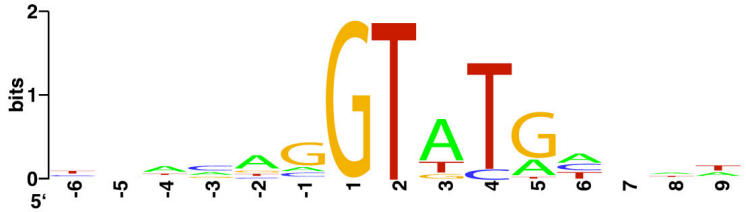
Geuvadis 2012-10-11

Everything you always wanted to know about splicing in Geuvadis

Michael Sammeth, summarizing results obtained so far also by
Pedro Ferreira
Matthias Barann
Anna Esteve

Splice Site Scores



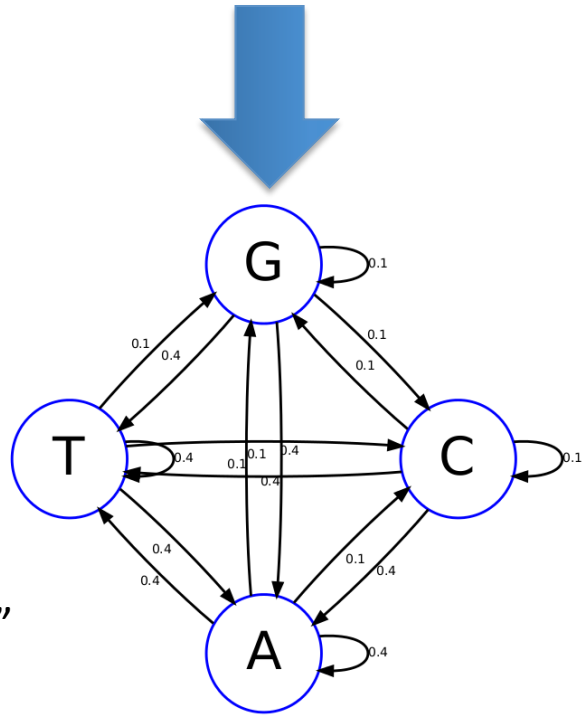
- are numerical condensations of a model that represents the frequency of observations of substrings in a splice site motif

- here: a first order Markov model captures *tansition probabilities* of di-nucleotides (16 x 16 matrix)

- scores are sums of log-likelihoods of underlying probabilities

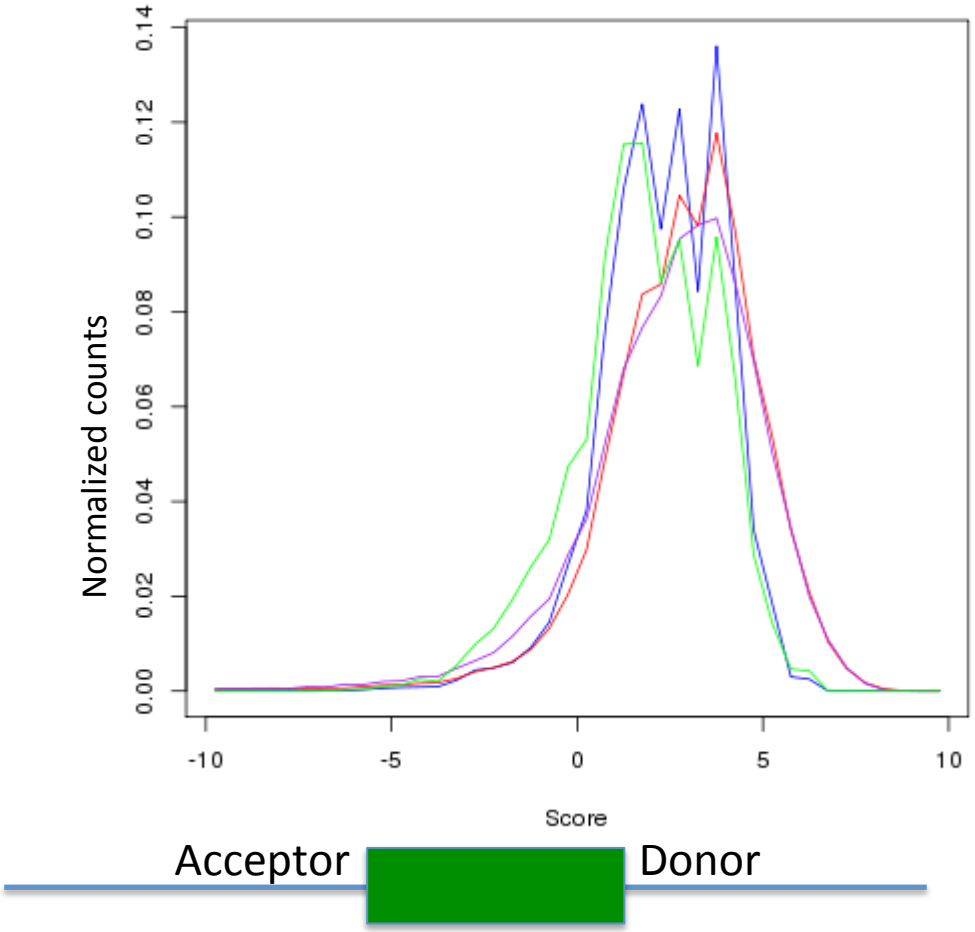
- biological hypothesis: higher scores represent “stronger” splice sites that are thermodynamically efficient

- Scores of –Inf indicate sites that are inactive in their splicing functionality according to the model



scores

Gencode (v12) Score Distributions

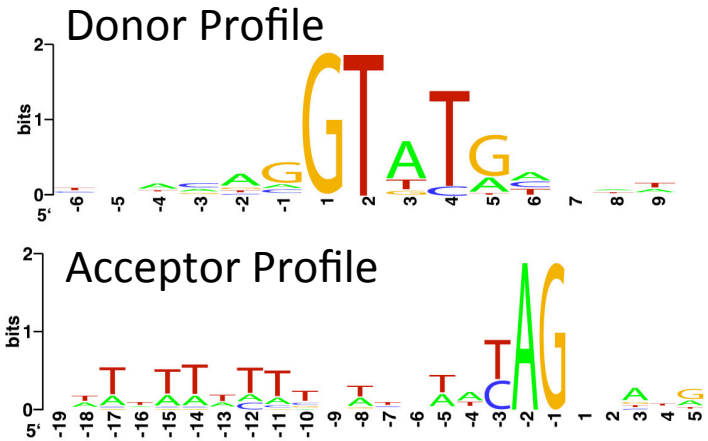


score reference donor

score variant donors

score reference acceptors

score variant acceptors

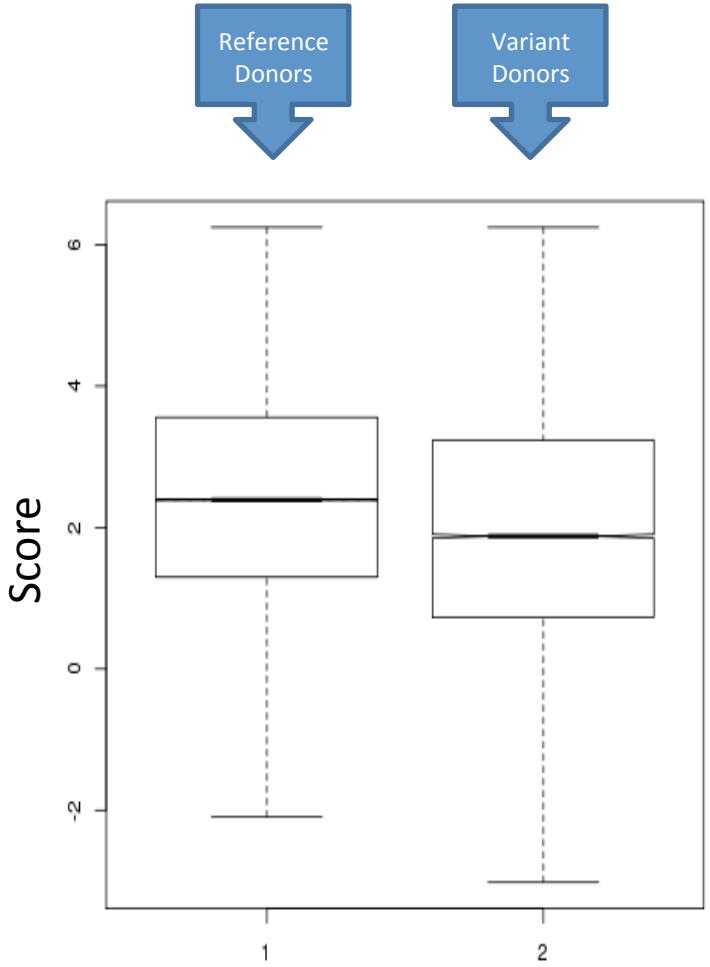


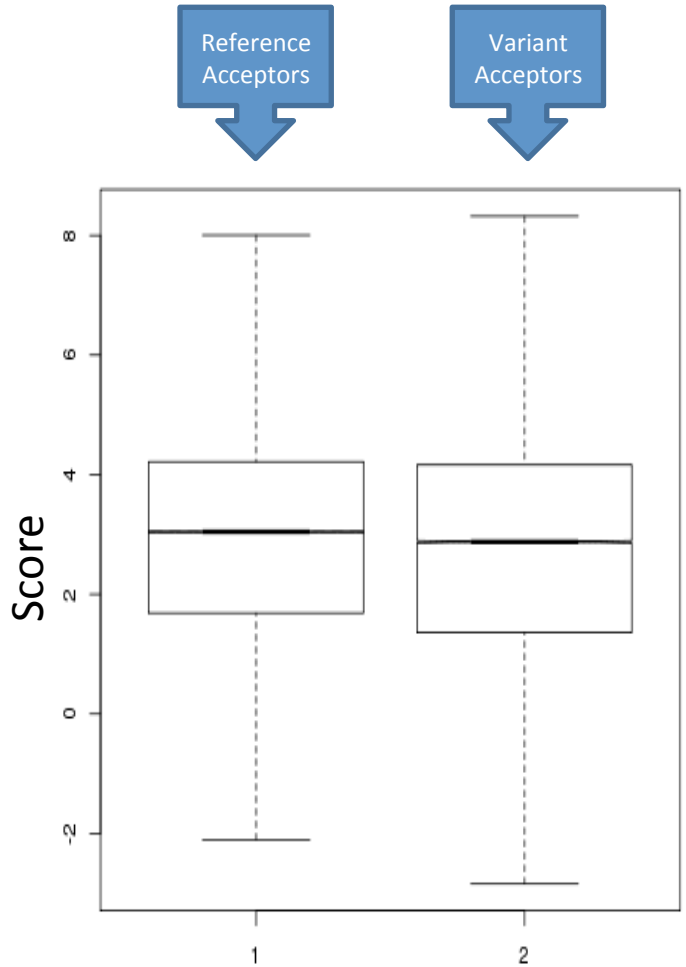
- the score distributions differ between donor and acceptor sites,

as a natural consequence of differences in the number of informative positions

- on average, sites score slightly better in the reference

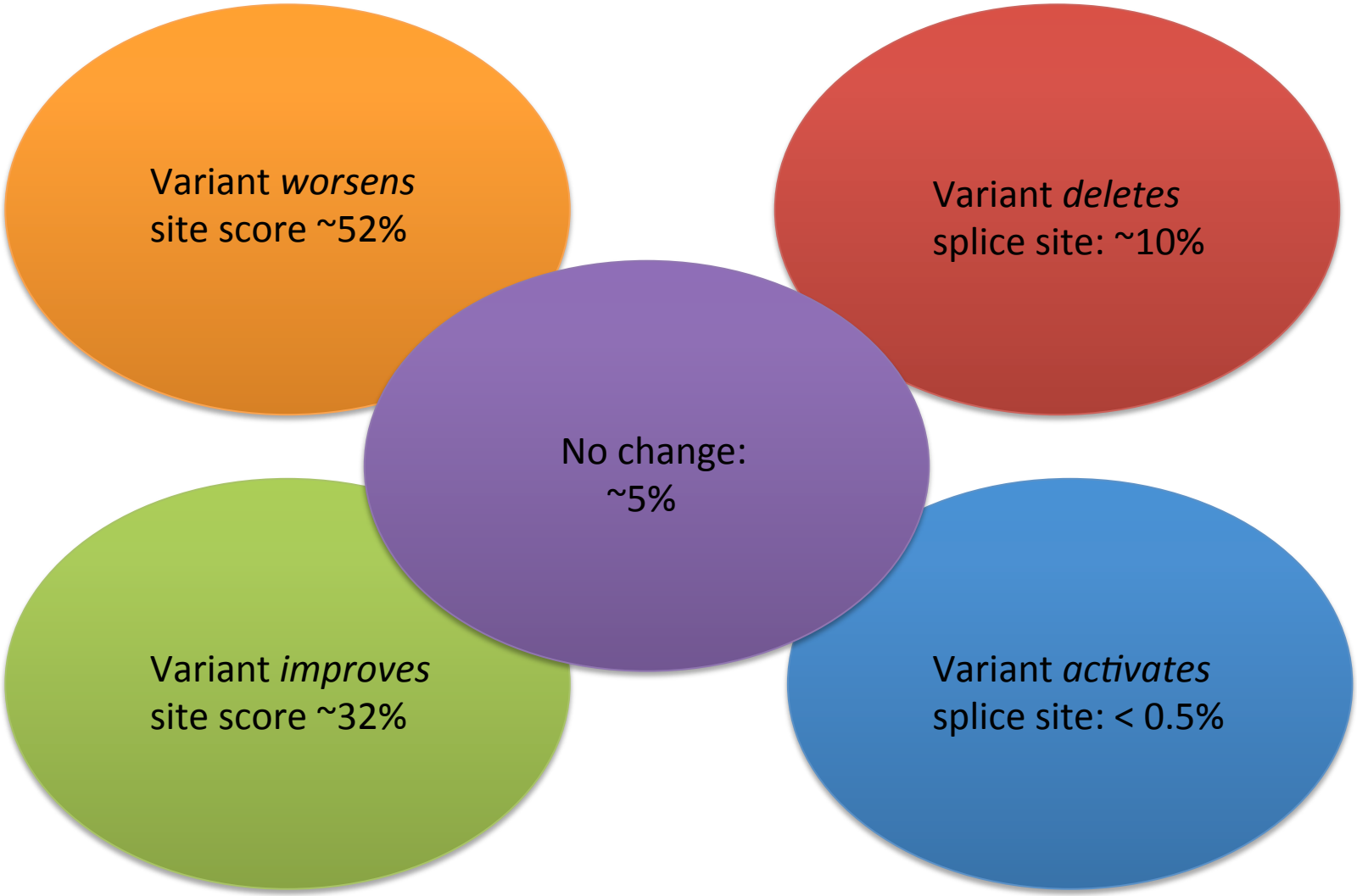
Score Distributions: reference vs. variant



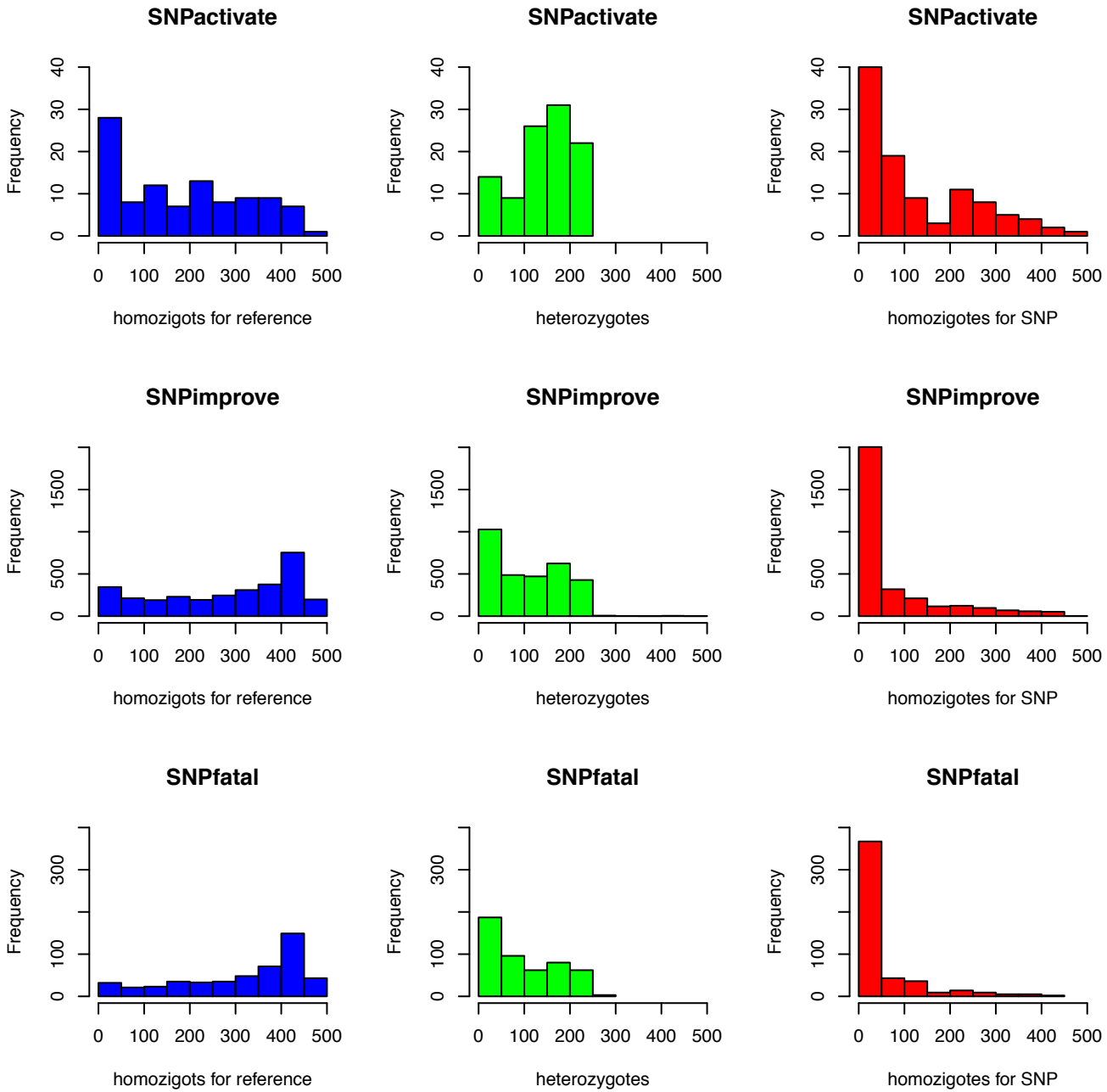


Ks.test significant

Splice Site Variants grouped by Predicted Effect

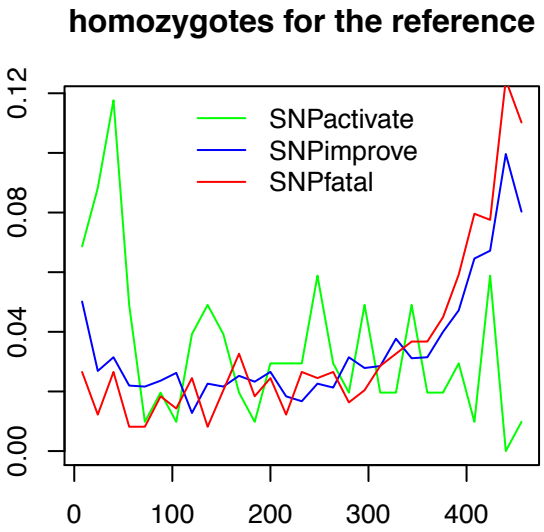


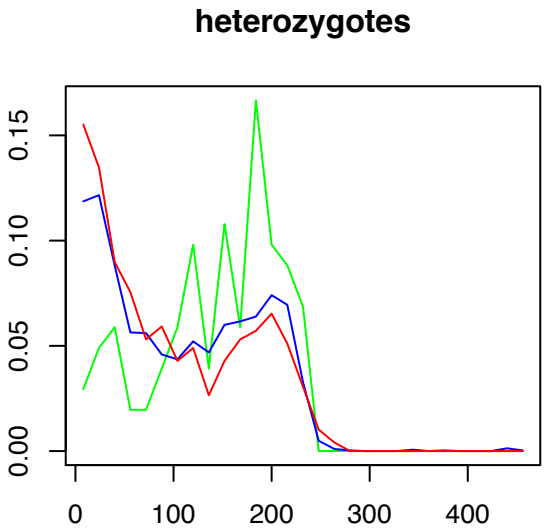
Splicesite Variants and Allele Frequencies



- variants with no non-hz Reference individuals Removed
- still a strong tendency for (homozyg) reference all.
- Overall similar trends, but subtantial differences between variant classes

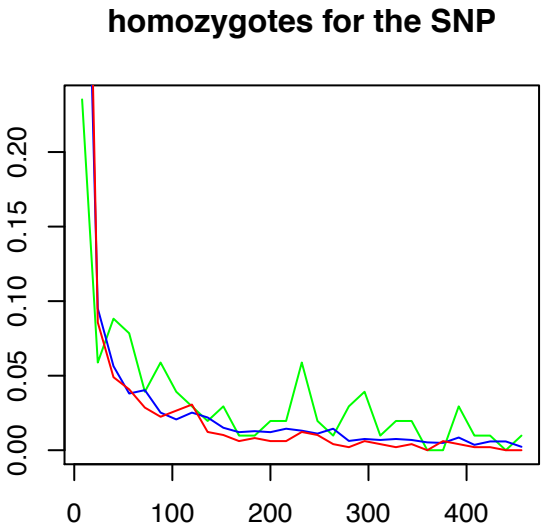
Splicesite Variants and Allele Frequencies





- gradual shift of hz-ref bias for deleterious > improving > activating SNPs

- heterozygotes of deleterious variants are less represented than improving or activating SNPs
- clear signal for activating variants
- hunchback at ~ half the population size



- activating SNP homozygotes are relatively more spread in the population

Splice Site Variants

employing the Gencode v12 transcriptome, and the Loss-of-Function variant annotations (v2) we (together with TL) propose the following VCF-specification for splice site variants

CHR	POS	ID	REF	ALT	QUAL	FILT	INFO	FORMAT	I1	I2	…
20	14370							GT	0|1	0|0	…
…	…							…			

splice site ID
(AStalavista)
`-14370^20`

genomic sequence hg19
(reference), soft-masked
lower-case = low complexity
`TTGTACGTG`

variant sequences
(comma-separated)
upper-case = variant positions
`ttgtaGgtg,ttgtCcgtg,…`

score of reference site
`1.5277311`

PASS or q-1000

MOD=ALT/CON;
ALT1=<cs list of SNP-IDs>;
VAR_SCORES=<cs list of var scores>
SNPS=<cs list of SNP-IDs>

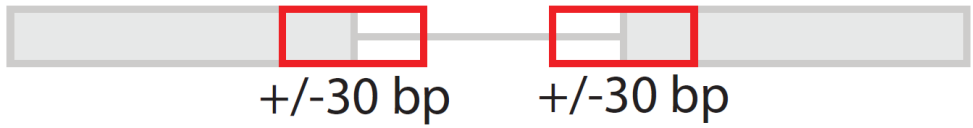
Variant-nr (not SNP!)
for each allele
(phased, missing if 1 SNP ".")

Intron Variations (Matthias)

1)

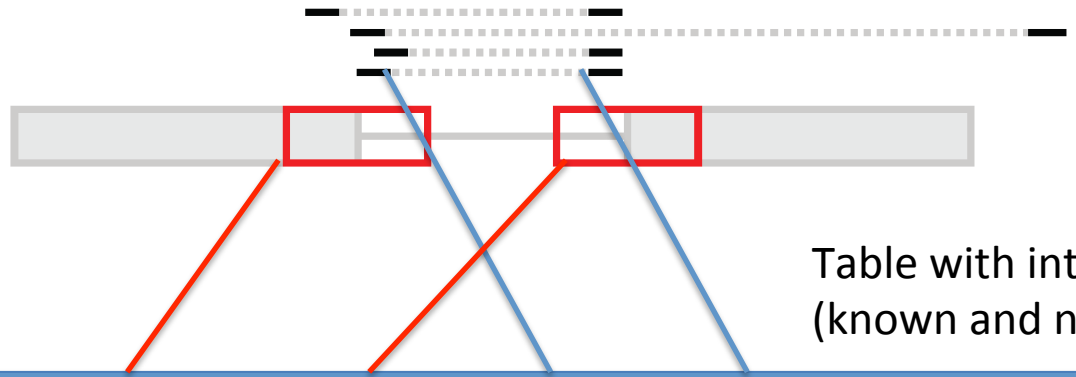
In the Gencode reference

2)



in the neighborhood of known splice sites

3)

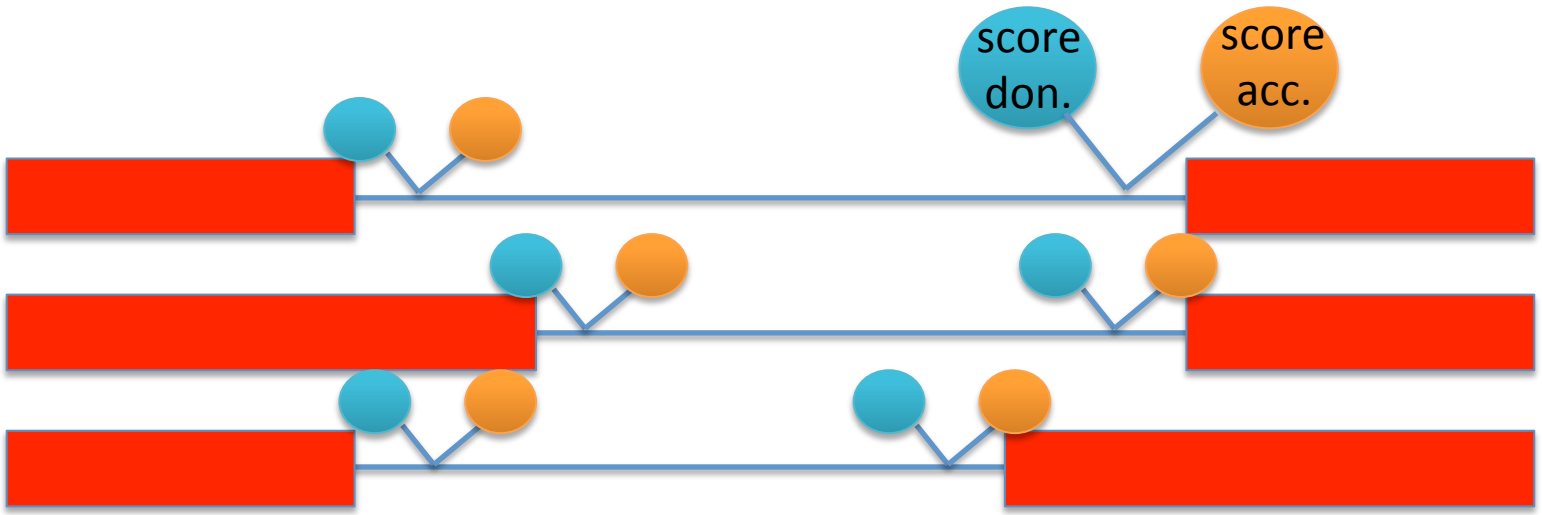


use split-mappings with novel sites that align to known sites across intron (filters applied)

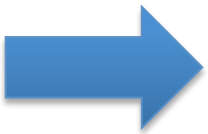
Table with intron candidates (known and novel)

Chr	qstart	qend	vstart	vend	HG00105	NA20828	...
1	123	345	128	281	13	0	...
...	...	...	...	...	...	...	...

Stranding the Introns



For an intron complex, compare sum of scores for the one or the other directionality.



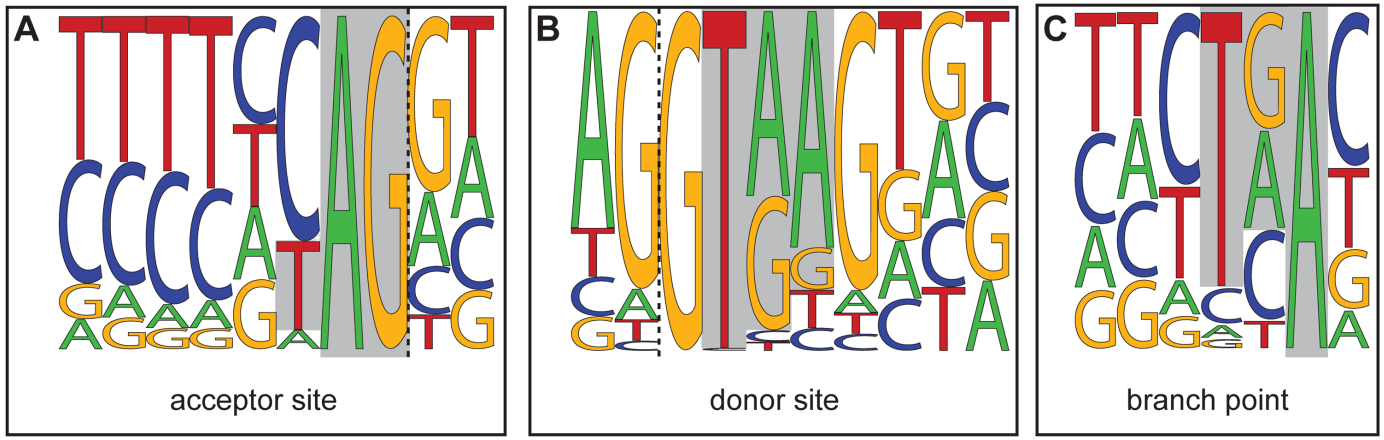
decide on directionality jointly for an “intron complex”, i.e. a group of splice sites connected by introns

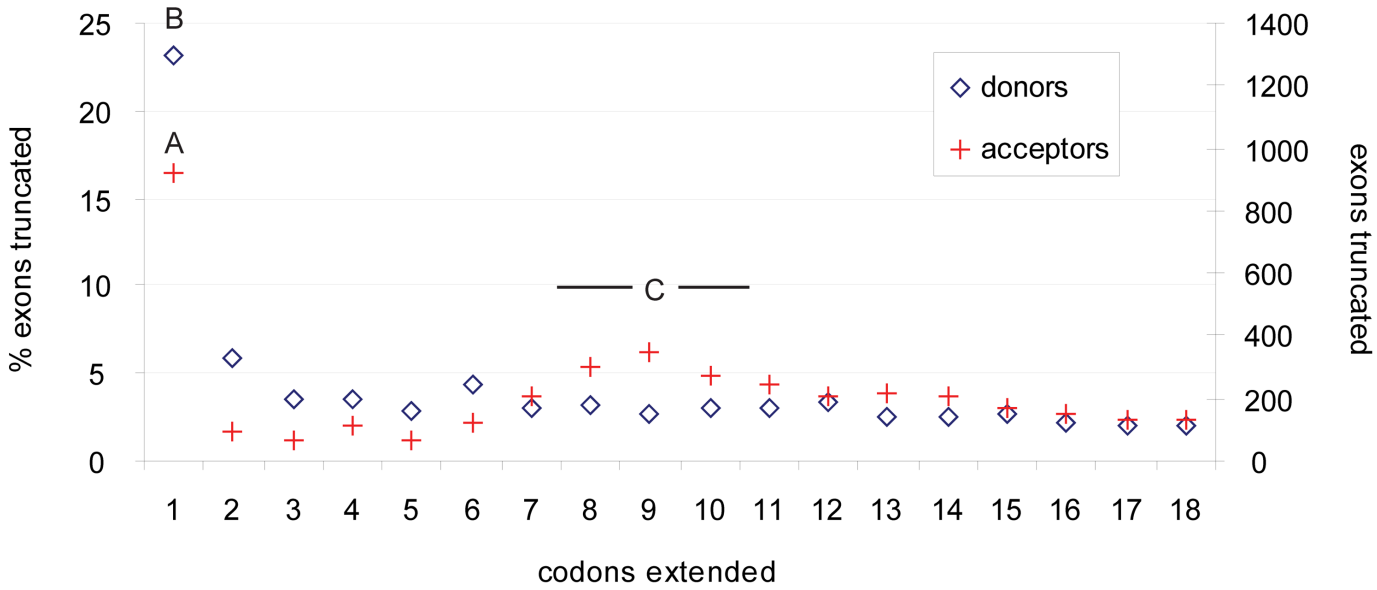
372k novel donor sites

12.5% more donor sites

418k novel acceptor sites

The Stop Codon Bias of Splice Sites





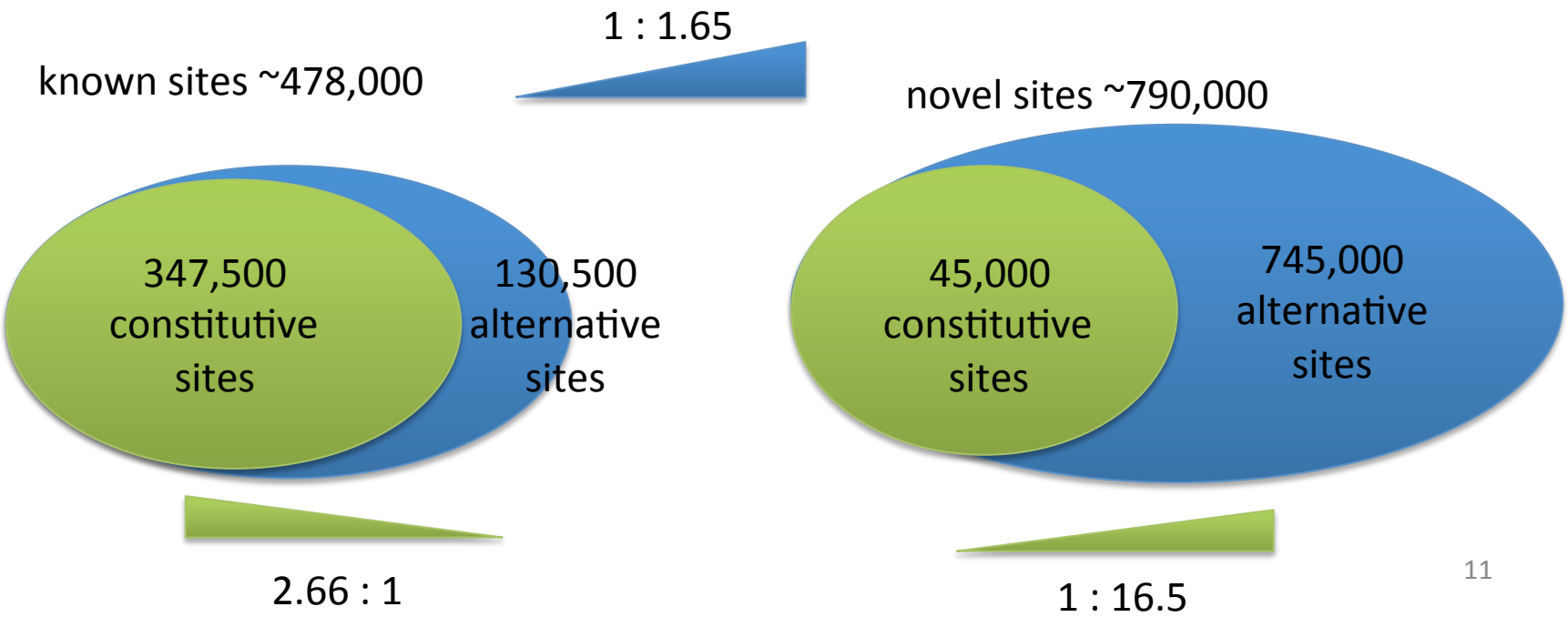
- Splice/branch site consensus harbours potential stop codons
- when extending the annotated frame of exons into the adjacent intron
- 7%-8% more exons truncated when extending CDS in donor sequences
- additional biases from differences in Information content / # of informative bases

[Sammeth, Foissac, Guigó 2008]

What adds *de novo* split-mapping to our knowledge about new splice sites of known introns ?

classify splice sites

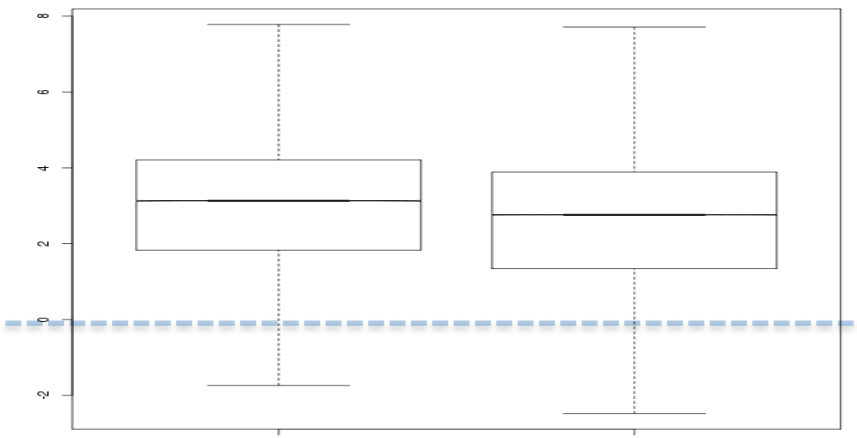
- according to the Gencode reference into *known/novel*
- according to other transcripts into *constitutive* / *alternative* (see earlier)



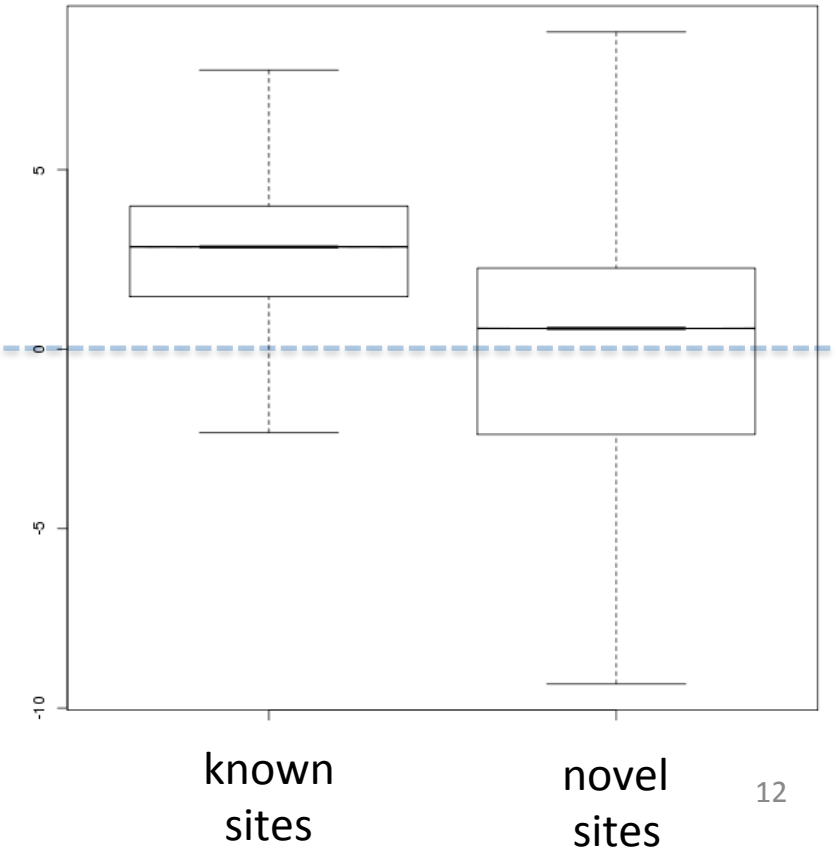
Splice Site Scores of Novel Sites

1. most of the novel splice sites are *alternative* by finding procedure (as seen before)
2. not yet annotated sites should score lower as they are likely to happen more rarely (i.e., they are likely to have escaped sufficient previous observations to be annotated)

score distribution known sites (cf. before)



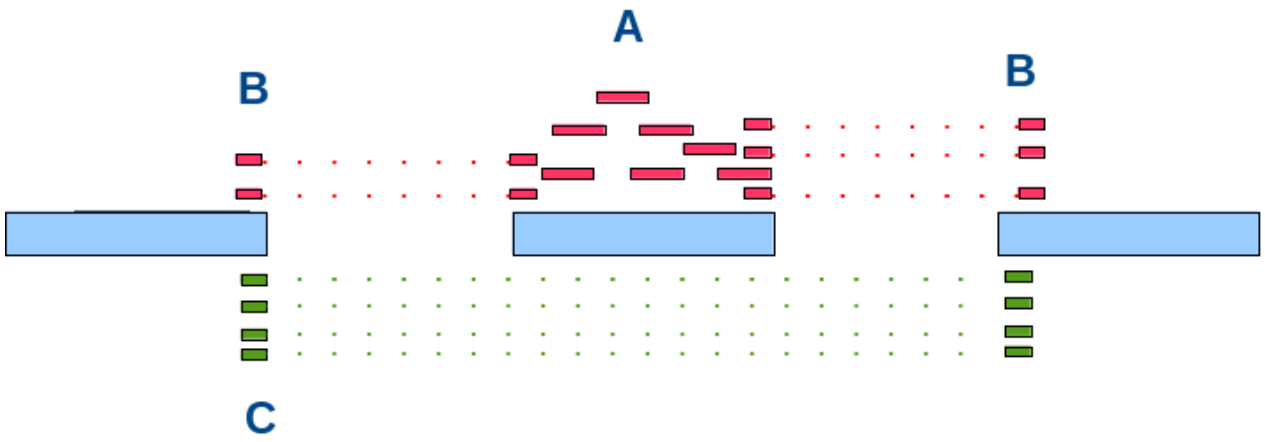
score distribution known vs novel sites



novel sites really exhibit significantly lower scores, even lower than the one of alternative sites

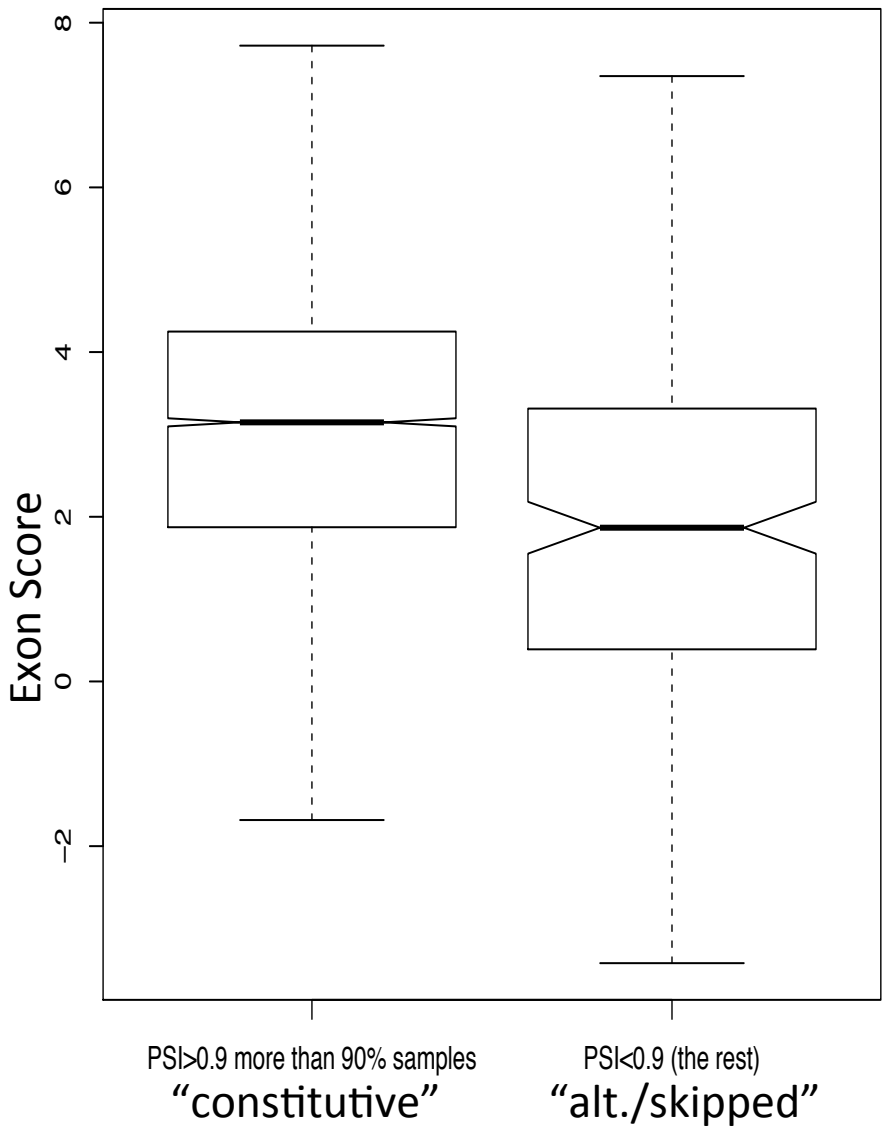
PSI calculation (Pedro)

- PSI = # inclusion_reads / (# inclusion_reads + # exclusion_reads) or PSI = A + B / (A +B + C)



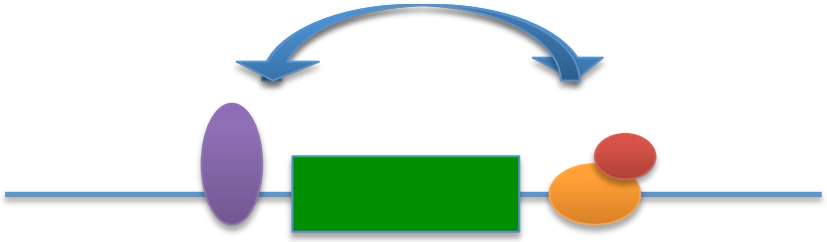
- **a** number of reads that map in the exon body (GD667.ExonQuantCount.txt) and **b** and **c** from flux files.
- A PSI value of **1** means that the exons is fully included and the other extreme a value of **0** means that the exon is not included.

Constitutive exons: PSI-Scores and Splice Site Scores



- constitutive Splice Sites show higher Splice Site Scores (see before)

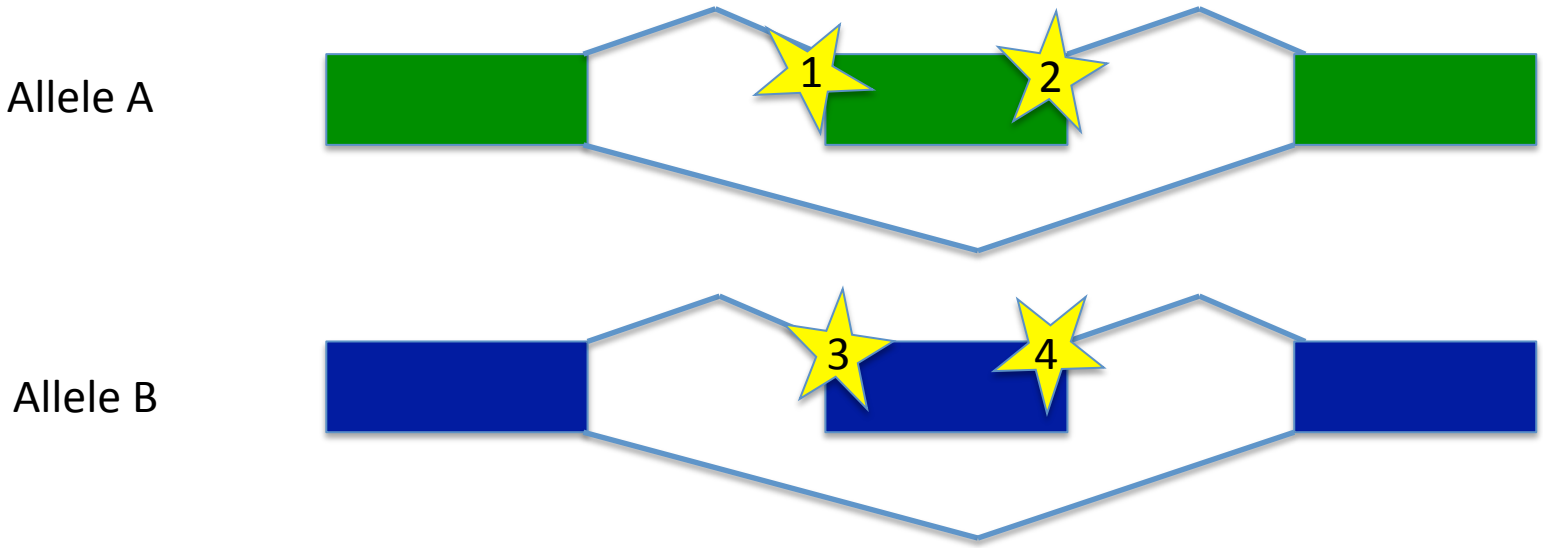
- Exon Scores from Splice Site Scores: assume *exon definition*, sum log-likelihoods



*score(*exon*)* = *score(*donor*)* + *score(*acceptor*)*

Exon splicing scores of "*constitutive exons*" (PSI>0.9 in >90% of the individuals) are *higher than* of splicing scores of *alternative* and *skipped exons* (PSI> 0.9 in < 90% pop.)

PSI Scores Across Different Alleles

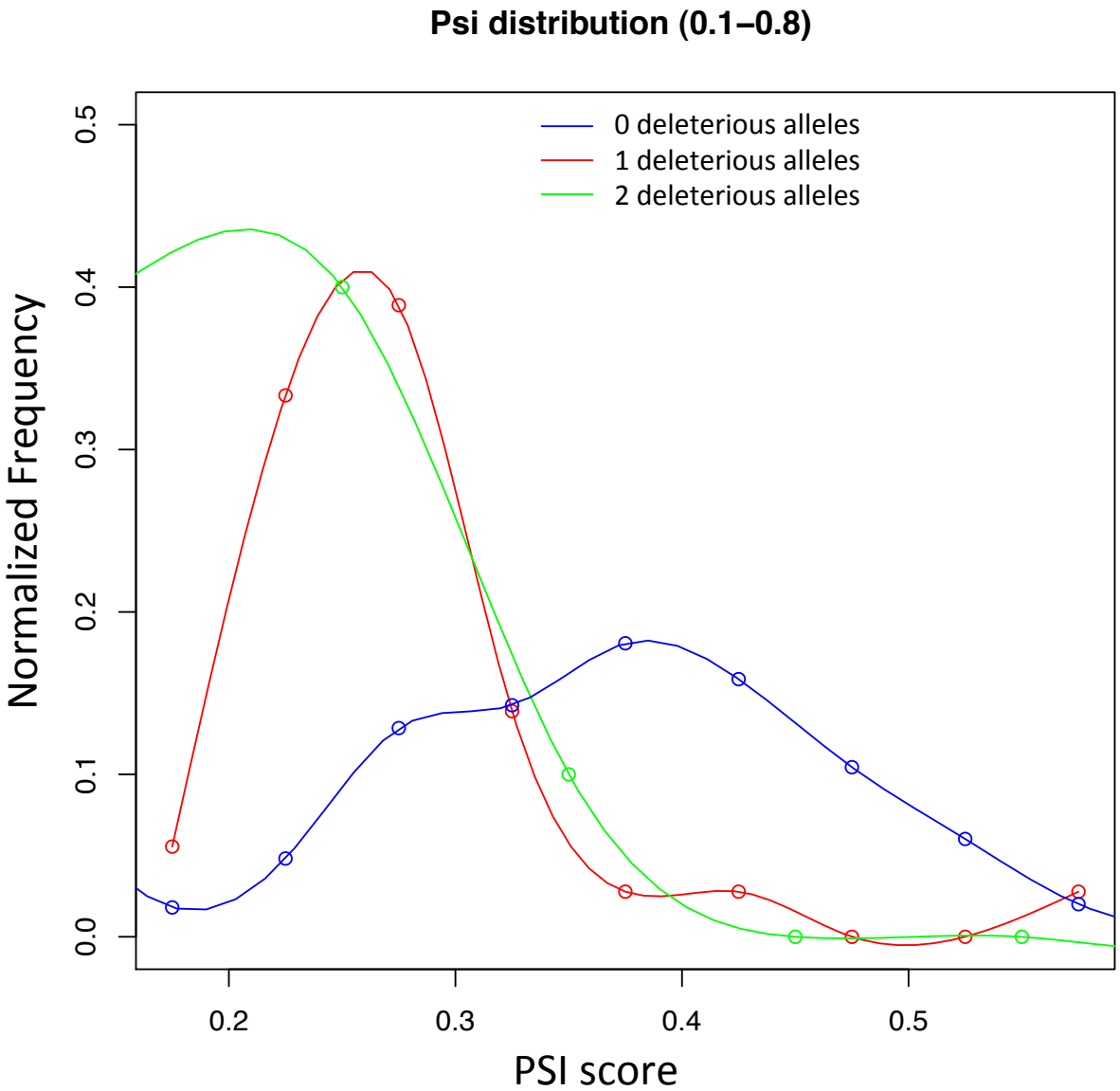


- hypothesis: splice site variants that are predicted to be deleterious by splice site scores should be reflected by PSI score inclusion levels

- each diploid individual can <u>in theory have *between 0 and 4 variants*</u> affecting an exon: maximally 2 delet. alleles at the splice acceptor, and also max. 2 delet. alleles at the donor

- in practice only up to 2 deleterious variants are detected for alternative exons (0.2 < PSI < 0.8):

Deleterious SNPs lower PSI score



- exons with deleterious SNPs in flanking splicesites
- exons that are alternatively spliced in this tissue: 0.2 < PSI < 0.8
- classify individuals by the nr of deleterious alleles
  blue = 0 SNP-alleles
  red = 1 delet. SNP
  green = 2 delet. SNPs
  (3 and 4 numerically too low)

Higher number of delet. SNPs shifts the histogram to lower PSI-scores

To Do

- provide final resources: vcf file, comprise also novel sites, add geneIDs…

- include minor spliceosome in the anlysis?

- coordinate splicing paper