### Module 1 Bioinformatics Basics

Taking a look behind the curtain



17 February 2021

Marco van Zwetselaar

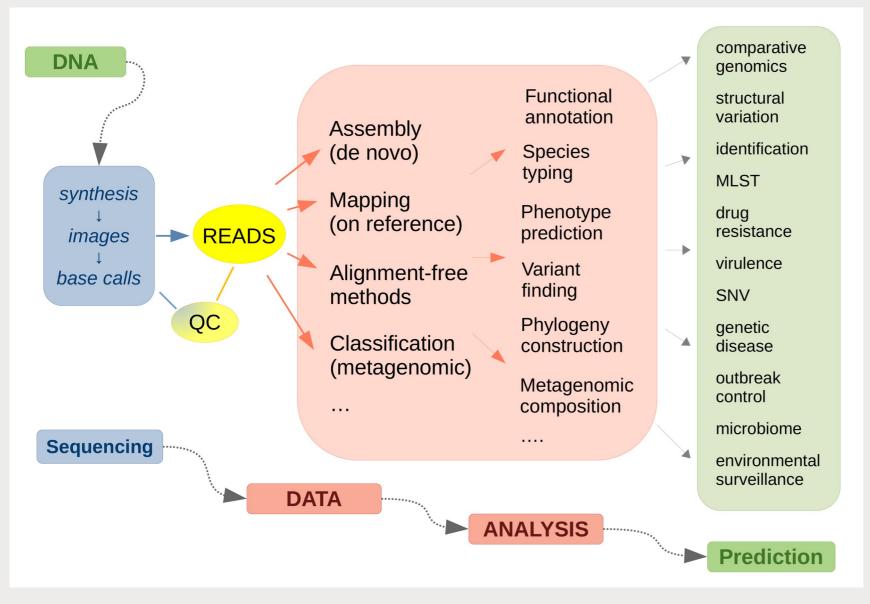
Kilimanjaro Clinical Research Institute





### **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?



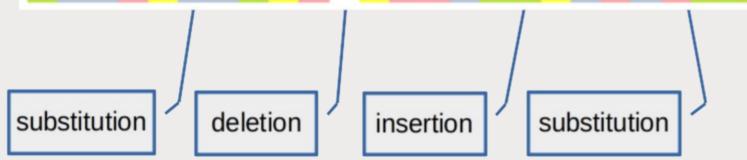


#### **Bioinformatician Bird's Eye View**

- 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'



## CTTAGATCGACGAATC-GTATGCCA CTTAGTTCGA-GAATCCGTATACCA



#### 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
  - • •

AST <sup>®</sup> » blas	tn suite				Home Recent	Results Saved S
olastn b	lastp blastx	tblastn t	tblastx	Standard Nucleotide BLAST		
		BLASTN	programs s	search nucleotide databases using a nucleotide query.	nore	Res
Enter Query Enter accession	Sequence number(s), gi(s), or FAST	A sequence(s) 🛱 <u>cle</u>	ar	Query subrange		
	AATCCCTATAGGCACTTGTG GGCCTAATCGGCTACAAAG/ Browse No			From To	Descript	umns added to t ion Table ct Columns' or 'Manag
Job Title	Enter a descriptive title for	or your BLAST search 🖥	-0 ;9			
-	nore sequences					
<ul> <li>Align two or n</li> <li>Choose Sea</li> <li>Database</li> </ul>	rch Set Standard database	•		ases O Genomic + transcript databases O Bet	acoronavirus	
Choose Sea	Standard database     RefSeq Representa	tive genomes (refseq e or idcompletions v	L_represen	tative_genomes		
Choose Sea Database Organism Optional Exclude	Standard database     RefSeq Representa     Enter organism nam	tive genomes (refseq e or idcompletions v name, binomial, or tax i	<b>represen</b> will be sugg id. Only 20	tative_genomes		
Choose Sea Database Organism	<ul> <li>rch Set</li> <li>Standard database</li> <li>RefSeq Representa</li> <li>Enter organism nam</li> <li>Enter organism common</li> </ul>	tive genomes (refseq e or idcompletions v name, binomial, or tax i Uncultured/environ vpe material	<b>represen</b> will be sugg id. Only 20	tative_genomes	)	



#### **BLAST**

- <u>https://blast.ncbi.nlm.nih.gov</u>
- 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%)

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

• Percentage identity

AATCCTTAGTGGGTCGAGATCGTCGATCCGTAAAAATACCACGTATACCAGGGTAATCCGTCGC TCTTAGTTCGAGAATCGGTAGTGGGTCGAGATCGTCGATCCGTAAAAATAC-ACGTATACCACGGTCATCCGTCGCGTAG

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



Multi	ble A	lignme	ent

- 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

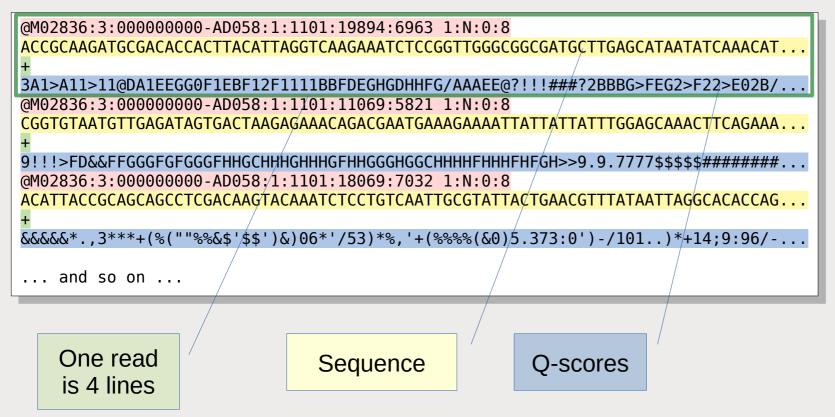
		* .	:		. *	: : :		
Q5E940 BOVIN	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDDYPK	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> QIRMSLRGK·	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 HUMAN	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDDYPK	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> Q IRMS LRGK ·	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 MOUSE	M <mark>P</mark> REDR <mark>A</mark> TW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDDYPK	CFIV <mark>G</mark> A <mark>D</mark> NV <mark>GS</mark> I	K <mark>QMQ</mark> Q IRMS LRGK-	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 RAT	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDD <mark>YP</mark> K	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> QIRMSLRGK-	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 CHICK	M <mark>P</mark> REDR <mark>A</mark> TW	KSNY <mark>F</mark> MK <mark>I</mark> I	QLLDDYPK	CFVV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> QIRMSLRGK-	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLAO RANSY	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDDYPK	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> QIRMSLRGK·	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNSALE	76
Q7ZUG3_BRARE	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLDDYPK	CFIV <mark>G</mark> A <mark>D</mark> NV <mark>GS</mark> I	K <mark>QMQ</mark> T IRLS LRGK·	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 ICTPU	MPREDRATW	KSNY <mark>F</mark> LK <mark>I</mark> I	QLLND <mark>YP</mark> K	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> T IRLS LRGK·	- AIV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPALE	76
RLA0 DROME	MVRENKAAW	K <mark>aqy</mark> fik <mark>v</mark> v	ELFDEF <mark>P</mark> K	CFIV <mark>G</mark> ADNV <mark>GS</mark> I	K <mark>QMQ</mark> N IRTS LRG <mark>L</mark> ·	- AVV LM <mark>GKNT</mark> MMR <mark>K</mark>	AIRGHLENNPQLE	76
RLA0 DICDI	MS <mark>G</mark> A <mark>G</mark> -SKR	KKLF <mark>I</mark> EKA <mark>I</mark>	KLFTT YDK	MIVAEA <mark>d</mark> fv <mark>gs</mark> :	S <mark>QLQ</mark> KIRKSIRGI	- <mark>GAVLMGK</mark> K <mark>T</mark> MIRK	VIRDLADSKPELD	75
Q54LP0_DICDI	MS <mark>G</mark> A <mark>G</mark> -SKR	KNVF <mark>I</mark> EKAI	KLFTT <mark>Y</mark> DK	MIVAEA <mark>d</mark> fV <mark>gs</mark> :	S <mark>QLQ</mark> KIRKSIRGI	- <mark>GAVLMGK</mark> K <mark>TMI</mark> RK	VIRDLADSKPELD	75
RLA0 PLAF8	MAKLSKQQK	KQMY <mark>I</mark> EKLS	S <mark>LI</mark> QQ <mark>Y</mark> SK	ILIVHV <mark>d</mark> nv <mark>gs</mark> i	N <mark>QM</mark> ASVRKSLRGK	- <mark>ATILM</mark> GKNTRIRI	ALKKNLQAVPQIE	76
RLA0_SULAC	<mark>MIG</mark> LAVTTTKK <mark>IA</mark> KW	KVDE <mark>V</mark> AE <mark>L</mark> I	EKLKTHKT	IIIAN I <mark>EG</mark> F <mark>P</mark> AI	DKLHE IRKKLRGK-	- ADI <mark>KVTKN</mark> NLFN J	ALKNAG YDTK	79
RLA0 SULTO	MRIMAVITQERKIAKW	K I E E <mark>V</mark> KE <mark>L</mark> e	Q <mark>K L</mark> RE <mark>Y</mark> HT	IIIAN I <mark>EG</mark> F <mark>P</mark> AI	DK <mark>L</mark> HD <mark>IR</mark> KK <mark>MRG</mark> M-	- AE IKVTKNTLFG J	AAKNAGLDVS	80
RLA0_SULSO	<mark>M</mark> KR <mark>L</mark> ALALKQRK <mark>VA</mark> SW	K LEE <mark>v</mark> ke L I	ELIKNSNT	ILI <mark>G</mark> NL <mark>EG</mark> F <mark>P</mark> AI	DK <mark>L</mark> HE <mark>IR</mark> KK <mark>LRG</mark> K ·	- A <mark>TIKVTKNT</mark> LFKI	AAKNAGIDIE	80
RLA0 AERPE	MSVVSLVGQMYKREK <mark>PIP</mark> EW	KTLM <mark>L</mark> RE LE	ELFSKHRV	VLF ADLT GTPT I	FV <mark>VQ</mark> RV <mark>R</mark> KKLWKK	- <mark>YP</mark> MMVAKKRIILF	AMKAAGLE LDDN	86
RLA0 PYRAE	-MMLAIGKRRYVRTRQYPAR	KVKI <mark>V</mark> SE <mark>A</mark> I	ELLQK <mark>YP</mark> Y	VFLFDLH <mark>G</mark> LS <mark>S</mark> I	RI <mark>LHE YR</mark> YR <mark>LR</mark> RY-	- <mark>GVIKIIKPT</mark> LFKJ	AFTKVYGGIPAE	85
RLA0 METAC	MAEERHHTEHIPQW	KKDE <mark>I</mark> EN <mark>I</mark> K	ELIQS <mark>h</mark> kvi	F <mark>GMVGIEG</mark> ILA	FK <mark>MQ</mark> K IRRD LKDV ·	- AVLKV <mark>SRNT</mark> LTEF	ALNQLGETIP	78
RLA0 METMA	MAEERHHTEH <mark>IP</mark> QW	KKDE <mark>I</mark> EN IK	E <mark>liqsh</mark> kv	F <mark>gmvrieg</mark> ila:	FK <mark>IQ</mark> KIRRDLKDV ·	-AVL <mark>KVSRNT</mark> LTEF	ALNQLGESIP	78
RLA0 ARCFU	PPEY	KVRA <mark>v</mark> ee ik	(RMISSK <mark>P</mark> V	VAIVSFRNV <mark>P</mark> A	G <mark>QMQ</mark> K IRRE FRGK	- AE IKVVKNTLLEF	ALDALGGDYL	75
RLA0 METKA	MAVKAKGQPPSGYEPKVAEW	KRRE <mark>v</mark> ke <mark>l</mark> k	ELMDE YEN	V <mark>G</mark> LVDL <mark>EGIP</mark> AI	PQLQE IRAK LRERI	OTIIRM <mark>SRNT</mark> LMRI	ALEEKLDERPELE	88
RLA0 METTH	MAHVAEW	KKKE <mark>V</mark> QE <mark>L</mark> H	ID <mark>LIK</mark> GYEV	VGIANLADIPAI	R <mark>QLQ</mark> KMRQT LRDS ·	-ALI <mark>RM<mark>SK</mark>KTLISI</mark>	ALEKAGRELENVD	74
RLA0 METTL	<mark>M</mark> ITAESEHK <mark>IAP</mark> W	K I E E <mark>V</mark> NK <mark>l</mark> k	ELLKNGQI	VALVDMMEV <mark>P</mark> AI	R <mark>QLQ</mark> E IRDK IR-GI	CM <mark>TLKMSRNT</mark> LIEF	AIKE VAEET GNPEFA	82
RLA0 METVA	<mark>M</mark> IDAKSEHK <mark>IAP</mark> W	K I E E <mark>V</mark> NA <mark>l</mark> k	ELLKS ANV	IALIDMME V <mark>P</mark> AV	VQLQE IRDK IR-DQ	OM <mark>tlkmsrnt</mark> likf	AVEEVAEETGNPEFA	82
RLA0 METJA	METKVKAH <mark>VAP</mark> W	K I E E <mark>V</mark> KT <mark>L</mark> K	G <mark>li</mark> ksk <mark>p</mark> v	VAIVDMMDV <mark>P</mark> AI	PQLQE IRDK IR-DI	KVKL <mark>RMSRNT</mark> LIIF	ALKE AAE E LNNPKLA	81
RLA0 PYRAB	MAHVAEW	KKKE <mark>v</mark> ee <mark>l</mark> a	NLIKS <mark>YP</mark> V	IALVDVSSM <mark>P</mark> A	Y <mark>PL</mark> SQ <mark>MR</mark> RL IREN(	GGLLRV <mark>SRNT</mark> LIEI	AIKKAAQELGKPELE	77
RLA0 PYRHO	MAHVAEW	KKKE <mark>v</mark> ee <mark>l</mark> a	K <mark>LIKS YP</mark> V	IALVDVSSM <mark>P</mark> A	Y <mark>PL</mark> SQ <mark>MR</mark> RL <mark>IR</mark> EN(	GGLLRV <mark>SRNT</mark> LIEI	AIKKAAKEL <mark>G</mark> KPELE	77
RLA0 PYRFU	MAHVAEW	KKKE <mark>v</mark> ee <mark>l</mark> a	N <mark>LI</mark> KS <mark>YP</mark> V	VALVDVSSM <mark>P</mark> A	Y <mark>PL</mark> SQ <mark>MR</mark> RL <mark>IR</mark> ENI	N <mark>GLLRV<mark>SRNT</mark>LIEI</mark>	AIKKVAQEL <mark>G</mark> KPELE	77
RLA0 PYRKO	MAHVAEW	KKKE <mark>v</mark> ee <b>l</b> a	N IIKS <mark>YP</mark> V	IALVDVA <mark>G</mark> V <mark>P</mark> A	Y <mark>PL</mark> SKMRDKLR-GI	KALL <mark>RVSRNT</mark> LIEI	AIKRAAQELGQPELE	76
RLA0 HALMA	<mark>MSA</mark> ESERKTET <mark>IP</mark> EW	KQEE <mark>V</mark> DAIV	EMIESYES	V <mark>G</mark> VVNIA <mark>gip</mark> si	R <mark>QLQDMR</mark> RDLHGT ·	- AEL <mark>RV SRNT</mark> LLE F	ALDDVDDGLE	79
RLA0 HALVO	<mark>MSESEVRQTEVIP</mark> QW	KREE <mark>V</mark> DELV	DFIESYES	V <mark>g v v g v ag i p s</mark> i	R <mark>QLQ</mark> SM <mark>R</mark> RE LH <mark>G</mark> S·	- AAV RM <mark>SRNT</mark> LVN F	ALDEVNDGFE	79
RLA0 HALSA	<mark>MSA</mark> EEQRTTEE <mark>VP</mark> EW	KRQE <mark>V</mark> AE L V	DLLET YDS	V <mark>GVVNVTGIPS</mark> I	K <mark>QLQ</mark> DMRR <mark>GL</mark> HGQ∙	-AAL <mark>RM<mark>SRNT</mark>LLVF</mark>	ALEE AGDGLD	79
RLA0 THE AC	MKE VSQQ	KKE L <mark>V</mark> NE <mark>I</mark> T	QRIKASRS	VAIVDTAGIRTI	R <mark>QIQ</mark> D <mark>IRG</mark> KN <mark>RG</mark> K·	- INLKVIKKTLLFK	ALENLGDEKLS	72
RLA0 <sup>THEVO</sup>	MRKINPK	KKE I <mark>v</mark> se <mark>l</mark> a	QD ITKSKA	VAIVDIK <mark>G</mark> VR <mark>T</mark> I	R <mark>QMQ</mark> DIRAKNRDK	-VKIKVVKK <mark>T</mark> LLF	ALDSINDEKLT	72
RLA0 PICTO	MTE <mark>PA</mark> QW	KIDF <mark>V</mark> KNLE	NEINSRKV	AAIVSIK <mark>G</mark> LRNI	NEFOKIRNSIRDK	-ARIKV <mark>SR</mark> ARLLRI	AIENTGKNNIV	72
ruler	1	3		. 40	50	70		

https://commons.wikimedia.org/wiki/File:RPLP0\_90\_ClustalW\_aln.gif (CC BY-SA 3.0)



### **Reads and Assemblies**

- $\bullet \hdots$  and FASTQ & FASTA
- De novo assembly
- Hybrid assembly



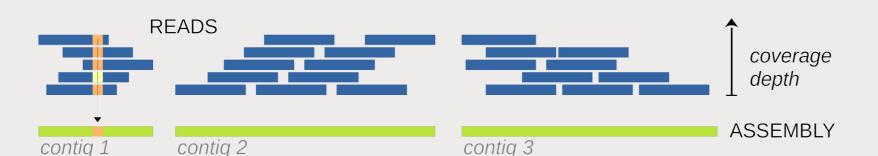


#### FASTQ

- 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



The Fleming Fund Regional Grants

#### 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)?

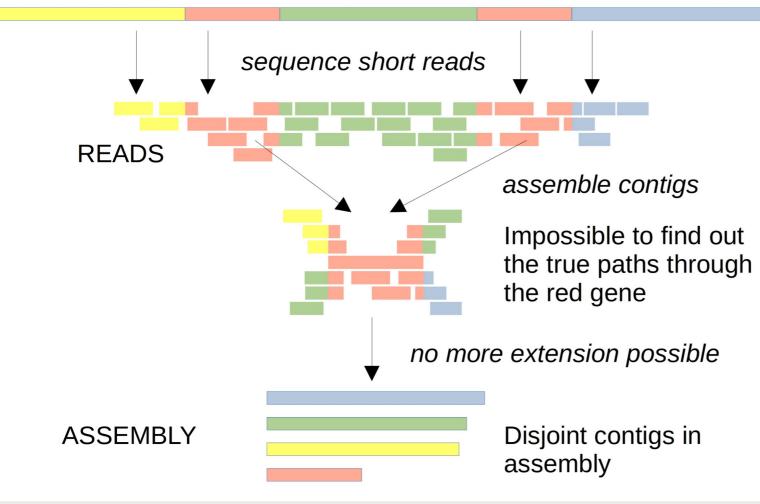
identifier description sequence data > >pKCRI-43 Acinetobacter baumannii circular plasmid pKCRI-43-1 **GGAAATTCTAGAAAATCTCTATGATGAAAATATTCAGATAGAGATTACAGCTATAGCCAAATAG** TGTCTTCAACCTATTAATTCCAAGTCATGTATGAAGCCAAAGAAGAAATAATTTATTATTGAGT TTCGCTAAAATTAACATAGTACATGTTATACGAAGTCAAAAATGGGAGCGTAAGCTCCCATTT AGCCTTTGCCTTGAGCTTTAAACCATGCAAGTACATCAGCATCTAAACGAATGGAAGTCTGTTG CTTCACTGGGCGATAGAATCGATTATGGCGTACAG... >R0015 43 1 Acinetobacter baumannii KCRI strain 43 contig 1 TGAACTCTTCATCTTTTTTTTTTTATTAAAGAGTCAGATACCTGAAACACACGAATTTTTGGTTTATT ACGAACTCTTCATCTTTTTTTTTTATTAAAGAGTCAGATACCTGAAACACACGAATTTTTGGTTTAT TACCTCTAAAGTTGCACTCGCCGCCTTAAAATTCTCACTCGTAAAATGGGTAAACGTTTTACCT ACCGCATTATGATAAACCAAAGCATCCAAATCAGCTGCTTCAAGACTTGCTGTTAAATCAGCAT CATAGCCATGCGTTTGATATGGAAATAAAGCAAATGTTGGCAATAATGAAGCCCGAATTGCTGT TGCTCTAAAACCTCAGCACGCTTAGATTCTTCAGTTACAGGCAGATTTTTATACCCACCACAGA ACATACGGTTTACATCGTCATGTACATAACGCTTGCCTTGACGCATCGCATAGGGATTACCCAA 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

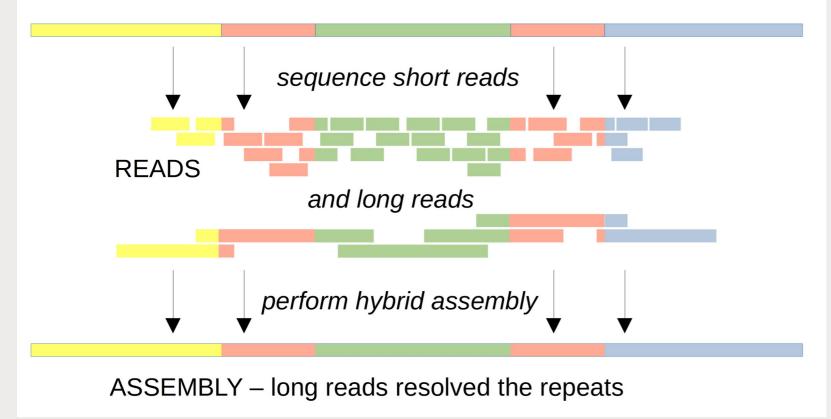
#### Ground truth: genome with two (near-identical) copies of the red gene





#### 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!

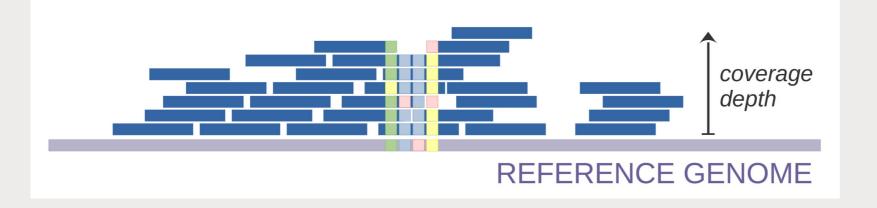
- Assemble the genome of your isolate
- Obtain FASTA files with the sequences of known AMR genes
  - Freely downloadable, e.g. DTU CGE (genomicepidemiology.org)
- BLAST the genes against your assembled genome
- List all high scoring alignments with sufficient coverage and identity
- Voilà, your own AMR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

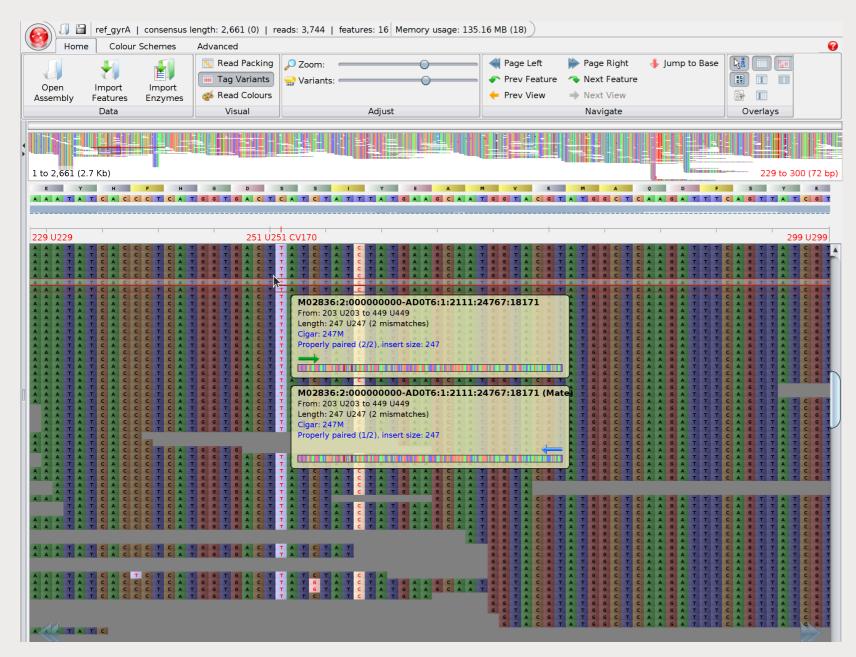
### • But ... what about point mutations?





- 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 SNPs (e.g. in phylogeny)
- The basis for variant calling and obtaining consensus sequence







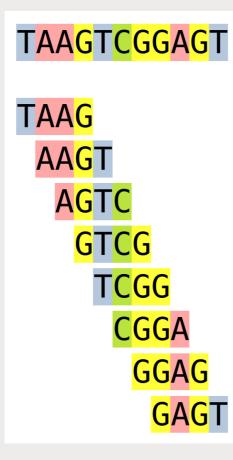
#### **Illustration (tool: Tablet)**

- 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 TCA (S) on reference is TTA (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





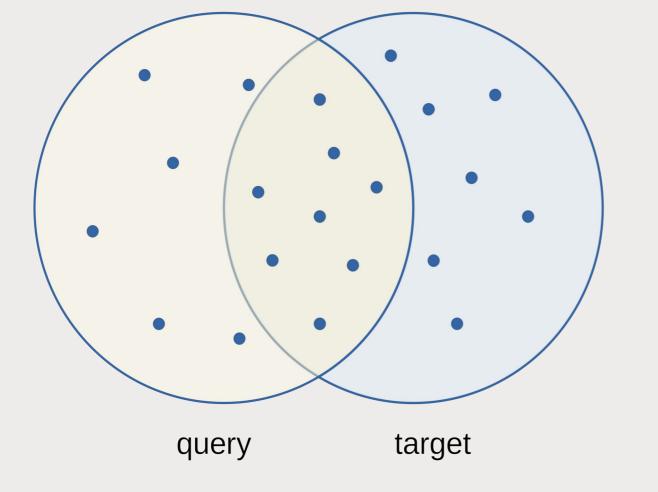
#### What is a k-mer?

- 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: 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





### **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



# Thank you



NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH UNIVERSITY OF GHANA, LEGON





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



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.