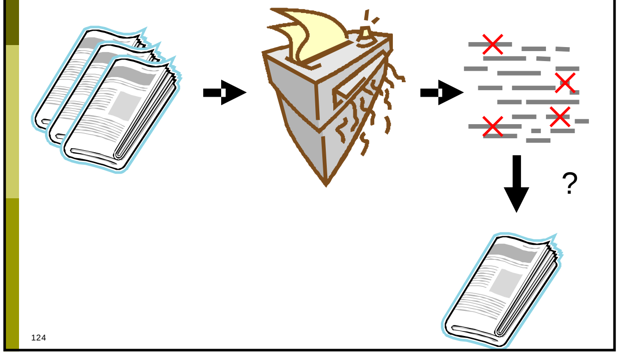
Introduction to Bioinformatics

Genome sequencing & assembly

Genome sequencing & assembly

- DNA sequencing
  - How do we obtain DNA sequence information from organisms?
- Genome assembly
  - What is needed to put together DNA sequence information from sequencing?
- First statement of sequence assembly problem (according to G. Myers):
  - Peltola, Söderlund, Tarhio, Ukkonen: Algorithms for some string matching problems arising in molecular genetics. Proc. 9th IFIP World Computer Congress, 1983

Recovery of shredded newspaper



DNA sequencing

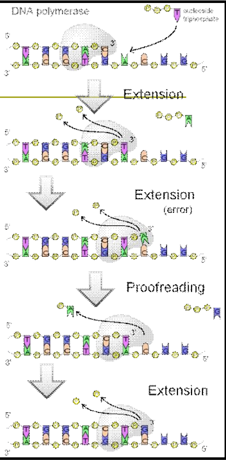
- DNA sequencing: resolving a nucleotide sequence (whole-genome or less)
- Many different methods developed
  - Maxam-Gilbert method (1977)
  - Sanger method (1977)
  - High-throughput methods

Sanger sequencing: sequencing by synthesis

- A sequencing technique developed by Fred Sanger
- Also called *dideoxy sequencing*

DNA polymerase

- A *DNA polymerase* is an enzyme that catalyzes DNA synthesis
- DNA polymerase needs a *primer*
  - Synthesis proceeds always in 5'->3' direction



http://en.wikipedia.org/wiki/DNA_polymerase

Dideoxy sequencing

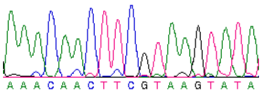
- In Sanger sequencing, chain-terminating dideoxynucleoside triphosphates (ddXTPs) are employed
  - ddATP, ddCTP, ddGTP, ddTTP lack the 3’-OH tail of dXTPs
- A mixture of dXTPs with small amount of ddXTPs is given to DNA polymerase with DNA template and primer
- ddXTPs are given fluorescent labels

Dideoxy sequencing

- When DNA polymerase encounters a ddXTP, the synthesis cannot proceed
- The process yields copied sequences of different lengths
- Each sequence is terminated by a labeled ddXTP

Determining the sequence

- Sequences are sorted according to length by capillary electrophoresis
- Fluorescent signals corresponding to labels are registered
- *Base calling*: identifying which base corresponds to each position in a read
  - Non-trivial problem!



Output sequences from base calling are called reads

Reads are short!

- Modern Sanger sequencers can produce quality reads up to ~750 bases[1]
  - Instruments provide you with a quality file for bases in reads, in addition to actual sequence data
- Compare the read length against the size of the human genome (2.9x10^9 bases)
- Reads have to be assembled!

[1] *Nature Methods* - 5, 16 - 18 (2008)

Problems with sequencing

- Sanger sequencing error rate per base varies from 1% to 3%[1]
- Repeats in DNA
  - For example, ~300 base *Alu* sequence repeated is over million times in human genome
  - Repeats occur in different scales
- What happens if repeat length is longer than read length?
  - We will get back to this problem later

[1] Jones, Pevzner (2004)

Shortest superstring problem

- Find the shortest string that ”explains” the reads
- *Given a set of strings (reads), find a shortest string that contains all of them*

Example: Shortest superstring

Set of strings: {000, 001, 010, 011, 100, 101, 110, 111}

Concetenation of strings: 000001010011100101110111

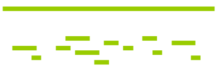
```
                              010
                            110
                          011
                        000
Shortest superstring: 0001110100
                         001
                           111
                             101
                               100
```

Shortest superstrings: issues

- NP-complete problem: unlike to have an efficient (exact) algorithm
- Reads may be from either strand of DNA
- Is the shortest string necessarily the correct assembly?
- What about errors in reads?
- Low *coverage* -> gaps in assembly
  - Coverage: average number of times each base occurs in the set of reads (e.g., 5x coverage)

Sequence assembly and combination locks

- What is common with sequence assembly and opening keypad locks?



Whole-genome shotgun sequence

- *Whole-genome shotgun sequence assembly* starts with a large sample of genomic DNA
  1. Sample is randomly partitioned into *inserts* of length > 500 bases
  2. Inserts are multiplied by cloning them into *a vector* which is used to infect bacteria
  3. DNA is collected from bacteria and sequenced
  4. Reads are assembled

Assembly of reads with Overlap-Layout-Consensus algorithm

- Overlap
  - Finding potentially overlapping reads
- Layout
  - Finding the order of reads along DNA
- Consensus (Multiple alignment)
  - Deriving the DNA sequence from the layout

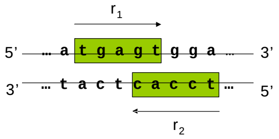
- Next, the method is described at a very abstract level, skipping a lot of details

Kececioglu, J.D. and E.W. Myers. 1995. Combinatorial algorithms for DNA sequence assembly. *Algorithmica* 13: 7-51.

Finding overlaps

- First, pairwise overlap alignment of reads is resolved
- Reads can be from either DNA strand: The *reverse complement* r* of each read r has to be considered

```
acggagtcc
     agtccgcgctt
```



r_1: tgagt, r_1^*: actca
r_2: tccac, r_2^*: gtgga

Example sequence to assemble

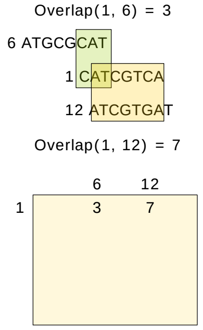
5’ – CAGCGCGCTGCGTGACGAGTCTGACAAAGACGGTATGCGCATCG
TGATTGAAGTGAAACGCGATGCGGTCGGTCGGTGAAGTTGTGCT - 3’

- 20 reads:

#	Read	Read*	#	Read	Read*
1	CATCGTCA	TCACGATG	11	GGTCGGTG	CACCGACC
2	CGGTGAAG	CTTCACCG	12	ATCGTGAT	ATCACGAT
3	TATGCGCA	TGCGCATA	13	GCGCTGCG	CGCAGCGC
4	GACGAGTC	GACTCGTC	14	GCATCGTG	CACGATGC
5	CTGACAAA	TTTGTCAG	15	**AGCGCGCT**	**AGCGCGCT**
6	ATGCGCAT	ATGCGCAT	16	GAAGTTGT	ACAACTTC
7	**ATGCGGTC**	GACCGCAT	17	AGTGAAAC	GTTTCACT
8	CTGCGTGA	TCACGCAG	18	ACGCGATG	CATCGCGT
9	GCGTGACG	CGTCACGC	19	GCGCATCG	CGATGCGC
10	GTCGGTGA	TCACCGAC	20	AAGTGAAA	TTTCACTT

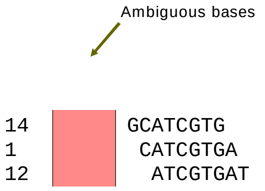
Finding overlaps

- Overlap between two reads can be found with a *dynamic programming* algorithm
  - Errors can be taken into account
- Dynamic programming will be discussed more on next lecture
- Overlap scores stored into the overlap matrix
  - Entries (i, j) below the diagonal denote overlap of read r_i and r_j^*



Finding layout & consensus

- Method extends the assembly *greedily* by choosing the best overlaps
- Both orientations are considered
- Sequence is extended as far as possible



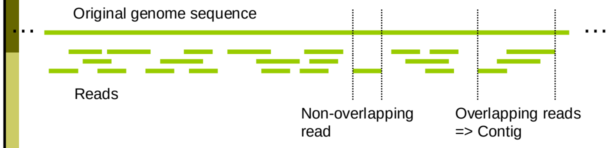
Consen

Finding layout & consensus

- We move on to next best overlaps and extend the sequence from there
- The method stops when there are no more overlaps to consider
- A number of contigs is produced
- Contig stands for contiguous sequence, resulting from merging reads

```
2              CGGTGAAG
10           GTCGGTGA
11          GGTCGGTG
7      ATGCGGTC
---------------------
       ATGCGGTCGGTGAAG
```

Whole-genome shotgun sequencing: summary



- Ordering of the reads is initially unknown
- Overlaps resolved by aligning the reads
- In a 3x10^9 bp genome with 500 bp reads and 5x coverage, there are ~10^7 reads and ~10^7(10^7-1)/2 = ~5x10^{13} pairwise sequence comparisons

Repeats in DNA and genome assembly

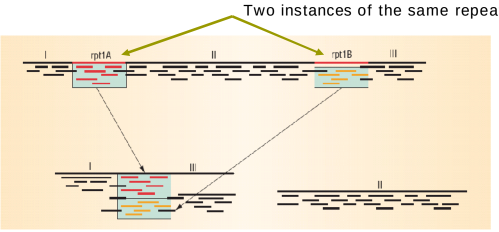


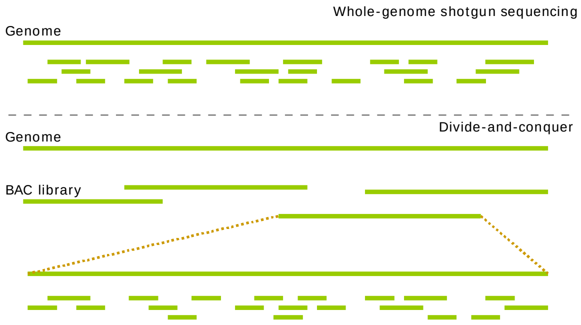
Figure 2. Repeat sequence. The top represents the correct layout of three DNA sequences. The bottom shows a repeat collapsed in a misassembly.

Pop, Salzberg, Shumway (2002)

Repeats in DNA cause problems in sequence assembly

- Recap: if repeat length exceeds read length, we might not get the correct assembly
- This is a problem especially in eukaryotes
  - ~3.1% of genome consists of repeats in Drosophila, ~45% in human
- Possible solutions
  1. Increase read length – feasible?
  2. Divide genome into smaller parts, with known order, and sequence parts individually

”Divide and conquer” sequencing approaches: BAC-by-BAC



BAC-by-BAC sequencing

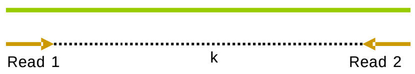
- Each BAC (Bacterial Artificial Chromosome) is about 150 kbp
- Covering the human genome requires ~30000 BACs
- BACs shotgun-sequenced separately
  - Number of repeats in each BAC is significantly smaller than in the whole genome...
  - ...needs much more manual work compared to whole-genome shotgun sequencing

Hybrid method

- Divide-and-conquer and whole-genome shotgun approaches can be combined
  - Obtain high coverage from whole-genome shotgun sequencing for short contigs
  - Generate of a set of BAC contigs with low coverage
  - Use BAC contigs to ”bin” short contigs to correct places
- This approach was used to sequence the brown Norway rat genome in 2004

Paired end sequencing

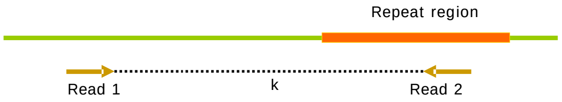
- *Paired end* (or *mate-pair*) sequencing is technique where
  - both ends of an insert are sequenced
  - For each insert, we get two reads
  - We know the distance between reads, and that they are in opposite orientation



- Typically read length < insert length

Paired end sequencing

- The key idea of paired end sequencing:
  - Both reads from an insert are unlikely to be in repeat regions
  - If we know where the first read is, we know also second’s location



- This technique helps to WGSS higher organisms

First whole-genome shotgun sequencing project: Drosophila melanogaster

- Fruit fly is a common *model organism* in biological studies
- Whole-genome assembly reported in Eugene Myers, *et al.*, A Whole-Genome Assembly of *Drosophila*, *Science* 24, 2000
- Genome size 120 Mbp

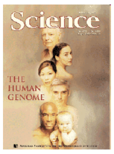
152

http://en.wikipedia.org/wiki/Drosophila_melanogaster

Sequencing of the Human Genome

- The (draft) human genome was published in 2001
- Two efforts:
  - Human Genome Project (public consortium)
  - Celera (private company)
- HGP: BAC-by-BAC approach
- Celera: whole-genome shotgun sequencing

HGP: Nature 15 February 2001
Vol 409 Number 6822



Celera: Science 16 February 2001
Vol 291, Issue 5507

153

Genome assembly software

- phrap (Phil’s revised assembly program)
- AMOS (A Modular, Open-Source whole-genome assembler)
- CAP3 / PCAP
- TIGR assembler

154

Next generation sequencing techniques

- Sanger sequencing is the prominent first-generation sequencing method
- Many new sequencing methods are emerging
- See Lars Paulin’s slides (course web page) for details

155

Next-gen sequencing: 454

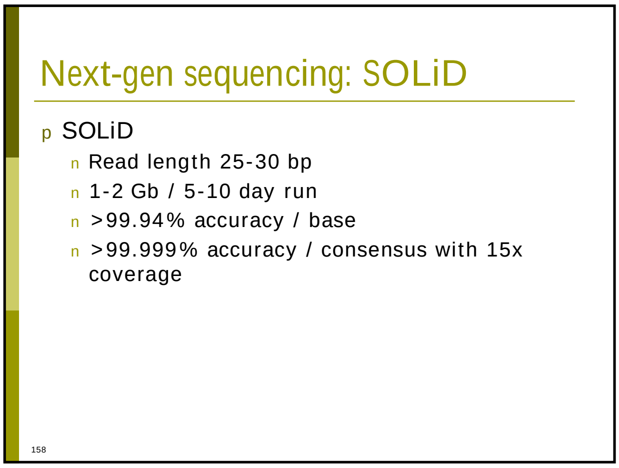
- Genome Sequencer FLX (454 Life Science / Roche)
  - >100 Mb / 7.5 h run
  - Read length 250-300 bp
  - >99.5% accuracy / base in a single run
  - >99.99% accuracy / base in consensus

156

Next-gen sequencing: Illumina Solexa

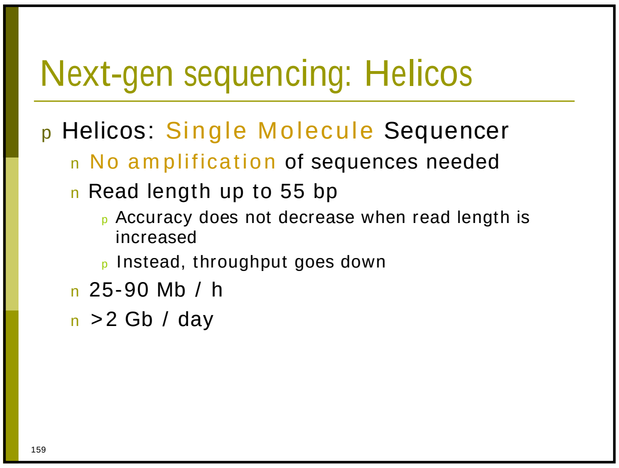
- Illumina / Solexa Genome Analyzer
  - Read length 35 - 50 bp
  - 1-2 Gb / 3-6 day run
  - > 98.5% accuracy / base in a single run
  - 99.99% accuracy / consensus with 3x coverage

157



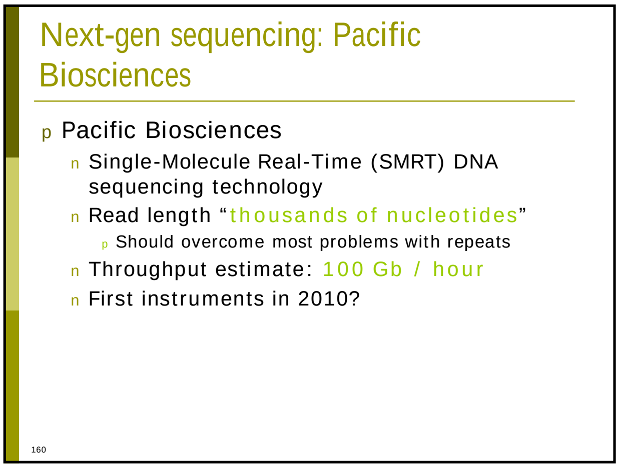
Next-gen sequencing: SOLiD

- SOLiD
  - Read length 25-30 bp
  - 1-2 Gb / 5-10 day run
  - >99.94% accuracy / base
  - >99.999% accuracy / consensus with 15x coverage



Next-gen sequencing: Helicos

- Helicos: Single Molecule Sequencer
  - No amplification of sequences needed
  - Read length up to 55 bp
    - Accuracy does not decrease when read length is increased
    - Instead, throughput goes down
  - 25-90 Mb / h
  - >2 Gb / day



Next-gen sequencing: Pacific Biosciences

- Pacific Biosciences
  - Single-Molecule Real-Time (SMRT) DNA sequencing technology
  - Read length “thousands of nucleotides”
    - Should overcome most problems with repeats
  - Throughput estimate: 100 Gb / hour
  - First instruments in 2010?