



wellcome
connecting
science

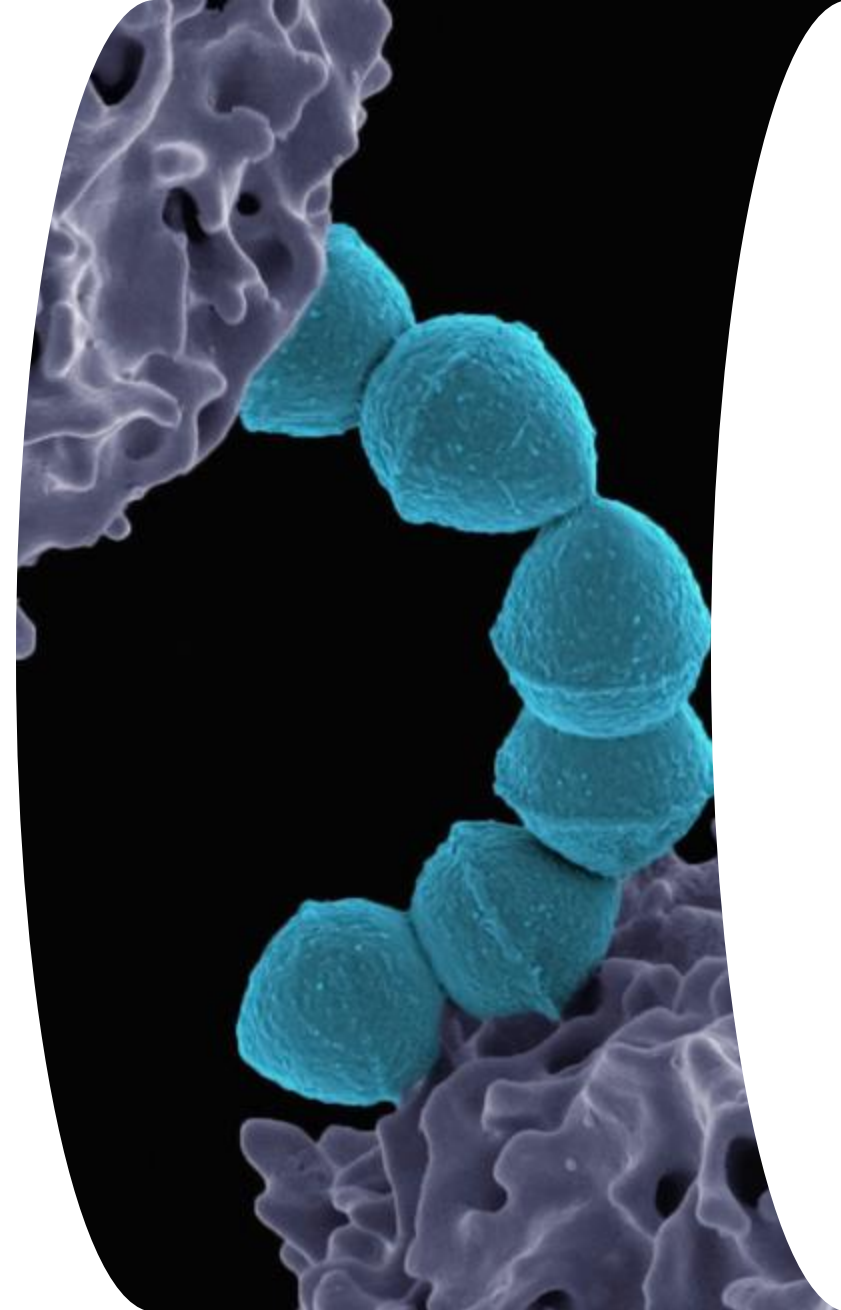
ACORN.

VARIANT CALLING: Detecting Mutations in Bacterial Genomes

15-20 September 2024

KEMRI, Kilifi, Kenya

Arun Gonzales Decano
Senior Bioinformatician



Background

- Next-generation sequencing (NGS) technologies have transformed genomic research by providing high-throughput sequencing capabilities with unparalleled speed and accuracy.
- The generated sequencing reads hold the key to unlocking valuable insights into genetic variations and disease mechanisms.
- The raw sequencing data from NGS platforms requires sophisticated computational analysis to extract meaningful information.

<https://pharmafeatures.com/cracking-genetic-ciphers-secondary-ngs-analysis/>

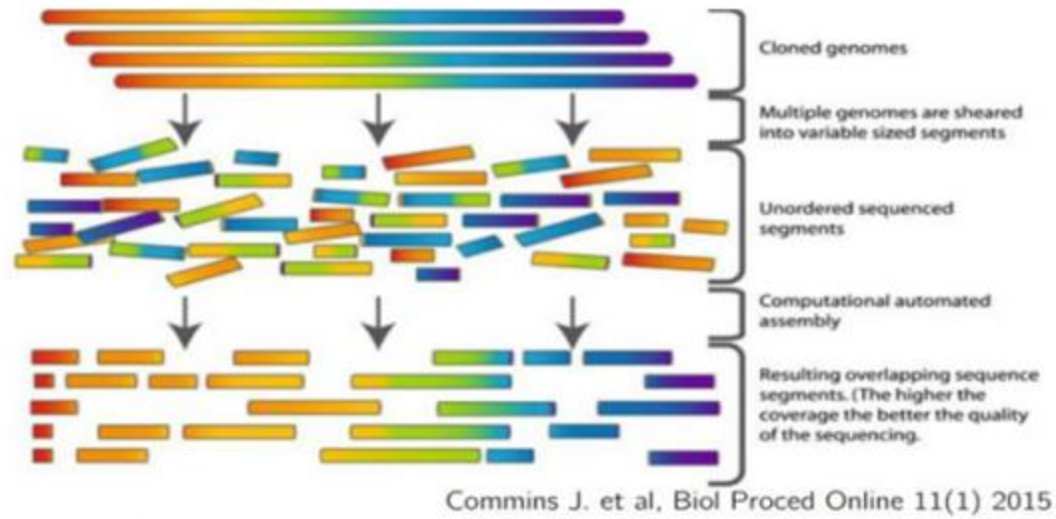
Background

- Secondary or Downstream Analysis
- Read alignment and variant calling are *critical steps that bridge the gap between raw data and actionable genetic insights*.
- By understanding the complexities and nuances of these analysis techniques, researchers can make informed decisions to harness the full potential of NGS data.

<https://pharmafeatures.com/cracking-genetic-ciphers-secondary-ngs-analysis/>

Learning Objectives

- Learn the principles and methods of detecting genetic variations in bacterial genomes.
- Explore tools and software for variant calling
- Gain knowledge of variant calling and its applications.
- Learn about challenges in variant calling.



Mapping to reference sequence

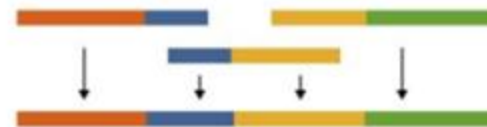
Recreate the genome with using prior knowledge as reference



Mapping is as good as reference used

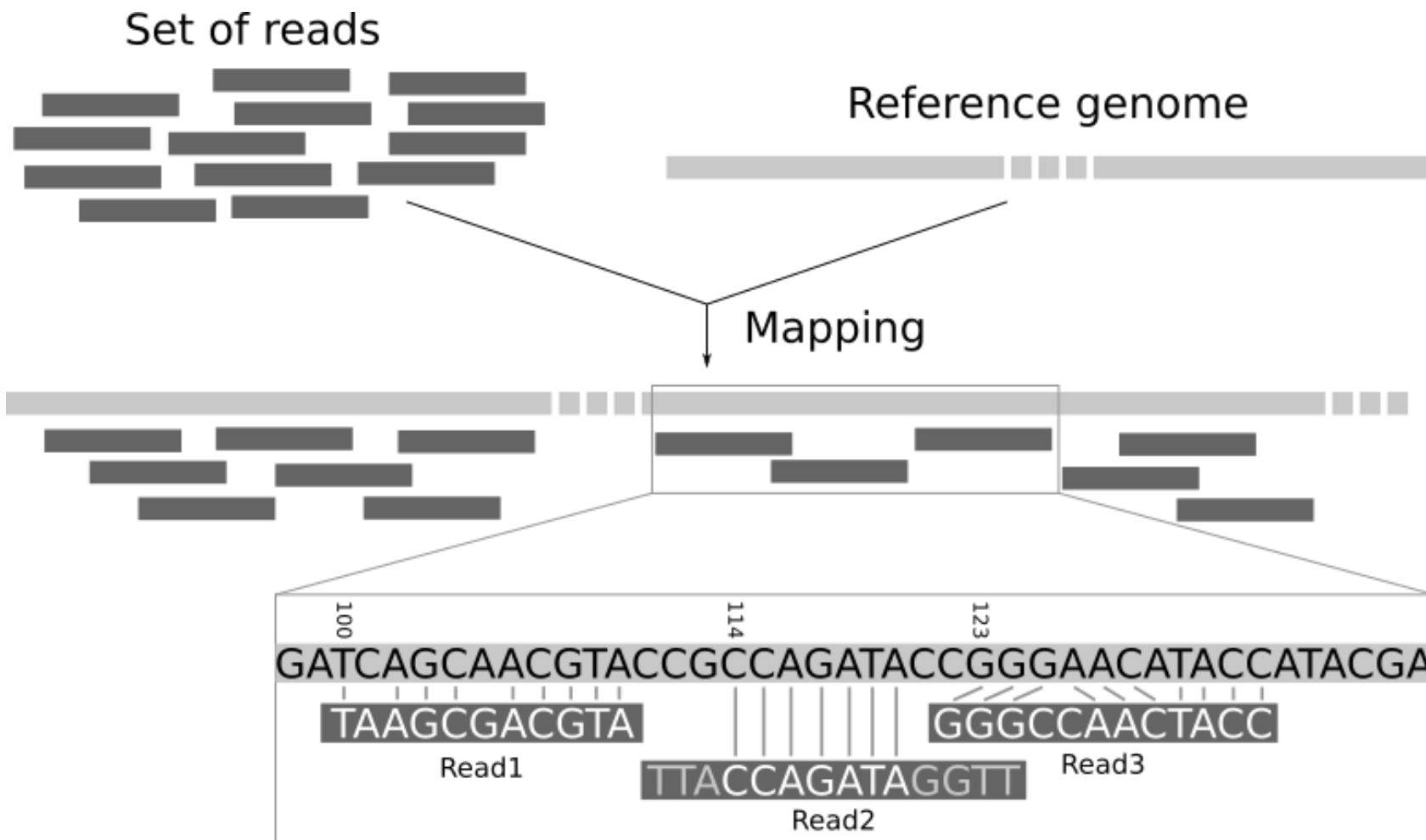
De Novo assembly

Recreate the genome with no prior knowledge



Problem with repeated regions, high coverage and long

<https://pharmafeatures.com/cracking-genetic-ciphers-secondary-ngs-analysis/>

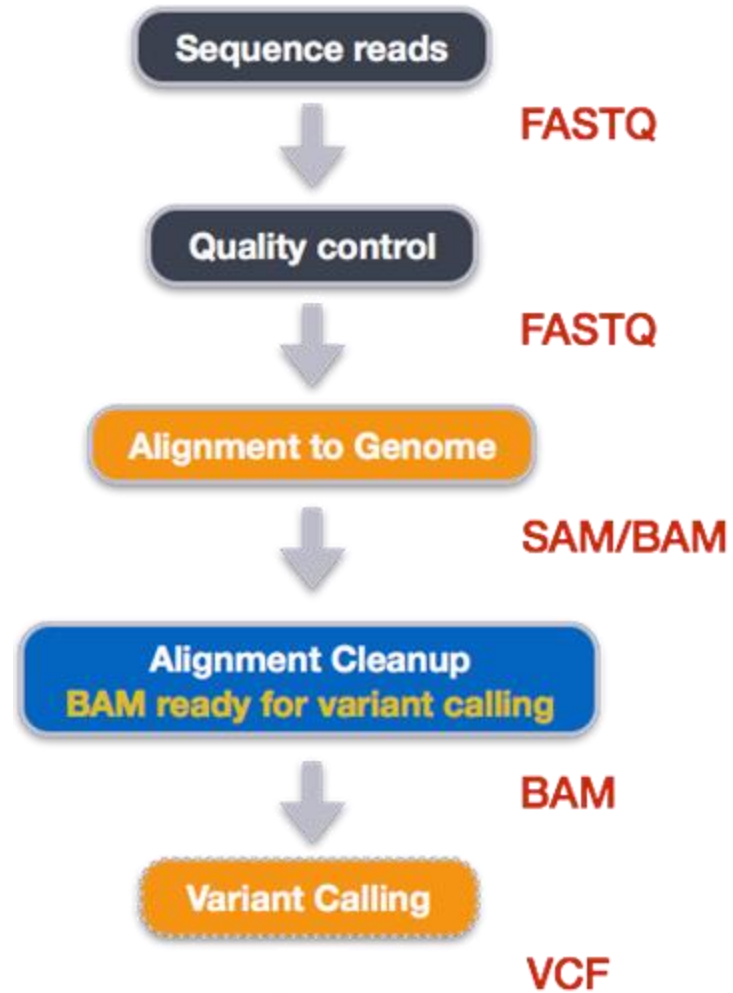


Review Read Mapping principles from our earlier ACORN course:

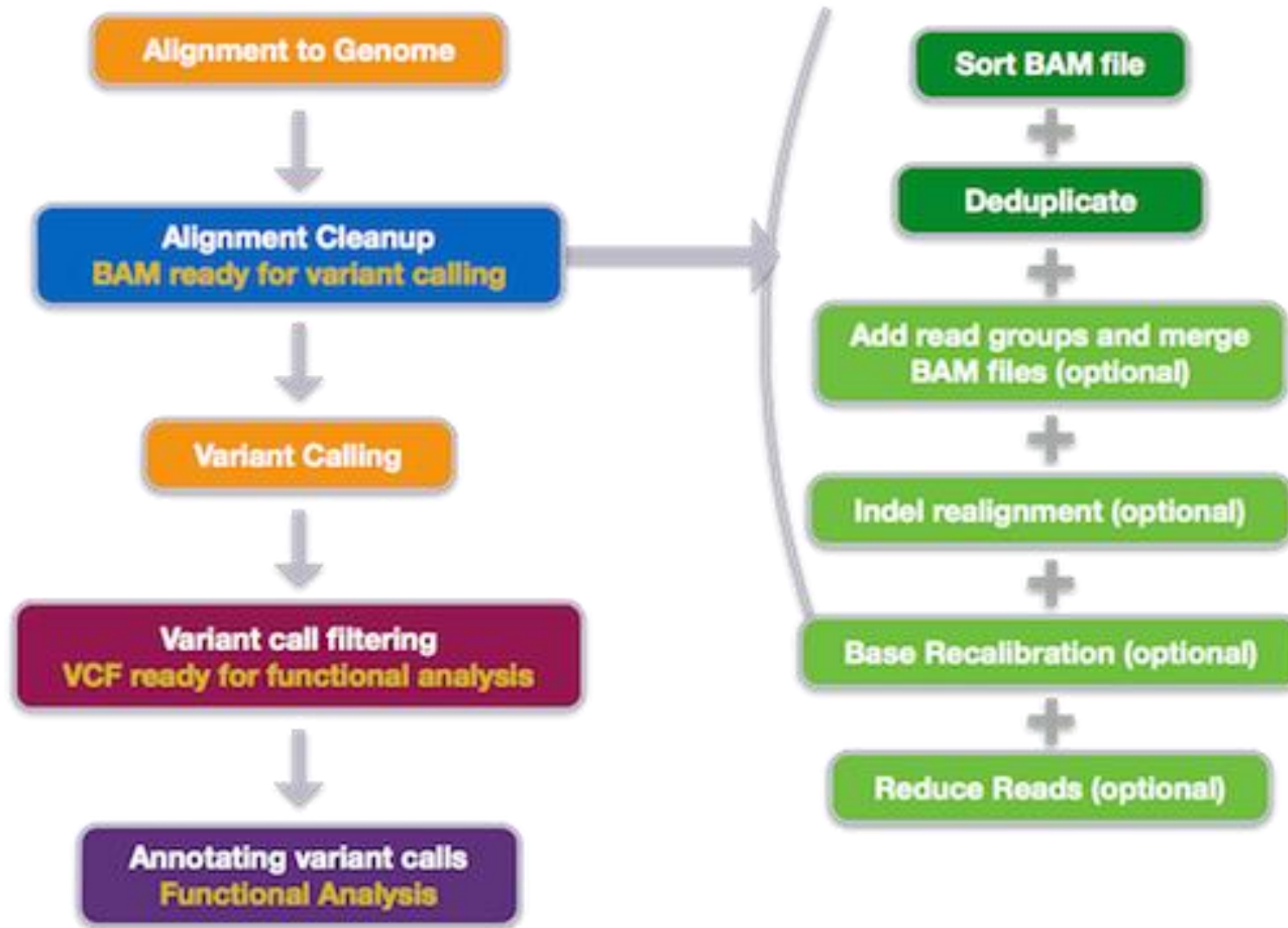
<https://zenodo.org/records/12805691>

<https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html>

Variant Calling Workflow

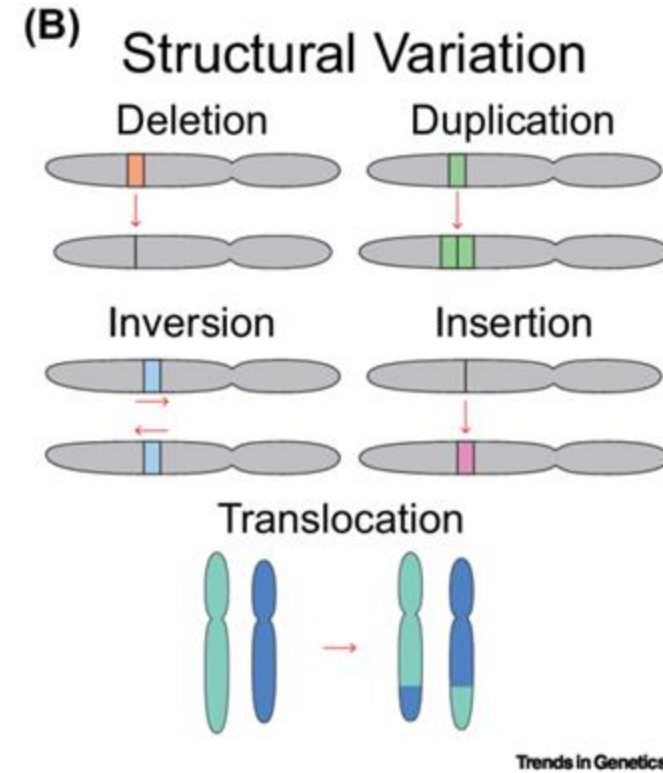
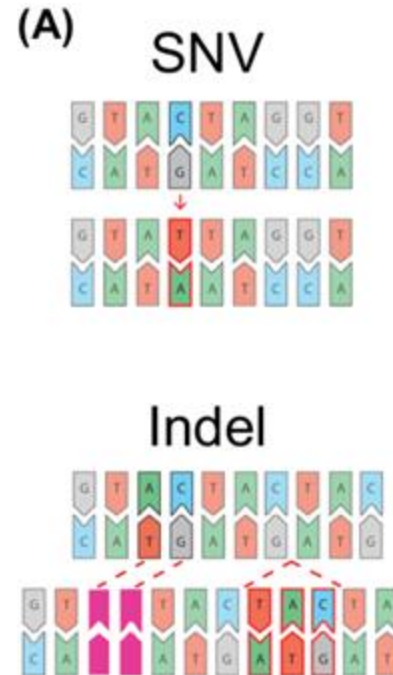
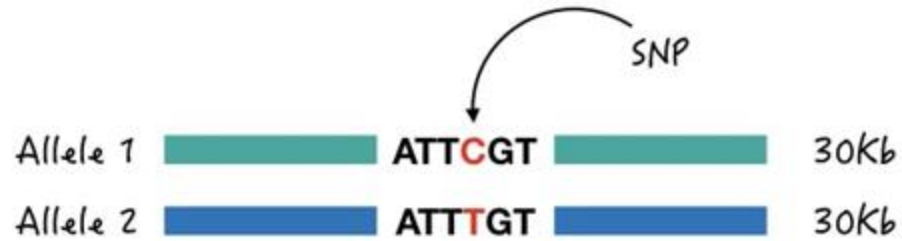


Variant Calling Workflow



https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/01_alignment.html

SNV/SNP vs SV



SNPs

Transitions (Ti): Point mutations where a purine (adenine, A, or guanine, G) is replaced by another purine, or a pyrimidine (cytosine, C, or thymine, T) is replaced by another pyrimidine. Thus, there are two types of transitions:

- **A ↔ G** (purine to purine)
- **C ↔ T** (pyrimidine to pyrimidine)

Transversions (Tv): Point mutations where a purine is replaced by a pyrimidine or vice versa. There are four possible transversions:

- **A ↔ C**
- **A ↔ T**
- **G ↔ C**
- **G ↔ T**

$$\text{Ti/Tv ratio} = \frac{\text{Number of Transitions (Ti)}}{\text{Number of Transversions (Tv)}}$$

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.

$$T_i/T_v > 1$$

Evolutionary Implications

- **Indicator of Evolutionary Conservation:** A high T_i/T_v ratio might indicate that a sequence has undergone relatively few mutational changes over time or that there is strong selective pressure to maintain certain genetic sequences. In such cases, transitions are more likely to be retained because they often result in less functional change (e.g., synonymous substitutions or conservative amino acid changes).
- **Reflects Mutation and Repair Biases:** The ratio greater than 1 suggests that the natural mutation processes and DNA repair mechanisms in the organism tend to favor the occurrence or retention of transitions over transversions.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.

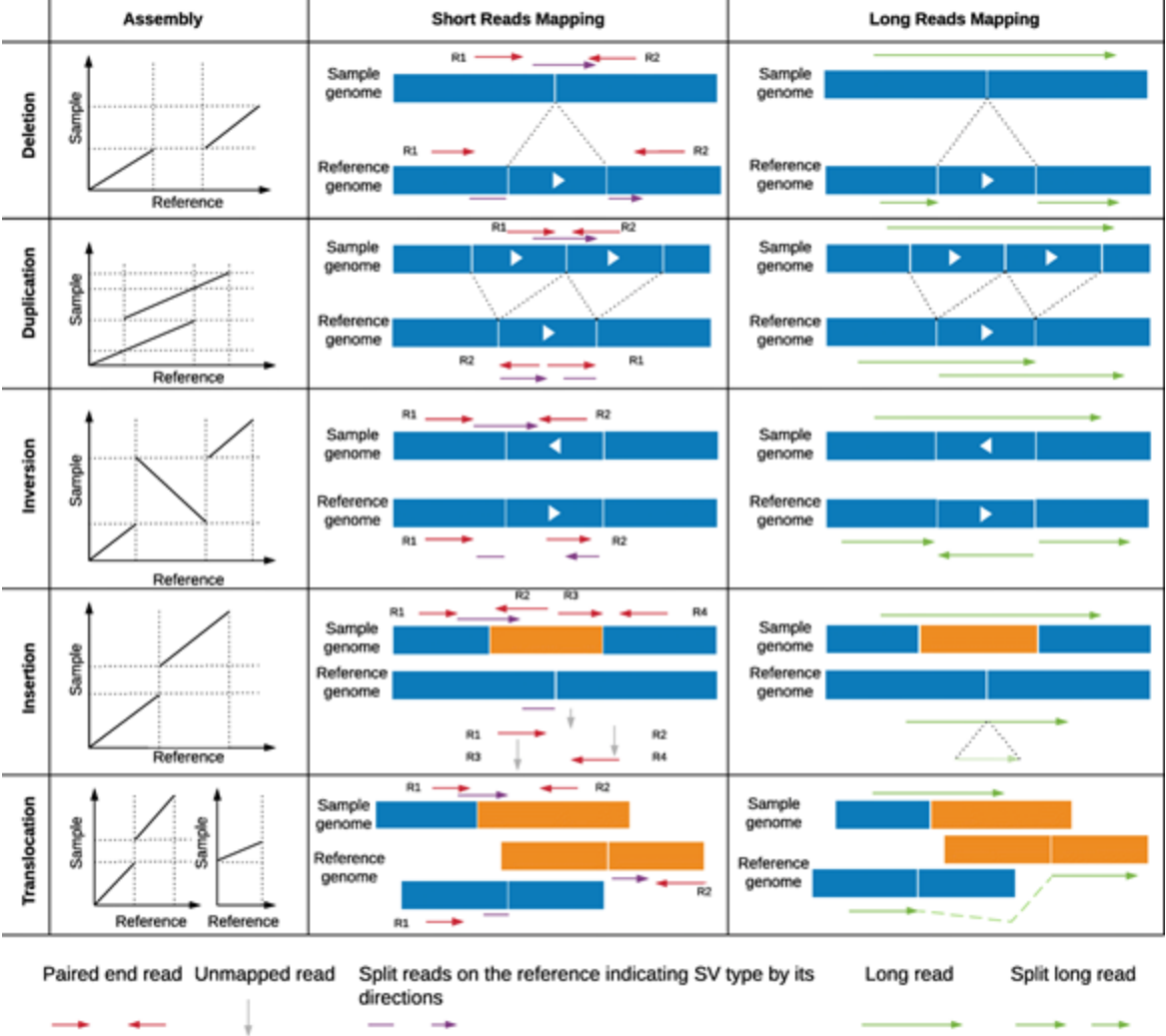
$Ti/Tv < 1$

Evolutionary Implications

- **Unusual Evolutionary Scenarios:** A Ti/Tv ratio less than 1 might be seen in unusual evolutionary scenarios, such as:
 - **Species or Population with Unusual Mutation Patterns:** Some organisms or specific populations may have mutational biases that lead to a higher rate of transversions. For example, some viruses or bacteria that have been exposed to certain mutagens may show such patterns.
 - **Extreme Selective Pressures:** Environments with extreme selective pressures (e.g., high levels of radiation or chemical exposure) may lead to increased transversion rates if those mutations provide some survival advantage or occur more frequently due to specific types of DNA damage.
- **Recent Rapid Evolution or Divergence:** If the sequence under study represents a lineage that has recently undergone rapid evolution or divergence, the Ti/Tv ratio might temporarily be lower than expected due to unique evolutionary pressures or recent changes in mutational processes.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.

De novo assembly, short-read and long-read mapping approaches to identify structural variants



Variant Calling Workflow: Summary

1. Get Illumina / Nanopore data – processed BAM files
2. Call SNPs & indels with **FreeBayes**
3. Call SNPs & indels with **BCFtools**
4. Screening SNPs & indels with **BCFtools** & **vcfutils.pl**
5. Visualise data with **IGV**
6. Evaluate VCF files

Metrics

Tech	Illumina				Nanopore			
Sample	ERR9975619	ERR9975619	ERR9975619	ERR9975619	ERR6319263	ERR6319263	ERR6319263	ERR6319263
Mapper	Minimap2	Minimap2	Bwa	Bwa	Minimap2	Minimap2	Bwa	Bwa
SNPcaller	BCFtools	FreeBayes	BCFtools	FreeBayes	BCFtools	FreeBayes	BCFtools	FreeBayes
Ti	41	36	43	37	24	1438	24	1469
Tv	15	18	17	18	14	104	14	100
Ti/Tv	2.7	2.0	2.5	2.1	1.7	13.8	1.7	14.7
Insertion	2	3	2	2	4	427	3	410
Deletion		8		8		1076		1053
SNP	56	56	60	57	38	1681	38	1053
Complex	NA	3	NA	3	NA	102	NA	98

Example results for 2 samples with differing mappers & SNP callers

Evaluate VCF files

Variant Call Format (VCF) files are the standard way of representing mutation data

Developed originally in 2008-10 for 1000 Genomes Project

VCF files vary – exact info depends on SNP caller used to make them, and era (VCF file versions)

Evaluate VCF files

VCF header

Body

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Evaluate VCF files

The header info has some valuable info, eg for depth DP:

```
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth of reads passing MAPQ filter">
```

```
##INFO=<ID=AC,Number=R,Type=Integer,Description="Total number  
of alleles in called genotypes">
```

Evaluate VCF files

First 34 lines = column info for SNPs
lines 35+ = SNPs (one per line)

"DP" at start of lines with SNPs

```
grep -c "DP" *.vcf
```

Sample

#SNPs

G3682.vcf

6

G3735.vcf

18

G3771.vcf

18

NoName.vcf

17

Visualisation: IGV

Download IGV (with Java)
Install it on your own
computer

See: <https://software.broadinstitute.org/software/igv/download>

The screenshot shows the 'Downloads' page for the Integrative Genomics Viewer (IGV) on the Broad Institute's software website. The page layout includes a top navigation bar with the URL 'software.broadinstitute.org/software/igv/download'. Below this is a sidebar with the IGV logo and a menu containing links to Home, Downloads, Documents, IGV User Guide, File Formats, Tutorial Videos, Hosted Genomes, FAQ, Release Notes, Credits, and Contact. The main content area features a 'Downloads' section with a note about IGV-Web, a prominent 'Install IGV 2.16.0' button, and links to Release Notes. It also includes specific instructions for 'Users of the new M1 Mac', 'Linux users', and 'About log4j'. At the bottom, there are two large yellow buttons for downloading the 'IGV MacOS App' and 'IGV for Windows', both noting that 'Java included'. A search bar and a copyright notice '© 2013-2021' are also visible.

software.broadinstitute.org/software/igv/download

Home » Downloads

Downloads

Did you know that there is also requires no downloads? See [ht IGV-Web](#).

Install IGV 2.16.0

See the [Release Notes](#) for what

Users of the new M1 Mac: you run IGV with your own Java

Linux users: The 'IGV for Lin Linux. See [their list of supported line IGV for all platforms](#)' and us

About log4j: IGV versions 2. recommend using version 2.11.

IGV MacOS App
Java included

IGV for Windows
Java included

Search website

search

© 2013-2021

Visualisation: IGV

Alternatively:

Install it locally in your server area:

```
wget https://data.broadinstitute.org/igv/projects/downloads/2.16/IGV_2.16.1.zip
```

```
unzip IGV_2.16.1.zip
```

```
cd IGV_2.16.1/
```

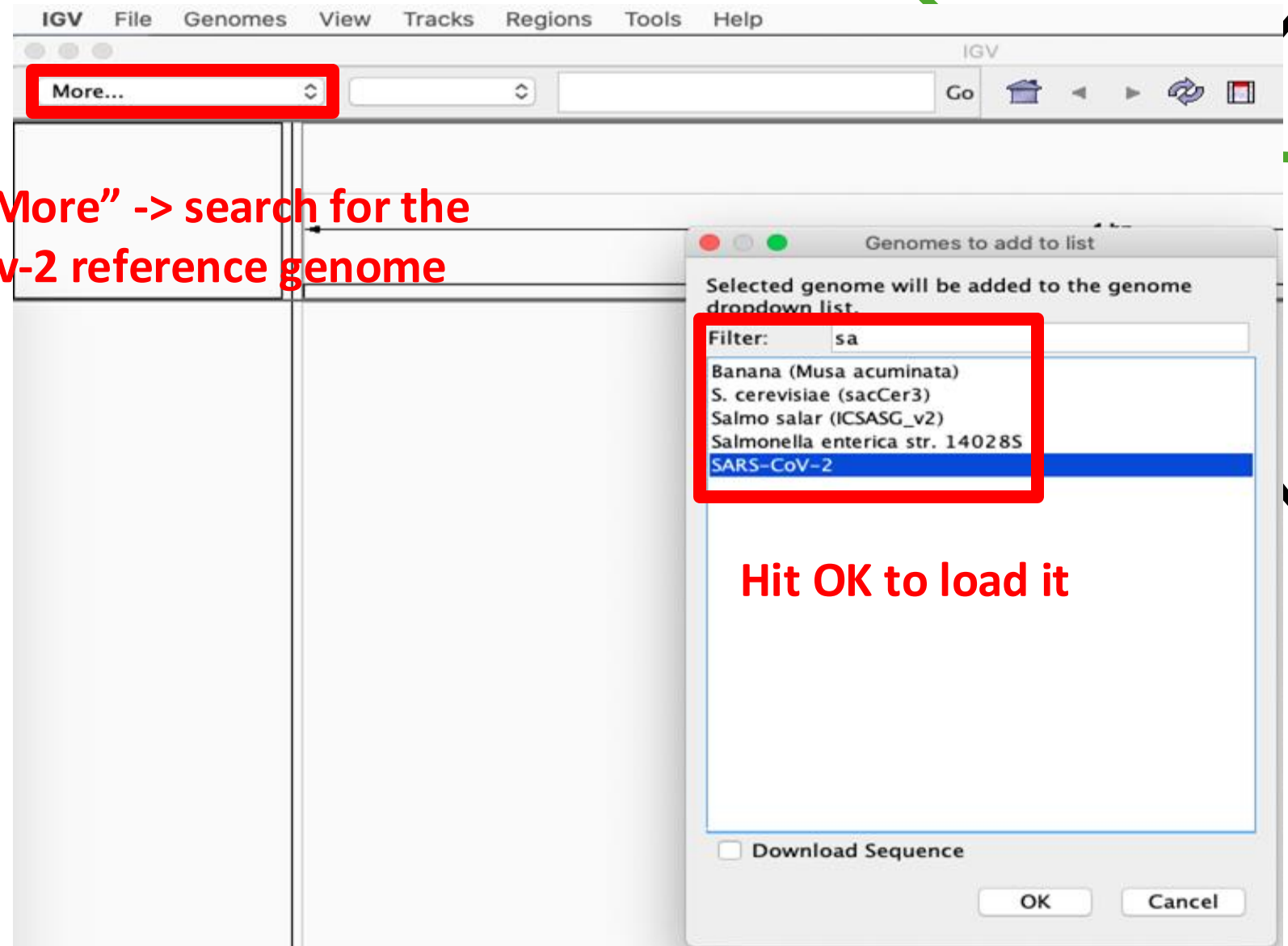
```
./igv.sh
```



Visualisation: IGV

Launch IGV
application

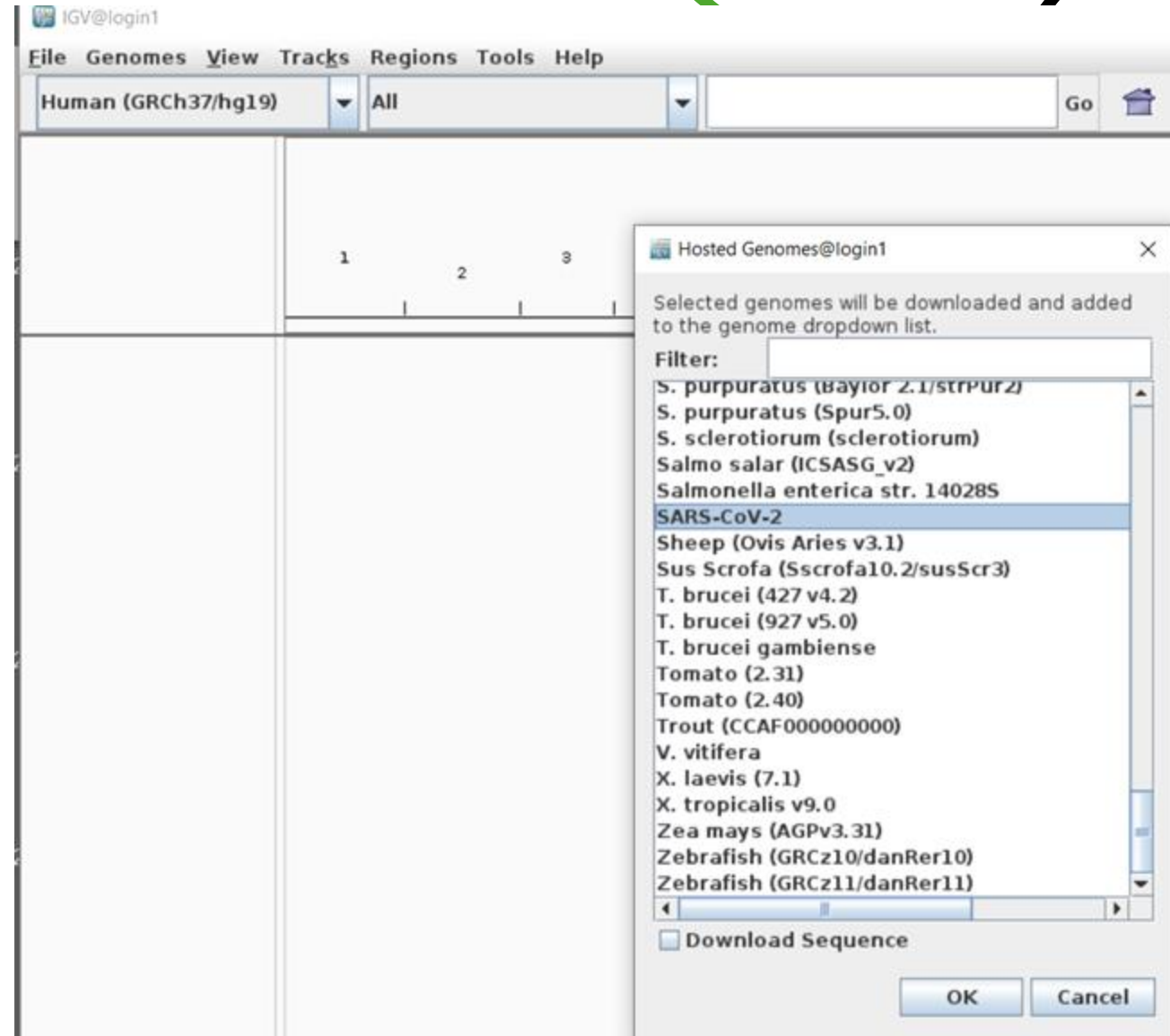
Go to "More" -> search for the
Sars-Cov-2 reference genome



Visualisation: IGV

Go to "File" and
select the reference
genome you want

e.g. Sars-Cov-2

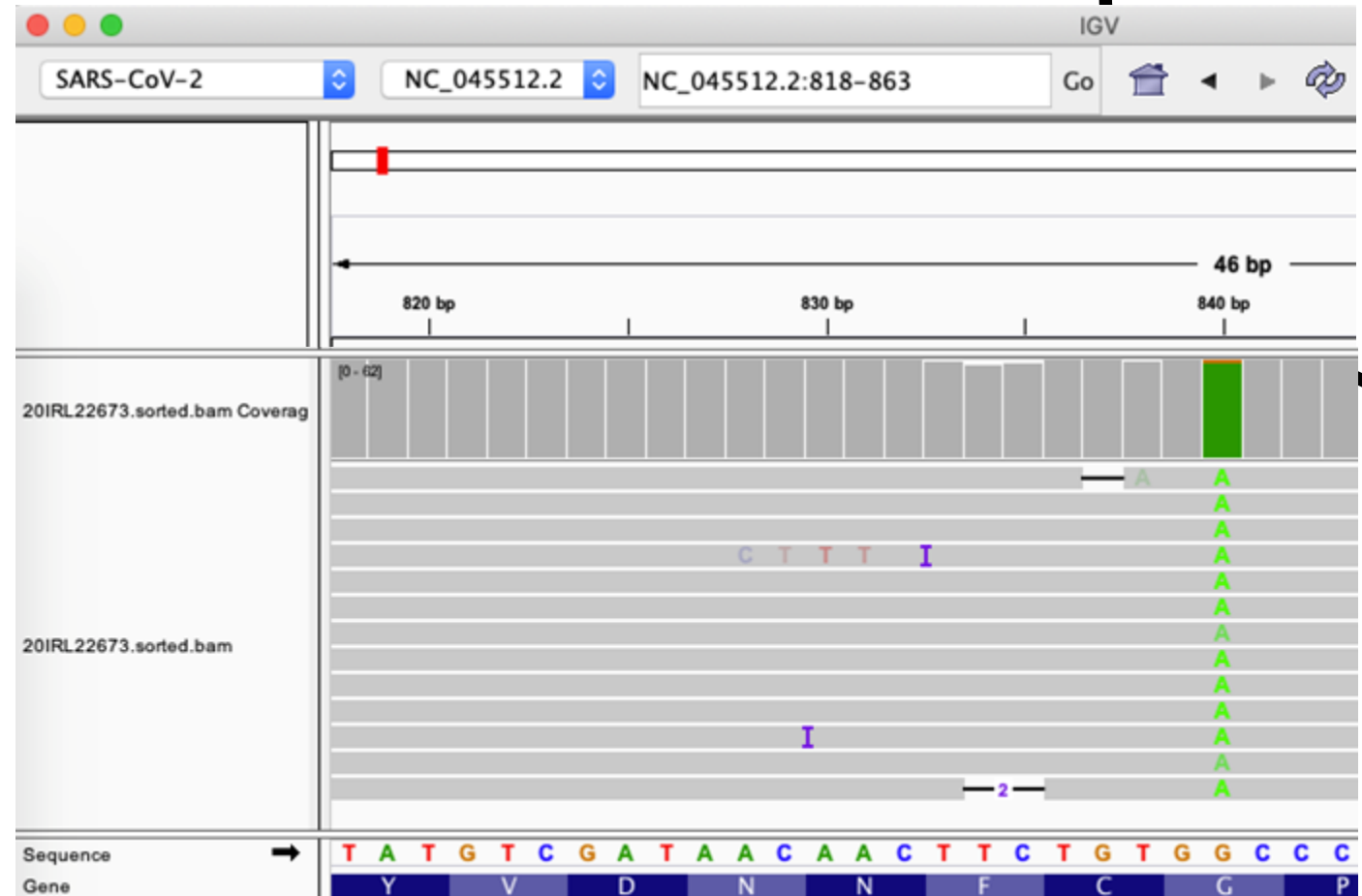


Visualisation: IGV

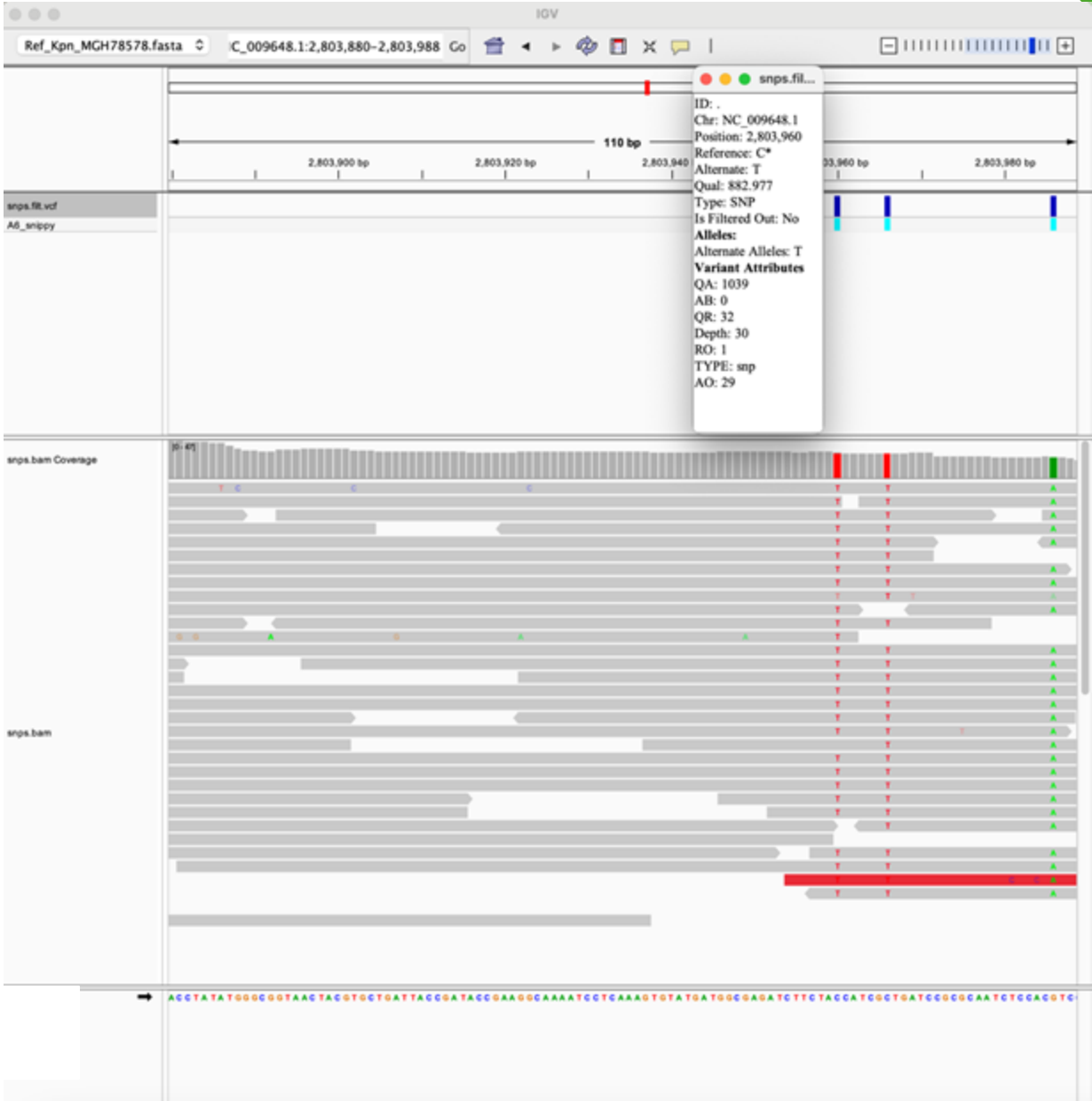
Check some mutations, e.g.:

Go to the sites
eg 840 here
GGC -> GAC
Gly -> Asp

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
MN908947.3	840	.	G	A	500	PASS	DP=62;A



Visualisation: IGV



Visualisation: IGV

More details during Tutorial Session

Tutorial (Make a copy or Download) :

<https://colab.research.google.com/drive/1zorsPRI2ioDcfH6QgBVDAspm2yVtcjtk#scrollTo=593c0two04R0>

Quick guide:

<https://github.com/aedecano/ACORN> Variant Calling