Genome-scale technologies 2 / Algorithmic and statistical aspects of DNA sequencing

Introduction to next generation sequencing and its applications

Ewa Szczurek
szczurek@mimuw.edu.pl

Instytut Informatyki
Uniwersytet Warszawski

Organisational remarks

Organisation of this course

- In English
- 1.5 hrs lecture
- Up to 2×1.5 hrs lab
- Homework 15% of your grade

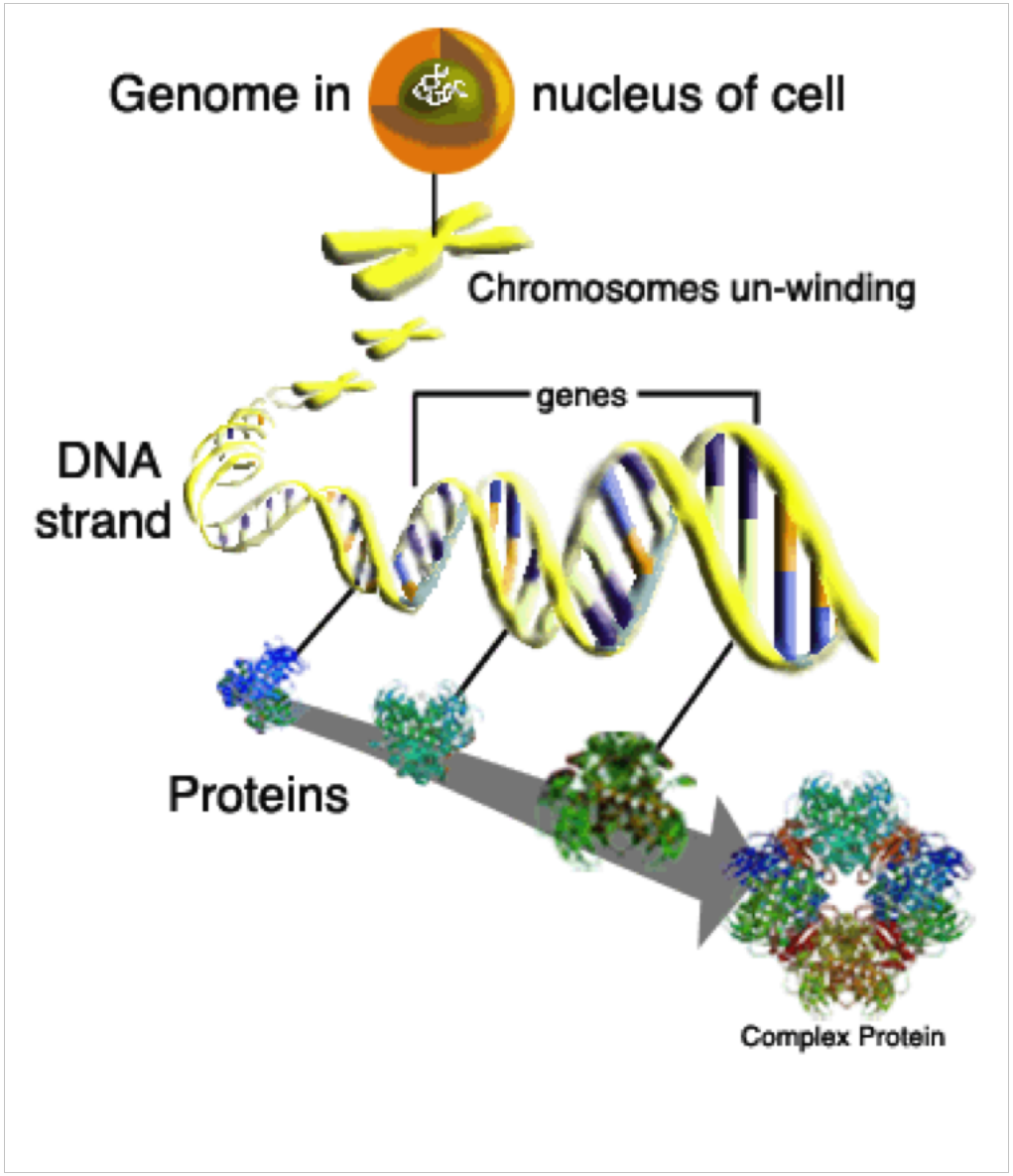
Final assesment

- Data analysis project
- Successful completion of the project $\Rightarrow$ 85% of the final grade
- Oral exam

Scope of this course

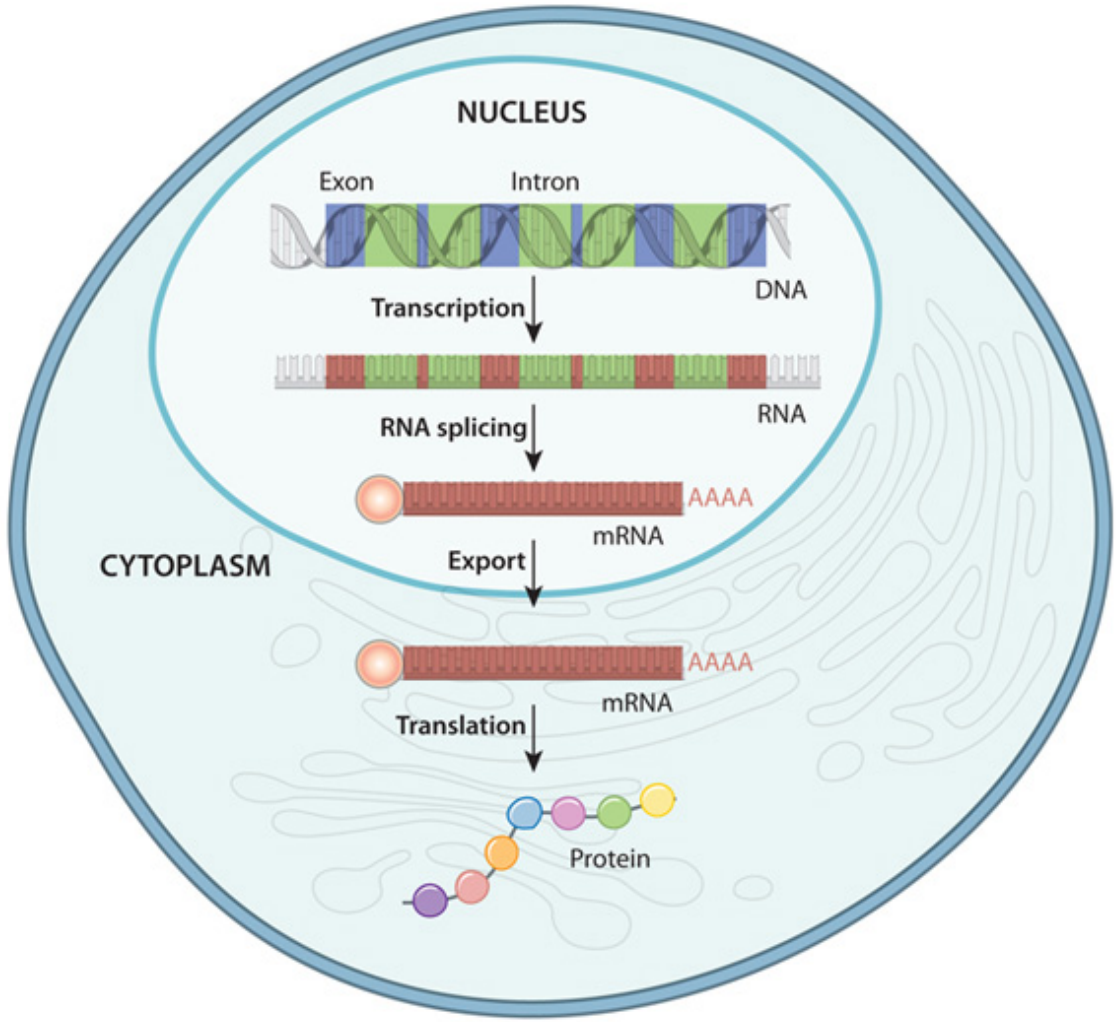
- Next generation sequencing (NGS) data
- Its analysis and applications

Short biological introduction



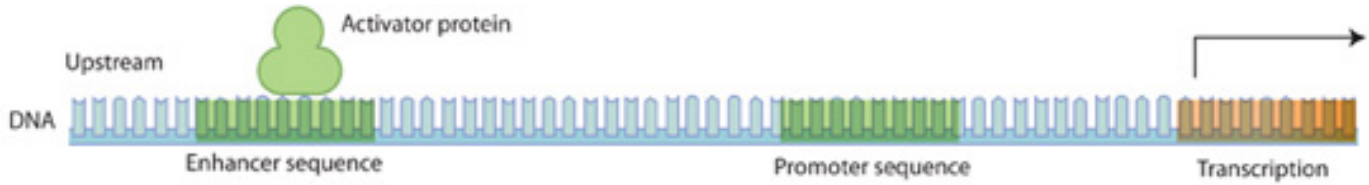
DNA sequence determines how cells work

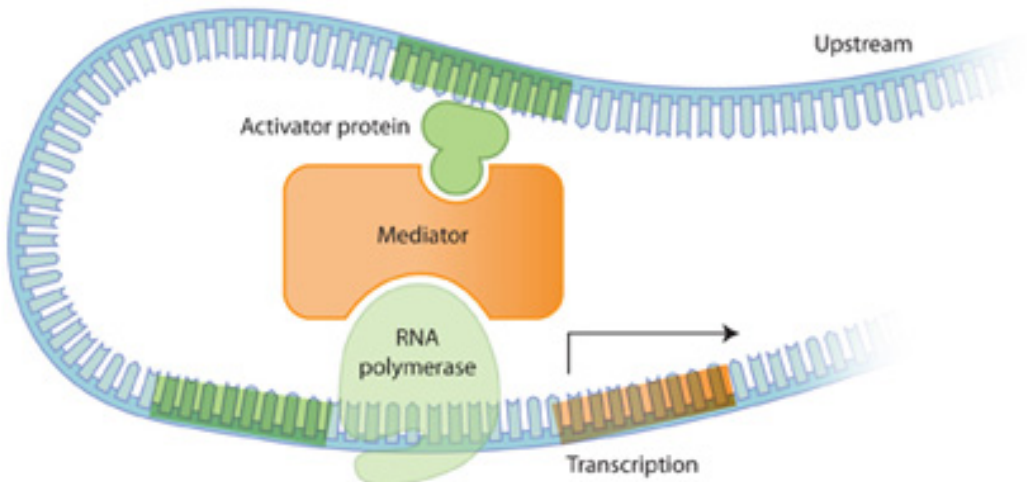
- DNA sequence → Gene expression → Protein function.



http://www.nature.com/scitable/topicpage/gene-expression-14121669

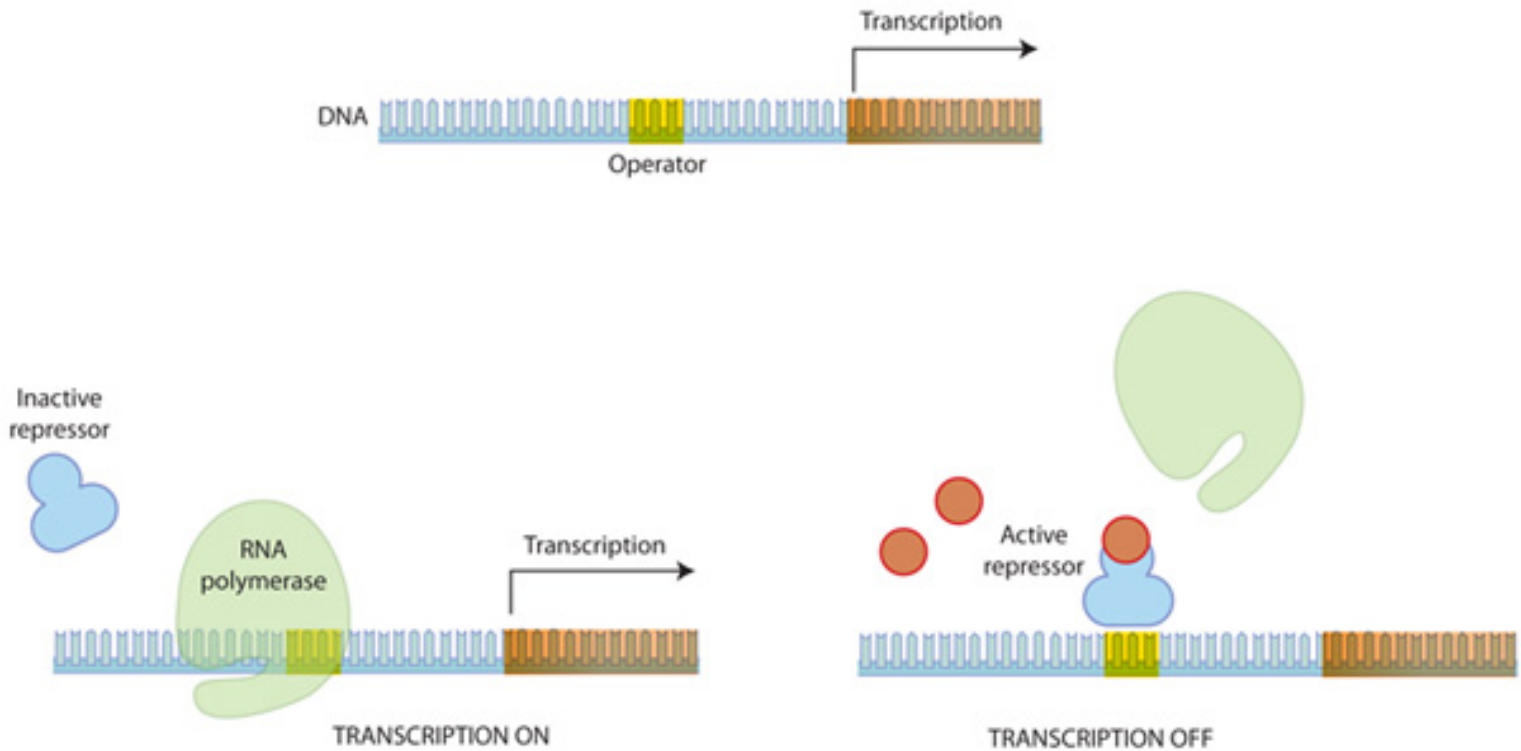
Short biological introduction





http://www.nature.com/scitable/topicpage/gene-expression-14121669

Short biological introduction



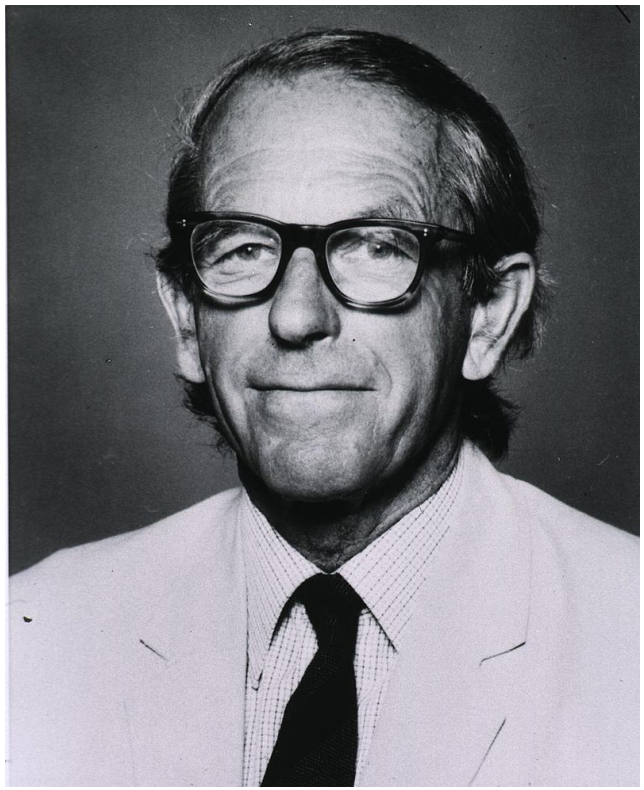
http://www.nature.com/scitable/topicpage/gene-expression-14121669

Sequencing methods

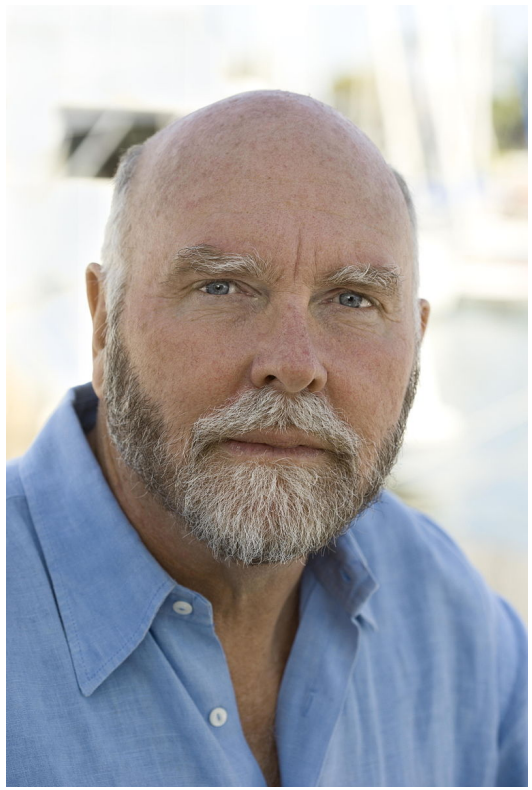
1. First generation sequencing: Sanger sequencing
2. Next (second) generation sequencing: 454, Solid, Illumina/Solexa
3. Third generation sequencing: IonTorrent, Single molecule sequencing

The focus of this course: NGS

Two big names in sequencing history



Frederick Sanger



Craig Venter

Sequencing history: the race for throughput and money

1958 Sanger's 1st Nobel prize for the sequence of insulin (51 bp)

1980 Sanger's 2nd Nobel prize for the dideoxy method of sequencing DNA

1990 The Human Genome Project plan for 15 yrs and $3 billion

1998 Craig Venter (Celera Genomics) plan for 3 yrs and $300 million

2001 HGP and Celera genome drafts published (3 billion bp; 3GB)

2003 HGP publishes the final sequence

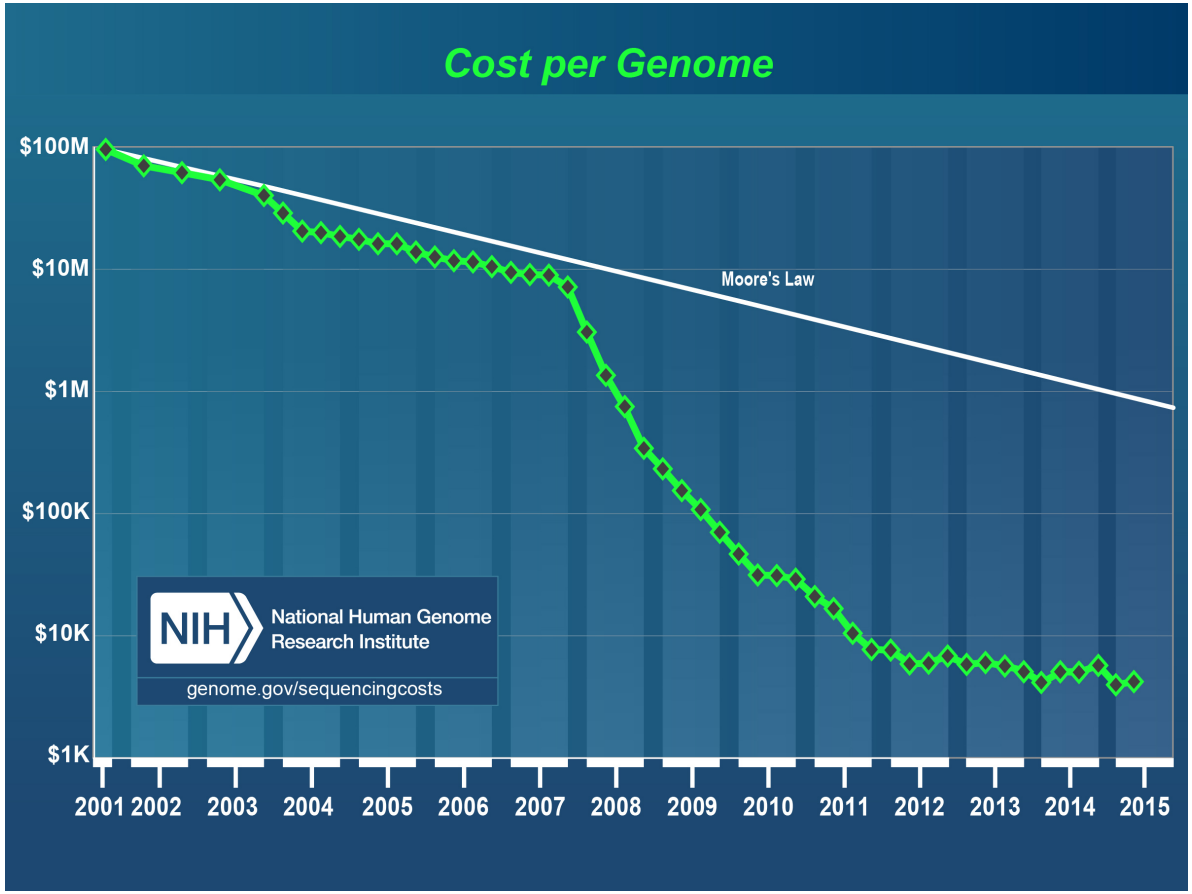
2008 1000 Genomes project launched

2012 1092 Human genome sequences published

2014 Illumina HiSeq X Ten Sequencer $1,000 genome

First individuals sequenced include Craig Venter, James Watson, Seong-Jin Kim, and Steve Jobs ($100K).

The ever dropping cost of sequencing



http://www.genome.gov/sequencingcosts/

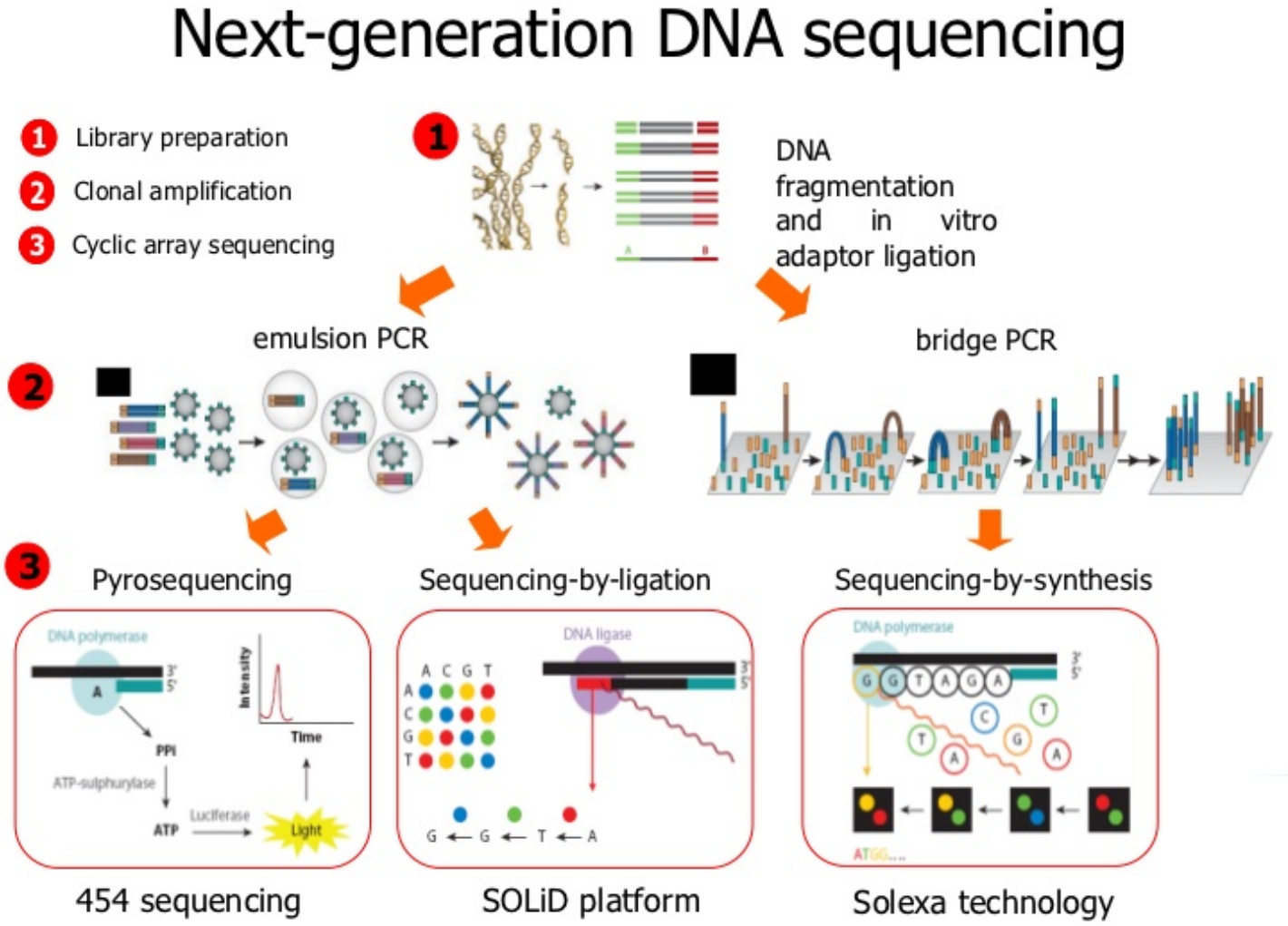
Applications of NGS

- *de novo* sequencing, e.g. genome assembly
- calling single nucleotide and structural variants in genomes
- personalized medicine, knowledge discovery, e.g. cancer genomics
- metagenomics
- ChIP-Seq - measuring expression control (TF binding)
- RNA-Seq - measuring gene expression
- DNASE-Seq - marking open chromatin regions
- CLIP- Seq, FAIRE-Seq, BiSulfite-Seq.. and many more

Plan of the course (but don't feel attached to it)

6.X.15	**01.** NGS, quality control
13.X.15	**02.** Genome assembly
20.X.15	**03.** Genome assembly
27.X.15	**04.** Read mapping
3.XI.15	**05.** Read mapping
10.XI.15	**06.** Variant calling
17.XI.15	**07.** Variant calling
24.XI.15	**08.** Cancer genomics
01.XII.15	**09.** Metagenomics
08.XII.15	**10.** ChIP-Seq
15.XII.15	**11.** RNA-Seq
22.XII.15	**12.** RNA-Seq
12.I.16	**13.** DNASE-Seq
19.I.16	**14.** Hi-C
26.I.16	**15.** Project presentations

How does NGS work?



How does NGS work?
A focus on Illumina (Solexa) sequencing.

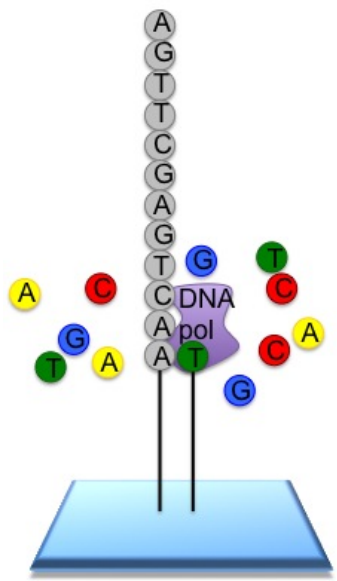
Preprocessing steps

1. Cleave the input sample into short fragments
2. Ligate the fragments to generic adaptors
3. Anneal the fragments to a slide using the adaptors
4. Amplify the reads with PCR $\rightarrow$ many copies of each read
5. Separate into single strands for sequencing

How does NGS work?
A focus on Illumina (Solexa) sequencing.

Sequencing steps

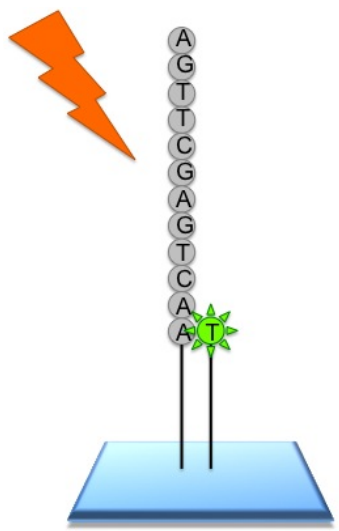
1. Flood the slide with nucleotides and DNA polymerase.
2. Nucleotides:
   - fluorescently labelled, with colour $\sim$ base.
   - have a terminator $\rightarrow$ only one base added at a time.



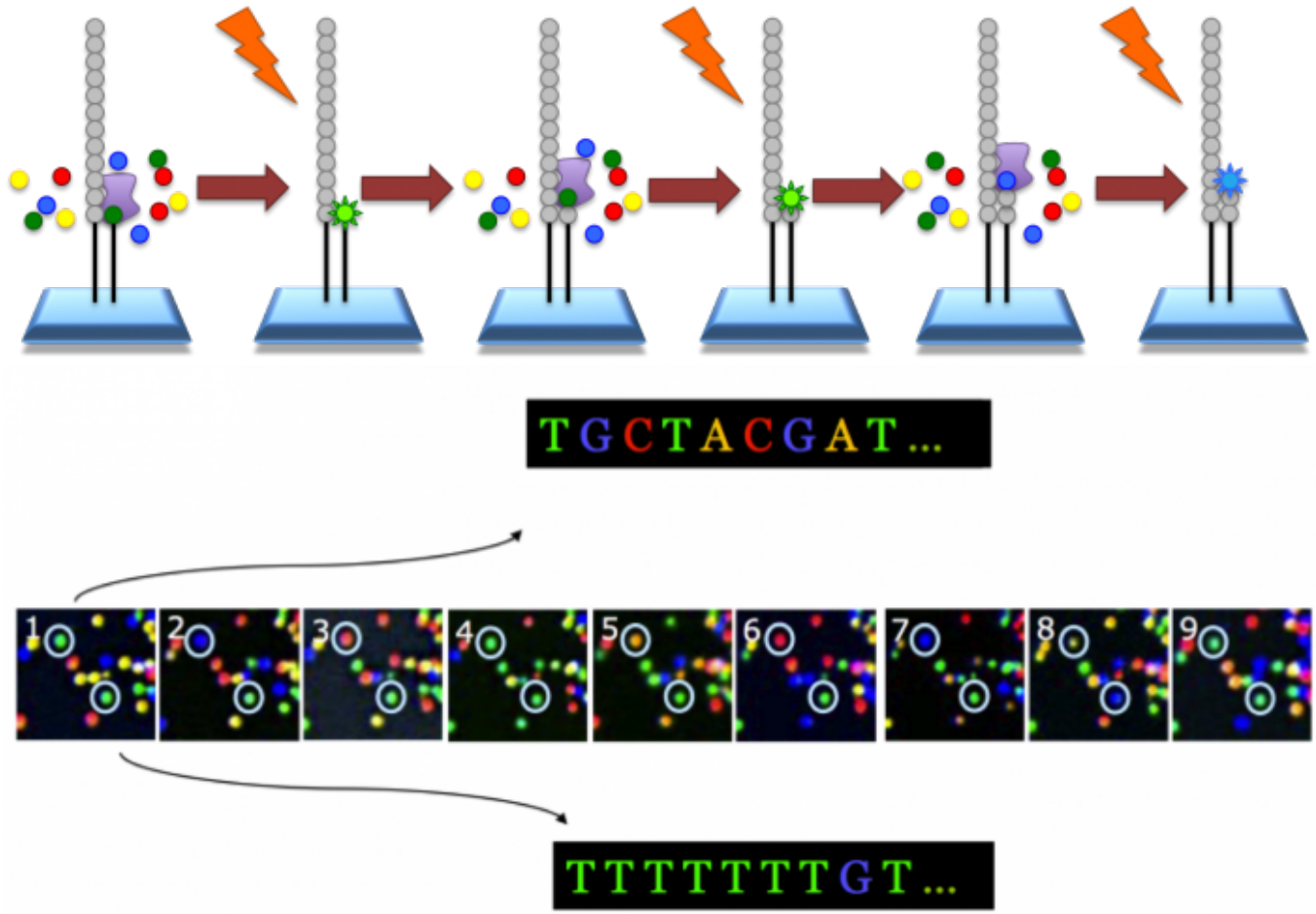
How does NGS work?
A focus on Illumina (Solexa) sequencing.

Sequencing steps cd

1. Take an image of each slide. Each read location $\leftrightarrow$ fluorescent signal $\leftrightarrow$ added base.
2. Remove the terminators, allowing the next base to be added
3. Remove the fluorescent signal, preventing contamination of the next image.



How does NGS work?
A focus on Illumina sequencing.



How does NGS work?

- We will watch a very simplistic video
  - `https://www.youtube.com/watch?v=-7GK1HXwCtE`
- To watch at home
  - `https://www.youtube.com/watch?v=jFCD8Q6qSTM`

Parameters of sequencing technologies

First generation sequencing: Sanger

- large read length ($700 - 1000bp$),
- large sequencing cost ($500\$/Mb$),
- high accuracy (read error rate 0.001%),
- low throughput.

Parameters of sequencing technologies

Table 1 Technical specifications of Next Generation Sequencing platforms utilised in this study

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	$128K	$80K**	$695K	$256K	$654K
Sequence yield per run	1.5-2Gb	20-50Mb on 314 chip, 100-200Mb on 316 chip, 1Gb on 318 chip	100Mb	30Gb	600Gb
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80%	1.71%	12.86%	0.76%	0.26%
Read length	upto 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	upto 150 bases	upto 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	upto 700 bases	upto 250 bases	upto 10kb	upto 700 bases	upto 700 bases
Typical DNA requirements	50-1000ng	100-1000ng	~1µg	50-1000ng	50-1000ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated
** System price including PGM, server, OneTouch and OneTouch ES
*** Includes two hours of cluster generation
**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter

(Quail *et al. BMC Genom.* '12)

Key concepts: coverage

Theoretical coverage

$c = \frac{L \times N}{G}$,

where L is the read length, N is the number of reads, and G is the length of the genome.

Sims *et al.* *Nature Reviews Genetics*, 2014

Key concepts: coverage

Theoretical coverage

$c = \frac{L \times N}{G}$,
where L is the read length, N is the number of reads, and G is the length of the genome.

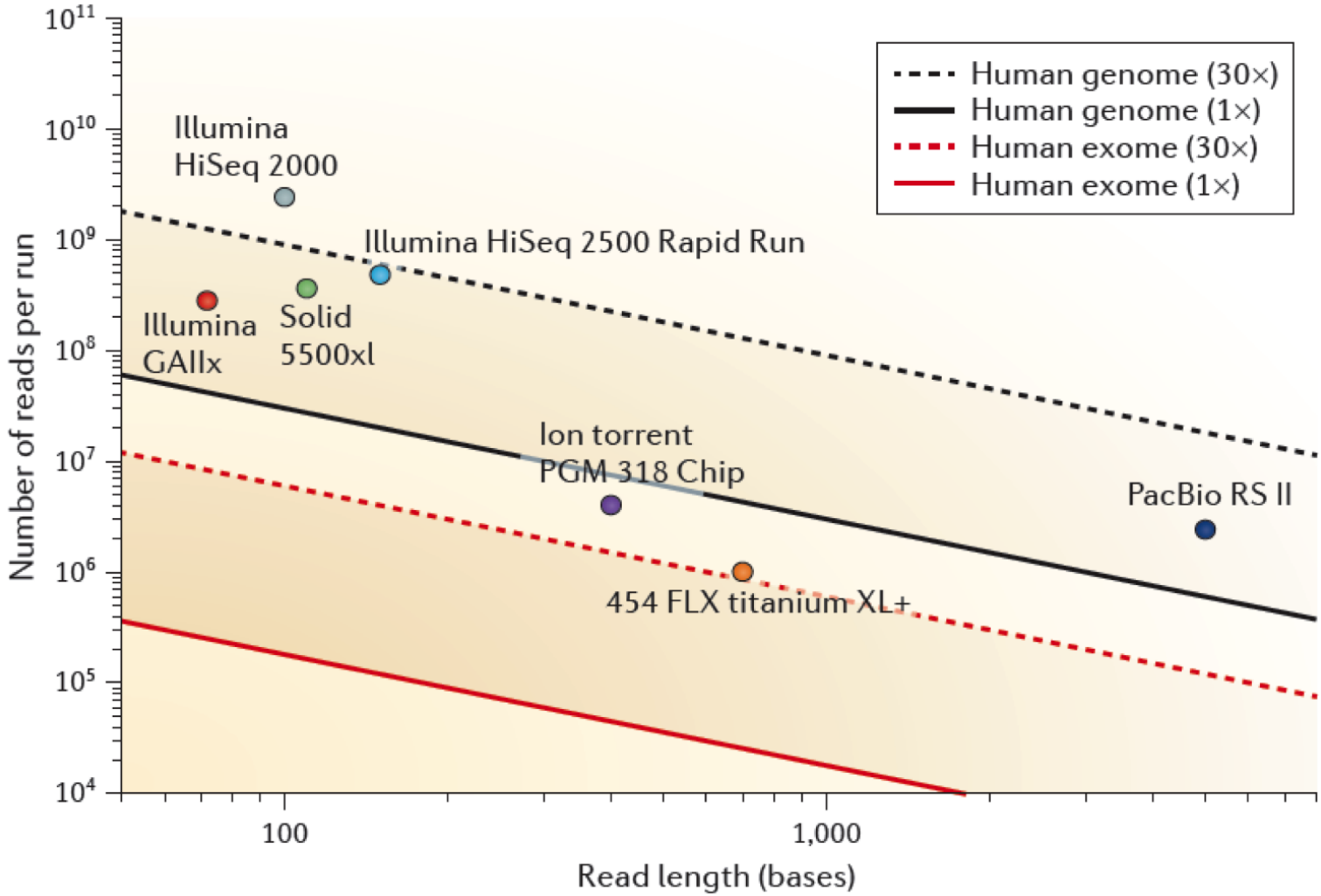
Empirical coverage

The exact number of times that a base in the reference is covered by a high-quality aligned read from a given sequencing experiment.

Both called simply coverage or depth and understood from the context.

Sims *et al.* *Nature Reviews Genetics*, 2014

Key concepts: coverage



Sims *et al.* *Nature Reviews Genetics*, 2014

Key concepts: single- and paired-end reads

Single-read sequencing

Sequencing DNA from only one end: simple, economical.

Paired-read sequencing

Sequencing both ends of a fragment: higher quality data. Facilitates detection of genomic rearrangements, repetitive sequence elements, gene fusions and novel transcripts.

The FastQ format

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Consecutive lines:

- Identifyier
- Sequence
- Identifyier
- Quality scores

Quality report

Phred score

- let P be the base-calling error probability,
- $Q = -10 \log_{10}(P) \Rightarrow P = 10^{-Q/10}$.

For example

- $Q = 10 \Rightarrow P = 0.1$
- $Q = 20 \Rightarrow P = 0.01$
- $Q = 30 \Rightarrow P = 0.001$
- $Q = 40 \Rightarrow P = 0.0001$

Phred scores reported by different platforms

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                             104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

https://en.wikipedia.org/wiki/FASTQ_format

Quality control

- FastQC
- Tracks errors both from the sequencer and of the material
- `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

Quality-based filtering

- Trimmomatic for Illumina data
- a variety of useful trimming tasks for illumina paired-end and single ended data.
- `http://www.usadellab.org/cms/?page=trimmomatic/`

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.

SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.

LEADING: Cut bases off the start of a read, if below a threshold quality

TRAILING: Cut bases off the end of a read, if below a threshold quality

CROP: Cut the read to a specified length

HEADCROP: Cut the specified number of bases from the start of the read

MINLEN: Drop the read if it is below a specified length

TOPHRED33: Convert quality scores to Phred-33

TOPHRED64: Convert quality scores to Phred-64

Acknowledgements

For input to these and the remaining slides of this course thanks to

- Norbert Dojer
- Jerzy Tiuryn
- Bartosz Wilczyński