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

Presented by Ben Shirley

©Cytognomix Inc 2013

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: $\Delta S = -k_{\rm B} \ln(2)R.$

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

constant $\rightarrow -k_BT \ln 2 < q / R_i$ (joules/bit)

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



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)



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



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

🦚 Applications Places System 🍪 🕐		😽 📢) 🖂 Mon 27 Aug, 5:22 PM 😣 bshirley 🖒	t t
CLC Genomics Workbench 5.5	i		
File Edit Download View Toolbox Wo	rkspace <u>H</u> elp	Workspace Plug-ins Download	Toom In Zoom In Zoom Out
Navigation Area	 Launch Pipeline Choose where to run 	Select Filters and Preferences	
CLC_ServerData	 Select input file containing variants Set filter options 	Genome Build hg19 Types Show Donors Show Natural Sites Show Acceptors Show Cryptic Sites Show Donors and Acceptors Show Cryptic and Natural Sites Output Format Show Delta Ri Plots Show Total Ri Plots Show Negative Strand Show Both Types of Plots Show Both Strands	
Q <		✓ Previous ✓ Einish X Cancel Import data New sequence Read tutorials	
🎯 🥅 Idle			1 element(s) are selected

Filtered cryptic splice site variants

	ß		2 m n	a f	7 🖸	E4										n k	Ŷ
Save Impo	n Export 6	iraphics Prot Undo Redo	Cut Copy Paste	Delete Wer	espace Plug-ins (Download								Fit Wi	dth 100% P	an Selection Z	oom In
tion	Pos 5	Strand Ac 😵 🗄 * T	rack List X	i 🔲 Inact	ivating ×	Com	plete Va	ari X	🖽 Сгур	tic Varia	×	1					
8 V							-										
mp 🔺	Rows:	: 134 / 22,197 Effect	of variants (on Ri and o	ther relevant	informati	ion		Filter:			Match	any	ģ	Match all	I	
acti	Ν				Di	st. from r	nearest	nat. site	- does	n't contain	-	-					1
аку												0					5
nor						u			• <u>></u>	· ·							_
ır 1 👘 🛛					Cr	yptic Ri r	elative t	o nat.			-	GREATER					
r 2 🚽					Ri	-final			• >	-		1.6					٦.
r 3																	Ap
ir 5																	
r 6	Chromoso	Coordinate Strand	Ri-initial	Ri-final	ARi	Type Ge	ene Name	Location	Location	n Loc. Rel	. t (Dist. from	Loc. of ne	. Ri of near.	. Cryptic Ri .	rsID if ava	Aver
7 8	1	1552560 +	-16.41	2.22	K 18.63 DOM	JOR HO	:N3	CRYPTICS	. INTRON	C 3'-FLAN	d	263	1552557	2.06	GREATER	rs14473	0
r 8 🔡	Z	2333068 +	-8.17	10.46	ARIB.63 DOM	IOR DI	S3LZP1	CRYPTICS	. INTRONI	C 3'+LAN	g	19	2333068	0.17	GREATER	rs790027	0.21
r9 🕴	5	1760831 +	-13,73	4.91	18.63 DOM	IOR IS	PANL7	CRYPTICS	INTRONI	C 3"+LAINI	d	48	1760831	4.24	GREATER	rs68/89//	0
10 8	0	44140320 +	-10.77	7.80	18.63 DOM	IOR US	APNII EDE1	CRYPTICS	INTRONI	C 3+LAN	a	102	44140158	0.0	GREATER	151418488	0.4
11 8	6	1424095 +	-12,80	5.77	18.63 DOM		00001	CRYPTICS	INTRONI	C 3+LAN	a. 1	80	1424094	3.57	GREATER	rs14071	0
12-18	9	1224650	-10.75	3.33	18.63 DOM	IOR CP	AMD1 D	CRYPTICS	INTRONI		a 1	140	1224649	-37.04	GREATER	1514201	U
	12	1234050 +	-10,25	0.39	18.63 DOM		CILLO	CRYPTICS	INTRONI	C 3+LAN	a 1	101	1234048	42.02	GREATER	-	0.4
	12	0303433 +	-13.77	2.07	18.63 DOM	IOR AL	Jorfe D	CRYPTICS	INTRON	C 3-FLAN	d	192	0303241	-43.03	GREATER	1929/0104	0.4
ea 🕆 🕴	12	26104042	12.07	5.45	18.63 DOM		201103	CRYPTICS.	INTRONU	C 3-FLAN	a 1	200	26104700	3.43	GREATER	19/300302	0.0
	10	12262067	15.06	3.00	18.63 DOM		DEA	COVETICE	INTRONU		1 1	199	12262060	1.45	GREATER	199301377	0.43
	10	56272579 4	-14.00	2.65	18.62 DOM		DDA	CRYPTICS	INTRONI	C 2' ELANI	1	50	56272526	2.65	GREATER	re10952	0.20
n Hu	20	61505537 +	-16.08	2.55	18.63 DOM		01740	CRYPTICS	INTRONI	C 31-ELANI	4	162	61505375	1.96	GREATER	rs2427464	0
och E	21	47410626 +	-11.00	7.63	18.63 DOM		1641	CRYPTICS	INTRONI	C 3'-ELANI	1	280	47410337	7.04	GREATER	rs1080081	0
al Se	1	1831005 +	-12.57	6.05	18.62 DOM		MC1	CRYPTICS	INTRONI	C 3'-FLAN	1	48	1831005	4.74	GREATER	rs4997370	0.4
lor Bi	2	1203962 +	-13.85	4.78	18.62 DOM	JOR AC	06915	CRVPTICS.	INTRONI	C 3'-FLAN	4	242	1203959	3.9	GREATER	rs939772	0.10
	2	1487160	-11.87	6.76	18.62 ACC	EPTOR OF	RC4	CRVPTICS.	INTRONI	C 3'-FLAN	d	145	1487159	4.54	GREATER	rs12463	0.47
re Tr	10	1274849	-10.57	8.05	18.62 ACC	EPTOR UP	105	CRYPTICS	INTRONI	C 3'-FLAN	4	198	1274847	6.38	GREATER	rs2281953	0.1
	10	1014193 +	-14.90	3.72	18.62 DOM	JOR EN	TPD7	CRYPTICS	INTRONI	C 3'-FLAN	a :	33	1014193	3.47	GREATER	rs3740080	0.4
	11	62749102 -	-13.18	5.45	18.62 ACC	EPTOR SL	C22A6	CRYPTICS	INTRONI	C 3'-FLANI	d 1	246	62748856	4.96	GREATER	rs4149173	0.4
otor	16	5038409 +	-11.74	6.89	18.62 DOM	JOR SE	C14L5	CRYPTICS	INTRONI	C 3'-FLANI	d 1	127	5038282	5.95	GREATER	rs2972261	0.4
pcor	17	40025263 -	-11.91	6.71	18.62 ACC	EPTOR AC	CLY.	CRYPTICS	INTRONI	C 3'-FLAN	d :	224	40025039	1.76	GREATER	rs9912300	0.36
miles a	22	16340610 +	-16.84	1.79	18.62 DOM	JOR LA	16c-59	CRYPTICS	INTRONI	C 3'-FLAN	d :	27	16340583	-25.97	GREATER	-	-
o sec	1	1006220	-4.91	9.81	14.71 ACC	EPTOR LR	RC39	CRYPTICS	INTRONI	C 3'-FLAN	d)	133	1006218	9.78	GREATER	rs14039	0
swo	2	98130015 -	-11.15	3.57	14.71 ACC	EPTOR AN	KRD36B	CRYPTICS	INTRONI	C 3'FLAN	d 1	188	98129827	1.37	GREATER	rs14355	0
	8	1439959	-10.38	4.34	14.71 ACC	EPTOR CY	P11B2	CRYPTICS	INTRONI	C 3'FLAN	d)	158	1439958	1.98	GREATER	rs79658	0
	9	38603383 -	-11.24	3.48	14.71 ACC	EPTOR AN	KRD18A	CRYPTICS	INTRONI	C 3'FLAN	d 1	173	38603210	2.05	GREATER	rs631327	0.43
- F E	12	57425145 -	-7.22	7.49	14.71 ACC	EPTOR MY	/01A	CRYPTICS	INTRONI	C 3'FLAN	d 1	273	57424959	0.98	GREATER	rs755221	0
	4					ii.					-						-



Mutation viewed in CLC-Bio genome browser



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

		Novel n	nutations	SNPs with		
				alle		
Cell line	Initial	Natural	Cryptic site	Natural	Cryptic site	Overall
	variants	site		site		Mutation
	analyzed					fraction
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; $6.89 \rightarrow -11.73$ 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; $10.24 \rightarrow 6.67$ bits).

Both alleles of *APIP*, an apoptosis associated gene, are inactivated in U251 (chr11:34905054G>C; $9.32 \rightarrow 0.54$ 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 \rightarrow -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/

CytognomiX

©Cytognomix Inc 2013 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



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 <u>4-9% enzymatic activity</u>. The reduction in information content is significant even though the Ri value is still much greater than Ri,min.



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 preexisting 8.7 bit cryptic site that is strengthened by the mutation to 10.2 bits (>=2.8 fold).