Module 2
Bioinformatics Recap

… and some new material

The Fleming Fund
Regional Grants

26 March 2021

Marco van Zwetselaar

Kilimanjaro Clinical Research Institute

Recap of the terms ...

- **Reads** are the nucleotide sequences of your library fragments
  - After trimming off the adapters, barcodes, and low quality ends
- Reads are stored in **FASTQ** files
  - For every nt a Q-score indicating probability of being correct
- **FASTA** files contain sequences of **nucleotides or amino acids**
  - Such as the **assembled** genome of your sample
  - The sequences are called **contigs**

```
LOCUS       NC_002505            2961149 bp    DNA     circular CON 03-AUG-2016
DEFINITION  Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I, complete
            sequence.
ACCESSION   NC_002505
VERSION     NC_002505.1
DBLINK      BioProject: PRJNA57623
            Assembly: GCF_000006745.1
KEYWORDS    RefSeq.
SOURCE      Vibrio cholerae O1 biovar El Tor str. N16961
  ORGANISM  Vibrio cholerae O1 biovar El Tor str. N16961
            Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales;
            Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 2961149)
  AUTHORS   Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L.,
            ...
FEATURES             Location/Qualifiers
     source          1..2961149
                     /organism="Vibrio cholerae O1 biovar El Tor str. N16961"
                     /mol_type="genomic DNA"
                     /strain="N16961"
                     /serotype="O1"
                     /biotype="El Tor"
     gene            7397..8815
                     /gene="dnaA"
     CDS             7397..8815
                     ...
                     /note="binds to the dnaA-box as an ATP-bound complex at
                     ...
                     /product="chromosome replication initiator DnaA"
                     /protein_id="NP_062596.1"
                     /translation="MSEGIVSSSLWLQCLQRLQEELPAAEFSMWVRPLQAELNDNTLT
                     LFAPNRFVLDWVRDKYLNNINRLLMEFSGNDVPNLRFEVGSRPVVAPKPAPVRTAADV
                     AAESSAPAQLAQRKPIHKTWDDDSAAADITHRSNVNPKHKFNNFVEGKSNQLGLAAAR
                     QVSDNPGAAYNPLFLYGGTGLGKTHLLHAVGNAIVDNNPNAKVVYMHSERFVQDMVKA
...
ORIGIN
        1 tggagtatta acagaaaatt gataccaaac gaacaaagtt aagtataaaa accgcgttta
       61 aataacccac atattcttcg ataaggagaa aacattttaa atattacagt gtcacttatt
      121 tacaatgtaa agccacgttt tgaagtgatg atgaataaat aaaagcgagc cgtaagcgga
      181 acgattaaac cgagccacta agttacggtg aatgccattc tgattgaaat gatgcgcagg
```

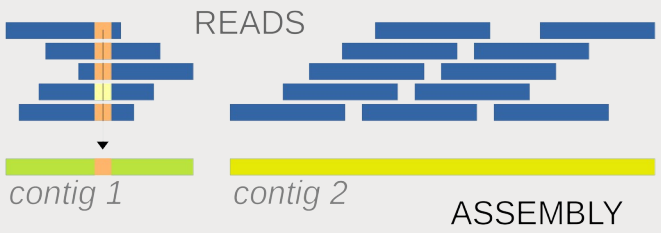
The Fleming Fund
Regional Grants

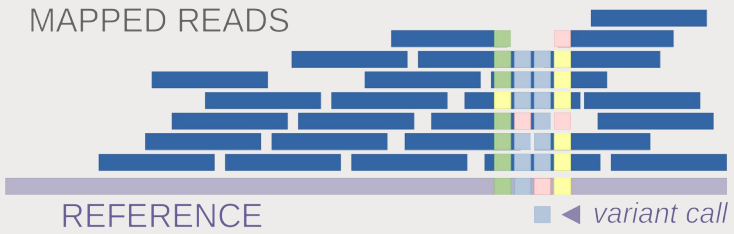
Not everything is FASTA/Q! GBFF & EMBL

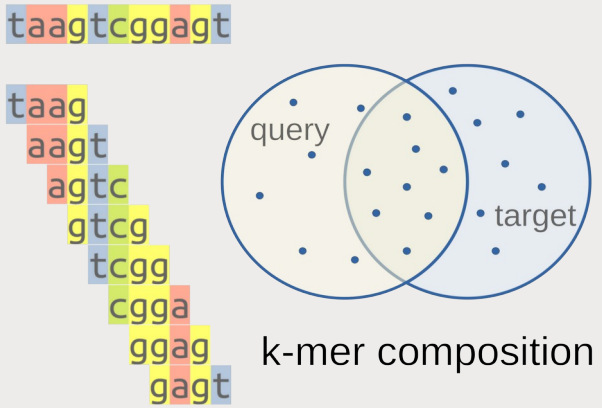
- Metadata and references
- Annotated genome
- Protein products
- Nucleotide sequence
- Very legible and (for reference sequences) can have lots of information

Recap of the core operations ...

- **Assembly** is reconstructing the genome(s) from reads
- **Mapping**: pile up aligning reads on a target sequence (SAM/BAM)
  - **Variant calling**: determine variants relative to reference (VCF)
- Alignment-free: match on **k-mer** composition of query and target



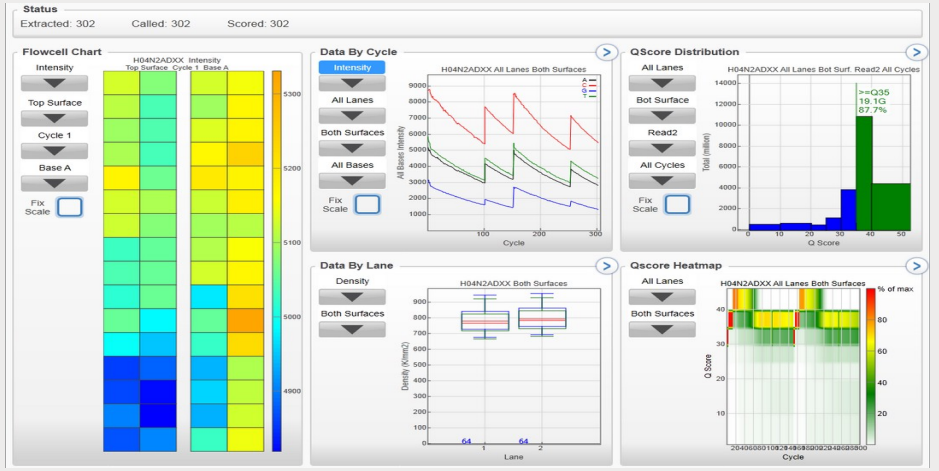


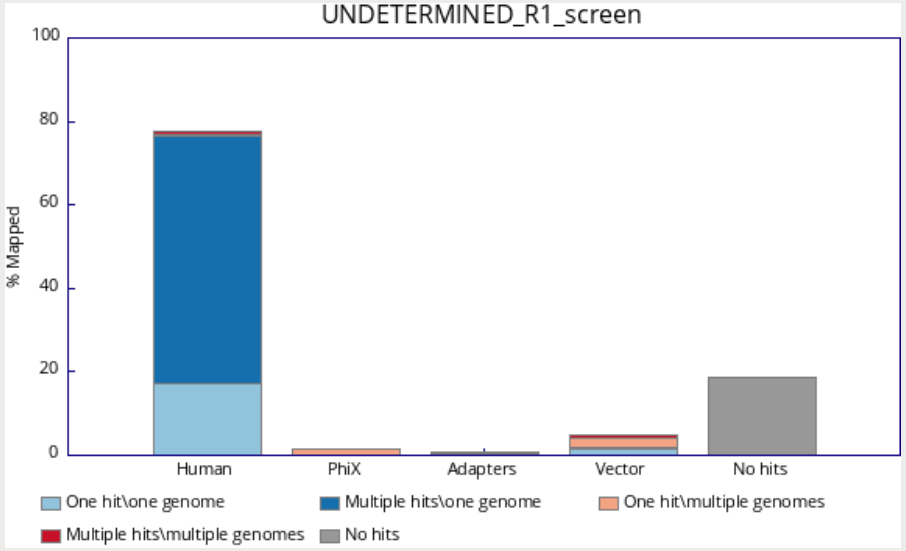


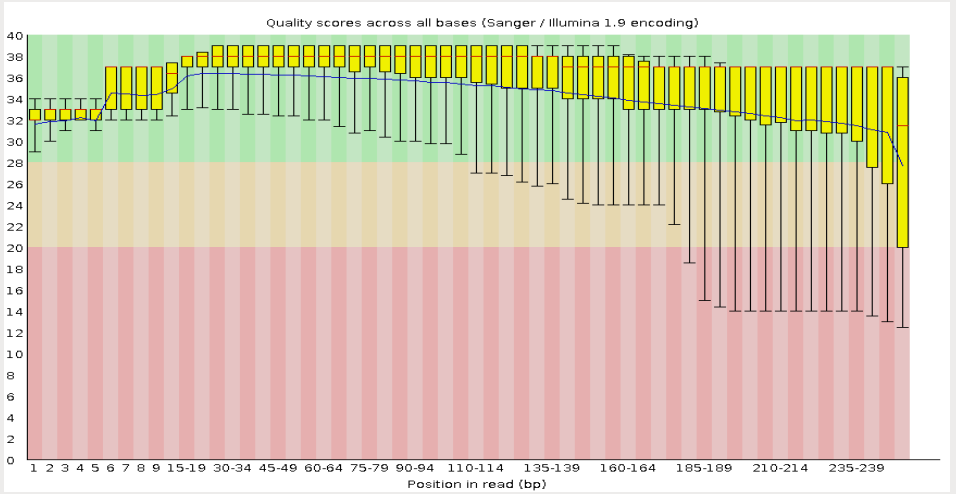
The Fleming Fund
Regional Grants

QA/QC on Reads

- MiSeq Reporter
- FastQC
- Fastq-Screen







The Fleming Fund
Regional Grants

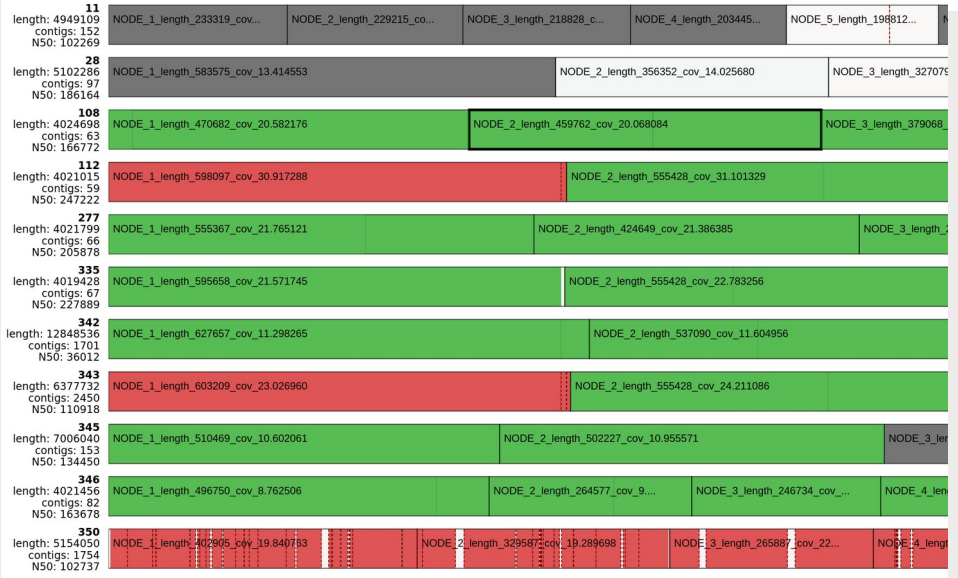
QA on Assemblies

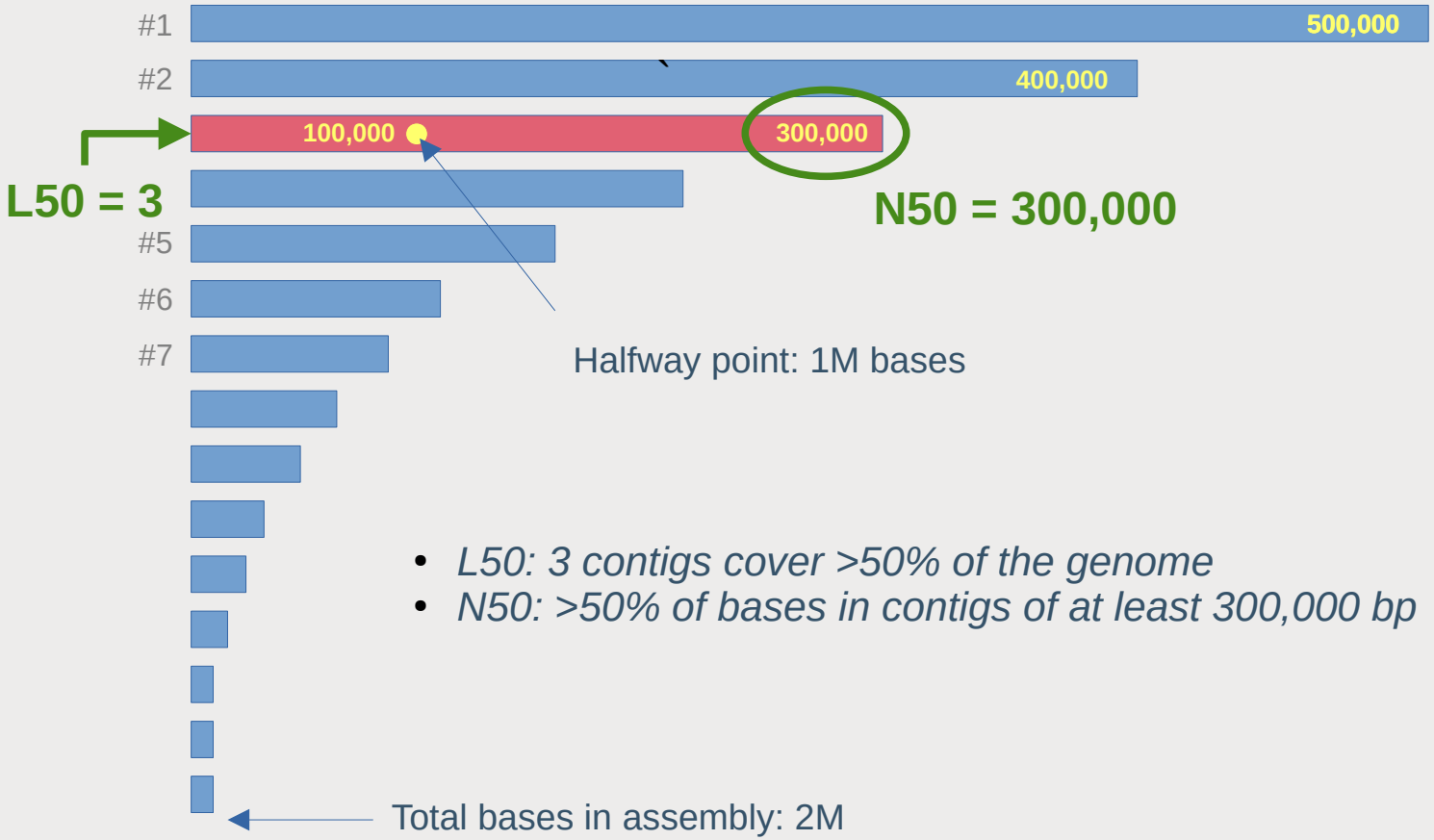
- Total assembly length
- Number of contigs
- Largest contig
- N50, L50 (N75, L75)
- Tool: Quast

Worst Median Best ☑ Show heatmap

Statistics without reference	116	123C	164C	186	28	309C	33	348C
# contigs	89	92	21	89	71	68	135	43
# contigs (>= 0 bp)	125	151	56	125	141	129	201	75
# contigs (>= 1000 bp)	75	80	17	75	57	55	118	42
# contigs (>= 5000 bp)	60	66	15	60	43	47	91	35
# contigs (>= 10000 bp)	49	52	14	52	40	38	70	32
# contigs (>= 25000 bp)	39	42	14	38	32	30	44	26
# contigs (>= 50000 bp)	25	24	11	24	22	25	26	23
Largest contig	478 004	389 111	935 746	320 898	462 529	389 111	258 355	304 793
Total length	4 093 638	4 086 582	3 727 924	3 900 552	3 966 414	4 091 248	3 823 598	3 283 148
Total length (>= 0 bp)	4 103 947	4 104 271	3 734 448	3 910 164	3 983 907	4 107 150	3 842 758	3 294 144
Total length (>= 1000 bp)	4 083 883	4 078 476	3 725 612	3 890 948	3 957 554	4 082 580	3 812 878	3 282 506
Total length (>= 5000 bp)	4 043 290	4 043 655	3 722 812	3 852 310	3 929 833	4 062 064	3 761 626	3 262 775
Total length (>= 10000 bp)	3 957 639	3 936 655	3 717 255	3 799 623	3 908 859	3 995 025	3 597 355	3 243 074
Total length (>= 25000 bp)	3 783 372	3 782 129	3 717 255	3 548 419	3 775 457	3 870 284	3 166 829	3 151 259
Total length (>= 50000 bp)	3 291 752	3 163 956	3 616 520	3 025 136	3 400 791	3 697 411	2 548 905	3 034 883
N50	121 778	117 340	464 597	124 509	154 847	176 800	75 931	156 290
N90	28 249	26 579	156 723	26 831	40 113	50 653	15 577	52 271
L50	9	10	3	11	8	9	16	9
L90	36	38	8	37	26	25	58	22
GC (%)	38.9	38.9	38.96	39.03	38.95	38.89	39.11	39.71
Mismatches								
# N's per 100 kbp	0	0	0	0	0	0	0	0
# N's	0	0	0	0	0	0	0	0

Contig size viewer. For better performance, only largest 1000 contigs of each assembly were loaded





- *L50: 3 contigs cover >50% of the genome*
- *N50: >50% of bases in contigs of at least 300,000 bp*

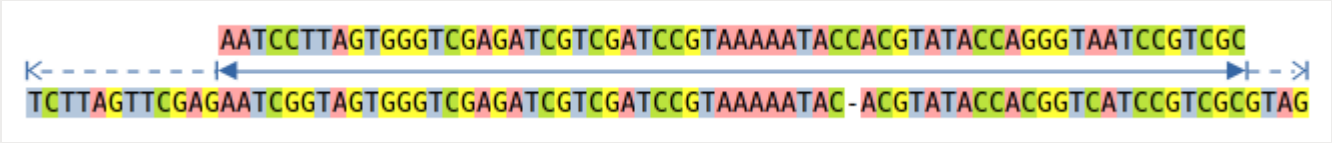
N50, L50: the “halfway contig”

- Order the contigs from longest to shortest, and start walking along the bases ...
- … once you have walked half the total assembly length, you are in the “halfway contig”.
- **N50** is **length** of the halfway contig (long is good)
- **L50** is **number** of the halfway contig (small is good)
- N75, L75, N90, L90: same for 75% and 90% ooints
- Quast adds: NA50, NG50, ...

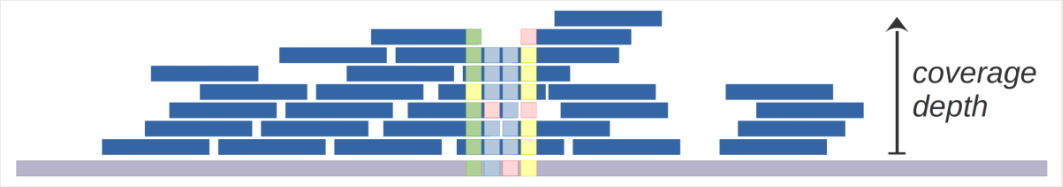
Core Metrics for “hits” (HSPs)

- Pct. Coverage

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

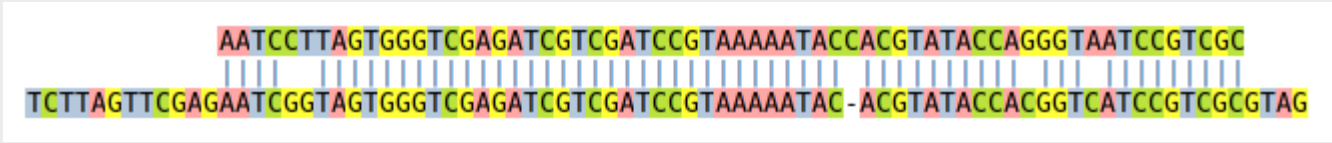


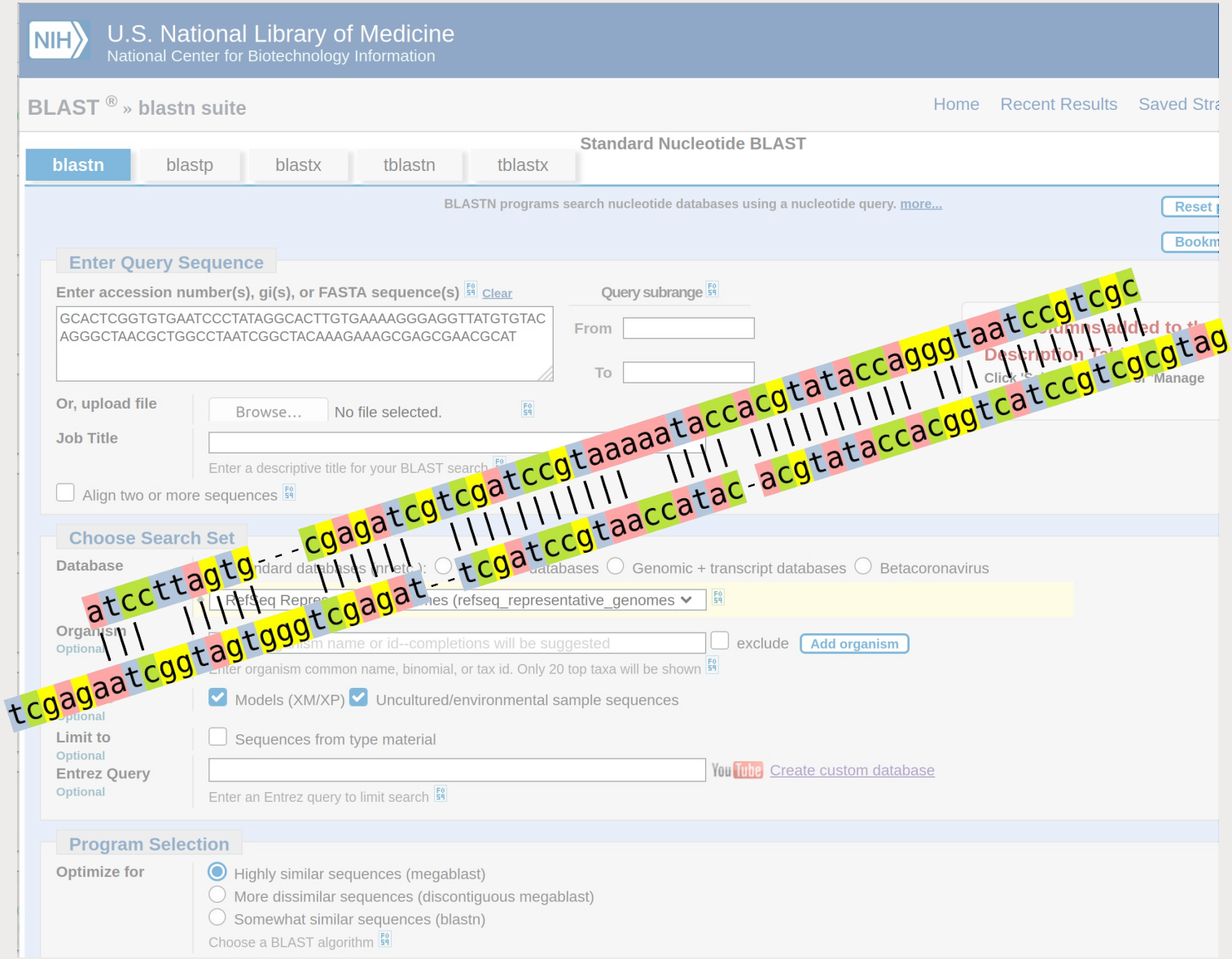
  - Not to be confused with coverage *depth* *(measured in “times”, e.g. 27x)*



- Pct. Identity

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





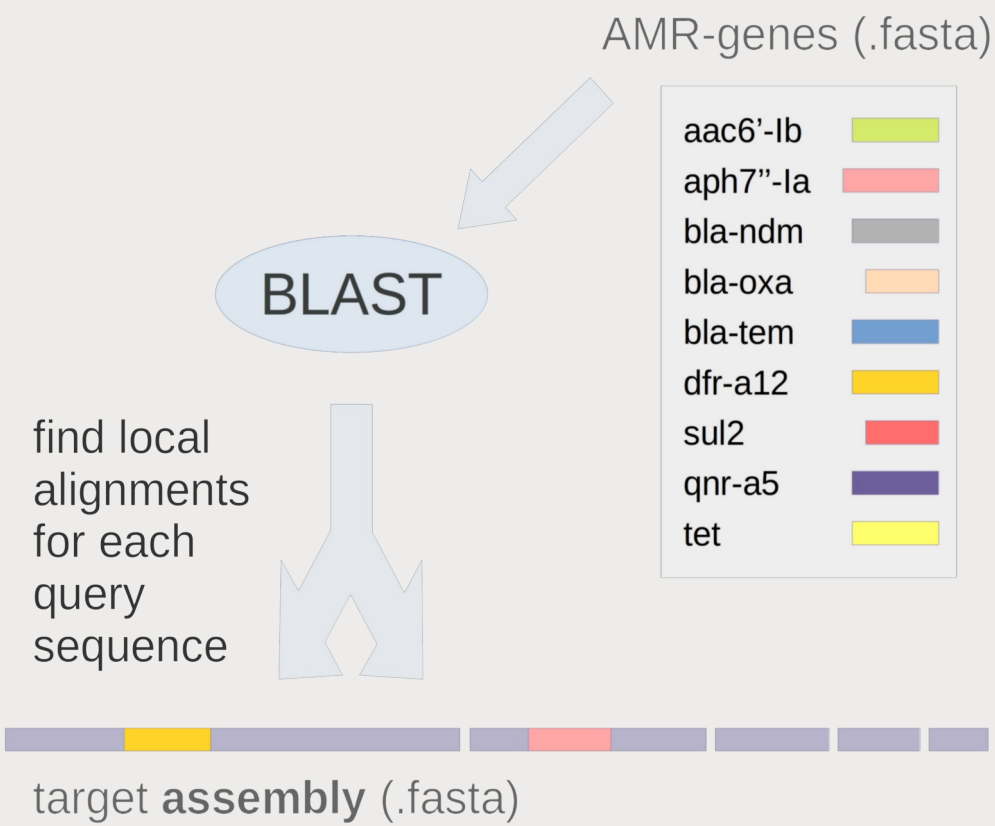
The Fleming Fund
Regional Grants

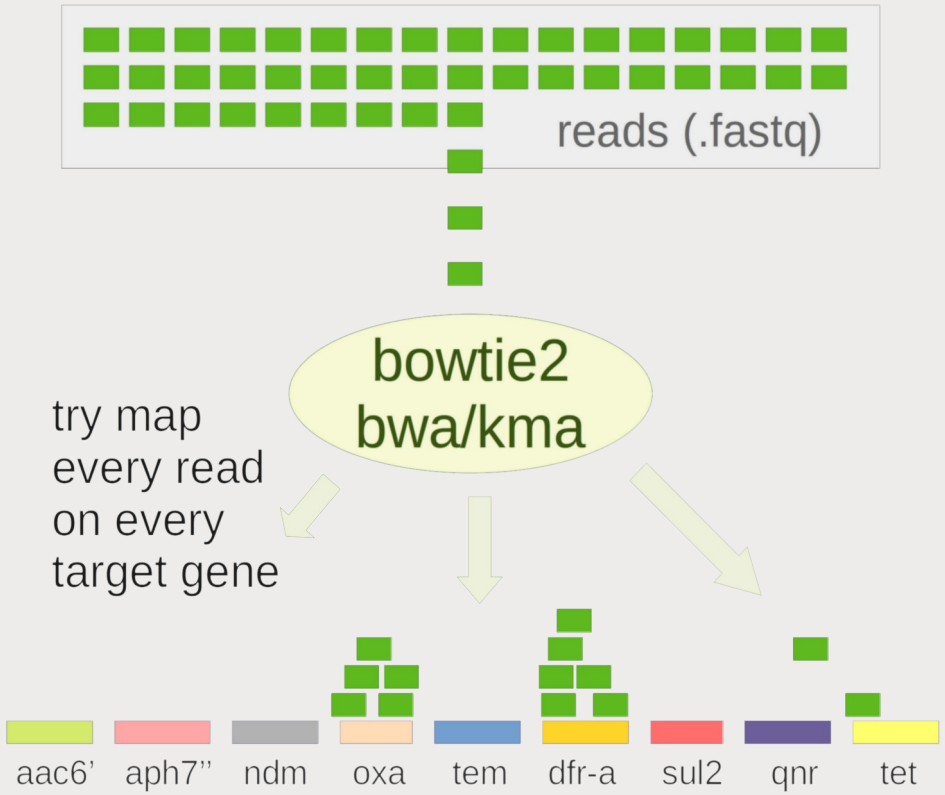
Alignment

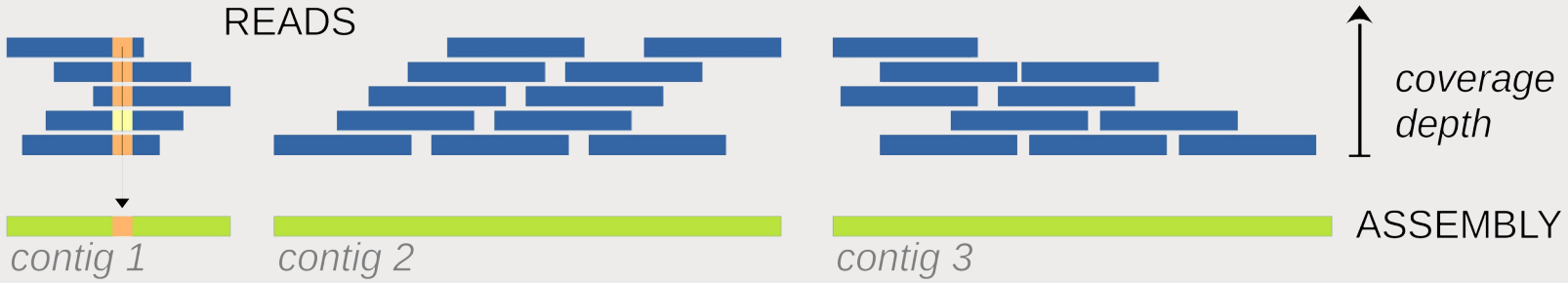
- Scored pairing of sequences
- Score: some weighted edit distance: what does it take to turn one sequence into the other?
  - Substitutions (SNPs)
  - Gaps (inserts & deletes)
- Most famous local alignment search tool: BLAST

Build Your Own *Finder

On FASTA (genome) input vs. On FASTQ (reads) input







Are Reads “Better”?

- Assembly discards a lot of information
- Reads:
  - Retain per base Q value
  - Retain per base read depth
  - Straddle edges of contigs
- But:
  - Assemblies are intuitive
  - Much more “lightweight” to handle
- Trade-off: quick vs thorough

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.