Joint Training Course of the inter EURLs Working Group on NGS:

Introduction to Bioinformatics for genomic data mining

Amplicon-based sequencing of viral genomes

Dr. Luca De Sabato
Istituto Superiore di Sanità ISS
Department of Food Safety Nutrition and Veterinary Public Health

Objectives

Understand main steps for NGS data analysis

1) why we check the quality of sequences (reads)

2) why we need to obtain contiguous sequence and long sequences

3) the difference between mapping and de-novo assemblies

4) application of mapping, de-novo or mixed methods

Viruses

•Compact

•typically ~3,000–200,000 bases

•Little wasted space

•DNA; RNA

•Single-stranded; double stranded

•Linear; circular

•Single; segmented

•Often highly variable

•Particularly true of ssRNA viruses

•Quasispecies. Example: hepatitis C virus

Why Next Generation Sequencing?

•Faster than classical methods

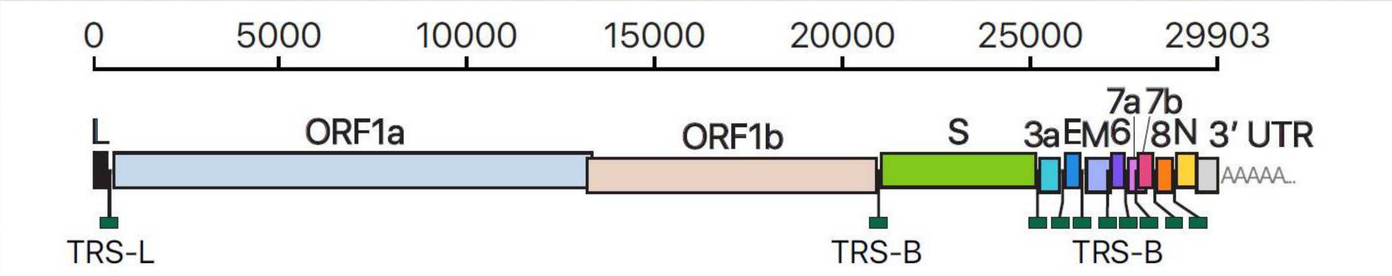
~24 h vs weeks

•Cheaper

•High Throughput: hundreds or thousands of viral genomes produced with a single NGS run

•Minor variants characterization, novel viral species discover etc.. etc…

Sars-CoV-2

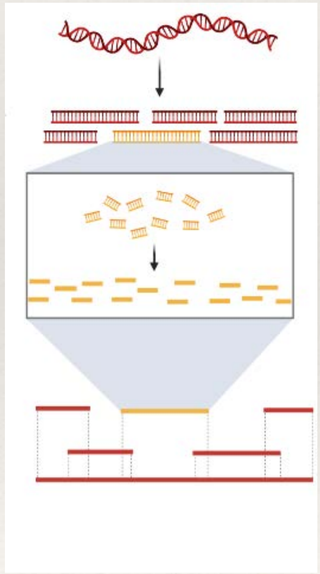


What to Produce

What?

- Sequence – generic name describing order of biological letters (DNA/RNA or amino acids).
- Reads – sequences produced by NGS that we are trying to assemble

```
ATCACAGTGGGACTCCATAAATTTTTCTGACTCCATAAATTTTTCT
```

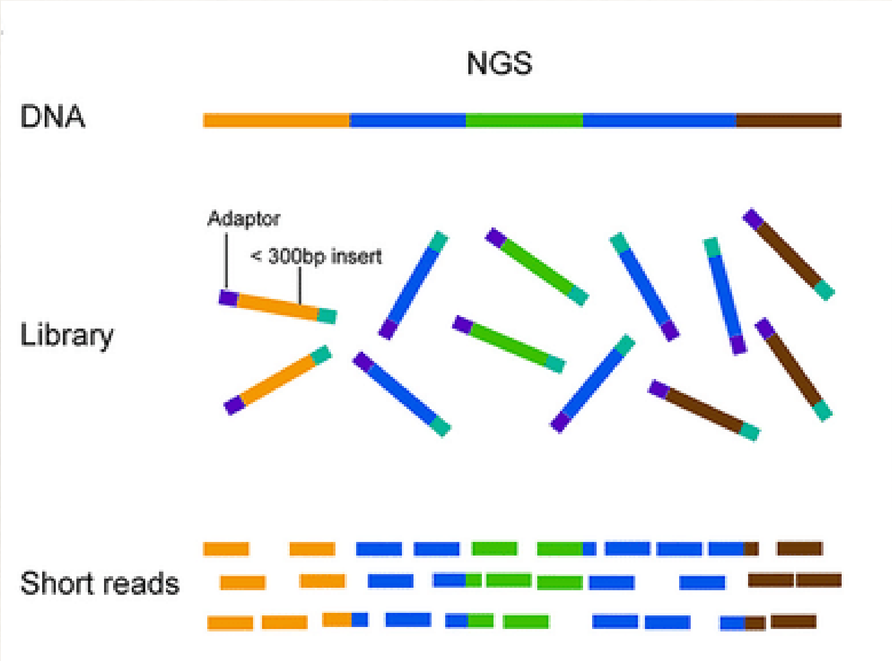


to perform →

phylogenetic analysis
variant calling
protein structure prediction
minor variants characterization
etc… etc…

Why do we need to assemble?

- Genomes are broken into pieces of about 250 nucleotides to sequence
- A large percentage of the sequence encode proteins (these regions are called genes)
- Those genomes then need to be aligned so they can be compared to each other



Longer sequences allow to perform more robust and better characterization of viral strains by phylogenetic analysis

How to Produce

- From fragments develop sequence reads
- Sequence reads are assembled into contiguous reads
- These are then compared to DB

Sample → Nucleic Acids → Library → Reads

Methods:

- Hybrid capture-enrichment sequencing
- Direct RNA sequencing
- Transcriptome
- Shotgun

But Amplicon-based sequencing…….

Amplicon-based amplification

An enrichment workflow consisting of first-strand cDNA synthesis (RNA viruses) and/or genome amplification with multiplex PCRs. The objective is to produce pools of amplicons that cover either the entire length or the discrete portions of the viral genome

Pro:

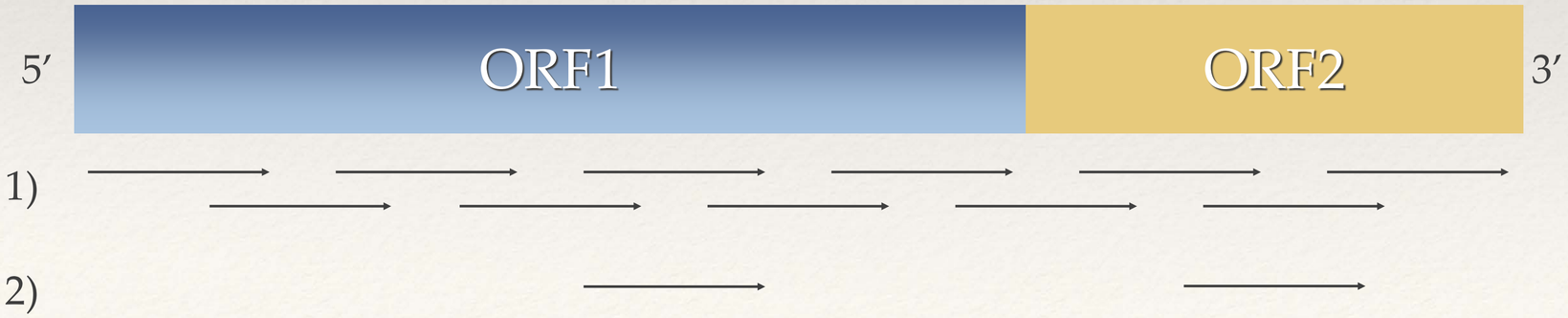
•is highly specific and robust

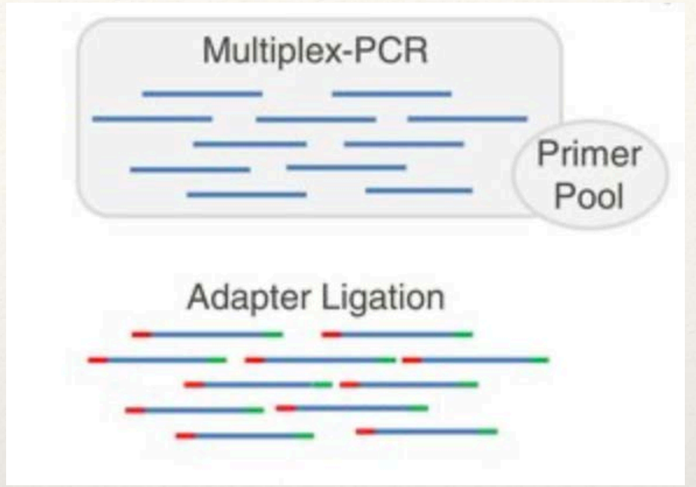
•less sequencing is required with respect to the metagenomic approaches

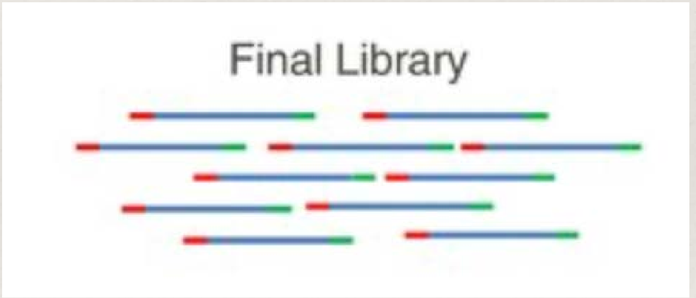
Limitations:

•differences in primer efficiency, or possible variants in the primer annealing regions, amplification across the genome can be biased, with decreased coverage in specific genomic regions leading to an incomplete assembly.

•since the primers are designed on the reference genome sequence or alignments of interesting sequences, this approach may not identify large structural variants, indels.



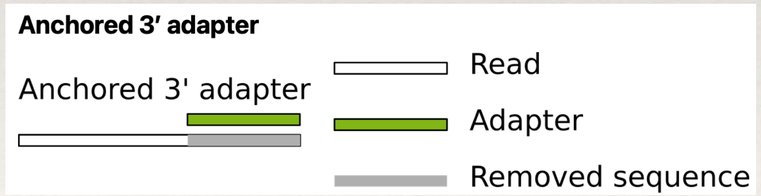




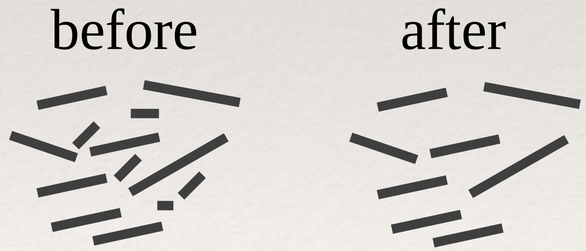
Quality check - TRIMMING

- Poor quality reads should be therefore excluded to improve the results of the downstream analysis.
- To speed up the downstream analysis (reducing the number of reads)
- To remove the shorter reads, usually <~50 bases
- Software used for read mapping/de novo assembly can be confounded by the presence of errors in reads
- To perform correct variant calling
- To speed up the the downstream analysis: the assembly phase become faster

1



2



3

AGCTGATGCTAGTCTAG
AGCTGATGCTAGTCTAG
AGCTGATGCTAG**AAAAT**
AGCTGATGCTAGTCTAG
AGCTGATGCTAGTCTAG
AGCTGATGCTAGTCTAG

Making an assemblage: Two different methods

•**Reference guided (mapping)**

When you have “map” of final product, and try to match your pieces together to look like final product

“you have picture and put pieces together”

•**De Novo assembly**

Put pieces together by what “makes sense”

“you may not have picture, but can put pieces together by what fits together”

Reference-Guided (Mapped) Assembly

When you conduct a reference guided assembly…For example – when you know that you are looking at a certain strain or similar strain to a referenze sequence – You might find some unmapped reads that may occur for a variety of reasons:

•The sequences are not actually in the reference
•Sequence strain may be a variant

ADVANTAGES: Relatively fast, well-suited to highly-conserved genomes.

DISADVANTAGES: Issues with high diversity, you can find unmapped reads

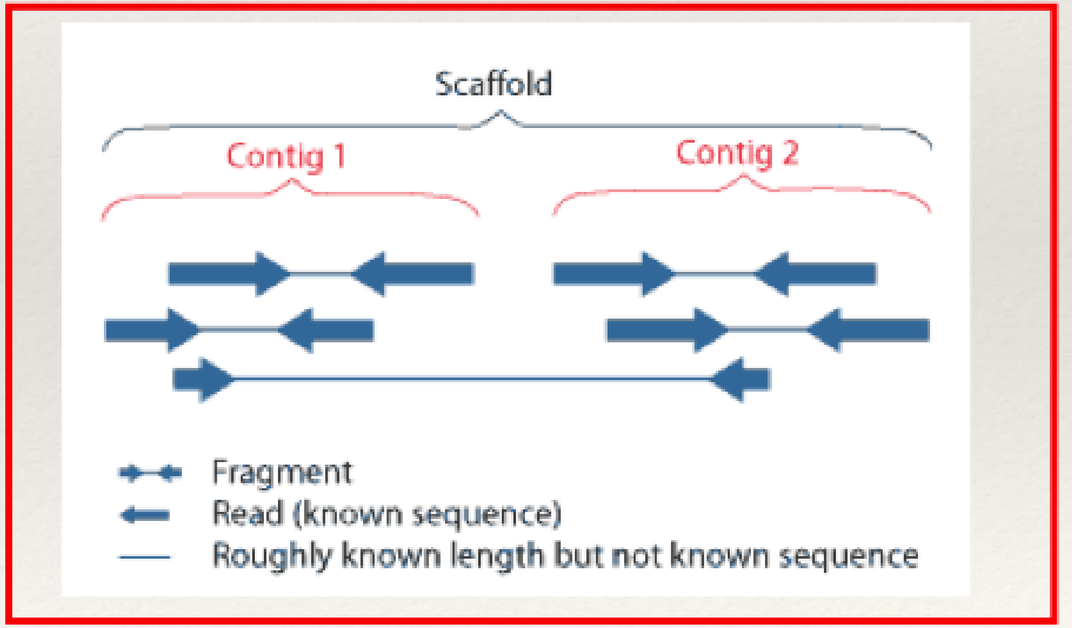
De Novo Assembly

- **Contig** (derived from contiguous): set of overlapping DNA segments that together represent a consensus region of DNA derived from a set of reads. A contig lacks gaps.

- **Scaffold**: ordered set of contigs and eventually gaps. Gap length can be guessed by using information from paired ends or mate pairs of different insert sizes.

ADVANTAGES: Reference agnostic: assembles all the reads it can.

DISADVANTAGES: Doesn't always get things right. Particularly with complex repeats.
Slow.



SARS-CoV-2
Severe Acute Respiratory Syndrome Coronavirus

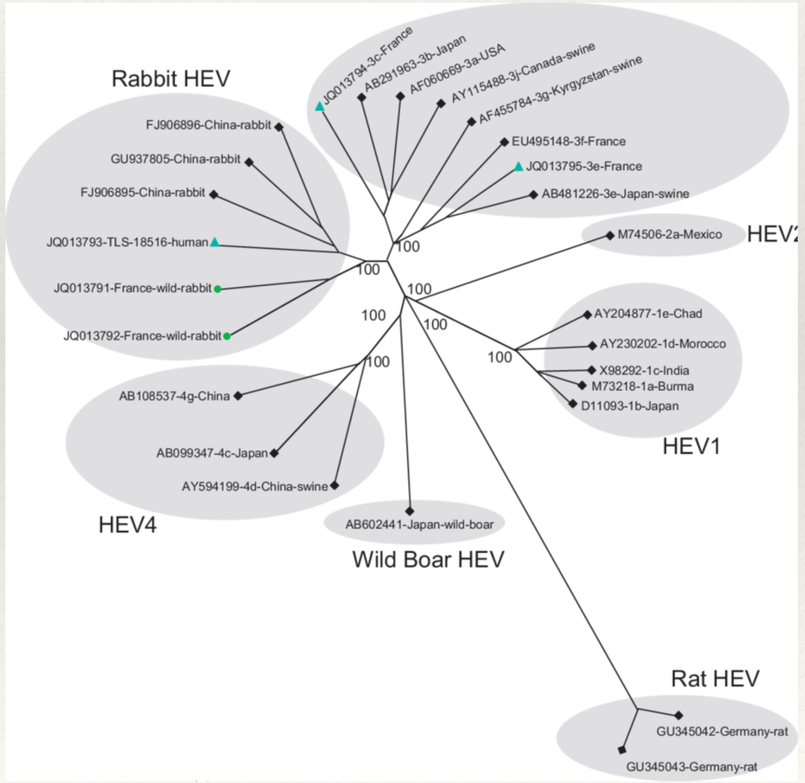
- RNA
- 30 kb
- Enveloped +ve stranded RNA
- mRNA encased in nucleocapsid
- 14 Open Reading Frames

Table. Currently designated variants of concern (VOCs)

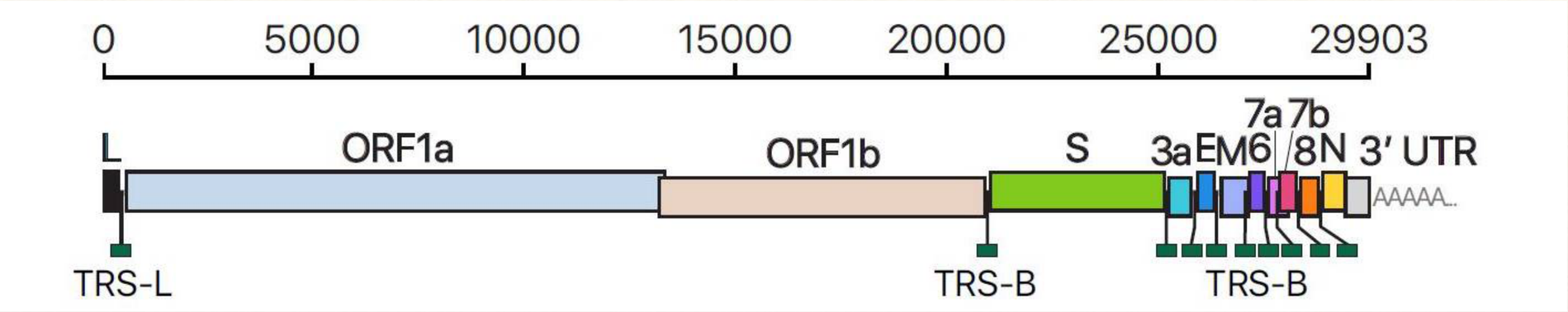
WHO label	Pango lineage	GISAID clade	Nextstrain clade	Additional amino acid changes monitored	Earliest documented samples	Data of designation
Alpha	B.1.1.7	GRY	20I (V1)	+S:484K +S:452R	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	+S:L18F	South Africa, May-2020	18-Dec-2021
Gamma	P.1	GR/501Y.V3	20J (V3)	+S:681H	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2	G/478K.V1	21A, 21I, 21J	+S:417N +S:484K	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May 2021
Omicron*	B.1.1.529	GR/484A	21K	-	Multiple countries, Nov-2021	VUM: 24-Nov-2021 VOC: 26-Nov-2021

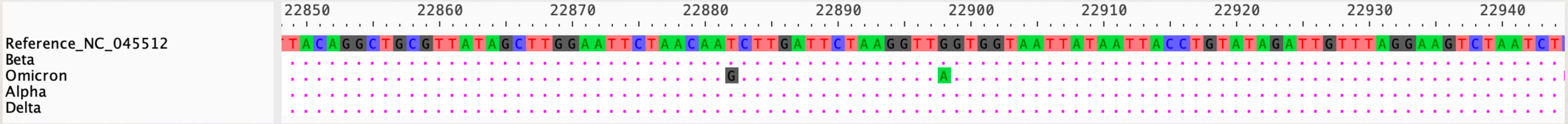
Hepatitis E virus (HEV)

- RNA
- 7 kb
- Quasi-Enveloped +ve stranded RNA
- 3 Open Reading Frames



SARS-CoV-2 sequencing strategy





What we know:

- small indels (~10 nt)
- genomes almost identical (~ 99% nucleotide identity)
- No novel ORFs acquisition

Which strategy to use?

Reference-guided

SARS-CoV-2 workflow (SARS-CoV-2 Recovery)

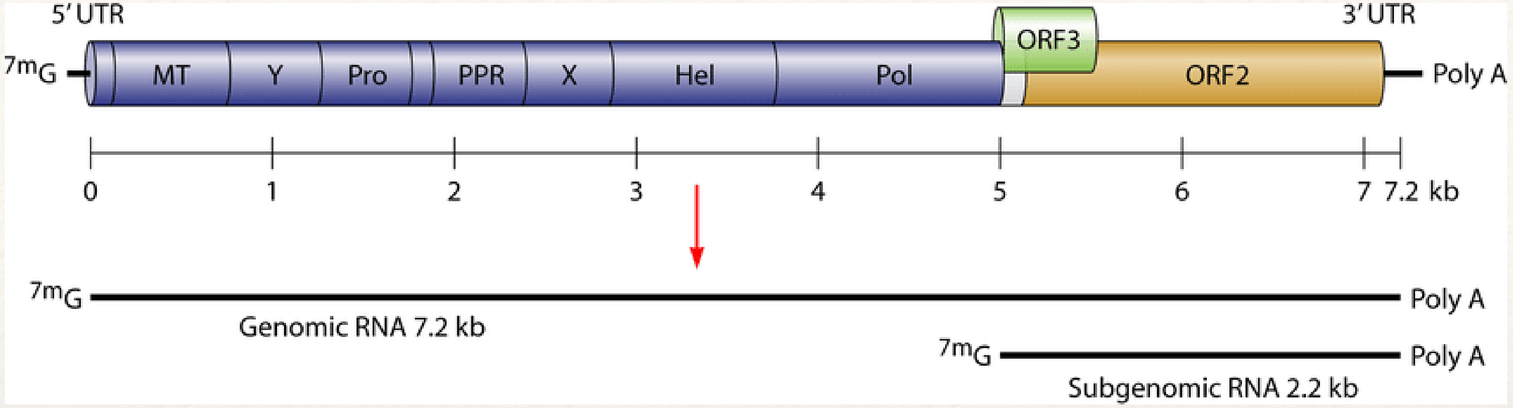
Trimming

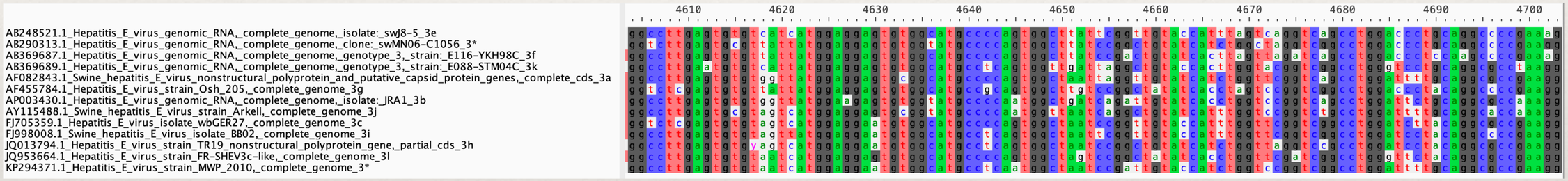
Mapping host genome

Mapping SARS-CoV-2 genome

Consensus

HEV sequencing strategy





What we know:

- Indels in hypervariable region (HVR) (from 3 to 200 nt)
- 4 main Genotype (HEV-1 to HEV-4)
- The reference is a HEV-1 sequence that share ~80% identity with HEV-2 to HEV-3
- Within a genotype are described the subtypes sharing >89% identity
- No novel ORFs acquisition

Which strategy to use?

Mixed strategy:
De novo and mapping

HEV workflow

Trimming

Mapping host genome

Unmapped reads

De novo

Unmapped reads

Mapping HEV Reference

I Consensus

BLASTn

DB subtype references

Mapping Most similar subtype ref

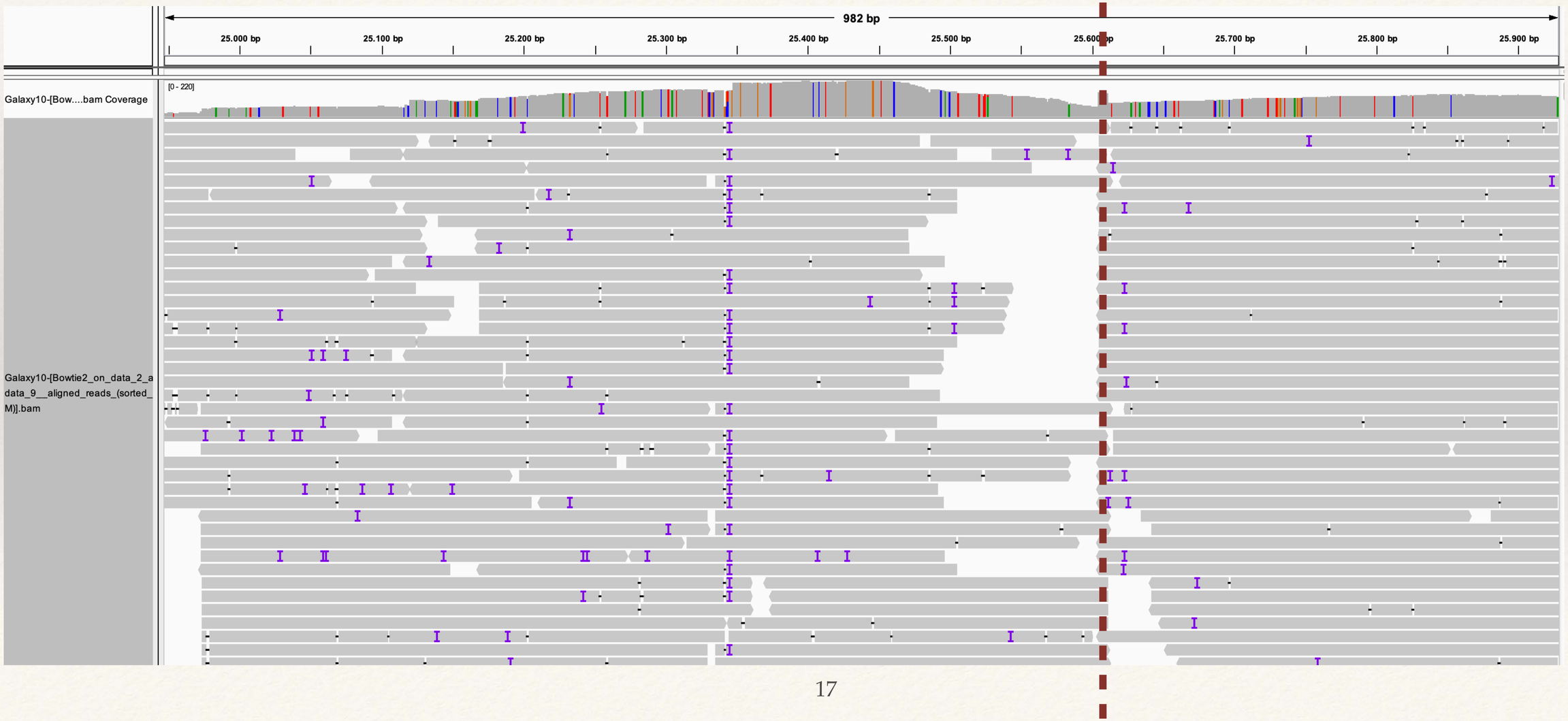
II Consensus

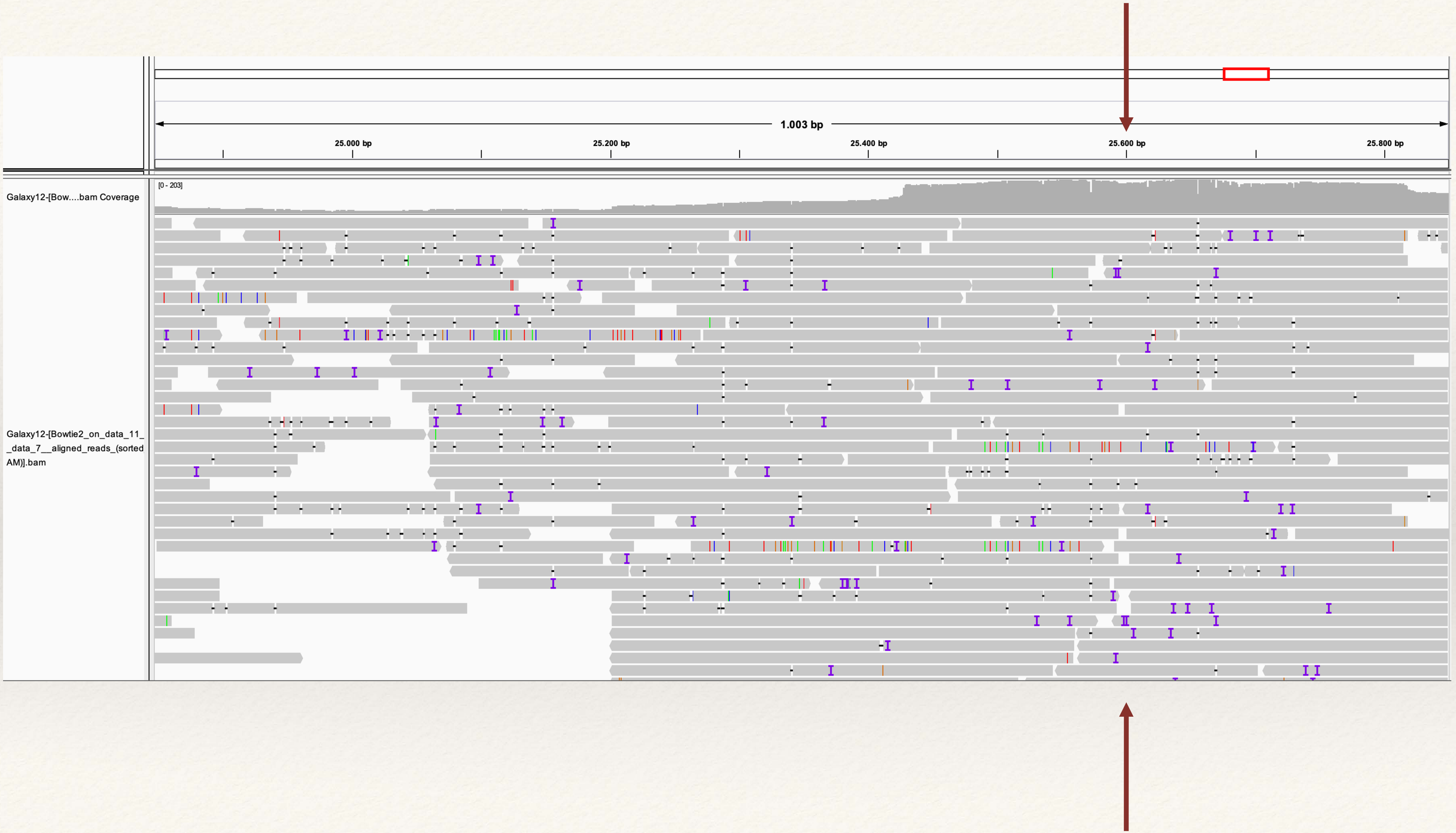
Final Consensus Alignment of consensus with scaffold

Exercise - novel ORF in hedgehogs Coronaviruses

Can Coronaviruses Steal Genes from the Host as Evidenced in Western European Hedgehogs by EriCoV Genetic Characterization?

Luca De Sabato [1], Ilaria Di Bartolo [1], Maria Alessandra De Marco [2,*], Ana Moreno [3,*], Davide Lelli [3], Claudia Cotti [4], Mauro Delogu [4,†] and Gabriele Vaccari [1,†]

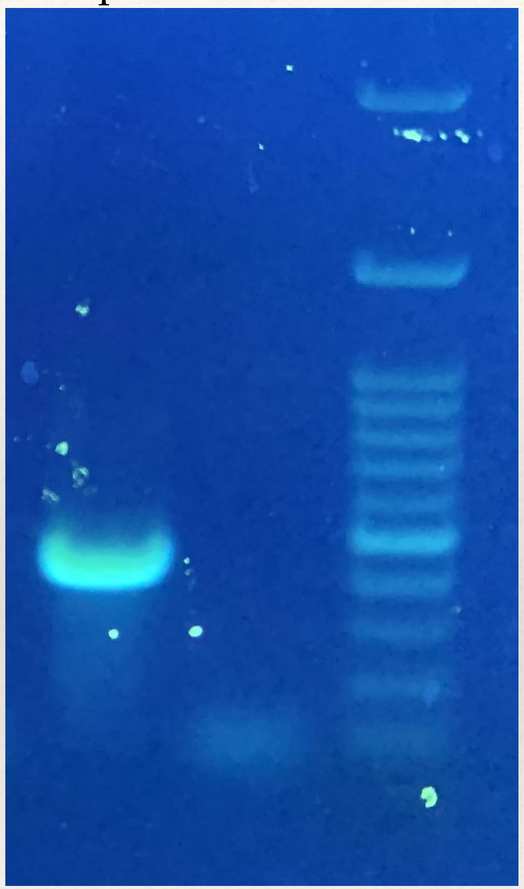



1.003 bp
25.000 bp
25.200 bp
25.400 bp
25.600 bp
25.800 bp
[0 - 203]
Galaxy12-[Bow....bam Coverage
Galaxy12-[Bowtie2_on_data_11_
_data_7__aligned_reads_(sorted
AM)].bam

RT-PCR

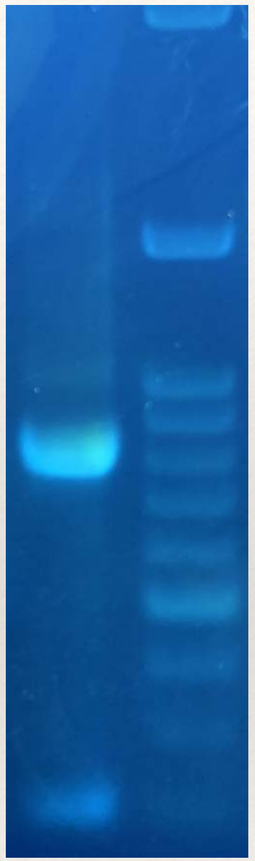
Without the novel ORF

Sample Marker



With the novel ORF

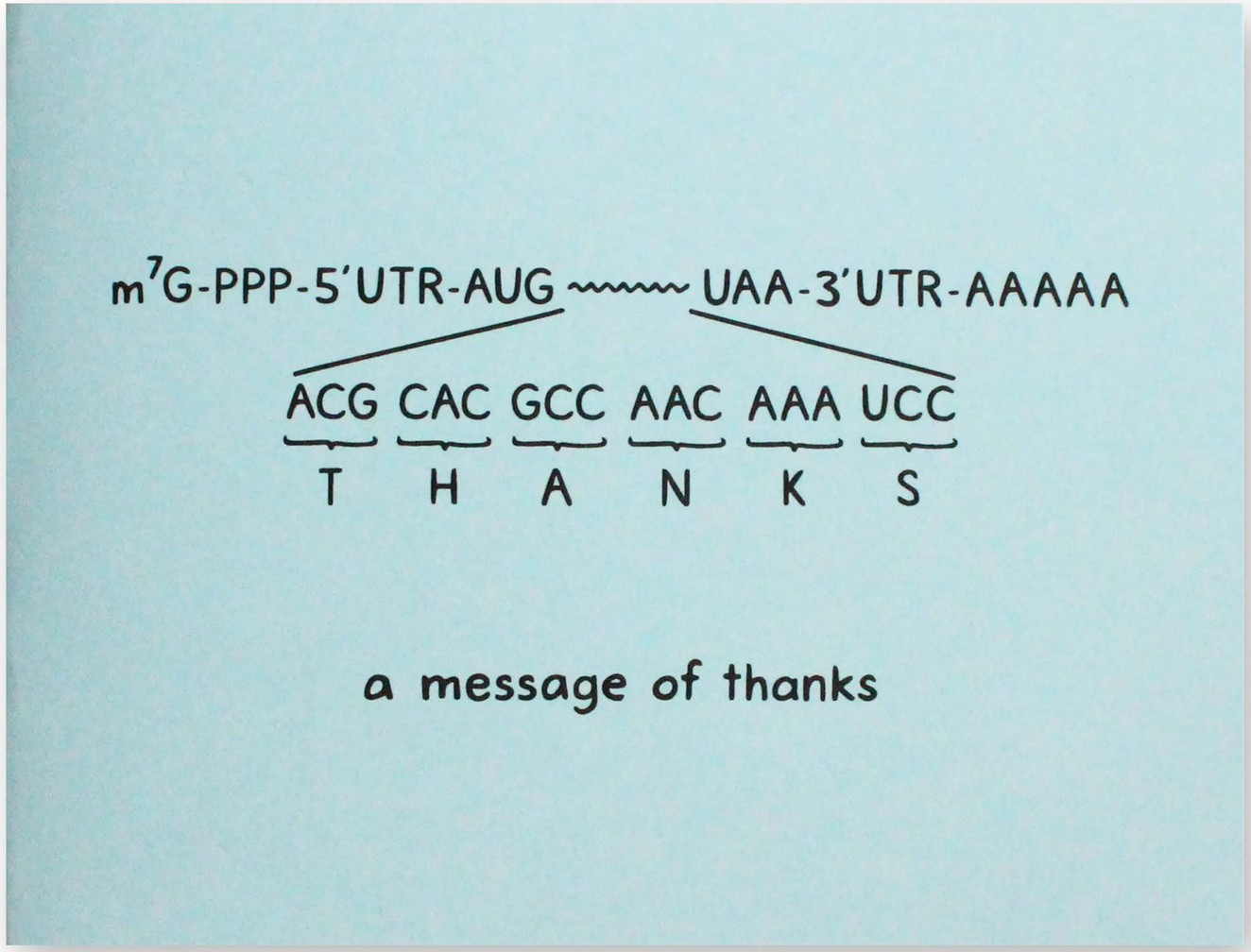
Sample Marker



Spike — CD200 — ORF3a

Take Home Messages

- NGS needs time, resources, and expertise for analysis
- Need to understand when and why NGS should be applied
- The several options used by tools and software can lead to different results
- In NGS analysis try different and several approaches finding the most suited for your analysis


m7G-PPP-5'UTR-AUG
UAA-3'UTR-AAAAA
ACG CAC GCC AAC AAA UCC
T H A N K S
a message of thanks