MCB 5472
Lecture 2 Feb 3/14

(1) GenBank continued

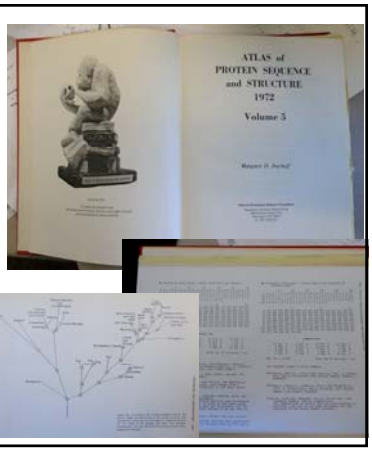
(2) Primer: Genome sequencing and assembly

Genbank

- Founded in 1982 at the Los Alamos National Laboratory
- Initially managed at Stanford in conjunction with the BIOSCI/Bionet news groups
- 1989-92 transition to the NCBI on the east coast
- One precursor was Margaret Dayhoff's Atlas of Protein Sequence and Structure
- In 1987 genbank fit onto a few 360 KB floppy disks.
- Genbank uses a flat file database format (see http://en.wikipedia.org/wiki/Flat_file_database)
- NCBI does not use a relational databank (as in Oracle, peoplesoft)
- NCBI stores data in ASN.1 format (61860050 http://en.wikipedia.org/wiki/Abstract_Syntax_Notation_One), which allows to hardwire crosslinks to other data bases, and makes retrieval of related information fast.
- NCBI's sample record (http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html) contains links to most the fields used in the gbk flatfile.
- In the genbank records at NCBI the links connect to the features (i.e. the pubmed record, or the encoded protein sequence) --- not easy to work with.

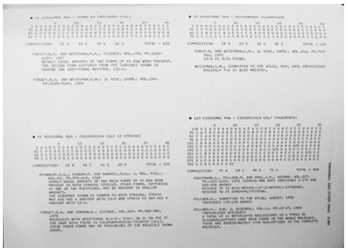
Dr. **Margaret** Belle (Oakley) **Dayhoff**
March 11, 1925 – February 5, 1983

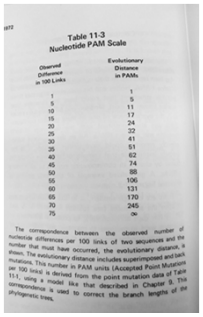
Among other things, we owe her the first nucleotide and protein data bank, the PAM substitution matrix, and the single letter amino acid code. (Image from wikipedia)



Atlas of Protein Sequences 1972 (cont)

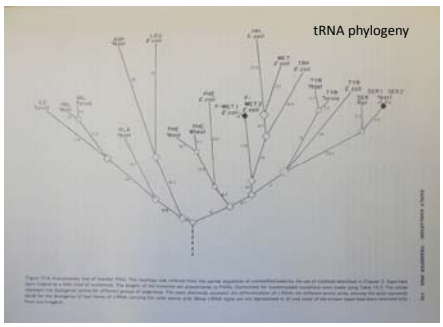
The Atlas also contained RNA sequences, and PAM matrix for nucleotides





Atlas of Protein Sequences 1972 (cont)

Contained phylogenetic reconstructions that went back in time to far before the Last Unversal Common Ancestor (LUCA) aka the cenancestor of all living cellular organisms alive today.



Relational *vs* flat-file

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

(B) GenBank Flat-File Format

```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and
            Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
            Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
```

data tables

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYF…
.....	

SQL can be used to connect/join and search tables

Example: GI numbers -> sequence and GI numbers to taxonomic information

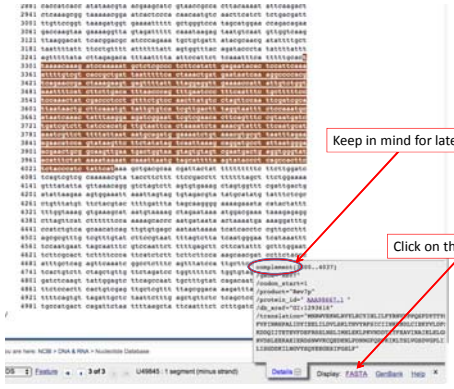
Taxonomy at the NCBI

- The taxonomy browser at NCBI is well maintained and useful, despite sometimes using strange labels (domains are labeled as superkingdoms)
- The taxonomic categories are linked to available sequences (genomes, proteins, nucleotide)
- The FTP site at the NCBI is a taxonomic wasteland: the archaeal genomes are stored in the folder labeled Bacteria.

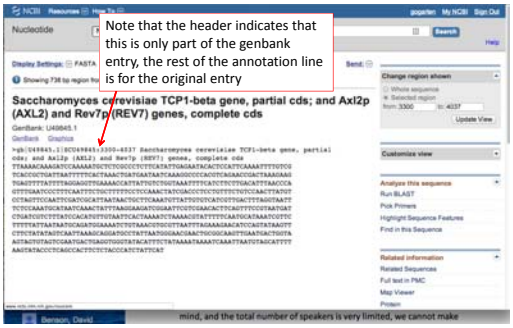
To obtain a CDS from a gene at NCBI

Click on the CDS you are interested in

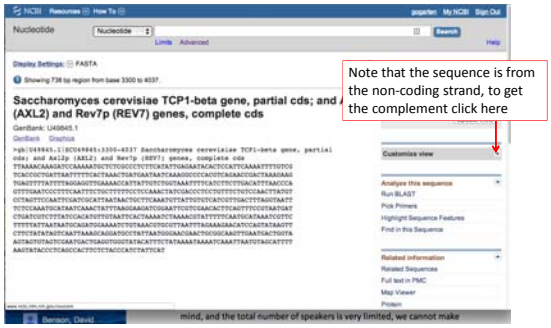
To obtain a CDS from a gene at NCBI



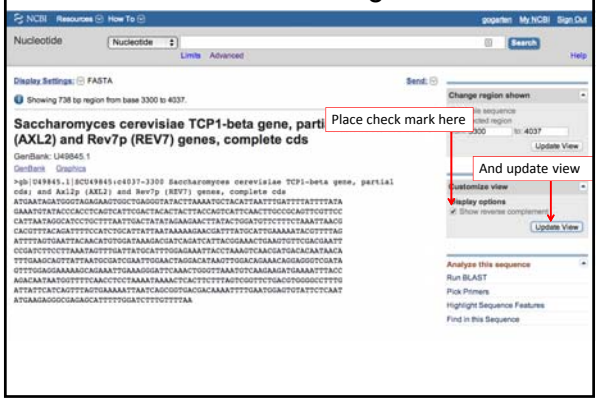
To obtain a CDS from a gene at NCBI



To obtain a CDS from a gene at NCBI



To obtain a CDS from a gene at NCBI

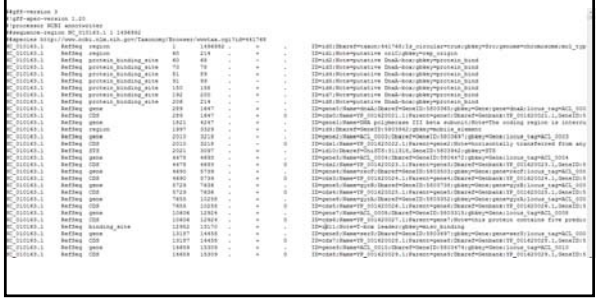


gff (Genome Feature Format)

- A compact tabular annotation format, especially common for larger eukaryotic genomes
- Does not include sequences
- An input file type for many genomics programs (particularly sequence viewers)

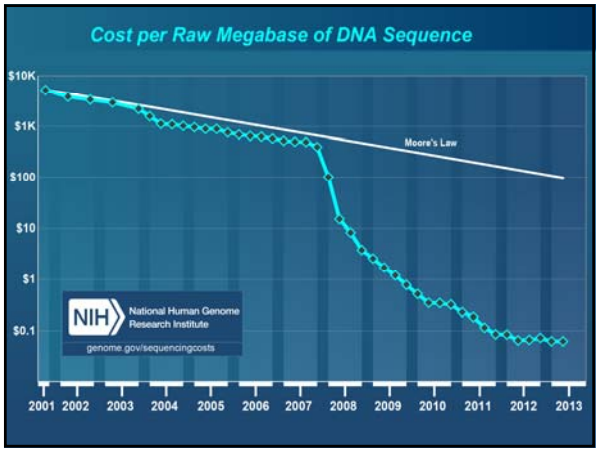
Columns:

1: Seq ID
2: Source
3: Type
4: Start
5: End
6: Score
7: Strand
8: Phase
9: Attributes (itself subdivided)



Genome sequencing and assembly: a primer

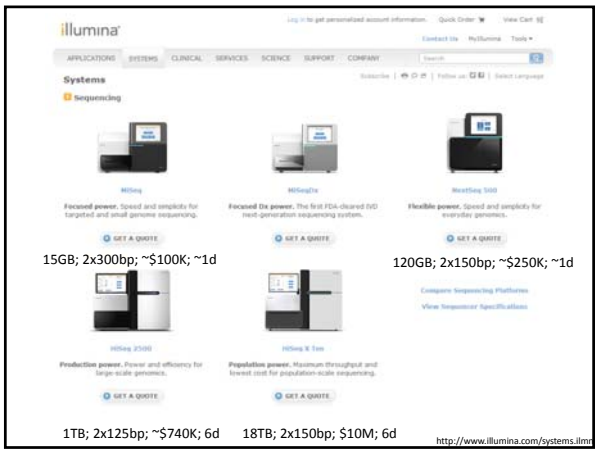
Analyses are only as good as their input data…



Changes in instrument capacity over the past decade, and the timing of major sequencing projects

ER Mardis. *Nature* **470**, 198-203 (2011) doi:10.1038/nature09796

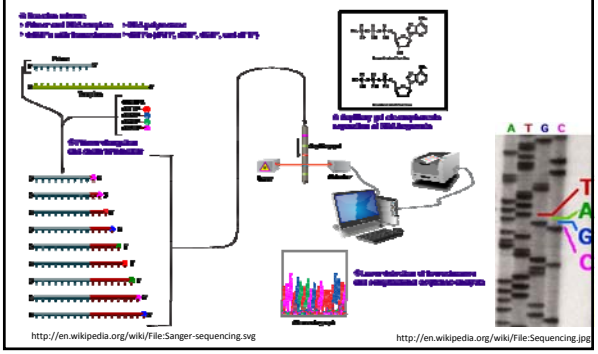
nature



Why it matters

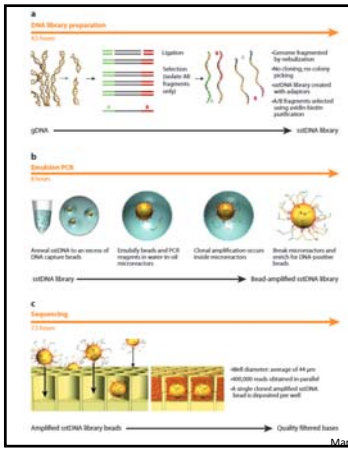
- How a genome was sequenced matters for molecular evolution studies
  - Different sequencing methods have different error profiles
  - Different sequencing methods require different assembly methods, each with different biases and error profiles

Sanger sequencing



Sanger sequencing

- High quality, especially because often manually examined
- Low throughput, high cost
- Read lengths 900-1000bp
- Still gold standard method for DNA sequencing (and most common!)



454 sequencing

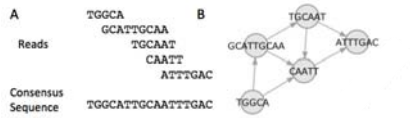
- Introduced in 2005
- Moderate throughput, cost
- Up to ~700bp read lengths
- Homopolymer errors problematic
  - e.g., “AAAAA” vs “AAAA”

Mardis (2008) Ann. Rev. Genom. Hum. Genet. 9:387-402

Overlap/layout/consensus genome assembly

1. Compare all reads to each other to find those that overlap
2. Create overlap graph arranging reads according to their overlaps
3. Find unique path through the graph
4. Assemble overlapping reads by aligning the reads and deriving consensus

Overlap/layout/consensus genome assembly



Nodes: reads

Edges: alignments

Only one unique path

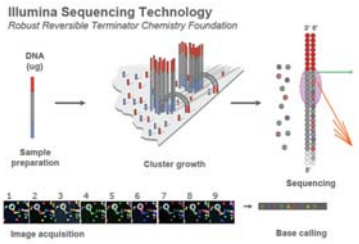
Leverage alignment probabilities

http://gcat.davidson.edu/phast/index.html

Overlap/layout/consensus genome assembly

- Requires all-vs-all comparison of reads
  - becomes computationally intensive as the number of reads increases
- Developed and applied for Sanger and 454 sequencing

Illumina sequencing



- Introduced 2006
- Short reads
- High throughput
- Substitutions are main error

http://openwetware.org/images/7/76/BMC_IlluminaFlowcell.png

De Bruijn graph assembly

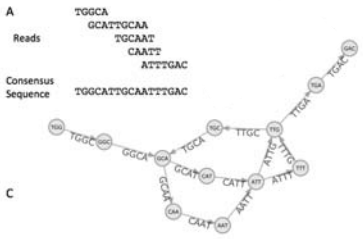
- Instead of comparing all reads with each other, split reads up into kmers
  - i.e., subsets of each read of a given length

```
                k=4
TGGCA           TGGC, GGCA
  GCATTGCAA     GCAT, CATT, ATTG, TTGC, TGCA, GCAA
    TGCAAT      TGCA, GCAA, CAAT
      CAATT     CAAT, AATT
        ATTTGAC ATTT, TTTG, TTGA, TGAC
```

http://gcat.davidson.edu/phast/index.html

De Bruijn graph assembly

- Draw a graph of kmer overlap
- Find unique path through graph
  - Leverage kmers next to each other in reads



http://gcat.davidson.edu/phast/index.html

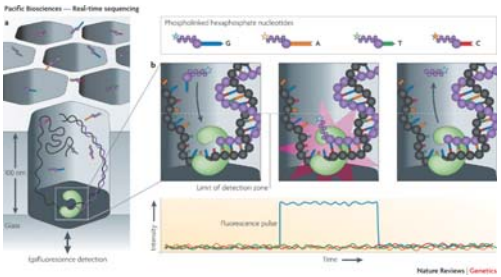
De Bruijn graph assembly

- Doesn't need all-vs-all comparison so is much faster
- Can handle large numbers of reads, e.g., as generated by Illumina technology
- Graph is much more complicated, RAM intensive
- More sensitive to errors

Other technologies

- SOLiD: different technology but similar data to Illumina, i.e., short reads, high throughput
- Ion Torrent: different technology but similar data to 454, i.e., moderately long reads, moderate throughput, homopolymer errors

PacBio sequencing

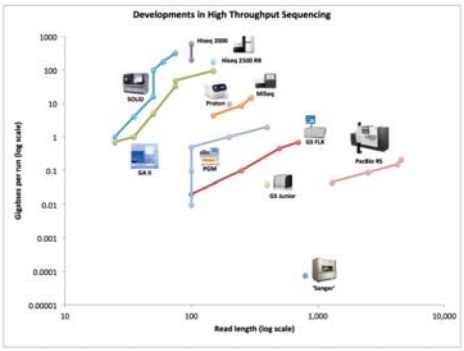


Metzker (2010) Nat. Rev. Genet. 11: 31-46

PacBio sequencing

- Single molecule (no PCR needed)
- Long reads (up to 30KB)
- Can read modified bases
- High error rate (~12%), mostly substitutions
  - Can overcome using high coverage
- Assembly via overlap/layout/consensus methods

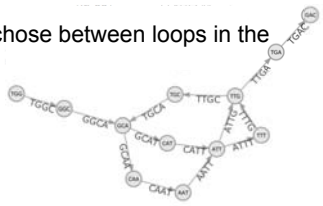
Trade off between read length & throughput



Lex Nederbragt (2013): developments in NGS. figshare. http://dx.doi.org/10.6084/m9.figshare.100940.

Why is read length important?

- Consider a genomic repeat
  - Each repeat is larger than the read length
  - Where should reads inside the repeat be placed?
  - OR: who do you chose between loops in the De Bruijn graph?



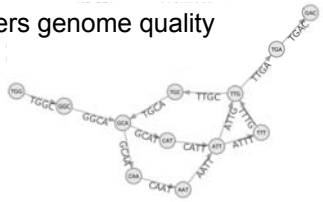
http://gcat.davidson.edu/phast/index.html

Define: contig

- A contig is a stretch of DNA without any gaps
- The result of sequencing read assembly

How do assemblers respond to repeats that they can't resolve?

- Break the graph into multiple contigs
- Some reads are never included in draft genomes
- Lowers genome quality

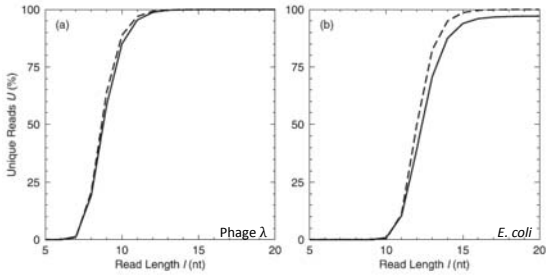


http://gcat.davidson.edu/phast/index.html

Does it matter? Are real genomes too complex to assemble using short reads?

Experiments using simulated reads from model genomes

Bacteriophage λ & *E. coli*

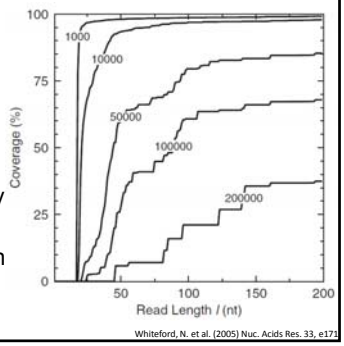


- As read lengths increase, more of the genome becomes unique
- Reads from real genomes (solid lines) do not become completely unique
  - Dashed is random reads for same sized genome

Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

E. coli

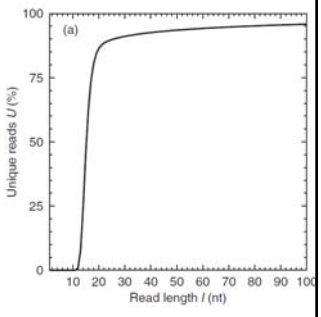
- Each line is % contigs > that size in bp
- *E. coli* genome is ~4.6 Mbp
- Still far from complete assembly at 200bp
- Assembly largely in gene-sized fragments



Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

C. elegans

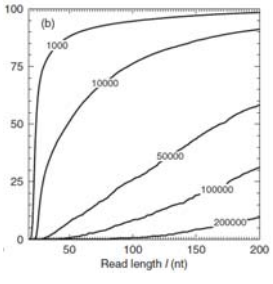
- Still unassembled at 100bp



Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

C. elegans

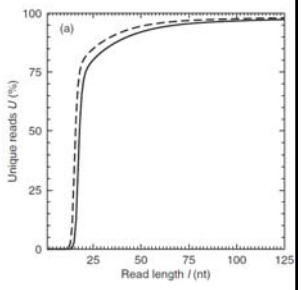
- *C. elegans* genome size: ~100 Mb
- Still far from complete at 200 bp
- Gene sized fragments reasonable (note genes have introns and so are larger)



Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

Human chromosome 1

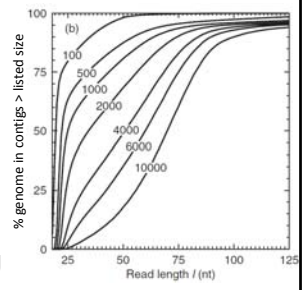
- Dashed: chr 1
- Solid: whole genome
- Still many 125bp reads that can't be uniquely mapped



Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

Human chromosome 1

- Chr 1 is 249 Mb
- Still far from complete assembly at 125 bp
- Gene-sized fragments still possible, but getting harder

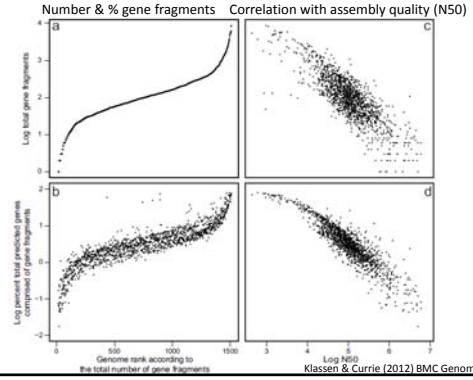


Whiteford, N. et al. (2005) Nuc. Acids Res. 33, e171

Problem

- Recall: Sanger sequencing has long read lengths, but is low-throughput and expensive
- Illumina etc. has short read lengths but is high-throughput and cheap
- Lots of low quality genomes therefore have appeared
  - Short-read Illumina etc.
  - Low coverage Sanger

One result: gene fragmentation



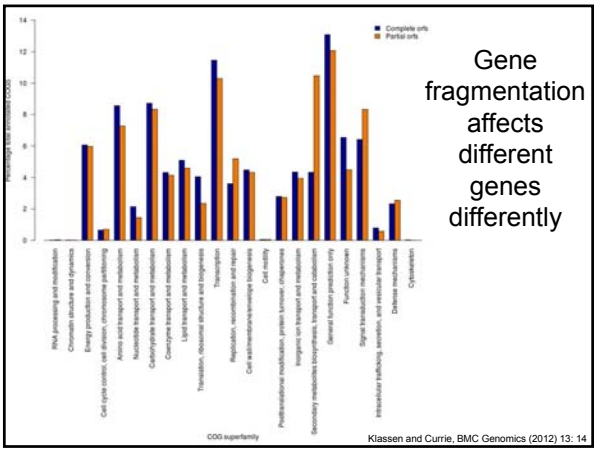
Klassen & Currie (2012) BMC Genomics 13:14

Definition: N50

- Order contigs from longest to shortest
- Sum lengths of all contigs
- N50 is contig size where you reach 50% of the total assembly size
- Other analogous measures, N80 etc.

Gene fragmentation

- Can cause your gene to be missed
- Confounds gene content analyses
  - Some genes counted as duplicates
  - Some genes falsely annotated



Gene fragmentation affects different genes differently

Klassen and Currie, BMC Genomics (2012) 13: 14

One solution: increase read length

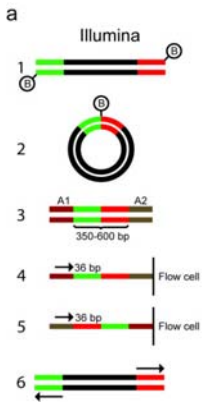
- Increasing read length is a focus of several sequencing platforms (PacBio, MiSeq)
  - These especially (but not exclusively) target bacterial genomes where they are most effective
- Not all technologies do this: less applicable for counting applications (e.g., RNAseq) and resequencing

Definition: scaffold

- Scaffolds are a series of contigs connected by gaps
  - i.e., an assembly of contigs
- Often the gaps are of known length

Scaffold increase genome quality

- Allow some contigs to be merged
- Often gaps are small limiting information loss for a genomic region
- Allow gross genome structure to be better revealed
- Gene fragmentation still exists because contigs are still broken



Mate pair libraries

Illumina method:

1. Biotinylate fragment
2. Circularize
3. Fragment, ligate sequencing adaptors

4 & 5. Sequence from both ends

6. Reconstruct original fragment computationally

Berglund et al. 2011 Invest. Genet. 2:23

Paired libraries

- Provide sequence from 2 chromosomal regions
- Paired-end: ~300bp apart
  - Same principle as mate-pair but fancy PCR instead of ligation (cheaper libraries)
- Mate-pairs: at least 3kb, often 8kb, 20kb, 40kb
  - Larger libraries span larger repeats, but can be tricky to make
  - Costly, lower throughput

Paired libraries

- In *de novo* genome assembly, nearly all read assemblers only use read pairing information AFTER contig assembly during scaffolding
  - This is starting to change as algorithms mature
- Read pairings are often used during read mapping to a reference genome

Resequencing

- If you have a high-quality reference genome already, it is often efficient to map sequencing reads to that genome instead of assembling it de novo
  - Computationally more tractable (restricted search space)
  - Common for epidemiology, population-level studies
- Caveat: you only get what you look for!

Other scaffolding methods

- Optical maps: create restriction maps of chromosome, link to genome sequences
  - Requires reasonable genome assembly to start with
- Genetic linkage maps: more classical experimental method of estimating gene location, can be linked to genome sequences

Outlook for sequencing

- Two themes:
  - Illumina increasing throughput, often short reads
    - Most important for resequencing, counting applications, clinical application
  - PacBio is recently taking over the de novo assembly niche
    - Watch for Oxford Nanopore in this space soon

Discuss:

(1)What are some different errors encountered during DNA sequencing?

(2)What effect do they have on molecular evolution studies?

(3)What can be done to mitigate them?