NORWEGIAN SEQUENCING CENTRE

Planning a sequencing project

Adapted from Nick Loman

University of Birmingham

https://github.com/lexnederbragt/denovo-assembly-tutorial

CEES
Centre for Ecological and Evolutionary Synthesis

Generating sequence is easy

Assembly is difficult … because of

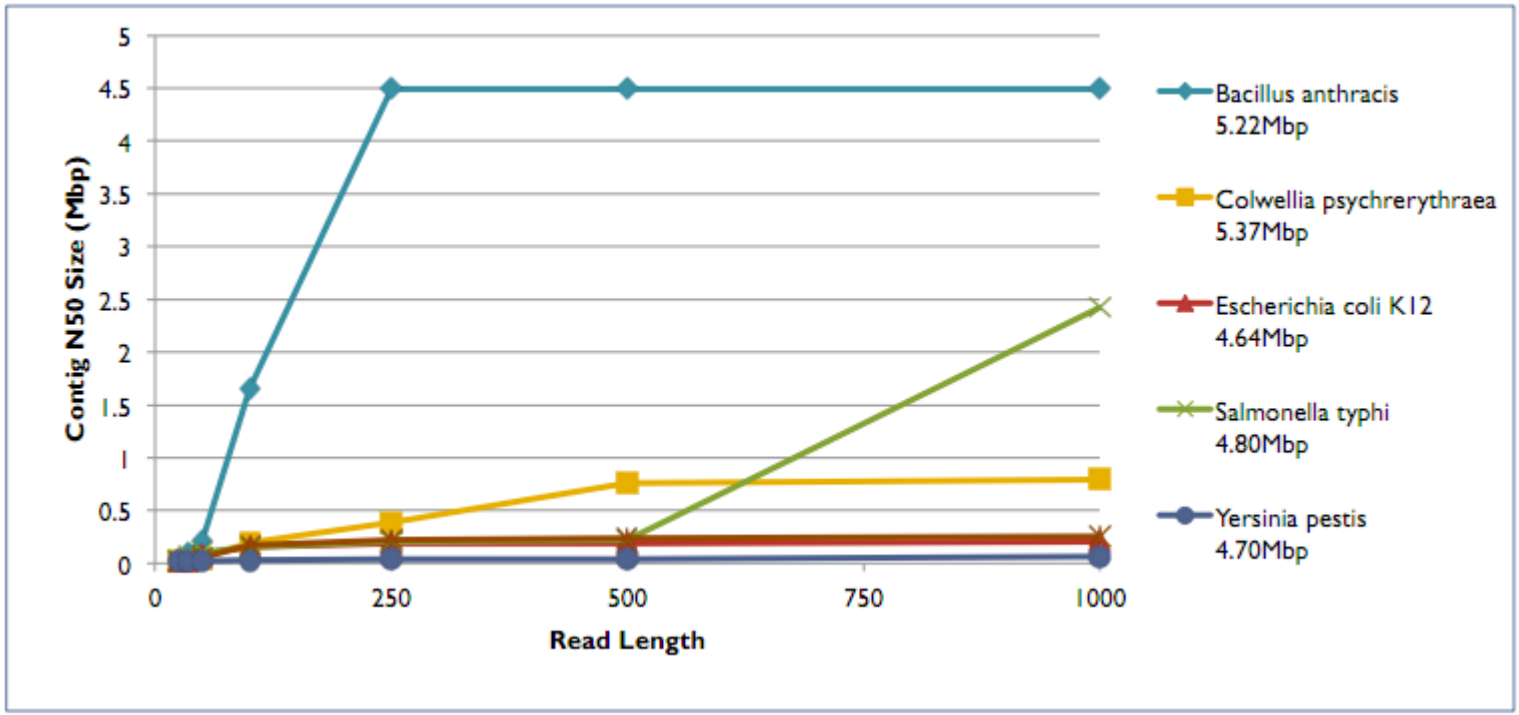
Repeats

Ploidy

Size

Contamination

Repeats and read length



Mike Schatz

Difficulty levels

Easy

- Chlamydia genome (small, no repeats)
- *E. coli* (small, some repeats)
- Small eukaryotic
- Large, diploid (*Homo sapiens*)
- Large, polyploid (wheat)

Hard

Biology comes first

- How long do you want to spend on this?
- Where's the value to you?

- Gene finding only (RNA-Seq, very draft WGS)
- Draft incomplete
- Draft complete
- Contiguous / circular complete
- Finished – *the holy grail*

Resources to help you decide

- What have others done already?
- Close reference available?
- What's your budget?
- What's your timescale?
- What are your bioinformatics resources?
- What choice of sequencing technology?

Choosing technologies

- Strengths/weaknesses
- “Cut your cloth to suit your purse”

- 454
  - Long reads
  - Low throughput
  - Homopolymeric tract errors
  - Expensive!

Choosing technologies

- Illumina, short reads (but getting longer)
  - High throughput
  - Cheap
  - Short reads (2x150 best)
- SOLiD – NO!
- Ion Torrent – short reads, low throughput, medium-cost

Choosing technologies

- PacBio
  - Looooooong reads
  - Low quality per base
  - Low throughput
  - Finished bacterial genomes!
  - Large genomes:
    - Need short reads for correction
    - (or need better software)

Mate-pairs

- 454
  - 3 to 20kb protocols
- Illumina
  - 3 to 10 kb (20?) protocols
- IonTorrent
  - demonstrated 3 and 9 kb protocols

What size is it?

- How many sequencing reads to you need?
  - Rule of thumb: 50x for short reads, 10x for long reads

What is the repeat structure?

- Transposons
  - LINEs
  - SINEs
  - IS elements
- Number, length, size

- Best read length
- Determine mate-pair strategy, or decide to give up on completeness!

What is the minimum information I need?

- Genes
- Regulatory regions
- Contiguous chromosomes
- Finished

Bioinformatics

- Large assemblies require large amount of memory
- Estimating requirements may be difficult
- Sometimes still not possible – *wheat*
- Hybrid assemblies not always straight-forward

Hybrid assembly

- Challenge: short read + long reads

→use de Bruijn graph?

→use Overlap Layout Consensus?

Hybrid assembly

- Use a single program with all data

OR

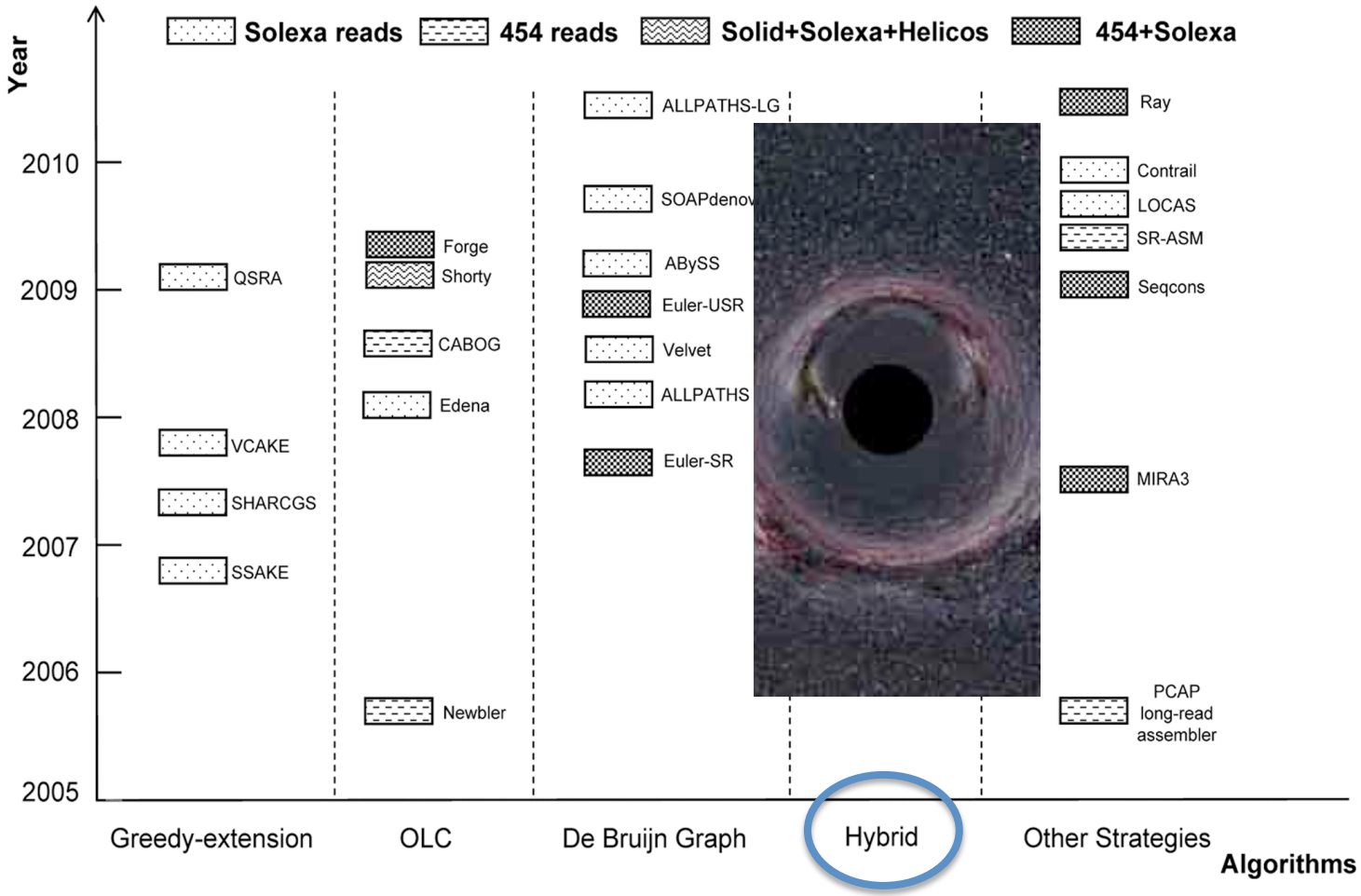
- Correct assembly with data from other technology/technologies

OR

- Use different programs for each dataset

→ merge into one

Hybrid assembly



http://www.crystalinks.com/blackholeeye350.jpg

Zhang et al. PLoSOne 2011

Hybrid assembly options

- Celera
  - sanger
  - 454
  - Illumina
  - PacBio

Hybrid assembly options

- Ray, Abyss, ...
  - → 454 and Illumina
- Newbler
  - → 454
  - → Illumina?

Hybrid assembly

- Does it work?

- Are we there yet?

Non-sequence data

Closely related genome

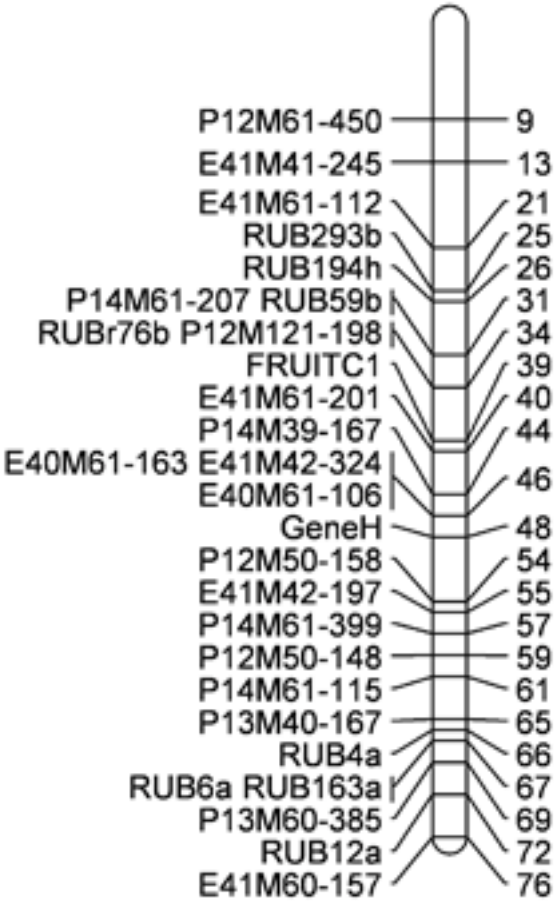
→ requires synteny

Non-sequence data

Linkage map

→superscaffolds

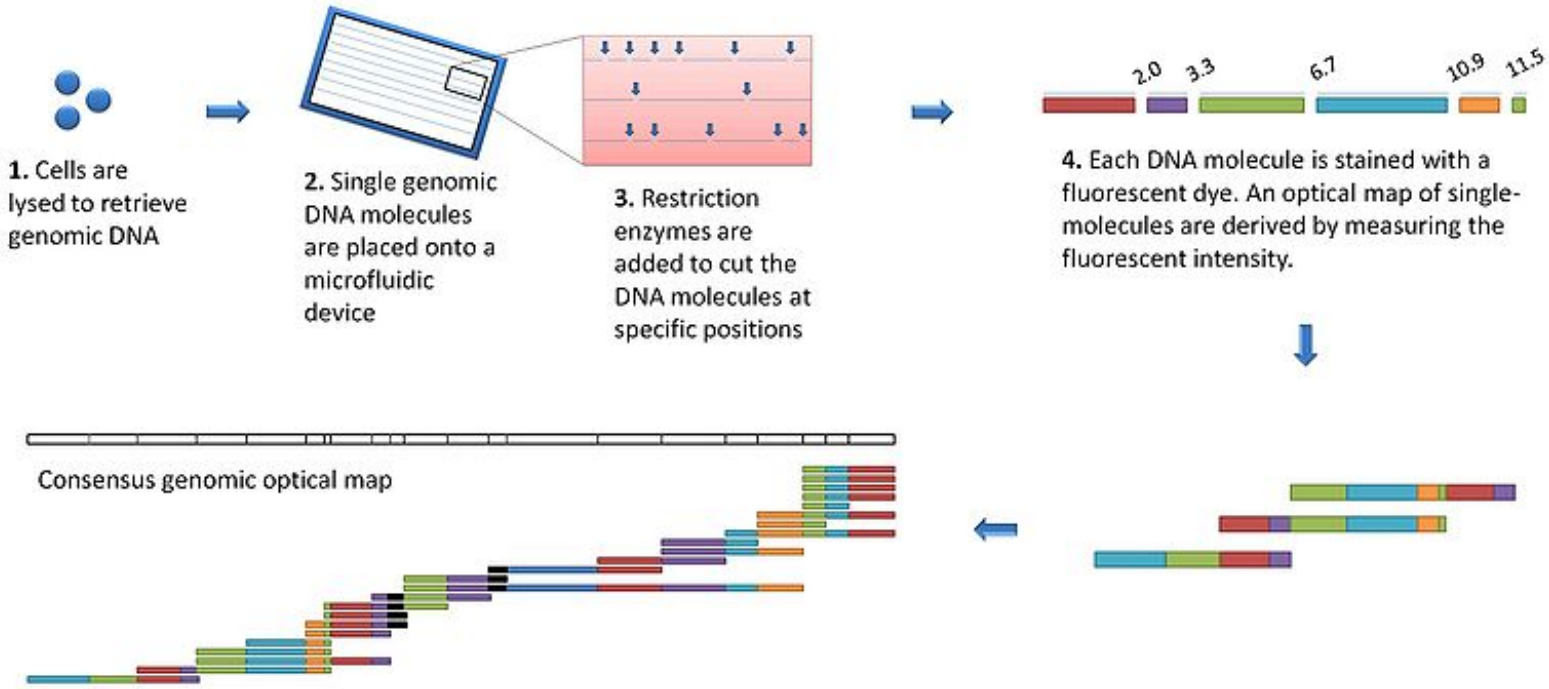
→ pseudochromosomes



http://www.hutton.ac.uk

Non-sequence data

Optical maps



http://en.wikipedia.org/wiki/File:Optical_mapping.jpg

Basic recipes for genome

Bacterial – Illumina

- Paired-end reads required
- the more mate pair libraries, the better

Basic recipes for genome

Bacterial – PacBio

- 60-100x in long reads
  - 2 SMRTcells (chips)
- May give 1 scaffold per chromosome
- Very high quality
- Can determine base modifications

How to sequence a genome

Eukaryote

Foundation of Illumina data

- 100x coverage Paired End reads (2x100bp)
- several Mate Pair libraries
  - 2kb, 3kb, 8k, 10kb, bigger?
- this is now very cheap!

Fill gaps with long reads

- PacBio

How to sequence a genome

- Add lots of bioinformatics...

http://cores.montana.edu/index.php?page=bioinformatics-core-facility