Next Generation DNA Sequencing - Introduction

MVE360 – Bioinformatics, 2012

Erik Kristiansson, erik.kristiansson@chalmers.se

Agenda

- Part 1: Sequencing Techniques
  - The history of DNA sequencing
  - Sanger sequencing (first generation sequencing)
  - Next generation sequencing
    - Massively parallel pyrosequencing
    - Sequencing by synthesis
    - Sequencing by ligation
  - Data analysis
- Next lecture: Applications

History of DNA sequencing

- Structure of the DNA discovered in 1953.
- First sequences in 1965.
- Rapid DNA sequencing developed by Frederick Sanger 1977.


Watson & Crick


Fred Sanger

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown*, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III†, P. M. Slocombe‡ & M. Smith*

History of DNA sequencing

- **Bacteriophage Phi X 174**
  - First sequenced genome. Done by Fred Sanger.
  - 11 genes, 5,386 bases
  - Published 1977

- **Haemophilus influenzae**
  - First sequenced free living organism
  - 1800 genes, 1.8 million bases
  - Published 1995

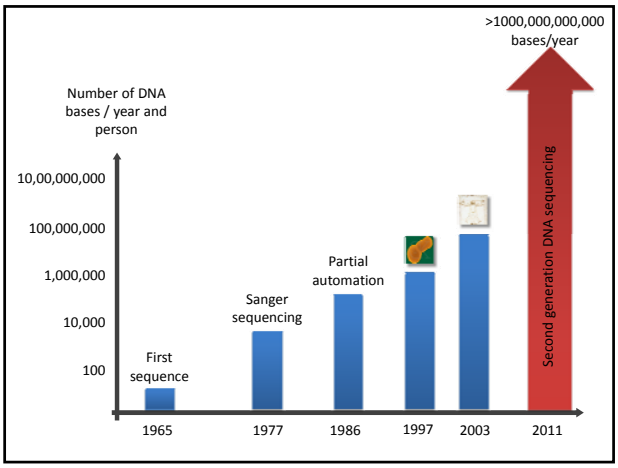
History of DNA sequencing

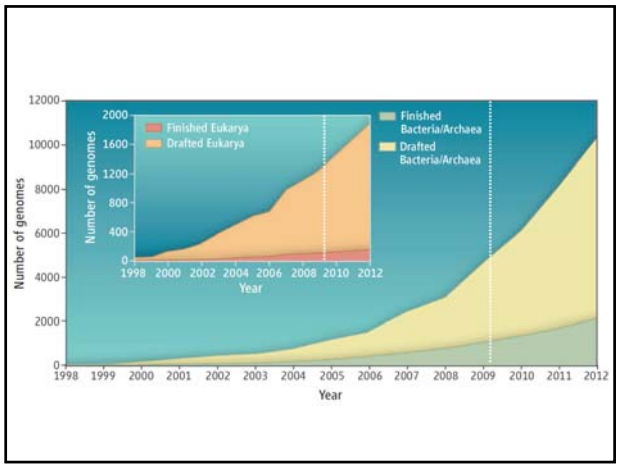
- **Saccharomyces cerevisiae**
  - First sequenced eukaryote
  - Genome consists of 6000 genes and 12 million bases
  - Published 1997 – the project took 7 years

- **Homo sapiens**
  - The Human Genome Project
  - Genome consists of ~21.000 genes and 3.25 billion bases

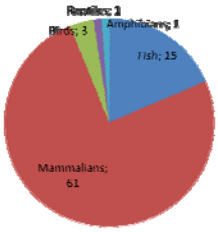
HGP – The Human Genome Project

- Initiated 1990 – finished 13 year later
- Largest research project
  - 200 research groups worldwide
  - Total cost was $3 billion
- Sequence still not 100% complete
- The Holy Grail: $1000 genome!

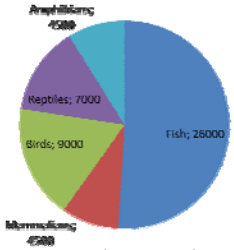




Genome sequencing



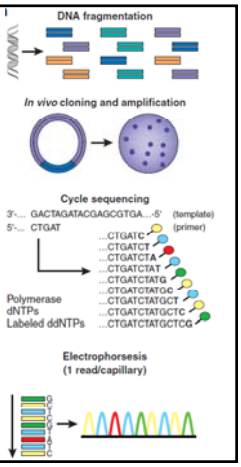
Planned whole-genome sequencing projects for vertebrates



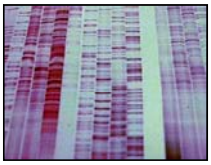
Vertebrate species distribution

Sanger sequencing

- "First generation" sequencing
  - Introduced 1977
  - Cloning in bacteria
  - Chain-termination combined with gel electrophoresis
  - Originally 80 bases fragments, today up to 1000 bases
  - High accuracy

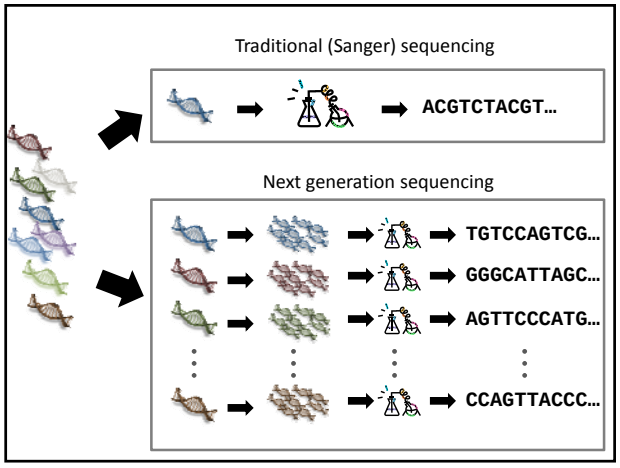


Sanger sequencing



Gel used for base-calling

Multiple sequencing machines at the Sanger institute



Next generation DNA sequencing

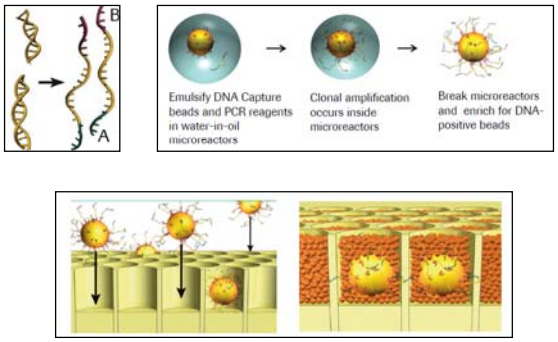
- Introduced in 2005
- From serial to parallel – multiple fragments of DNA sequenced simultaneously
- Many techniques on the market
  - Massively parallel pyrosequencing (454)
  - Sequencing by synthesis (Illumina)
  - Sequencing by ligation (SOLiD)

454 sequencing

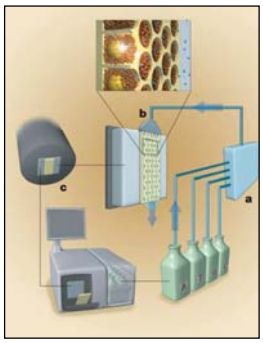
- Massively parallel pyrosequencing
- Introduced 2005 by 454 Life Sciences/Roche
- Read length around 1000 bases
- One sequencing run
  - Generates 1,2 million reads, 500 million bases
  - Takes four hours

GS FLX Pyrosequencer

454 sequencing

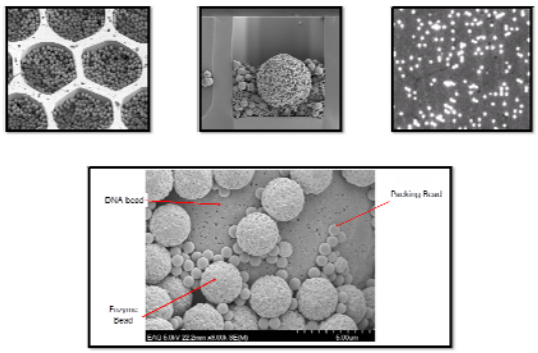


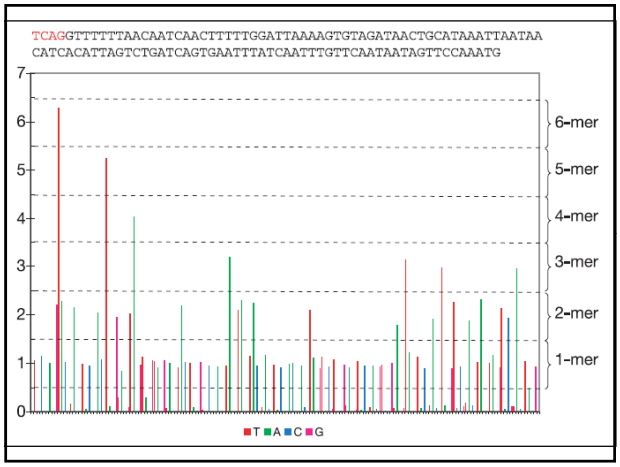
454 sequencing



- Nucleotides are flowed sequentially (**a**)
- A signal is generated for each nucleotide incorporation (**b**)
- A CCD camera is generating an image after each flow (**c**)
- The signal strength is proportional to the number of incorporated nucleotides.

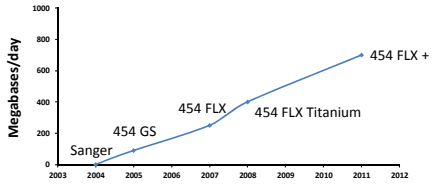
454 sequencing





454 sequencing

- Advantages
  - Handles GC-rich regions (fairly) well
  - Long reads
  - Fast sequencing runs
- Disadvantages
  - Low throughput
  - Error rate at 1%
  - Homopolymeric regions are problematic

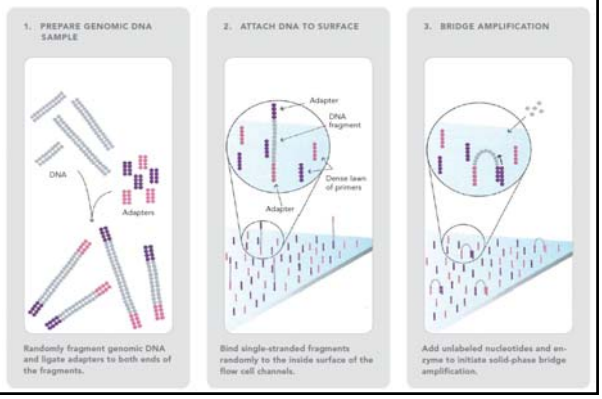


Illumina sequencing

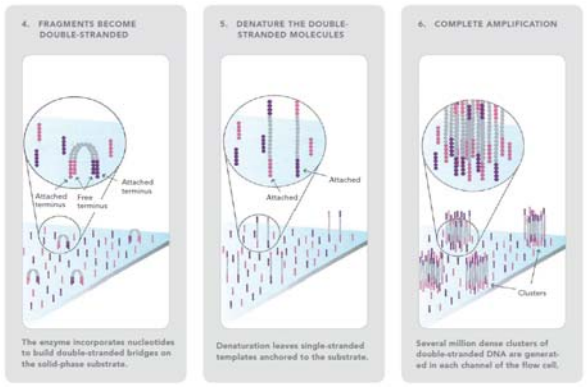
- Developed by Solexa (acquired by Illumina)
- Sequencing by synthesis – cyclic reverse termination
- HiSeq 2000
  - 3 billion reads, 35-100 bases/read
  - Up to 600 billion bases/run, 25 gigabases/day
- Long sequencing times

HiSeq 2000

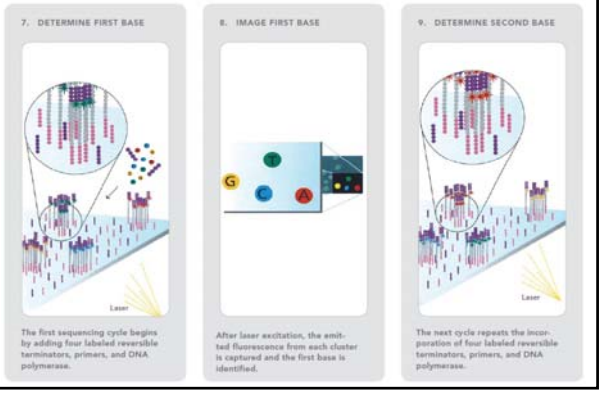
Illumina sequencing

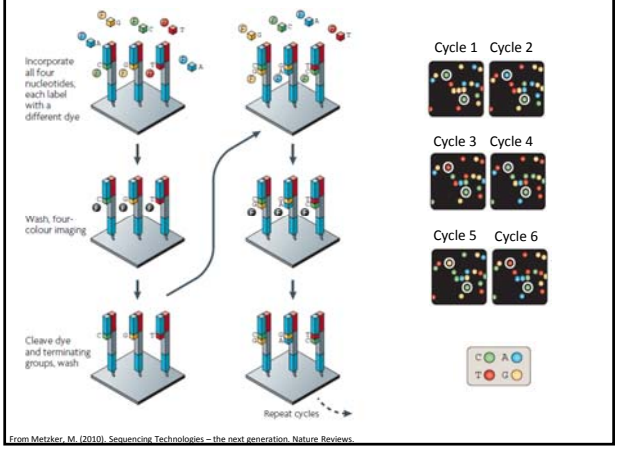


Illumina sequencing

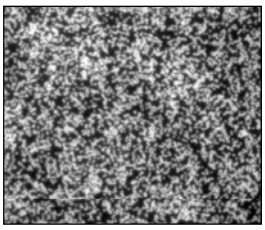


Illumina sequencing

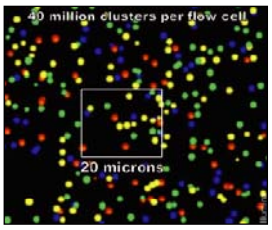




From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.



Raw image



Pseudo-colored image

- Around 100,000 high-resolution images are analyzed during an sequencing run (terabytes of data).

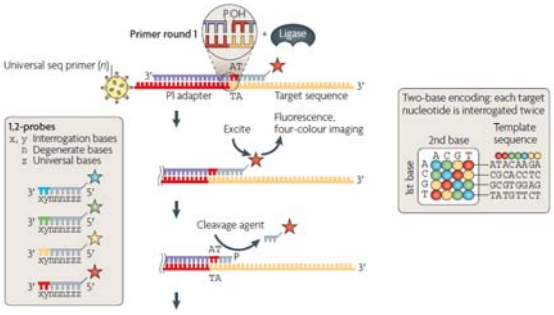
Illumina sequencing

- Advantages
  - High throughput (max 600 gigabases/run)
  - Low cost per base
- Disadvantages
  - Error rate at 1% - only substitutions
  - Problems with AT- and GC-rich regions
  - Long sequencing times (dependent on the read length)

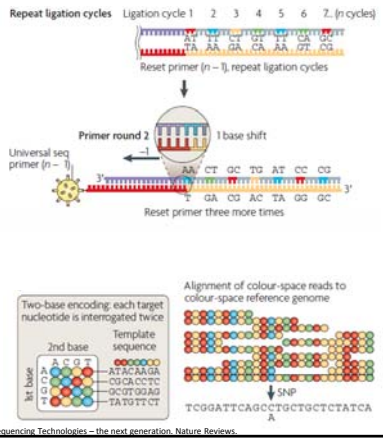
SOLiD sequencing

- Life Technologies/ABI
- Amplification with emulsion PCR
- Sequencing by ligation
- 5500xl
  - 5 billion reads, 35-75 bases/read
  - Up to 300 billion bases/run, 30 gigabases/day
- Long sequencing times

Sequencing by ligation



From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.



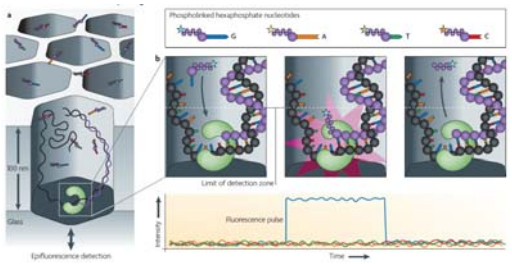
From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.

SOLiD sequencing

- Advantages
  - High throughput (max 300 gigabases/run)
  - Low cost per base
  - High accuracy when a reference genome is available
- Disadvantages
  - Few software working with “color space”
  - Problems with AT- and GC-rich regions
  - Long sequencing times (dependent on the read length)

Pacific Bioscience

- Real-time sequencing
- Long fragments – average length 1000 bases
- High error rate (up to 10%)

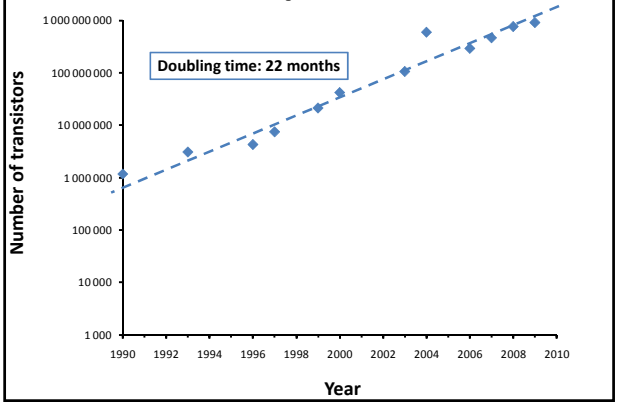


From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.

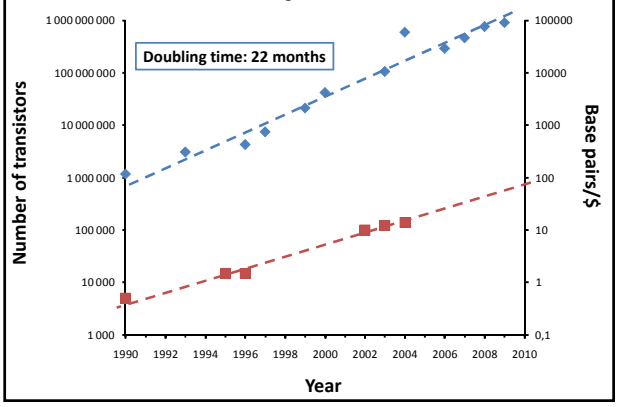
Overview of NGS technologies

Technology	Company	Throughput	Cost/base	Read length
Parallel pyrosequencing	454 Life Sciences /Roche	0.5Gb/day	$0.01	~700 bases
Illumina sequencing	Solexa/Illumina	100 Gb/day	$0.000001	35-100 bases
SOLiD	Life Technologies	100 Gb/day	$0.000001	35-75 bases
HeliScope	Helicos Biosciences	5 Gb/day	$0.005	25-55 bases
PacBio RS	Pacific Biosciences	Unknown	Unknown	~1000 bases
Traditional sequencing		0.00005Gb/day	0.5$	up to 1000 bases

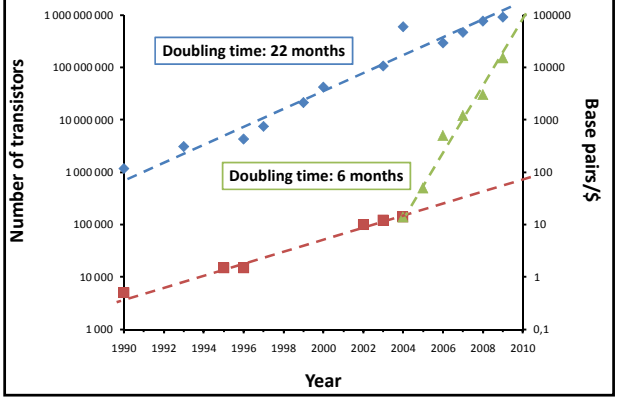
The development of NGS



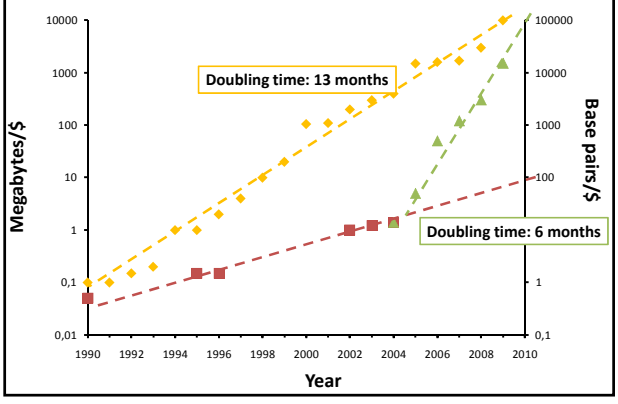
The development of NGS

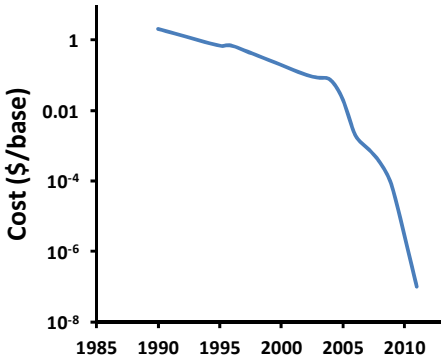


The development of NGS



Kryder's Law





Will 2012 be the year of the holy grail?

- Yes, according to Life Technologies
  - Ion Torrent Proton: Sequencing of a human genome in 2 hours for less than $1000.

Next generation sequencing

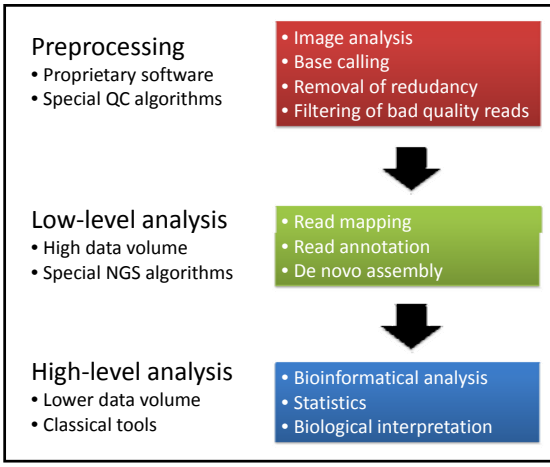
- Unprecedented amount of data
- Computationally efficient methods needed
- Lack of biostatisticians/bioinformaticians for the future?

Spruce genome sequencing project

Genome: 20 billion bases
Collaboration between UU, KTH and SU.
Sequenced using Illumina and 454
Supercomputers needed for the analysis

Data analysis

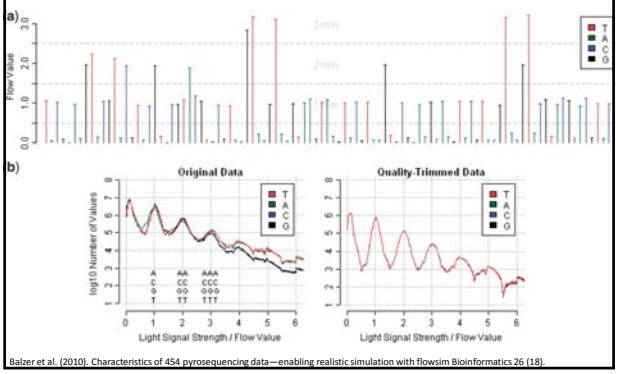
- Large data volumes
  - Optimized algorithms
  - Computationally heavy – high performance computing and eScience
- Different algorithms for different read lengths
- Different platforms have different error patterns
- Many classical bioinformatical tools are still useful



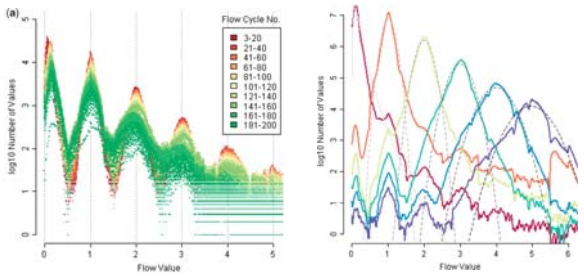
Preprocessing

- Quality assessment of NGS data is essential!
  - High error rate
  - Problematic regions
- Actions to increase quality
  - Removing bad sequencing runs
  - Filtering/trimming bad reads
  - Removing redundant reads
  - Multiplexing: Reads without interpretable barcode
- Low cost/base – possible to throw away more

Error patterns in 454 sequencing



Balzer et al. (2010). Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim Bioinformatics 26 (18).



- Sequencing quality decreases with read length
- Homopolymeric errors are more likely for long repeats

DNA alignment

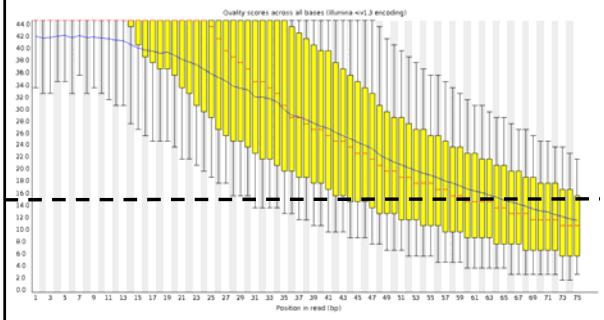
```
CAAGACAAATCTTAGTTACGCTAATTTTT--CTAGAGTCATATTAGAAAAAGTGCGAACTGTGGGAAAATCGC
CAAGACAAATCTTAGTTACGCTAATTTTT--CTAGAGTCATATTAGAAAAA-GTGCGAACTGTGGGAAAATCGC
CAAGACAAATCTTAGTTACGCTAATTTTT--CTAGAGTCATATTAGAAA---GTGCGAACTGTGGGAAAATCGC
CAAGACAAATCTTAGTTACGCTAATTTTT--CTAGAGTCATATTAGAAAA---GTGCGAACTGTGGGAAAATCGC
CAAGACAAATCTTAGTTACGCTAATTTTTTCTAGAGTCATATTAGAAAAA-GTGCGAACTGTGGGAAAATCGC
**************************** ************************ ********************
```

Protein alignment

```
KTNLSYANRSRVILEKSANCGKI
KTNLSYANRSRVILEKVRTVGKS
KTNLSYANRSRVILE-SANCGKI
KTNLSYANRSRVILEKCELWENR
KTNLSYANRSRVILEKCELWENR
KTNLSYANRFLESYKKCELWENR
********
```

- Around 50% of all errors in 454 data are related to homopolymers
- Substitutions should not be present in theory but exist (they are rare).

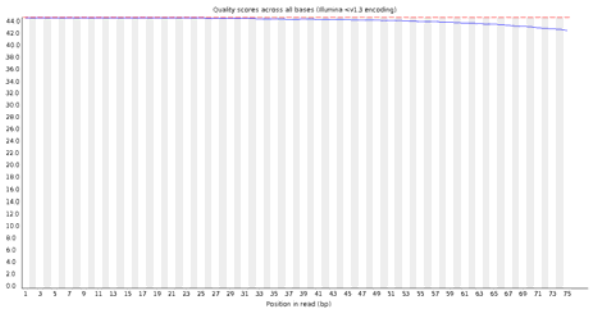
Error patterns in Illumina sequencing



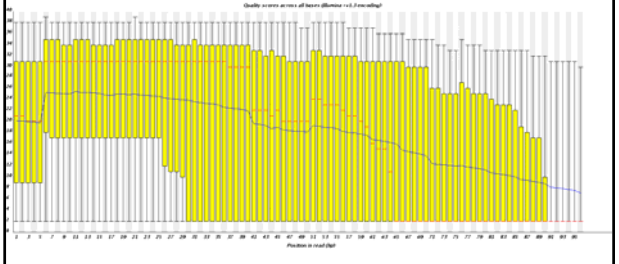
Based on 10 million reads from one sequencing run on a Genome Analyzer IIx.

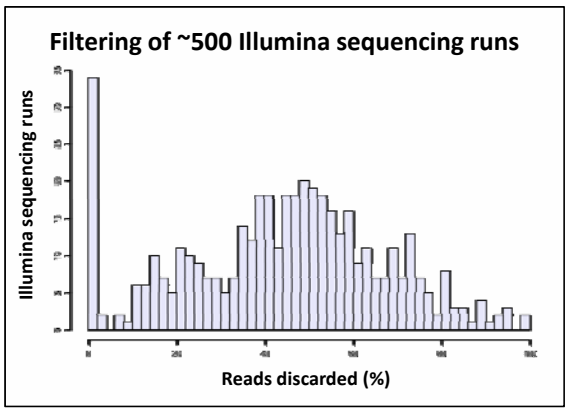
Error patterns in Illumina sequencing

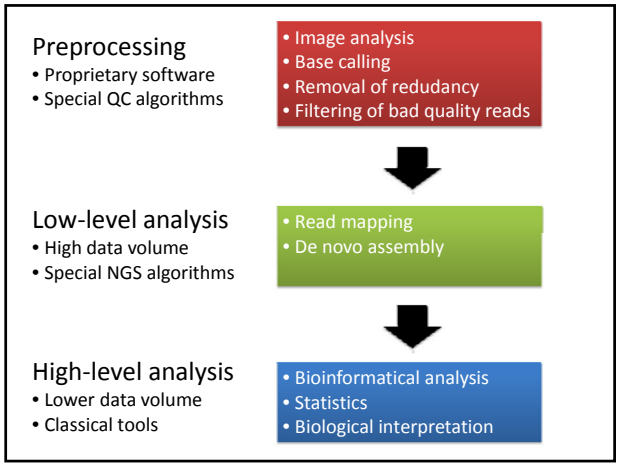
- High variability between runs!



Error patterns in Illumina sequencing



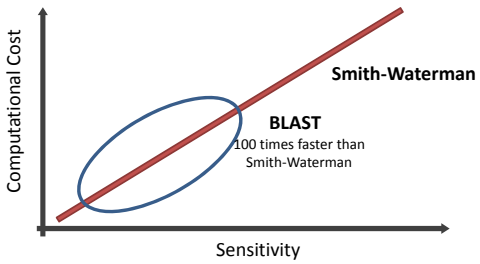




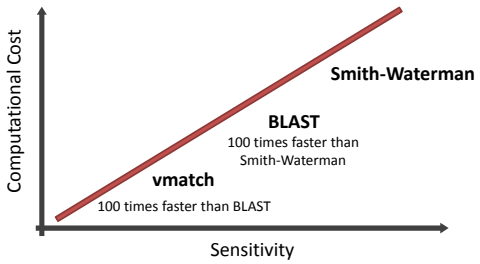
Read Mapping

- Comparison of reads against a reference genome
- Traditional algorithm: Smith-Waterman (e.g. BLAST)
- Faster algorithms
  - Hash tables of k-meres (e.g. SSAHA2)
  - Burrows-Wheeler transform (e.g. bwa)
  - Suffix-arrays (e.g. vmatch)
- Complexity scales linear to the amount of data

Read Mapping



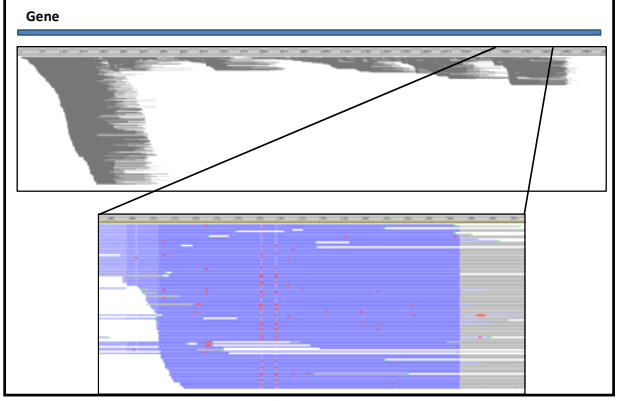
Read Mapping



Read Mapping

- Applications
  - Genome/exome resequencing
    - Genetically linked diseases
    - Cancer research
    - Infectious diseases
  - Transcriptomics (RNA-seq)
    - Large-scale gene expression analysis

RNA-seq - RNAseq

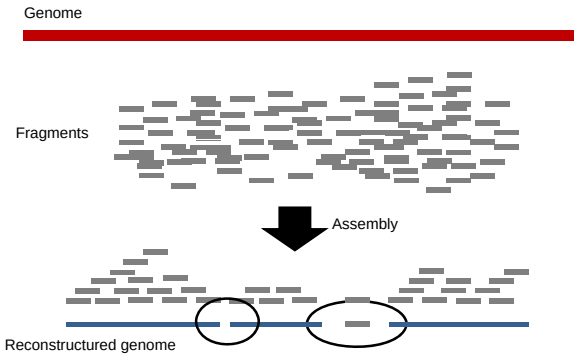


De novo assembly

- Form the sequenced fragments into a contiguous stretch of DNA
- Applications:
  - Genome sequencing
  - Transcriptome sequencing

- Naïve algorithm
  1. Compare all fragments with each other using pairwise alignments.
  2. Identify the fragments with the best overlap - merge
  3. Repeat

Genome assembly



Genome assembly - challenges

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

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 are today generated in short time
- Novel bioinformatical approaches are need to handle and analyze the produced data
- High applicability in many areas of biology and medicine