wellcome connecting science

ACORN

Module 3 - Genome Assembly

28-05-2024

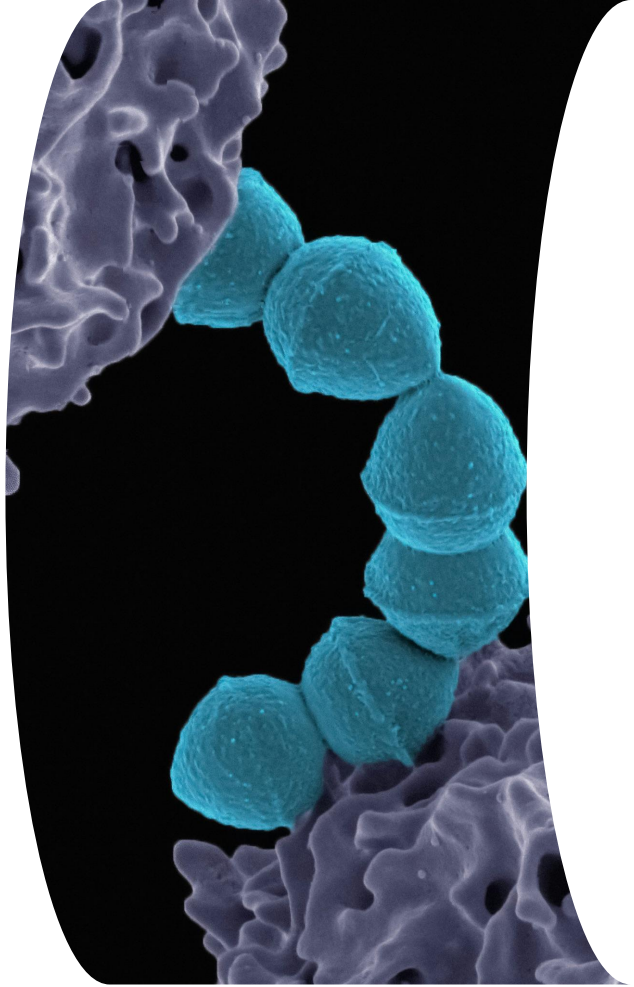
Virtual, Across Africa and Asia

Dr Patrick Musicha

Lecturer and Associated Group Head
Malawi Liverpool Wellcome Programme & LSTM

Genome Assembly

WCS ACORN - Bioinformatics for Antimicrobial Resistance - Virtual Course



Streptococcus Pyogenes
Photo by National Institute of Allergy and Infectious Diseases on Unsplash

What is genome assembly

- Sequence reads-small fragments of DNA sequence output from a sequencing machine.
- Could be
  - -Short reads (Next generation sequencing-Illumina)
  - -Long reads (PacBio SMRT or Oxford Nanopore sequencing)

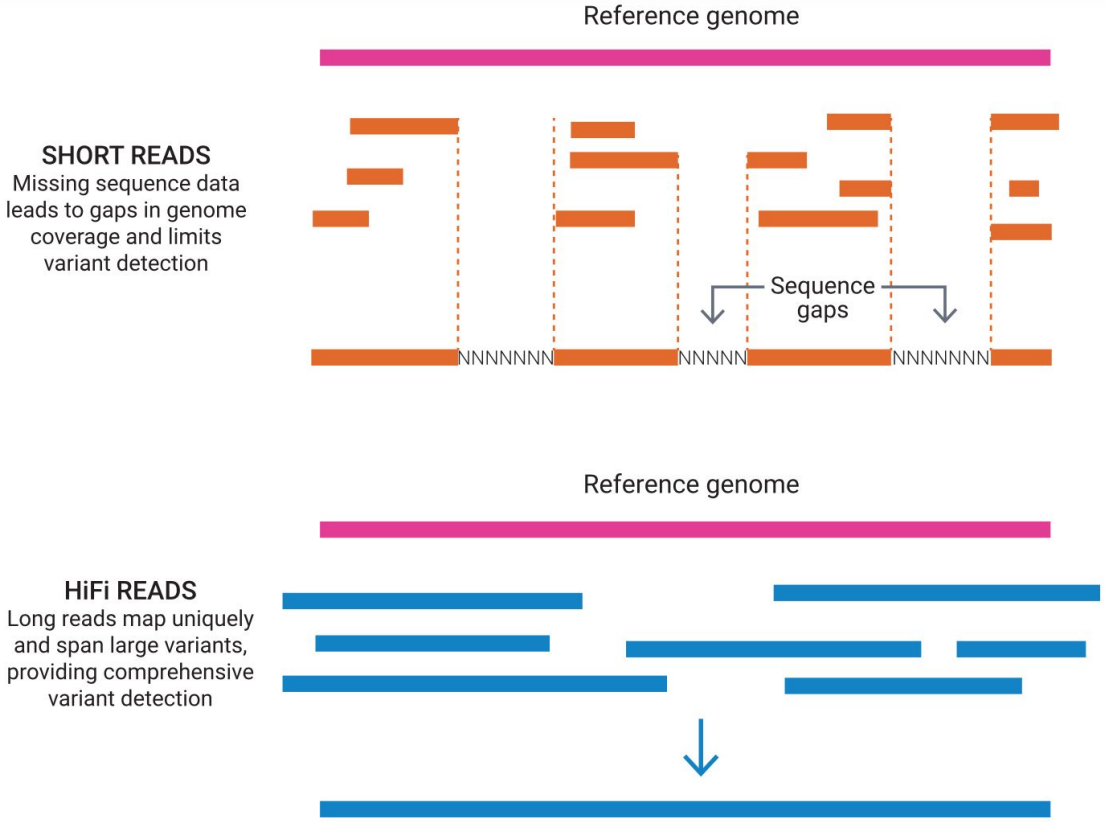
Genome assembly

- Piecing these sequence fragments into a genome (chromosome + any plasmids) sequence is a challenge
  - Reads are somewhat random relative to the genome sequence
  - Repeat sequences and MGE e.g. IS elements and plasmids

How to assemble raw reads

- Can be achieved by
  - Mapping to a reference genome
  - **- *de novo* assembly (without a reference)**.

Genome Assembly



De novo assembly

- Several tools available to perform de novo assembly
  - Velvet (NGS short reads)
  - **Spades** (NGS short reads but can also support hybrid mode)+
    - **Unicycler**
    - **Shovill**
    - **Flye (long reads)**

De novo Assembly

```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@@FF
NATGGCACCATTAAAAAGAATGTTTTATATGGTGTTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIIGCCHIIIGIIIFFHIIDGHIGIIIIIIICHGIIIGGCECEGICFHCECDEFFFFDEEEEEDDDDDCDDCDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATTGTCTTTAATGAAGTACAGTCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGGAAACTTTAAAGTTTGTAATG
+
#4A2<AADHIIIIIIIIIHHIIIIIIIIIIFGIII@GIIFIIIGIIIIEIDHEHIIIHIIIIIIIIIIICHIIIHHEEDFFFFFEEEECEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCCTATCTAAAAAAGAATCAAAAAAGGGATAGTACAGAGGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

```
>EP38001 (+) Ce hist. H1 his-24; range -299 to 100.
GAGAGTCAGGTCGTGTGAAAACCAATGCGTCGACTTCAGGGCCCAATTACTCGGTCATTT
ATAATCGTTTTCTCTCGAATTTTGAGCACAATGTAGATAATGTCTTCAGCTATCAGATGT
TATCAGGAAATTTCATAAAAATTGATCCGGAGTATCCAAATTGTCAGCGCCCGACACCTC
CTCCTTTCGAGACCTGCTATCTTATTCGGTGCAGTAAGGGAGAGGCGGGATGTGTCCCCG
CAGGGTGGTAGAAATTGGGTATATAAGAGAACGAGAGGACTCGCACAGTCATCACTTTTC
AAGTGTCACCCAACCAACCAAACCGCCGTCGAACGATGTCTGATTCCGCTGTTGTTGCCG
CCGCTGTCGAGCCAAAGGTCCCAAAGGCTAAGGCCGCCAA

>EP33004 (+) Ce hist.H2A-A his-12; range -299 to 100.
ATGATTCCTTACGGGCATGACGTCTCTTCTTTCCGTCCTTTGGCTTCGTAACGGTCTTGG
CGGCCTTCTTGGCTCCCTTGGCAGATGGCTTTGGTGGCATGTTGAGAGTTGGTGACTTGA
AACAAGTGTGAGGAGACCCTTGTCTCCCTTCTCTTTTATTTGTGTCTGTGGTGGGAAGGA
GGAGTCATTGAAGGGACAGGTGACATTCGGTCTGATGCTTATCGCTTGAAATGTGTCCCC
GCAGTGTCTCCGCCTACCCACCACAGAAATTGTATATAATAGTGTCTCTGCAGTTGCCTC
ATCAGATTCGATTCTATCAATCAAACAATGTCTGGACGTGGAAAGGGAGGCAAAGCCAAG
ACCGGAGGAAAGGCCAAGTCCCGCTCATCAAGAGCCGGAC
```

Assessing Assembly quality

- Assemblies need to be evaluated for quality using some assembly stats
- Assembly stats can be generated with tools such as **Quast**

Example assembly stats

B	C	D	E	F	G	H	I	J	K	L	M	N
total_length	number	mean_length	longest	shortest	N_count	Gaps	N50	N50n	N70	N70n	N90	N90n
4,991,645.00	60	83,194.08	610,445.00	524	1	1	232,462.00	8	131421	14	43876	26
4,589,899.00	86	53,370.92	357,165.00	308	250	5	119,264.00	12	80380	22	42122	38
4,655,125.00	103	45,195.39	328,684.00	311	347	4	141,947.00	12	97965	20	31187	37
4,708,681.00	54	87,197.80	804,746.00	502	406	2	341,170.00	5	223220	8	74907	16
4,736,275.00	72	65,781.60	473,243.00	381	365	3	155,594.00	9	95828	16	42687	29
4,847,918.00	50	96,958.36	724,413.00	313	433	4	267,536.00	6	185247	11	55448	18
4,895,618.00	42	116,562.33	903,697.00	429	237	2	313,077.00	5	214654	8	72117	17
4,822,679.00	83	58,104.57	322,552.00	388	422	3	128,833.00	13	82743	22	32215	39
4,981,853.00	37	134,644.68	537,027.00	450	119	2	482,573.00	5	365130	8	120866	12
4,598,065.00	118	38,966.65	250,513.00	373	818	4	86,000.00	17	57543	30	20651	55
5,028,650.00	94	53,496.28	393,261.00	363	220	5	136,834.00	11	73943	21	32883	40
4,873,756.00	50	97,475.12	919,335.00	330	1022	6	515,053.00	4	229640	7	87020	13
5,832,872.00	1666	3,501.12	116,024.00	126	64024	165	19,520.00	79	8734	160	915	632
4,565,615.00	120	38,046.79	161,592.00	415	383	2	76,534.00	21	51590	35	19559	63
4,849,614.00	80	60,620.18	254,757.00	308	417	2	125,968.00	13	85633	23	37842	40
4,682,411.00	83	56,414.59	505,722.00	354	635	3	138,005.00	9	79629	17	33506	35
4,713,616.00	120	39,280.13	328,731.00	461	1	1	95,601.00	15	48747	28	22081	57
4,674,998.00	64	73,046.84	623,293.00	308	548	3	185,601.00	7	106685	14	40445	27
4,821,497.00	34	141,808.74	659,473.00	360	407	1	294,404.00	6	163583	10	88712	17
4,766,021.00	87	54,781.85	609,140.00	307	483	7	111,295.00	13	63582	23	32122	43
4,814,643.00	40	120,366.07	772,555.00	359	1	1	280,338.00	5	163175	10	75168	19
5,144,517.00	65	79,146.42	437,938.00	419	979	5	210,154.00	9	173826	14	54012	26
4,902,416.00	58	84,524.41	523,264.00	395	539	3	224,746.00	7	145254	13	48816	23

Using Assembly stats-summary

Metrics	Description
N50	N50 means, half of the genome sequence is larger than or equal the N50 contig size (↑).
NG50	The length of the scaffold at which 50% of the genome length is covered (↑).
Coverage	If 90% of the bases have at least 5X read coverage, the genome is considered accurate (↑).
N90	An assembly is considered to have continuity provided its N 90 > 5 Kb (↑).
Average contig length	The average contig length should be longer than 5000 bases (5 Kb) (↑).
Number of genes	If an assembly that identifies most of the known genes is considered the better assembly (↑).
Number of gaps	The gaps in an assembly decreases the quality (↓).
Validity	An assembly can be validated by the reference sequence (↑).

Adapted from https://www.cd-genomics.com/an-overview-of-genome-assembly.html

Post-Assembly analyses

- MLST
- AMR screening
- Plasmid typing
- Annotation
- Phylogenetic analyses

Adapted from https://www.cd-genomics.com/an-overview-of-genome-assembly.html

wellcome connecting science

ACORN

questions?