Variant Annotation

M7 Bioinformatics

MSc Genomic Medicine

Tues 16th Feb 2016

Simon Topp

simon.topp@kcl.ac.uk

Basic & Clinical Neuroscience

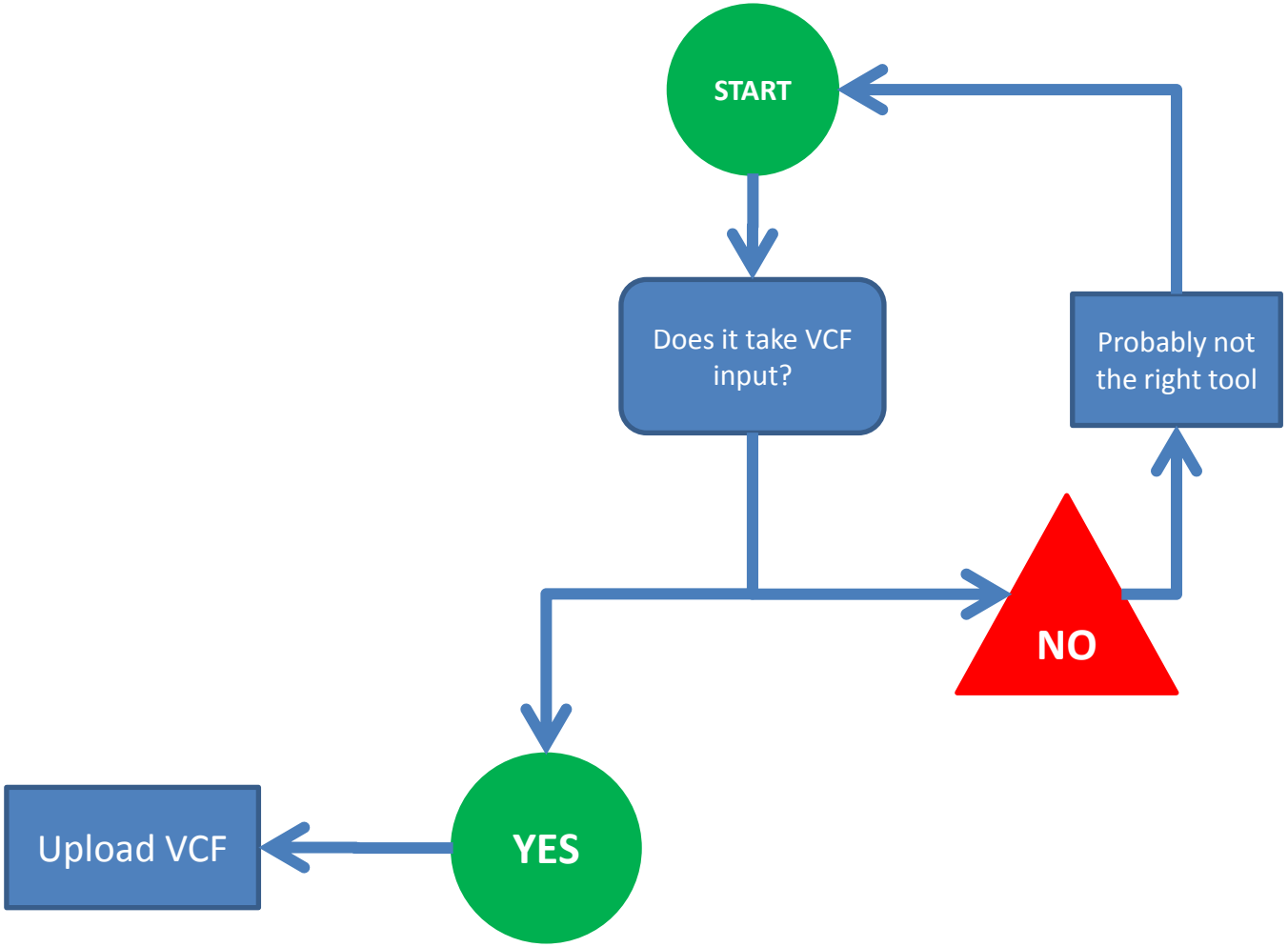
Institute of Psychiatry, Psychology & Neuroscience

Overview

- Annotation Tools
- Transcript choice
- Variant Nomenclature
- Left vs Right alignment issues
- Population Databases
- Pathogenicity Predictors
- Non-coding Variants

Annotation Tools

http://omictools.com/variant-annotation-category



Annotation Tools

- **3 most popular tools:**
  - **wANNOVAR**  
  - http://wannovar.usc.edu/

  - **SNPeff** 
  - http://snpeff.sourceforge.net/

  - **Variant Effect Predictor** 
  - http://www.ensembl.org/Tools/VEP

Also recommended...

(command line only)

- **VAT – Variant Annotation Tool**
  - http://vat.gersteinlab.org/index.php

- **VAAST - Variant Annotation, Analysis and Search Tool**
  - http://www.yandell-lab.org/software/vaast.html

- **SNPAAMapper - A SNP Amino Acid Mapping tool**
  - http://www.ccmb.med.umich.edu/ccdu/SNPAAMapper

Genome Version

- **hg18 (NCBI36) Mar 2006**
  - Deprecated, but used in many older publications and resources
  - If you can't find the variant they refer to, check the small print in the methods and the publication date!
- **hg19 (GRCh37) Feb 2009**
  - Most commonly used (still)
- **hg38 (GRCh38) Dec 2013**
  - Better for HLA and repeat-rich regions
  - Very little uptake by scientific community
  - Fewer annotations available mapped to it

Transcripts - Refseq

- Curated by the NCBI
- http://ncbi.nlm.nih.gov/gene
  - NM_nnnnn.n mRNA
  - NR_nnnnn.n non-coding
  - NP_nnnnn.n protein
  - XM_/XR_/XP_ predicted

- Manually curated, high quality, not comprehensive. Some issues in genome mapping have been reported (incorrect location of splice sites in 3% of transcripts)

GAPDH in Refseq

4 Alternately spliced isoforms
3 of them encode the same protein (splicing only affects the UTR)

NM_001256799.2 → NP_001243728.1 **glyceraldehyde-3-phosphate dehydrogenase isoform 2**

See identical proteins and their annotated locations for NP_001243728.1

Status: REVIEWED

Description	Transcript Variant: This variant (2) differs in the 5' UTR and coding region compared to variant 1. These differences cause translation initiation at a downstream AUG and result in an isoform (2) with a shorter N-terminus compared to isoform 1.
Source sequence(s)	AF261085, BE893087, HY000136
Consensus CDS	CCDS58201.1
UniProtKB/Swiss-Prot	P04406
Related	ENSP00000380067, OTTHUMP00000174434, ENST00000396858, OTTHUMT00000268066

NM_001289745.1 → NP_001276674.1 **glyceraldehyde-3-phosphate dehydrogenase isoform 1**

See identical proteins and their annotated locations for NP_001276674.1

Status: REVIEWED

Description	Transcript Variant: This variant (3) differs in the 5' UTR, compared to variant 1. Variants 1, 3, and 4 encode the same isoform (1).
Source sequence(s)	BE893087, BM763361, HY046784, M33197
Consensus CDS	CCDS8549.1
UniProtKB/Swiss-Prot	P04406
UniProtKB/TrEMBL	V9HVZ4
Related	ENSP00000380070, OTTHUMP00000174431, ENST00000396861, OTTHUMT00000268060

NM_001289746.1 → NP_001276675.1 **glyceraldehyde-3-phosphate dehydrogenase isoform 1**

See identical proteins and their annotated locations for NP_001276675.1

Status: REVIEWED

Description	Transcript Variant: This variant (4) differs in the 5' UTR, compared to variant 1. Variants 1, 3, and 4 encode the same isoform (1).
Source sequence(s)	BC023632, BE893087, HY004110, HY022295
Consensus CDS	CCDS8549.1
UniProtKB/Swiss-Prot	P04406
UniProtKB/TrEMBL	V9HVZ4

NM_002046.5 → NP_002037.2 **glyceraldehyde-3-phosphate dehydrogenase isoform 1**

See identical proteins and their annotated locations for NP_002037.2

Status: REVIEWED

Description	Transcript Variant: This variant (1) encodes the longer isoform (1).
Source sequence(s)	BC009081, BE893087, HY046784
Consensus CDS	CCDS8549.1
UniProtKB/Swiss-Prot	P04406
UniProtKB/TrEMBL	V9HVZ4

Transcripts - Ensembl

- Curated by the EBI (www.ensembl.org)
  - ENSTnnnnnnnnnnnn – mRNA
  - ENSPnnnnnnnnnnnn – protein

- Not manually curated. Usually one Ensembl transcript for every uniquely observed splicing pattern. Many are partial fragments of full length transcripts or pre-mRNA with unspliced introns (junk).
- Can result in annotation-overload.

GAPDH in Ensembl

11 Transcripts

Show All entries | Show/hide columns (1 hidden) | Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
GAPDH-001	ENST00000229239	1875	335aa	Protein coding	CCDS8549	P04406 V9HVZ4	NM_002046 NP_002037	TSL:1 GENCODE basic APPRIS P1
GAPDH-002	ENST00000396861	1348	335aa	Protein coding	CCDS8549	P04406 V9HVZ4	NM_001289745 NP_001276674	TSL:5 GENCODE basic APPRIS P1
GAPDH-008	ENST00000396858	1292	293aa	Protein coding	CCDS58201	P04406	NM_001256799 NP_001243728	TSL:5 GENCODE basic
GAPDH-003	ENST00000396859	1256	335aa	Protein coding	CCDS8549	P04406 V9HVZ4	NM_001289746 NP_001276675	TSL:1 GENCODE basic APPRIS P1
GAPDH-004	ENST00000396856	1266	260aa	Protein coding	-	E7EUT5	-	TSL:5 GENCODE basic
GAPDH-201	ENST00000619601	1086	293aa	Protein coding	-	P04406	-	TSL:5 GENCODE basic
GAPDH-007	ENST00000466525	1720	No protein	Retained intron	-	-	-	TSL:5
GAPDH-005	ENST00000466588	1363	No protein	Retained intron	-	-	-	TSL:5
GAPDH-006	ENST00000474249	1333	No protein	Retained intron	-	-	-	TSL:5
GAPDH-011	ENST00000492719	930	No protein	Retained intron	-	-	-	TSL:3
GAPDH-010	ENST00000496049	390	No protein	Retained intron	-	-	-	TSL:2

4 overlap with Refseq

5 have retained introns – probably unspliced pre-mRNA

LRG – Locus Reference Genomic

- “LRG sequences provide a stable genomic DNA framework for reporting mutations with a permanent ID and core content that never changes.”
- “Only transcripts with good biological understanding and essential for reporting variants will be included.”
- not in widespread use (yet). Not available for every gene (yet) - 1070 to date.
- http://www.lrg-sequence.org/

Canonical Transcripts

- Usually defined as the splice isoform with the longest Open Reading Frame (translated protein)
- Not always clear which isoform is canonical
- NOT necessarily the most abundant isoform
- NOR the one first discovered / most cited in the literature
- NOR the primary sequence in UniProt.
- May not be present in your tissue of interest or relevant to your disease
- Therefore variants must be mapped to all isoforms.

Popular Terminology

SNP – Single Nucleotide Polymorphism
SNV – Single Nucleotide Variant (preferred)
CNV – Copy Number Variant

Mutation – a variant likely to contribute to disease
SNP – A (usually common) benign variant

Missense – a single amino acid change
Nonsense – A truncated (or extended) protein sequence

HGVS Nomenclature

- Human Genome Variation Society (hgvs.org)
- Standardised naming system for genomic / transcript / protein variants
- Always cite the genome / transcript / protein unique identifier it refers to (with version numbers).
  - **g.nnnnnnn** co-ordinate on chromosome
    - hg19:chr1:g.1234567
  - **c.nnnn** position in cDNA (A of ATG start codon = c.1)
    - NM_012345.1:c.128A>T / ENST00000012345:c.128A>T
  - **p.nnn** position in amino acid sequence
    - NP_012534.2:p.P20L / ENSP000000012354:p.P20L

HGVS cDNA variants

- **Substitution (SNV only!)**
  - c.123A>G
- **deletion**
  - c.123del
  - c.123delA
  - c.586_591del
  - c.586_591delTGGTCA
- **duplication**
  - c.123dup
  - c.123dupA
  - c.586_591dup
  - c.586_591dupTGGTCA
- **insertion**
  - c.123_124insC
  - !NOT! c.123insC
  - c.1086_1087insGCGTGA
- **complex indel**
  - c.123_136delinsAGT
  - c.123_125delTGAinsACC

- **protein coding region**
  - c.1637A>G
- **in intron (5' half)**
  - c.859+12T>C
- **in intron (3' half)**
  - c.2396-6G>A
- **5' of protein coding region (5' UTR)**
  - c.-23C>G
- **3' of protein coding region (3' UTR)**
  - c.*143A>T
- **intron in 5' UTR (5' of ATG)**
  - c.-89-12T>G
- **intron in 3' UTR (3' of stop)**
  - c.-649+79G>C

HGVS Amino Acid Variants

- **substitution**
  - p.Ala23Thr
  - p.A23T
- **stop-gain**
  - p.Arg105*
  - p.R105X * is preferred to X for stop codons
- **stop-lost**
  - p.*673R
  - p.X673R
- **frameshift**
  - p.Arg83fs
  - p.Arg83Serfs*15 frameshift changes Pro to Ser, and new stop codon introduced 15 residues downstream.
- **indel**
  - p.R123delinsKKK
  - p.R123_K136delinsGGQQQQGG

Left vs Right Alignment

HGVS standard is to report variant at most 3' position, relative to transcript. For forward strand genes:

```
          M  K  K  K   *
REF:  CCCATGAAAAAAAAA-TGACCC          g.15dupA   g.15_16insA
VAR:  CCCATGAAAAAAAAAATGACCC          c.12dupA   c.12_13insA
          M  K  K  K  M  T            p.*5Mfs*?
```

But for most read aligners/variant callers (not all) the standard is to left-shift variants, relative to the genome reference forward strand.

```
          M   K  K  K  *
REF:  CCCATG-AAAAAAAAATGACCC          g.6_7insA
VAR:  CCCATGAAAAAAAAAATGACCC          c.3_4insA
          M  K  K  K  M  T            p.*5Mfs*?
```

Some annotation pipelines correct for this, some don't.

For reverse strand genes, This issue should cancel out

```
FORWARD VAR:   5' CCCTCATTTTTTTTTTTGAGGG 3'       g.6_7insT
FORWARD REF:   5' CCCTCA-TTTTTTTTTTGAGGG 3'
                  |||||<---------||||||
REVERSE REF:   3' GGGAGT-AAAAAAAAAGTACCC 5'
REVERSE VAR:   3' GGGAGTAAAAAAAAAAGTACCC 5'       c.12_13insA
                    T  M  K  K  K  M              p.*5Mfs*?
```

VCF Normalisation Examples

Right-Aligned

```
       REF:ATCTTTTTCTA
           ||||||| |||
       VAR:ATCTTTT-CTA
```

VCF: 7 TT/T
ANNOVAR: 8 T/-
HGVS: c.8del

```
       REF:ACTCTCTC--CA
           ||||||||  ||
       VAR:ACTCTCTCTCCA
```

VCF: 8 C/CTC
ANNOVAR: 9 -/TC
HGVS: c.8_9insTC

Left-Aligned

```
ATCTTTTTCTA
||| |||||||
ATC-TTTTCTA
```

VCF: 3 CT/C
ANNOVAR: 4 T/-
HGVS: c.4del

```
A--CTCTCTCCA
|  |||||||||
ACTCTCTCTCCA
```

VCF: 1 A/ACT
ANNOVAR: 1 -/CT
HGVS: c.1_2insCT

Multi-allelic Loci

- VCFs have one row per *locus*
- Annotation output has one row per *variant*

eg

➢ VCF:	A/G,ATC,ATCTC	ATCTC/A,ATC,G
➢ ANNOVAR1:	A/G	A/G
➢ ANNOVAR2:	-/TCTC	TCTC/-
➢ ANNOVAR3:	-/TC	TC/-

Can’t find your annotated variant in the original VCF? This is probably why.

Detecting False Positives

- Low Read depth (DP)
- Low GQ value (Genotype Quality)
  - Misaligned reads
  - Homopolymer runs (sequencing errors)
  - Left/right alignment
  - Polymorphic InDels – microsatellites or longer repeats
  - Close paralogues – wrong gene
  - Contamination – wrong species

A real-world False Positive

From our exome capture cohort – after annotation and filtering I had the following NOVEL coding variants, giving a highly significant result to the gene in burden tests.

2x chr10:3208567 T / TGCACGCTAGGGAAGAGAGAG
2x chr10:3208567 T / TGCACGCTAGGGAAGAGAGAGG
1x chr10:3208567 T / TGCACGCTAGGGAAGAGAGAGGA
4x chr10:3208567 T / TGCACGCTAGGGAAGAGAGAGGAA

However, in ExAC:
Variant: chr10:3208567 T / TGCACGCTAGGGAAGAGAGAGGAATG

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	524	3472	15	0.1509
European (Non-Finnish)	375	16208	6	0.02314
European (Finnish)	27	1184	0	0.0228
Latino	44	3608	0	0.0122
Other	1	276	0	0.003623
South Asian	27	8630	0	0.003129
East Asian	1	2538	0	0.000394
Total	**999**	**35916**	**21**	**0.02781**

Other alt alleles in ExAC:
chr10-3208567-T-TGCACACTAGGGAAGAGAGAGGAATG
chr10-3208567-T-TGCATGCTAGGGAAGAGAGAGGAATG
chr10-3208567-T-TAGGGAAGAGAGAGGAATG
chr10-3208567-T-TGCACGCTAAGGAAGAGAGAGGAATG

Common polymorphic insertion

Overlapping Genes

Most annotation tools have a heirarchy of significance
eg,
stop-lost > splicing > nonsynonymous > UTR > synonymous > intronic > intergenic

There isn't always a clear 'winner' if two genes share exons

Population Databases

- **dbSNP**
  - NCBI repository for all reported small variants.
  - 87 million human variants in version 146
  - dbSNP ids: **rsnnnnnn**
  - One entry can encompass multiple variants
  - http://ncbi.nlm.nih.gov/snp

- **1000 genomes** (1000g)
  - Entire genome sequence and variants from ~2,500 people across 25 selected population groups.
  - http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes

- **ESP/EVS** (Exome Sequencing Project / Exome Variant Server)
  - Exome variants from 6,300 people (4,300 European-American & 2,000 African-American)
  - http://evs.gs.washington.edu/EVS

Population Databases

- **ExAC** (Exome Aggregation Consortium)
  - Exome variants from ~61,000 people. 50% Are Non-Finnish Europeans. Remainder divided between South Asian, East Asian, Finnish, African, Latino.
  - http://exac.broadinstitute.org
- **UK10K**
  - Exome variants from 3,700 UK individuals
  - http://www.uk10k.org
- **Cosmic**
  - Somatic variants from cancer studies
  - http://cancer.sanger.ac.uk
- **ClinVar**
  - Variants believed to be pathogenic
  - http://www.ncbi.nlm.nih.gov/clinvar

dbNSFP

database for Nonsynonymous SNPs' Functional Predictions

Originally every competing annotation tool used its own derived databases for variant annotations.

Now most obtain them from a single 3rd party source – dbNSFP
http://sites.google.com/site/jpopgen/dbNSFP

Pathogenicity Predictors

	ANNOVAR	VEP	SNPEff	Data
GERP++	X	X	?	DNA Conservation
LRT	X	X	?	DNA Conservation
phastCons	X	X	?	DNA Conservation
phyloP	X	X	?	DNA Conservation
SIFT	X	X	?	DNA Conservation
SiPhy	X	X	?	DNA Conservation
BLOSUM		X	?	AA Conservation
FATHMM	X	X	?	AA Conservation
MutationAssessor	X	X	?	AA Conservation
PolyPhen	X	X	?	AA Conservation
Provean	X	X	?	AA Conservation
CADD	X	X	?	Meta
ConDel		X	?	Meta
RadialSVM	X	X	?	Meta
MetaLR	X	X	?	Meta
VEST3	X	X	?	Meta
DANN	X	X	?	Meta
MutationTaster	X	X	?	AI classifier
MaxEnt		X	?	Splicing
ADA	X	?	?	Splicing
RF	X	?	?	Splicing
LoFtool		X	?	NMD
miRNA	X	X	?	miRNA

Prediction Assessment

- **Independent benchmarking of 20 Annovar-derived prediction tools (Simon Topp, 2016).**
- **Percentage of variants classed as "Damaging", based on 950 validated pathogenic nonsynonymous mutations (True Damaging) vs ~40,000 private ExAC variants (Random).**
- **FATHMM did best, followed by several meta-analysis methods.**
- **PolyPhen and SIFT performed poorly.**

	FATHMM	SVM	LR	VEST3	MutationAssessor	CADD	Provean	SIFT	MutationTaster	PP2_HDIV	LRT	PP2_HVAR	DANN	FATHMM_MKL	SIPHY	PHASTCONS7	PHASTCONS20	GERPrs	PHYLOP7	PHYLOP20
Optimised Threshold	**<= -1.21**	**>= -0.12**	**>= 0.43**	**>= 0.73**	**>= 2.39**	**>= 24.8**	**>= -2.9**	**<= 0.014**	**A or D**	**>= 0.99**	**D**	**>= 0.78**	>= 0.99	**>= 0.84**	**>= 12.23**	**>= 0.83**	**>= 0.93**	**>= 3.84**	**>= 0.77**	**>= 0.84**
True Damaging (%)	87.1	75.8	78.6	77.5	52.5	66.1	64.5	64.6	92.4	61.4	76.7	65.5	77.2	79.5	69.9	87.1	77.5	76.3	81.5	81.1
Random (%)	28.1	19.4	22.2	25.3	19.9	33.6	32.6	35.6	65.8	36.7	52.3	41.2	54.4	58.0	50.2	68.2	59.7	61.2	68.3	69.0
Difference (%)	**59.0**	**56.4**	**56.4**	**52.2**	**32.6**	**32.5**	**31.9**	**29.0**	**26.6**	**24.7**	**24.4**	**24.3**	**22.8**	**21.5**	**19.7**	**18.9**	**17.8**	**15.1**	**13.2**	**12.1**

Other benchmarks have shown DANN > CADD > FATHMM

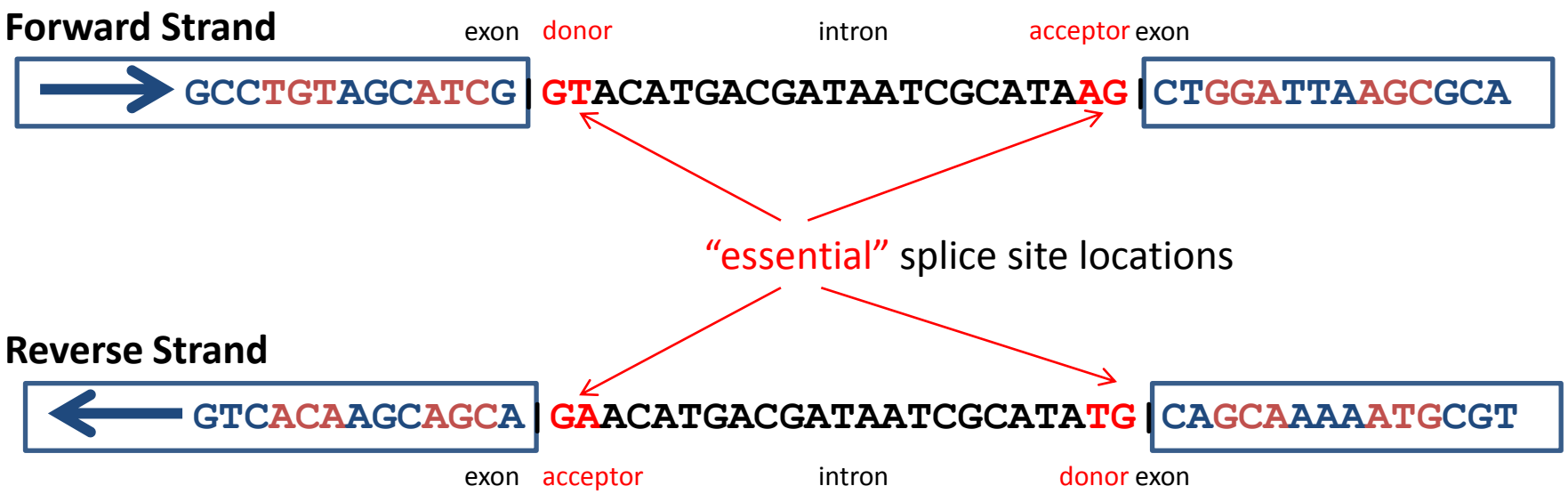
http://www.enlis.com/blog/2015/03/17/the-best-variant-prediction-method-that-no-one-is-using

Depends on your test set data!

Also need to consider…

- Active Sites.
- Post Translational Modification Sites.
- Amyloidogenic / Aggregation-Prone Regions.
- Secondary structure elements (creation and destruction).
- Protease cleavage sites.
- Transmembrane regions.
- Subcellular targeting signals.
- Etc...

Splice Sites



Many tools only annotate "essential splice" or "near splice".
Some incorporate more sophisticated predictions.

Splicing can also be affected by Exonic Splicing Enhancer (ESE) motifs or Exonic Splicing Suppressor (ESS) motifs, .

Splice Site Predictors

dbscSNV

assesses variants:

- 11-base region near the 5' ("donor") end of each intron
- 14-base region near the 3' ("acceptor") end of each intron
- "Essential" splice sites excluded

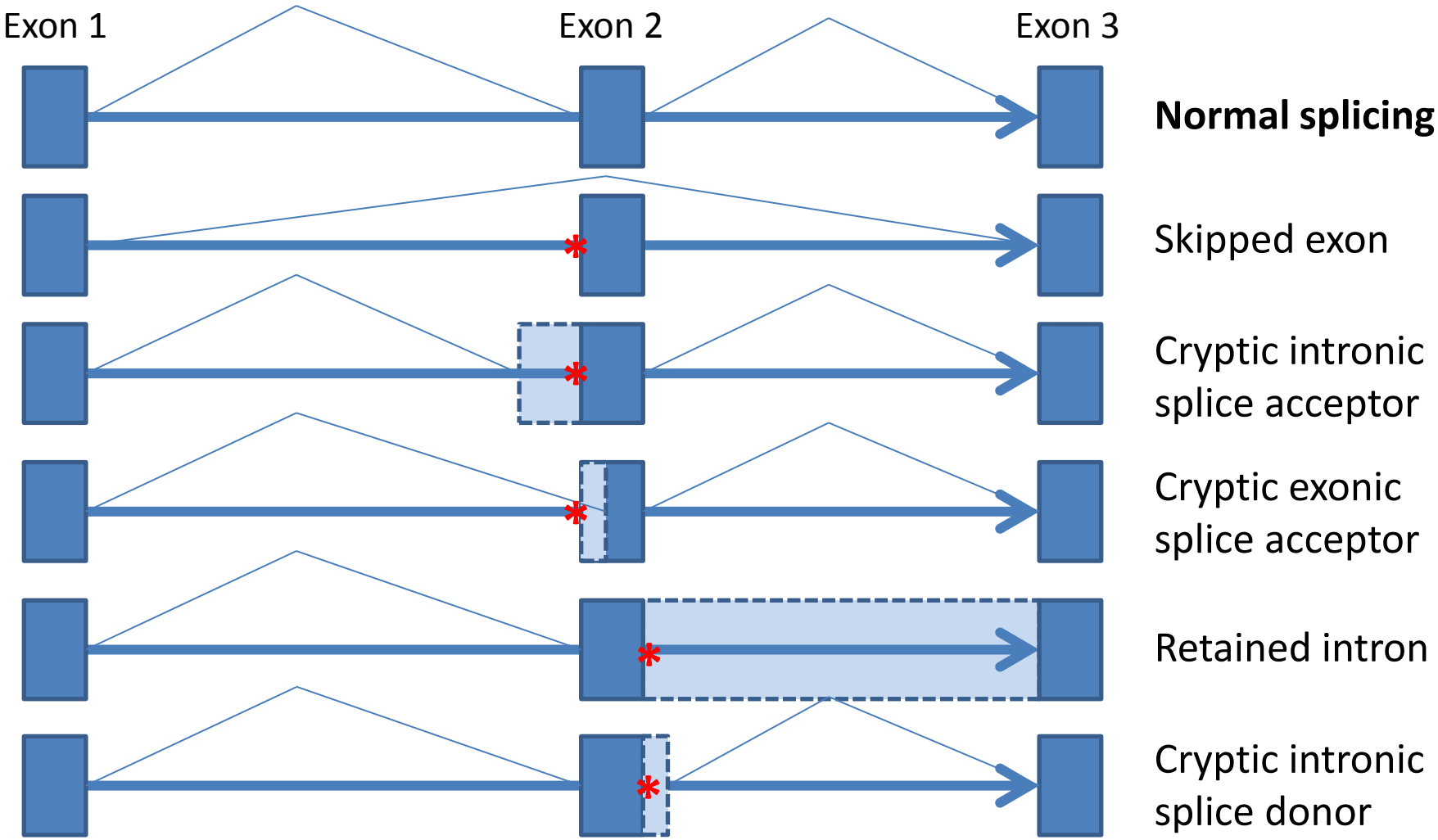
AI consensus of 6 splice site prediction tools (including **MaxEnt**) comparing reference vs variant derives 2 output scores:

RF (Random Forest AI)

ADA (AdaBoost AI)

Score > 0.6 implies probable alteration of a splice site

Splice Site variants



Impact depends heavily on reading frame (insertion/deletion, frameshift?)

Nonsense Mediated Decay

- Premature Termination Codons (PTC) are normally degraded by the NMD pathway.
- Complete knock-out of the gene from one allele.
- Cells can compensate via feedback loops, and increase expression of remaining allele.
- Hence Nonsense mutations are more likely to cause recessive rather than dominant disorders.
- Can bypass NMD if the PTC is in the last exon, possibly producing a misfolded or non functional protein.

Promoters

- Region upstream of (and often overlapping) the 5’ exon of the gene.
- Binding sites for transcription factors, that control the expression of the gene.
- Some annotation tools attempt to map to these
  - Predicted, conserved across species.
  - Validated, ENCODE experimental results.
- Impact of a variant often uncertain and difficult to validate experimentally

3’UTR

- Most common variant of interest in 3’ UTRs are those impacting micro RNA (miRNA) binding sites.
- miRNAs target transcript for degradation, lowering absolute expression levels.
- Some annotation tools attempt to map to these binding sites, but most are only *predicted,* and the impact of a variant uncertain.

Non-coding RNAs

- Many thousands discovered to date.
- Often within introns of coding genes.
- Very poorly understood.
- Can be spliced.
- Opposite strand to ‘host’ gene could have a regulatory function
- Very few have had pathogenic variants identified within them.