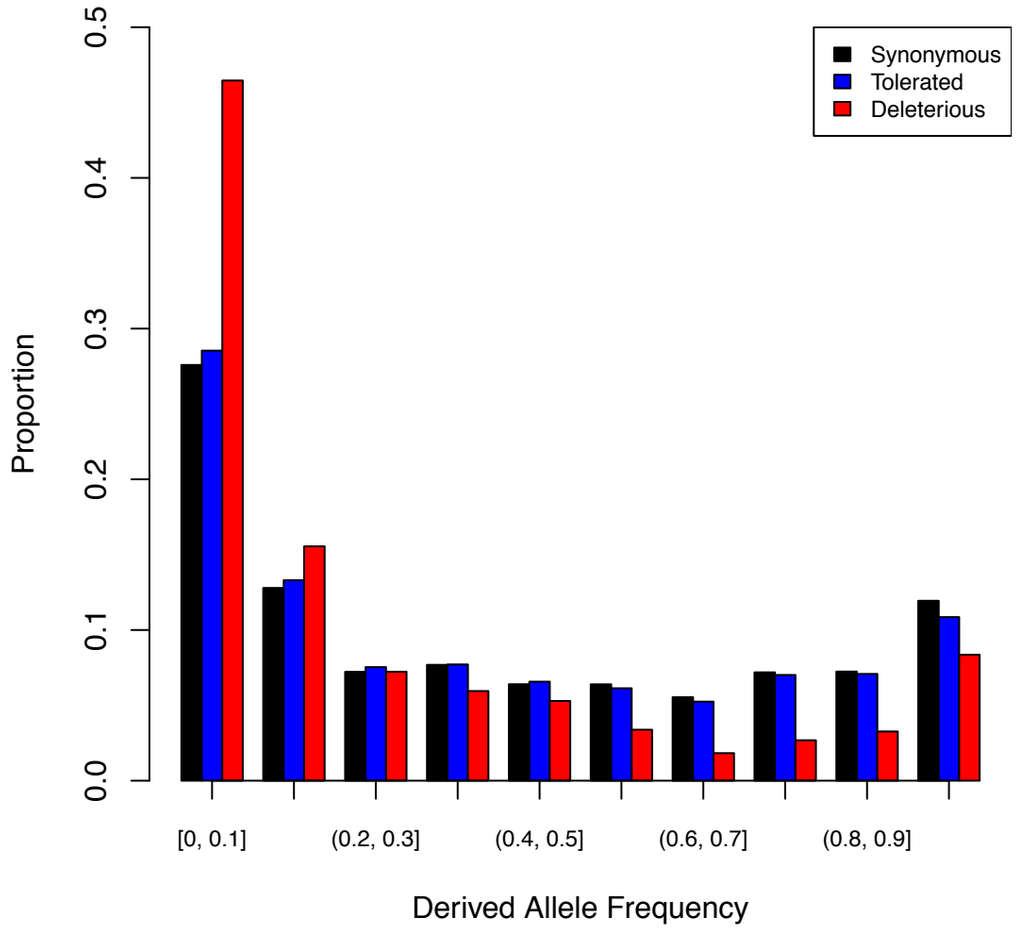
Variant Annotation

2017-02-28

Why Annotate Variants?

- Candidates for trait mapping…
- Putatively neutral variants for population genetics

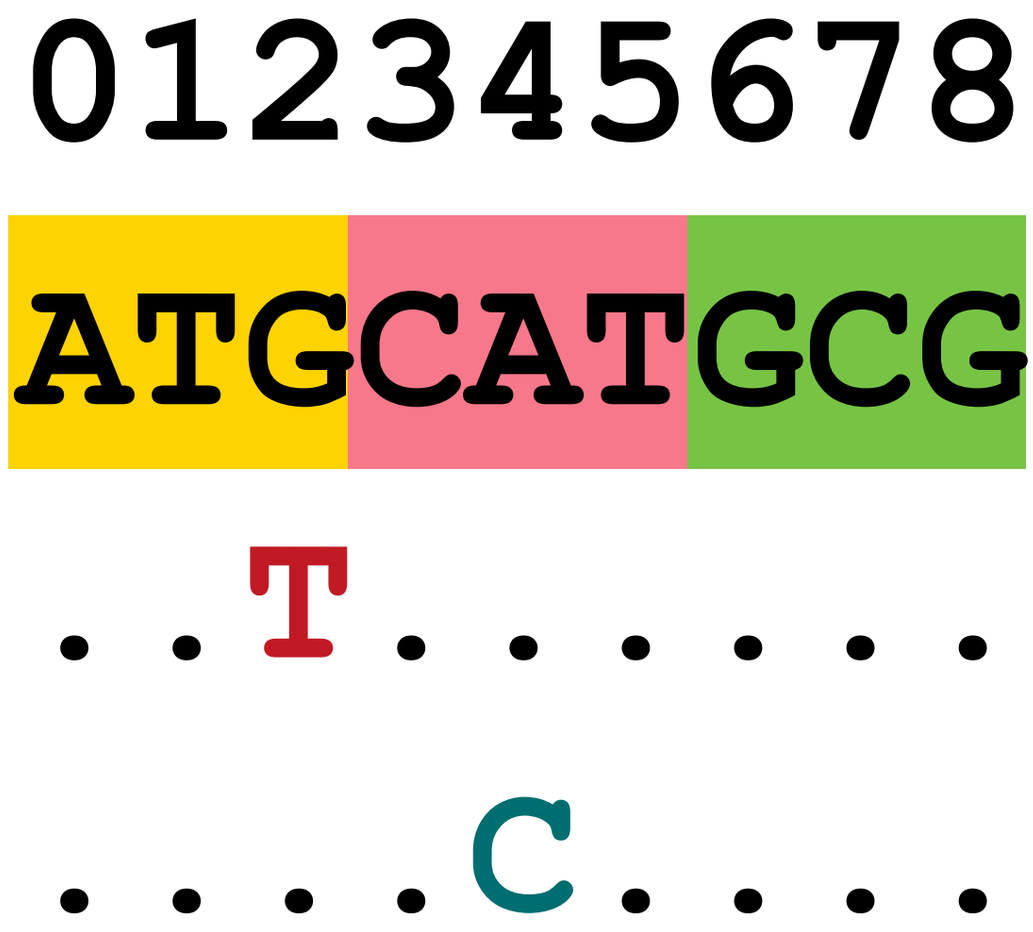


A Photo Analogy

- Photo data: this nice duck
- EXIF data (metadata):
  - Date/time
  - Exposure settings
  - Geographic coordinates

Example Variants

- Variant data: Position, reference, alternate
- Annotation data:
  - Synonymous/ Nonsynonymous
  - Gene name
  - Functional impact



Example Variants

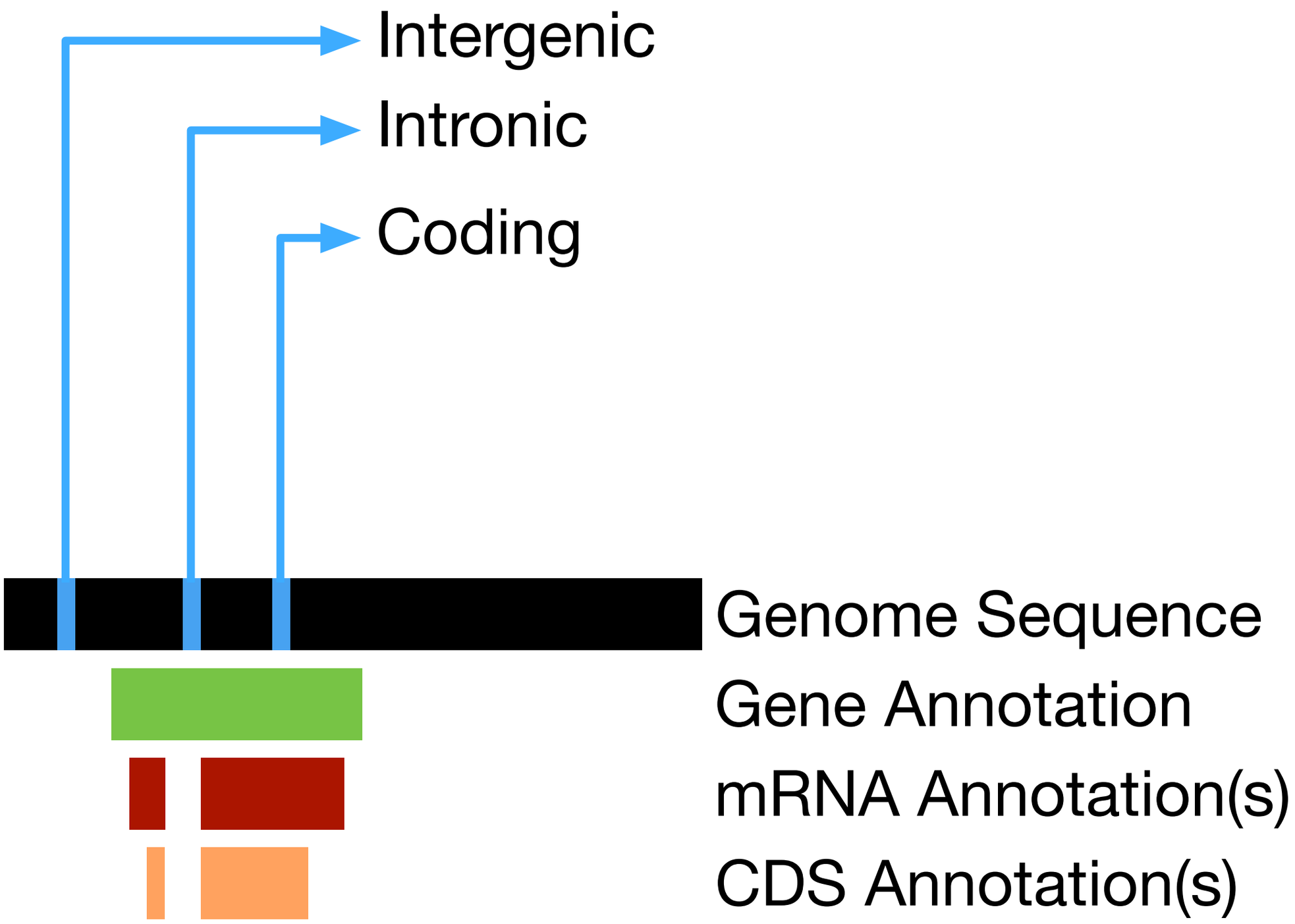
- Variant data: Position, reference, alternate
- Annotation data:
  - Synonymous/ Nonsynonymous
  - Gene name
  - Functional impact

```
01234567
ATGCATGCG
..T......
....C....
```

BAD_Mutations etc.

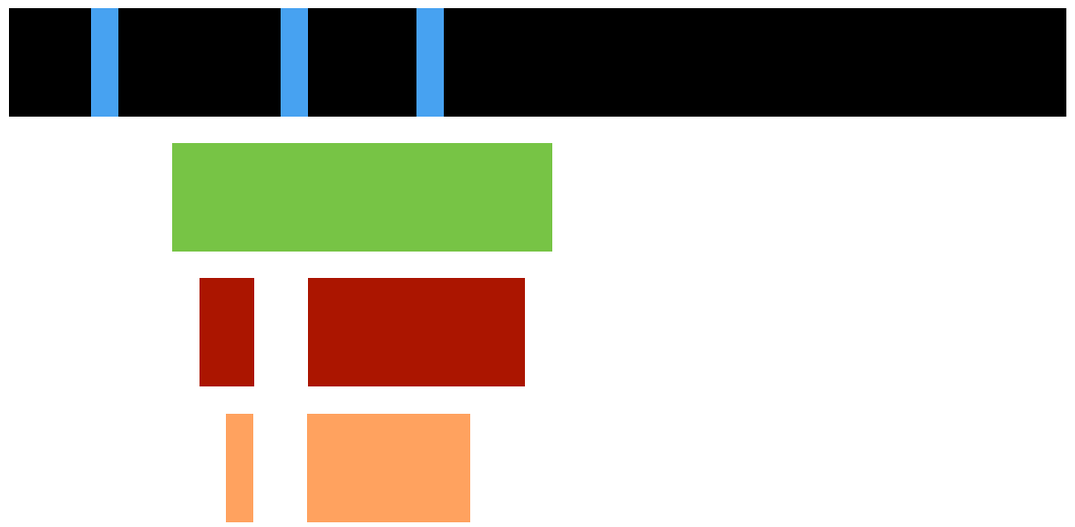
ANNOVAR etc.

Annotation Schematic



What You'll Need

- Variants (VCF)
- Reference assembly (FASTA)
- Gene annotations (GFF)
- Annotation software
  - Custom script
  - ANNOVAR, SNPEff…



ANNOVAR

- Homepage: http://annovar.openbioinformatics.org/en/latest/

- Download requires registration (free)

- Need ‘gtfToGenePred’ from this link: http://hgdownload.cse.ucsc.edu/admin/exe/

  - Choose the platform that is correct for you

- Also need Perl

ANNOVAR Example

- Clone (or pull) the latest version of the repository
- Read the commands and comments in the ‘annovar_cmds.sh’ script
- If you downloaded the prerequisites and ANNOVAR software, edit the script and try it out.
- If you cannot, there are pre-built annotations in the ‘Annotations/‘ directory.

ANNOVAR Example

- Output files are complex, but consistent
  - Important fields in ‘exonic function’:
    2 - synon./nonsynon./nonsense/etc.
    3 - Transcript and amino acid states
    12 - SNP ID (rs identifier, e.g.)
- Other info included is frequency, transition/transversion, VCF metadata

ANNOVAR Example

SNP ID	Functional Class	CDS Impact	AA1	AA2	Residue
Barley_666866	Noncoding (UTR5)	NA	NA	NA	NA
Barley_271520	Exon	Nonsyn	Gly	Asp	8
Barley_271521	Exon	Nonsyn	Glu	Asp	26
Barley_271522	Exon	Syn	Ala	Ala	130
Barley_666867	Exon	Nonsyn	Thr	Ile	147
Barley_271523	Exon	Syn	Ala	Ala	150
Barley_271525	Noncoding (UTR 3)	NA	NA	NA	NA
Barley_666868	Noncoding (UTR 3)	NA	NA	NA	NA

ANNOVAR Quirks

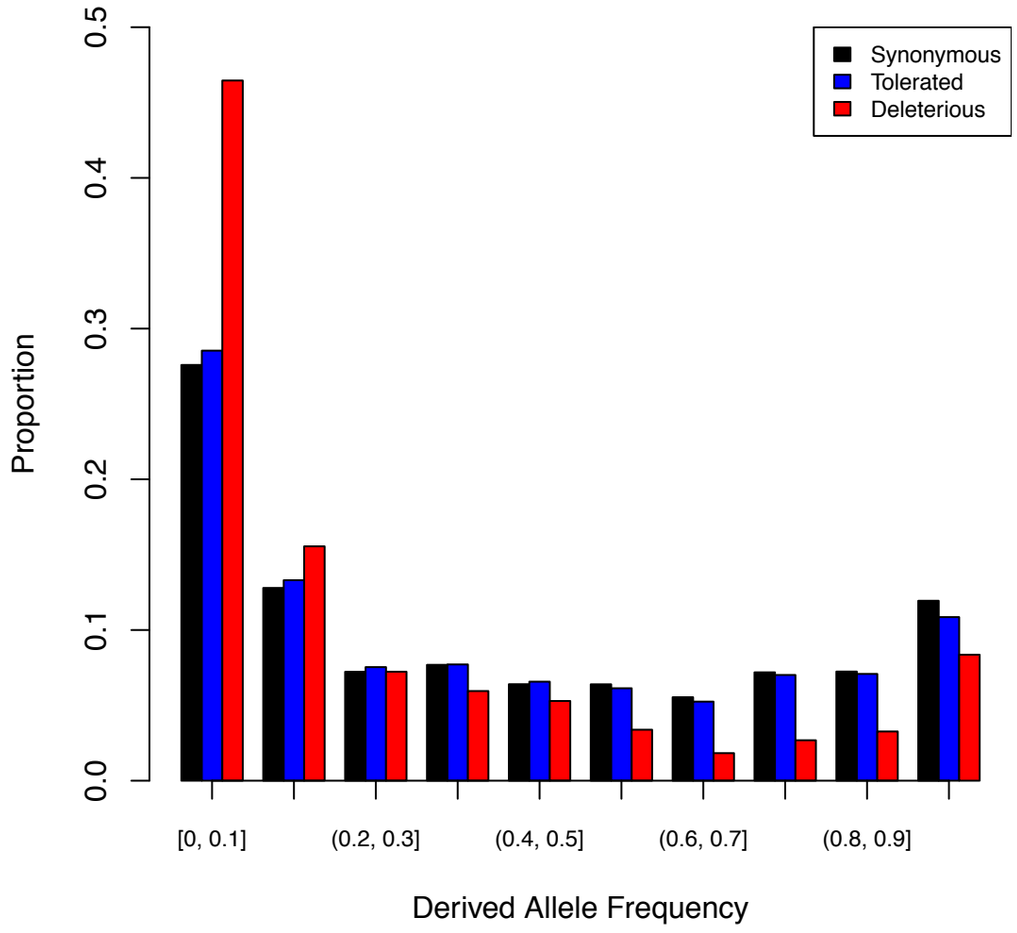
- Does not play nice with VCF format, especially from the GATK.
    - Script and command line are provided to convert to ANNOVAR-preferred format, though

- Uses **GTF** format, rather than **GFF** format

- Generates two output files - you will need to merge them, or link them in your scripts

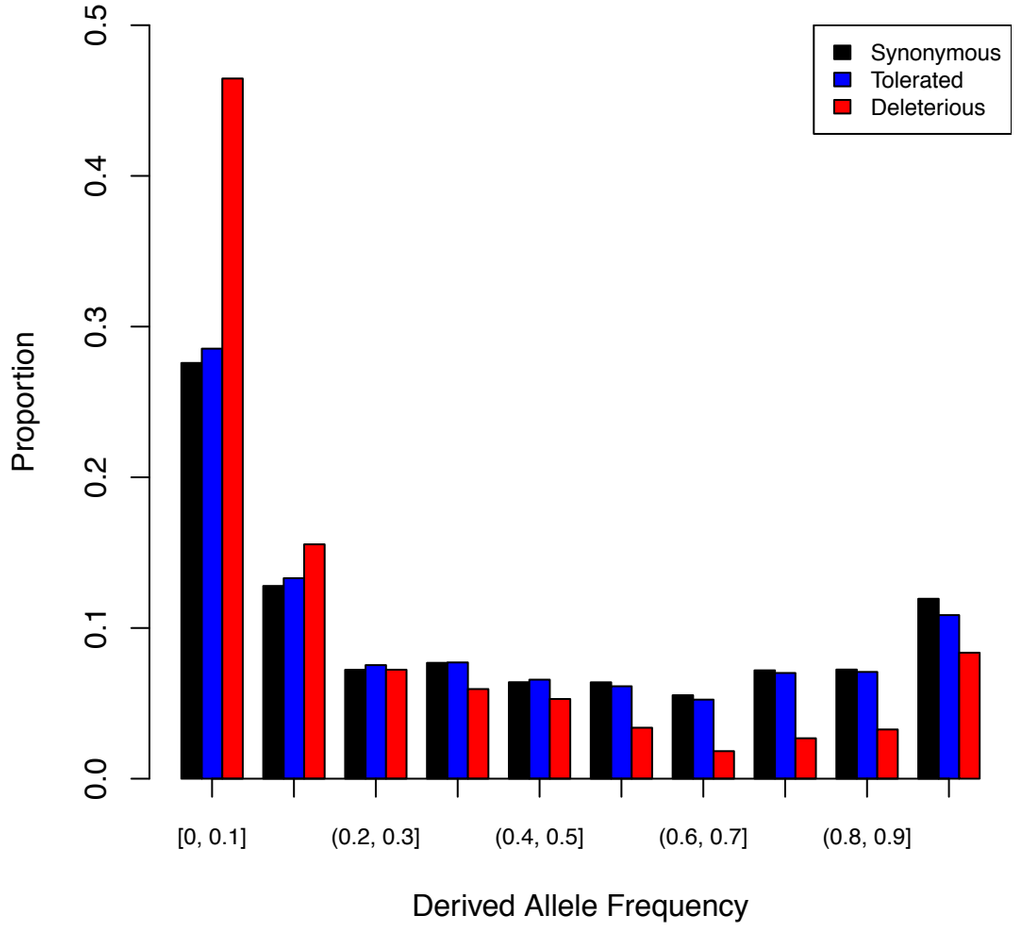
Why Annotate Variants?

- Candidates for trait mapping…
- Putatively neutral variants for population genetics



Why Annotate Variants?

- "*All variants are interesting, but some variants are more interesting than others.*"



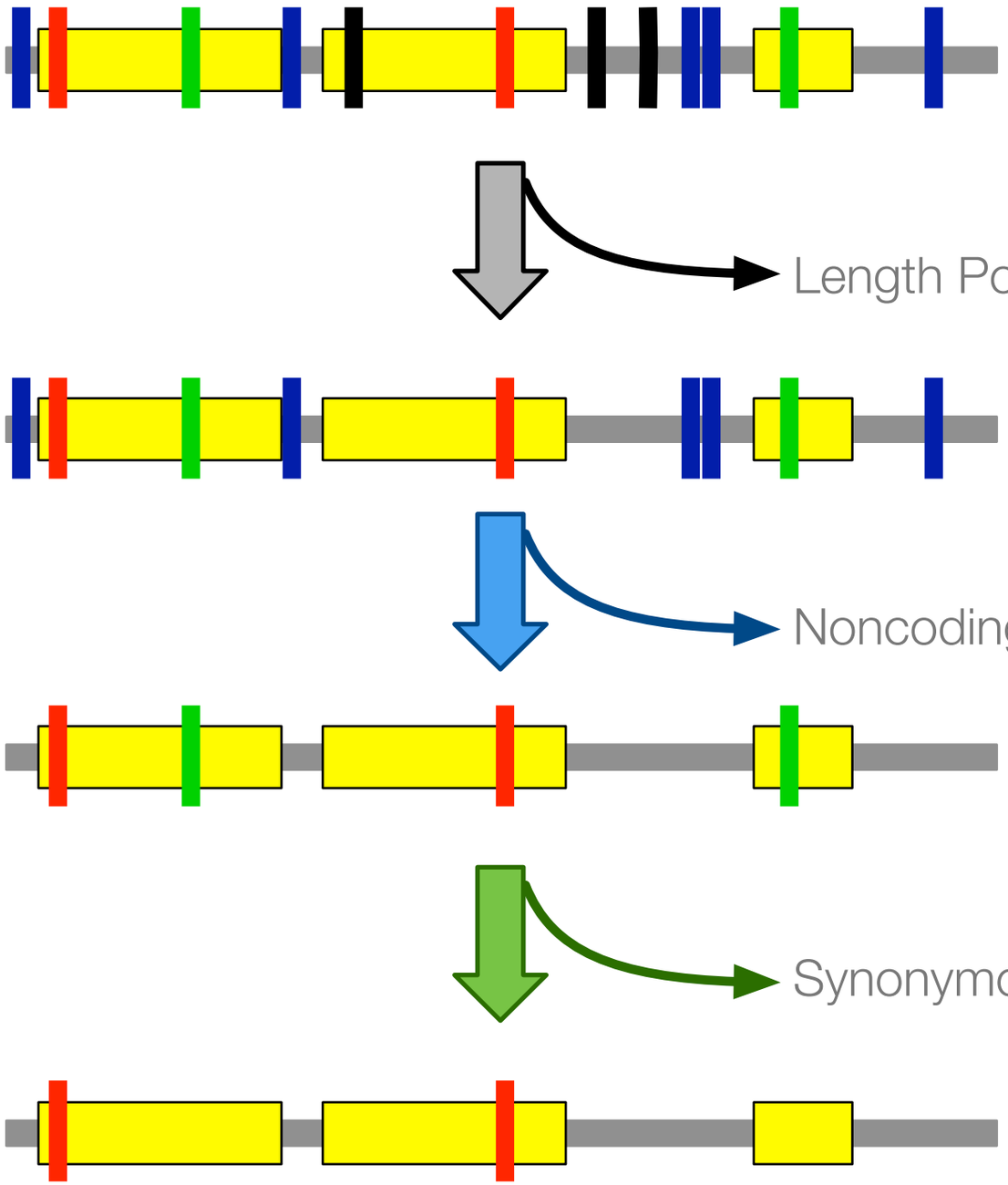
BAD_Mutations

- One class of ‘potentially interesting’ variants occur at phylogenetically conserved sites
  - Variants in sufficiently conserved sites are presumed to have deleterious effects

- A long debate about genetic load in bottlenecked populations… Not for here.*

*: see http://www.sciencedirect.com/science/article/pii/S0168952516000147 (Brandvain and Wright 2016)
and http://www.nature.com/ng/journal/v46/n3/full/ng.2896.html (Simons et al. 2014)
and https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-9-r84 (Marth et al. 2010)

BAD_Mutations



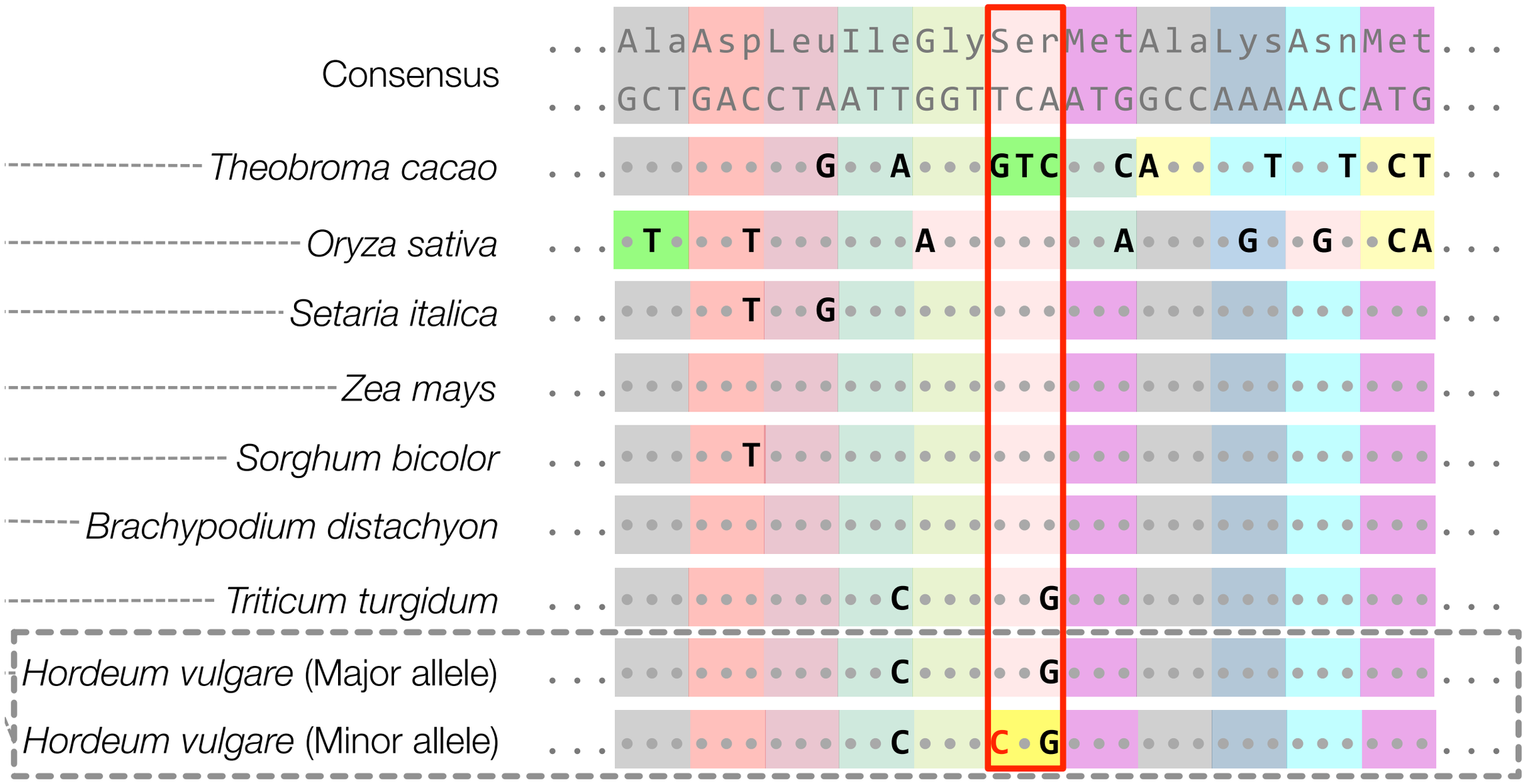
All Variants

SNPs

Coding SNPs

Nonsynonymous SNPs

BAD_Mutations



BAD_Mutations

Deleterious

```
N  D  DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
H  Y  CCCCYY--YYYCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C  F  FFFFFFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
R  T  TTTMTTTTTTTTTTTTTTTDTTTTTTTTTTTTTTTTTTTTT
A  T  VVAAVVVAVAAAVSVVVTVVVVIVAAAALATTTVVVVV
A  V  VVVVVIVVVIIIIIVVVV-VVV-VVVVVVVVVVVVVV
D  E  EDDEEEEEEEE-EEEEEEEEEEEEEEVVEEEEEEEEE
Q  L  LLLLLLL-LMLLLLLLLLLLL-LLLLLLLLLLLLLLL
```

Tolerated

```
D  N  NDDDNNNNNDDDDNDNNNNDDDDDDDDDNNNNNNDNN
V  I  VVLVVVMMVVVVVVVSVIIIIVMVVVVVIVVVVVIVVV
A  V  AAAAAAAAAAAAAAAAA-AVT-TA-SAAATTTTTAAT
D  E  DDDDDDDN-DDDDDDDDDNEDD-EEEEEEEDDDDDDDD
A  P  EKSKEEEKEENAENEEEEDAAPAAAEKKEEGGEEEEEE
E  D  RNKKKRK-KKKKKNSSKDPESSPS-KS-SN---KKHHN
E  D  EEEEEEQDDEEEEEEEDEEEDEEDDDDEEEEEEEEED
E  A  TATTMTTSTTTTATSSCA-SPPMASREECGSSSMTAAT
```

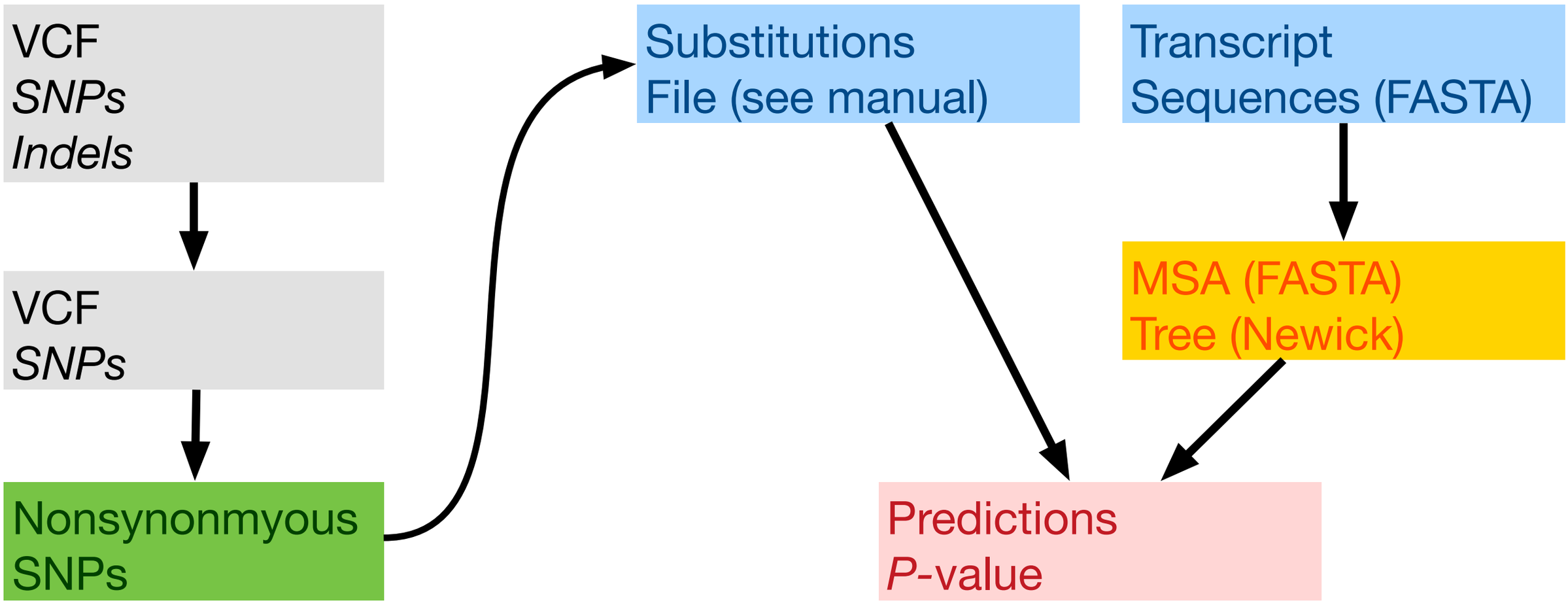
- Heuristically, deleterious variants are in positions with high conservation, and tolerated variants are in positions with low conservation

- But! It is a formal likelihood ratio test of sequence conservation

BAD_Mutations

- https://github.com/MorrellLAB/BAD_Mutations
- Python program, several dependencies. See the manual!
- Uses publicly available Angiosperm genomes from Phytozome and Ensembl Plants
- No example, has very high runtimes (~hours per gene)!

BAD_Mutations Workflow



Other Tools Exist

- SNPEff - less flexible and more error-prone than ANNOVAR
  VEP - from Ensembl. Works well on “nice” genomes.

- **S**orting **I**ntolerant **F**rom **T**olerant (SIFT)
  Polymorphism **Phen**otyping **2** (PPH2)
  Protein **V**ariation **E**ffect **An**alyzer (PROVEAN)
  Genomic **E**volutionary **R**ate **P**rofiling (GERP++)