From Infection to Genbank

How a pathogenic bacterium gets its genome to NCBI

Torsten Seemann
The steps

1. Sample collection
2. DNA purification
3. Prepare library
4. Sequencing
5. Read filtering
6. De novo assembly
7. Contig ordering
8. Genome closure
9. Annotation
10. Curation
11. Submission
Sample collection

● Take patient blood
  ○ Blood should not contain bacteria
  ○ bacteremia / septicemia / sepsis = bad = not good

● Centrifuge
  ○ slow spin to remove human cells
  ○ fast spin to pellet bacterial cells

● Streak onto agar media
  ○ emulsify the pellet first to make it spreadable
  ○ grow for 24 hours, likely to be monoculture
Purify DNA

● DNA extraction kit
  ○ lyse cells and digest (proteinaseK)
  ○ centrifuge to remove cell debris
  ○ pass lysate through column
    ■ DNA sticks to a DNA binding matrix
  ○ wash bound DNA
  ○ lower salt concentration - release bound DNA
  ○ precipitate: dubiously familiar stringy white pellet
    ■ salt and ethanol

● Extract DNA from strawberries at home!
  ○ detergent - breaks cells (*octoploid genome*)
  ○ strainer/pantyhose - remove particulate matter
  ○ salt - aids DNA precipitation
  ○ alcohol - precipitates DNA, keeps rest in solution
Library preparation

● Enough DNA?
  ○ each technology requires different amounts

● Library type
  ○ shotgun, short paired, or long paired reads?
  ○ different construction methods eg. circularization

● Size selection
  ○ nebulize, sonicate, enzymatic methods
  ○ run on gel + scalpel, or fancier methods

● Amplification
  ○ lots DNA - cluster generation, emulsion PCR
  ○ little DNA - multiple displacement amplification
Genome sequencing

- Lots of technologies at market
  - 454, Illumina, SOLiD, Ion, PacBio

- Each has its ups and downs
  - Speed, yield, read length, price, quality (pick 3)

- *De novo* assembly has particular needs
  - longer reads are always better
  - paired-end reads are even better
  - long mate pair reads are even more betterest!
  - preferably a mixture of insert sizes
  - desire > 8 kbp due to copies of rRNA locus
Read filtering

- **Why filter?**
  - reduce size of read set
  - improve average quality of reads
  - decrease RAM and CPU needs
  - improve assembly results

- **What to filter on?**
  - low Phred quality bases - note Q<20 still >1% error!
  - ambiguous bases ie. [^AGTCagtc]
  - reads that are too short
  - widowed reads
**De novo** genome assembly

- **De novo**
  - Latin - "from the beginning", "afresh", "anew"
  - Without reference to any other genomes

- Various types of assemblers
  - overlap graph, de Bruijn graph, string graph
  - but all doing essentially the same thing
Assembly algorithm

- Find all overlaps between all reads
  - naively this is O(N^2) for N reads
  - parameters are: min. overlap, min % identity
  - de Bruijn is fixed overlap (k) and 100% id required
- Build a graph from these overlaps
  - nodes/arcs <=> reads/overlaps <=> vertices/edges
- Simplify the graph
  - because real reads have errors
- Trace a single path through the graph
  - Just read off the consensus as you go
  - Elegant simplicity?
Assembly graph

Shared vertices are repeats....
Scaffolding

- Use paired reads to join contigs
  - reads with their mates in different contigs in a consistent manner suggests adjacency

- A difficult constraint problem
  - distance between mates ("insert size") variable
  - repeats cause ambiguous mate placement

- Assembler scaffolding support
  - Included: Velvet, Newbler, SGA, CAP3
  - Not included: Mira, Abyss
  - Separate scaffolders: Bambus, SSPACE
Contig ordering

- **Optical maps**
  - wet lab method, real experimental evidence
  - chromosome sized restriction site map

- **Align to reference genome**
  - fit contigs best as possible against known reference
  - some contigs will fit if split (DNA rearrangement)
  - expect orphan contigs (novel DNA)
Genome closure

- **Finished genome**
  - one contig per replicon in original sample
  - bacterial chromosomes/plasmids usually circular
- **Why bother?**
  - ensures you didn't miss anything
  - no close reference exists
  - simplifies all downstream analysis
- **Labour intensive**
  - design primers around gaps, PCR, Sanger
  - Fosmid/BAC libraries for larger inconsistencies
- **Satisfying when complete**
  - but you swear never to do another one...
Annotation

- Annotation is the process of identifying important features in a genome
  - **gene** - protein product, promoter, signal sequences
    - ~1000 per Mbp in bacteria, coding dense
  - **tRNA** - transfer RNA
    - ~30 per bacteria cover all codons (wobble base)
  - **rRNA** - ribosomal RNA locus
    - 1 to 9 per bacteria, fast vs slow growers
  - *And many more...*
    - small RNAs, ncRNA, binding sites, tx factors
Annotating proteins

● Homology vs. Similarity
  ○ homology means same biological function
  ○ we use sequence similarity as a proxy for homology
  ○ works well for most situations

● Sequence alignment methods
  ○ "Exact" - Needleman-Wunsch, Smith-Waterman
  ○ "Approx" - BLAST, FASTA, and many others!
  ○ Database is sequences: nr, RefSeq, UniProt

● Sequence profile methods
  ○ Build a HMM (model) of aligned sequence families
  ○ HMMer - scan profiles against query protein seq.
  ○ Database is profiles: Pfam, TIGRfams, FigFam
Curation

● Automatic annotation
  ○ better in recent years
  ○ more quality databases and models now
  ○ but still flawed

● Manual curation
  ○ Essential for a quality annotation
  ○ Find pseudo, missing, bogus, and broken genes
  ○ Discover mis-assemblies
  ○ Correct mis-annotated protein families
  ○ Fix incorrect start codons
    ■ Bacteria have 3-5 start codons, not just AUG (M)
Submission

● Start a "BioProject"
  ○ which has "Studies"
    ■ which have "Samples"
      ● which have "Runs" and "Other stuff"

● Assembled genomes
  ○ submit finished genome or "draft" contigs
  ○ can be unannotated, NCBI usually auto-annotates

● Raw reads
  ○ submit to NCBI Sequence Read Archive (SRA)
  ○ make sure to keep good records of meta-data!
    ■ they want machine ID, chemistry, Lane #, ....
Conclusion

- Easy to pipeline most of this
  - *Galaxy* anyone?

- I've developed key components
  - Neson (clean reads, correct indel assembly errors)
  - VelvetOptimiser (automate Illumina assemblies)
  - Prokka (rapid prokaryotic annotation)

- My LSCC EOI
  - Do every bacterial genome in SRA (~30,000)
  - Create a public database, with derived results