Finding subtle mutations with the Shannon human mRNA splicing pipeline

Presentation at the CLC bio Medical Genomics Workshop
American Society of Human Genetics Annual Meeting
November 9, 2012

Peter K Rogan
Western University, London, Ontario   Canada
progan@uwo.ca
info@cytognomix.com

Motivation

The Shannon pipeline created to address the vexing problem of assessing the many variants of unknown significance (VUS) that are detected in genetic testing, exome, and complete genome sequencing. Which are deleterious and which are benign?

Besides translational effects and modifications, variants may affect promoter regulation, mRNA splicing, long range effects.

Shannon pipeline reanalysis of the Breast Cancer Information Core identified 299 novel mRNA splicing mutations (Mucaki et al. Hum. Mutation 32:735-42, 2011).

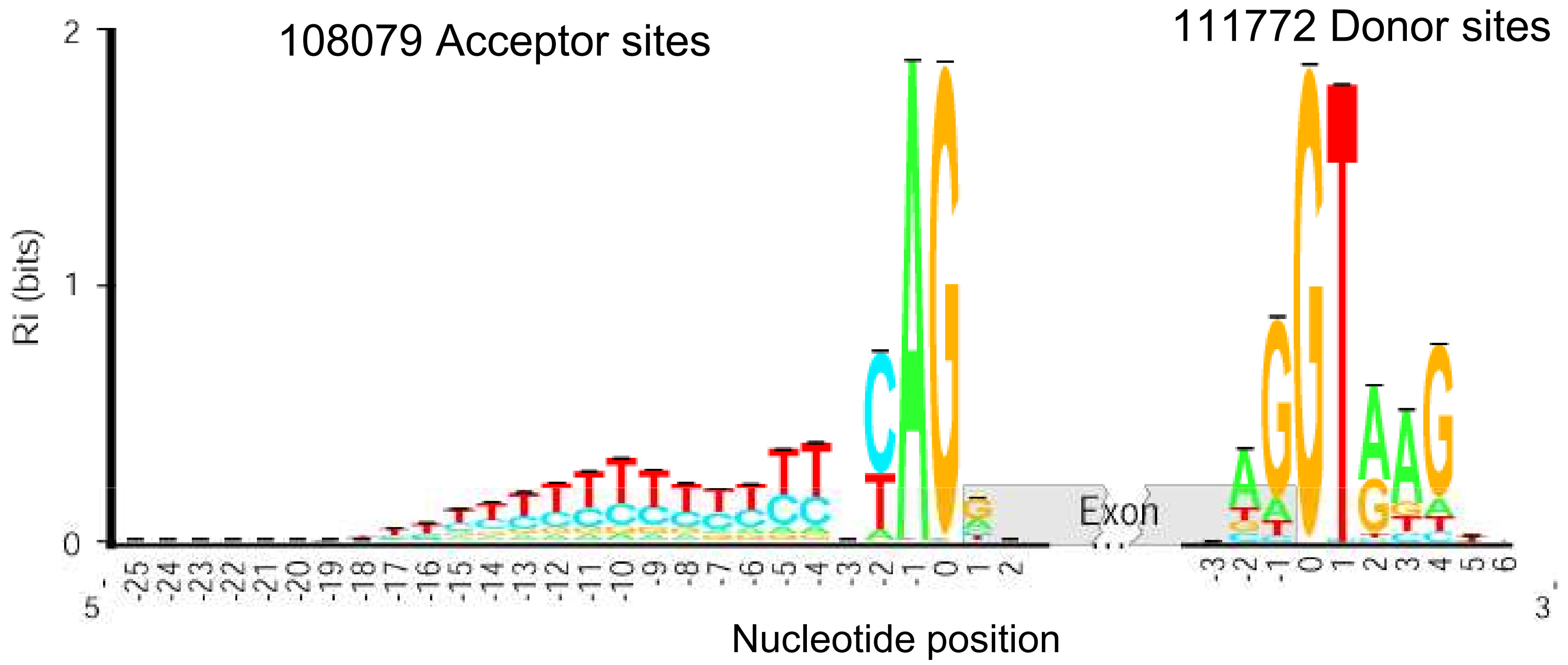
Our goal is to produce software capable of any large-scale, non-coding VUS analysis - regardless of disease, trait or phenotype.

Splicing signals *in vivo* and *in silico*

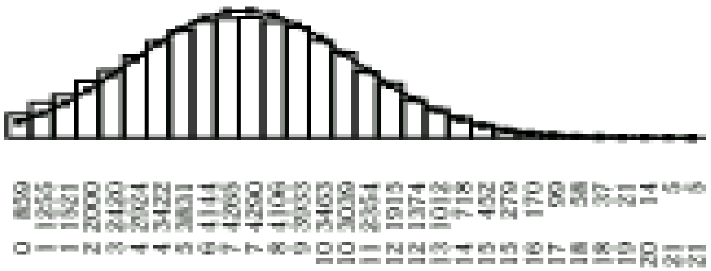
- Like most nucleic acid binding sites, sequences of splice donor, acceptor, and regulatory sites vary
- The splicing machinery recognizes and processes these combinations of splicing signals.
- **Information theory** provides a *thermodynamic* framework to recognize members of a group of structurally-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

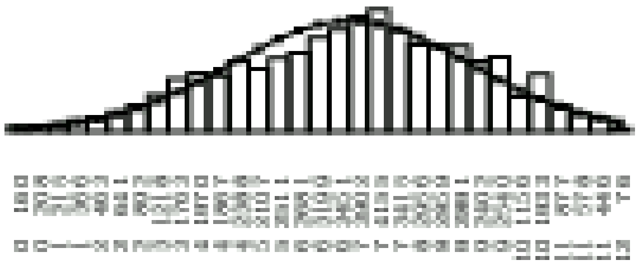
Human splice junction models

(refined from finished human reference sequence)



R_i histograms:





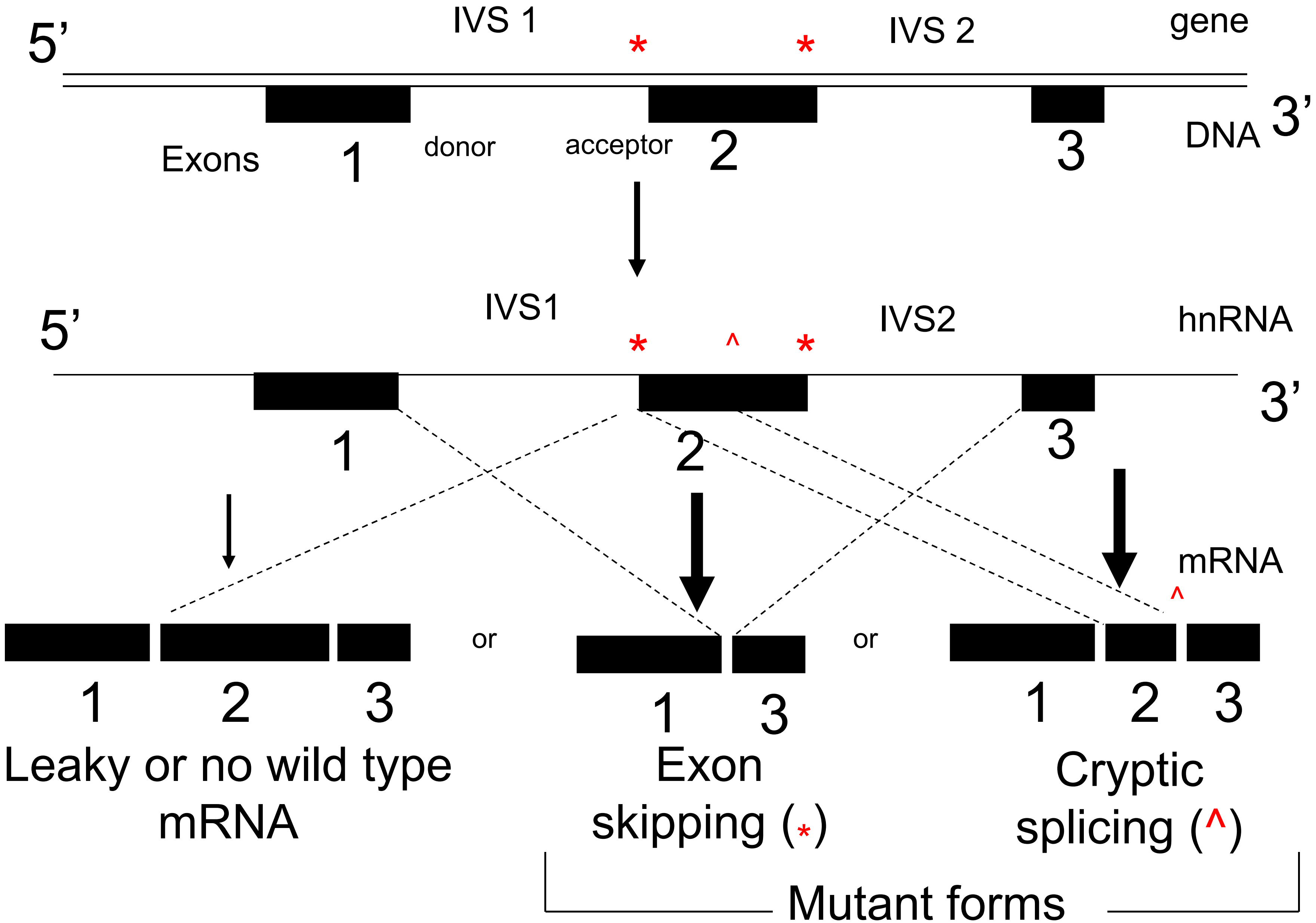
Frequency:

Bits:

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

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

Effects of mutations on mRNA splicing



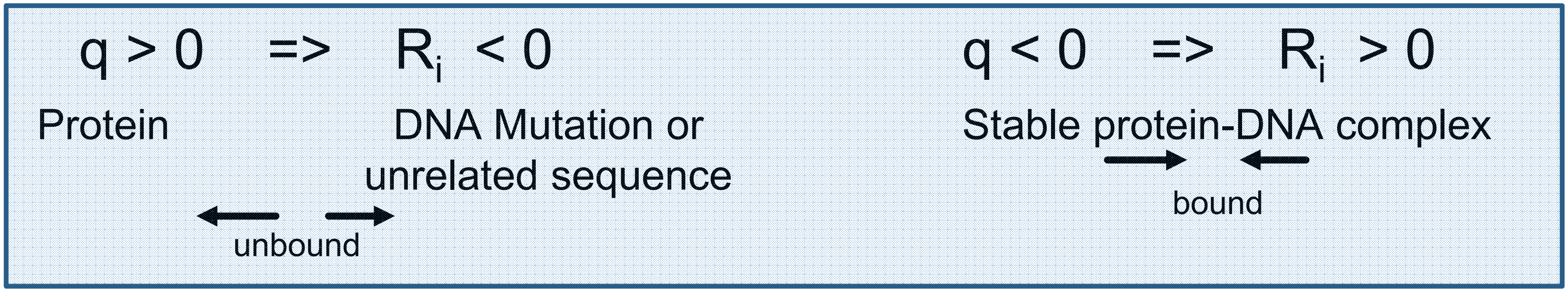
Molecular information theory

The information gained when a nucleic acid is recognized by a molecular machine (ie. the splicesome) is accompanied by a decrease in entropy that occurs upon binding: $\Delta S = -k_{\mathrm{B}} \ln(2) R.$

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

$$\text{constant} \rightarrow \; -k_B T \ln 2 \leq q / R_i \qquad \text{(joules/bit)}$$

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



$q > 0 \quad => \quad R_i < 0$

Protein

DNA Mutation or unrelated sequence

unbound

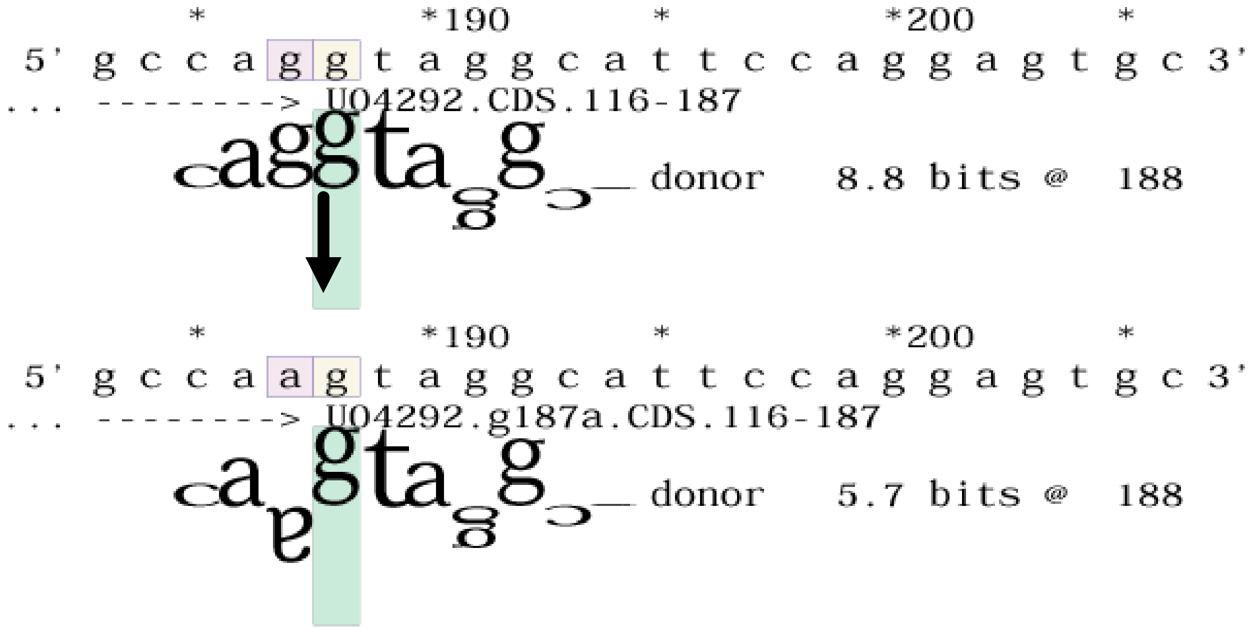
$q < 0 \quad => \quad R_i > 0$

Stable protein-DNA complex

bound

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

Mild (or leaky) splicing mutation


* *190 * *200 *
5' g c c a g g t a g g c a t t c c a g g a g t g c 3'
... ---------> U04292.CDS.116-187
donor 8.8 bits @ 188
* *190 * *200 *
5' g c c a a g t a g g c a t t c c a g g a g t g c 3'
... ---------> U04292.g187a.CDS.116-187
donor 5.7 bits @ 188

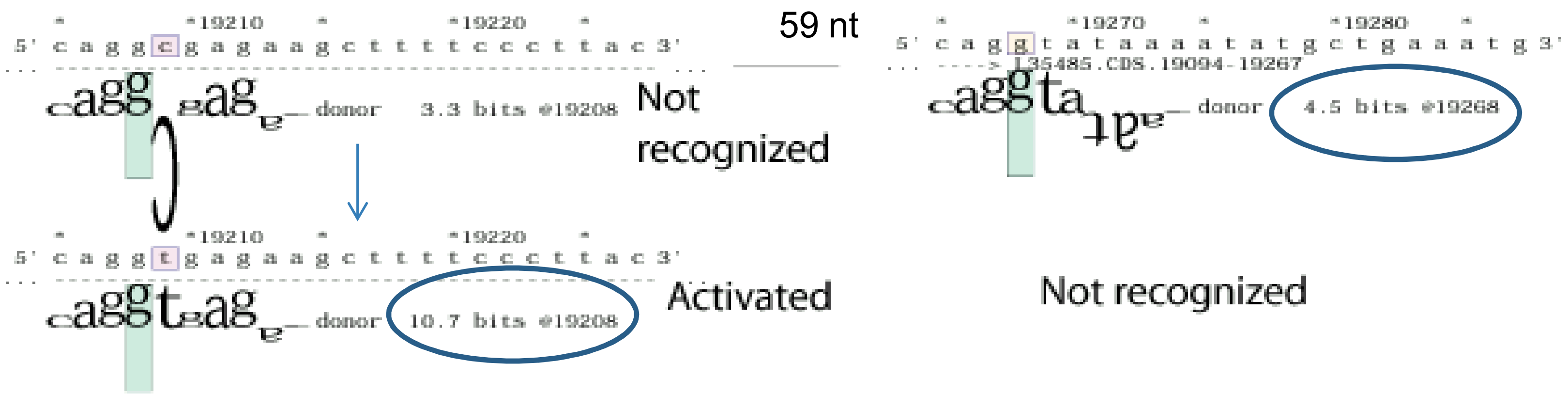
ΔR_i (bits)	Fold	%
3.1	**8.6**	**11**

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

Cryptic site

Natural Donor site


*19210 * *19220 *
5' c a g g c g a g a a g c t t t t t c c c c t t a c 3'
donor 3.3 bits @19208
*19210 * *19220 *
5' c a g g t g a g a a g c t t t t t c c c c t t a c 3'
donor 10.7 bits @19208
*19270 * *19280 *
5' c a g g g t a t a a a a t a t g c t g a a a t g 3'
----> L35485.CDS.19094-19267
donor 4.5 bits @19268

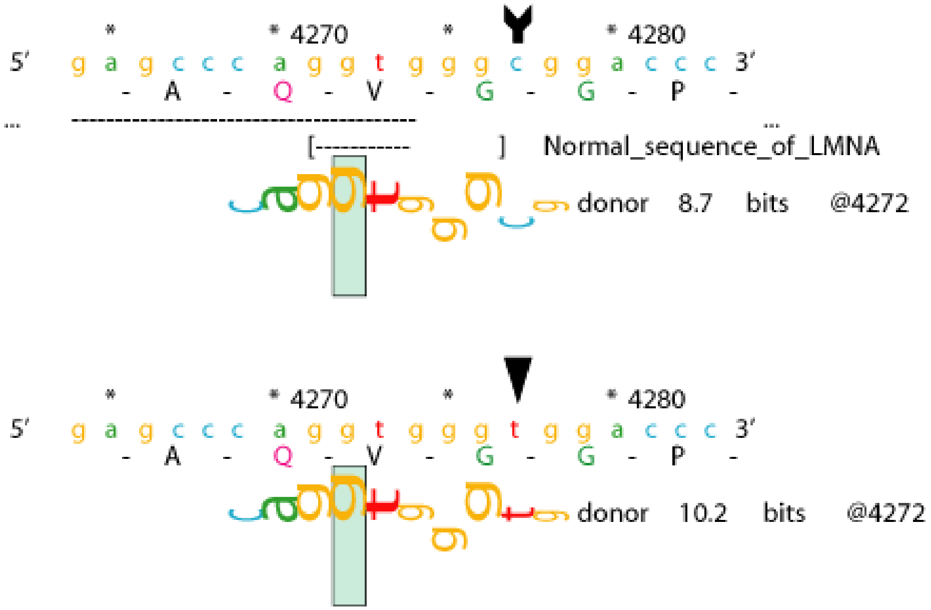
59 nt

Not recognized

Activated

Not recognized

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).


* * 4270 * * 4280
5' g a g c c c a g g t g g g c g g a c c c 3'
- A - Q - V - G - G - P -
Normal_sequence_of_LMNA
donor 8.7 bits @4272
* * 4270 * * 4280
5' g a g c c c a g g t g g g t g g a c c c 3'
- A - Q - V - G - G - P -
donor 10.2 bits @4272

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).

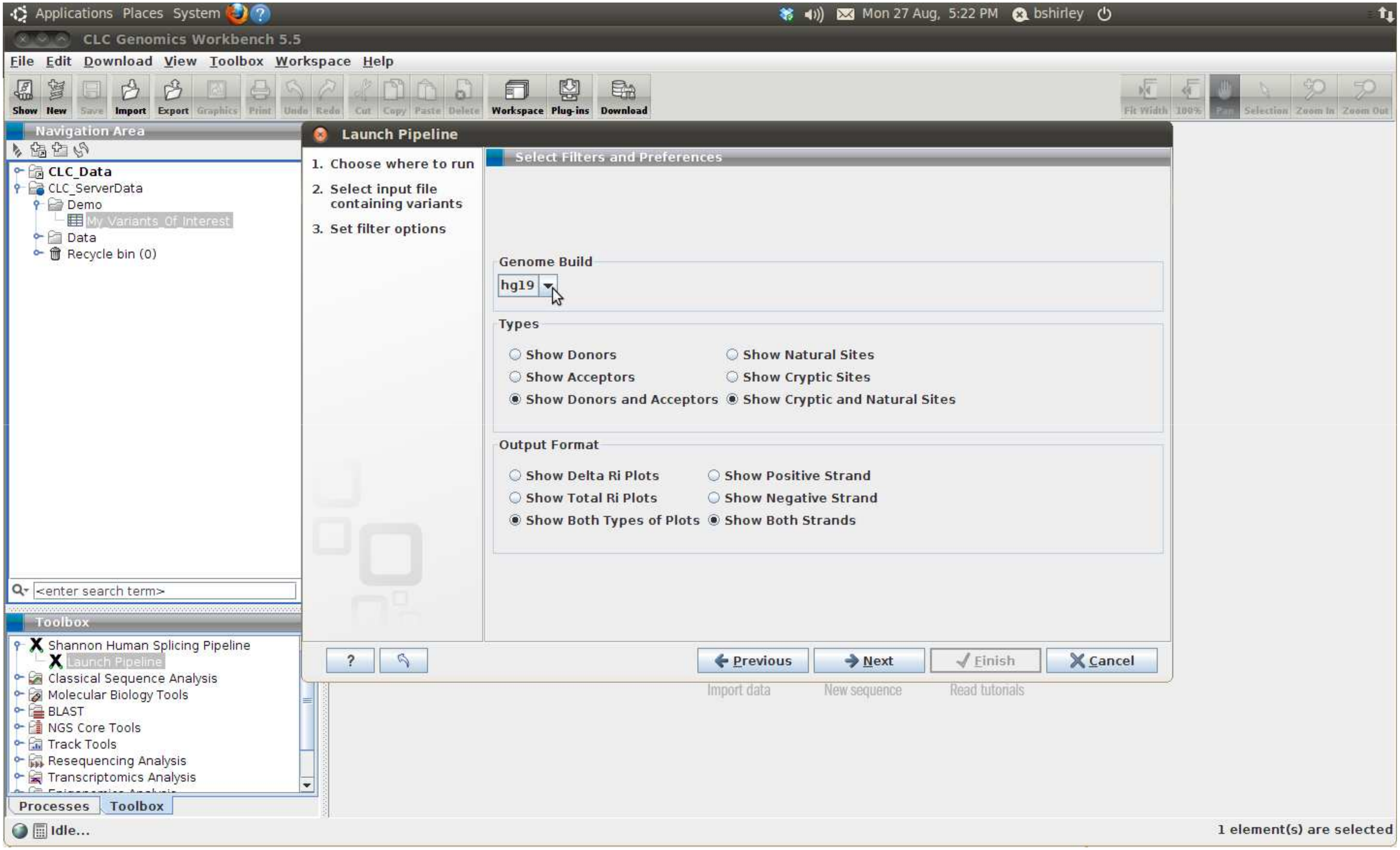
Shannon Pipeline for mutation analysis

- 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)
- 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, 571–583)
- Predecessor software for single mutation analysis has been designated as a medical device by US FDA (not approved for clinical diagnostics)

How the plug-in works

- CLC-Genomics Workbench retrieves lists of variants, and either processes the data itself or funnels it to the Shannon pipeline on the Genomics Server.
- Genome-wide information analysis is performed for all variants
- Variants with changes in information content (in bits) are annotated against standard databases (Ensembl Refseq, dbSNP)
- Prospective mutations are categorized and filtered
- Results displayed as exportable chromosome plots, sortable tables, and genome browser tracks.

Choose genome reference and results to display ...



Launch Pipeline

1. Choose where to run
2. Select input file containing variants
3. Set filter options

Select Filters and Preferences

Genome Build: hg19

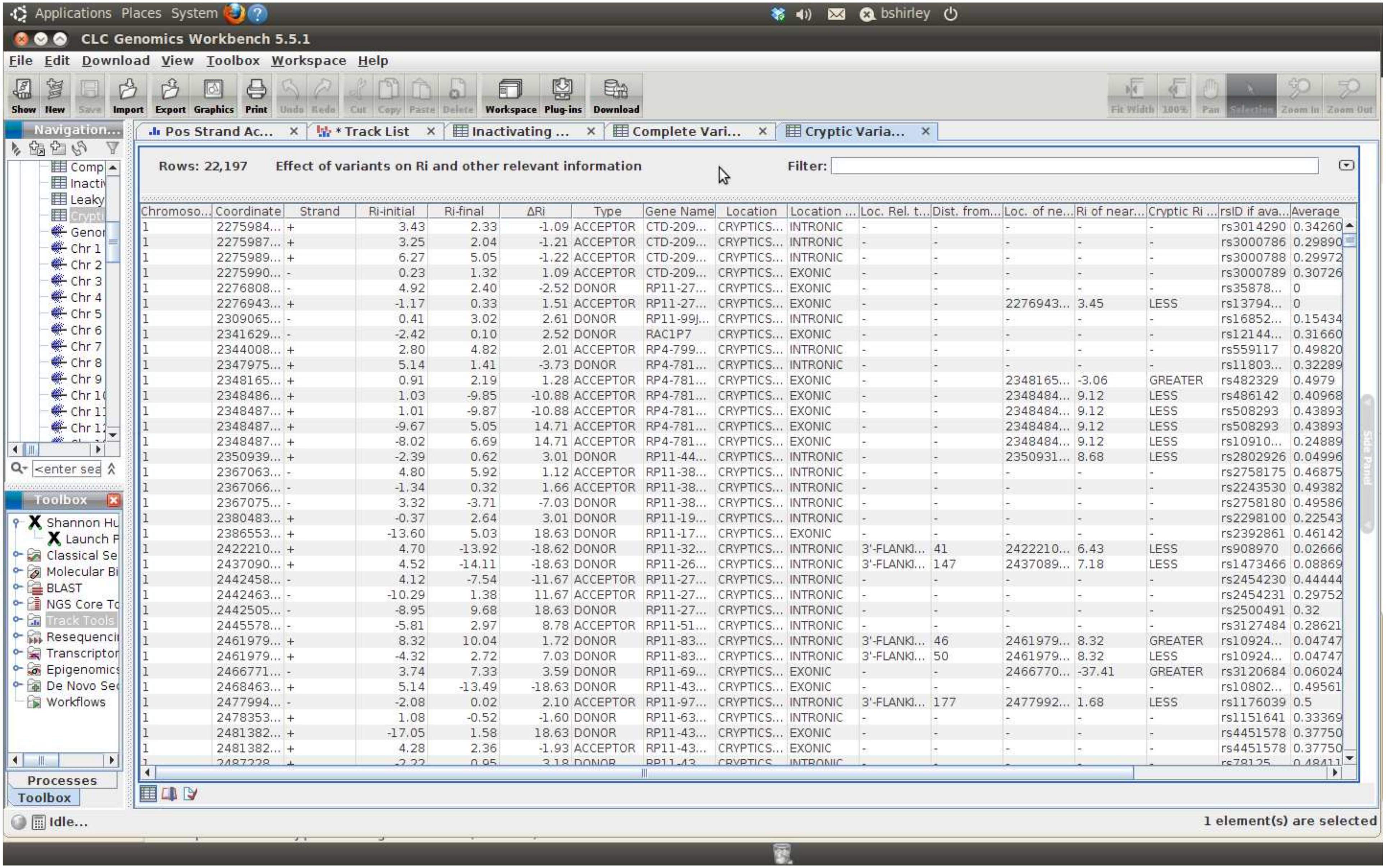
Types

○ Show Donors
○ Show Acceptors
● Show Donors and Acceptors
○ Show Natural Sites
○ Show Cryptic Sites
● Show Cryptic and Natural Sites

Output Format

○ Show Delta Ri Plots
○ Show Total Ri Plots
● Show Both Types of Plots
○ Show Positive Strand
○ Show Negative Strand
● Show Both Strands

Completed run ... Tabular output for cryptic sites below,

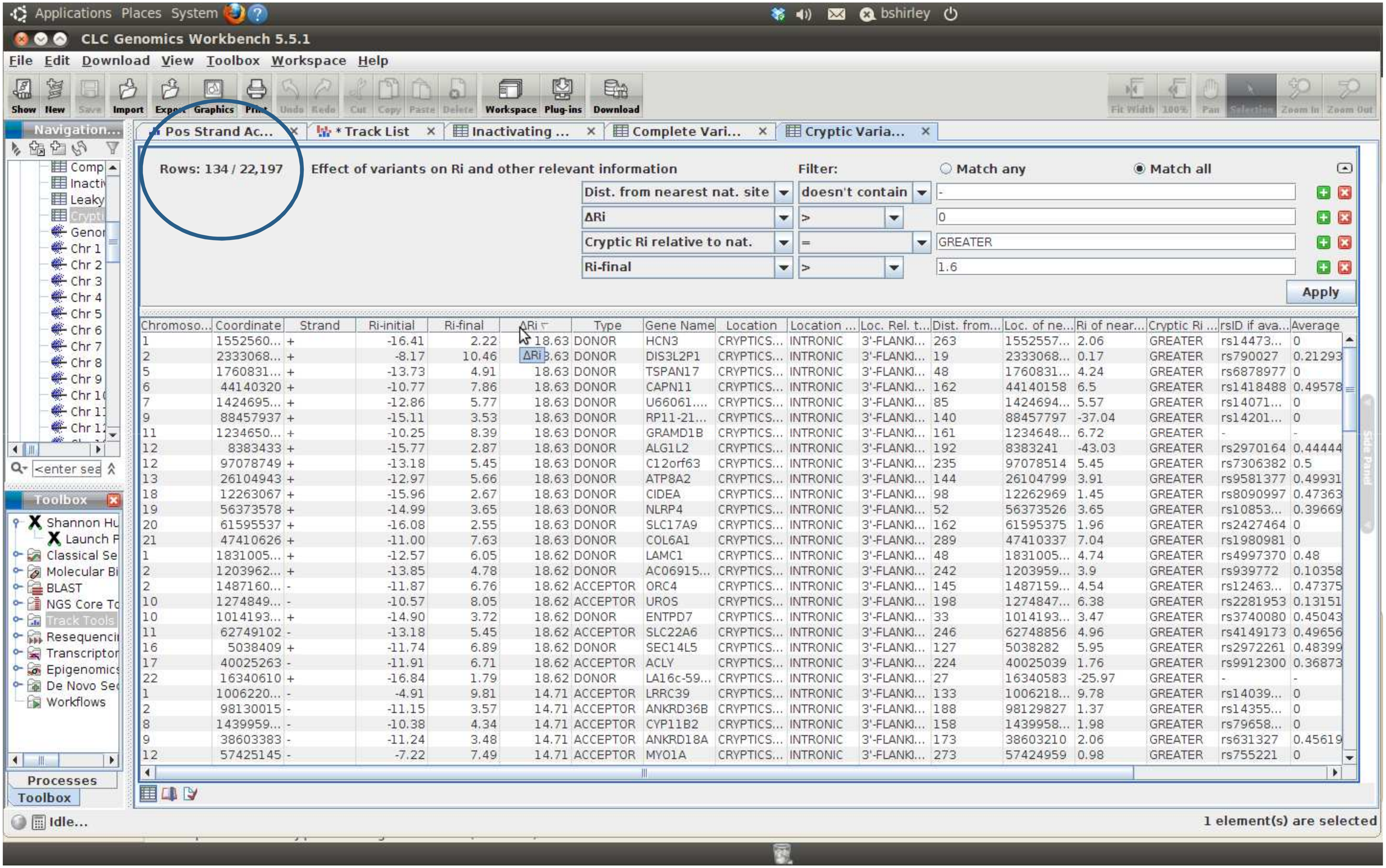


Rows: 22,197   Effect of variants on Ri and other relevant information

Chromoso...	Coordinate	Strand	Ri-initial	Ri-final	ΔRi	Type	Gene Name	Location	Location ...	Loc. Rel. t...	Dist. from...	Loc. of ne...	Ri of near...	Cryptic Ri ...	rsID if ava...	Average
1	2275984...	+	3.43	2.33	-1.09	ACCEPTOR	CTD-209...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs3014290	0.34260
1	2275987...	+	3.25	2.04	-1.21	ACCEPTOR	CTD-209...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs3000786	0.29890
1	2275989...	+	6.27	5.05	-1.22	ACCEPTOR	CTD-209...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs3000788	0.29972
1	2275990...	-	0.23	1.32	1.09	ACCEPTOR	CTD-209...	CRYPTICS...	EXONIC	-	-	-	-	-	rs3000789	0.30726
1	2276808...	-	4.92	2.40	-2.52	DONOR	RP11-27...	CRYPTICS...	EXONIC	-	-	-	-	-	rs35878...	0
1	2276943...	+	-1.17	0.33	1.51	ACCEPTOR	RP11-27...	CRYPTICS...	EXONIC	-	-	2276943...	3.45	LESS	rs13794...	0
1	2309065...	-	0.41	3.02	2.61	DONOR	RP11-99J...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs16852...	0.15434
1	2341629...	-	-2.42	0.10	2.52	DONOR	RAC1P7	CRYPTICS...	EXONIC	-	-	-	-	-	rs12144...	0.31660
1	2344008...	+	2.80	4.82	2.01	ACCEPTOR	RP4-799...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs559117	0.49820
1	2347975...	+	5.14	1.41	-3.73	DONOR	RP4-781...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs11803...	0.32289
1	2348165...	+	0.91	2.19	1.28	ACCEPTOR	RP4-781...	CRYPTICS...	EXONIC	-	-	2348165...	-3.06	GREATER	rs482329	0.4979
1	2348486...	+	1.03	-9.85	-10.88	ACCEPTOR	RP4-781...	CRYPTICS...	EXONIC	-	-	2348484...	9.12	LESS	rs486142	0.40968
1	2348487...	+	1.01	-9.87	-10.88	ACCEPTOR	RP4-781...	CRYPTICS...	EXONIC	-	-	2348484...	9.12	LESS	rs508293	0.43893
1	2348487...	+	-9.67	5.05	14.71	ACCEPTOR	RP4-781...	CRYPTICS...	EXONIC	-	-	2348484...	9.12	LESS	rs508293	0.43893
1	2348487...	+	-8.02	6.69	14.71	ACCEPTOR	RP4-781...	CRYPTICS...	EXONIC	-	-	2348484...	9.12	LESS	rs10910...	0.24889
1	2350939...	+	-2.39	0.62	3.01	DONOR	RP11-44...	CRYPTICS...	INTRONIC	-	-	2350931...	8.68	LESS	rs2802926	0.04996
1	2367063...	-	4.80	5.92	1.12	ACCEPTOR	RP11-38...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2758175	0.46875
1	2367066...	-	-1.34	0.32	1.66	ACCEPTOR	RP11-38...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2243530	0.49382
1	2367075...	-	3.32	-3.71	-7.03	DONOR	RP11-38...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2758180	0.49586
1	2380483...	+	-0.37	2.64	3.01	DONOR	RP11-19...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2298100	0.22543
1	2386553...	+	-13.60	5.03	18.63	DONOR	RP11-17...	CRYPTICS...	EXONIC	-	-	-	-	-	rs2392861	0.46142
1	2422210...	+	4.70	-13.92	-18.62	DONOR	RP11-32...	CRYPTICS...	INTRONIC	3'-FLANKI...	41	2422210...	6.43	LESS	rs908970	0.02666
1	2437090...	+	4.52	-14.11	-18.63	DONOR	RP11-26...	CRYPTICS...	INTRONIC	3'-FLANKI...	147	2437089...	7.18	LESS	rs1473466	0.08869
1	2442458...	-	4.12	-7.54	-11.67	ACCEPTOR	RP11-27...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2454230	0.44444
1	2442463...	-	-10.29	1.38	11.67	ACCEPTOR	RP11-27...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2454231	0.29752
1	2442505...	-	-8.95	9.68	18.63	DONOR	RP11-27...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs2500491	0.32
1	2445578...	-	-5.81	2.97	8.78	ACCEPTOR	RP11-51...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs3127484	0.28621
1	2461979...	+	8.32	10.04	1.72	DONOR	RP11-83...	CRYPTICS...	INTRONIC	3'-FLANKI...	46	2461979...	8.32	GREATER	rs10924...	0.04747
1	2461979...	+	-4.32	2.72	7.03	DONOR	RP11-83...	CRYPTICS...	INTRONIC	3'-FLANKI...	50	2461979...	8.32	LESS	rs10924...	0.04747
1	2466771...	-	3.74	7.33	3.59	DONOR	RP11-69...	CRYPTICS...	EXONIC	-	-	2466770...	-37.41	GREATER	rs3120684	0.06024
1	2468463...	+	5.14	-13.49	-18.63	DONOR	RP11-43...	CRYPTICS...	EXONIC	-	-	-	-	-	rs10802...	0.49561
1	2477994...	-	-2.08	0.02	2.10	ACCEPTOR	RP11-97...	CRYPTICS...	INTRONIC	3'-FLANKI...	177	2477992...	1.68	LESS	rs1176039	0.5
1	2478353...	+	1.08	-0.52	-1.60	DONOR	RP11-63...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs1151641	0.33369
1	2481382...	+	-17.05	1.58	18.63	DONOR	RP11-43...	CRYPTICS...	EXONIC	-	-	-	-	-	rs4451578	0.37750
1	2481382...	+	4.28	2.36	-1.93	ACCEPTOR	RP11-43...	CRYPTICS...	EXONIC	-	-	-	-	-	rs4451578	0.37750
1	2487228...	+	-2.22	0.95	3.18	DONOR	RP11-43...	CRYPTICS...	INTRONIC	-	-	-	-	-	rs78125...	0.48411

1 element(s) are selected

Filtering cryptic splice site variants hones in on likely mutations



Applications Places System
bshirley
CLC Genomics Workbench 5.5.1
File Edit Download View Toolbox Workspace Help
Show New Save Import Export Graphics Print Undo Redo Cut Copy Paste Delete Workspace Plug-ins Download
Fit Width 100% Pan Selection Zoom In Zoom Out
Navigation...
Pos Strand Ac... × * Track List × Inactivating ... × Complete Vari... × Cryptic Varia... ×
Comp
Inacti
Leaky
Crypti
Genor
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 1(
Chr 1]
Chr 1;
<enter sea
Toolbox
Shannon Hu
Launch F
Classical Se
Molecular Bi
BLAST
NGS Core To
Track Tools
Resequencii
Transcriptor
Epigenomics
De Novo Sec
Workflows
Processes
Toolbox
Side Panel
Idle...
1 element(s) are selected

Rows: 134 / 22,197

Effect of variants on Ri and other relevant information

Filter: ○ Match any ● Match all

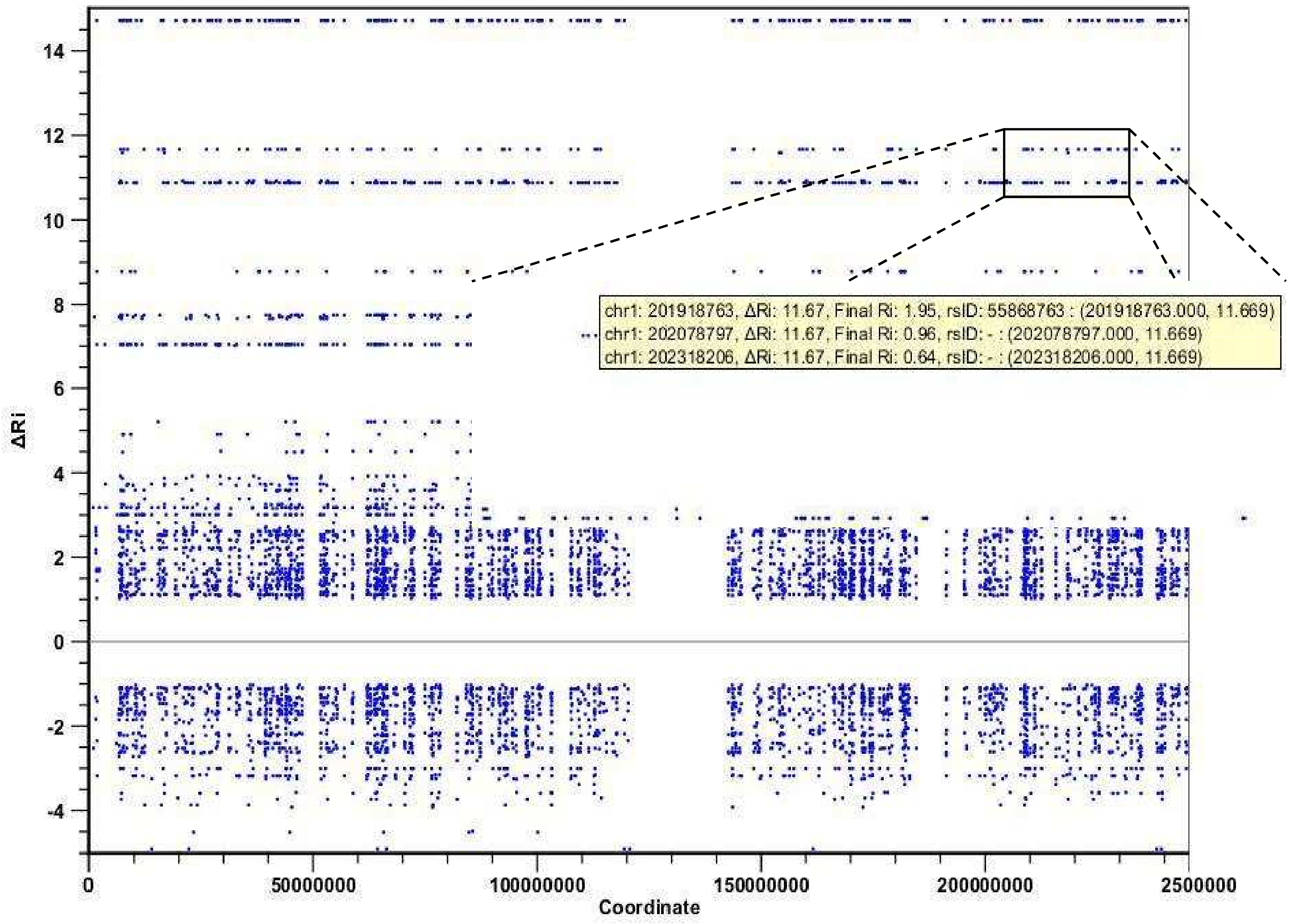
Dist. from nearest nat. site	doesn't contain	-
ΔRi	>	0
Cryptic Ri relative to nat.	=	GREATER
Ri-final	>	1.6

Apply

Chromoso...	Coordinate	Strand	Ri-initial	Ri-final	ΔRi	Type	Gene Name	Location	Location ...	Loc. Rel. t...	Dist. from...	Loc. of ne...	Ri of near...	Cryptic Ri ...	rsID if ava...	Average
1	1552560...	+	-16.41	2.22	18.63	DONOR	HCN3	CRYPTICS...	INTRONIC	3'-FLANKI...	263	1552557...	2.06	GREATER	rs14473...	0
2	2333068...	+	-8.17	10.46	ΔRi 3.63	DONOR	DIS3L2P1	CRYPTICS...	INTRONIC	3'-FLANKI...	19	2333068...	0.17	GREATER	rs790027	0.21293
5	1760831...	+	-13.73	4.91	18.63	DONOR	TSPAN17	CRYPTICS...	INTRONIC	3'-FLANKI...	48	1760831...	4.24	GREATER	rs6878977	0
6	44140320	+	-10.77	7.86	18.63	DONOR	CAPN11	CRYPTICS...	INTRONIC	3'-FLANKI...	162	44140158	6.5	GREATER	rs1418488	0.49578
7	1424695...	+	-12.86	5.77	18.63	DONOR	U66061....	CRYPTICS...	INTRONIC	3'-FLANKI...	85	1424694...	5.57	GREATER	rs14071...	0
9	88457937	+	-15.11	3.53	18.63	DONOR	RP11-21...	CRYPTICS...	INTRONIC	3'-FLANKI...	140	88457797	-37.04	GREATER	rs14201...	0
11	1234650...	+	-10.25	8.39	18.63	DONOR	GRAMD1B	CRYPTICS...	INTRONIC	3'-FLANKI...	161	1234648...	6.72	GREATER	-	-
12	8383433	+	-15.77	2.87	18.63	DONOR	ALG1L2	CRYPTICS...	INTRONIC	3'-FLANKI...	192	8383241	-43.03	GREATER	rs2970164	0.44444
12	97078749	+	-13.18	5.45	18.63	DONOR	C12orf63	CRYPTICS...	INTRONIC	3'-FLANKI...	235	97078514	5.45	GREATER	rs7306382	0.5
13	26104943	+	-12.97	5.66	18.63	DONOR	ATP8A2	CRYPTICS...	INTRONIC	3'-FLANKI...	144	26104799	3.91	GREATER	rs9581377	0.49931
18	12263067	+	-15.96	2.67	18.63	DONOR	CIDEA	CRYPTICS...	INTRONIC	3'-FLANKI...	98	12262969	1.45	GREATER	rs8090997	0.47363
19	56373578	+	-14.99	3.65	18.63	DONOR	NLRP4	CRYPTICS...	INTRONIC	3'-FLANKI...	52	56373526	3.65	GREATER	rs10853...	0.39669
20	61595537	+	-16.08	2.55	18.63	DONOR	SLC17A9	CRYPTICS...	INTRONIC	3'-FLANKI...	162	61595375	1.96	GREATER	rs2427464	0
21	47410626	+	-11.00	7.63	18.63	DONOR	COL6A1	CRYPTICS...	INTRONIC	3'-FLANKI...	289	47410337	7.04	GREATER	rs1980981	0
1	1831005...	+	-12.57	6.05	18.62	DONOR	LAMC1	CRYPTICS...	INTRONIC	3'-FLANKI...	48	1831005...	4.74	GREATER	rs4997370	0.48
2	1203962...	+	-13.85	4.78	18.62	DONOR	AC06915...	CRYPTICS...	INTRONIC	3'-FLANKI...	242	1203959...	3.9	GREATER	rs939772	0.10358
2	1487160...	-	-11.87	6.76	18.62	ACCEPTOR	ORC4	CRYPTICS...	INTRONIC	3'-FLANKI...	145	1487159...	4.54	GREATER	rs12463...	0.47375
10	1274849...	-	-10.57	8.05	18.62	ACCEPTOR	UROS	CRYPTICS...	INTRONIC	3'-FLANKI...	198	1274847...	6.38	GREATER	rs2281953	0.13151
10	1014193...	+	-14.90	3.72	18.62	DONOR	ENTPD7	CRYPTICS...	INTRONIC	3'-FLANKI...	33	1014193...	3.47	GREATER	rs3740080	0.45043
11	62749102	-	-13.18	5.45	18.62	ACCEPTOR	SLC22A6	CRYPTICS...	INTRONIC	3'-FLANKI...	246	62748856	4.96	GREATER	rs4149173	0.49656
16	5038409	+	-11.74	6.89	18.62	DONOR	SEC14L5	CRYPTICS...	INTRONIC	3'-FLANKI...	127	5038282	5.95	GREATER	rs2972261	0.48399
17	40025263	-	-11.91	6.71	18.62	ACCEPTOR	ACLY	CRYPTICS...	INTRONIC	3'-FLANKI...	224	40025039	1.76	GREATER	rs9912300	0.36873
22	16340610	+	-16.84	1.79	18.62	DONOR	LA16c-59...	CRYPTICS...	INTRONIC	3'-FLANKI...	27	16340583	-25.97	GREATER	-	-
1	1006220...	-	-4.91	9.81	14.71	ACCEPTOR	LRRC39	CRYPTICS...	INTRONIC	3'-FLANKI...	133	1006218...	9.78	GREATER	rs14039...	0
2	98130015	-	-11.15	3.57	14.71	ACCEPTOR	ANKRD36B	CRYPTICS...	INTRONIC	3'-FLANKI...	188	98129827	1.37	GREATER	rs14355...	0
8	1439959...	-	-10.38	4.34	14.71	ACCEPTOR	CYP11B2	CRYPTICS...	INTRONIC	3'-FLANKI...	158	1439958...	1.98	GREATER	rs79658...	0
9	38603383	-	-11.24	3.48	14.71	ACCEPTOR	ANKRD18A	CRYPTICS...	INTRONIC	3'-FLANKI...	173	38603210	2.06	GREATER	rs631327	0.45619
12	57425145	-	-7.22	7.49	14.71	ACCEPTOR	MYO1A	CRYPTICS...	INTRONIC	3'-FLANKI...	273	57424959	0.98	GREATER	rs755221	0

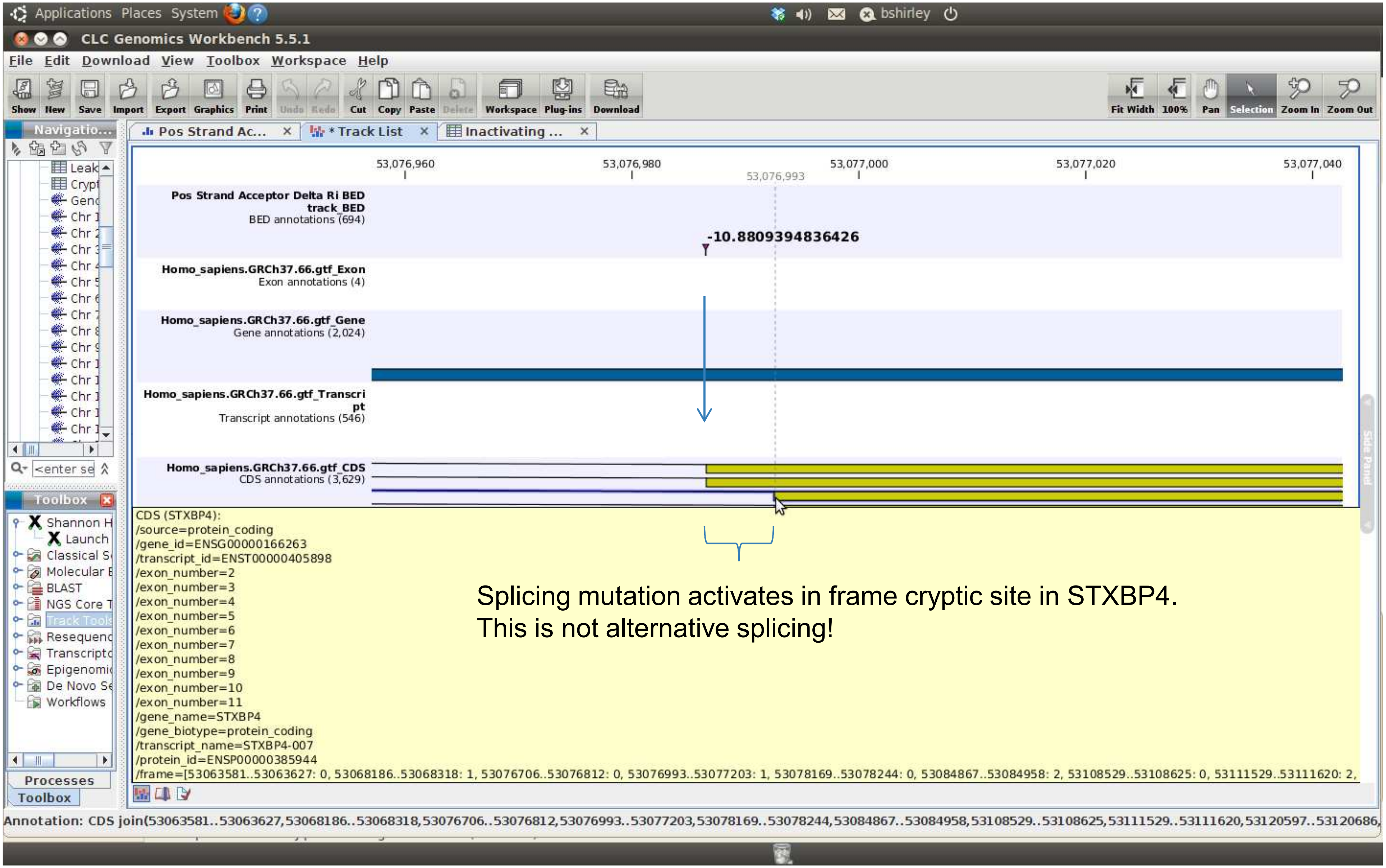
“Manhattan-like” plots to survey all potential splicing related information changes by chromosome

Chromosome 1



```
chr1: 201918763, ΔRi: 11.67, Final Ri: 1.95, rsID: 55868763 : (201918763.000, 11.669)
chr1: 202078797, ΔRi: 11.67, Final Ri: 0.96, rsID: - : (202078797.000, 11.669)
chr1: 202318206, ΔRi: 11.67, Final Ri: 0.64, rsID: - : (202318206.000, 11.669)
```

Displaying custom track of ΔR_i values with Genome Workbench Browser



```
CDS (STXBP4):
/source=protein_coding
/gene_id=ENSG00000166263
/transcript_id=ENST00000405898
/exon_number=2
/exon_number=3
/exon_number=4
/exon_number=5
/exon_number=6
/exon_number=7
/exon_number=8
/exon_number=9
/exon_number=10
/exon_number=11
/gene_name=STXBP4
/gene_biotype=protein_coding
/transcript_name=STXBP4-007
/protein_id=ENSP00000385944
/frame=[53063581..53063627: 0, 53068186..53068318: 1, 53076706..53076812: 0, 53076993..53077203: 1, 53078169..53078244: 0, 53084867..53084958: 2, 53108529..53108625: 0, 53111529..53111620: 2,
```

```
Annotation: CDS join(53063581..53063627,53068186..53068318,53076706..53076812,53076993..53077203,53078169..53078244,53084867..53084958,53108529..53108625,53111529..53111620,53120597..53120686,
```

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

Implementation

1. *Shannon Human Splicing Pipeline* has been released for **Linux** and **MacOSX** operating systems supporting Perl and gcc.
2. Installation has been verified with Perl v.5.8.8 and 5.10.1 and gcc v.4.1.2 and v.4.4.3 with the Ubuntu 2.6.32-27 (32 and 64 bit), CentOS 2.6.18-238 (64 bit), and 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).
3. 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).
4. 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.
5. 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.

Performance

Number of variants	Running time
100,000	37m
211,049	1h 12m
290,589	1h 17m
314,637	1h 20m

I7 Processor,16 Gb RAM, Ubuntu Linux

Contributors/ Support

Western University
Ben Shirley
John Mucaki
Tyson Whitehead
Stephanie Dorman
Coby Viner
Eddie Dovigi
Yanxin Li
Asanka Wickramasinghe
Joan Knoll

Science for Life Laboratory,
Karolinska Institute
Paul Costea
Pelin Akan

National Cancer Institute
Thomas Schneider

Funding
Cytognomix Inc.
Canadian Breast Cancer Foundation
National Science and Engineering Research Council
Compute Canada (SHARCNET)
Canada Research Chairs
Canada Foundation For Innovation