Pathogen Genomics – working with pathogen genomes

Adam Reid

Wellcome Sanger Institute

LSHTM Pathogen Genomics 2020

Summary

- What is the point of a genome sequence?
- Genome sequencing technologies
- Sequence data files
- Viewing genomes

Why do genome sequencing?

- Reference for molecular biology
  - *Tropheryma whipplei* causes the potentially fatal Whipple's disease. Could not easily be grown. Genome revealed it had lost genes involved in producing amino acids.
- Identify all the genes that determine the function of the organism
  - *Neisseria meningitidis*, a major cause of meningitis. The first vaccine for a particular form of meningitis for identified by looking for candidates in its genome.
  - *Rickettsia prowazekii* is the cause of epidemic typhus, which killed millions in the early 20th century. It cannot reproduce outside of these cells. It was found to have just over 800 genes.
- Examine evolution by comparative genomics
- Track spread of pathogens
- Identify antimicrobial/drug resistance genes and drug targets
  - Mtb researchers made bacteria resistant to a new drug. Genome sequencing identified the gene involved in resistance.
- Basis for other omics technologies – RNA-seq, ChIP-seq, Methylome etc.

Why do genome sequencing? - Video of Wellcome Sanger Institute researchers

Technology overview

- Sanger sequencing produces ~500bp reads
  - Pros: Highly accurate
  - Cons: Expensive, laborious
  - Uses: High quality reference genomes
- Illumina's sequencing-by-synthesis 75-250bp
  - Pros: cheap, lots of reads (e.g. 500 million per run)
  - Cons: short reads
  - Uses: Resequencing, draft genomes, RNA-seq
- Pacific Biosciences Single-Molecule Real Time (SMRT) reads of 5000bp-40000bp
  - Pros: long reads
  - Cons: Fewer reads than Illumina - ~1 million, low accuracy
  - Uses: Reference genomes

Genome sequencing technologies – Interview with Mike Quail

Sequence data

Fasta

```
>yneN
TTAATGCCTCTTCTCATTTCTTCTGCTGTCATCCGCACAGCAGAAGAATTCCTCATTGAC
TATTATTTCGCAATTTGCTCACATGGATTAAATTAAACTACATACTATAAGATATAAACT
TCTGCCTACAGCTGTAAGAAACTCCGCTCAGTACTGAAGCACCAGTCCTATTTCCTCTTT
TCTCCAGCCTGTTATATTAAGCATACTGATTAACGATTTTTAACGTTATCCGCTAAATAA
ACATATTTGAAATGCATGCGACCACAGTGAAAAACAAAATCACGCAAAGAGACAACTATA
A
>yegR
ACTAACGGCTGCCACCGATAAATTTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCCATGGAGTTACCCTGAAGGGCCTCAATG
TCCGTAATTCCTACTTATGTAGGAAATGTTGTACAGAACATTTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATATTTAAACACTAGAGAGTGTCGTTGGTATTTAATGG
GGGAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAATTAGTATTTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTTATAGTTTCAAGTTGGCACTATAAGTCTTCTTACTA
ATCCTACAGGCGTAAGAATTGTATTGCAAAAGCCACGGTTTAGTCCTCTGTTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCCTGGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTTCATTCAATTCATAAATTTCATCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAAACATTCTAACAGAAGAAAATACT
T
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTTTATTGTTGACATTTCATTTATGCCGACTATTTATA
TGGTATACTTGTCGAATTATCTTAAAGGAAGCTCAGATTTTCTTATTTTTATTGAGAAAA
TGAGATGACGCCTTATGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCGCAGCAATTCGT
>yfdX
TGGCTGTATTTACATTTAATTAATCAGTATTTACATCGATATAATAAATGACATCTCTTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCGATAATAACTACAACTGTAAGATAACCCTTTCAAAATGACCGTTGCTCT
CTGATTTCTCATTTCATGCTCACCCAATATGATGGCGGCGTTTTCTAAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTTAATTATGGCCACGATGGTCACAGCAATTCTGGCATCTT
C
```

Fastq

```
@HS34_24228:8:1101:1116:7158/2
ATGCAGATTTTTTTACTATAAAAAATCATCATAAGGATAATANNGATAACAATGANNNNNATGTGAATGTAATAA
+
BBBBBFFFBFFFFF<FFFFFFFFFFBFFFFFFFFFFFFFFF!!<<<FFFFFFFF!!!!!<<<<FFFFFFFFFF
@HS34_24228:8:1101:1116:7461/2
AATATTAAAAAAATGGTTGAAAATCCAGAAGCATTTGCCGCANNTGCTGCACCTGNNNNNGCAGCCAAAGCTGAT
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFB<F/BBBBFB<F!!<<</<FB/B<!!!!!///<<B<FFFF/</F/
@HS34_24228:8:1101:1116:8637/2
TGGTGTTTTTATTTATTTTAAATATATTCTAATAATAATATANNATAAATAAAAANNNNNAATATATATATATAT
+
BB/BBFFFF</F<FFFFFF<F/FFBBFFBF//FFFFFBFBBF!!<<<</<<F/FF!!!!!<</</</<FBFFF/<
@HS34_24228:8:1101:1116:52646/2
CCGAATTATGCCTAGAAATTCACGAAGATGAGGAGGAAGGCGNNGGAAACGACGANNNNNACGACGACCACGAAG
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF!!/<<BFFFFFFF!!!!!<<<FFFFFFFFFF
@HS34_24228:8:1101:1116:52943/2
AATATAAAACATTGATTGAAGTGGTATTAAAACCAAGCGAACNNACATATGATATNNNNNTTACACATGCGTATT
+
BBBBBFBBFBFFFF<FFF/FFFFFFFFFFFF/<FFFFFFFF!!<<FFFFFBFFF!!!!!/<<FFF/FFFFFFFF
@HS34_24228:8:1101:1116:65353/2
ACGAAATAAATAAAAAGGTATTTAAAACCAAAAATGATAATANNCAATATGTTTANNNNNCATTTAATATTTATT
+
//<BB/FFFFFFFF/FFFFFFFFBFBFB<BBFFFFFFFB<<<!!//<<<BFFBB/!!!!!<<<<FFBBFFFBFFB
@HS34_24228:8:1101:1117:7618/2
TGCTCACCTGTTGATAAAAAATATAATAAAGTTAGAGTTACANNACAACAACAACNNNNNCAACAAGAAGCATAC
+
B/!</F//////B/BFF/FF/F/B//BFF///</B///B//!!<<<F/FF/<F/!!!!!/<</FFFFFFFF/F/
```

EMBL

```
ID   AF115338   standard; DNA; PRO; 591 BP.
AC   AF115338;
SV   AF115338.1
DT   03-JUN-1999 (Rel. 59, Created)
DT   23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE   Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW   .
OS   Pseudomonas fluorescens
OC   Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN   [1]
RP   1-591
RX   MEDLINE; 99369842.
RA   Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT   "Influence of a putative ECF sigma factor on expression of the major outer
RT   membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT   fluorescens";
RL   J. Bacteriol. 181(16):4746-4754(1999).
RN   [2]
RP   1-591
RA   De Mot R.;
RT   ;
RL   Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL   F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL   Mercierlaan 92, Heverlee B-3001, Belgium
DR   SPTREMBL; Q9X4L7; Q9X4L7.
FH   Key             Location/Qualifiers
FH
FT   source 1..591
FT                   /db_xref="taxon:294"
FT                   /organism="Pseudomonas fluorescens"
FT                   /strain="M114"
FT   CDS 1..591
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:Q9X4L7"
FT                   /transl_table=11
FT                   /gene="sigX"
FT                   /product="ECF sigma factor SigX"
FT                   /protein_id="AAD34329.1"
FT                   /translation="MNKAQTLSTRYDPRELSDEELVARSNTELFWVTRAYEELMRRYQR
FT                   TLPRVCARYLGNIRDADDVVQEVMLKVLVGLKNLESKRKFKTWLYSITVNECITQYRKE
FT                   RRKRRLMDALSLDPLEEASEEKALQPEEKGGLDRWLVTVNPIDRGILVLRFVAELEPQE
FT                   IADIMHMSLSATKMRYKRALLKLREKFAGETET"
SQ   Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
     atgaataaag ccoaaacgct atcoacgcgc tacgaccccc gcgggctctc tgatgaggag        60
     ttggtcggcc gctcgcatac cgagcttttt cacgtaacgc gcgcctatga agaactgatg       120
     cggcgttacc agcgaacatt atttaacgtt tgtgcgagat atcttgggaa cgatcgcgac       180
     gcagaagatg tctgtcagga agtcatgttg aagtgctgt atggcgtgaa gaacctcgag       240
     gggaaatcga agttcaaaac gtggctctac agcatcacgt acaacgaatg tattacgcag       300
     tatcggaagg aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga cccctcgag        360
     ...                                                                    ...
```

EMBL Flat File

Header
- Title
- Taxonomy
- Citation

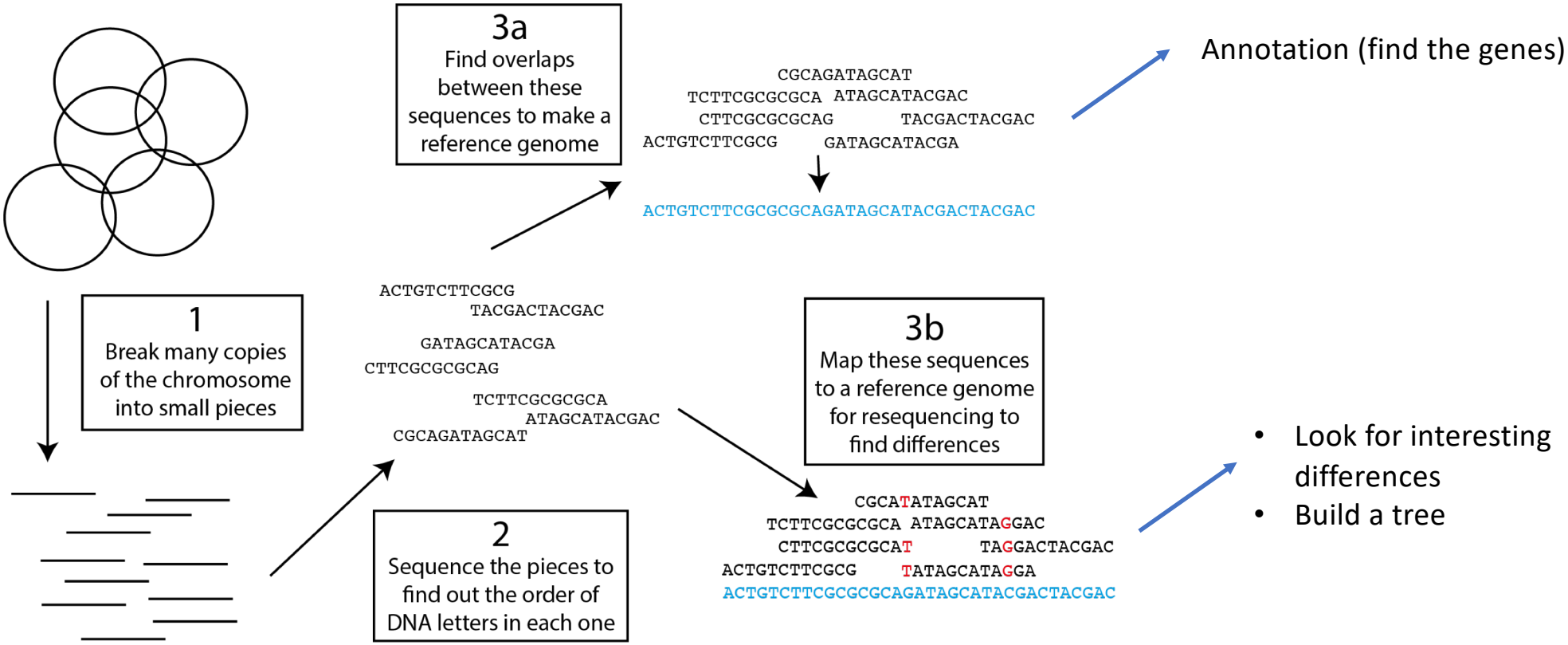
Features (AA seq)

DNA Sequence

SAM/BAM

```
HS34_24228:8:2109:7462:34726    65      Pf3D7_01_v3     21      0       23M140N52M      Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:1213:10197:86106   65      Pf3D7_01_v3     48      255     43M229N32M      Pf3D7_07_v3     1444731 0       CCTGAACCCTAAACCCTAAA
CCCTAAACCCTAAACCCTAAACCCTAAACCTAAAACCTAAAACCCAAAACCCTGA BBBBBFFFFFFFFFFFFFFF<FFFFFFFFFF<FFFFFFBFBFFFFF</FFFBBB/FFFFFF/FBF/<<</FFB///     AS:i
:-10    XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:24G39T10   YT:Z:UP XS:A:-  NH:i:1
HS34_24228:8:2109:7462:34726    321     Pf3D7_01_v3     56      0       23M105N52M      Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:2109:7462:34726    321     Pf3D7_01_v3     77      0       23M84N52M       Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:1209:16347:84469   99      Pf3D7_01_v3     144     255     66M9S   =       242     173     ACCCTAACCCTAAACCCTAAACCTAAAACCCTGAAC
CCTAAACCCTGAACCCTGAACCCTAAACCCAAACCGCCA BBBBBFFFFFF</FFFFFBFFFFBFF</FFBFBFFFFB//FFFF/<//</FFFBBBFFFFB/<BBF/B//<////     AS:i:-9 XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:66 YS:i:0  YT:Z:CP XS:A:-  NH:i:1
HS34_24228:8:1209:16347:84469   147     Pf3D7_01_v3     242     255     75M     =       144     -173    CCCTGAACCCTAAACCCTAAACCCTGAACCCTGAAC
CCTAAAACCTAAACCCTAAACCCTAAACCCTAAACCCTG <F<///BF<////B<////</BB<//FBFBC<BF<FBF/FFF/FB</FBF<FBBFFB//BFFFFF<<FB/<BB<B     AS:i:0  XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:-9 YT:Z:CP XS:A:-  NH:i:1
HS34_24228:8:2105:17855:88452   99      Pf3D7_01_v3     265     255     75M     =       675     485     CTGAACCCTGAACCCTAAAACCTAAACCCTAAACCC
TAAACCCTAAACCCTGAACCCTAAACCTAAAACCTAAAAC BBBBBCBFFB<F//F//BFFFF/</FFFFF<F</</</BFF<F<<BFFF///FFFF//F/F//BB/BBF/<</F     AS:i:0  XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:-9 YT:Z:CP XS:A:-  NH:i:1
```

What do we do with these data?



Doing bioinformatics

- Has anyone here done any bioinformatics?
- How are we going to do our bioinformatics?
  - Virtual machine with Linux
  - Artemis for viewing genomes
  - Various command line tools for mapping, assembling etc.
  - Web-based applications

What will we do today?

- Get familiar with the Virtual Machine
- Use Artemis to get intimate with genomes (morning)
- Map Illumina genome sequencing data to understand differences between closely related bacteria (afternoon)