Finding non-coding mRNA splicing variants using the Shannon Human Splicing Pipeline

Presented by Ben Shirley

www.cytognomix.com

Shirley BC, Mucaki EJ, Whitehead T, Costea PI, Akan P, Rogan PK. Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. Genomics Proteomics Bioinformatics. 2013 Mar 14. PubMed PMID: 23499923.

Cytognomix Inc.

- Cytognomix is a London Ontario-based biotechnology company developing software-based solutions and reagents for clinical genomics.

- The company's intellectual property portfolio emphasizes products for clinical diagnostic applications.

- Current products include mutation interpretation software, DNA hybridization reagents, and companion software to detect human chromosome abnormalities.

Motivation

- Widespread adoption of Next Generation Sequencing (NGS) has facilitated sequencing of large numbers of exomes and full genomes.

- Sequencing detects an overwhelming number of variants of unknown significance (VUS).
  - Which are deleterious and which are benign?

- In addition to translational effects and modifications, variants may affect mRNA splicing.

- mRNA splicing mutations are common in Mendelian (single gene) diseases, and it is likely that they contribute to many complex disorders.

Background

- Like most nucleic acid binding sites, sequences of splice donor, acceptor, and regulatory sites vary.
- The splicing machinery recognizes and processes these combinations of splicing signals.
- Information theory provides a framework to recognize members of a group of structurally and functionally related sequences.
- It models sequence variability inherent in these signals, then predicts which sequence variants may be deleterious.

Molecular information theory

The information gained when a nucleic acid is recognized by a molecular machine (ie. the spliceosome) is accompanied by a decrease in entropy that occurs upon binding:

$$\Delta S = -k_B \ln(2) R.$$

Second law relates information to the enthalpy of the molecular machine:

constant → $-k_B T \ln 2 < q / R_i$ (joules/bit)

q: work; T: temperature; R_i: individual information

$q > 0 \Rightarrow R_i < 0$

Protein

DNA Mutation or unrelated sequence

unbound

$q < 0 \Rightarrow R_i > 0$

Stable protein-DNA complex

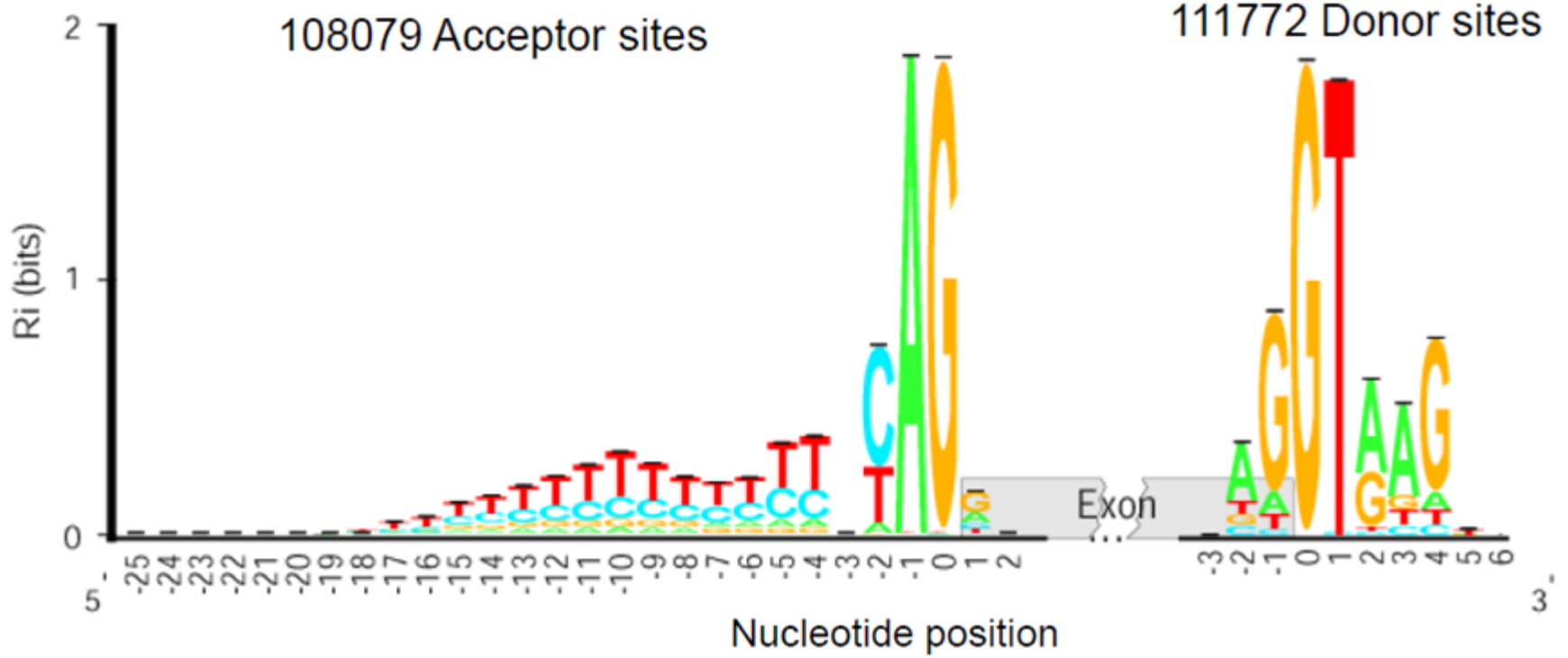
bound

Fold change in affinity < $100/2^{\Delta Ri}$

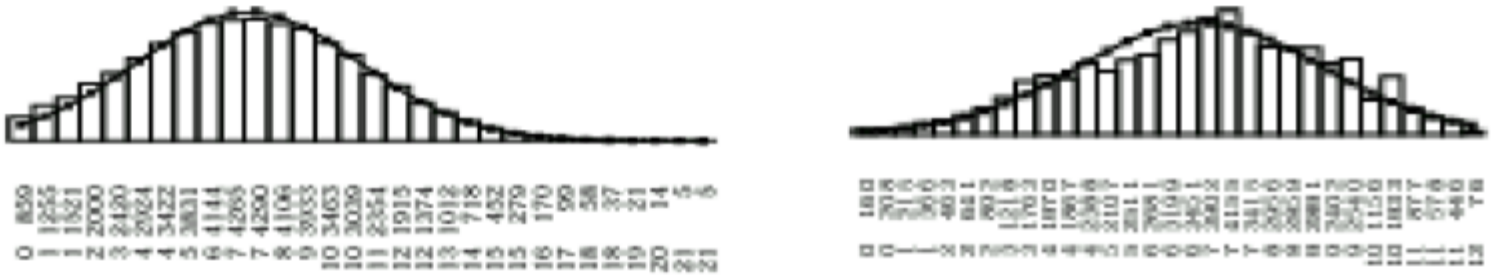
Rogan et al. J Theor Biol **189**:427-41, 1997.

Human splice junction models

(refined from finished human reference sequence)



R_i histograms:

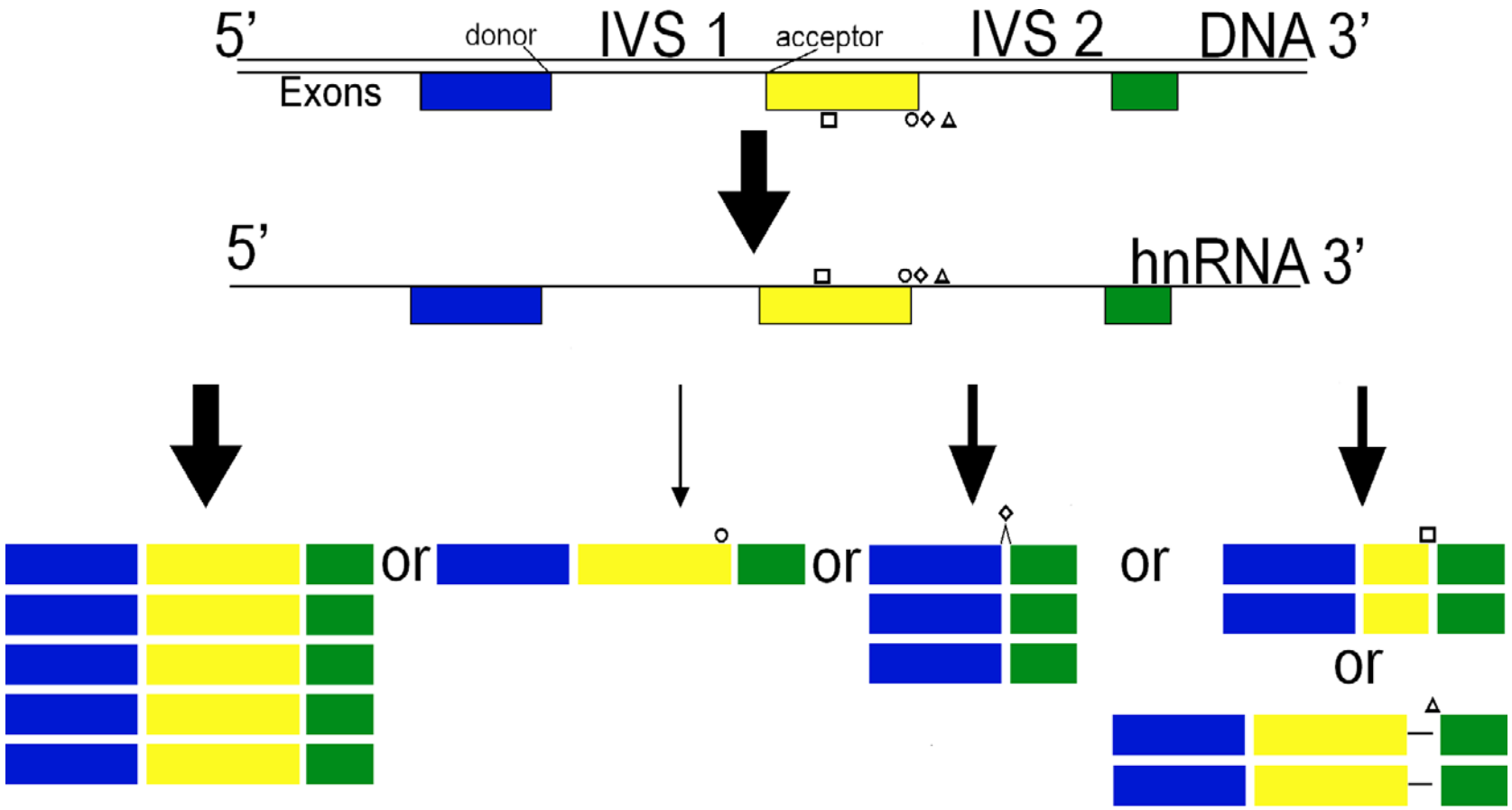


Frequency:

Bits:

$R_{sequence}$ (avg information): 8.87 8.00 bits

Rogan et al. Pharmacogenetics **13**:207-213, 2003



$R_i = R_{i,wildtype}$

$R_i < R_{i,wildtype}$ and $R_i > 0$

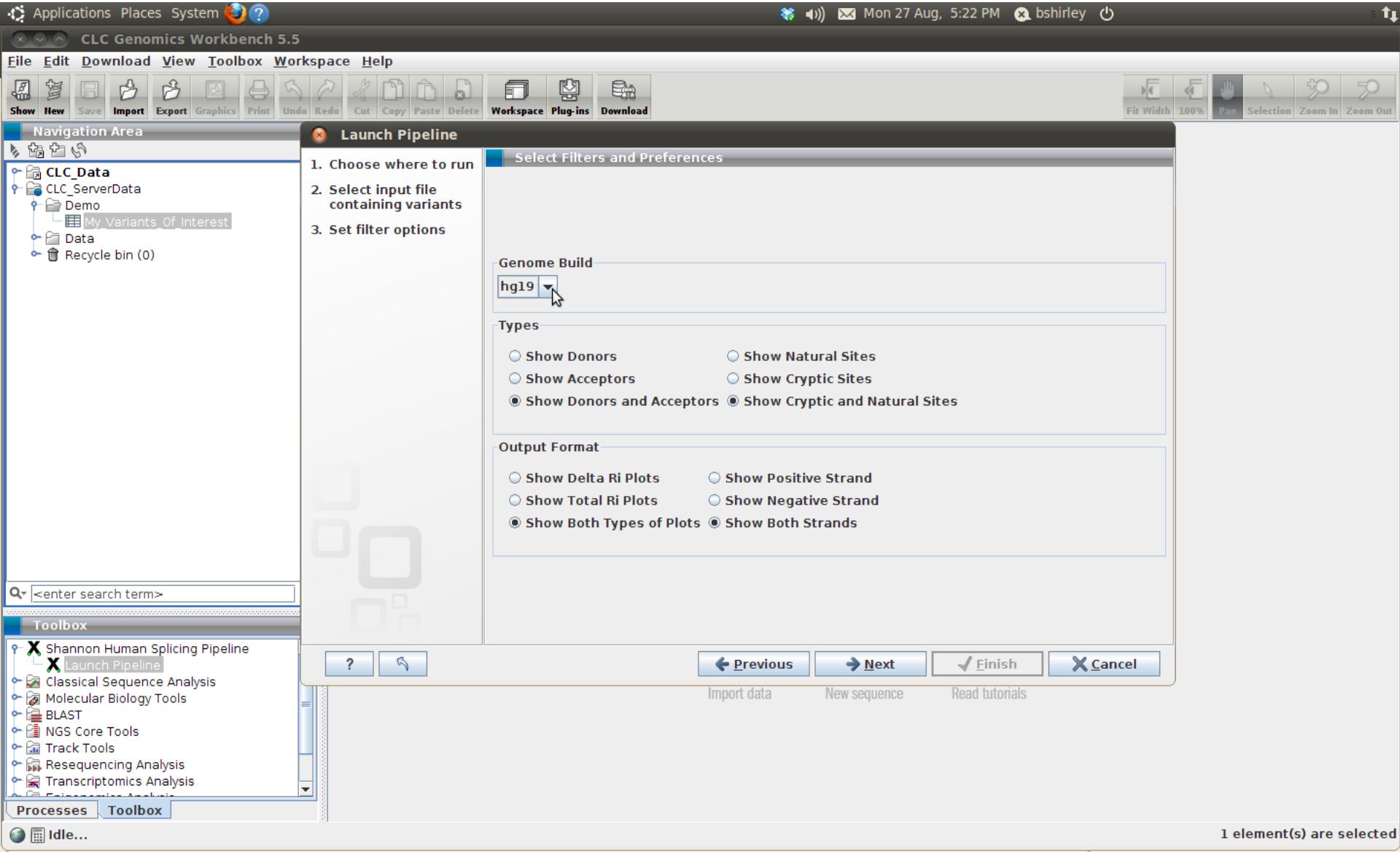
$R_i < R_{i,wildtype}$ and $R_i < 0$

$R_{i,cryptic} > R_{i,natural}$

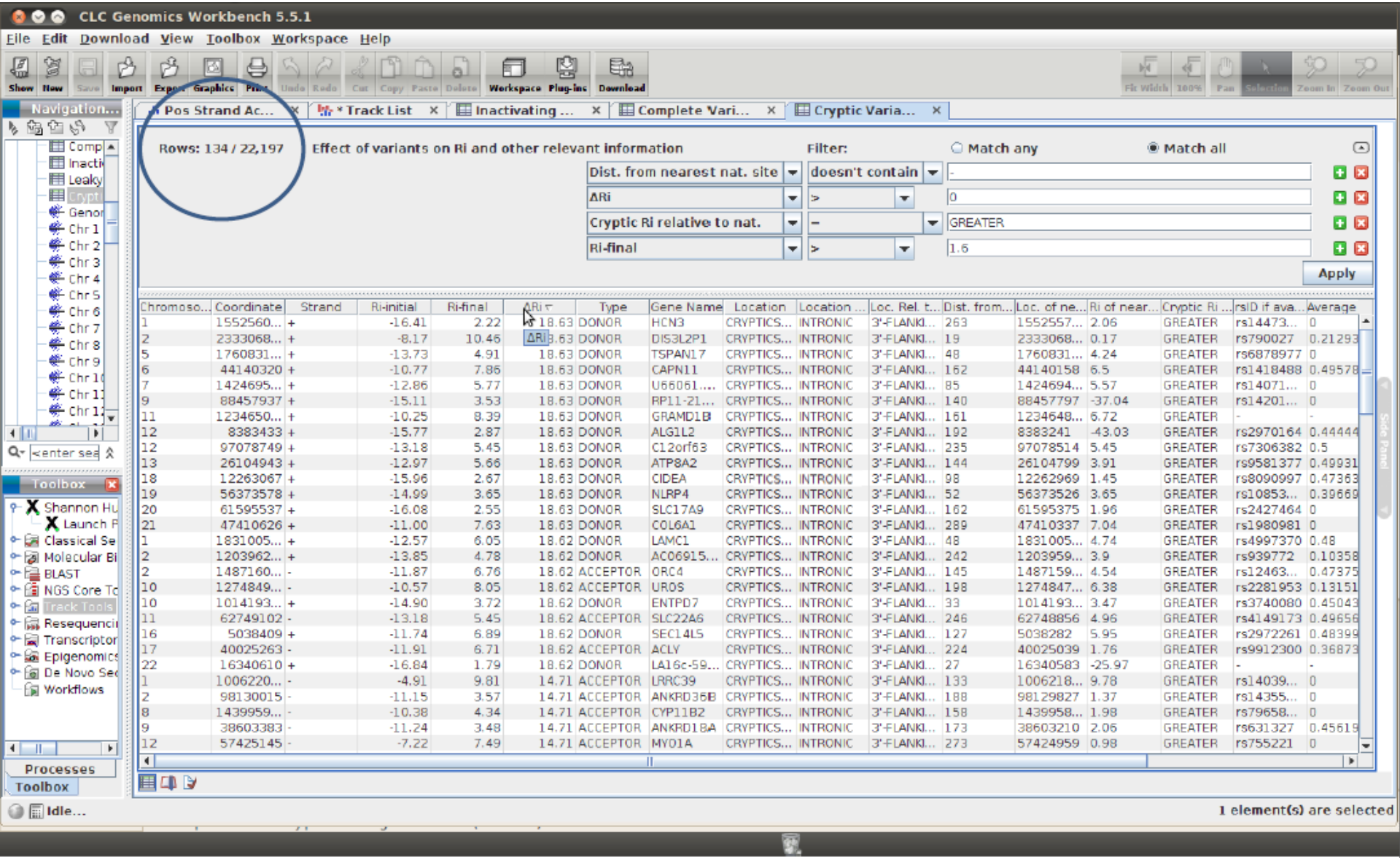
Shannon Pipeline for mutation analysis

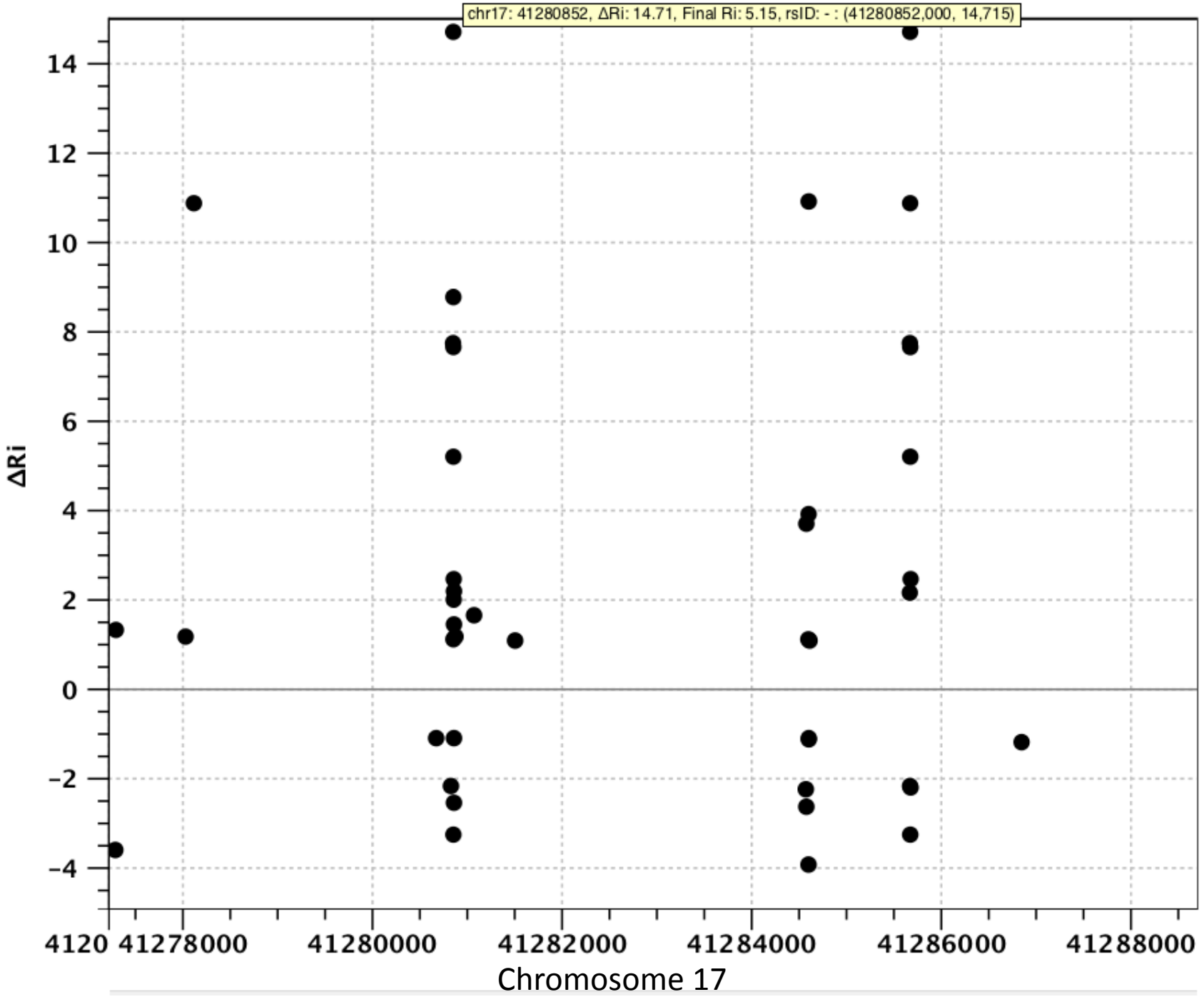
- For prediction of functionally-significant, non-coding variants in genome or exome sequences.
- Mutation analysis on a genome-scale.
- Patented and proven information theory-based binding site analysis (US Pat. 5,867,402):
  - Make quantitative predictions - information related to binding affinity.
  - Can distinguish benign, fully and partially inactivating binding site variants.
  - Common paradigm for all types of nucleic binding sites (eg. splicing, transcription factors).
- Commercial software plugin for the CLC-Bio Genomics Workbench/Server/Grid.
- Algorithm has been validated in hundreds of peer-reviewed research studies of splicing mutations.
- Algorithm recommended by the American College of Medical Genetics and Genomics in their published guidelines and standards (*Genet. Med*. 7, 571-582).
- Predecessor software for single mutation analysis has been designated as a medical device by US FDA (not approved for clinical diagnostics).

Select genome build and filtering options

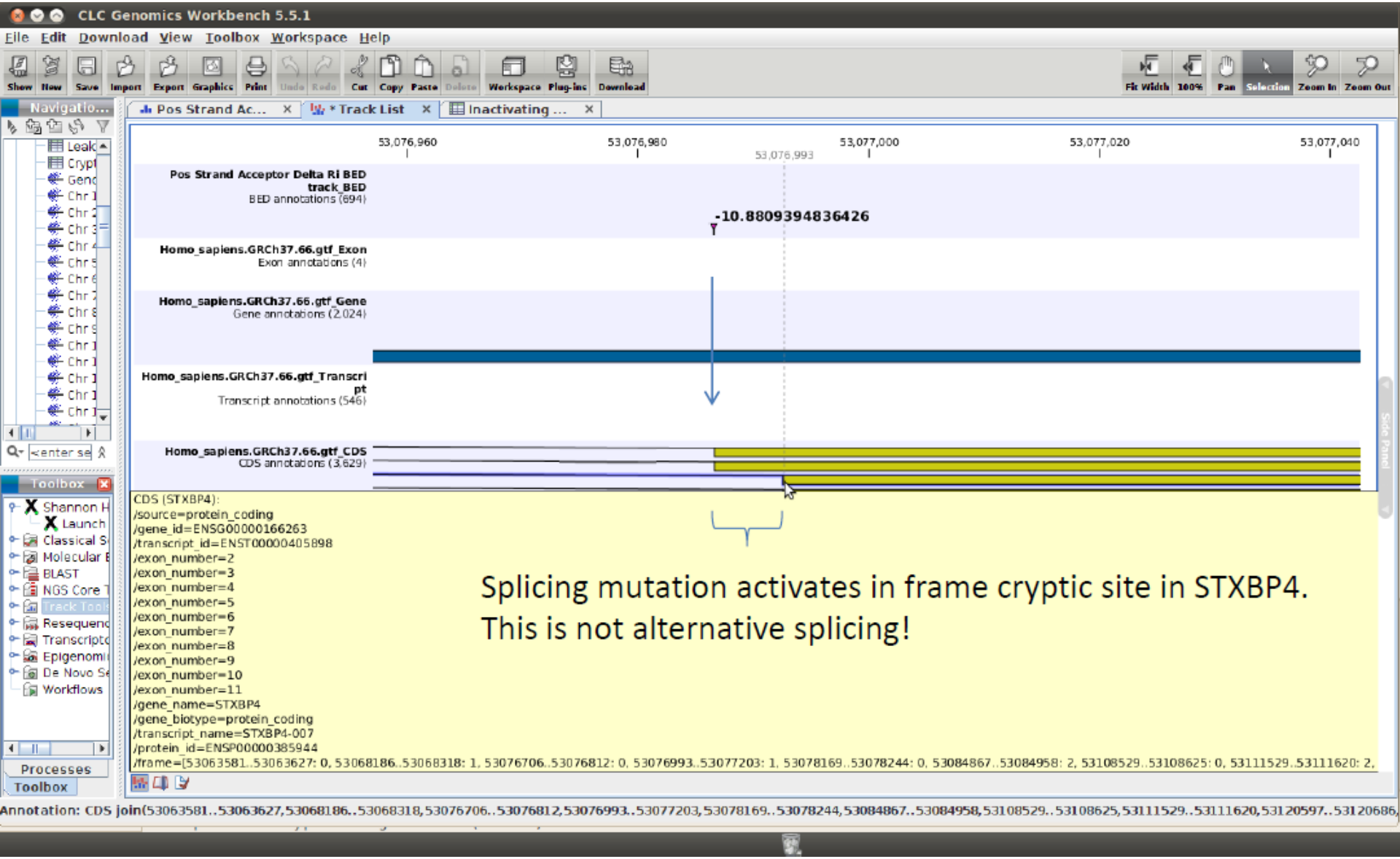


Filtered cryptic splice site variants




chr17: 41280852, ΔRi: 14.71, Final Ri: 5.15, rsID: - : (41280852,000, 14,715)
ΔRi
14
12
10
8
6
4
2
0
-2
-4
4120 41278000
41280000
41282000
41284000
41286000
41288000
Chromosome 17

Mutation viewed in CLC-Bio genome browser



Splicing mutation activates in frame cryptic site in STXBP4.
This is not alternative splicing!

As part of a larger Pipeline

- Whole-genome (or exome) next generation sequencing.
- Variant caller determines variants.
- Obtain List of variants in VCF format.

✓ Genome-wide information analysis is performed for all variants

✓ Variants contributing to changes in information content of binding sites are annotated against standard databases

  - Ensembl, Refseq, dbSNP

✓ Prospective deleterious mutations are categorized as inactivating, leaky, or cryptic site variants.

✓ Results displayed as exportable tables, plots, and genome browser custom tracks.

✓ Results can be filtered to further reduce the number of potentially deleterious variants.

- Examine potential deleterious variants in the laboratory.

Performance of Shannon Pipeline for human mRNA splicing mutation prediction

Source of variants	Number of variants analyzed	Running time*
U2OS cell line	211,049	1h 12m
A431 cell line	290,589	1h 17m
U251 cell line	314,637	1h 20m
ESP 6500 Exomes	1,872,893	2h 35m

Note *Intel I7 CPU with 16 Gb RAM

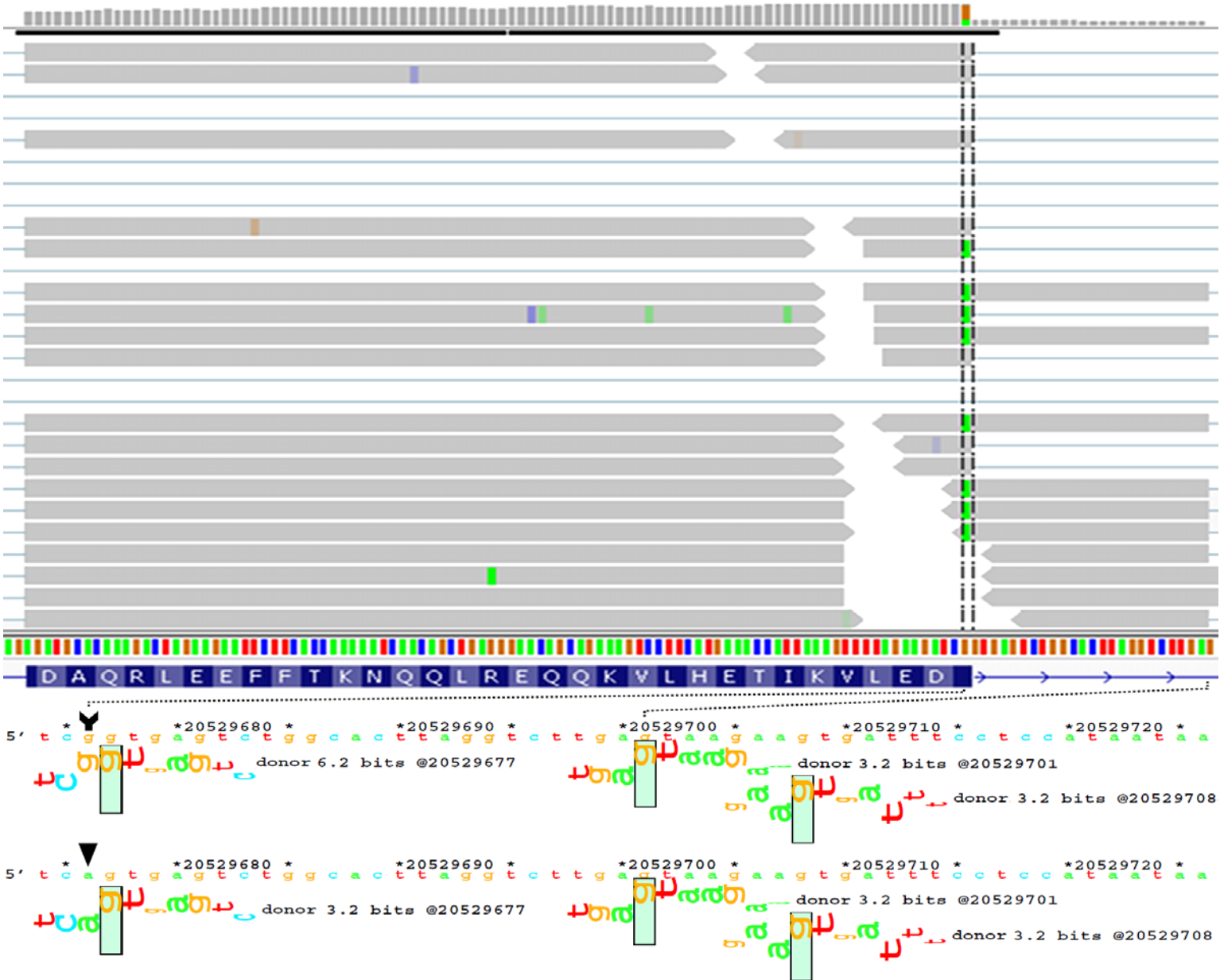
Shannon Pipeline execution time is heavily dependent on the number of different chromosomes represented in the data set to be examined. Adding additional variants to data sets containing variants on all chromosomes results in a negligible execution time increase.

Enrichment for predicted splicing mutations after processing and filtering

Cell line	Initial variants analyzed	Novel mutations		SNPs with rare minor alleles*		Overall Mutation fraction
		Natural site	Cryptic site	Natural site	Cryptic site	
A431	290,589	16	13	13	3	0.015%
U251	314,637	7	10	18	3	0.012%
U2OS	211,049	22	9	13	4	0.022%
Total	816,275	45	32	44	10	0.016%

*dbSNP135; <1% heterozygosity

RBBP8 (a tumour suppressor gene) splicing mutation viewed in IGV with sequence walker. The mutation reduces the strength of the natural donor site from 6.2 to 3.2 bits.



Interesting variants found by the pipeline

DDX11 is inactivated in U2OS (chr12:31242087T>G; 6.89 → -11.73 bits). DDX11 is a component of the cohesin complex which has a crucial role in chromosome segregation, and is essential for survival of advanced melanoma.

In U2OS, *WWOX*, a tumor suppressor gene in osteosarcoma, contains a leaky mutation (chr16:78312497C>A; 10.24 → 6.67 bits).

Both alleles of *APIP*, an apoptosis associated gene, are inactivated in U251 (chr11:34905054G>C; 9.32 → 0.54 bits). Gene expression of *APIP* is down regulated in non-small cell lung carcinoma.

SMARCD1, encoding a chromatin modulator that interacts with nuclear receptor transcription factors, is also inactivated in A431 (chr12:50480538G>C; 8.46 → -3.21 bits), and has been shown to be mutated in carcinomas.

Implementation

- *Shannon Human Splicing Pipeline* has been released for **Linux** and **MacOSX** operating systems and requires Perl and gcc.
- Installation has been verified with Perl v.5.8.8 and 5.10.1, gcc v.4.1.2 and v.4.4.3, Ubuntu 2.6.32-27 (32 and 64 bit), CentOS 2.6.18-238 (64 bit), Fedora 16 (32 bit) kernels, and MacOSX (Mountain Lion release version 10.8, Lion release version 10.7.4; gcc v.4.2.1 and Perl 5.12.3 and 5.12.4).
- Several C libraries determine the information content of a position in the genome before and after a variant is introduced using convolution-style sliding-window computation. Changes in R_i introduced by genomic variation are computed by subtracting the initial R_i value of a position by the sum over a surrounding window, then adding the new value for each position (ΔR_i).
- Perl scripts wrap these C libraries and annotate data pipeline results. Integration with the CLC-Bio workbench environment was achieved through code written in Java utilizing the CLC-Bio developer API.
- This software is assembled as a client plugin requiring a connection to the server to execute, a server plugin, and a standalone client plugin. Two additional plugins contain a modified dbSNP135 (Indels and extraneous data removed), Ensembl Exon Data (Build 66), and GRCh37/ NCBI36 respectively.

Benefits of the Shannon Pipeline

- Several other score-based methods exist to detect individual splicing mutations (such as HSF, NNSplice, MaxEntScan, etc.). These methods are not scaled to handle the large numbers of variants generated by next generation sequencing.
- We have observed at least 30 previously unrecognized mutations per cancer cell line genome in our analyses.
- Laboratories interested genome-wide NGS or gene deep sequencing may discover previously undocumented splicing mutations using this software
- The Shannon Pipeline provides a method to detect small changes in mRNA splicing due to sequence variation. This is essential for mutation discovery in exomes, complete genomes, and high density targeted deep sequencing projects.

Dr. Peter K. Rogan
President of Cytognomix Inc.
info@cytognomix.com

Ben C. Shirley
Chief software architect
Ben.Shirley@scprobe.info

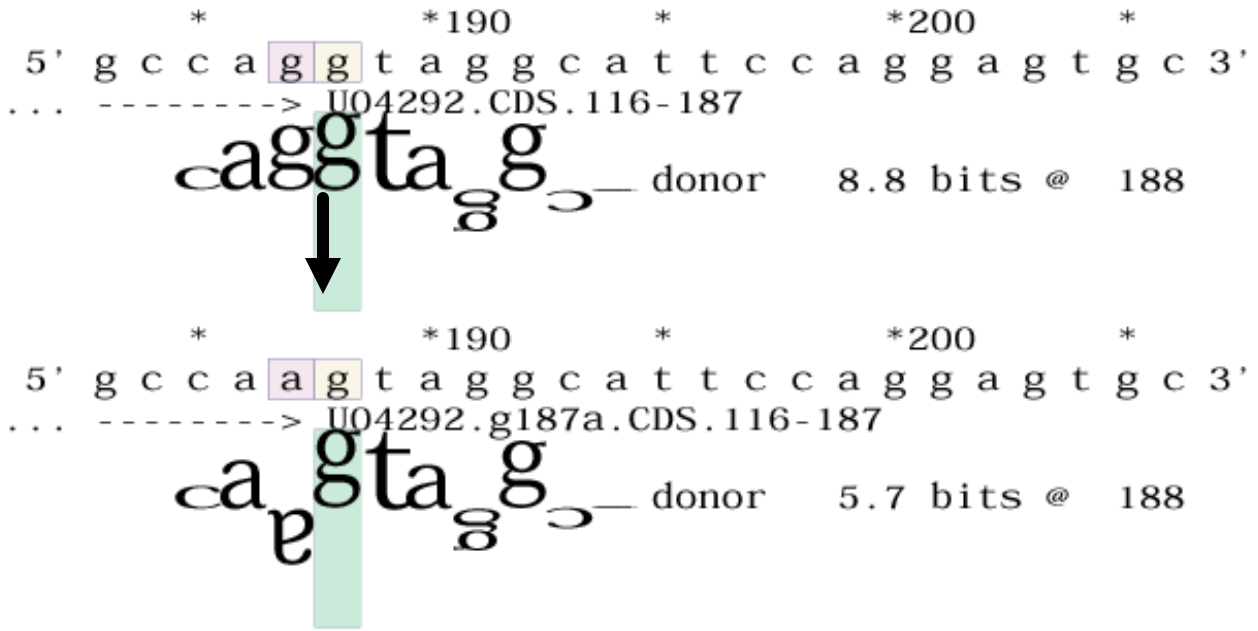
Download the full plugin or a trial version:

http://www.clcbio.com/clc-plugin/shannon-human-splicing-pipeline/

700 Collip Circle #150
London Ontario N6G 4X8 Canada
info@cytognomix.com
www.cytognomix.com

Shirley BC, Mucaki EJ, Whitehead T, Costea PI, Akan P, Rogan PK. Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. Genomics Proteomics Bioinformatics. 2013 Mar 14. PubMed PMID: 23499923.

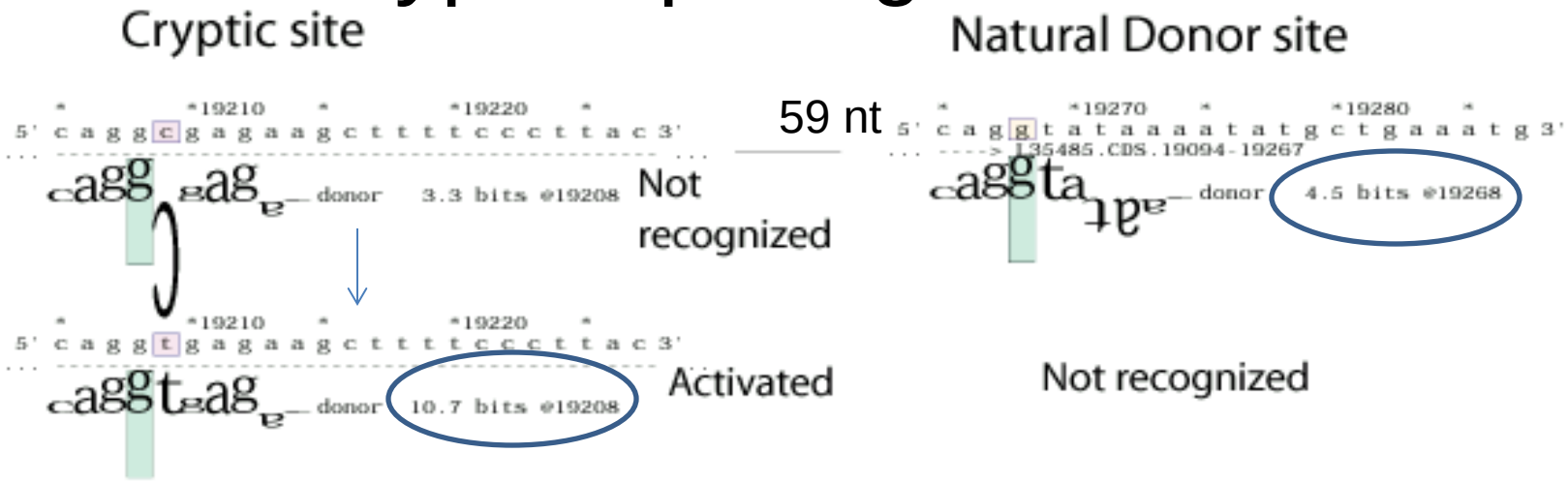
Mild (or leaky) splicing mutation



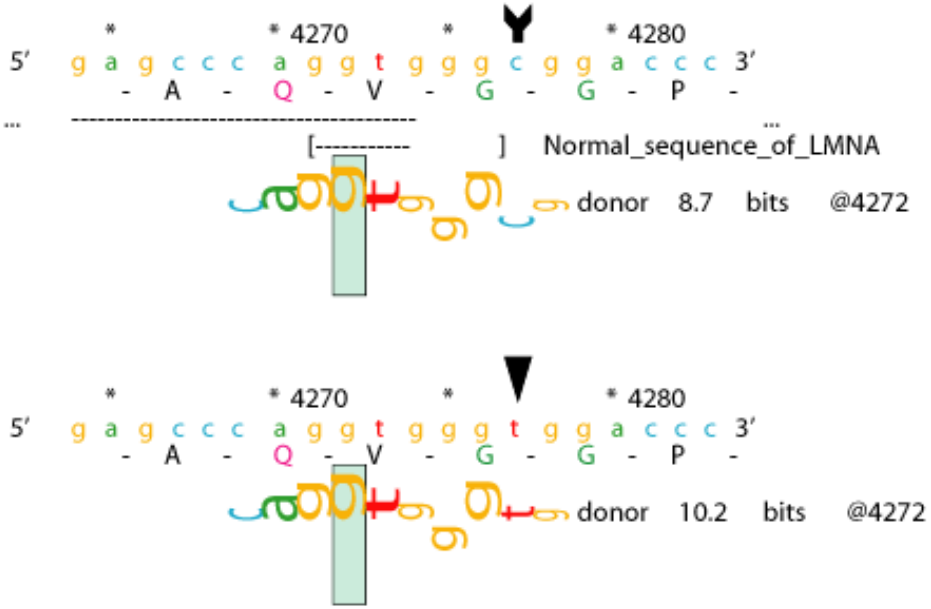
ΔR_i (bits)	Fold	%
3.1	**8.6**	**11**

A G-> A mutation 1 nucleotide upstream of the exon 8 donor site of the lysosomal lipase gene [LIPA; U04292] results in mild cholesterol ester storage disease with 4-9% enzymatic activity. The reduction in information content is significant even though the Ri value is still much greater than Ri,min.

Cryptic splicing mutations



A C->T mutation in intron 3 of the iduronidate sulfate synthetase (Mucopolysaccharidosis type II) gene strengthens and activates a cryptic donor site in exon 3 of the gene (Rogan et al. 1998).



A synonymous C>T substitution at codon 608 **strengthens** a cryptic donor splice site in exon 11 the *LMNA* gene in patients with Hutchinson-Gifford progeria (Ericksson et al. Science 2003). The walker, shown below the sequence, indicates a pre-existing 8.7 bit cryptic site that is strengthened by the mutation to 10.2 bits (>=2.8 fold).