Module 1
Bioinformatics Basics

Taking a look behind the curtain

The Fleming Fund
Regional Grants

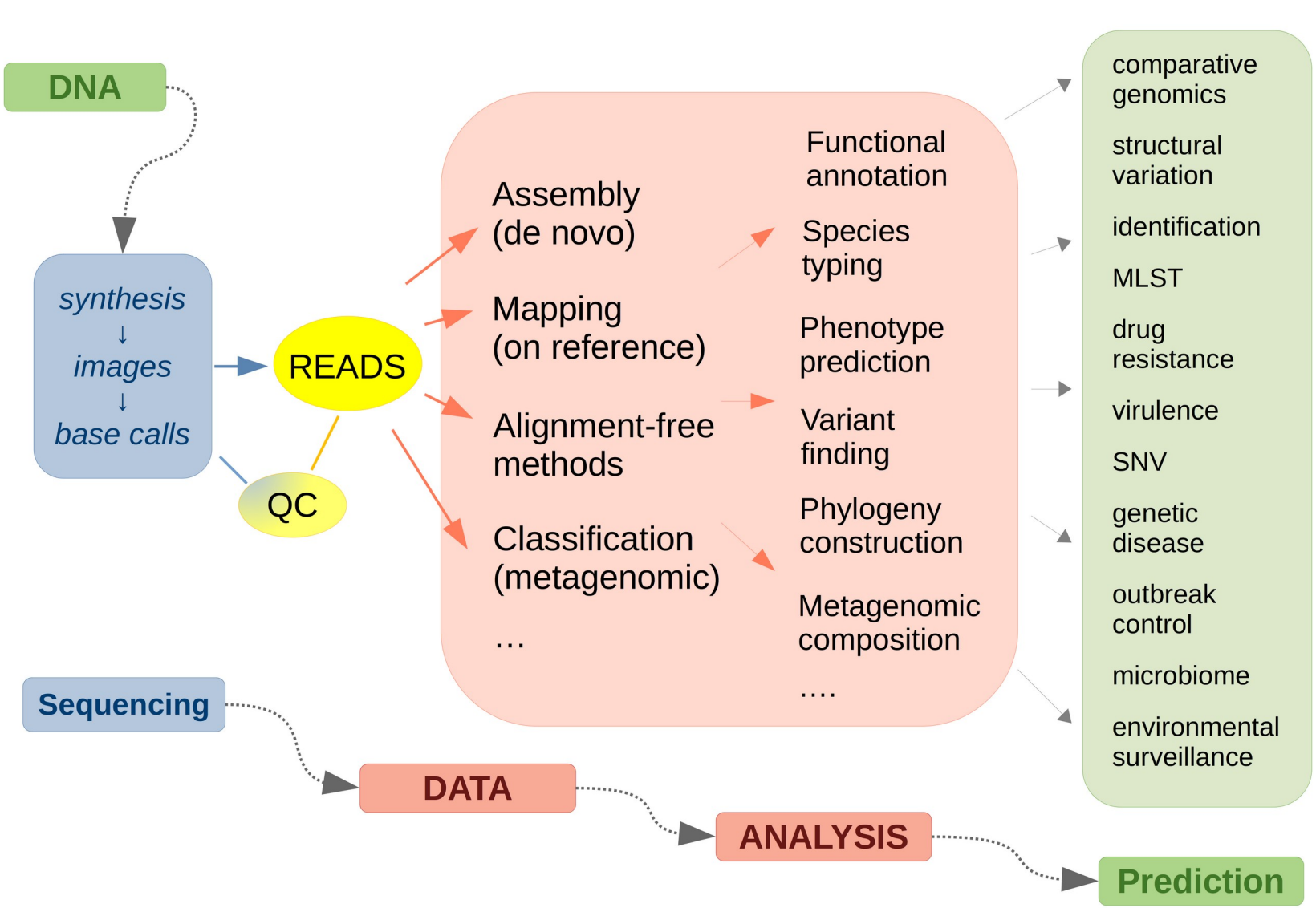
17 February 2021

Marco van Zwetselaar

Kilimanjaro Clinical Research Institute

Topics

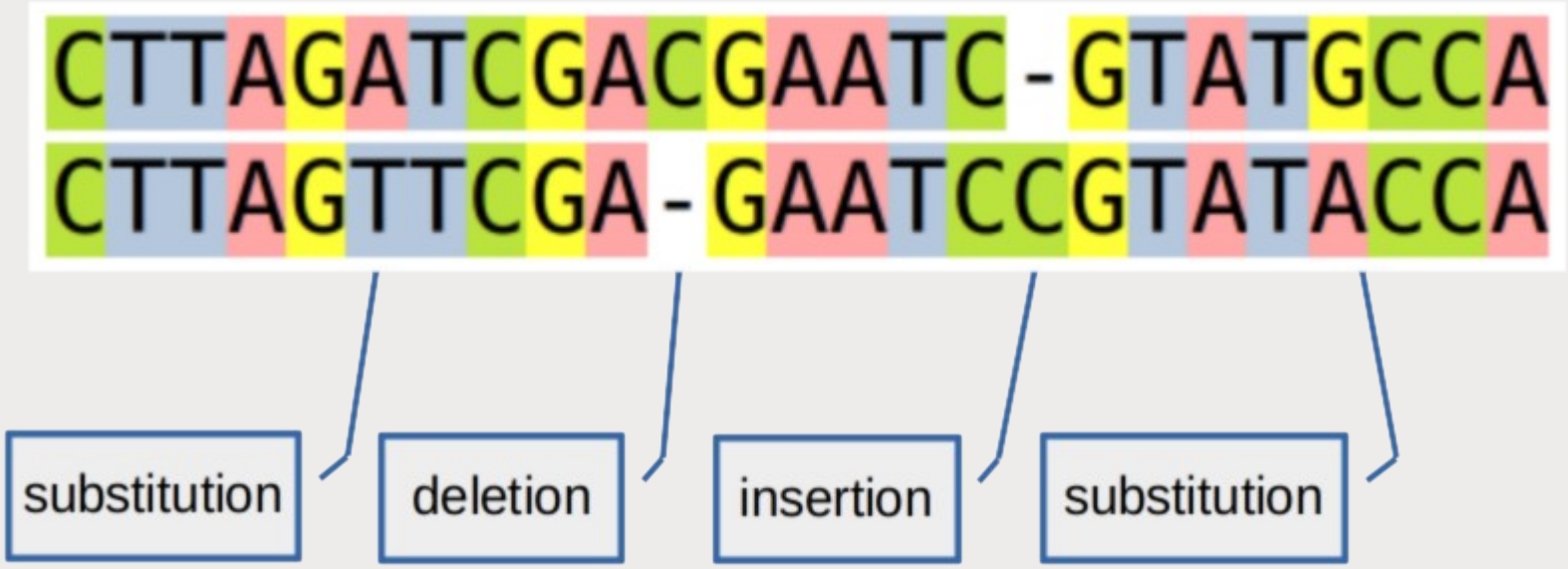
- What do all these technical terms mean?
  - What are reads, assembly, FASTQ, FASTA?
  - More terminology: alignment, quality scores, coverage, depth
- How is sequencer output turned into a genome?
- How does species identification and typing work?
- How do we find AMR genes and mutations?
- What is happening “behind the curtain” in the tools we use?



The Fleming Fund
Regional Grants

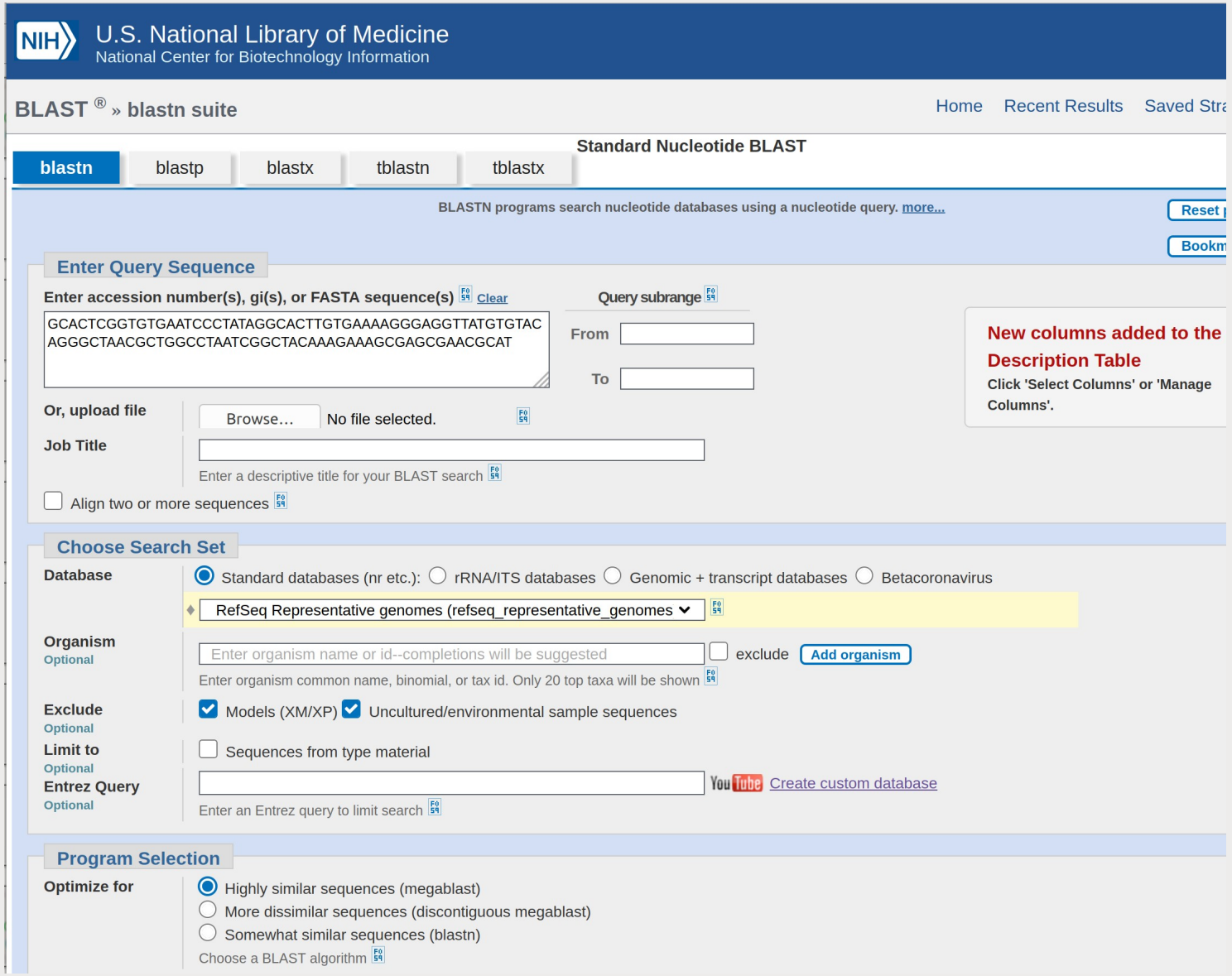
Bioinformatician Bird's Eye View

- Everything starts with **reads** ...
- … and ends at (just before) prediction
- Trend toward end user operable **pipelines** that perform a workflow of analyses
- These analyses are built using a still fast-growing toolset
- Rapid innovation continues – keep abreast of the "data deluge"
- But at the basis are a relatively small number of 'core operations'



Alignment

- At the heart of bioinformatics
- Alignments can be **scored** to give an alignment quality
- Dissimilarity of two sequences: **edit distance.** How many changes turn one into the other?
- Edit penalties can be **weighted**, e.g:
  - Gap vs substitution
  - Transversion vs transition
  - Observed substitution rates
  - ...



The Fleming Fund
Regional Grants

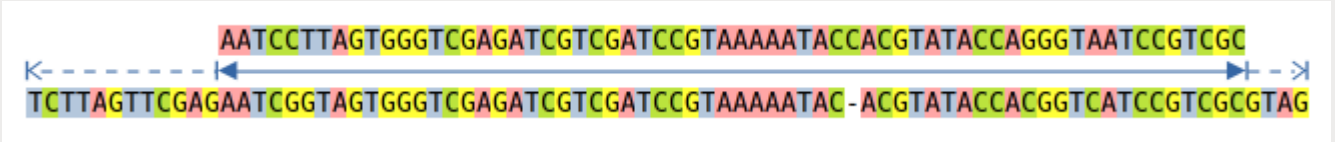
BLAST

- https://blast.ncbi.nlm.nih.gov
- Basic **Local Alignment** Search Tool
- Search for matches of a query sequence in (huge) sequence databases
  - But can be used offline too
- Matches come with metrics that express alignment quality

Metrics you will likely encounter

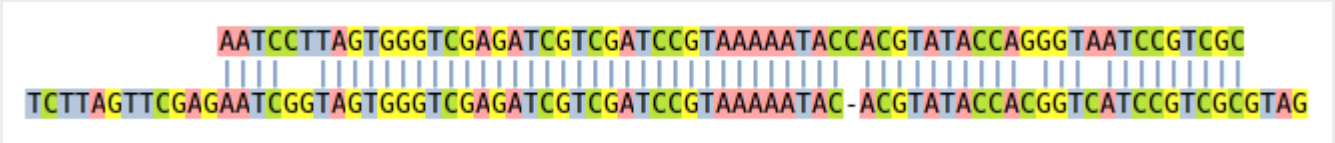
- Coverage
  - Do not confuse with coverage *depth*

Coverage: percentage of target region covered by query (here 80%)

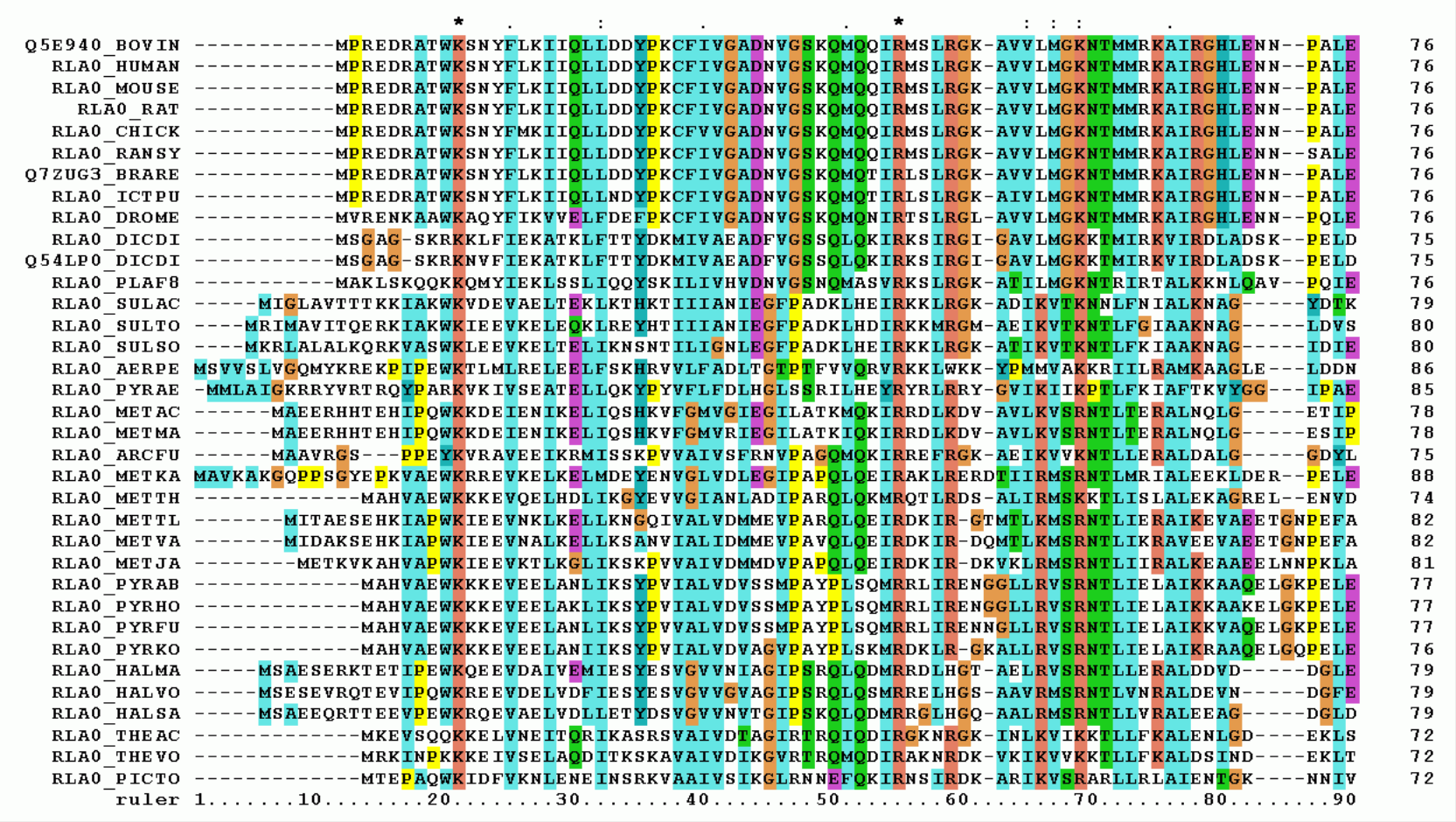


- Percentage identity

Identity: percentage of bases in the alignment that match exactly (here 92%)



- *Bit score, E-value, p-value*: related to the probability of attaining at least the alignment score by chance



https://commons.wikimedia.org/wiki/File:RPLP0_90_ClustalW_aln.gif (CC BY-SA 3.0)

Multiple Alignment

- Same concept, multiple sequences
- Here to illustrate homology of ribosomal protein P0 across species
- Basis for phylogenetic analyses:
  - Multiple-align genomes of a collection of isolates
  - Compute edit distance between every pair
  - Assume edit distance ~ evolutionary distance
  - “Evolutionary distance matrix” is basis for inferring phylogenetic tree

Reads and Assemblies

- … and FASTQ & FASTA
- *De novo* assembly
- Hybrid assembly

```
@M02836:3:000000000-AD058:1:1101:19894:6963 1:N:0:8
ACCGCAAGATGCGACACCACTTACATTAGGTCAAGAAATCTCCGGTTGGGCGGCGATGCTTGAGCATAATATCAAACAT...
+
3A1>A11>11@DA1EEGG0F1EBF12F1111BBFDEGHGDHHFG/AAAEE@?!!!###?2BBBG>FEG2>F22>E02B/...
@M02836:3:000000000-AD058:1:1101:11069:5821 1:N:0:8
CGGTGTAATGTTGAGATAGTGACTAAGAGAAACAGACGAATGAAAGAAAATTATTATTATTTGGAGCAAACTTCAGAAA...
+
9!!!>FD&&FFGGGFGFGGGFHHGCHHHGHHHGFHHGGGHGGCHHHHFHHHFHFGH>>9.9.7777$$$$$########...
@M02836:3:000000000-AD058:1:1101:18069:7032 1:N:0:8
ACATTACCGCAGCAGCCTCGACAAGTACAAATCTCCTGTCAATTGCGTATTACTGAACGTTTATAATTAGGCACACCAG...
+
&&&&&*.,3***+(%(""%%&$'$$')&)06*'/53)*%,'+(%%%%(&0)5.373:0')-/101..)*+14;9:96/-...

... and so on ...
```

One read is 4 lines

Sequence

Q-scores

The Fleming Fund
Regional Grants

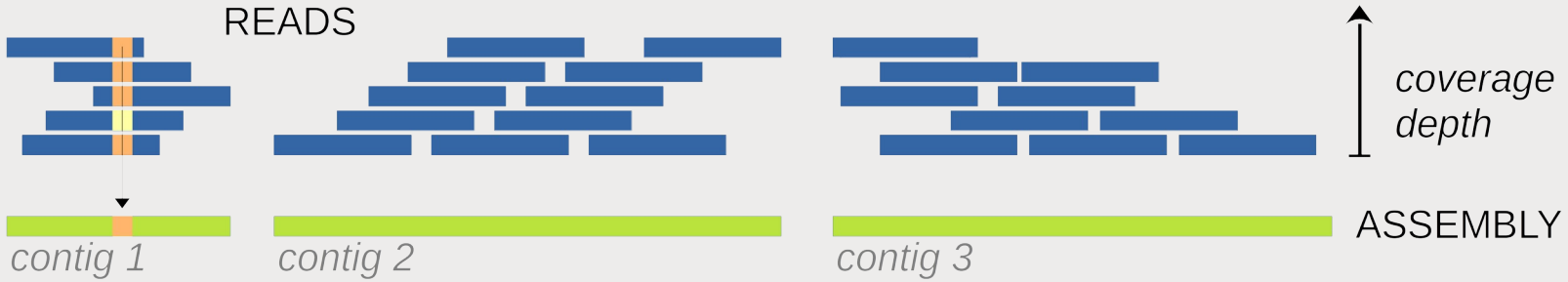
FASTQ

- Produced by the sequencer, one or two files per sample
- Contains (often millions) of **reads**: the nucleotide sequences of the fragments in your library
- … for *each* base an estimate of its **accuracy**, its Q-score:

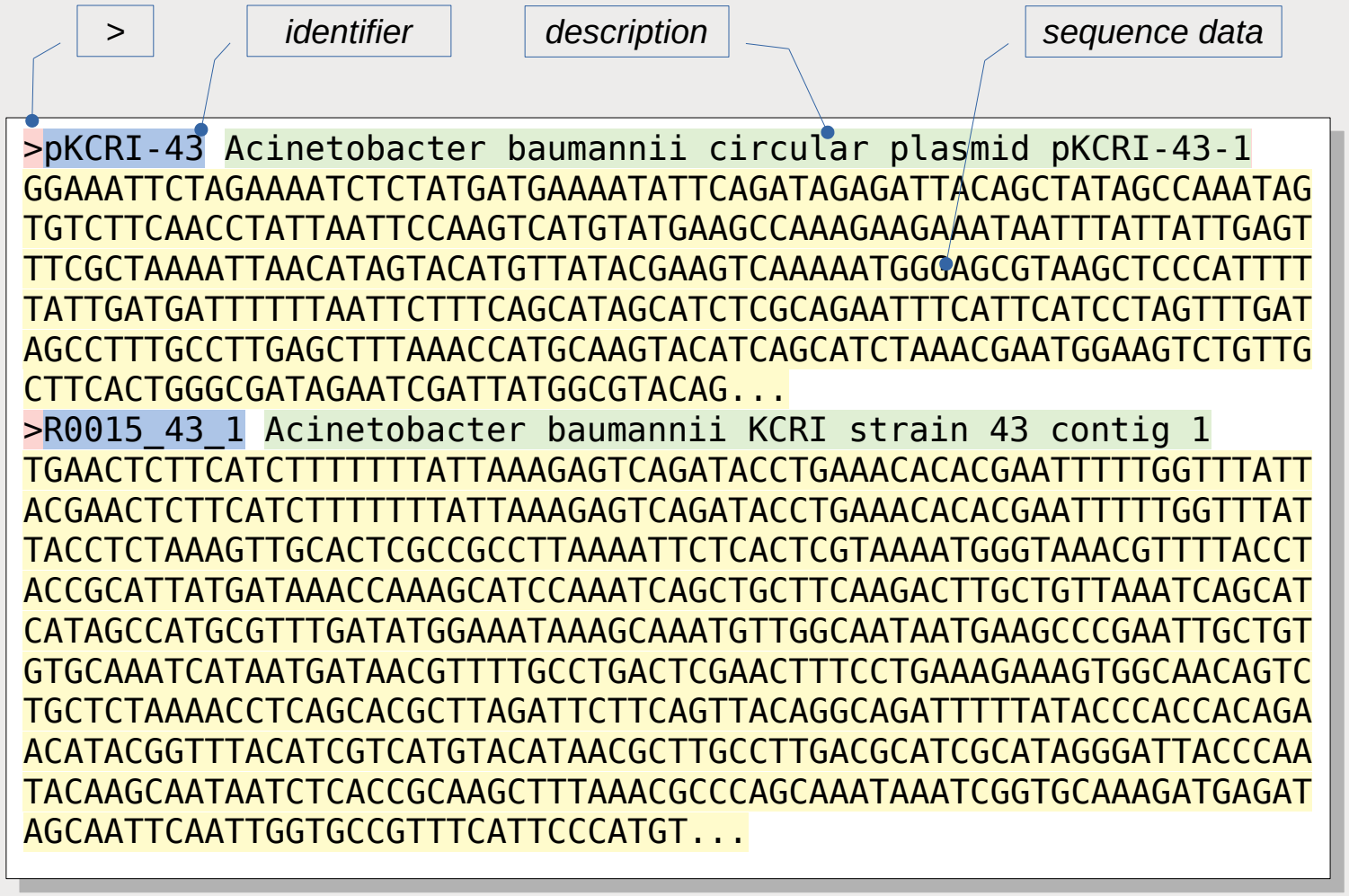
  Q30 (on the Phred scale) means 1:1000 probability of being incorrect.

- The read headers have technical metadata, relevant for QA
- File extension usually `.fastq.gz` or `.fq.gz`

Assembly *(de novo)*



- Goal: reconstruct the genome of the organism from the reads
- The “exploding newspapers” analogy (Pevzner & Compeau)
- Puzzle together increasingly longer contigs by joining ones with overlapping edges
- Result: a set of contigs (unitigs) that can’t be joined further as
  - no other contig overlaps, or
  - multiple overlap (so which to choose)?



identifier

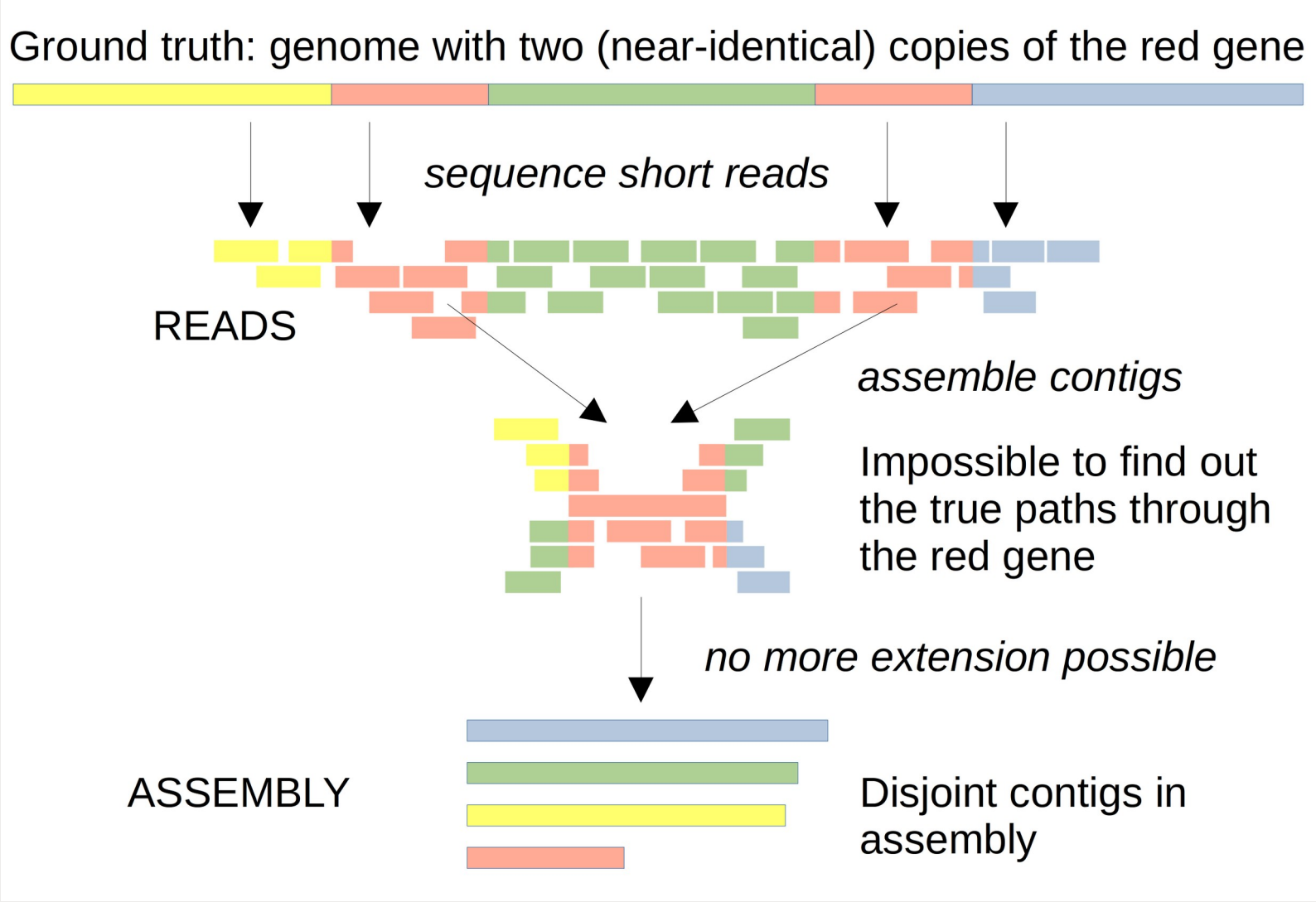
description

sequence data

```
>pKCRI-43 Acinetobacter baumannii circular plasmid pKCRI-43-1
GGAAATTCTAGAAAATCTCTATGATGAAAATATTCAGATAGAGATTACAGCTATAGCCAAATAG
TGTCTTCAACCTATTAATTCCAAGTCATGTATGAAGCCAAAGAAGAAATAATTTATTATTGAGT
TTCGCTAAAATTAACATAGTACATGTTATACGAAGTCAAAAATGGCAGCGTAAGCTCCCATTTT
TATTGATGATTTTTTAATTCTTTCAGCATAGCATCTCGCAGAATTTCATTCATCCTAGTTTGAT
AGCCTTTGCCTTGAGCTTTAAACCATGCAAGTACATCAGCATCTAAACGAATGGAAGTCTGTTG
CTTCACTGGGCGATAGAATCGATTATGGCGTACAG...
>R0015_43_1 Acinetobacter baumannii KCRI strain 43 contig 1
TGAACTCTTCATCTTTTTTTATTAAAGAGTCAGATACCTGAAACACACGAATTTTTGGTTTATT
ACGAACTCTTCATCTTTTTTTATTAAAGAGTCAGATACCTGAAACACACGAATTTTTGGTTTAT
TACCTCTAAAGTTGCACTCGCCGCCTTAAAATTCTCACTCGTAAAATGGGTAAACGTTTTACCT
ACCGCATTATGATAAACCAAAGCATCCAAATCAGCTGCTTCAAGACTTGCTGTTAAATCAGCAT
CATAGCCATGCGTTTGATATGGAAATAAAGCAAATGTTGGCAATAATGAAGCCCGAATTGCTGT
GTGCAAATCATAATGATAACGTTTTGCCTGACTCGAACTTTCCTGAAAGAAAGTGGCAACAGTC
TGCTCTAAAACCTCAGCACGCTTAGATTCTTCAGTTACAGGCAGATTTTTATACCCACCACAGA
ACATACGGTTTACATCGTCATGTACATAACGCTTGCCTTGACGCATCGCATAGGGATTACCCAA
TACAAGCAATAATCTCACCGCAAGCTTTAAACGCCCAGCAAATAAATCGGTGCAAAGATGAGAT
AGCAATTCAATTGGTGCCGTTTCATTCCCATGT...
```

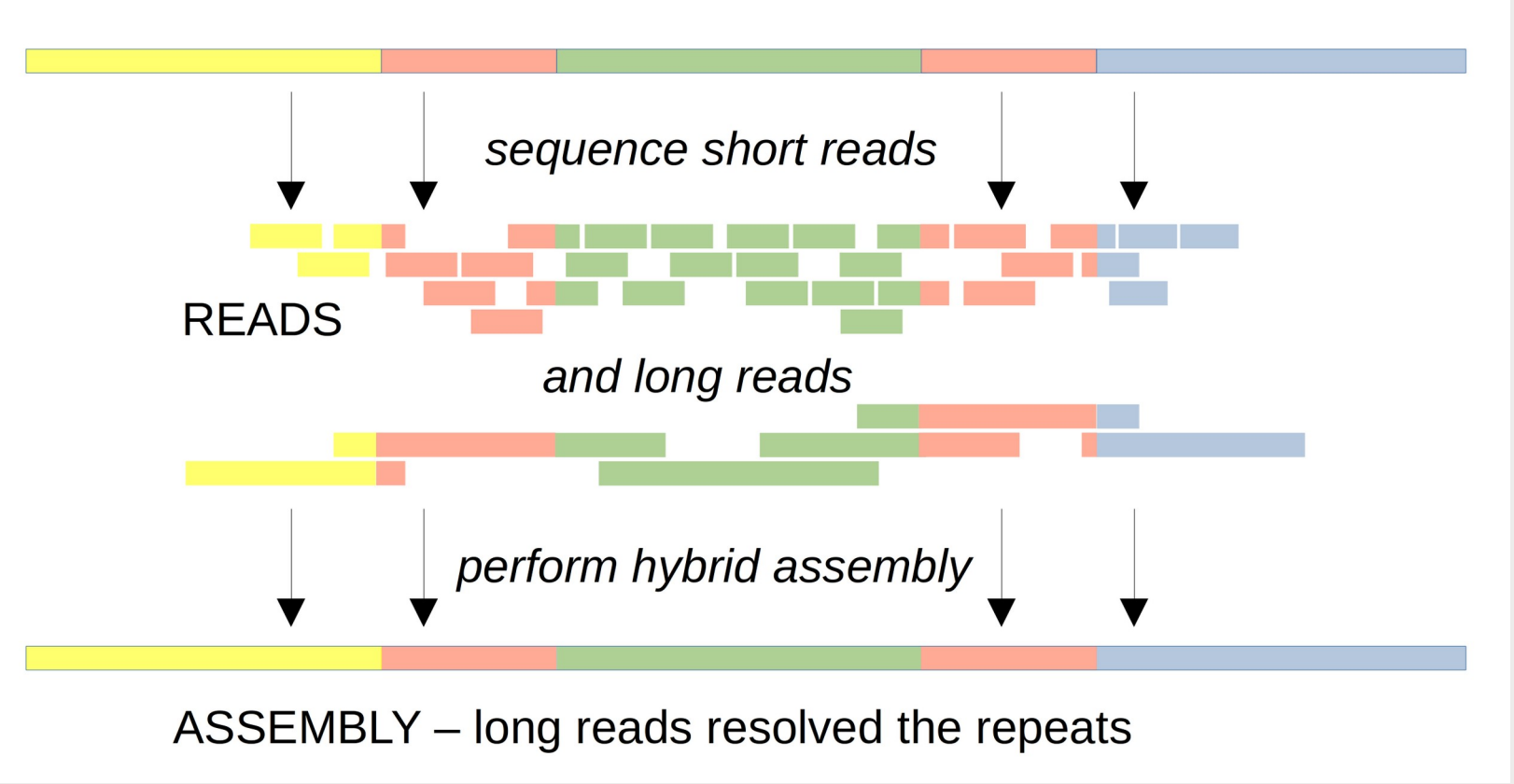
FASTA

- Contains one or more sequences
  - For instance the contigs produced by an assembler
  - But can be any nucleotide or amino acid sequence
- Each sequence preceded by an identifying header
- Common file name extensions: `.fas`, `.fasta`, `fsa`, `.fna`, `.faa`; compressed `.gz`
-



The problem with repeats

- When there are near-identical repeats of a region that is larger than the read length …
- … then the assembler cannot tell from which of the copies the reads were read
- … so it produces a single contig covering either repeat region
- … and it gets in trouble at the edges of the contig, where there are two possible continuations
- So it must split the contigs there, and we can't know their order



Hybrid assembly

- Combine short and long reads
- “Best of both worlds”:
  - Short reads provide accuracy
  - Long reads for structure

Brief Recap

- FASTQ contains reads
  - Nucleotide sequences of your library fragments
  - With a quality score for every base read
- FASTA files contain sequences
  - Typically an assembled genome (broken into contigs)
  - But can be any collection of sequences: AMR genes, alleles of some gene, protein products, etc.

You know enough to do species detection

- Assemble the genome of your isolate
- Download the *Microbial 16S rRNA* database from NCBI
- BLAST your assembled genome against this database
- Pick the highest scoring alignment
- Check that its coverage and identity percentages are good
- Voilà, your own species finder!

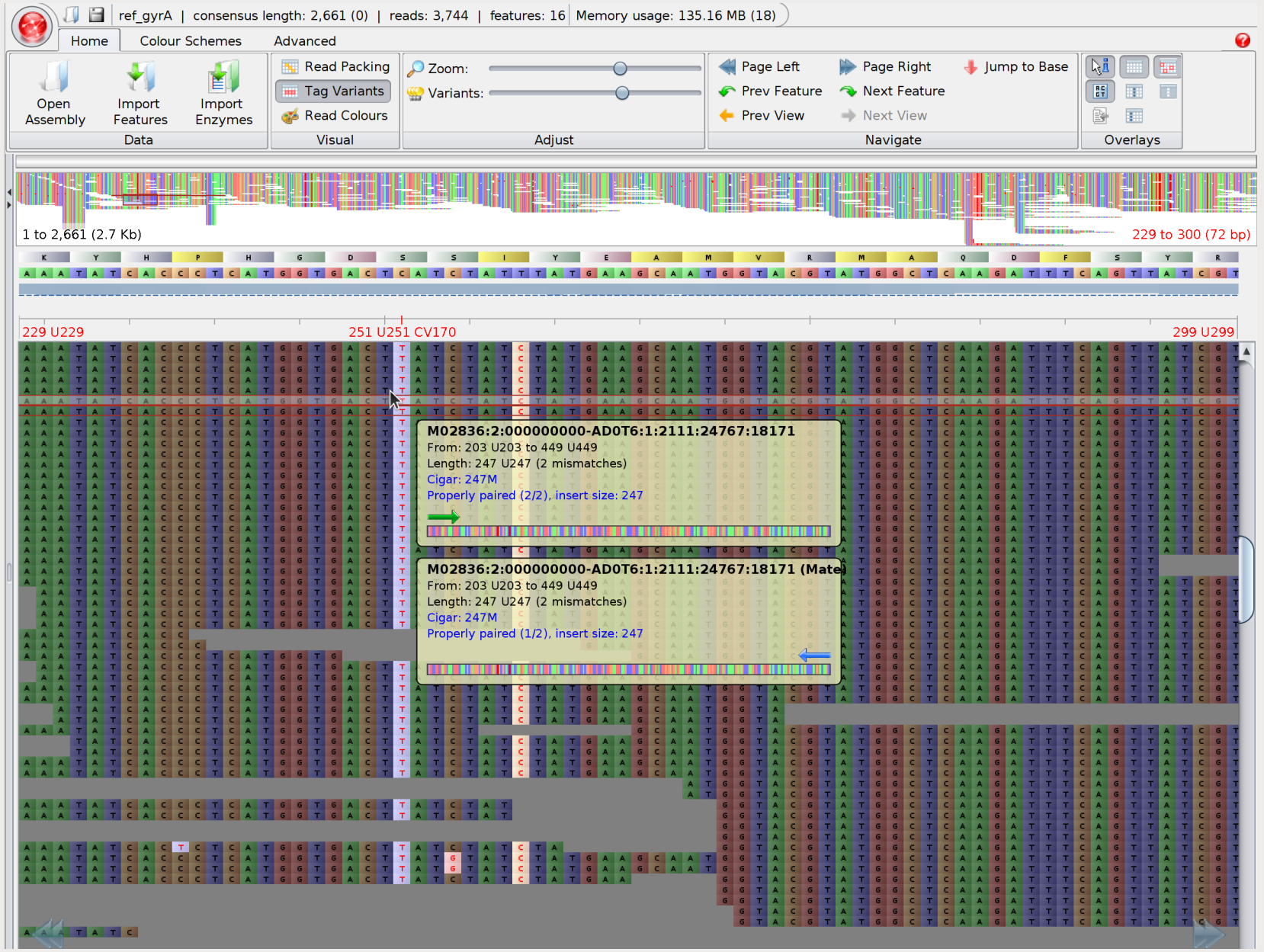
And you can do AMR detection too!

- Assemble the genome of your isolate
- Obtain FASTA files with the sequences of known AMR genes
  - Freely downloadable, e.g. DTU CGE (genomicepidemiology.org)
- BLAST the genes against your assembled genome
- List all high scoring alignments with sufficient coverage and identity
- Voilà, your own AMR Finder!

- But … what about point mutations?

Mapping

- Core bioinformatics procedure
- Mapping has a **target**, e.g. **reference** gene or genome
- Map all **reads** for an isolate on the target – dropping unmapped ones
- Alternative for *de novo* assembly when we know the organism: map on a **reference genome**
- Particularly appropriate when the goal is finding **SNP**s (e.g. in phylogeny)
- The basis for **variant calling** and obtaining **consensus sequence**



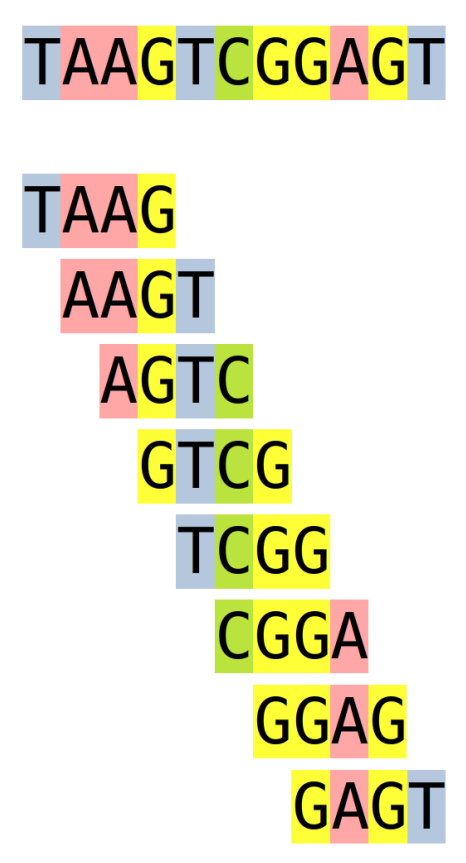
The Fleming Fund
Regional Grants

Illustration (tool: Tablet)

- Mapping of the reads of a *Staph aureus* isolate on reference *gyrA* gene
- Mutation S84L on *gyrA* is known to contribute to Quinolone resistance
- We spot C>T in nearly all 170 reads covering nt pos 251, thus call variant T with confidence
- Meaning that codon T**C**A (S) on reference is T**T**A (L) on isolate

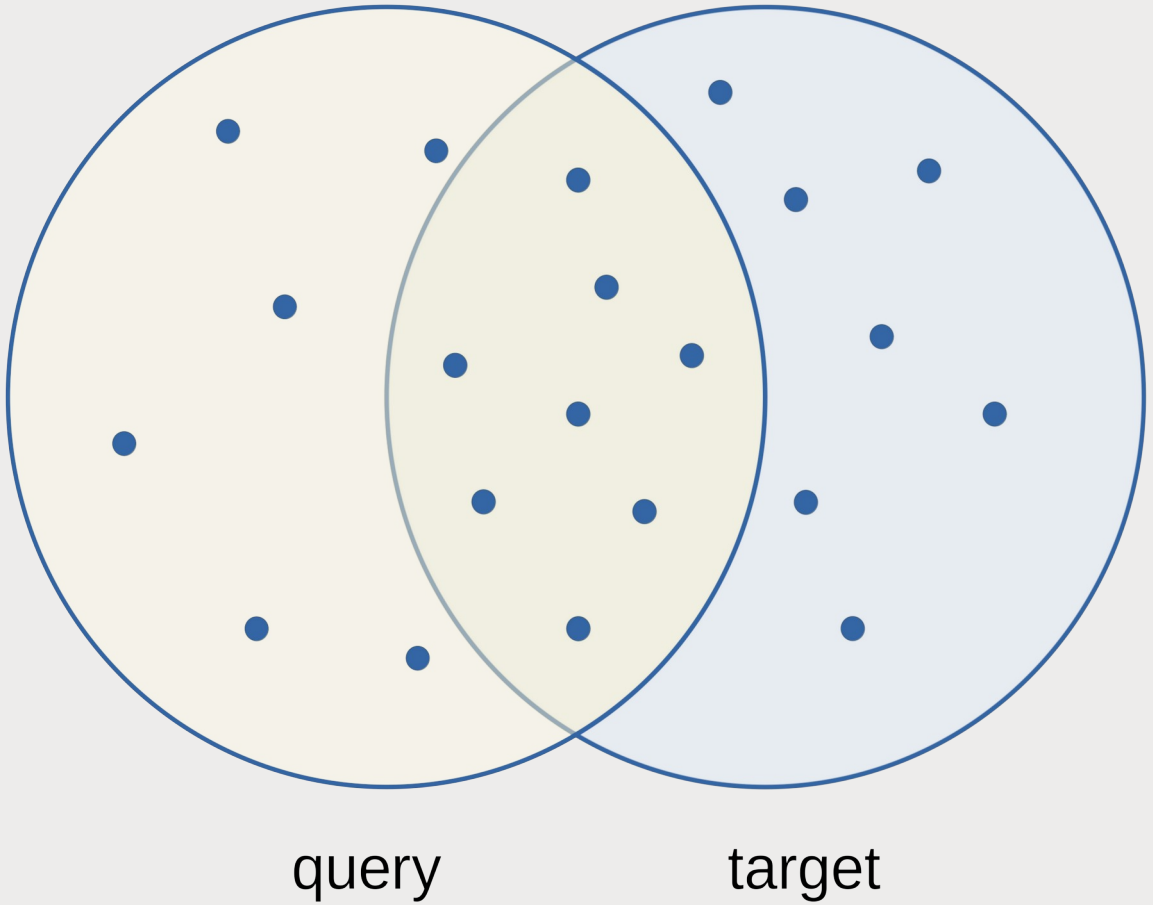
What can your do with the extended tool box?

- Detect AMR caused by point mutations
- MLST by mapping reads on profile alleles
- Discover and submit novel MLST alleles
- Accurate SNP detection for phylogenetic analysis
- Analyse outbreaks by assessing relatedness of isolates
- Identify virus strains, detect novel variants



What is a k-mer?

- A *k*-mer is a subsequence of length *k*
- The *k*-mers of a sequence are *all* its subsequences of length *k*
- The *k*-mer composition of a sequence is like a “spectrum”
  - Can be used to identify a sequence
- Computers can deal with k-mers very efficiently
  - Assemblers, mappers, binners all make use of k-mers
  - Can scale to extremely large databases



Alignment-free methods

- Alignment-free: no assembly, no mapping, no reference – just count *k*-mers
- For instance:
  - Tally every distinct *k*-mer in the query
  - Compare with the *k*-mer composition of the targets
  - Pick the target that shares the most *k*-mers with the query
- Applicable to reads and contigs
- Fast even with huge databases

The Fleming Fund Regional Grants

Summarising ...

- FASTQ files contain reads and a quality score for every base
- FASTA files contain plain sequences (genes, genomes, ...)
- Assembly reconstructs the genome from reads
- Mapping piles up aligning reads on a reference sequence
- Alignment and mapping underlie many genomic analyses
- K-mers enable rapid search through large datasets
- Alignment-free methods combine speed and huge data sets

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.