Next Generation Sequencing

Tobias Österlund tobiaso@chalmers.se

NGS part of the course

| Week 4	Friday 13/2	15.15-17.00	NGS lecture 1: Introduction to NGS, alignment, assembly
Week 6	Thursday 26/2	08.00-09.45	NGS lecture 2: RNA-seq, metagenomics
Week 6	Thursday 26/2	10.00-11.45	NGS computer lab: Resequencing analysis
Week 7	Thursday 5/3	10.00-11.45	Marcela: Exome sequencing
Week 8	Monday 9/3	17.00	Deadline: Essay on NGS and metagenomics
Week 8	Thursday	08.00-09.45	Fredrik: HMMer and Metagenomics

Essay on Next Generation Sequencing and Metagenomics

Task description

- Write a short essay (roughly 2+2 pages long) introducing *next generation sequenceing* and metagenomics, and discussing their connection, i.e. how next generation sequencing is used as a key technology in metagenomics.

1. The essay should start with a short introduction to a couple of the main next generation sequencing platforms and some of the challenges that arise in the interpretation of data from these. Focus either on:

    - **resequencing**, that is the sequencing of a genome with a previously sequenced reference genome. A major application of resequencing is identification of disease-causing mutations in the human genome.

- or

    - *de novo* **sequencing**, that is the characterization of a completely new genome for an organism where there is no or little previously knowledge.

2. Discuss metagenomics, its dependence on next generation DNA sequencing and the challenges in analyzing the resulting metagenomic data. Identify also (at least) one application of metagenomics and next generation sequencing and give a brief summary of it (them).

Send the essay to tobiaso@chalmers.se at latest Monday 9 March, at 17.00

History of sequencing
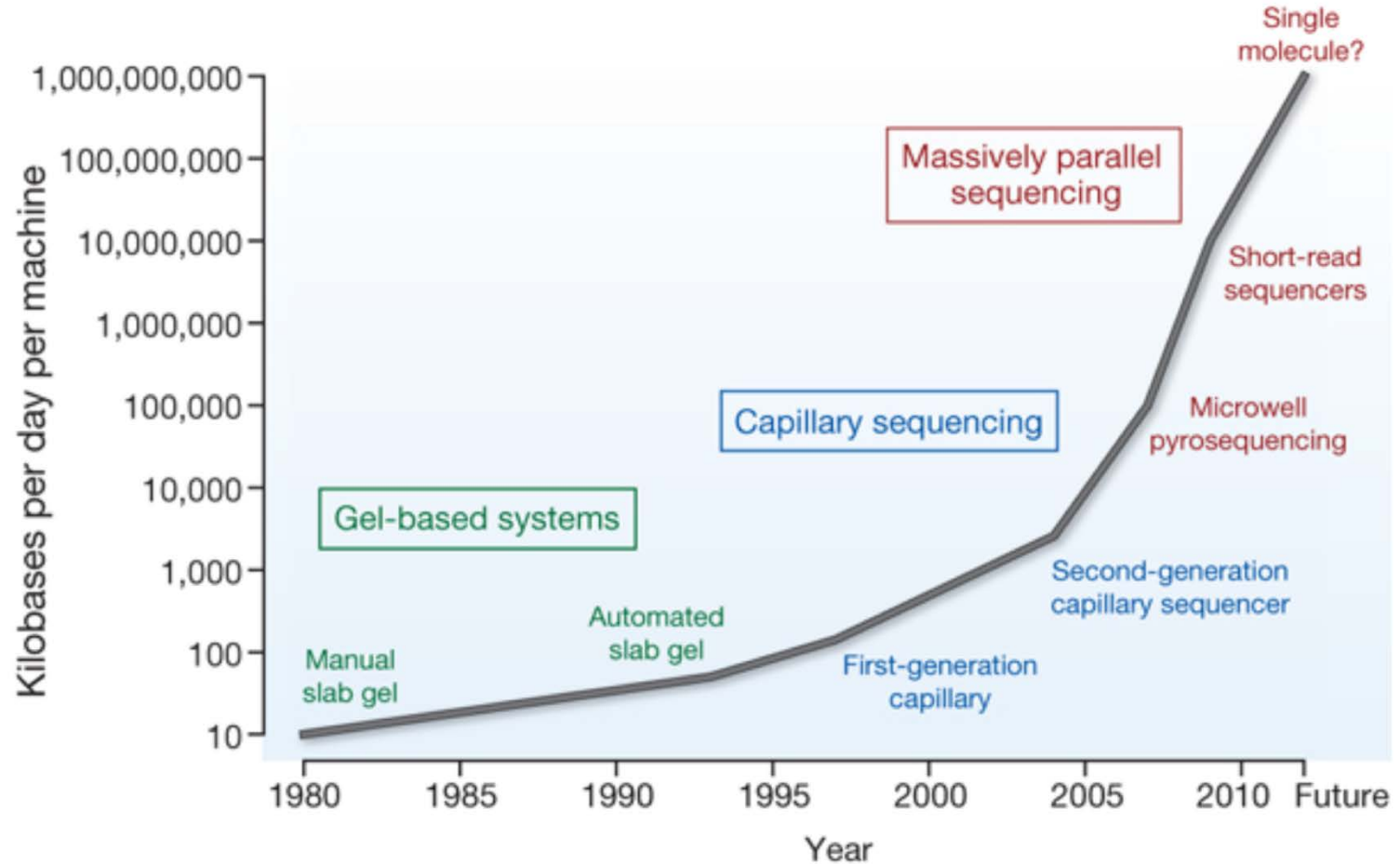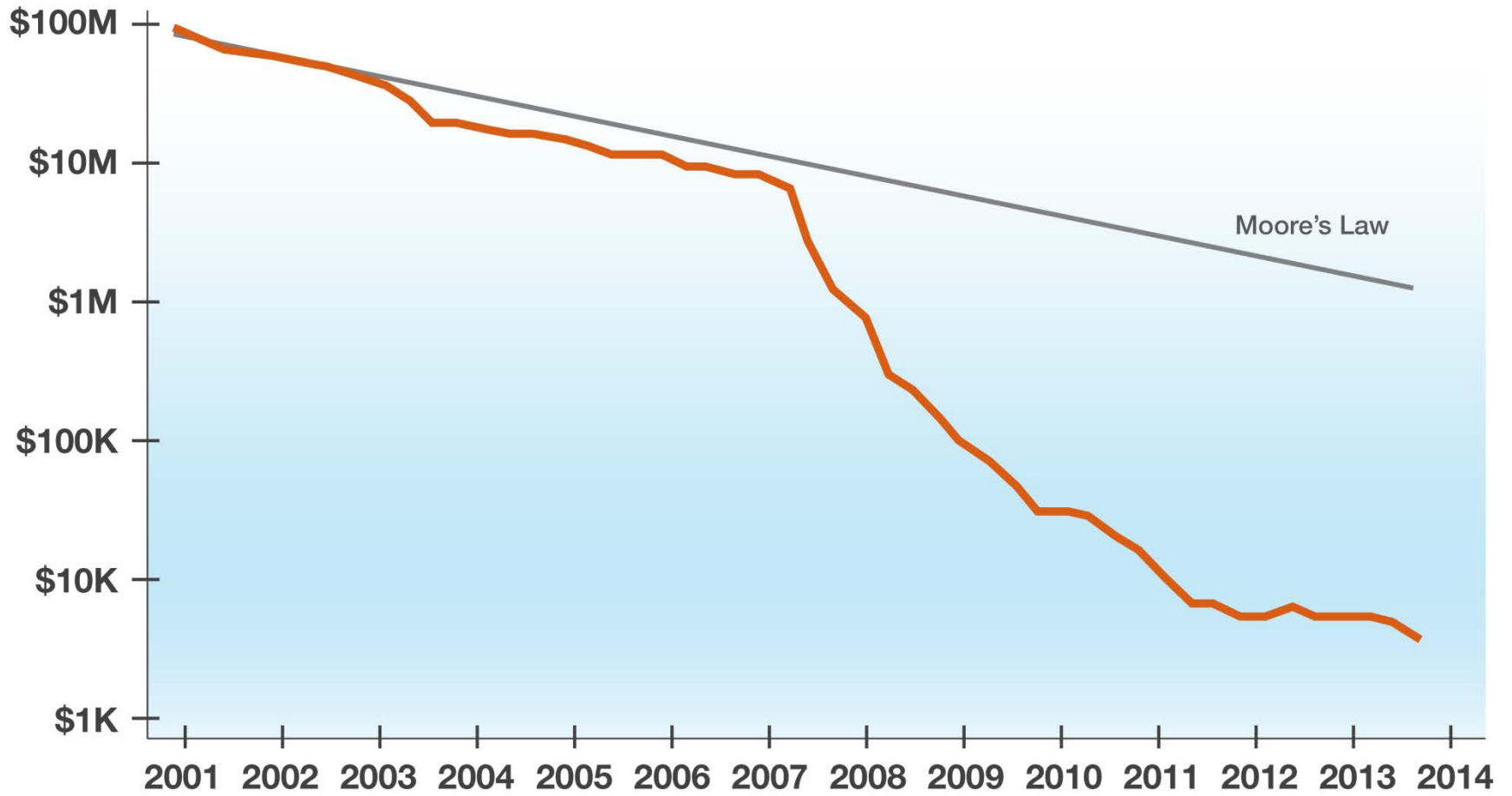
# A paradigm shift

Thanks to NGS:
- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From microarrays to RNA-seq
- From single organisms to complex pools (e.g. metagenomes)
- From model organisms to the species you are actually studying

Generates huge amounts of data
- A bioinformatics challenge

NGS Applications

- Whole genome sequencing
  - De novo sequencing
    - No reference genome available
    - *De novo* assembly
  - Resequencing
    - High quality reference sequence available
    - SNP/Indel detection (whole genome)
    - Genomic rearrangements
  - Application examples:
    - Biodiversity, epidemiology, pathogen detection, evolution
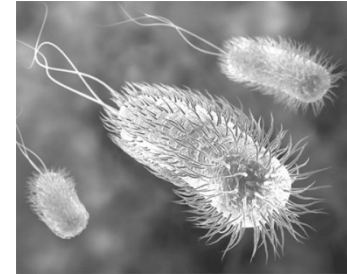
NGS Applications

- Transcriptome sequencing (RNA-seq)
  - Gene expression (differential gene expression)
  - Novel splice sites/splice variants
  - *De novo* transcriptome assembly (find new genes)
  - Non-coding RNA
- Exome sequencing (amplicon sequencing)
  - SNP/Indel detection (only exons)
  - Genomic rearrangements
  - Examples: Medicine, human genetics, cancer research
- Metagenomics
 - Microbial communities (human gut, environment etc.)
  - Species composition
  - Genes/functions

Sequencing techniques

- First generation sequencing
  - Sanger sequencing
- Next generation sequencing (massively parallel sequencing)
  - Illumina sequencing
  - 454 pyrosequencing
  - SOLiD sequencing
- Single molecule sequencing (3rd generation sequencing)
  - PacBio
  - Oxford Nanopore

Sanger sequencing (1st generation sequencing)

– Frederick Sanger 1977

– First sequence: Bacteriophage Phi X 174

– Shearing of DNA

– Cloning in bacteria

– To each sequence reaction dNTP's (dATP,dGTP,dTTP, dCTP) and one of the four ddNTP's are added

– The ddNTP's are incorporated randomly by the DNA polymerase.

– Determine the sequence by gel electrophoresis or fluorescence.

AGTC**A**-
AGTCA**A**-
AGTCAA**G**-
AGTCAAG**T**-
AGTCAAGT**C**-
AGTCAAGTC**G**-
AGTCAAGTCG**G**-
AGTCAAGTCGG**T**-
AGTCAAGTCGGT**C**-
AGTCAAGTCGGTC**A**-

G A T C
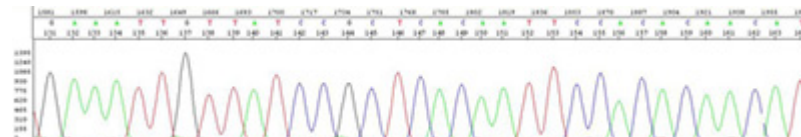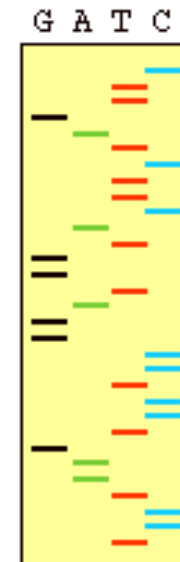
Sanger sequencing (1st generation sequencing)

The "Gold standard" sequencing for single genes

+ High accuracy

+ Long fragments (800-1000 bp)

- Low throughput

Illumina sequencing

- High throughput sequencing
- Sequencing by synthesis
- HiSeq 2500 (high output mode)
  - 2x100 bp reads
  - 2x125 bp reads
  - Paired end sequencing
  - ~50 Gbp per lane
  - 8 lanes per flowcell
  - Run time 5-7 days
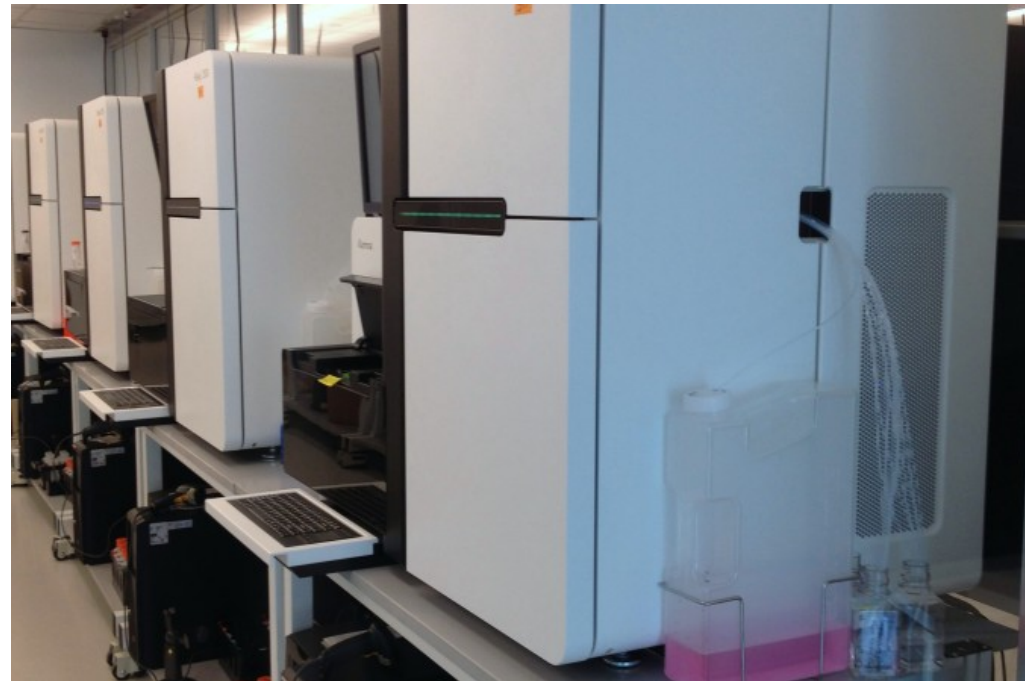
Sample preparation (Illumina)

- DNA extraction (or RNA extraction)
- 1-2 µg of total DNA per sample as starting material

- Shear the DNA to 300-600 bp long fragments (typically 350 bp)
  - Sonication (ultrasound)
  - Random enzymatic digestion
  - Nebulization

Library preparation

- Ligation of adapter sequences to the fragments.
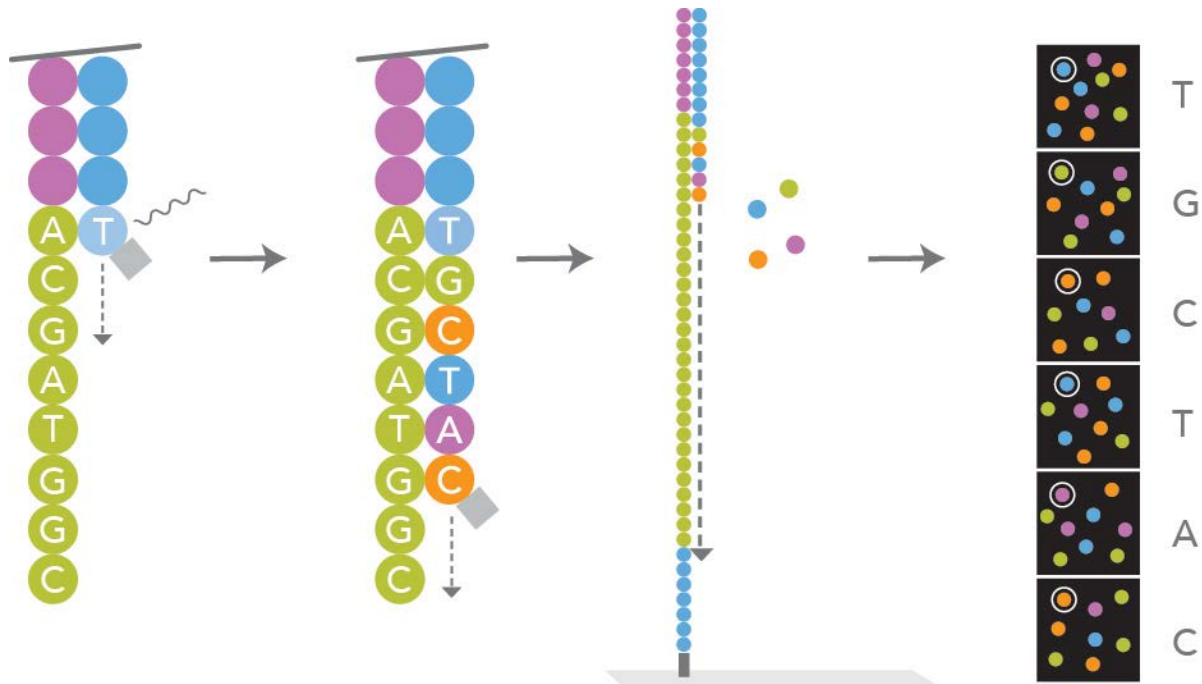- Barcoding possible (makes it possible to run several samples in the same lane)

DNA Fragment (~350 bp)

p5   adapter   fragment   adapter   barcode   p7
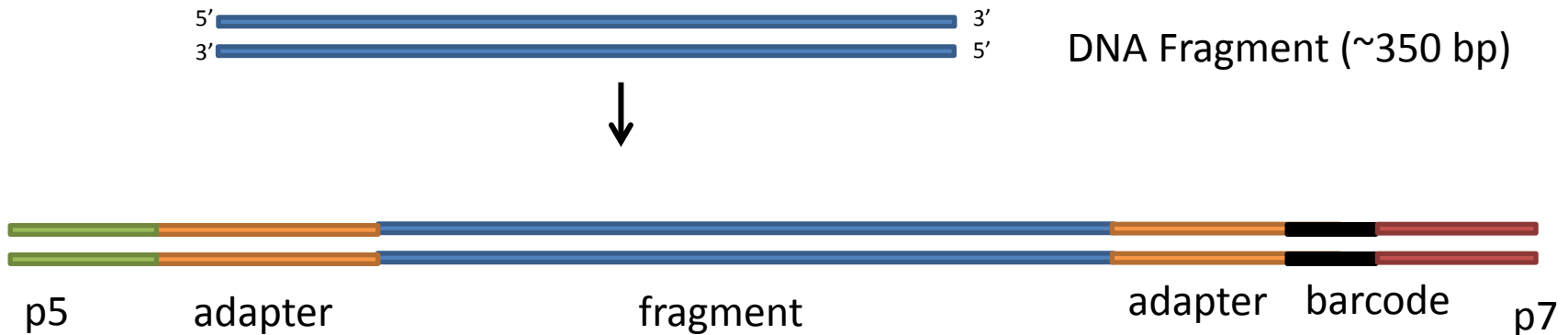
Library preparation for RNA-seq
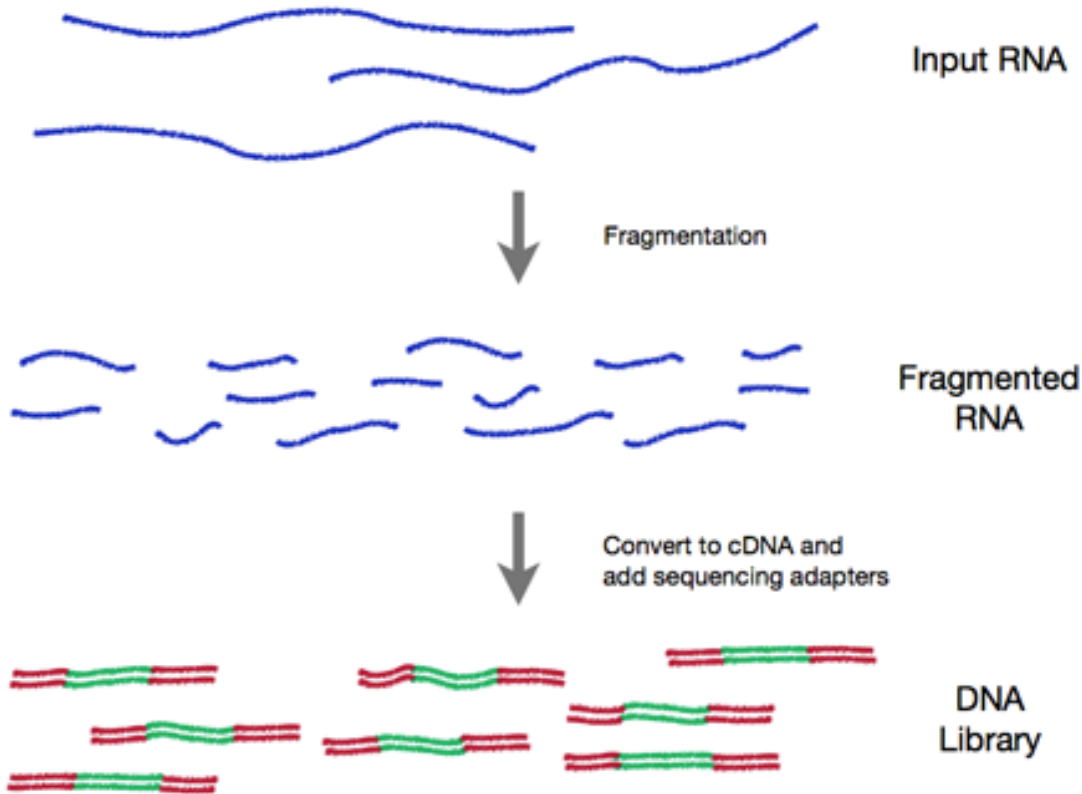
Cluster generation

- PCR amplification step (bridge amplification)
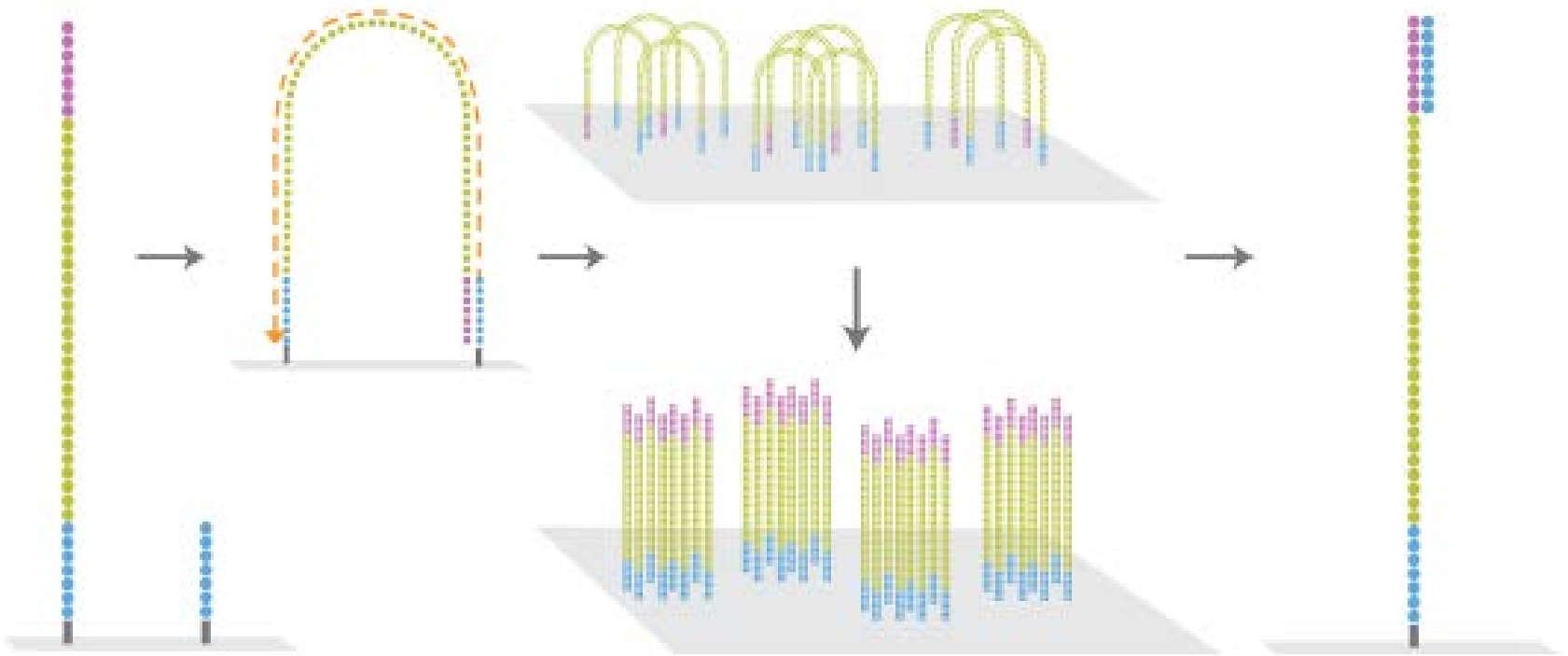
Sequencing by synthesis

Paired end sequencing
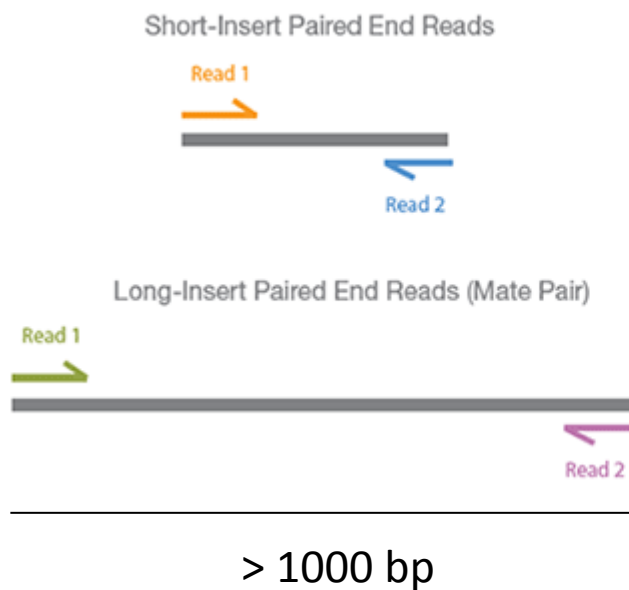
- Information about pairs can help both in alignment and assembly
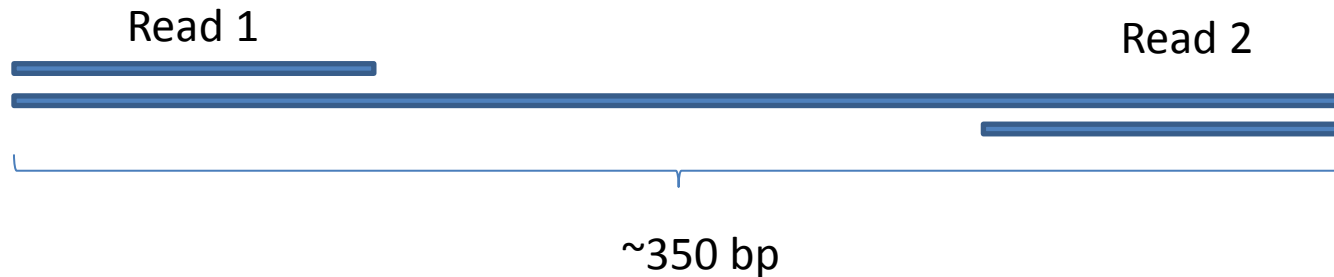
Mate pair reads

- Paired end reads with long (1000-10000 bp) insert size

- Can help in *de novo* assembly

> 1000 bp

Illumina sequencing summary

- Read length ~ 2 x 100 bp

- Advantages
  - High Throughput (~400 gigabases per flowcell)
  - Low cost per base

- Disadvantages
  - Error rate up to 1% (Phred score 20) (only substitutions)
  - Problems with AT- and GC-rich regions
  - Long sequencing times (depending on the read length)

454 sequencing

- Massively parallel pyrosequencing
- Introduced by Roche in 2005
- Read length at least 400 bp
- 1 sequencing round ~500 million bases, takes 4 hours
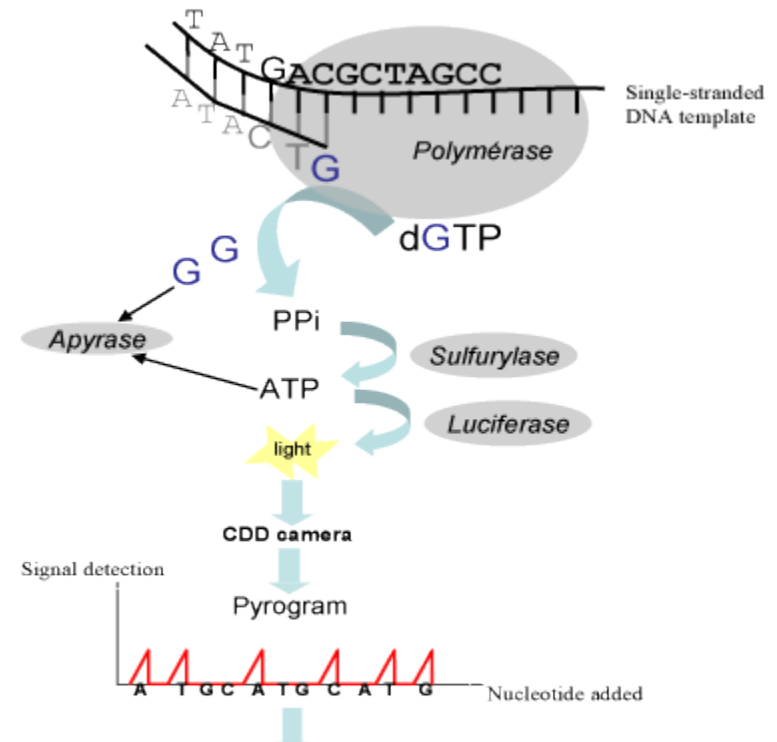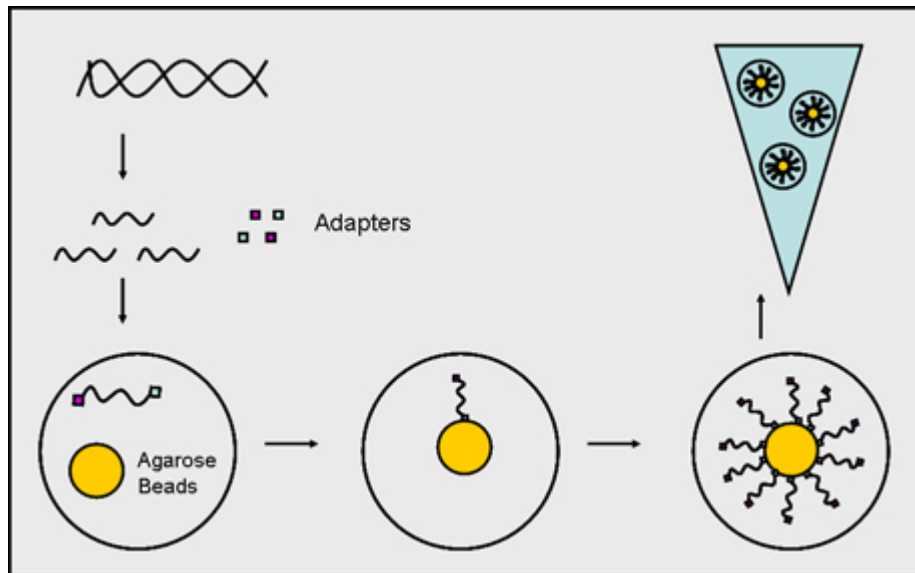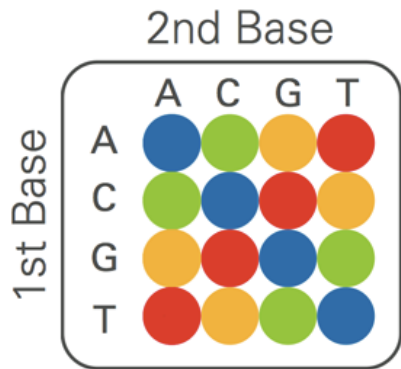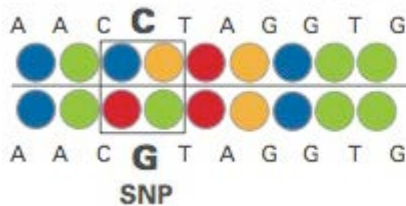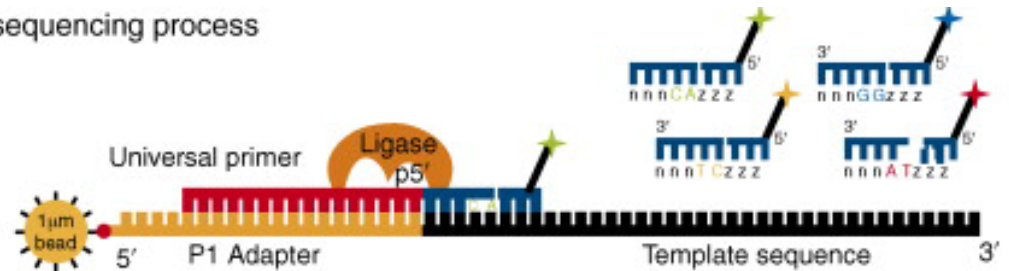
454 sequencing

- Advantages
  - Fast run time
  - Long reads 400-750 bp
- Disadvantages
  - Low throughput (compared to Illumina)
  - Error rate at 1%. Type of errors: Indels
  - Problematic at homopolymeric regions, e.g. TAAAAAAA
  - Relatively high running costs

SOLiD sequencing

- Life technologies/ABI
- Sequencing by ligation (ligase)
- "Color space" dinucleotides
- Paired end, read length 75 + 35 bp

SOLiD sequencing

- Advantages
  - High throughput
  - Low cost per base
  - High accuracy when reference genome is available (resequencing)
- Disadvantages
  - Few software working with color space
  - Problems with AT- and GC-rich regions
  - Long sequencing times

Pacific Biosciences (PacBio)

- Single molecule sequencing (no PCR amplification step)

Fragment (1000-3000 bp)

Adapter ligation
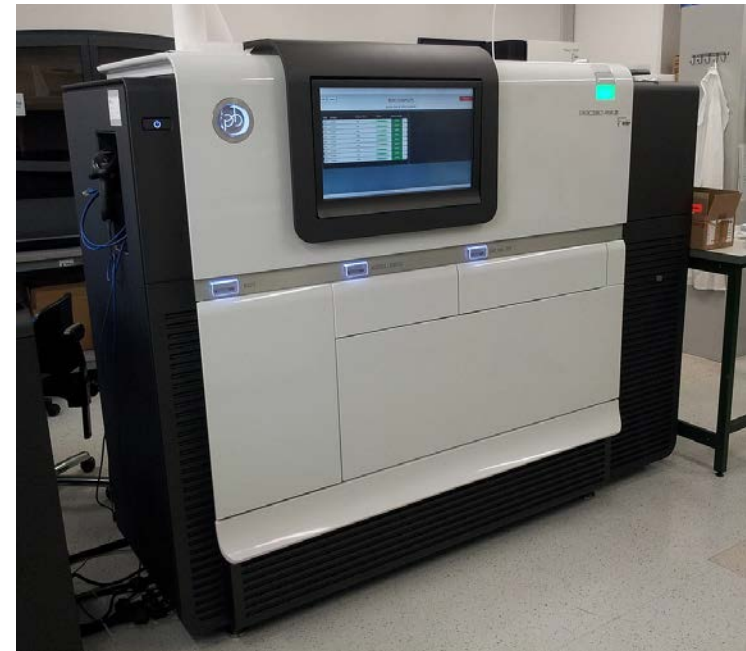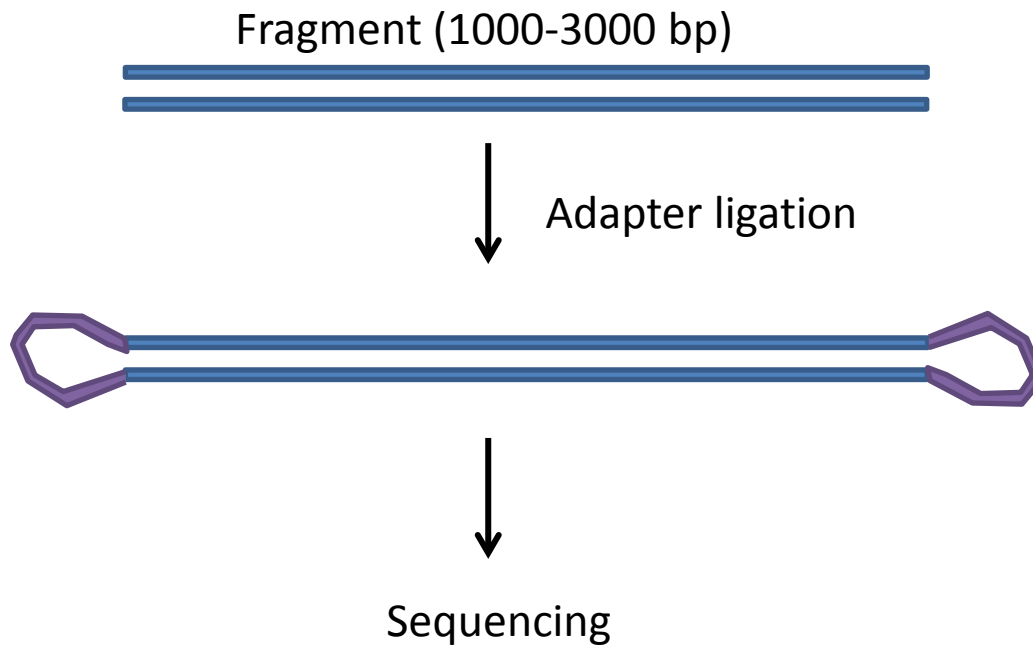
Sequencing
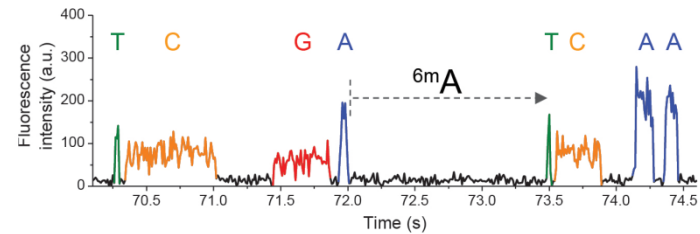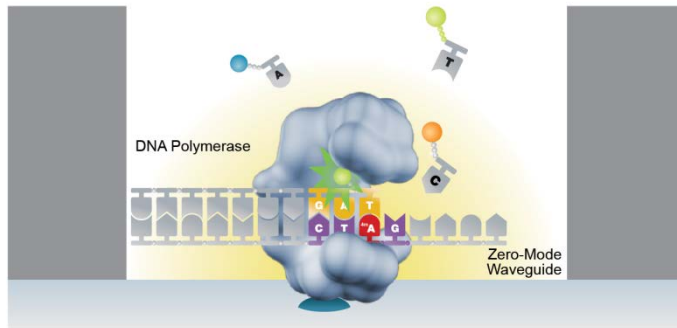
PacBio sequencing

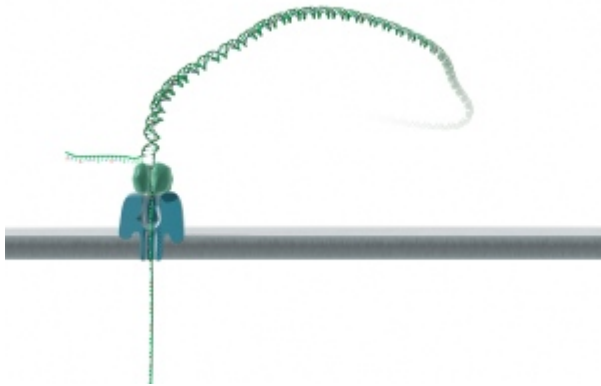

- SMRT – Single Molecule Real Time

- High error rate (up to 10 %)

- Random errors

- Sequence ~10,000 bases – Same fragment sequenced multiple times

PacBio sequencing

- Advantages:
  - Long fragments – average length 1000 bases
  - No PCR amplification step (no PCR bias)

- Disadvantages:
  - High error rate (but random errors)
  - Expensive
  - Low throughput

Oxford Nanopore

"Technique promises it will produce a human genome in 15 minutes"

NGS data analysis

- Pre-processing of raw sequencing reads
  - Remove bad quality data
- For resequencing:
  - Alignment to reference genome
  - Variant detection
- For *de novo* DNA sequencing
  - Assembly of the reads
  - Genome annotation
- RNA-seq

Data analysis

- Huge amount of data to handle
- Many tools are command-based tools
- Working in a Linux environment
- Make use of a bioinformatics server or a computer cluster
- Run several CPUs in parallel

What do the sequence files look like?

- Sample1_2.fastq
- Each sequence file can contain ~20 million reads
- For paired end sequencing: 2 files per sample
- Many samples

Read ID

If the reads are paired, then the first in pair will end with a /1 and the second with a /2
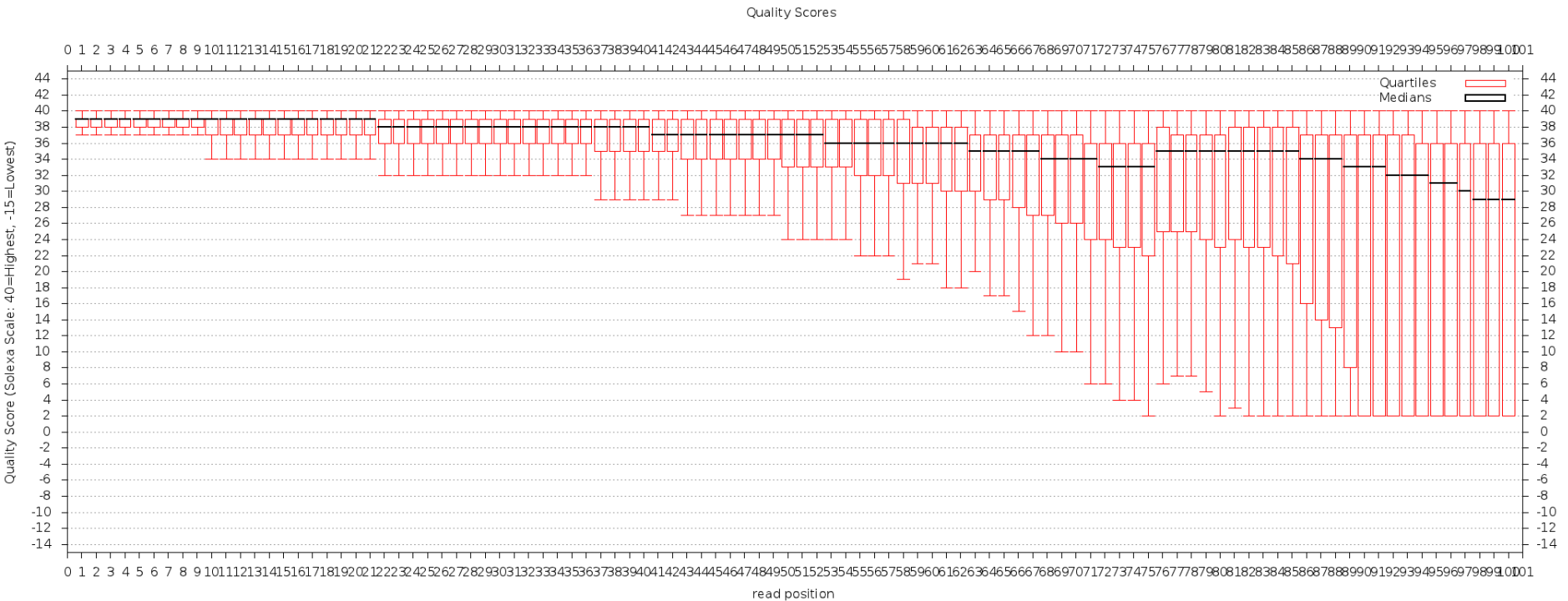
Sequence

```
@HWI-ST951:8:1101:2172:2242#0/2
CTTGGTATCCATTTTGGGTAAGTCATATTCAATGAACTAGGTTTCGCAAACTTTTTGTTCGAAAAGCGGTAGTGCATAGTTATGCT
+
B@CDFBDFHHDFHHHJIJJJJJHIJEGIIIGGII9<CGIGIJJIH=DG8?D>FGGCGC3=@;.;@*.:449>3>;;@6@)55;>:@
```

Quality scores in hexadecimal numbers
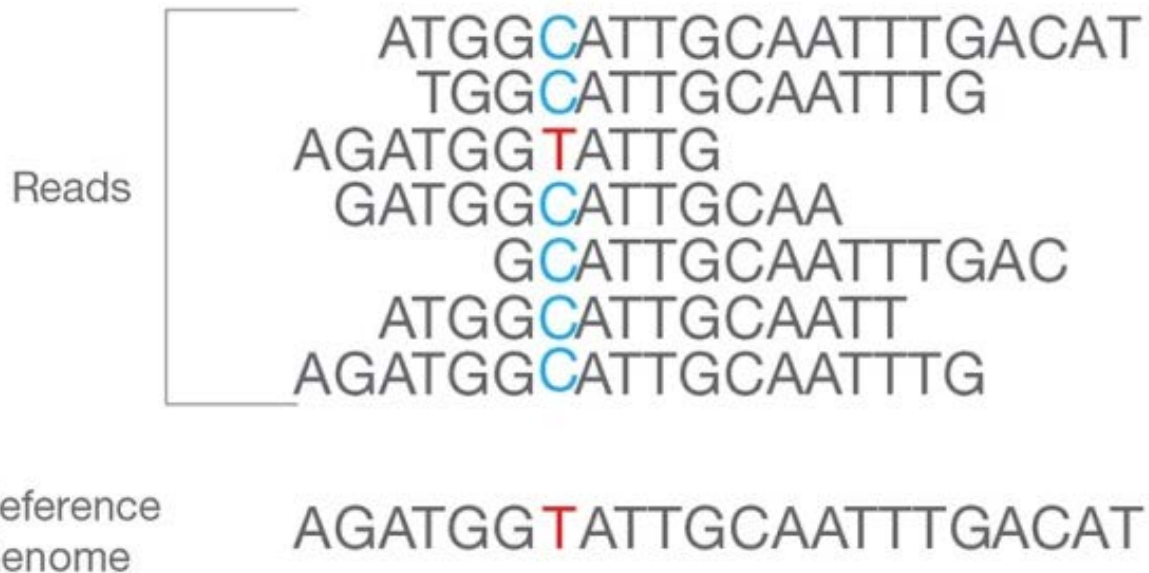
Read quality per base

Example: Using fastx toolkit

Command line:

```
~/fastq_quality_boxplot_graph.sh -i R1_output.txt -o R1_box_plot.png
```

Alignment and data analysis

Example of alignment

Read:

TCAACTCTGCCAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCA

Genome:

ATAAAATGGCCAAAATTAACTAGAAGGTGAGTAGAAACTTAAATAAACTAATTACCATTGATGAGAAAAAAATC
TGCCACTGAAAAAGGCACCCGGTCCAGAGGGTTTCATGAGCGGGAACTGTAGAAACCTTTCGAATTCAACTCTGC
CAACACCTTCCTCCTCCAGGAAGCACTCCTGGATTTCCCTCTTGCCAACAAGATTCTGGGAGGGCAGCTCCTCCA
ACATGCCCCCAACAGCTCTCTGCAGACATATCATATCATATCATATCTTCCATACCATAACTGCCATGCCATACA

How would you find that?

- Brute force

  TCGATCC

  GACCTCATCGATCCCACTG

- Smith-Waterman alignment

- Blast (local alignment)

- Suffix tree

- Burrows-Wheeler transform

Burrows wheeler transform

Read: GATC   Reference: TCGATCC

- Let $ represent the end of the reference string
  - TCGATCC$
- Perform all possible rotations of the string
- Sort the rotations in alphabetical order
- Save only the last column and back-transform when needed

TCGATCC$	$TCGATCC
$TCGATCC	ATCC$TCG
C$TCGATC	C$TCGATC
CC$TCGAT	CC$TCGAT
TCC$TCGA	CGATCC$T
ATCC$TCG	GATCC$TC
GATCC$TC	TCC$TCGA
CGATCC$T	TCGATCC$

Alignment of reads

- Smith-Waterman
- BLAST (100 times faster)
- Vmatch (suffix trees) 100 times faster
- Bwa, bowtie etc. (based on Burrows-Wheeler transform) fast alignment

Alignment of reads to protein reference

Fast and sensitive protein alignment using DIAMOND

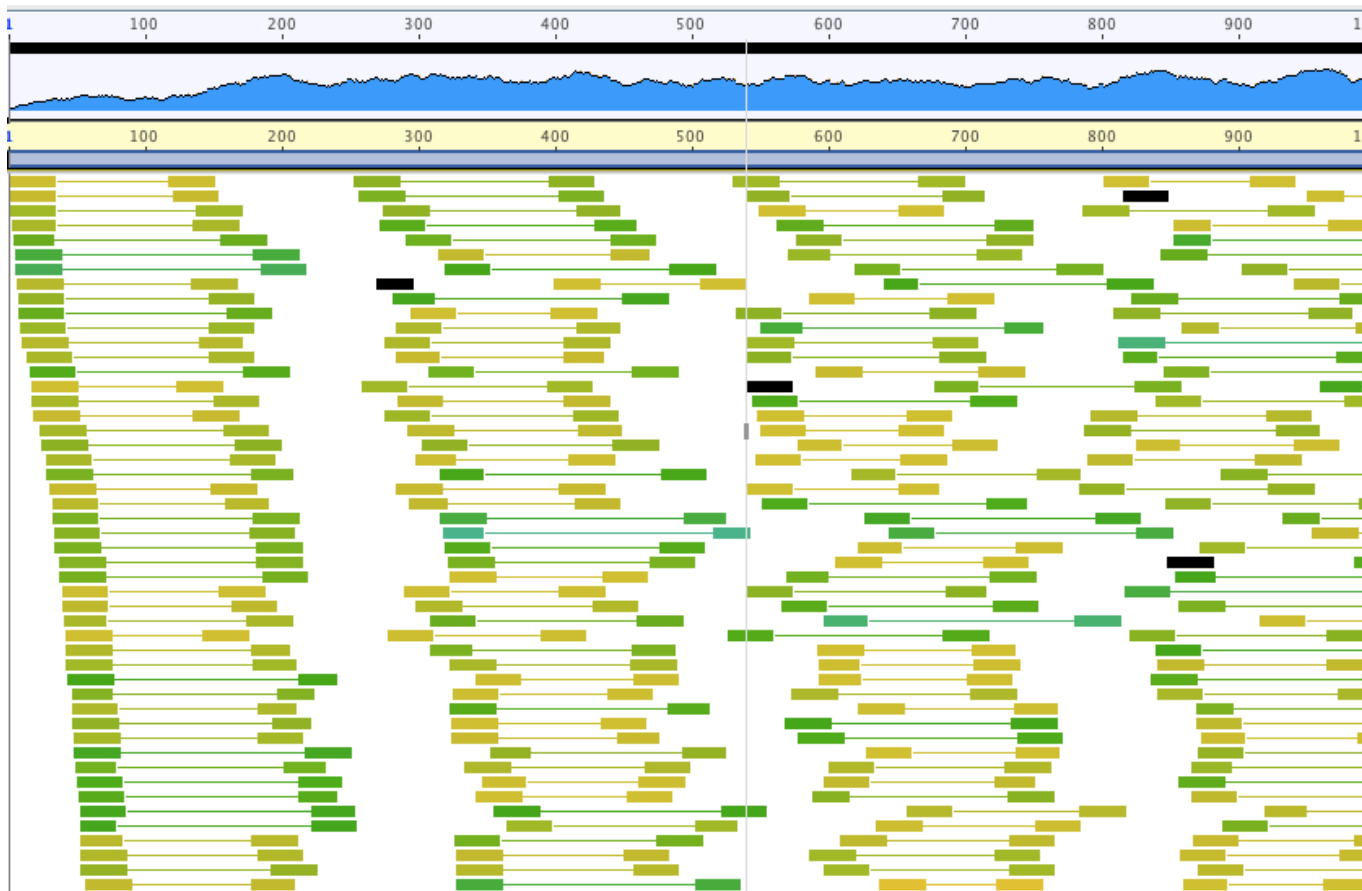Benjamin Buchfink[1], Chao Xie[2,3] &
Daniel H Huson[1,2]

Nature Methods, January 2015

Claimed by the authors to be 20,000 times faster than blastx in mapping short reads to a protein database.
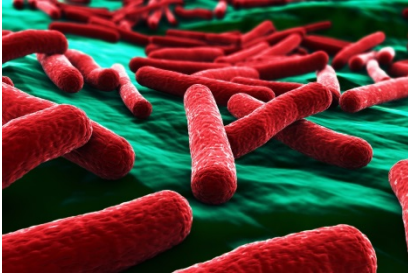
De novo assembly

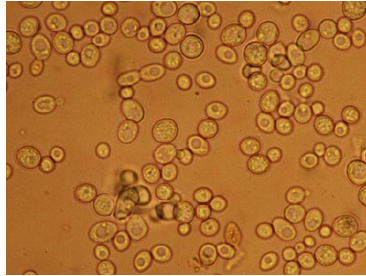- Puzzle reads together to longer fragments (contigs)

Genome assembly – challenges

- Computationally heavy
  - Computational complexity: $o(n^2)$
  - Memory complexity: $o(n^2)$
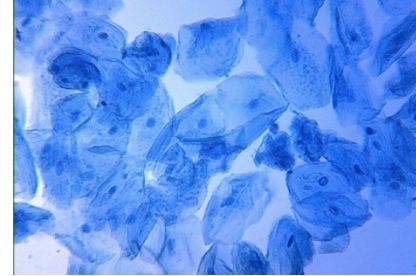- Sequencing errors
- Repetitive regions

Genome complexity

E. coli
~4.5 Mbases
~4300 genes

S.cerevisiae
~12 Mbases
~6000 genes

Human
~3 Gbases
~20000 genes

Spruce
~20 Gbases
~30000 genes

Increased genome complexity

Assembly of the spruce genome

- Large and complex genome
  - 20 gigabases (6 times as big as the human genome)
  - Many repetitive regions
- Assembly statistics
  - 1 terabases sequenced (mainly Illumina)
  - 3 million contigs longer than 1000 bases
  - 30 % of the genome
  - Assembly had to be done on a supercomputer with 1 TB RAM.

Summary – Next generation sequencing

- Next generation sequencing enables sequencing of billions of DNA fragments simultaneously

- Huge amount of sequence data in a short time

- Highly applicability in many areas of biology and medicine

- Needs bioinformatics to handle and analyze the produced data