CYTOGNOMIX

# Shannon pipeline plug-in:

## For human mRNA splicing mutations

CLC bio Genomics Workbench plug-in
CLC bio Genomics Server plug-in

*Features and Benefits*

*Cytognomix* introduces a line of Shannon pipeline plug-ins for prediction of functionally-significant, non-coding variants in genome or exome sequences. Comprehensive genome-scale analysis is now possible for mutations which completely or partially inactivate mRNA splice sites or activate cryptic splicing. The Shannon pipeline plug-in uses our patented and proven information theory-based binding site analysis. Its algorithm has been validated in hundreds of peer-reviewed research studies of splicing mutations, and has been recognized by the American College of Medical Genetics and Genomics in their published guidelines and standards (*Genet. Med.* 7, 571–583).

The CLC Bio Genomics Workbench is used to visualize, export, and further analyze the results of the Shannon pipeline suite.

## Cytognomix Tools

The 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, then annotated against standard databases, and resulting mutations are filtered. Results are displayed as exportable Manhattan-style plots, sortable tables, or browser tracks.

## User Benefits

- The Shannon Human Splicing Pipeline starts with hundreds of thousands of variants, then hones in on the very limited number that potentially alter mRNA splicing.
- Variants are categorized by whether they fully or partially inactivate natural sites or activate cryptic sites proximate to exons or within them.
- Predicts variants missed by many other techniques
- Splicing-related changes are displayed graphically either as “Manhattan-like” plots or as BED tracks.
- Mutations can be sorted according to:
  - the change in information content,
  - the proximity to a natural splice site,
  - the relative strength of cryptic *vs* adjacent natural site,
  - gene and chromosome coordinate.
- Variants that affect known single nucleotide polymorphisms (SNPs) are identified and can be filtered according to allele frequency.
- Results for genome-wide high throughput sequence data obtained in ≤2 hours.

- Fully integrated with CLC-Bio Genomics Workbench and Genomics Server

## Features

- Identifies likely mutations with industry-leading sensitivity and specificity. Trusted legacy of experimentally validated mutation predictions.
- Input accepts variants in VCF, in *Cytognomix*'s simple indexed format, or as CLC Bio variant objects.
- Can accept variants and output results with either hg18 (NCBI36) or hg19 (GRCh37) coordinates.
- Intuitive exportable results based on sensible defaults.
- Run as a standalone application on the Genomics Workbench or configured as a client-server with both the Workbench and Server.
- Fully compatible output for other CLC Bio Workbench Tools.

## Innovative platform for rapid, non-coding mutation analysis

*Cytognomix* uses information theory-based models of mRNA splicing to analyze mutations that alter transcript structure and abundance[1-3]. Information models rank sequences according to their individual information content ($R_i$ in bits). Functional binding sites have $R_i > 0$, corresponding to $\Delta G < 0$ kcal/mol. Strong binding sites have $R_i > R_{sequence}$ while weak sites have $R_i < R_{sequence}$. Variations which alter the affinity of a protein to bind there modify the $R_i$ of the site. A 1 bit change in information content ($\Delta R_i$) corresponds to a ≥ 2 fold change in binding affinity. This approach is applicable to any type of nucleic acid binding site, including transcription factors and other conserved non-coding sequences.

Predictions from these models are accurate[2-4], as differences in individual information contents ($\Delta R_{i,}$ in bits) are related to the splicesomal affinities of natural and variant sequences[1,5]. Pathogenicity is related to $\Delta R_i$, which is decreased at natural splice sites and/or increased at cryptic sites [6-7]. Sites with negative $R_i$ values are not recognized. Leaky mutations have modestly reduced $R_i$ values. Cryptic splice sites with $R_i$ values exceeding adjacent natural splice sites are activated[2]. The plug-in reports minimally detectable expression changes of >2 fold [8], or $\Delta R_i > 1$ bit, as significant.

The Shannon pipeline was initially created to address the vexing problem of assessing the many variants of unknown significance that are detected in cancer genetic testing. The pipeline has been used to reanalyze the Breast Cancer Information Core identifying many splicing mutations, most of which were previously unrecognized[9].

The Shannon software pipeline was developed and implemented in C and Perl to perform information analysis fast on a genome-wide scale. Determining $R_i$ values of donor and acceptor sites along a nucleotide sequence is carried out using a convolution-style, sliding-window computation on chromosomes or subsets of chromosomes. $R_i$ values are computed with $R_i(b,l)$ information weight matrix (based on a genome-wide set of verified donor and acceptor splice sites). Variants with significant $\Delta R_i$ values are then filtered, and annotated based on the gene they reside within. For cryptic splice sites, the distance and relative location of the adjacent natural splice site of the same polarity is reported. Inactivating and leaky natural splice sites and cryptic splicing variants are categorized. Variants with known SNP designations and respective allele frequencies are also reported.

The CLC Bio implementation allows for additional filtering of input, and produces a graphical display of both $\Delta R_i$ and final $R_i$ values for each variant on each chromosome, a table output which can be dynamically sorted that is categorized as separate tab for each type of mutation consequence, and BEDGRAPH output suitable for genome browsing. Results may be exported to a spreadsheet format for further data exploration.

References:
1.Schneider TD. J. Theor. Biol. 189: 427-41, 1997; 2. Rogan PK et al. Hum Mutat 12:153-171, 1998; 3. Rogan PK et al. Pharmacogenetics. 13:207-18, 2003; 4. Rogan PK, Schneider TD.. Hum Mutat 6:74-76, 1995; 5. Gadiraju S, et al. BMC Bioinformatics 4:38, 2003; 6. von Kodolitsch et al Circulation 100:693-9, 1999; 7. von Kodolitsch et al. 12: 258-262, 2006; 8. Nalla VK, Rogan PK. Hum Mut. 25(4):334-42, 2012.; 9. Mucaki et al. Hum. Mut. 32:735-742, 2012.

## Benchmarks

The *Shannon Human Splicing Pipeline* analyzes an average of 3198 variants/min on an I7-based server:

**Performance of Shannon pipeline plug-in on complete genome sequence data***

| Number of variants | Complete analysis time |
|---|---|
| 100,000 | 37 m |
| 211,049 | 1h 12 m |
| 290,589 | 1h 22 m |
| 314,637 | 1h 27 m |

***hg18/NCBI36**

## Requirements and validation

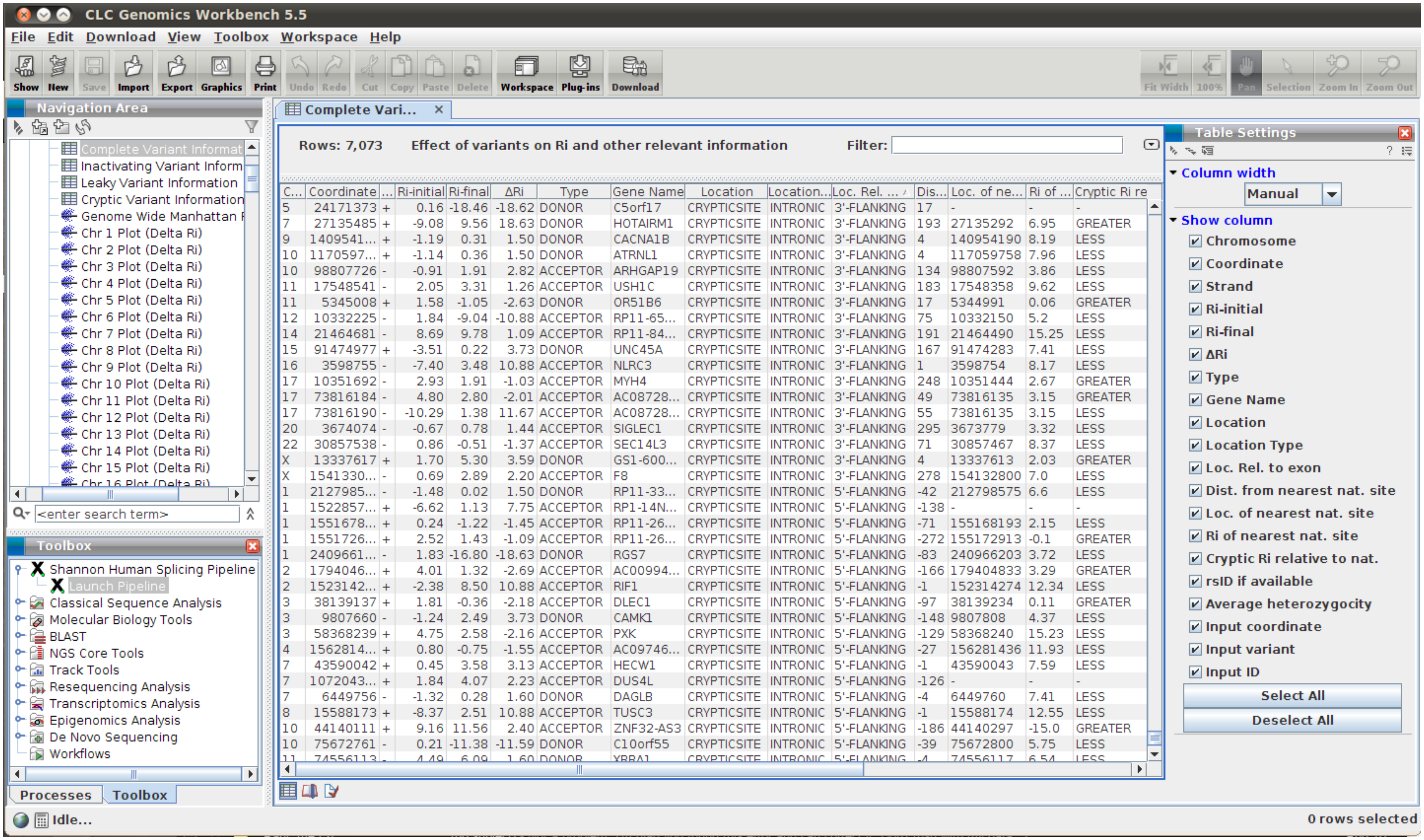
The Cytognomix Shannon human mRNA splicing plug-in runs in standalone mode on the CLC Genomics Workbench V5.5 or with both the Workbench and CLC Genomics Server V.4.5 (as a standalone server or running Gridworks). Released for **Linux** and **MacOSX** Operating systems supporting Perl and gcc. Installation 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 (Lion release version 10.7.4; gcc v.4.2.1 and Perl 5.12.3) on hardware equipped with an Intel I7 processor and at least 4Gb RAM.

## Support

CLC Bio customer support (primary)
*Cytognomix* Inc. (secondary)
700 Collip Circle #150, London ON N6G 4X8 Canada
info@cytognomix.com ♦ www.cytognomix.com

Notice: Exclusive marketing rights for products based on **US Pat. No. #5,867,402** have been granted to *CytognomiX*.

Tabular results from genome-wide mutation analysis produced by *CytognomiX*'s Shannon mutation pipeline:



CLC Genomics Workbench 5.5
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 Area
Complete Vari...
Table Settings
Toolbox
Processes Toolbox

Navigation Area:

- Complete Variant Informat
- Inactivating Variant Inform
- Leaky Variant Information
- Cryptic Variant Information
- Genome Wide Manhattan F
- Chr 1 Plot (Delta Ri)
- Chr 2 Plot (Delta Ri)
- Chr 3 Plot (Delta Ri)
- Chr 4 Plot (Delta Ri)
- Chr 5 Plot (Delta Ri)
- Chr 6 Plot (Delta Ri)
- Chr 7 Plot (Delta Ri)
- Chr 8 Plot (Delta Ri)
- Chr 9 Plot (Delta Ri)
- Chr 10 Plot (Delta Ri)
- Chr 11 Plot (Delta Ri)
- Chr 12 Plot (Delta Ri)
- Chr 13 Plot (Delta Ri)
- Chr 14 Plot (Delta Ri)
- Chr 15 Plot (Delta Ri)
- Chr 16 Plot (Delta Ri)

<enter search term>

Toolbox:

- Shannon Human Splicing Pipeline
  - Launch Pipeline
- Classical Sequence Analysis
- Molecular Biology Tools
- BLAST
- NGS Core Tools
- Track Tools
- Resequencing Analysis
- Transcriptomics Analysis
- Epigenomics Analysis
- De Novo Sequencing
- Workflows

Rows: 7,073 Effect of variants on Ri and other relevant information Filter:

| C... | Coordinate | ... | Ri-initial | Ri-final | ΔRi | Type | Gene Name | Location | Location... | Loc. Rel. ... | Dis... | Loc. of ne... | Ri of ... | Cryptic Ri re |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 24171373 | + | 0.16 | -18.46 | -18.62 | DONOR | C5orf17 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 17 | - | - | - |
| 7 | 27135485 | + | -9.08 | 9.56 | 18.63 | DONOR | HOTAIRM1 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 193 | 27135292 | 6.95 | GREATER |
| 9 | 1409541... | + | -1.19 | 0.31 | 1.50 | DONOR | CACNA1B | CRYPTICSITE | INTRONIC | 3'-FLANKING | 4 | 140954190 | 8.19 | LESS |
| 10 | 1170597... | + | -1.14 | 0.36 | 1.50 | DONOR | ATRNL1 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 4 | 117059758 | 7.96 | LESS |
| 10 | 98807726 | - | -0.91 | 1.91 | 2.82 | ACCEPTOR | ARHGAP19 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 134 | 98807592 | 3.86 | LESS |
| 11 | 17548541 | - | 2.05 | 3.31 | 1.26 | ACCEPTOR | USH1C | CRYPTICSITE | INTRONIC | 3'-FLANKING | 183 | 17548358 | 9.62 | LESS |
| 11 | 5345008 | + | 1.58 | -1.05 | -2.63 | DONOR | OR51B6 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 17 | 5344991 | 0.06 | GREATER |
| 12 | 10332225 | - | 1.84 | -9.04 | -10.88 | ACCEPTOR | RP11-65... | CRYPTICSITE | INTRONIC | 3'-FLANKING | 75 | 10332150 | 5.2 | LESS |
| 14 | 21464681 | - | 8.69 | 9.78 | 1.09 | ACCEPTOR | RP11-84... | CRYPTICSITE | INTRONIC | 3'-FLANKING | 191 | 21464490 | 15.25 | LESS |
| 15 | 91474977 | + | -3.51 | 0.22 | 3.73 | DONOR | UNC45A | CRYPTICSITE | INTRONIC | 3'-FLANKING | 167 | 91474283 | 7.41 | LESS |
| 16 | 3598755 | - | -7.40 | 3.48 | 10.88 | ACCEPTOR | NLRC3 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 1 | 3598754 | 8.17 | LESS |
| 17 | 10351692 | - | 2.93 | 1.91 | -1.03 | ACCEPTOR | MYH4 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 248 | 10351444 | 2.67 | GREATER |
| 17 | 73816184 | - | 4.80 | 2.80 | -2.01 | ACCEPTOR | AC08728... | CRYPTICSITE | INTRONIC | 3'-FLANKING | 49 | 73816135 | 3.15 | GREATER |
| 17 | 73816190 | - | -10.29 | 1.38 | 11.67 | ACCEPTOR | AC08728... | CRYPTICSITE | INTRONIC | 3'-FLANKING | 55 | 73816135 | 3.15 | LESS |
| 20 | 3674074 | - | -0.67 | 0.78 | 1.44 | ACCEPTOR | SIGLEC1 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 295 | 3673779 | 3.32 | LESS |
| 22 | 30857538 | - | 0.86 | -0.51 | -1.37 | ACCEPTOR | SEC14L3 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 71 | 30857467 | 8.37 | LESS |
| X | 13337617 | + | 1.70 | 5.30 | 3.59 | DONOR | GS1-600... | CRYPTICSITE | INTRONIC | 3'-FLANKING | 4 | 13337613 | 2.03 | GREATER |
| X | 1541330... | - | 0.69 | 2.89 | 2.20 | ACCEPTOR | F8 | CRYPTICSITE | INTRONIC | 3'-FLANKING | 278 | 154132800 | 7.0 | LESS |
| 1 | 2127985... | - | -1.48 | 0.02 | 1.50 | DONOR | RP11-33... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -42 | 212798575 | 6.6 | LESS |
| 1 | 1522857... | + | -6.62 | 1.13 | 7.75 | ACCEPTOR | RP1-14N... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -138 | - | - | - |
| 1 | 1551678... | + | 0.24 | -1.22 | -1.45 | ACCEPTOR | RP11-26... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -71 | 155168193 | 2.15 | LESS |
| 1 | 1551726... | + | 2.52 | 1.43 | -1.09 | ACCEPTOR | RP11-26... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -272 | 155172913 | -0.1 | GREATER |
| 1 | 2409661... | - | 1.83 | -16.80 | -18.63 | DONOR | RGS7 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -83 | 240966203 | 3.72 | LESS |
| 2 | 1794046... | + | 4.01 | 1.32 | -2.69 | ACCEPTOR | AC00994... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -166 | 179404833 | 3.29 | GREATER |
| 2 | 1523142... | + | -2.38 | 8.50 | 10.88 | ACCEPTOR | RIF1 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -1 | 152314274 | 12.34 | LESS |
| 3 | 38139137 | + | 1.81 | -0.36 | -2.18 | ACCEPTOR | DLEC1 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -97 | 38139234 | 0.11 | GREATER |
| 3 | 9807660 | - | -1.24 | 2.49 | 3.73 | DONOR | CAMK1 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -148 | 9807808 | 4.37 | LESS |
| 3 | 58368239 | + | 4.75 | 2.58 | -2.16 | ACCEPTOR | PXK | CRYPTICSITE | INTRONIC | 5'-FLANKING | -129 | 58368240 | 15.23 | LESS |
| 4 | 1562814... | + | 0.80 | -0.75 | -1.55 | ACCEPTOR | AC09746... | CRYPTICSITE | INTRONIC | 5'-FLANKING | -27 | 156281436 | 11.93 | LESS |
| 7 | 43590042 | + | 0.45 | 3.58 | 3.13 | ACCEPTOR | HECW1 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -1 | 43590043 | 7.59 | LESS |
| 7 | 1072043... | + | 1.84 | 4.07 | 2.23 | ACCEPTOR | DUS4L | CRYPTICSITE | INTRONIC | 5'-FLANKING | -126 | - | - | - |
| 7 | 6449756 | - | -1.32 | 0.28 | 1.60 | DONOR | DAGLB | CRYPTICSITE | INTRONIC | 5'-FLANKING | -4 | 6449760 | 7.41 | LESS |
| 8 | 15588173 | + | -8.37 | 2.51 | 10.88 | ACCEPTOR | TUSC3 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -1 | 15588174 | 12.55 | LESS |
| 10 | 44140111 | + | 9.16 | 11.56 | 2.40 | ACCEPTOR | ZNF32-AS3 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -186 | 44140297 | -15.0 | GREATER |
| 10 | 75672761 | - | 0.21 | -11.38 | -11.59 | DONOR | C10orf55 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -39 | 75672800 | 5.75 | LESS |
| 11 | 74556113 | - | 4.49 | 6.09 | 1.60 | DONOR | XRRA1 | CRYPTICSITE | INTRONIC | 5'-FLANKING | -4 | 74556117 | 6.54 | LESS |

Table Settings

Column width: Manual

Show column

- ☑ Chromosome
- ☑ Coordinate
- ☑ Strand
- ☑ Ri-initial
- ☑ Ri-final
- ☑ ΔRi
- ☑ Type
- ☑ Gene Name
- ☑ Location
- ☑ Location Type
- ☑ Loc. Rel. to exon
- ☑ Dist. from nearest nat. site
- ☑ Loc. of nearest nat. site
- ☑ Ri of nearest nat. site
- ☑ Cryptic Ri relative to nat.
- ☑ rsID if available
- ☑ Average heterozygocity
- ☑ Input coordinate
- ☑ Input variant
- ☑ Input ID

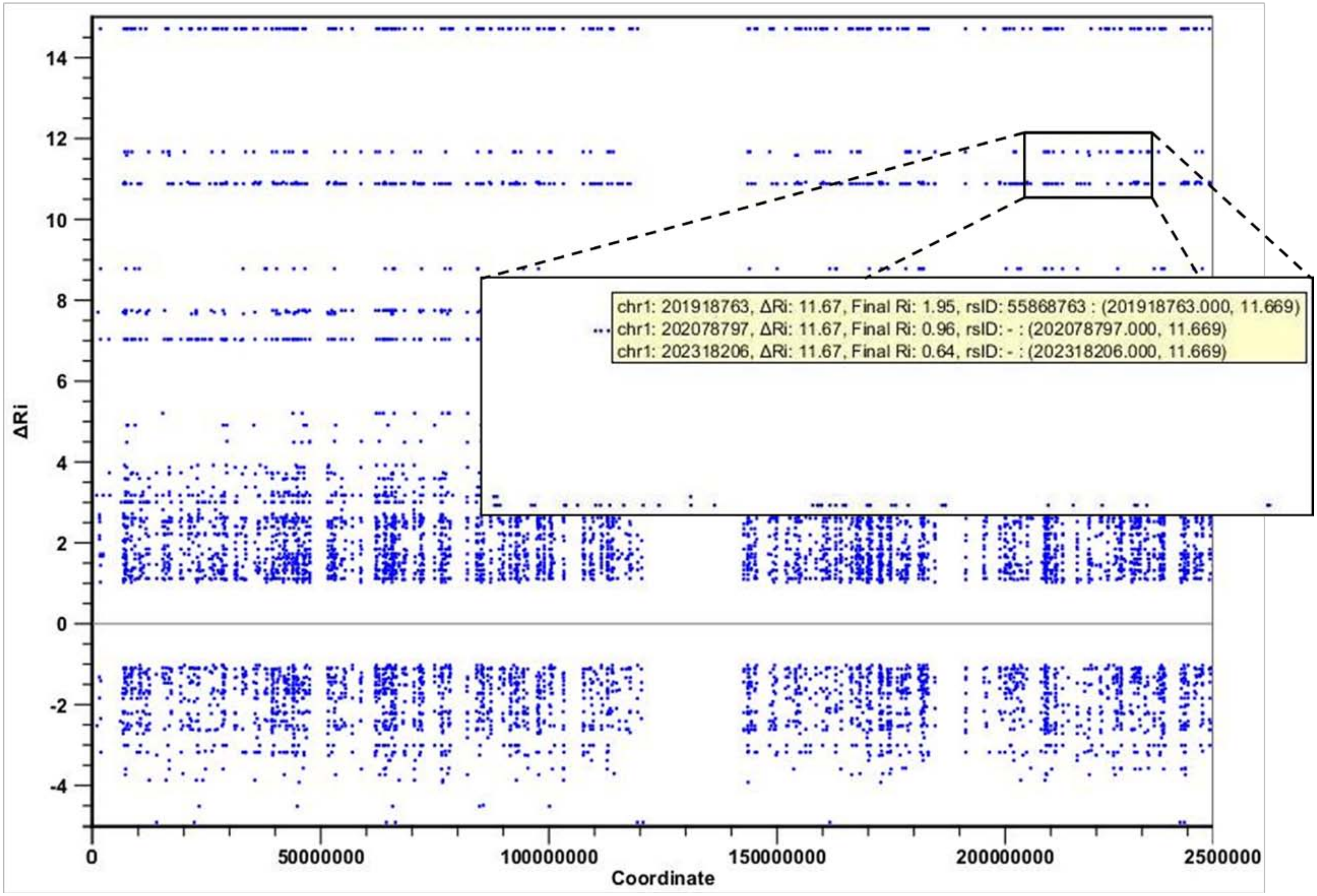
Select All

Deselect All

Idle...

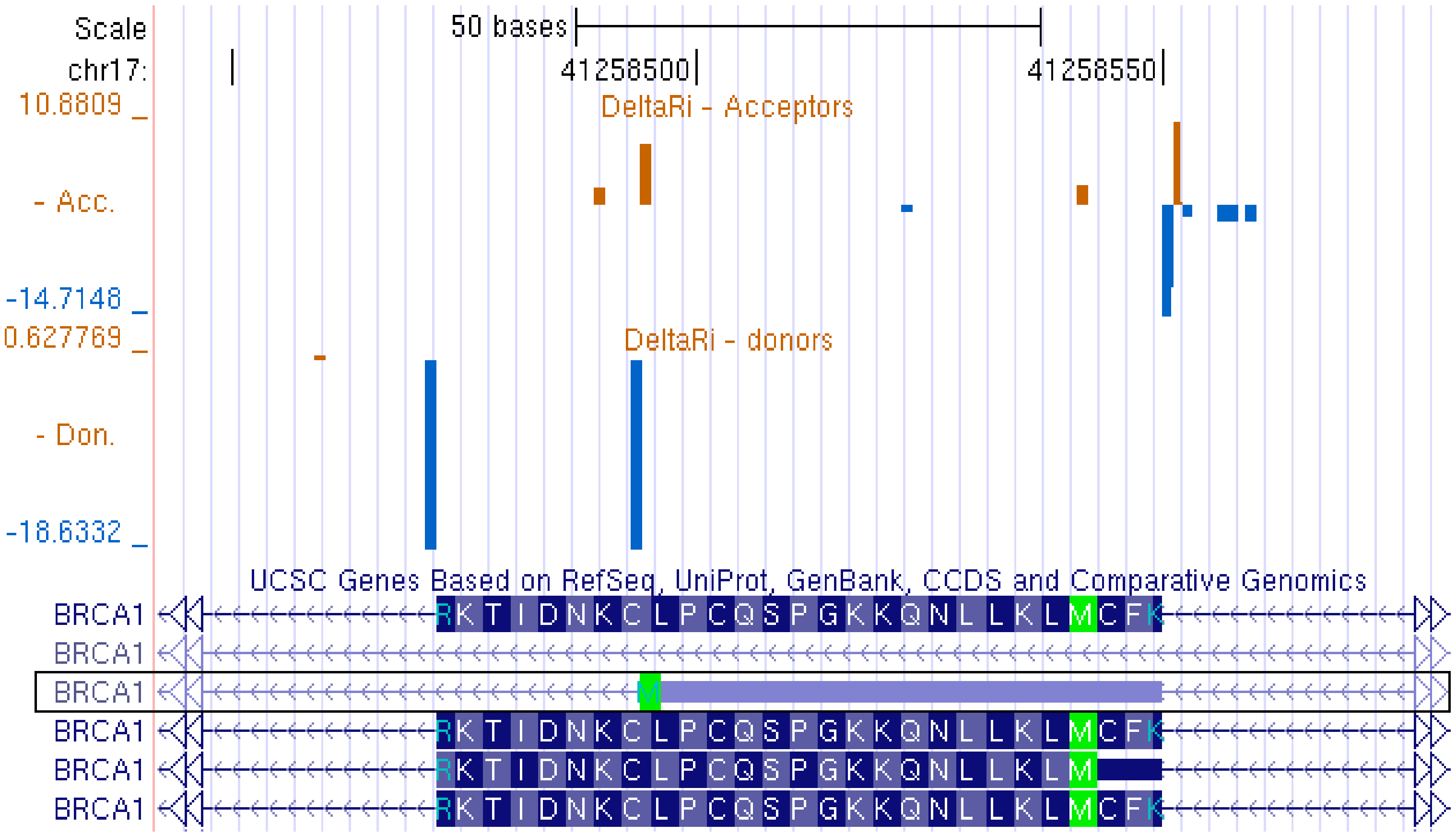
0 rows selected

``Manhattan``-like plot output of potential splicing mutations (unfiltered) on chromosome 1 generated by *CytognomiX*'s Shannon mutation pipeline. Inset shows context-dependent mutation detail generated by mouse-over:



```
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)
```

Graphical custom genome browser track output produced by *Cytognomix*'s Shannon mutation pipeline indicating information changes of a series of mutations occurring in the *BRCA1* gene (Mucaki et al. Hum. Mut. 32:735-742, 2012):


Scale
50 bases
chr17:
41258500
41258550
10.8809 _
DeltaRi - Acceptors
- Acc.
-14.7148 _
0.627769 _
DeltaRi - donors
- Don.
-18.6332 _
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BRCA1
BRCA1
BRCA1
BRCA1
BRCA1
BRCA1
R K T I D N K C L P C Q S P G K K Q N L L K L M C F K
M
R K T I D N K C L P C Q S P G K K Q N L L K L M C F K
R K T I D N K C L P C Q S P G K K Q N L L K L M
R K T I D N K C L P C Q S P G K K Q N L L K L M C F K