MCBBi | NOVA
UNIVERSIDADE NOVA DE LISBOA

Applied Computational Multi-Omics

Genetic Alterations and Functional Impact

NOVA
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE CIÊNCIAS DA VIDA
DEPARTMENT OF LIFE SCIENCES

Topics

https://onlinelibrary.wiley.com/doi/10.1002/humu.24311

- Finding Genetic Alterations
  - Single Nucleotide Variants and small Indels

    https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/
  - Copy Number Alterations and other Structural Variants

    https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300727/
  - Finding clonal vs subclonal variants

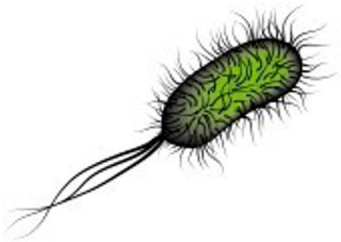
    https://www.sciencedirect.com/science/article/pii/S2001037017300946
- Variant annotation

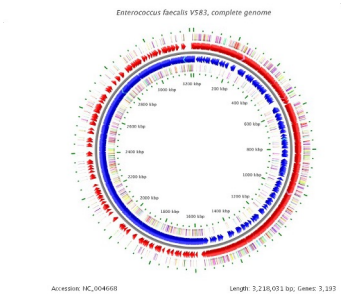
  https://www.nature.com/articles/nprot.2015.105

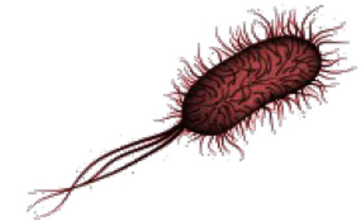
  https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

Application of Resequencing: variant calling

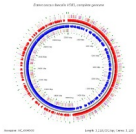
Most frequent biological question: find mutations causing certain phenotypes

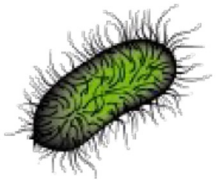




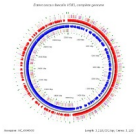


?

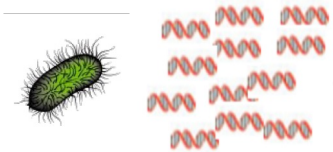




?



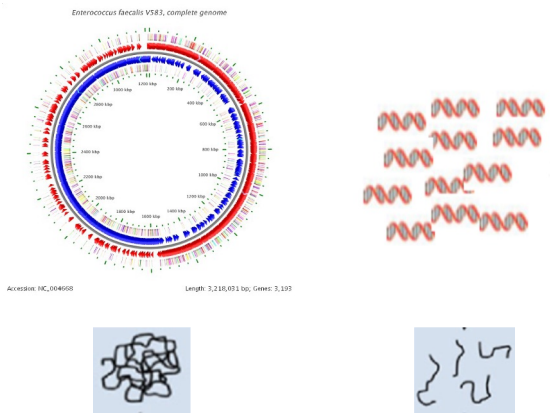
Application of Resequencing: variant calling



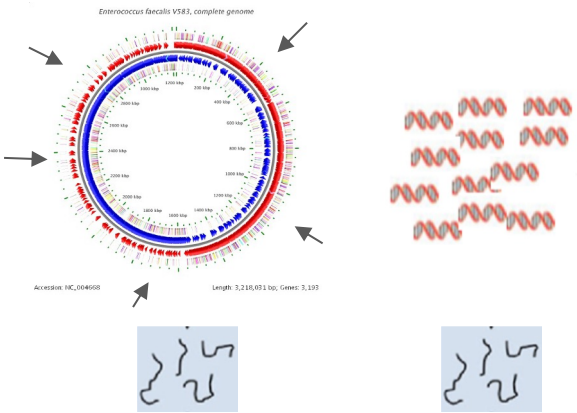
DNA Extraction

Application of Resequencing: variant calling

Whole Genome (WGS) or Targeted (eg. Whole Exome - WES)

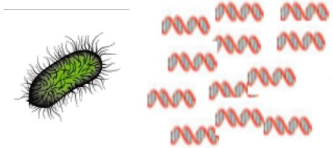


Whole Genome



Targeted

Application of Resequencing: variant calling



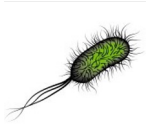
DNA Extraction

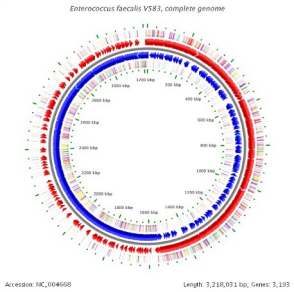
GTTGT TGCTCAGTT TGAC ACTCCAT

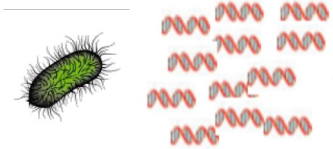
Sequencing

FastQ files

Application of Resequencing: variant calling







DNA Extraction

AGTTGTGT
TGCTCAGTT
AAGCGATG TCCATGC GTGTGTTT-CA
ACTCCAT GTTGT

geneA geneB geneC
AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

GTTGT TGCTCAGTT TGAC ACTCCAT

Sequencing

SAM/BAM files

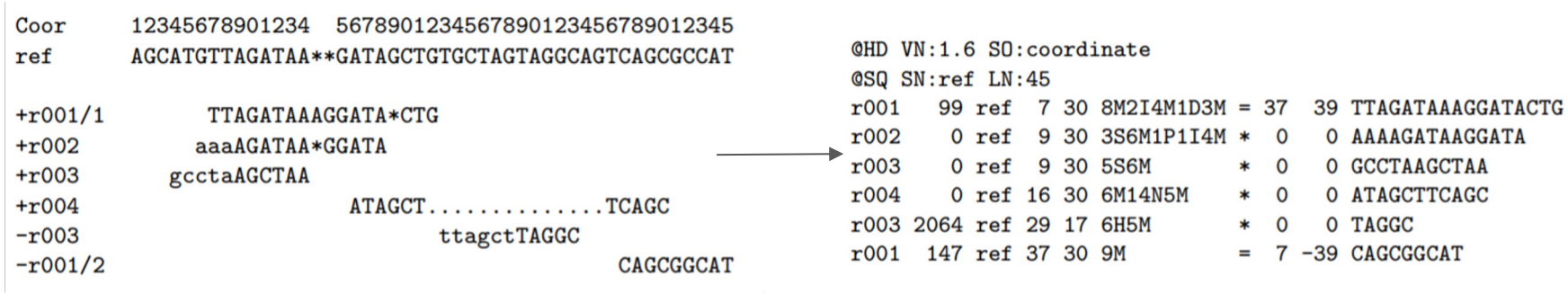
Alignment algorithms (most frequent is BWT)

FastQ files

HTS Data Analysis: Resequencing

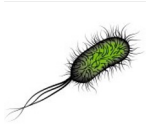
Sequence Alignment/Map Format (SAM): a file format to represent alignments

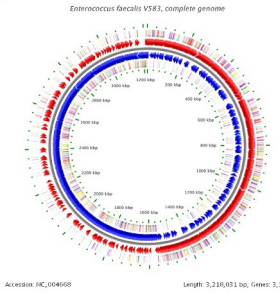
- Most often used for HTS alignments

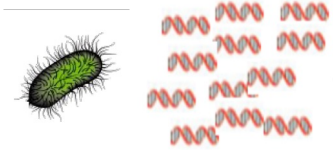


https://samtools.github.io/hts-specs/SAMv1.pdf

Application of Resequencing: variant calling







DNA Extraction

Amplification

SNP

SNP

AGTTGTGT

Indel

TGCTCAGTT

AAGCGATG TCCATGC

ACTCCAT

GTGTGTTT-CA Deletion

GTTGT

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

GTTGT TGCTCAGTT TGAC ACTCCAT

Sequencing

VCF files

Variant Detection algorithms

SAM/BAM files

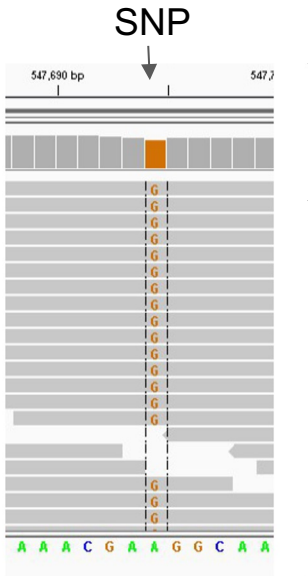
Alignment algorithms (most frequent is BWT)

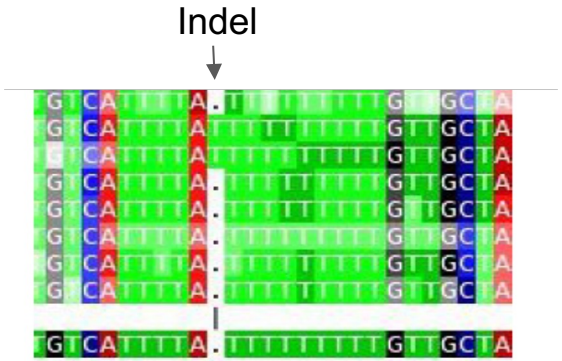
FastQ files

Single Nucleotide Variants (SNV) and small Indels

Variants detected within reads (smaller than size of read)

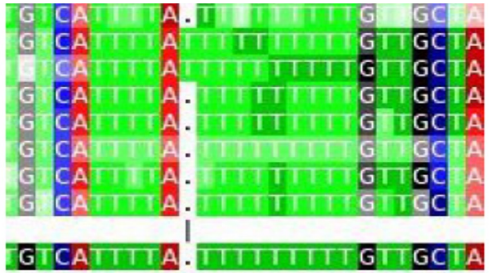
- SNVs:
  - Change of a single nucleotide
- Indels:
  - “Small” deletion or amplification (less than the size of a read)





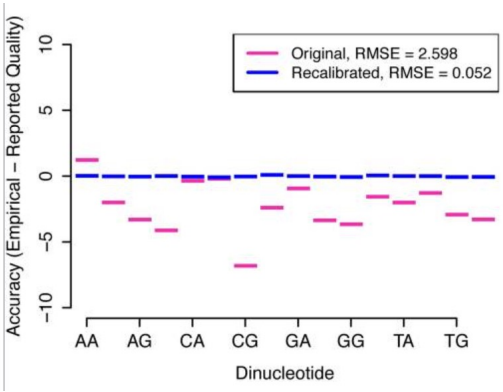
Single Nucleotide Variants (SNV) and small Indels

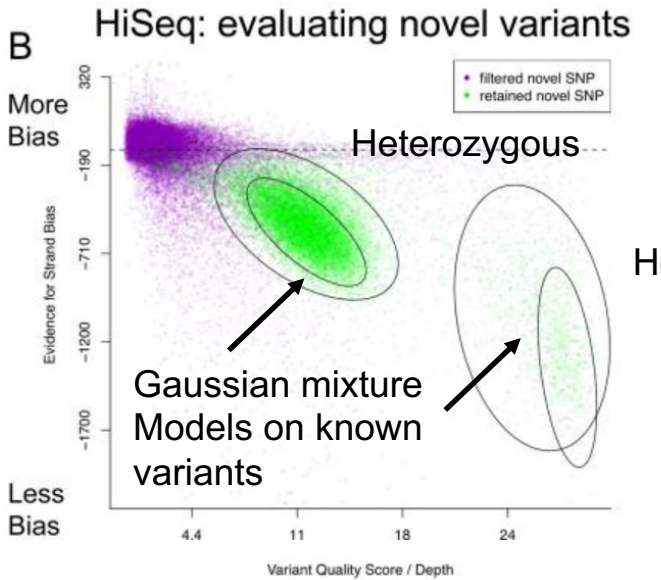
- Main factors affecting the detection of SNVs and indels
  - Number of reads (coverage)
  - base quality (mostly affects SNVs)
  - Duplicates (eg. Bias due to PCR)
  - Misalignments (mostly affects indels, but also affects SNVs)
  - Strand bias (mostly in the case of targeted sequencing and near repetitive regions)



Single Nucleotide Variants (SNV) and small Indels

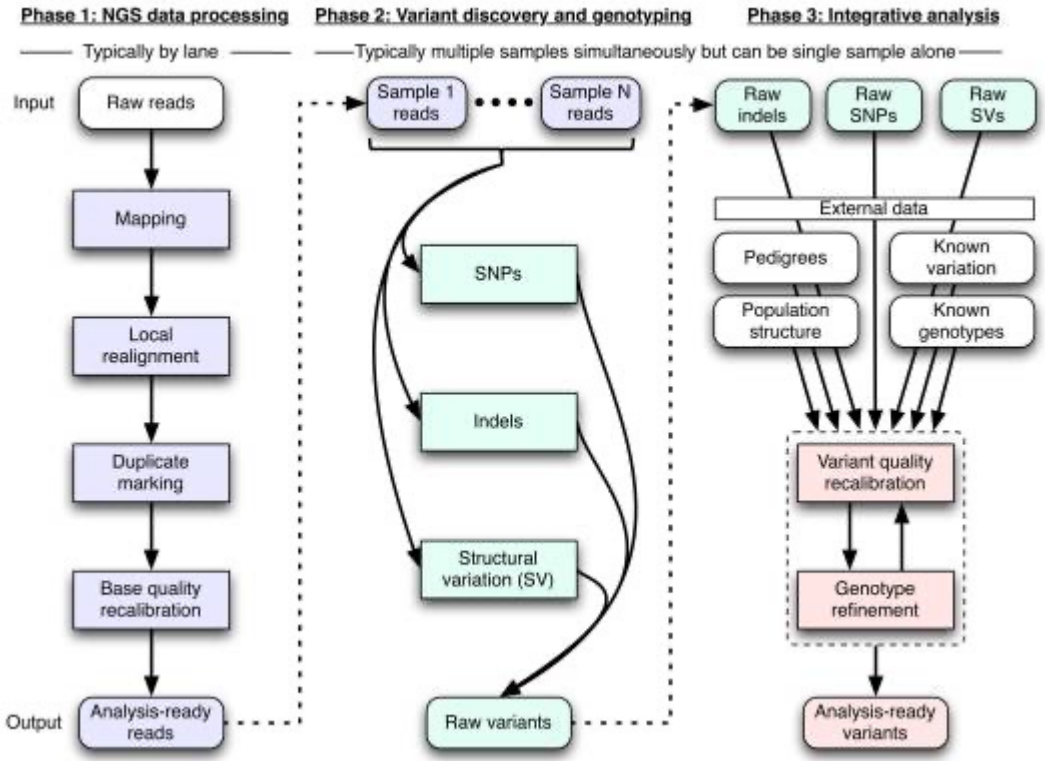
- In the case of Human and other well studied model organisms
  - Known variants (obtained and confirmed using other methods) can be used to
    - Perform base recalibration
    - Evaluate alignment problems
    - Perform variant recalibration





(DePristo et al.)

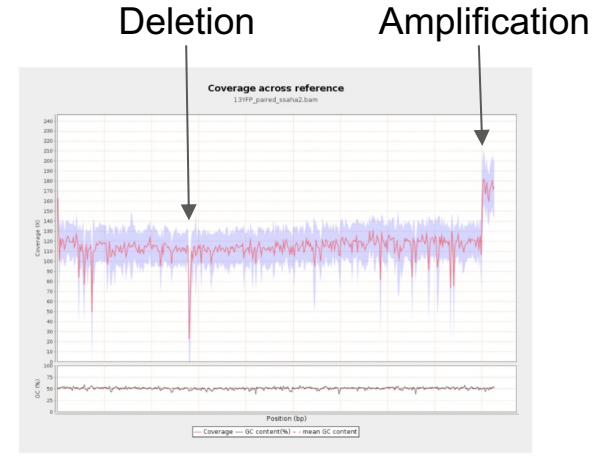
DePristo et al. (GATK)



Copy Number Variants and other Structural Variants

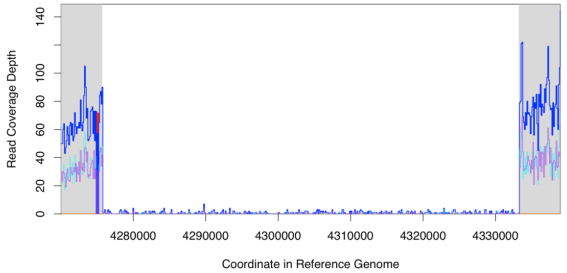
Variants larger than the size of reads

- Copy Number Variants
  - Large Deletions and Amplifications
- Other Structural Variants
  - Fusions; Inversions, etc...
- Horizontal Transfer
  - Needs to be analyzed separately

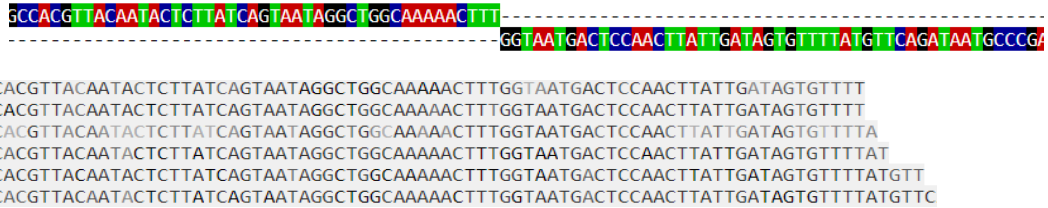


Copy Number Variants and other Structural Variants

- Evidence used to detect Structural Variants
  - Differences in Coverage
    - Most commonly used
    - Particularly with targeted sequencing
      - Although there's still amplification bias



  - Junction evidence (difficult in targeted sequencing)
    - Can use paired read information (namely, expected fragment length – noisier)
    - Can use information within reads (more precise - requires bigger reads)



Finding clonal vs subclonal variants

- Distinguish finding variants vs inferring genotype
  - Finding variants: not due to technical errros
  - Infer the genotype: what the most likely genotype (homozygote; heterozygote in diplody)

- Finding subclonal (population) variants require specific analysis
  - Many algorithms assume diplod organisms to infer a genotype
    - Eg. a “real” variant at 25% is most llkely associated to a heterozygous genotype
  - To find subclonal populations you can infer genotypes using a large ploidy
  - In case of haploid, “real” variants at less than 100% are considered subclonal

Variant Call Format (VCF):

A file format to represent variants and their properties

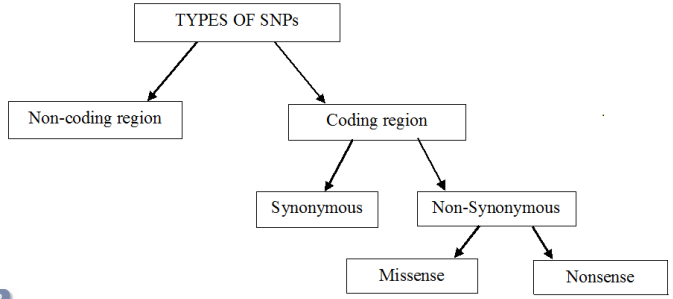
```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

From https://doi.org/10.1093/gigascience/giab007

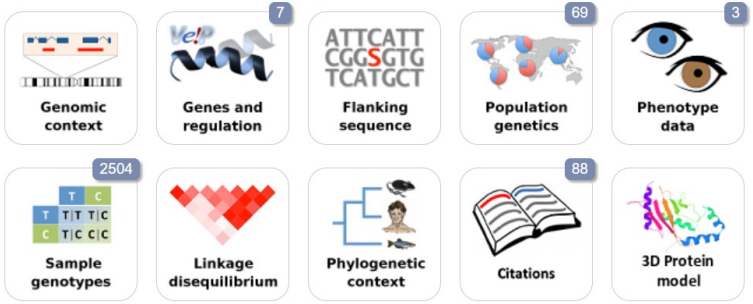
https://samtools.github.io/hts-specs/VCFv4.2.pdf

Variant Annotation

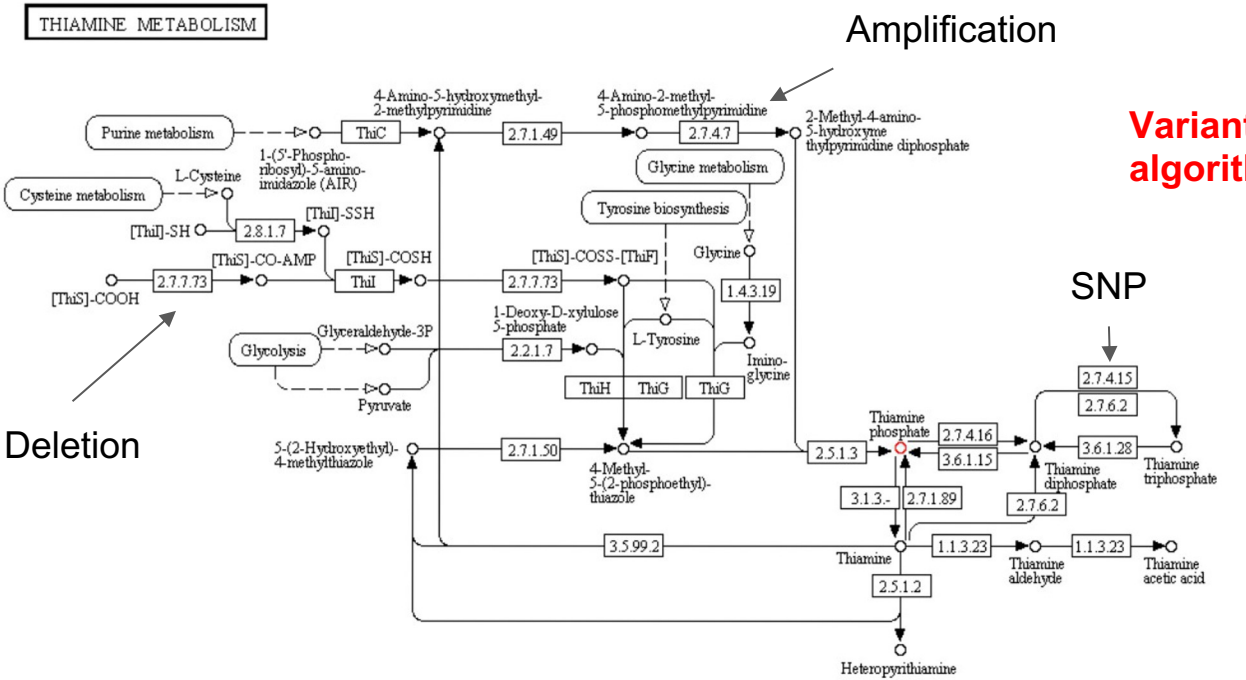
- Main goal is to uncover effect and relevance of variant
  - Coding versus non-coding
    - Coding: Silent versus non-silent
    - Non-coding: can be complex



- For Human and other model organisms
  - Population frequency; Disease-association; etc...



Variant Annotation



Variant Annotation algorithms

Summary

- Finding Genetic Alterations using High Throughput Sequencing
  - General process: from fastq to BAM to VCF
  - Distinguish beween small SNVs/indels and large structural variants
    - Different methods are used to find evidence for each of them
    - What are the main factors affecting their detection
  - Distinguish between uncovering clonal / subclonal (population) variants
    - Estimating most likely genotype versus finding a real (not artifactual) mutation

- Variant Annotation
  - Distinguish coding versus non-coding variants
  - Distinguish Coding (silent versus non silent)
  - Integrate other types of information (eg. population frequency)