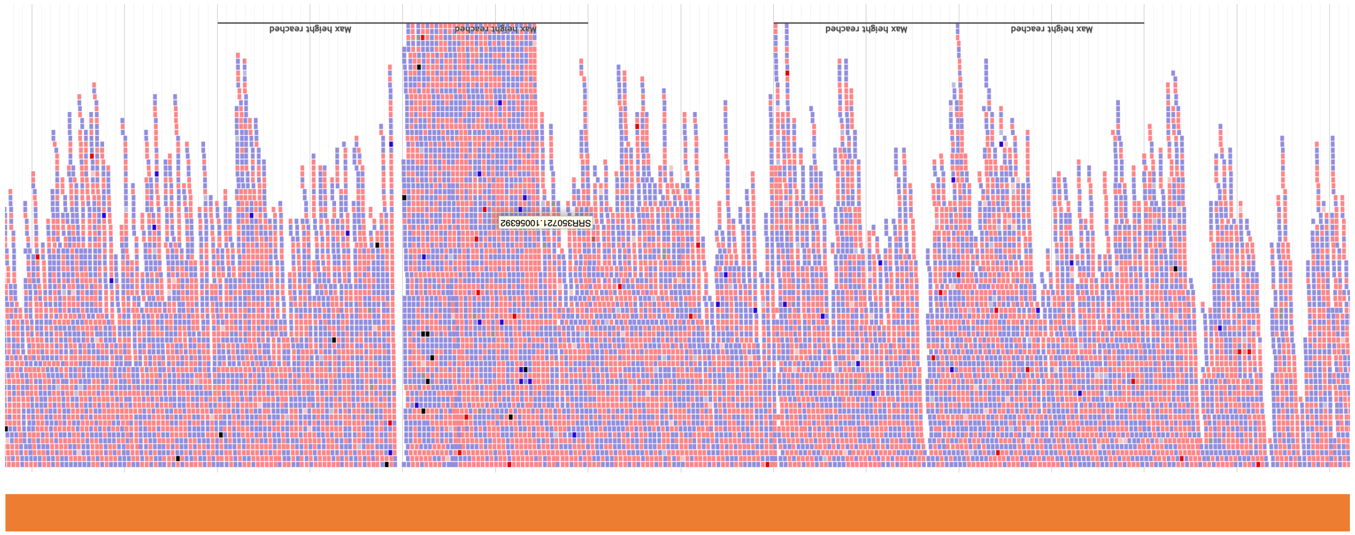
Variation data from high throughput sequencing

Omar S. Harb, PhD
University of Pennsylvania
VEuPathDB

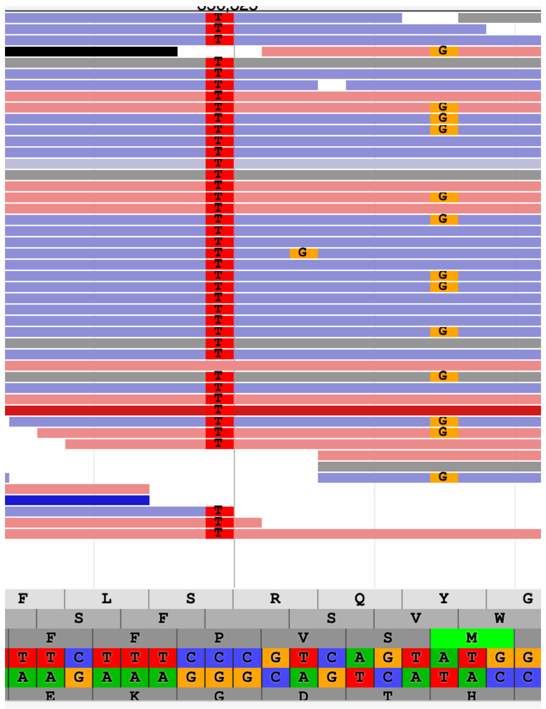
Experimental Sequencing Reads



Aligned Reads

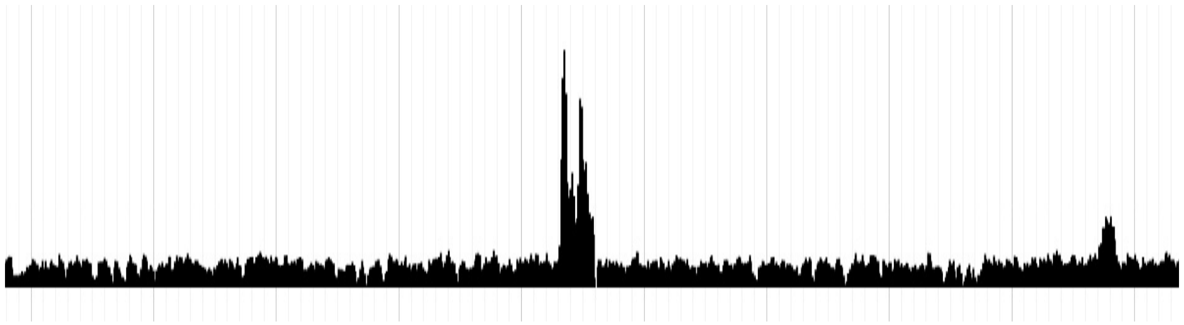
Reference Genome

Variants



F L S R Q Y G
S F S V W
F F P V S M
T T C T T T C C C G T C A G T A T G G
A A G A A A G G G C A G T C A T A C C
E K G D T H

CNV



Single Nucleotide Polymorphisms (SNPs)

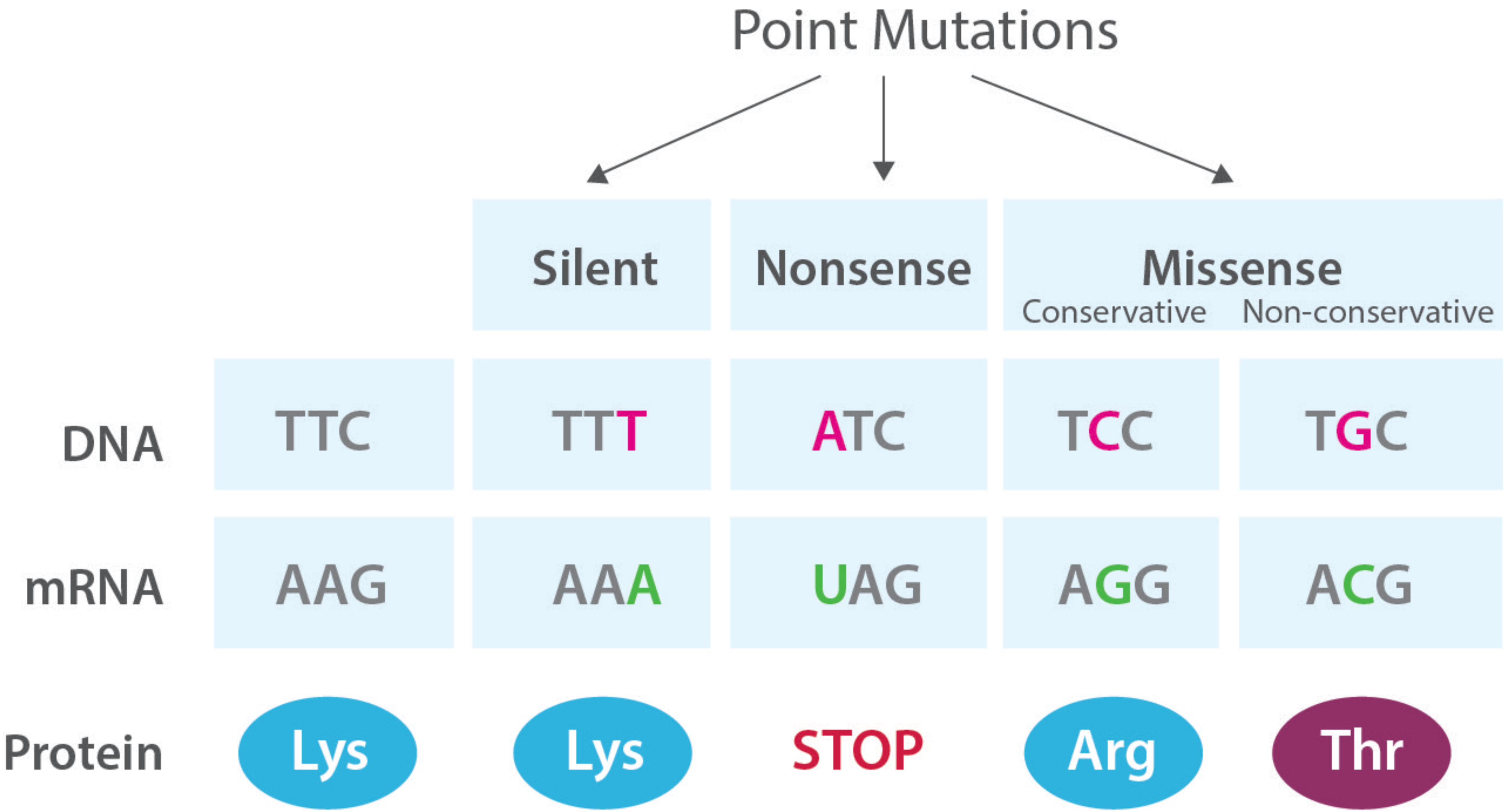
What are SNPs

SNPs are genetic variations. Each SNP represents a single change in one nucleotide.

There are different types of SNPs but broadly can be divided into synonymous, nonsynonymous non-coding.

Synonymous: a nucleotide change that does not change the amino acid sequence – also called silent mutation.

Nonsynonymous: a nucleotide change that results an amino acid change. These can be missense (results in a different amino acid) or nonsense (results in a stop codon) mutations.



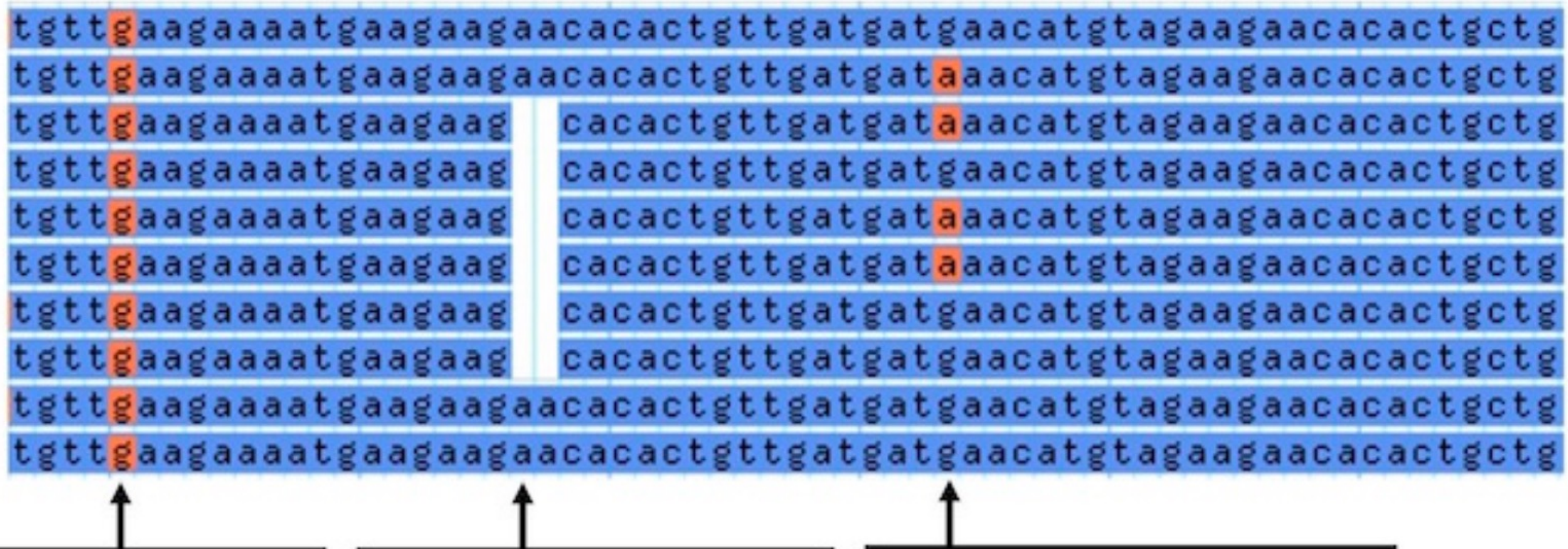
Point Mutations

		Silent	Nonsense	Missense	
				Conservative	Non-conservative
DNA	TTC	TTT	ATC	TCC	TGC
mRNA	AAG	AAA	UAG	AGG	ACG
Protein	Lys	Lys	STOP	Arg	Thr

Important factors to think about when considering SNPs:

- Can a SNP be called based on the data?
- How many of your sample had suffient reads to call a SNP at that position?
- What is the frequency of the SNP? What could affect read frequency?
- How many of your samples have that SNP? Minor/major allele frequency.

Isolate X aligned sequencing reads



```
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag  cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
```

Nucleotide position = 1567
Aligned reads = 10
Reference = A
Isolate reads = G (10)
Read Frequency = 100%

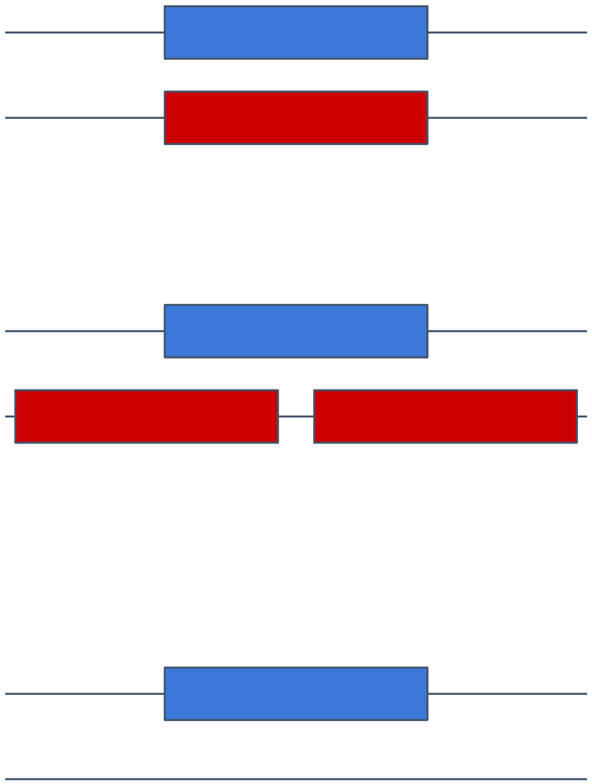
Nucleotide position = 1583
Aligned reads = 4
No Call: <5 reads

Nucleotide position = 1600
Aligned reads = 10
Reference = G
Isolate reads = A (4) and G (6)
Read Frequency G = 60%
Read Frequency A = 40%

Copy Number Variations

What are Copy Number Variations (CNVs)?

A phenomenon in which a region of the genome appears a different number of times in different individuals



No CNV

Duplication

Deletion

Types of CNV

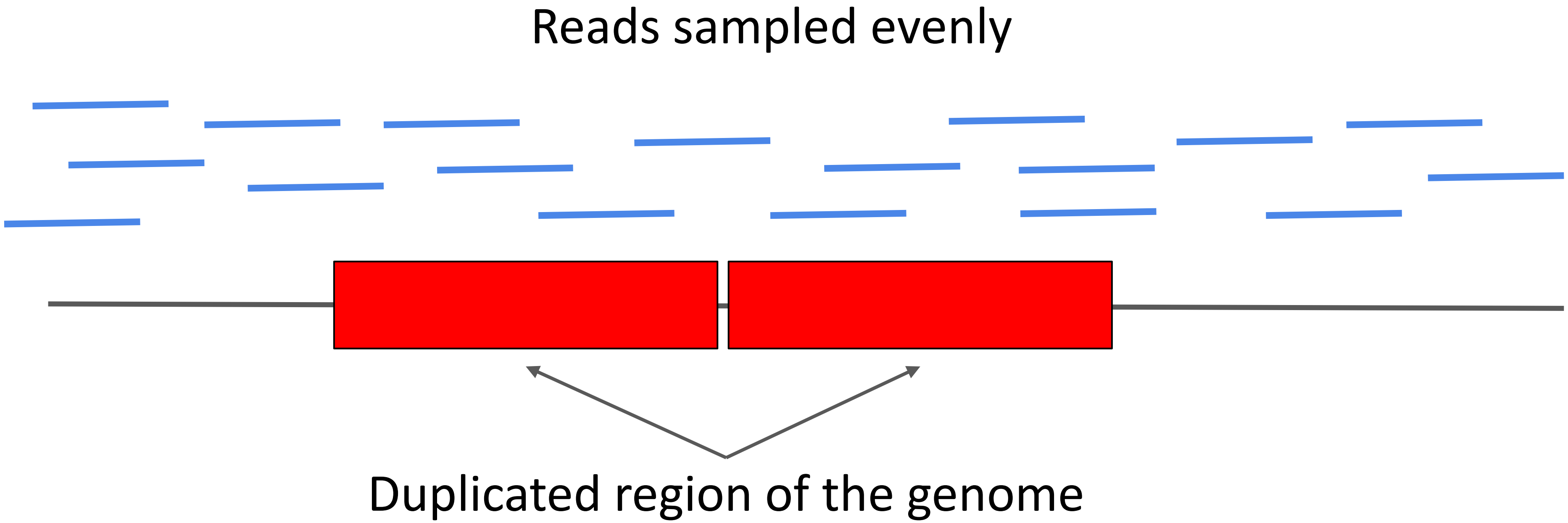
- **Polyploidy**
  - Duplication of the whole genome
  - Common in plants
  - Happened twice in the vertebrate lineage leading to humans
- **Aneuploidy**
  - Duplication of some but not all chromosomes
  - Common in *Leishmania* and *Trypanosoma cruzi*
- **Segmental duplications**
  - Duplication of part of a chromosome
  - We'll look at an example in *Candida*
- **Gene duplications**
  - Duplication of a single gene
  - *ROP5* in *Toxoplasma gondii*
- **Short CNVs**
  - Duplications of short regions, e.g., CAG repeats in Huntingdon's disease or BRC repeats in *BRCA2*

Types of CNV

- **Polyploidy**
  - Duplication of the whole genome
  - Common in plants
  - Happened twice in the vertebrate lineage leading to humans
- **Aneuploidy**
  - Duplication of some but not all chromosomes
  - Common in *Leishmania* and *Trypanosoma cruzi*
- **Segmental duplications**
  - Duplication of part of a chromosome
  - We'll look at an example in *Candida*
- **Gene duplications**
  - Duplication of a single gene
  - *ROP5* in *Toxoplasma gondii*
- **Short CNVs**
  - Duplications of short regions, e.g., CAG repeats in Huntingdon's disease or BRC repeats in *BRCA2*

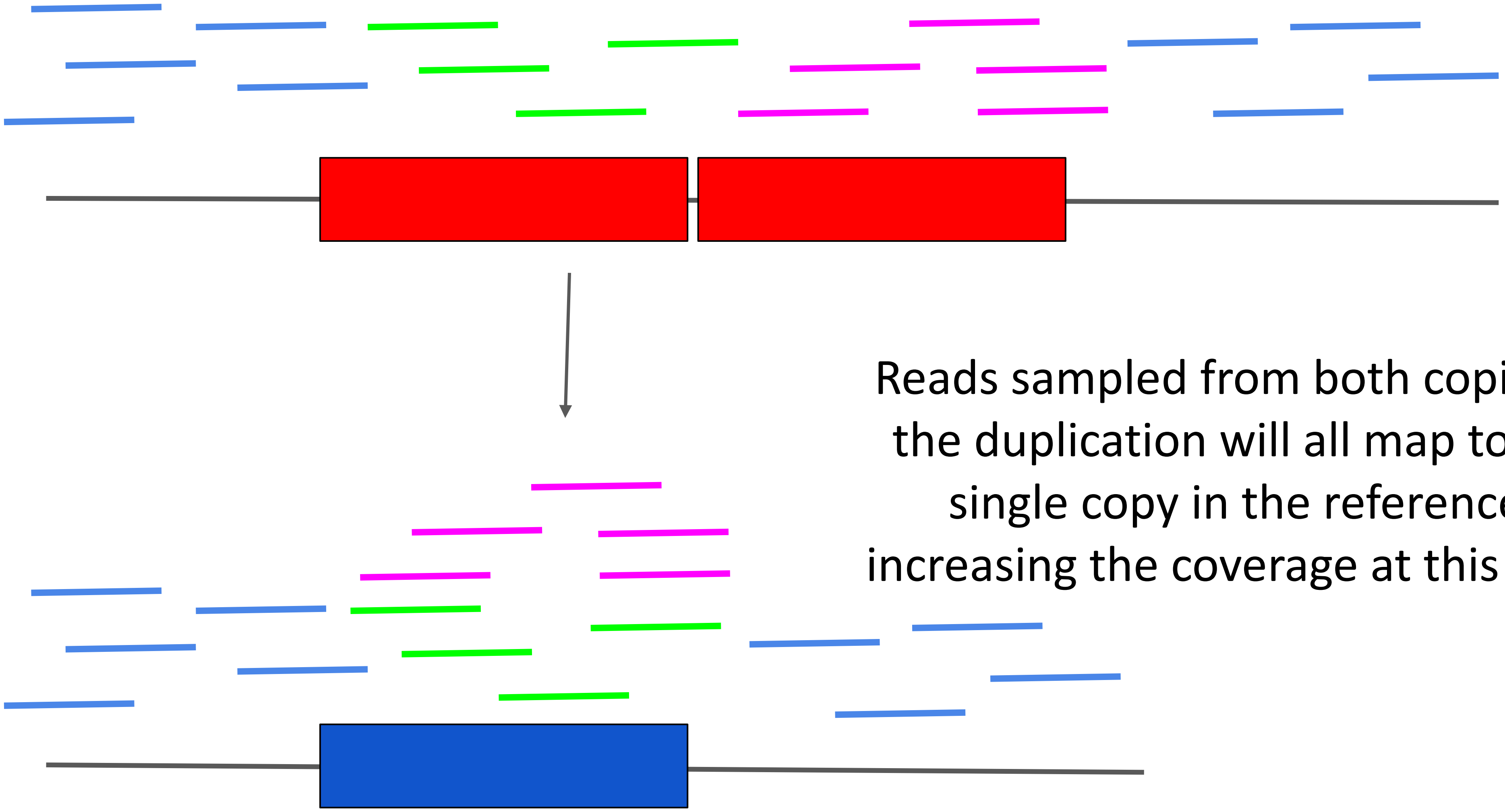
CNVs and Coverage

In whole genome sequencing (WGS) one would expect to sample evenly along the genome resulting in even coverage after alignment



This is subject to minor variation associated with GC content and stochastic variation

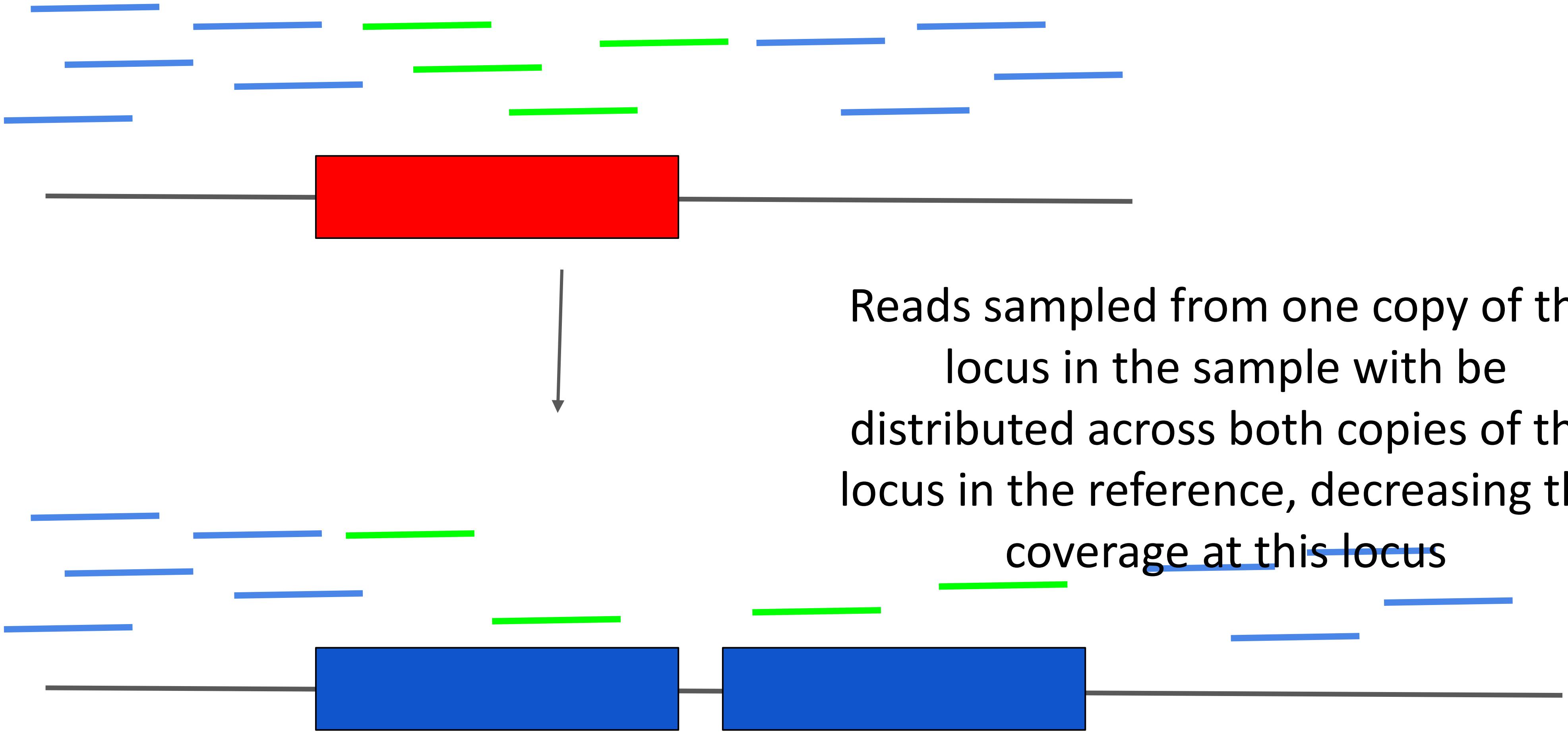
Duplication results in increased coverage



Reads sampled from both copies of the duplication will all map to the single copy in the reference, increasing the coverage at this locus

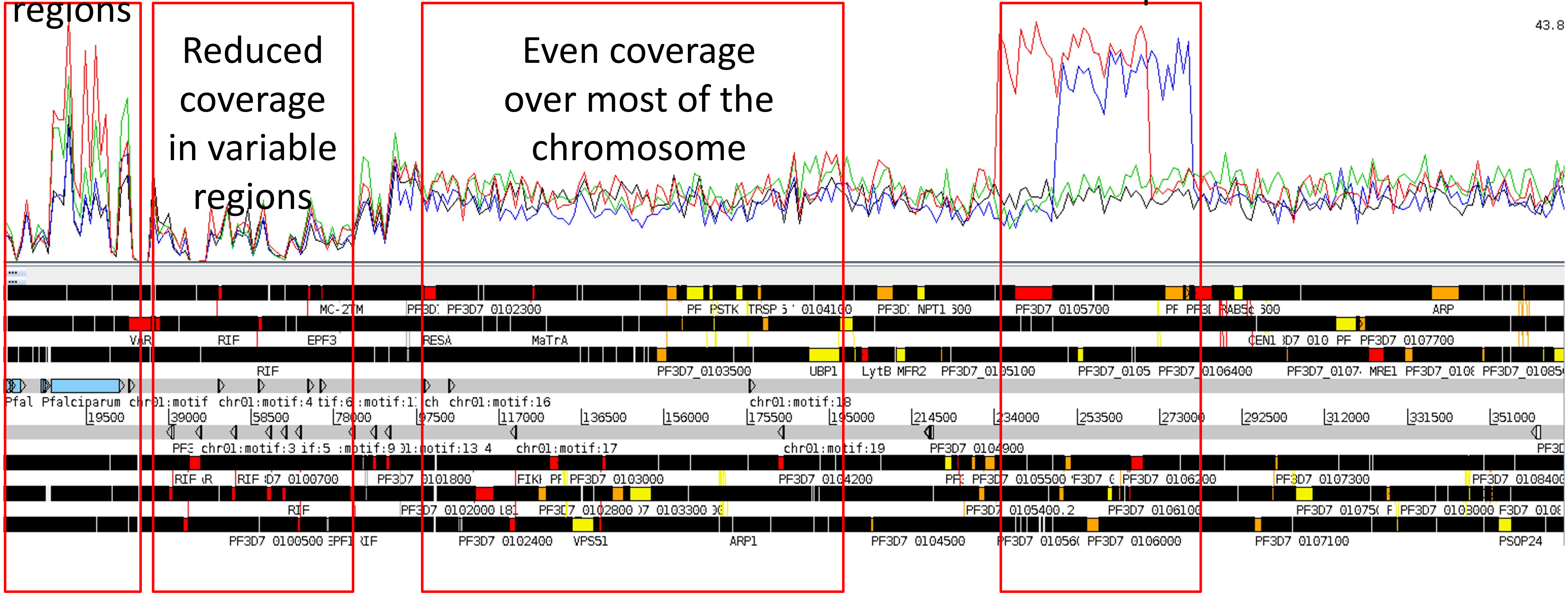
Reference genome without duplication

Deletion results in decreased coverage



Reads sampled from one copy of the locus in the sample with be distributed across both copies of the locus in the reference, decreasing the coverage at this locus

Duplications Associated with Drug Resistance



Variable coverage in repetitive regions

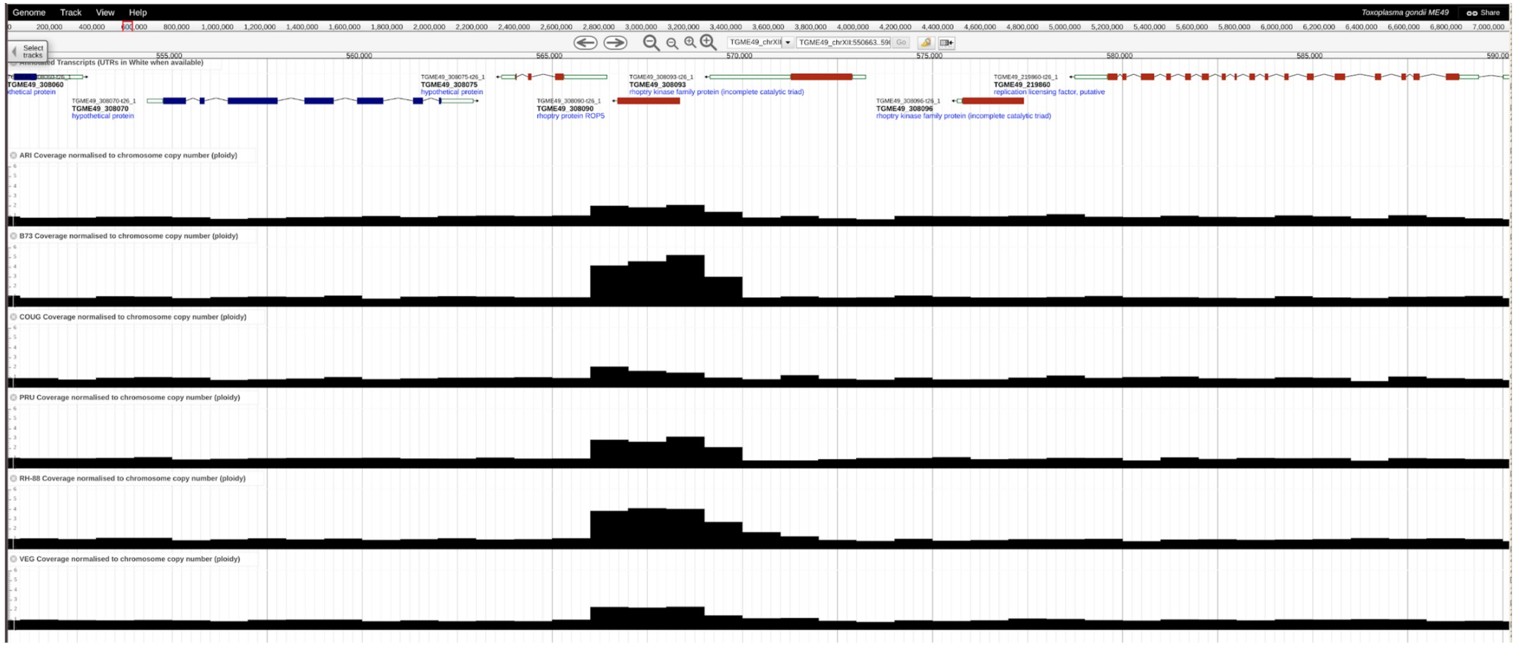
Reduced coverage in variable regions

Even coverage over most of the chromosome

Duplications in two samples

In VEuPathDB

- Search for supernumerary chromosomes in different isolates
- Search for genes with increased or decreased copy number in different isolates
- Explore coverage in JBrowse


Genome
Track
View
Help
Toxoplasma gondii ME49
Share
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
2,000,000
2,200,000
2,400,000
2,600,000
2,800,000
3,000,000
3,200,000
3,400,000
3,600,000
3,800,000
4,000,000
4,200,000
4,400,000
4,600,000
4,800,000
5,000,000
5,200,000
5,400,000
5,600,000
5,800,000
6,000,000
6,200,000
6,400,000
6,600,000
6,800,000
7,000,000
Select tracks
TGME49_chrXII
TGME49_chrXII:550663..59(
Go
555,000
560,000
565,000
570,000
575,000
580,000
585,000
590,00
Annotated Transcripts (UTRs in White when available)
TGME49_308060-t26_1
TGME49_308060
hypothetical protein
TGME49_308070-t26_1
TGME49_308070
hypothetical protein
TGME49_308075-t26_1
TGME49_308075
hypothetical protein
TGME49_308090-t26_1
TGME49_308090
rhoptry protein ROP5
TGME49_308093-t26_1
TGME49_308093
rhoptry kinase family protein (incomplete catalytic triad)
TGME49_308096-t26_1
TGME49_308096
rhoptry kinase family protein (incomplete catalytic triad)
TGME49_219860-t26_1
TGME49_219860
replication licensing factor, putative
ARI Coverage normalised to chromosome copy number (ploidy)
B73 Coverage normalised to chromosome copy number (ploidy)
COUG Coverage normalised to chromosome copy number (ploidy)
PRU Coverage normalised to chromosome copy number (ploidy)
RH-88 Coverage normalised to chromosome copy number (ploidy)
VEG Coverage normalised to chromosome copy number (ploidy)
6
5
4
3
2