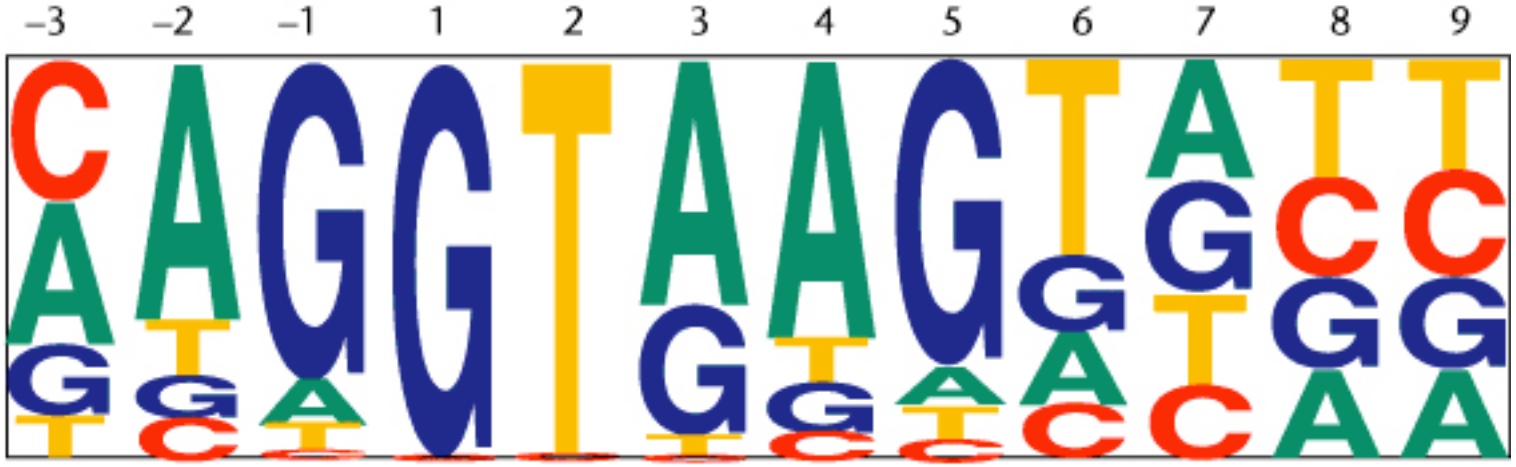
Splice Site Variants

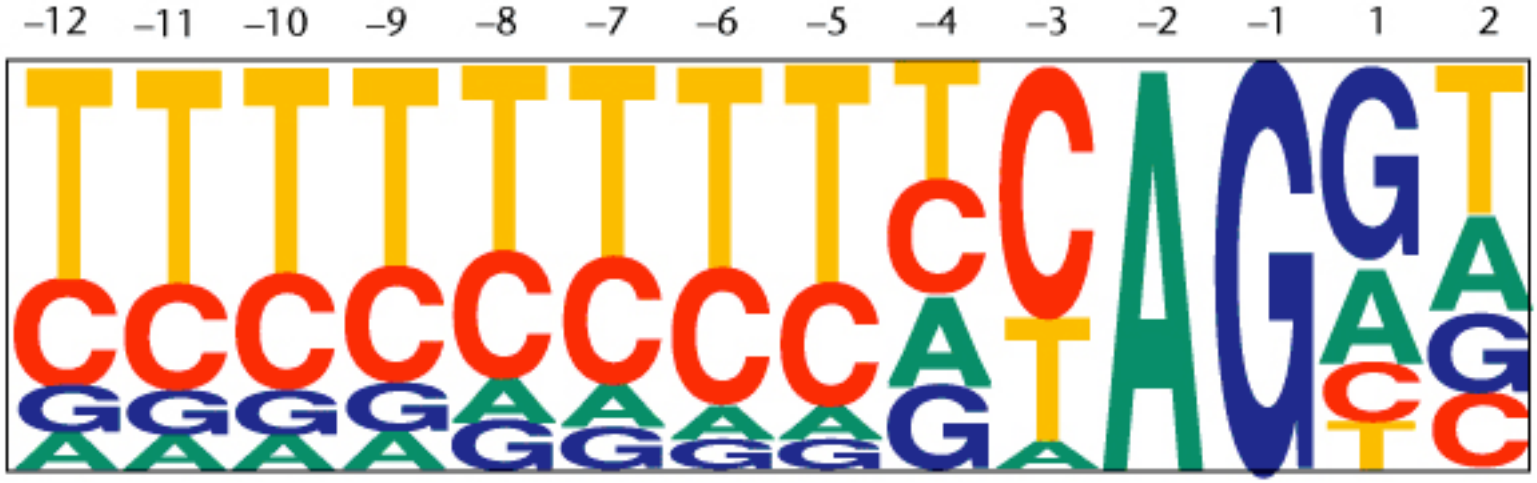
Michael Sammeth
Anna Esteve
CNAG-CRG Barcelona

Splice Site Motifs

Splice Donor Sites



Splice Acceptor Sites



Markov Chains

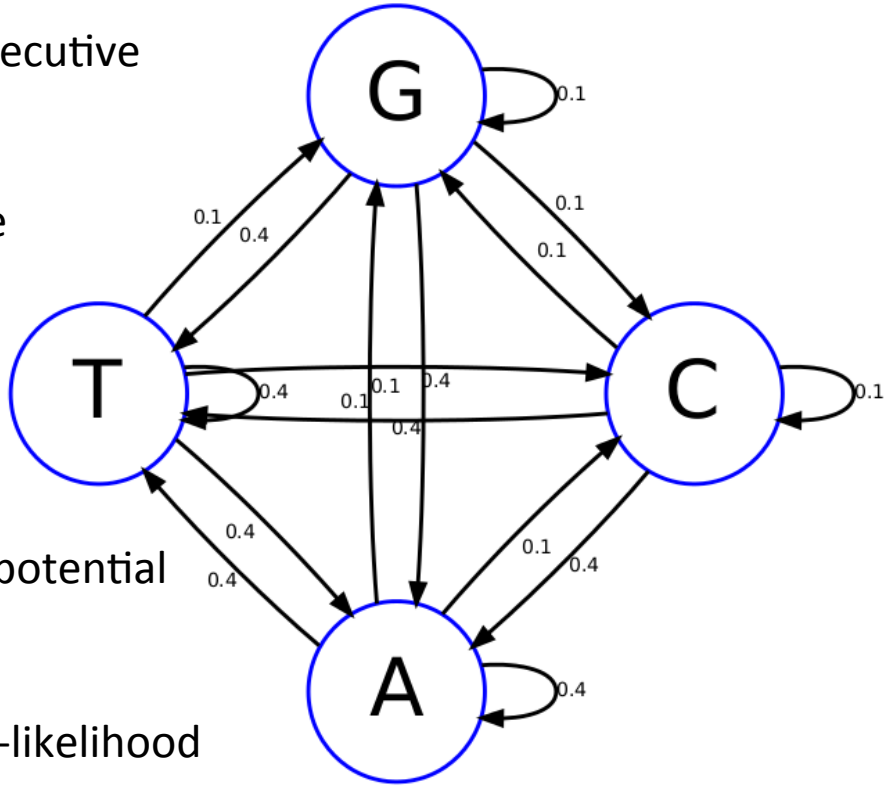
- capture *tansition probabilities* between consecutive substrings in a motif

- exist of different *orders*, i.e. the length of the substring; single base transitions are of order 0

- are *trained* with data available for TP / TN splice sites

- subsequently can be applied for *scoring* the potential of a motif to act as a splice site

- here, the score of a site is the sum of the log-likelihood transition probabilities (geneID)



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,…`

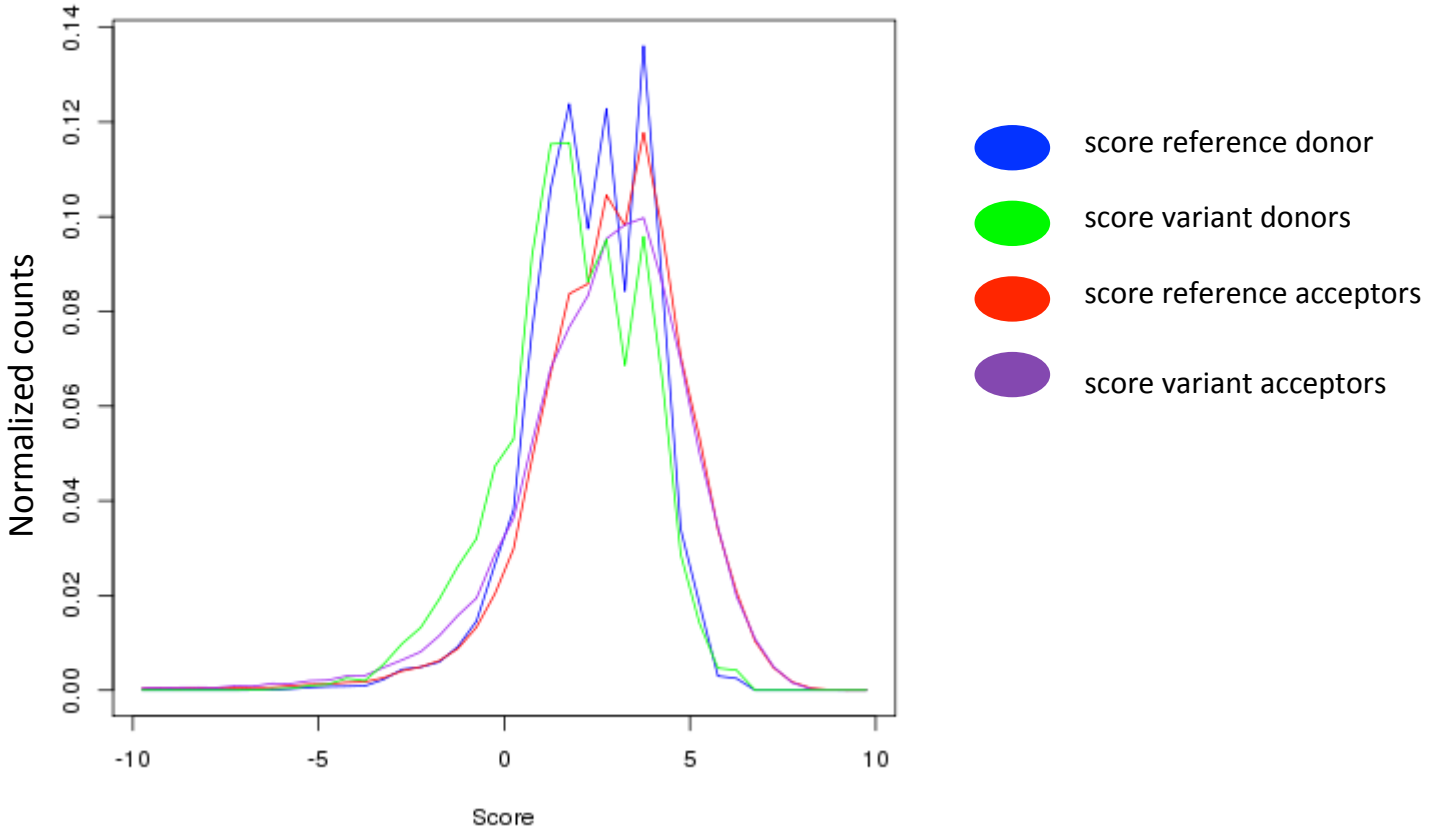
score of reference site
`1.5277311`

PASS or q-1000

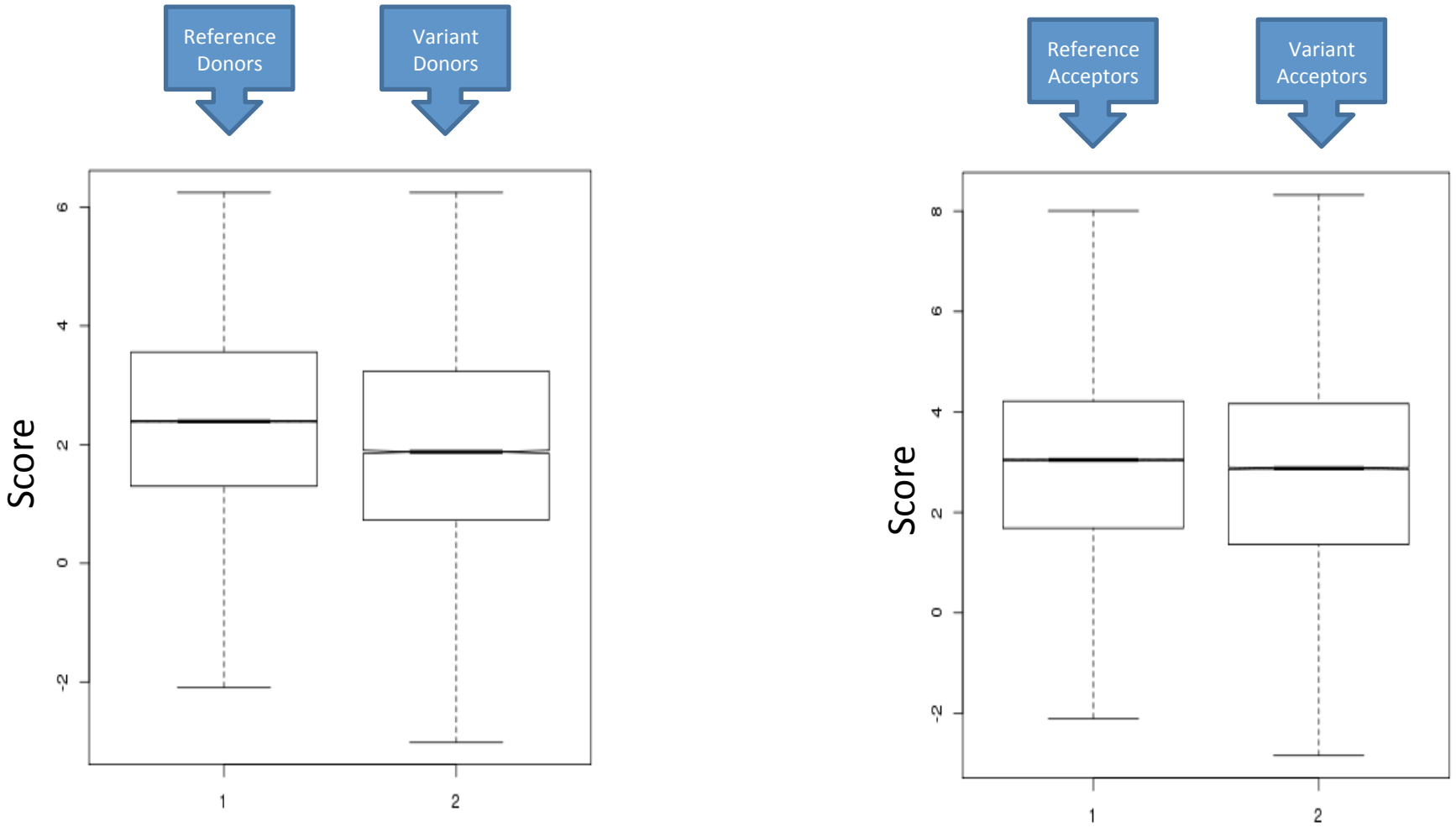
Variant-nr (not SNP!)
for each allele
(phased, missing if 1 SNP ".")

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

Gencode v12 Score Distributions



Score Distributions

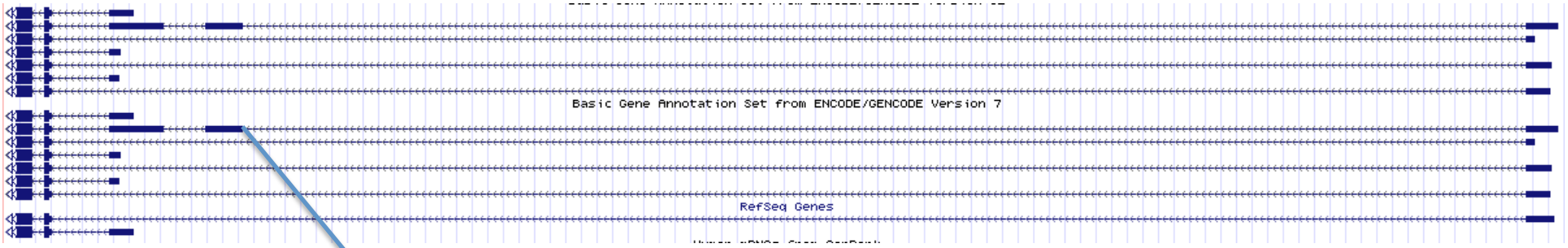


Ks.test significant

Numerical Analysis

- **Var_score** < **Ref_score**: 62.3%
  - Destruction: 10.2%
- **Var_score** > **Ref_score**: 32.1%
  - Activation: 0.44%
- Var_score = Ref_score: 5.6%

Example: 5’-UTR lengths



score reference: PASS (2.8)
score variant: q-1000

Population	
11	0|0
69	0|1
79	1|0
306	1|1