late Log2 Diff vs Control	Skip Exon	Gene Exons
91661.001.2	ZNF765	CEN	-2.43	1.10	5	12

RT-PCR validation of the splice events identified by the simple junction analysis is currently in progress. If these high confidence splice events are validated and if they are representative of the splicing changes caused by topoisomerase inhibition, then a few general conclusions may be draw. The affect of topoisomerase on alternative splicing has been shown on a few genes in previous studies [51, 52]. The 217 distinct splice events

on 204 different genes in this study demonstrate a broad impact of CPT treatment on splicing and provide further evidence for the role of topoisomerase in splicing regulation. The results also indicate that the majority of splicing events caused by CPT treatment are not observed until 15-20 hours after treatment. Only 30 of the 217 splice events were detected with greater than a 2 $\log_2$ difference in the 1-4 hour time points. This may indicate the time required for CPT treatment to affect the kinase activity of topoisomerase or may simple indicate that mRNA turn over takes many hours for most genes. Finally, the exon position of the splicing events showed that 48.5% of events occur on the 3' end of transcripts and 51.5% of events occur on the 5' end of transcripts indicating a relatively even distribution of the position of splicing changes in transcripts. SpliceCenter data and utilities were required in order to perform the high throughput analysis presented in this section.

REFERENCES

REFERENCES

1. Robinson, M.O. and M.I. Simon, *Determining transcript number using the polymerase chain reaction: Pgk-2, mP2, and PGK-2 transgene mRNA levels during spermatogenesis.* Nucleic Acids Res, 1991. **19**(7): p. 1557-62.

2. Hannon, G.J. and J.J. Rossi, *Unlocking the potential of the human genome with RNA interference.* Nature, 2004. **431**(7006): p. 371-8.

3. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.

4. Lequin, R.M., *Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA).* Clin Chem, 2005. **51**(12): p. 2415-8.

5. Lodish, H., et al., in *Molecular Cell Biology*. 2004, W.H. Freeman and Company: New York.

6. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science, 2003. **302**(5653): p. 2141-4.

7. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

8. Garcia-Blanco, M.A., A.P. Baraniak, and E.L. Lasda, *Alternative splicing in disease and therapy.* Nat Biotechnol, 2004. **22**(5): p. 535-46.

9. Hu, G.K., et al., *Predicting splice variant from DNA chip expression data.* Genome Res, 2001. **11**(7): p. 1237-45.

10. Blencowe, B.J., *Alternative splicing: new insights from global analyses.* Cell, 2006. **126**(1): p. 37-47.

11. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.

12. Buratti, E., M. Baralle, and F.E. Baralle, *Defective splicing, disease and therapy: searching for master checkpoints in exon definition.* Nucleic Acids Res, 2006. **34**(12): p. 3494-510.

13. Lee, C.J. and K. Irizarry, *Alternative splicing in the nervous system: an emerging source of diversity and regulation.* Biol Psychiatry, 2003. **54**(8): p. 771-6.

14. Venables, J.P., *Aberrant and alternative splicing in cancer.* Cancer Res, 2004. **64**(21): p. 7647-54.

15. Faustino, N.A. and T.A. Cooper, *Pre-mRNA splicing and human disease.* Genes Dev, 2003. **17**(4): p. 419-37.

16. Roy, M., Q. Xu, and C. Lee, *Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms.* Nucleic Acids Res, 2005. **33**(16): p. 5026-33.

17. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.

18. Kim, N., et al., *The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species.* Nucleic Acids Res, 2007. **35**(Database issue): p. D93-8.

19. Stamm, S., et al., *ASD: a bioinformatics resource on alternative splicing.* Nucleic Acids Res, 2006. **34**(Database issue): p. D46-55.

20. de la Grange, P., et al., *A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants.* BMC Bioinformatics, 2007. **8**: p. 180.

21. Holste, D., et al., *HOLLYWOOD: a comparative relational database of alternative splicing.* Nucleic Acids Res, 2006. **34**(Database issue): p. D56-62.

22. Bhasi, A., et al., *EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes.* Bioinformatics, 2007. **23**(14): p. 1815-23.

23. Kim, P., et al., *ECgene: genome annotation for alternative splicing.* Nucleic Acids Res, 2005. **33**(Database issue): p. D75-9.

24. Rambaldi, D., et al., *Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data.* BMC Bioinformatics, 2007. **8 Suppl 1**: p. S17.

25. Kahn, A.B., et al., *SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis.* BMC Bioinformatics, 2007. **8**: p. 75.

26. Wang, H., et al., *Gene structure-based splice variant deconvolution using a microarray platform.* Bioinformatics, 2003. **19 Suppl 1**: p. i315-22.

27. Srinivasan, K., et al., *Detection and measurement of alternative splicing using splicing-sensitive microarrays.* Methods, 2005. **37**(4): p. 345-59.

28. *Technical Note: Identifying and Validating Alternative Splicing Events*, Affymetrics Incorperated.

29. *Partek Genomics Suite*. Available from: http://www.partek.com/software.

30. Purdom, E., et al., *FIRMA: a method for detection of alternative splicing from exon array data.* Bioinformatics, 2008. **24**(15): p. 1707-14.

31. *NCBI Evidence Viewer*. Available from: <http://www.ncbi.nlm.nih.gov/sutils/static/evvdoc.html>.

32. Liu, H., et al., *AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets.* Bioinformatics, 2007. **23**(18): p. 2385-90.

33. *NCBI RefSeq*. Available from: <http://www.ncbi.nlm.nih.gov/RefSeq/>.

34. *NCBI GenBank*. Available from: <http://www.ncbi.nlm.nih.gov/Genbank/>.

35. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

36. Lewis, B.P., R.E. Green, and S.E. Brenner, *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A, 2003. **100**(1): p. 189-92.

37. Lee, C., *Generating consensus sequences from partial order multiple sequence alignment graphs.* Bioinformatics, 2003. **19**(8): p. 999-1008.

38. Xing, Y., A. Resch, and C. Lee, *The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures.* Genome Res, 2004. **14**(3): p. 426-41.

39. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.

40. *NCBI Conserved Domain Database*. Available from: <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

41. *Primer Match*. Available from: http://www.umiacs.umd.edu/~nedwards/research/primer_match.html.

42. Dallas, P.B., et al., *Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR -- how well do they correlate?* BMC Genomics, 2005. **6**(59).

43. Pei, Y. and T. Tuschl, *On the art of identifying effective and specific siRNAs.* Nat Methods, 2006. **3**(9): p. 670-6.

44. *NCBI RNAi*. Available from: http://www.ncbi.nlm.nih.gov/projects/genome/rnai/.

45. Martin, S.E., et al., *Multiplexing siRNAs to compress RNAi-based screen size in human cells.* Nucleic Acids Res, 2007. **35**(8): p. e57.

46. Lee, J.C., et al., *A detailed transcript-level probe annotation reveals alternative splicing based microarray platform differences.* BMC Genomics, 2007. **8**(284).

47. Bourdon, J.C., *p53 and its isoforms in cancer.* Br J Cancer, 2007. **97**(3): p. 277-82.

48. Bourdon, J.C., et al., *p53 isoforms can regulate p53 transcriptional activity.* Genes Dev, 2005. **19**(18): p. 2122-37.

49. Xu, C. and B. Ma, *Software for computational peptide identification from MS-MS data.* Drug Discov Today, 2006. **11**(13-14): p. 595-600.

50. *Kolmogorov-Smirnov* Available from: http://www.physics.csbsju.edu/stats/KS-test.html.

51. Pilch, B., et al., *Specific inhibition of serine- and arginine-rich splicing factors phosphorylation, spliceosome assembly, and splicing by the antitumor drug NB-506.* Cancer Res, 2001. **61**(18): p. 6876-84.

52. Solier, S., et al., *Topoisomerase I and II inhibitors control caspase-2 pre-messenger RNA splicing in human cells.* Mol Cancer Res, 2004. **2**(1): p. 53-61.

53. Zeeberg, B.R., et al., *High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID).* BMC Bioinformatics, 2005. **6**: p. 168.

54. Laiho, A., *Personal Communication - Finnish DNA Microarray Centre* 2008.

55. Aziz, M., *Personal Communication - Translational Genomics Research Institute (TGen)*. 2008.

56. Daemen, A., *Personal Communication - Katholieke Universiteit Leuven*. 2008.

57. Yera, E.R., *Personal Communication - University of California San Francisco*. 2008.

58. Collins, C., *Personal Communication - The Children's Hospital Boston*. 2008.

59. Reimers, M., *Personal Communication - Virginia Commonwealth University*. 2008.

60. Webster, D., *Personal Communication - University of Wisconsin*. 2007.

61. Guo, X., *Personal Communication - Oncogenomics group NCI Pediatric Oncology group*. 2008.

62. Jamison, C., *Personal Communciation - Cincinnati Children's Hospital Center*. 2009.

63. Christoforou, A., *Personal Communication - The University of Edinburgh*. 2009.

64. Ryan, M.C., et al., *SpliceCenter: a suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies.* BMC Bioinformatics, 2008. **9**: p. 313.

65. Anderson, E., Khvorova, A, Karpilow, J, *Identifing siRNA-Induced Off-Targets by Microarray Analysis*, in *Methods in Molecular Biology*, Humana Press: Totowa, NJ.

66. Nielsen, C.B., et al., *Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.* RNA, 2007. **13**(11): p. 1894-910.

67. Chalk, A.M. and E.L. Sonnhammer, *siRNA specificity searching incorporating mismatch tolerance data.* Bioinformatics, 2008. **24**(10): p. 1316-7.

68. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.

CURRICULUM VITAE

Michael Ryan was born on July 20, 1966, in Baltimore, Maryland, and is a citizen of the Unites States. He graduated from Calvert Hall High School, Towson, Maryland, in 1984. He was awarded a Master's of Science in Bioinformatics from George Mason University in 2006 and a Bachelor of Science in Computer Science with High Honors from the College of William and Mary in 1988.

Mr. Ryan has been employed as a professional software developer since 1988. He began working as an independent contractor in 1995 and in 1997 founded Tiger Team Consulting in order to establish a company that fosters a good working environment for senior software engineers. Tiger Team has grown to a 22 employee company and Mr. Ryan is an owner and managing partner of Tiger Team. In addition to company management, Mr. Ryan performs software architecture design, bioinformatics development, performance tuning, project management, implementation of critical system components, and technical mentoring. Key projects have included: a network management system for Sprint International, the Billing Data Entry System for a European telecom, an on-line brokerage system for FolioFn, the Advisen internet site for insurance professionals, a modernization project for US Customs and genetic sequence search system for the US Patent Office. Currently Mr. Ryan works with the National Cancer Institute on a variety of bioinformatics applications and with Johns Hopkins on genome-wide analysis of SNPs and protein structure.

Prior to founding Tiger Team, Mr. Ryan worked for NASA and American Management Systems. For NASA, he developed ground data systems for an earth observing satellite, TRMM. For AMS, he worked on a wide variety of large-scale software development projects for telecom, imaging, and hybrid mail applications.

Publications

Kahn AB, Ryan MC, Liu H, Zeeberg BR, Jamison DC, Weinstein JN (2007). SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis. *BMC Bioinformatics*, **8**:75.

Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC, Kane DW, Weinstein JN (2007). AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. *Bioinformatics*, **23**(18):2385-90.

Ryan M, Zeeberg B, Caplen N, Cleland J, Kahn A, Liu H, Weinstein J, SpliceCenter (2008). A suite of web-based bioinformatics applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. BMC Bioinformatics, **9**:313.