# Variant annotation

Approaches and resources

# Annotate variant statistics

How to compute and annotate variant statistics and flags

# Useful statistics to consider for variant annotation


## *Possible useful metrics to annotate in a multi-sample VCF for QC*

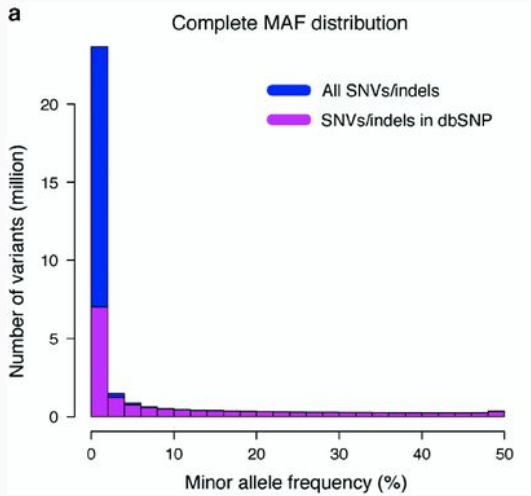
```
INFO/AC          Number:A  Type:Integer  |  Allele count in genotypes
INFO/AN          Number:1  Type:Integer  |  Total number of alleles in called genotypes
INFO/AF          Number:A  Type:Float    |  Allele frequency (AC/AN)
INFO/MAF         Number:A  Type:Float    |  Minor Allele frequency
INFO/ExcHet      Number:A  Type:Float    |  Test excess heterozygosity; 1=good, 0=bad
INFO/END         Number:1  Type:Integer  |  End position of the variant
INFO/F_MISSING   Number:1  Type:Float    |  Fraction of missing genotypes
INFO/HWE         Number:A  Type:Float    |  HWE test (PMID:15789306); 1=good, 0=bad
INFO/NS          Number:1  Type:Integer  |  Number of samples with data
INFO/TYPE        Number:.  Type:String   |  The record type (REF,SNP,MNP,INDEL,etc)
INFO/MEDIAN_DP   Number:1  Type:Float    |  The median depth observed across all genotypes
INFO/MEDIAN_GQ   Number:1  Type:Float    |  The median GQ observed across all genotypes
FORMAT/VAF       Number:A  Type:Float    |  The fraction of reads with the alternate allele
                                             requires FORMAT/AD
```

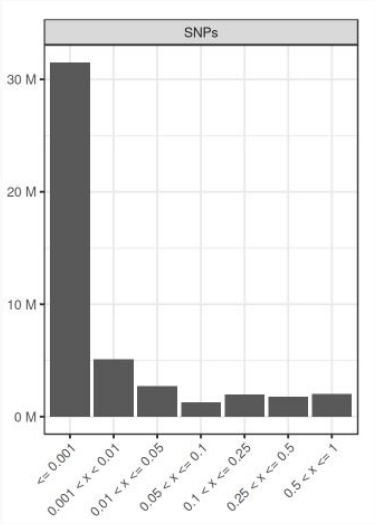
# Variant metrics - AF vs MAF

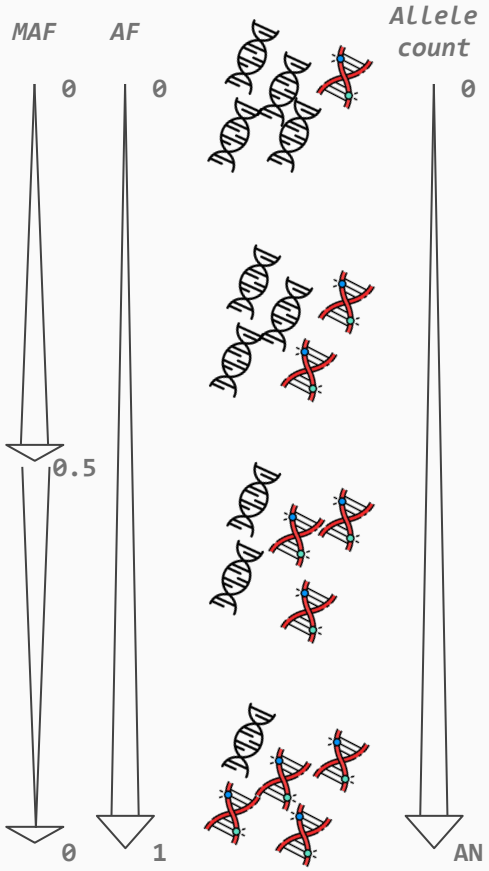
| CHR | POS | REF | ALT |
|---|---|---|---|
| chr1 | 125634728 | A | G |


**MAF**
based on the count of the allele representing less than 50% of total alleles (AN)

**AF**
based only on the ALT allele count



# BCFTOOLS +fill-tags plugin

```
Usage: bcftools +fill-tags [General Options] -- [Plugin Options]
```

*Basic usage*

```
bcftools +fill-tags input_file.vcf.gz -- -t AC,AN,AF,MAF
```

*Simply provide the desired tag(s) to annotate your VCF (-l to see a list of available tags)*

## Computation of custom tags

```
bcftools +fill-tags in.bcf -- -t 'DP=sum(FORMAT/DP)'
```

*Any custom value can be computed using the format 'TAG_NAME=formula'*

*This allows to combine information from multiple tags to generate a new value*

Many functions implemented like:

- *N_PASS*: number of samples passing a condition
- *F_PASS*: fraction of samples passing a condition
- *MEDIAN, MEAN*
- *SUM, COUNT*

See documentation on expression for a full list

See the official tool documentation here

# BCFTOOLS +fill-tags plugin

## Computation of per-group statistics

```
bcftools +fill-tags in.bcf -- -t AC,AN,AF -S groups.tsv
```

*Tags can be computed per group by providing a grouping table*

*Example grouping file*

```
Sample1     Group1
Sample2     Group1,Group2
Sample3     Group2
```

*Resulting INFO tags*

```
AC_Group1
AC_Group2
AN_Group1
AN_Group2
…
```

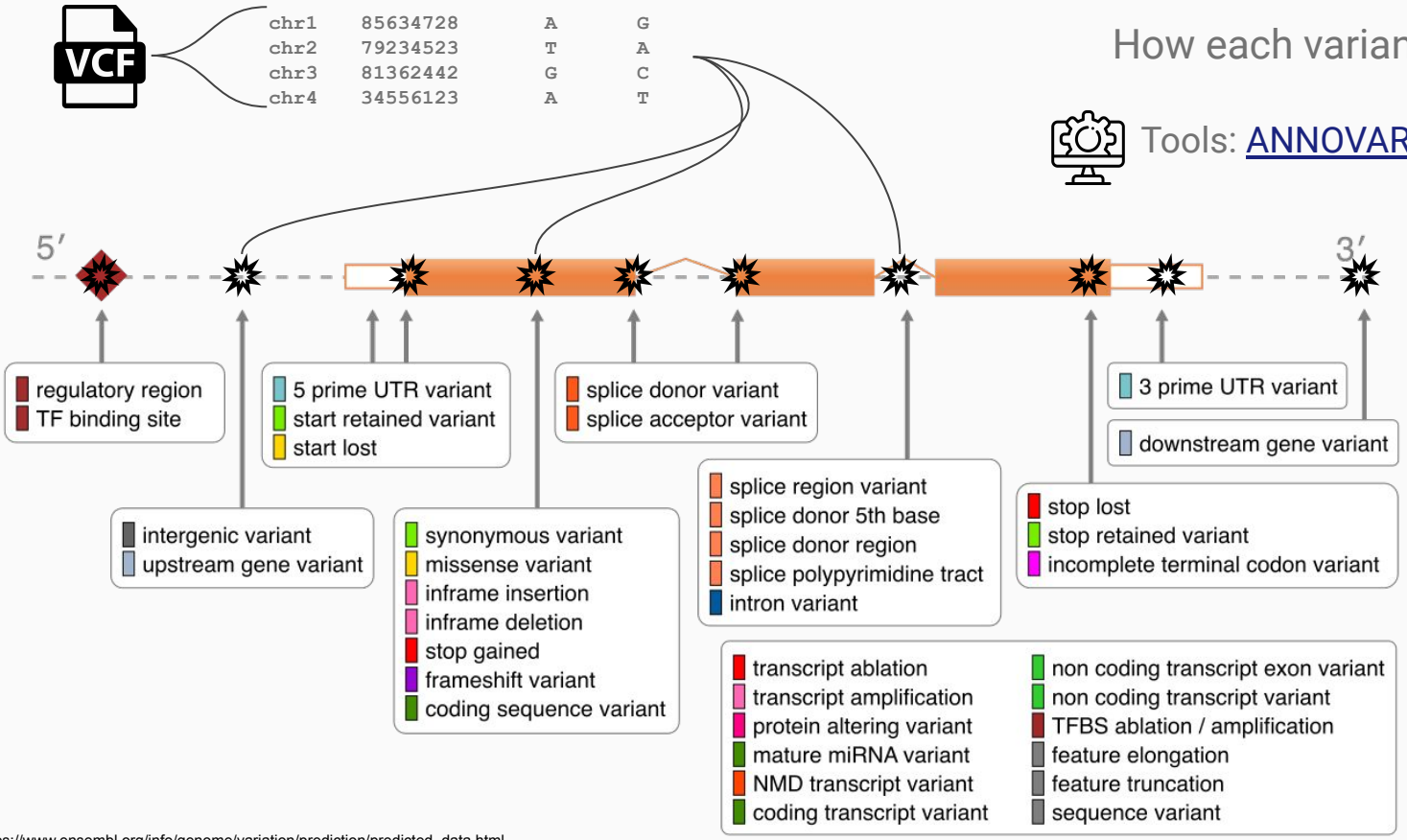
Useful to compute population specific values or specific values for cases and controls

See the official tool documentation here

# Consequences on gene

Understand the impact of variant on known genes / transcripts

# Gene consequences




How each variant impact on genes?

Tools: ANNOVAR, VeP, SnpEff, bcftools

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

# Gene consequences - available gene definition sources

| Database | Present Release | Number of transcripts | Type of data |
|---|---|---|---|
| **RefSeq** | 227 | 207,289 | Computational prediction, experimental validation |
| **Ensembl** | 113 | 387,944 | Computational prediction, experimental validation, large RNA-Seq, GENCODE Basic subset |
| **GENCODE** | 24 | Basic: 158,338<br>Complete: 385,659 | Manual and automated annotation from EST, RNA-seq, curated projects |

Transcript structure information define coordinates for exons, introns, and UTRs and influence how gene consequences are defined for a given genetic variant.

# Gene consequences - The GTF / GFF formats

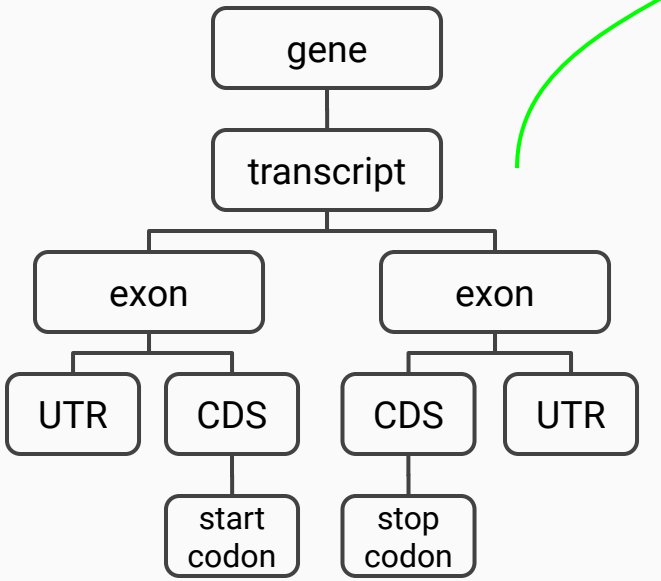
Tab-separated 9 columns:

```
chr  pos  feature  score  strand  phase  attributes
```

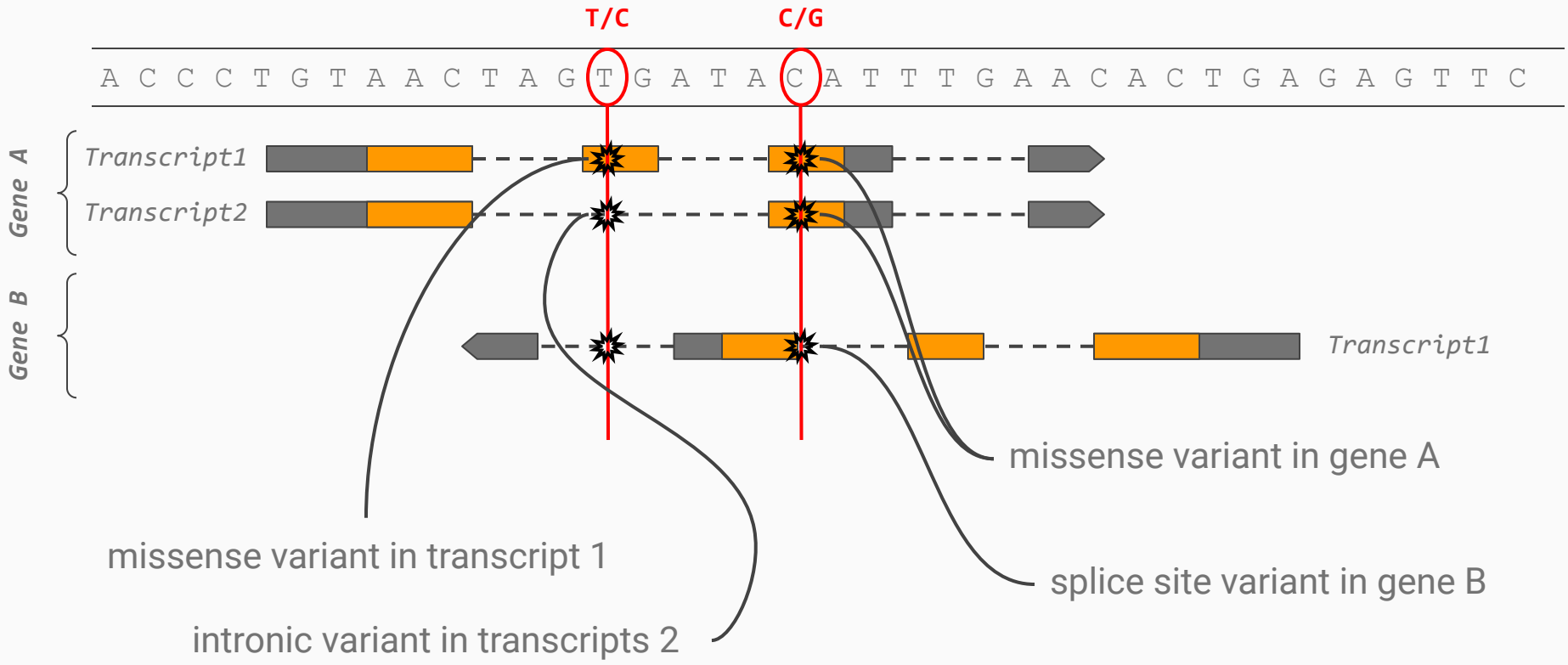
Genomic strand
plus or minus

Annotations
- gene_id
- transcript_id
- CDS_it
- exon_number
- gene_type
- transcript_type
- ...

Hierarchical structure

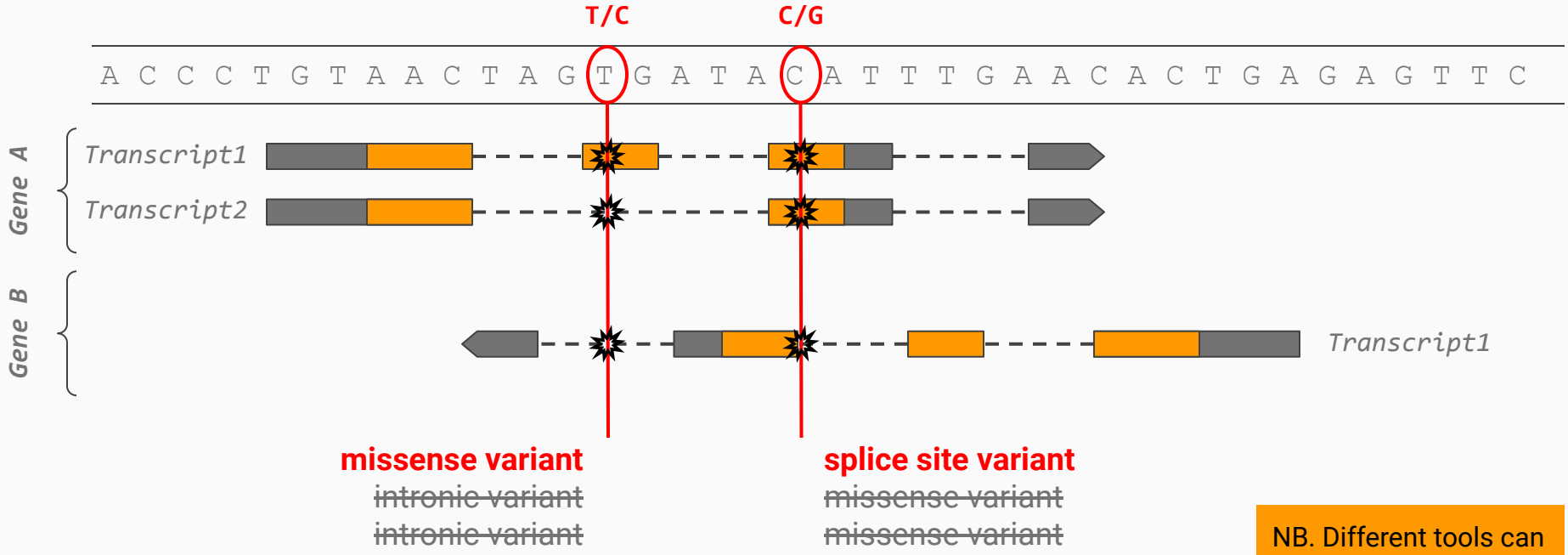


# Gene consequences - the problem of multiple impacts



# Gene consequences - most-severe approach

For each variant, pick the most severe consequence across all impacted transcripts

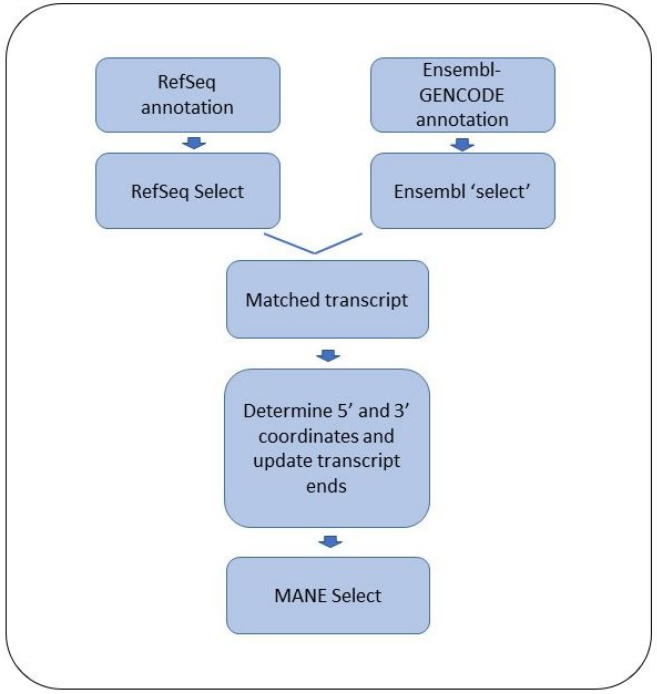



NB. Different tools can adopt slightly different rankings

# Gene consequences - Curated selection of transcripts

## MANE select

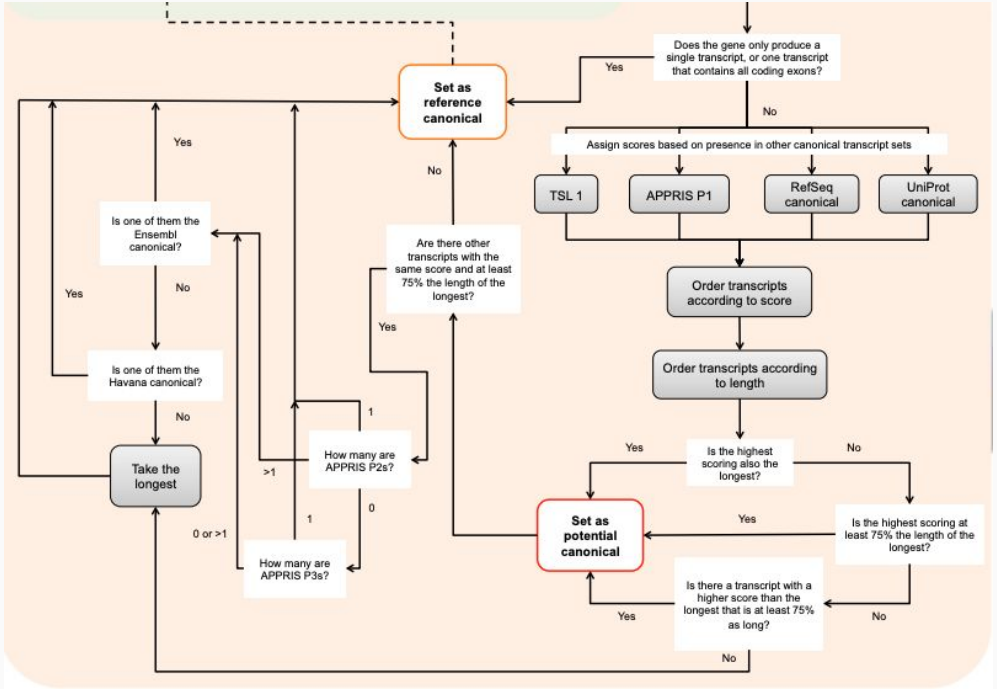
Very few curated transcripts per gene




https://www.ncbi.nlm.nih.gov/refseq/MANE/

## Ensembl canonical

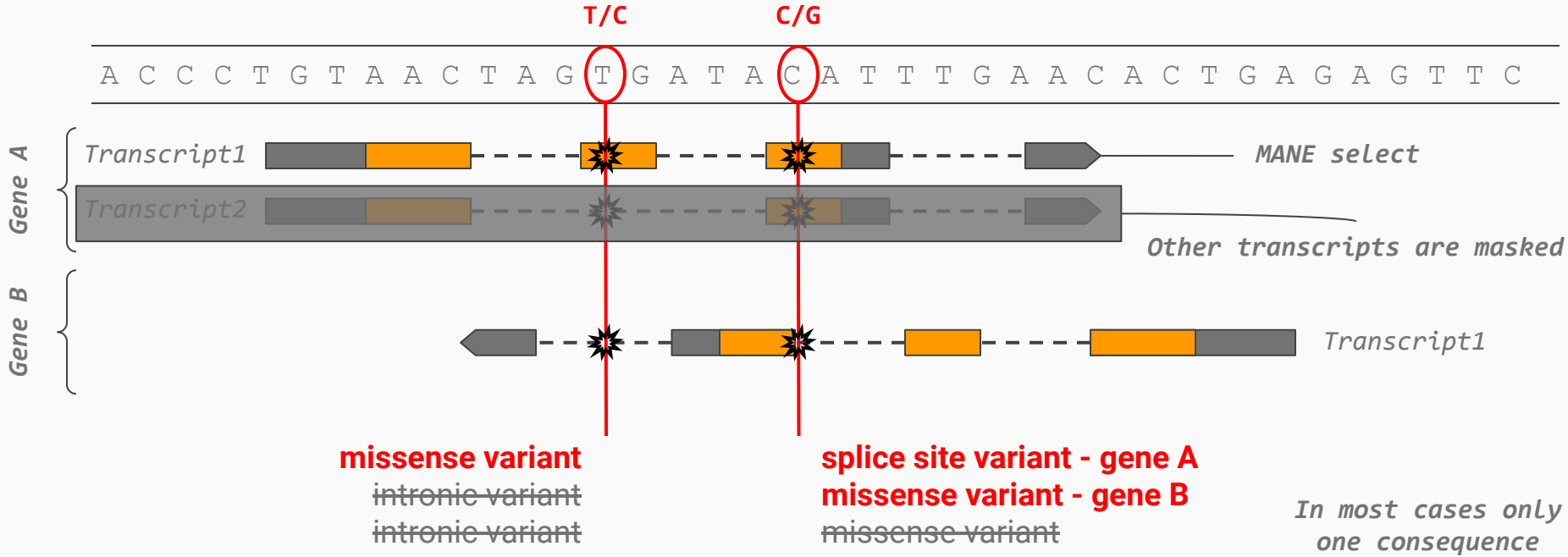
Only a single transcript per gene




DOI: 10.13140/RG.2.2.30201.26723

# Gene consequences - Filter by canonical or MANE transcripts

For each gene, a representative transcript is identified and tagged




*In most cases only one consequence per gene remains*

# Gene consequences - multi-nucleotide polymorphism (MNP)

*Multiple back-to-back SNVs*

MNP:

```
chr1    7653562  rsXXXXX  AT  CG  ...  0|1
```

Atomized MNP:

```
chr1    7653562  rsXXXXX  A   C   ...  0|1
chr1    7653563  rsXXXXX  T   G   ...  0|1
```

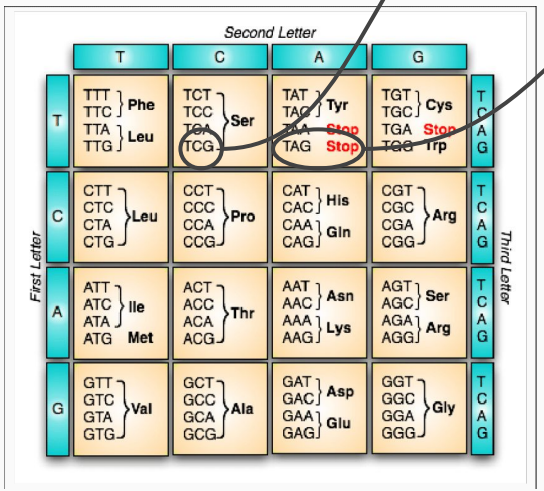
CTC TAT AAC

TCG

TCT

TAG

*Consequence properly understood looking at phase*



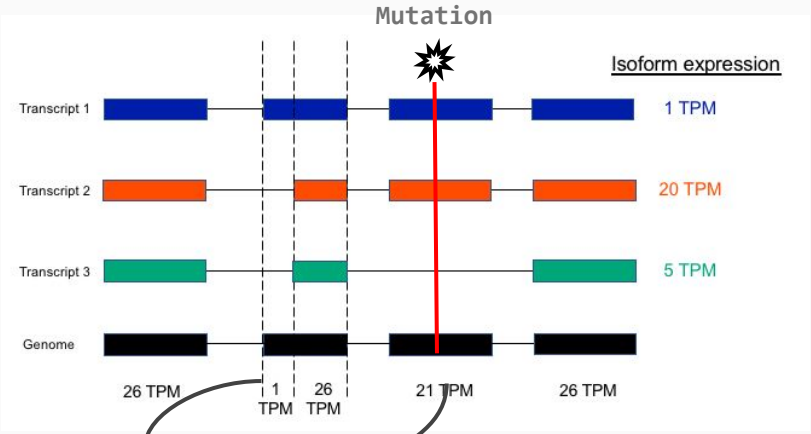

A wrong consequence may result from looking at atomized MNP

***Real consequence***
*missense Tyr>Ser*

***Atomized consequence***
*stop gained*

# Gene consequences - evaluate an impact in the context of transcript expression

Transcript expression varies across tissues and some transcripts may be spurious and rarely observed




pext=1/26=0.04

pext=21/26=0.80



Computational artifacts
"Fake" transcript always rarely detected

1 1 0

1 1 20

10 20 1

Tissue-relevant consequence
- intronic in brain/liver
- missense in uterus

**GnomAD pext**
For each base, in each tissue, divide the expression observed across transcripts by the total gene expression

*http://gnomad-sg.org/help/pext*

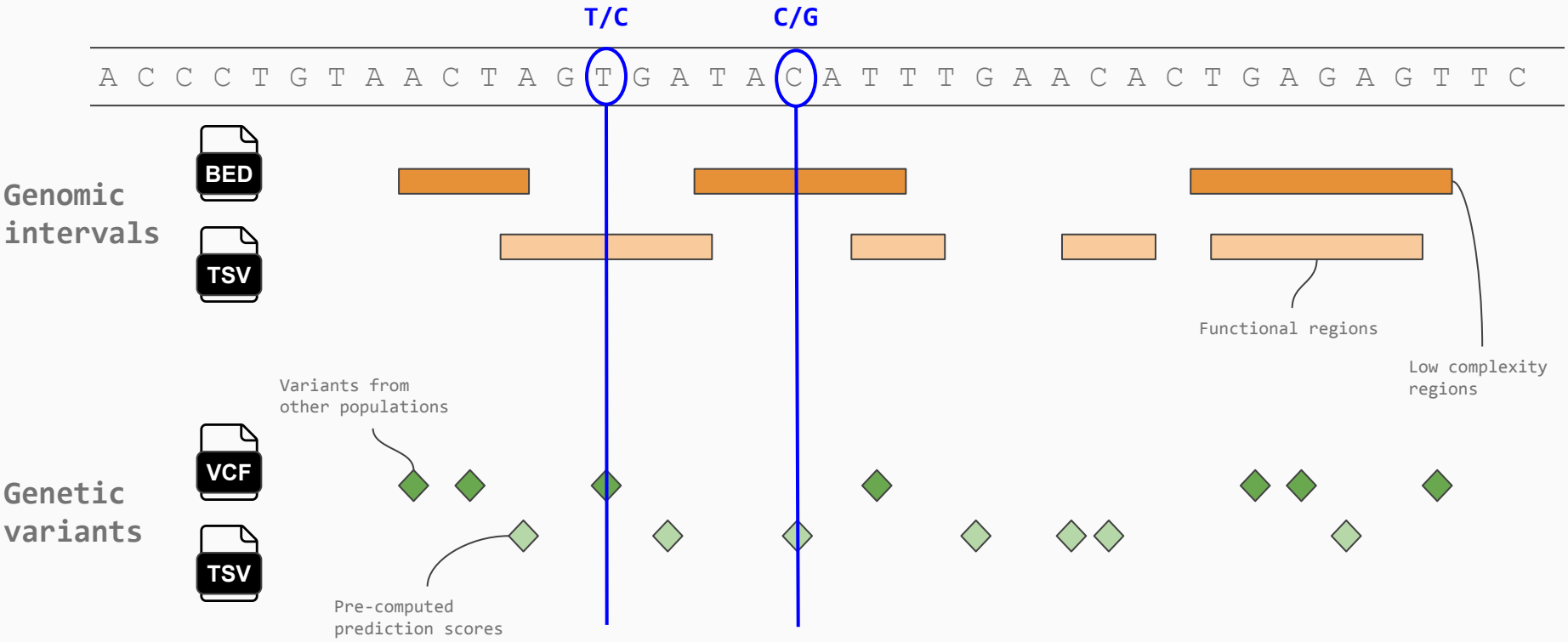
See the original pext paper (Cummings et al., Nature, 2020)

# Add custom annotations

How to annotate your variants with information from external sources

# Annotate a variant with information from external sources

Add useful annotations to variants based on the overlap or exact match with external annotations



# Annotate small variants - BCFTOOLS ANNOTATE

```
Usage: bcftools annotate [options] VCF/BCF
```

Accepted inputs


0-based coordinates
mandatory fields: CHROM, START, END


1-based coordinates
mandatory fields: CHROM, POS (or FROM/TO)


1-based coordinates

REF/ALT allele will be matched when REF and ALT columns are present

| Argument | Description |
|---|---|
| `-a` | *VCF file or tabix-indexed FILE with annotations* |
| `-c` | *Columns (TSV/BED) or INFO/FORMAT fields (VCF/BCF) from the annotation file to use in annotation* |
| `-I` | *Set ID column using a `bcftools query`-like expression* |
| `--min-overlap` | *Required overlap as a fraction of variant in the -a file (ANN), the VCF (:VCF), or reciprocal (ANN:VCF)* |
| `-x` | *List of annotations (e.g. ID,INFO/DP,FORMAT/DP,FILTER) to remove* |
| `-h` | *Lines which should be appended to the VCF header* |

See examples at http://samtools.github.io/bcftools/howtos/annotate.html

# Annotate small variants - VCFANNO

```
Usage: vcfanno config.toml input.vcf > annotated.vcf
```

Accepted inputs


0-based coordinates
mandatory fields: CHROM, START, END
Header ignored

REF/ALT allele will be matched when REF and ALT columns are present


1-based coordinates
mandatory fields: CHROM, POS (or FROM/TO)
Header mandatory


1-based coordinates

## config.toml

```
[[annotation]]
file="my_regions.bed.gz"
columns = [4,5]
ops=["flag","uniq"]
names=["ISREGION","REGION_NAME"]

[[annotation]]
file="ReMM_score.tsv.gz"
columns = [3]
ops=["max"]
names=["ReMM"]

[[annotation]]
file="gnomad.vcf.gz"
fields = ["AF","NFE_AF","AFR_AF"]
ops=["max", "max", "max"]
names=["gnomad_af","gnomad_eur_af""gnomad_afr_af"]
```

Indexes for columns to use in annotation

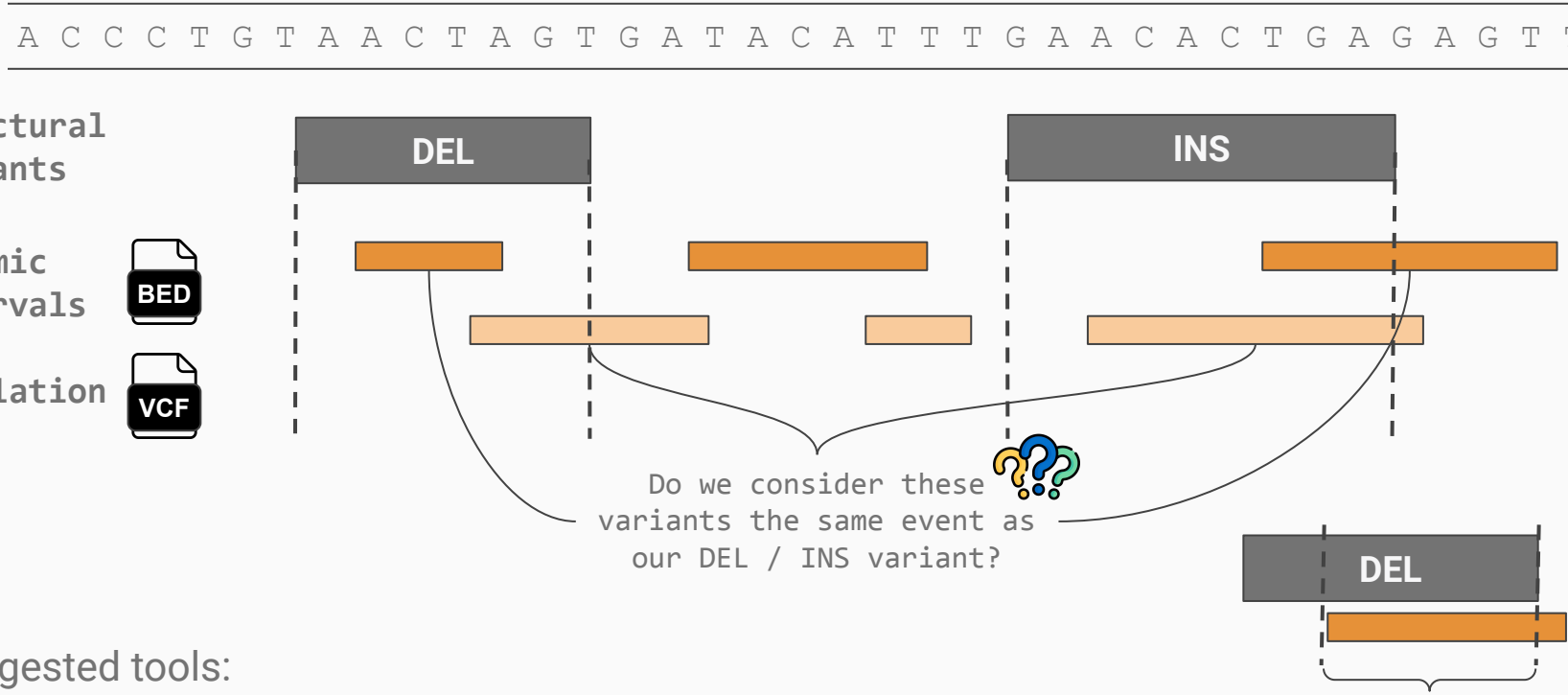
Names to use in the annotated file

INFO field to use in annotation

Operation to apply when multiple annotations are present
- max, min, mean, sum
- uniq: list of distinct values
- flag: true/false
- self: report value as it is

See documentation at https://github.com/brentp/vcfanno

# Annotate structural variants - AnnotSV and SVAfotate




Suggested tools:

- AnnotSV: annotate any bed like file based on overlap
- SVAFotate: annotate AF from reference populations

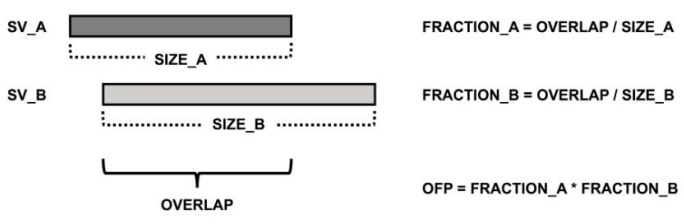
Reciprocal overlap is used to decide annotations
Generally, >= 50% accepted

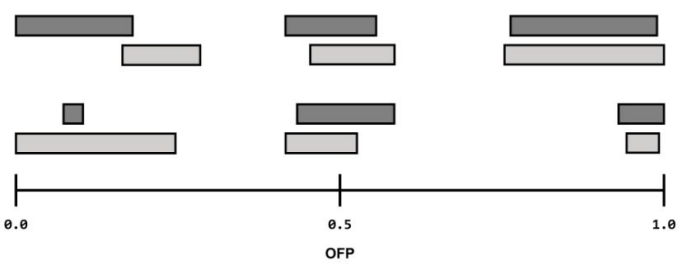
# Annotate structural variants - SVAFotate

Select annotation events based on a minimum overlap with annotation intervals

## Overlap fraction product (OFP)

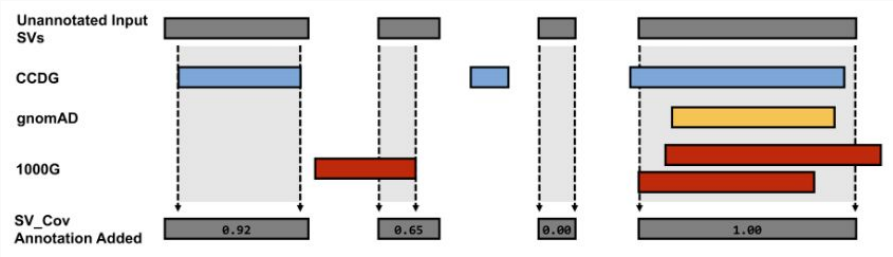
*Identify the most similar events*



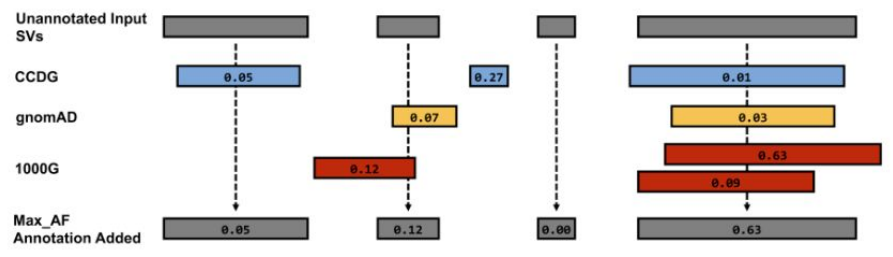

## Annotate fraction of variant covered

*How much of the variant is covered by annotation sources*




## Annotate maximum AF

*Select max AF across possible overlaps*



# Popular useful annotations

- population allele frequencies
- low-complexity regions
- deleteriousness predictors
- splicing impact predictions
- non-coding variants predictions
- conservation scores
- constrained regions
- regulatory regions

# Population allele frequencies

WGS/WES

Population variants

NHLBI TOPMed

1000 Genomes Project

gnomAD

Genome Aggregation Database

Large sequencing studies provide estimation on the variability of human genome and precise estimation of allele frequencies in different populations

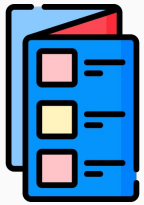
- 1000G — 2,504 whole-genome
- gnomAD v4 — 730,947 exomes and 76,215 genomes
- TopMed — 138,000 whole-genome
- Population-specific projects — UK10K, GO-NL, UKBB...

**Target**

- identify rare vars
- compare AF for QC
- find population specific variants

# Previously characterized variants

Is a variant already characterized and associated to a phenotype of interest?

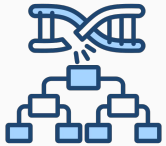


***Variant catalogs***

*GWAS associated variants*

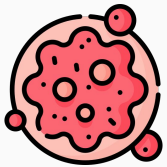
- GWAS catalog
- OpenGWAS
- GRASP
- GeneBass

*Variants involved in rare diseases*



- ClinVar
- HGMD

*Cancer related variants*



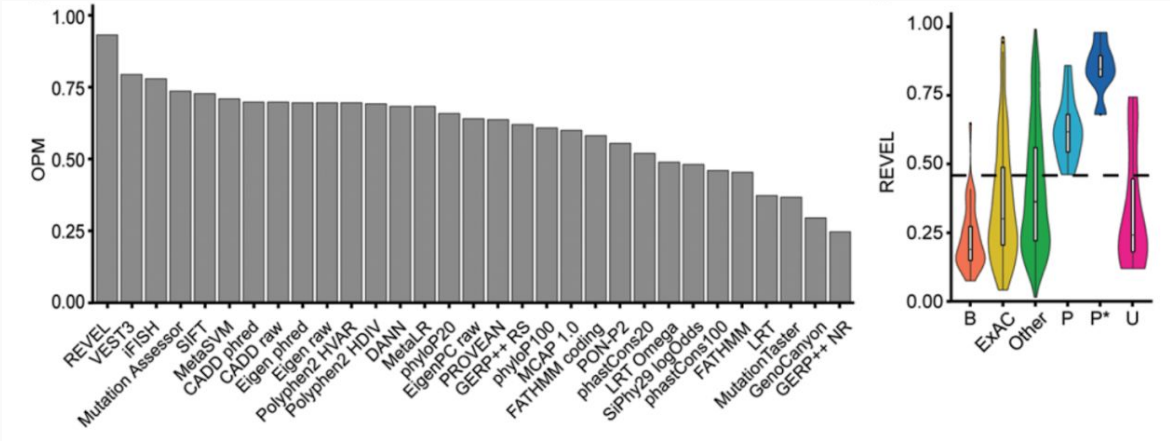
- COSMIC
- TCGA

**Target**

- interpretation of variants
- identify phenotype relevant variants

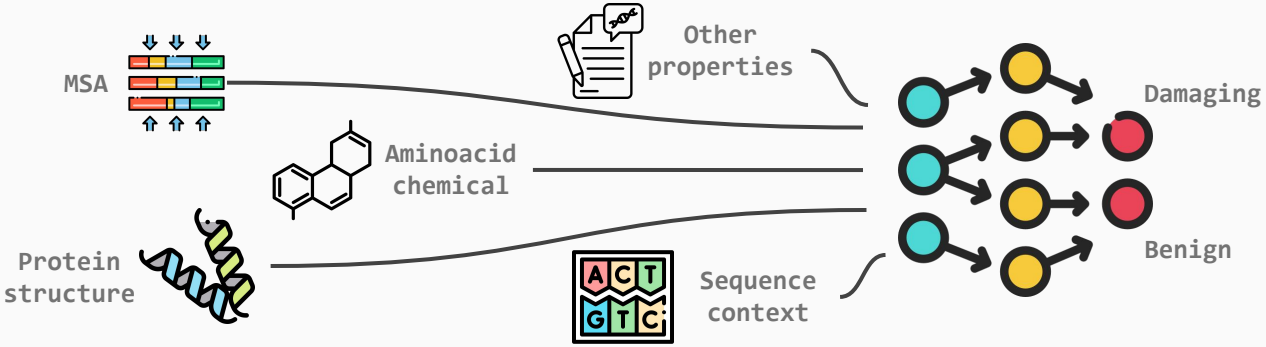
# Variant deleteriousness predictions

computational scores developed to predict the functional impact of variants, especially for missense




**Target**

- prioritize impactful variants
- remove neutral variants



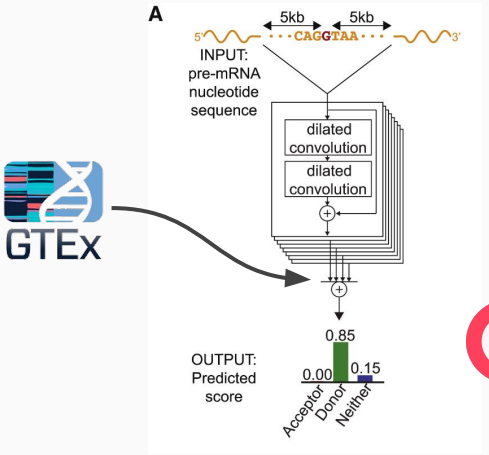

Popular available predictors:
CADD, REVEL, PolyPhen, M-CAP, EVE, PrimateAI-3D, AlphaMissense

# Splice impact predictions

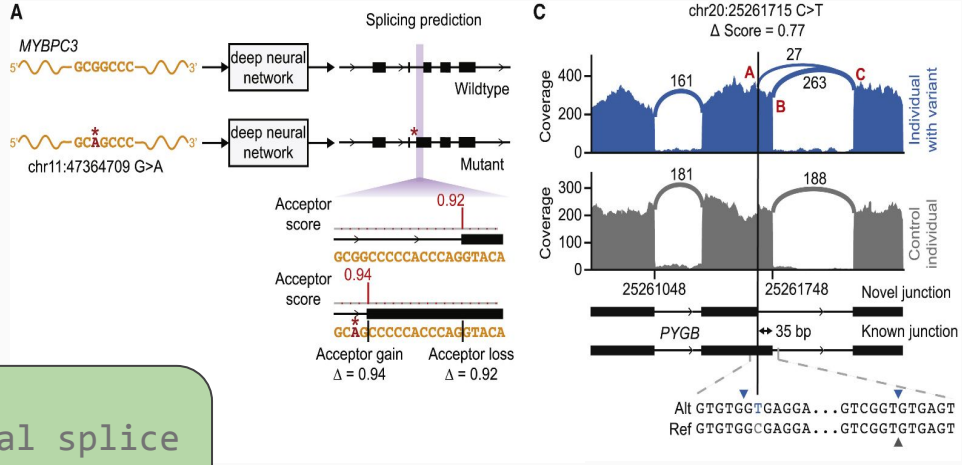
predict impact on splicing - disruption of existing splicing sites or creation of cryptic splicing events

## SpliceAI method

Training on sequence-gene expression paired dataset




Predict splicing impact for new SNVs or small INDELs




**Target**

find potential splice LoF variants in splice regions and introns

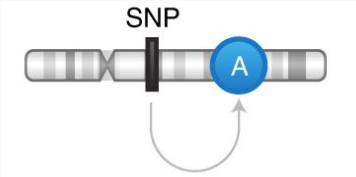
See the original paper and code repository

*Jaganathan et al., 2019*

# Non-coding variant impact prediction

ML models trained on non-coding variants, either disease-causing or eQTL variants

***Possible training sets*** ***Limitations***




*eQTLs*

- low-effect moderate frequency variants
- may be tagging variants
- mostly cis eQTLs

- DeepSEA
- Phen-Gen
- FIRE
- ncER
- ReMM
- LinSight
- FATHMM-NC

---


*Disease-associated variants*

- low number available
- mostly in promoters/introns

---

*Shared*

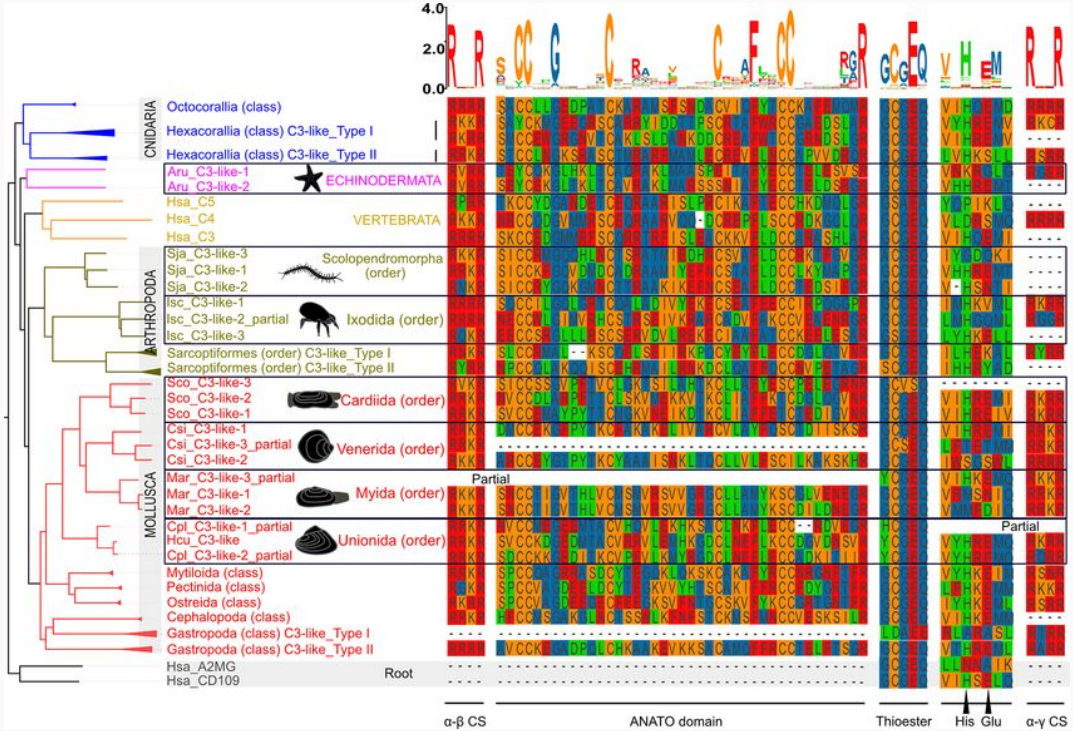
- effect is often context specific
- Regulatory elements are tolerant to point mutations

**Target**

- prioritize variants outside protein-coding genes
- identify variants more likely to have an effect on gene expression

# Evolutionary conservation of sequence

Sequence conservation based on large multi-species alignments identify important DNA positions




Peng et al., 2022

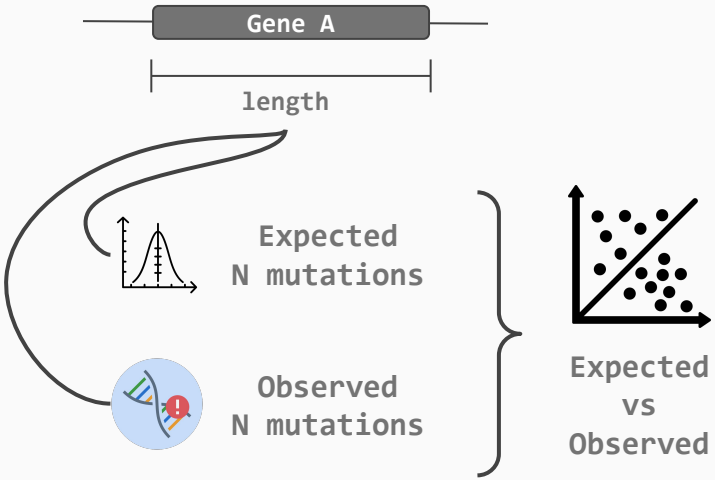
- **GERP++**
  Range -12 to +6 (Constraint site > 2)
  Higher value ⇒ More conserved.
- **PhyloP**
  Range -20 to +10 (Conserved site < 0)
  Lower value ⇒ More conserved.
- **PhastCons**
  Range 0 to 1.
  Higher value ⇒ More conserved.

**Target**

- prioritize functionally relevant variants
- identify species-specific relevant variants

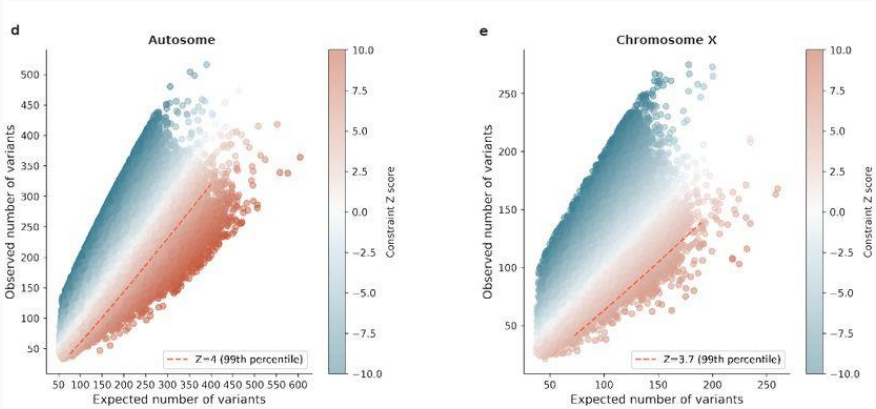
# Gene mutation constraint scores

Rank genes based on their tolerance to functional or regulatory variants. Intolerance indicates functionally relevant genes




**Target**

- prioritize medical relevant genes
- evaluate impact of selected variants

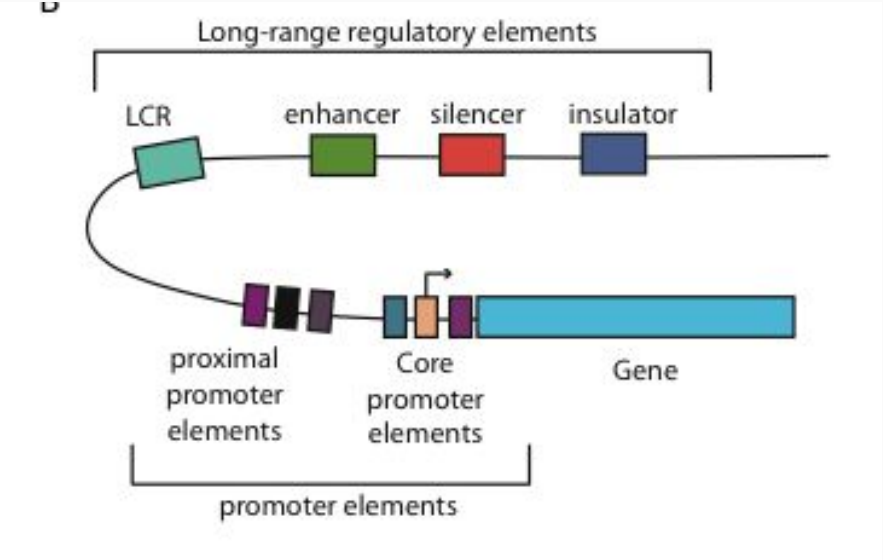



Popular scores: RVIS, GDI, LOEUF, pLOF

https://doi.org/10.1101/2022.03.20.485034

# Regulatory regions - functional elements that control gene expression

Transcription factor binding sites (TFBS), Enhancer / Silencer, Promoter regions, Insulators, DNaseI hypersensitive sites, CpG methylated regions, eQTLs.

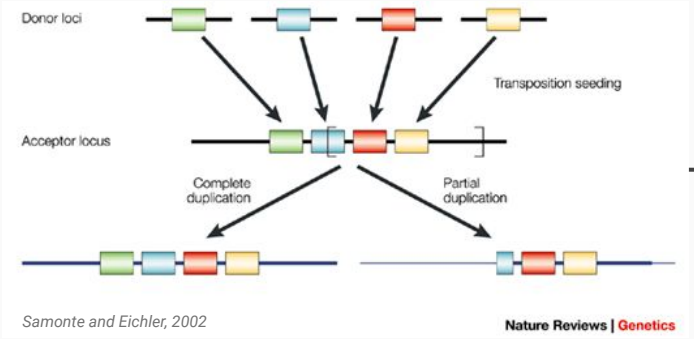



- ENCODE cCRE/DNase/TFBS
- FANTOM5 promoters/enhancers
- Roadmap epigenomics project
- GTeX eQTLs

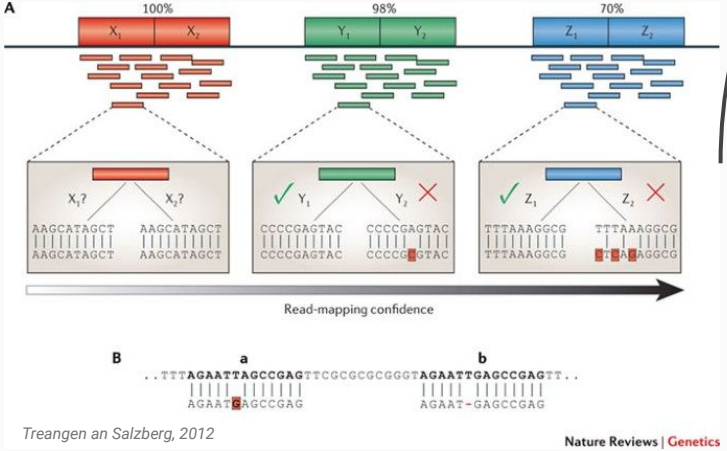
**Target**

- prioritize variants outside genes
- connect non-coding variants to genes (e.g. for burden tests)

# Additional region-based annotations




*Samonte and Eichler, 2002*




*Treangen an Salzberg, 2012*

Regions where mapping and variant calling may be unreliable, useful for QC:

- Segmental Duplications
- Low mappability and low complexity regions

Additional functional regions of interest

- TargetScan miRNA interaction
- InterPro and PFAM for protein domains

**Target**

- identify FP/FN variants
- prioritize project-specific functional regions