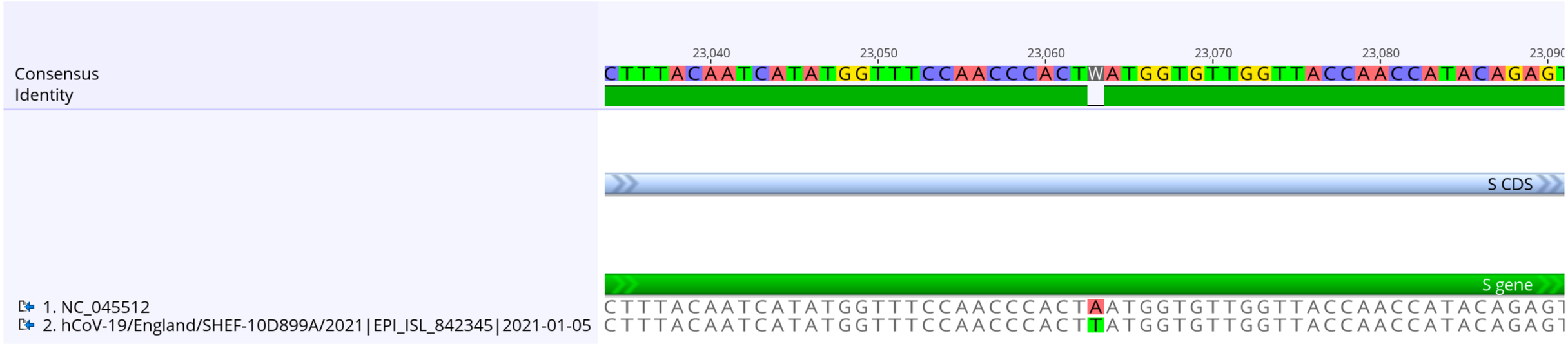
Detecting and quantifying variants

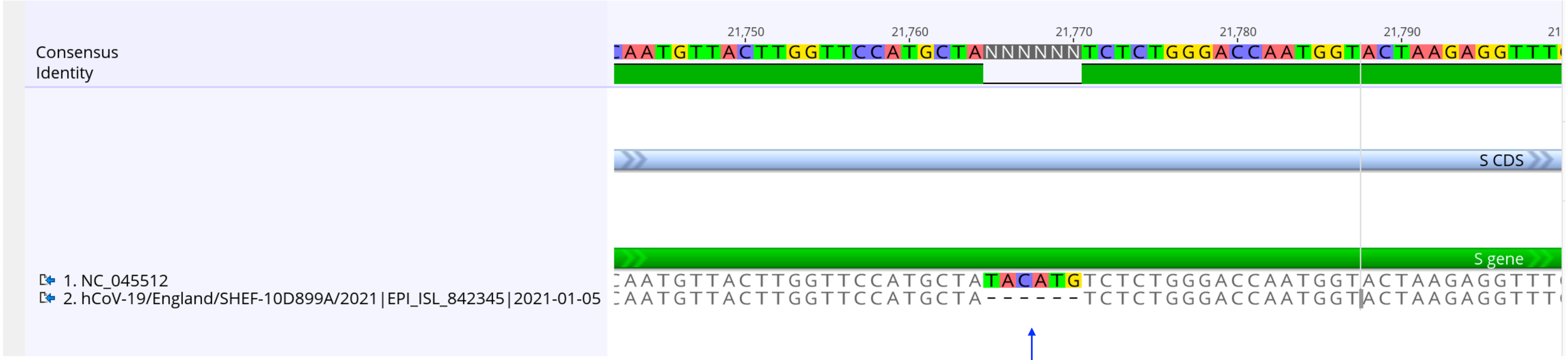
Mark Stenglein, GDW Workshop

A variant is a difference between a sequence and some reference sequence



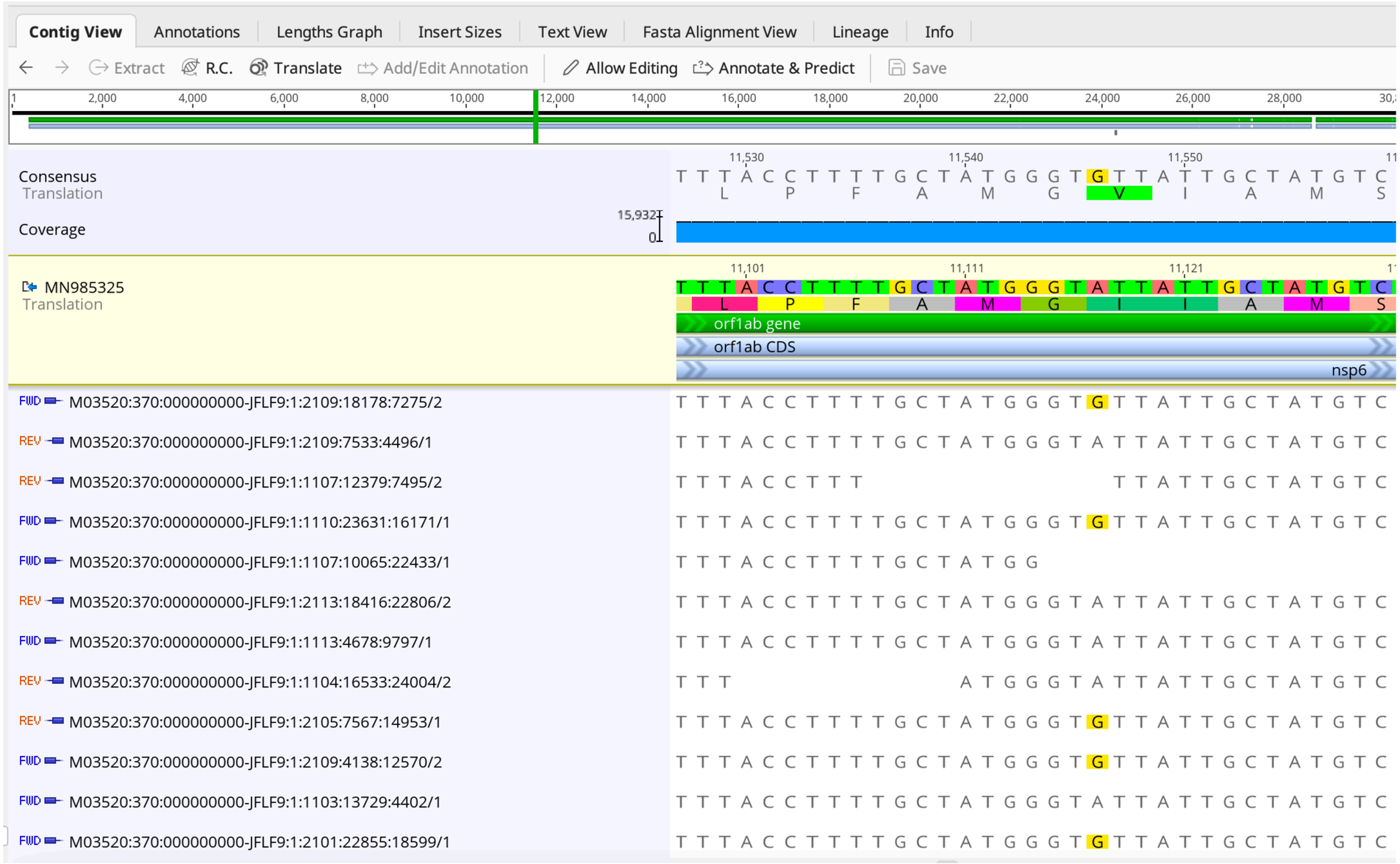
A variant in the SARS-CoV-2 spike sequence relative to the original Wuhan reference sequence

Variants can be single nucleotide variants (SNVs/SNPs) or insertions/deletions and other types of “structural variation



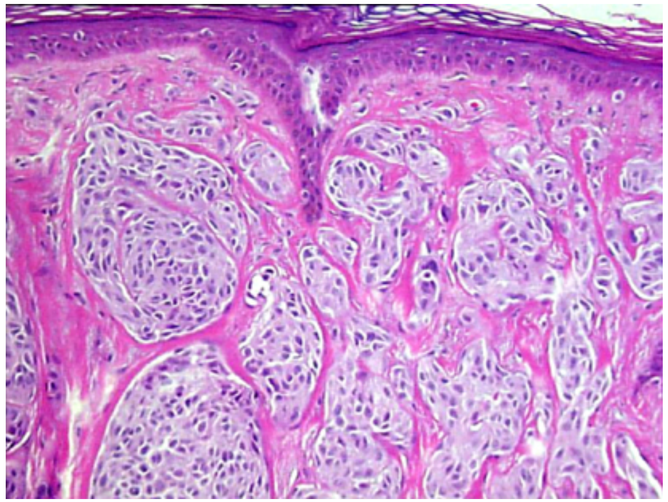
A 2 codon deletion in the SARS-CoV-2 spike sequence relative to the original Wuhan reference sequence

Variants can be detected in reads as mismatches relative to the reference sequence. The fraction of reads with the variant estimates the variant's frequency in a population

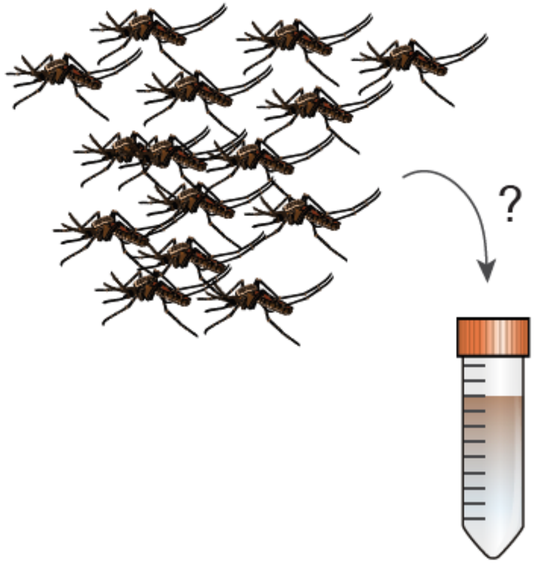


Sub-consensus genomic variation is important in a variety of contexts

Rare somatic variants in cancer (cancer subclones)



Population genomics using pools of individuals (Pool-Seq)



Intrahost viral variation

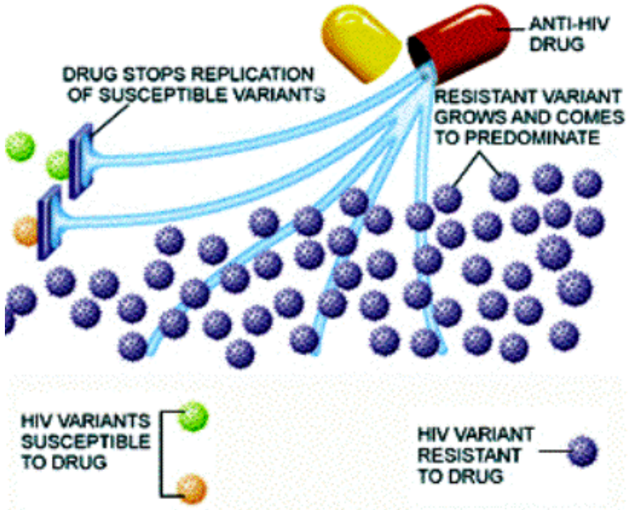
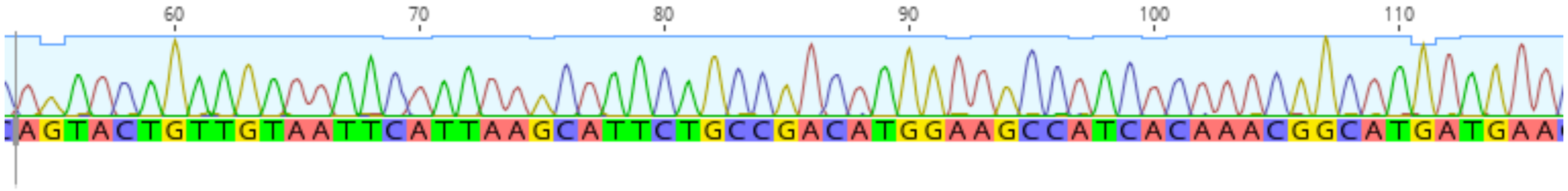
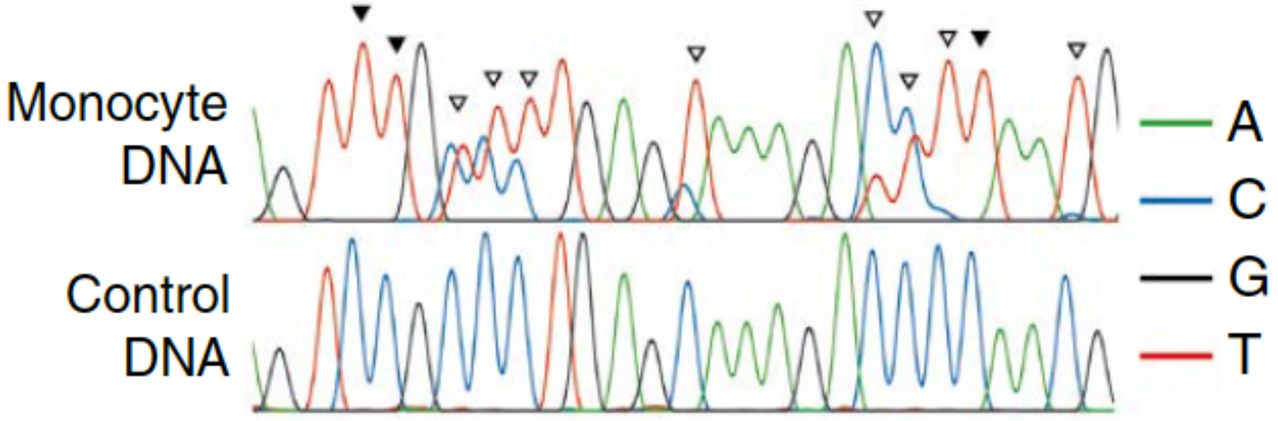


image: Magro et al (2006) Modern Path.

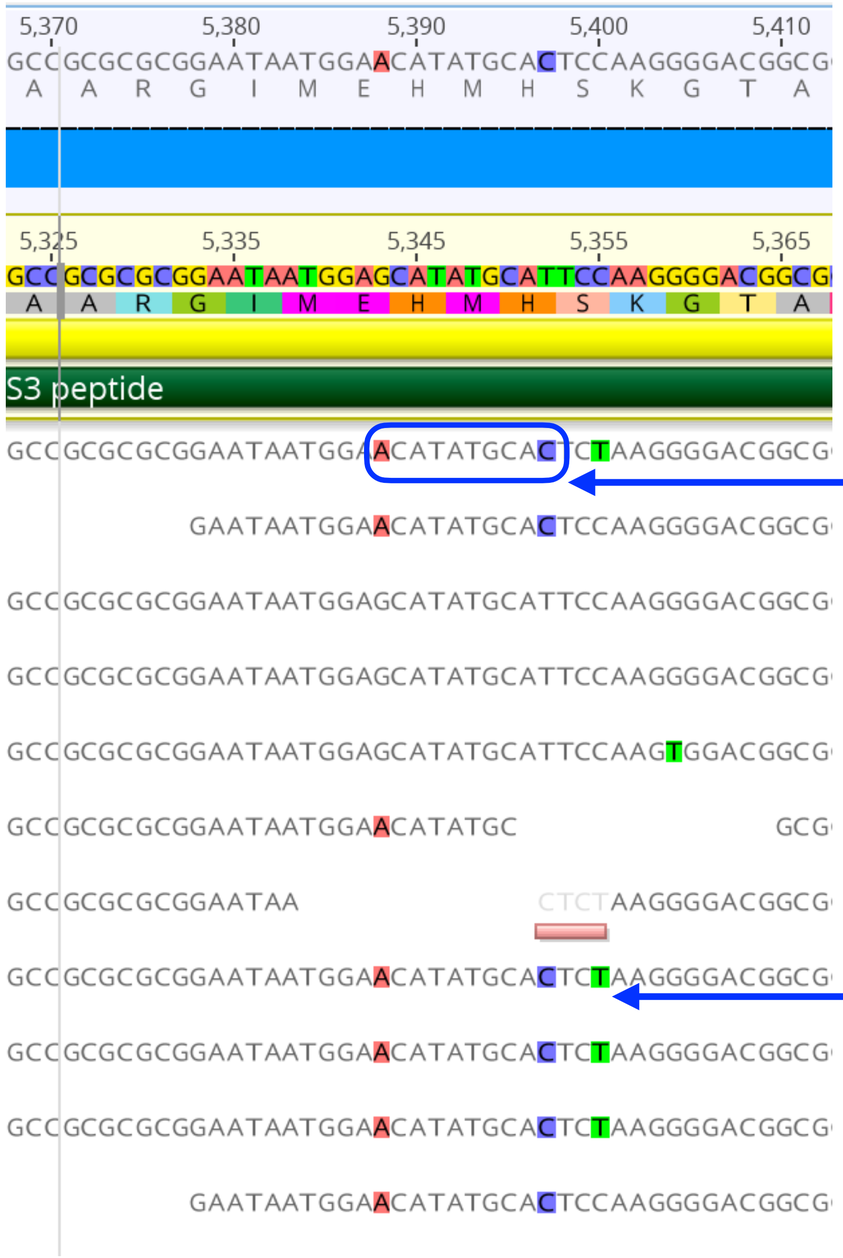
Sanger sequencing typically produces consensus sequence





It is possible to analyze chromatograms to obtain variant frequencies - but it's difficult to quantify and not really done

Goal: identify variants, their frequencies, and potential functional impact



Consensus sequence

Reference sequence

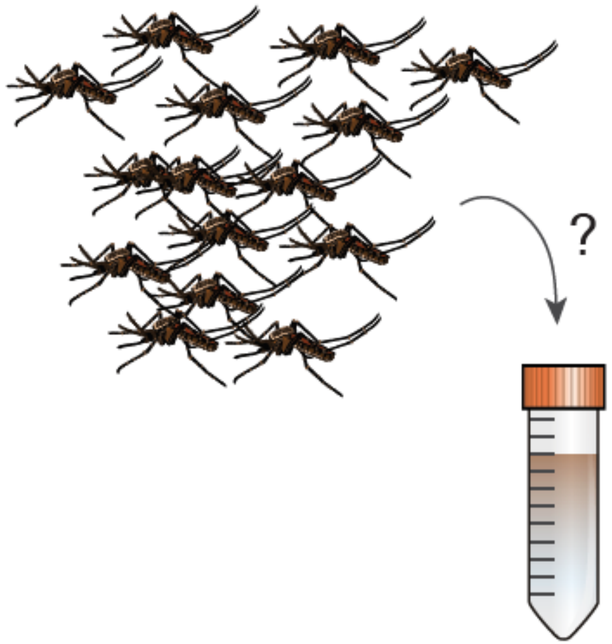
These 2 variants are at a >50% allele frequency and so are consensus changing variants

They are also both synonymous mutations

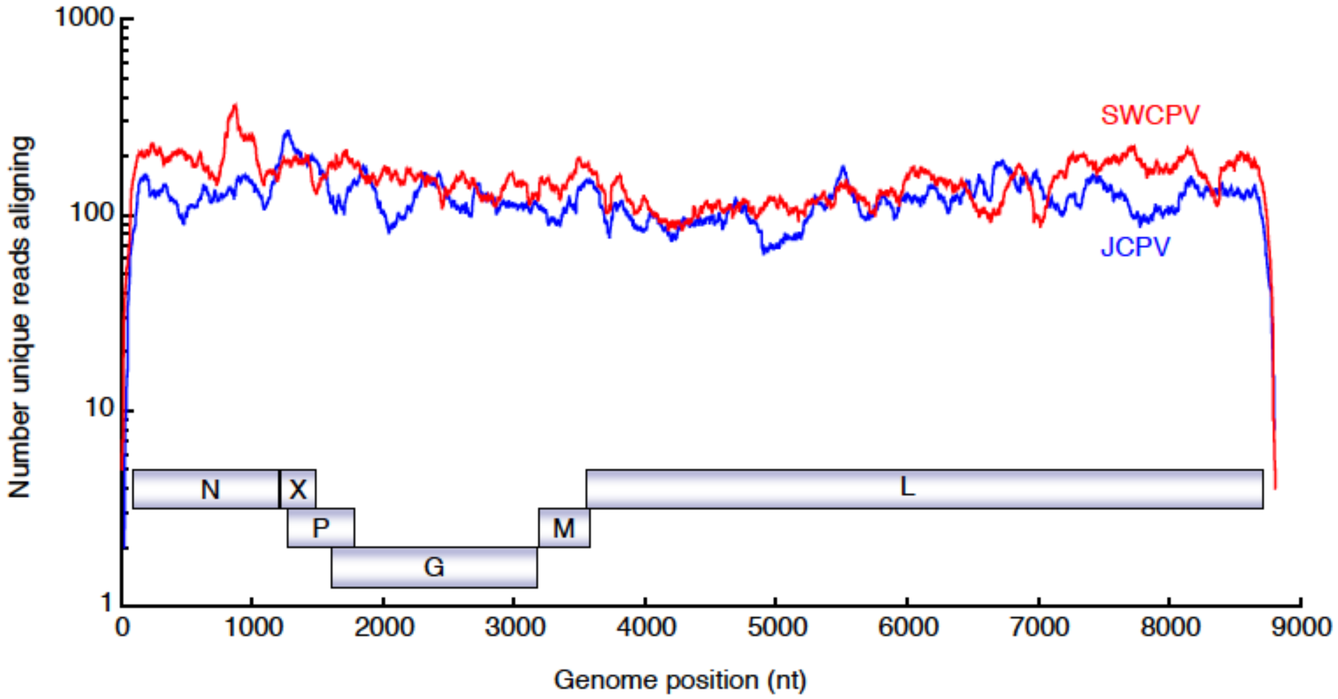
This T's allele frequency is 42% -> no consensus change

Biological and technical limitations to the ability to detect rare variants

Pool size could limit the ability to detect rare variants



unlikely to observe variants with frequency < 1% in these datasets

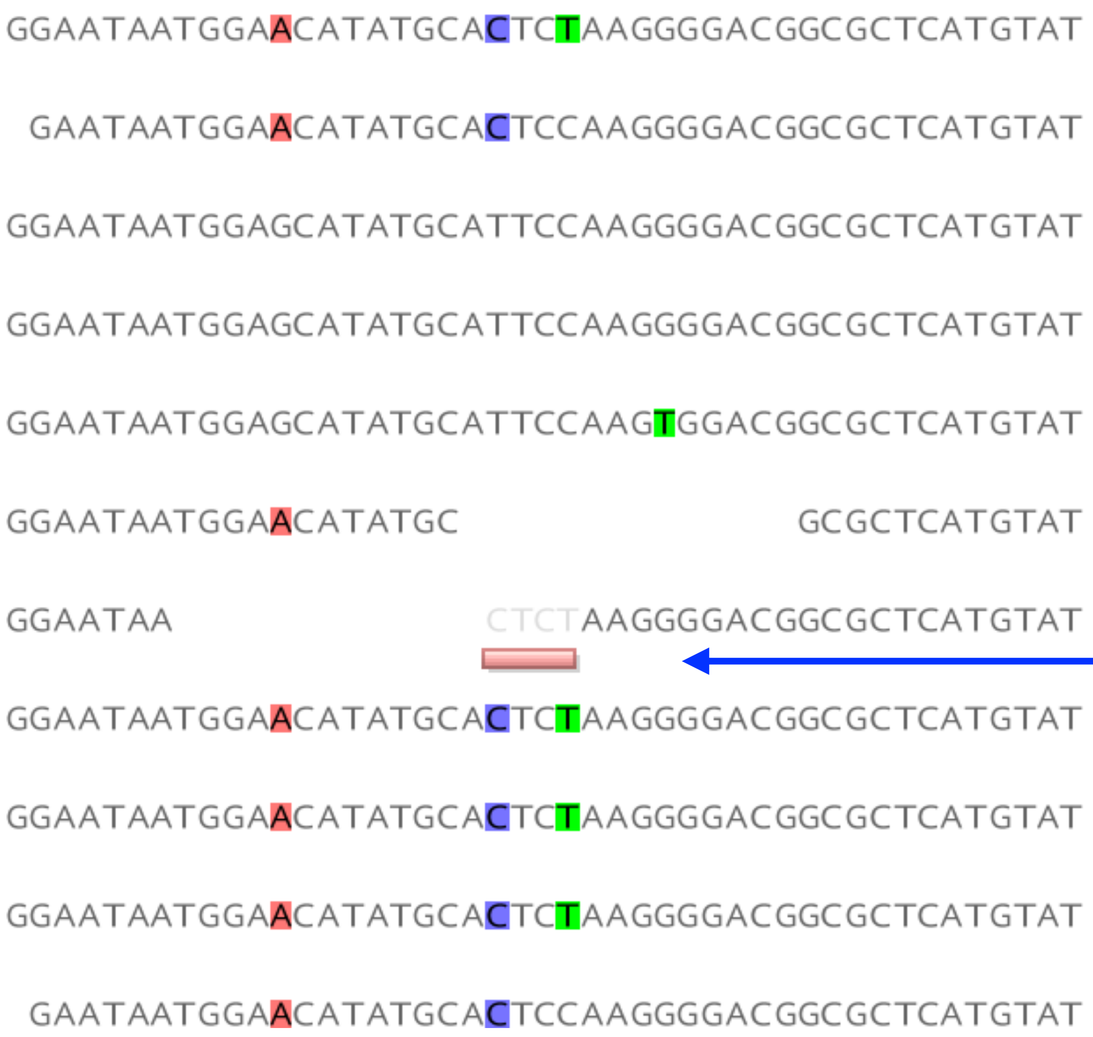


carpet python bornaviruses

Distinguishing sequencing errors from true rare variants can be a challenge

sequencing error, or real low frequency variant?

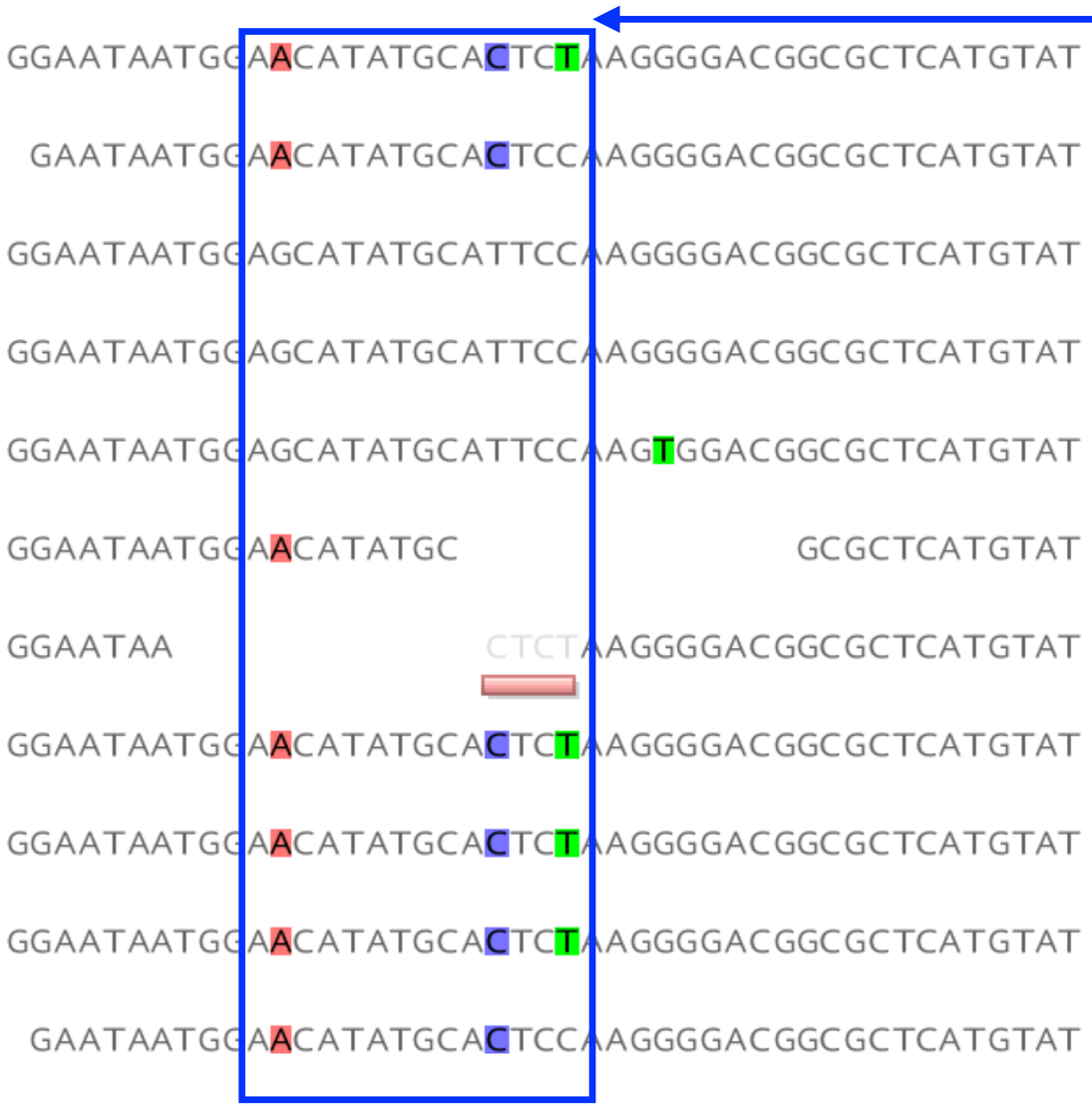
Variant calling is also sensitive to mapping



These bases were soft-trimmed (not aligned), but they support variant basecalls

Different mapping software could well produce different results.

Another issue is linking or ‘phasing’ variants (haplotype reconstruction)



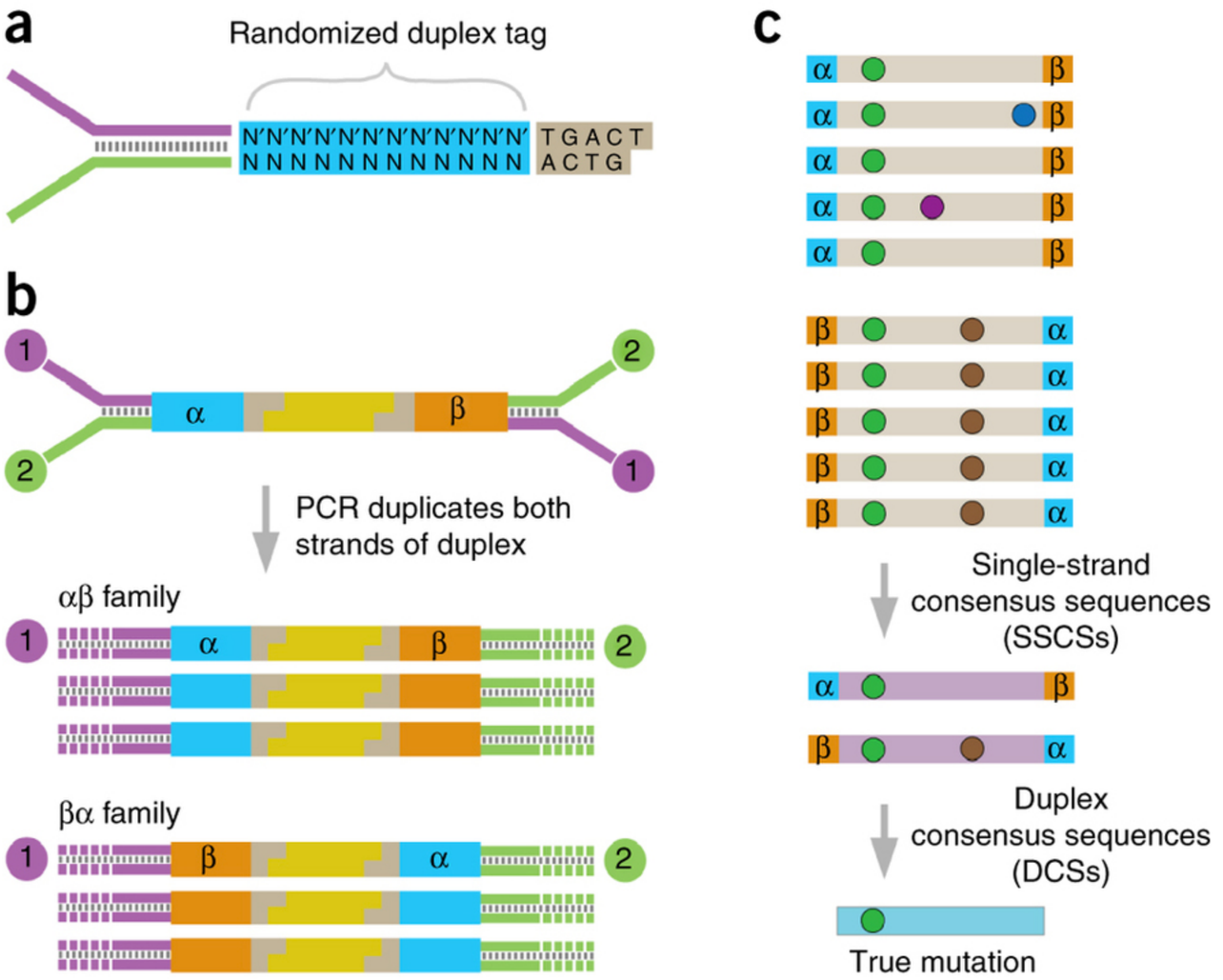
3 haplotypes evident here

G T C [reference sequence]
A C C [2 mutations]
A C T [3 mutations]

Much harder to link distant variants using short read data

Clever methods have been developed to get beyond the limit of detection due to sequencing errors

Some protocols take advantage of PCR duplicates to sequence the same original molecule multiple times



Kennedy et al (2014) Nature Protocols

The good news!

You don’t necessarily or even often need linked variants or ultra low frequency variants to infer population genetic parameters (or otherwise answer your question of interest)

A typical workflow for variant identification

Sample: DNA/RNA isolation

↓

Library prep / sequencing

↓

Mapping to a reference genome or an assembly

↓

Variant calling

↓

Downstream analysis of variants
(e.g. using SNPEff)

The standard format for variant data is the vcf file (variant call format)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://github.com/samtools/hts-specs

Let’s call some variants!