INF5380/INF9380 NORBIS course
High-Performance Computing in Bioinformatics

Introduction to sequencing & mapping

Torbjørn Rognes
torognes@ifi.uio.no

Department of Informatics, UiO
16 March 2022

UiO : Universitetet i Oslo

DNA sequencing

High-Throughput Sequencing (HTS), Deep sequencing, Next Generation Sequencing (NGS)

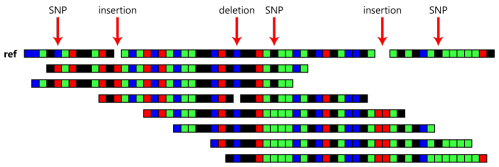
Obtain sample

DNA extraction, preparation

Sequencing

Raw data

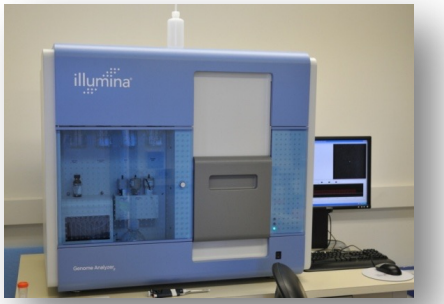
Sequence analysis



Interpretation

Illumina

- Sequencing by synthesis using fluorescence
- One fragment = one cluster = one read
- Read lengths up to 250bp, paired-end reads
- Dominant technology today
- Formerly known as Solexa
- HiSeq 2500 specifications:
  - Can sequence entire human genome in 27 hours at 30X coverage (2x100bp)
  - Up to 2x150 bp
  - Run time 7 hours to 11 days
  - Up to 6 billion 100bp reads in 11 days



GA IIx

HiSeq 2500

Sanger sequencing center

MiSeq

NovaSeq 6000

Other sequencing technologies

Roche (454)

ABI (SOLiD)

Ion Torrent

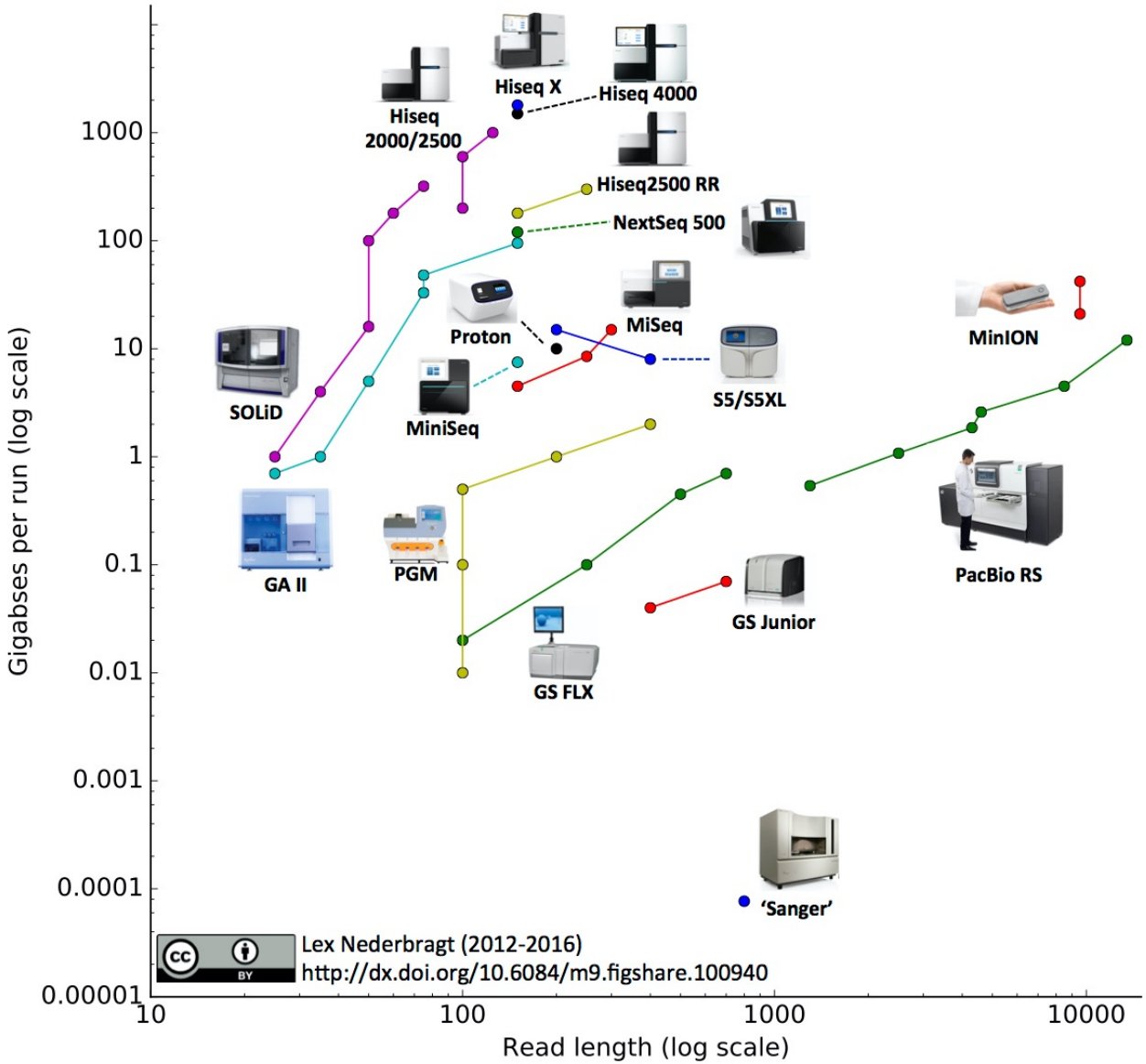
Pacific Biosciences SMRT and Sequel systems

Oxford Nanopore

Important technology properties

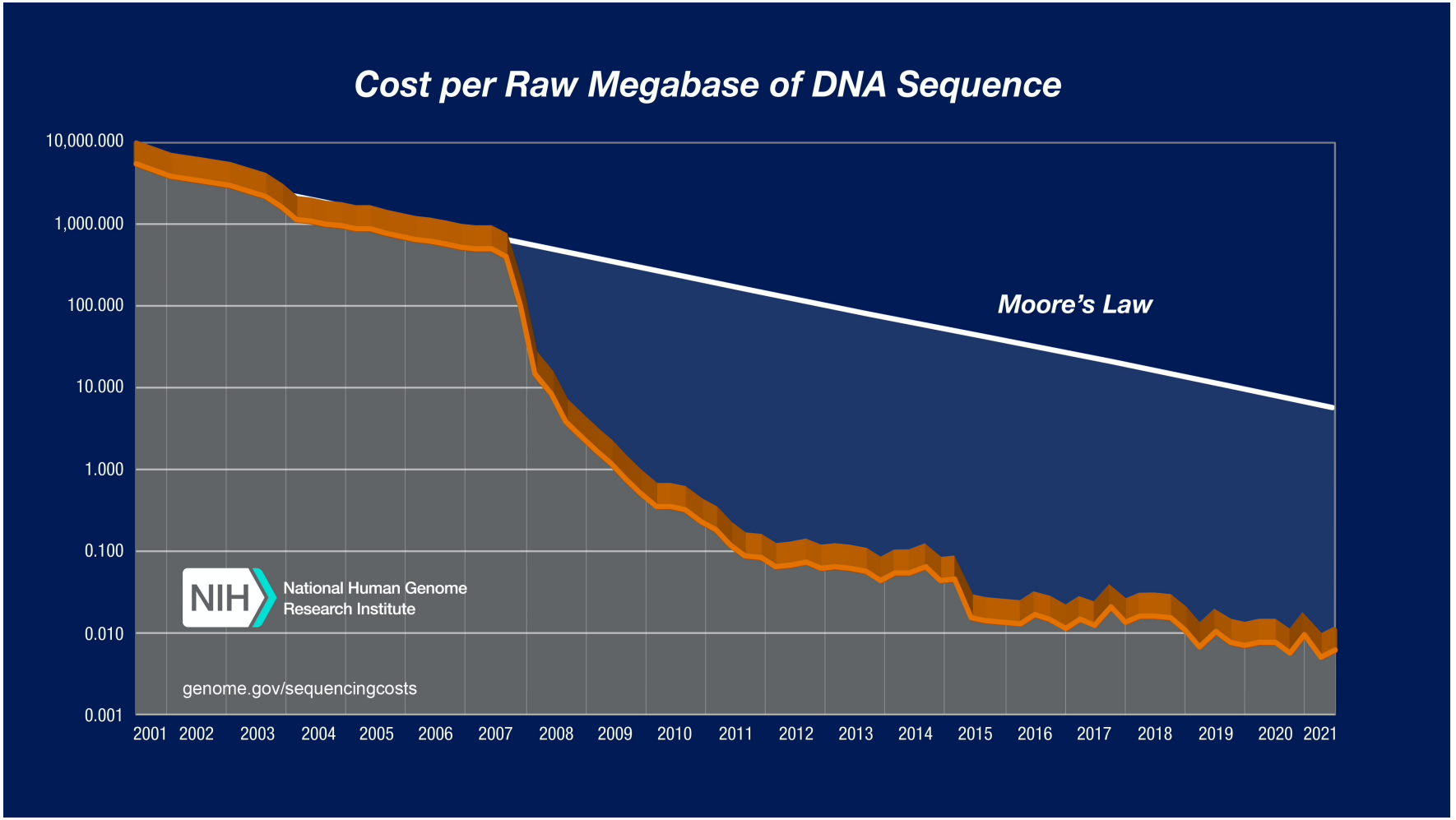
- Cost
  - Per base
  - Investment
- Read length
- Speed / capacity (bases per day)
- Errors
  - Frequency
  - Profile (indels, substitutions)
  - Random or systematic?
- Paired-end support
- PCR-based?
  - Single molecule
  - PCR amplification step
- Amount of lab work necessary

Sequencing technology development



Source: Lex Nederbragt (2012-2016) https://doi.org/10.6084/m9.figshare.100940

The cost of sequencing



The FASTQ format

- A sequence file format in plain text that includes quality scores for each nucleotide in the sequence
- Example:

```
@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
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- The first line starts with a '@' symbol followed by an identifier before the first space.
- The second line contains the actual sequence.
- The third line starts with a '+' symbol, optionally followed by the same identifier as the first line. Identifier rarely used.
- The fourth line contains characters that represent the quality scores for each nucleotide
- In principle, sequence and quality may span multiple lines, but rarely do.

FASTQ quality ranges

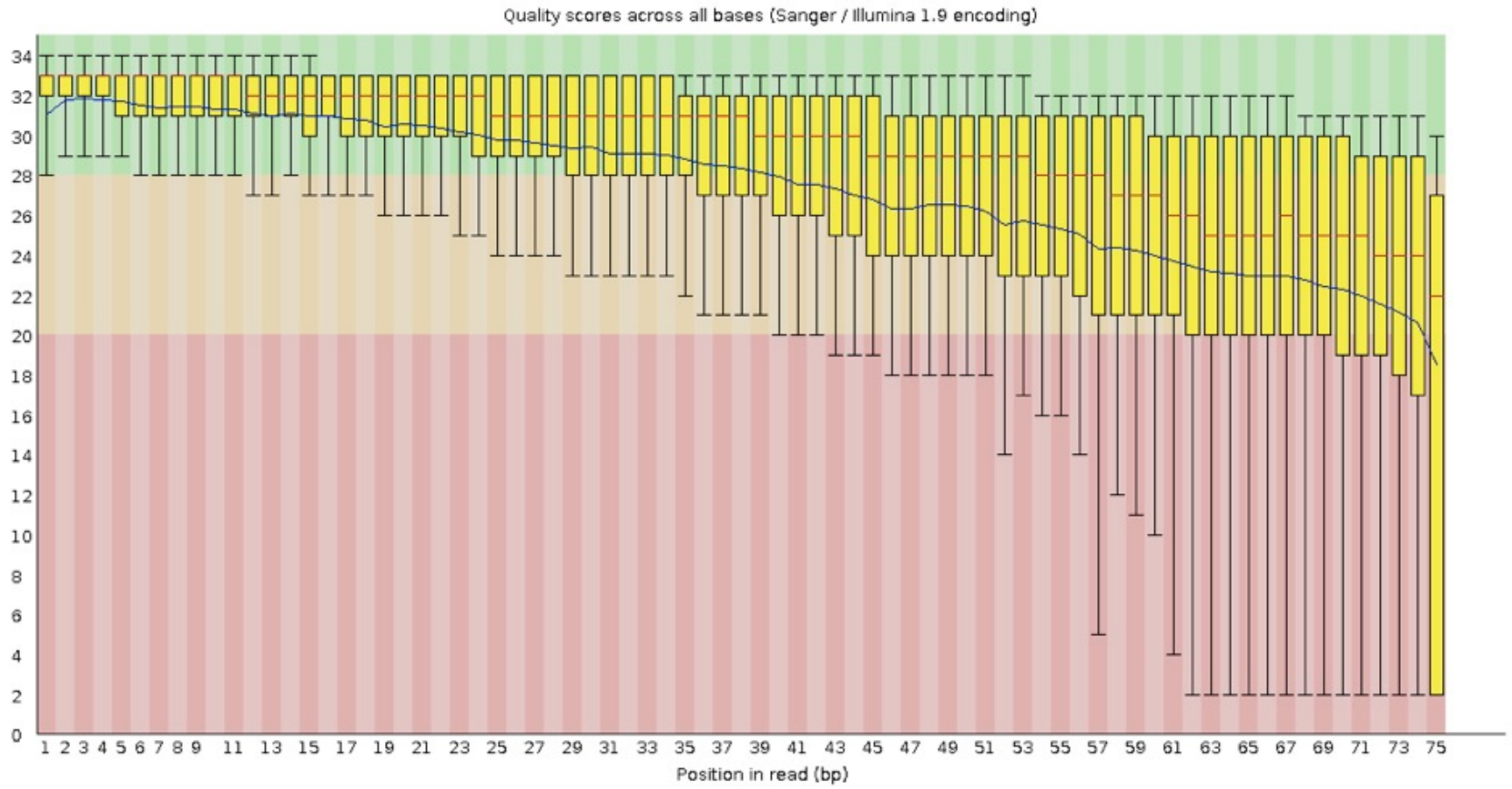
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2......................26...31........41
0..................20........30........40........50..........................................93

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, 41)
    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)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

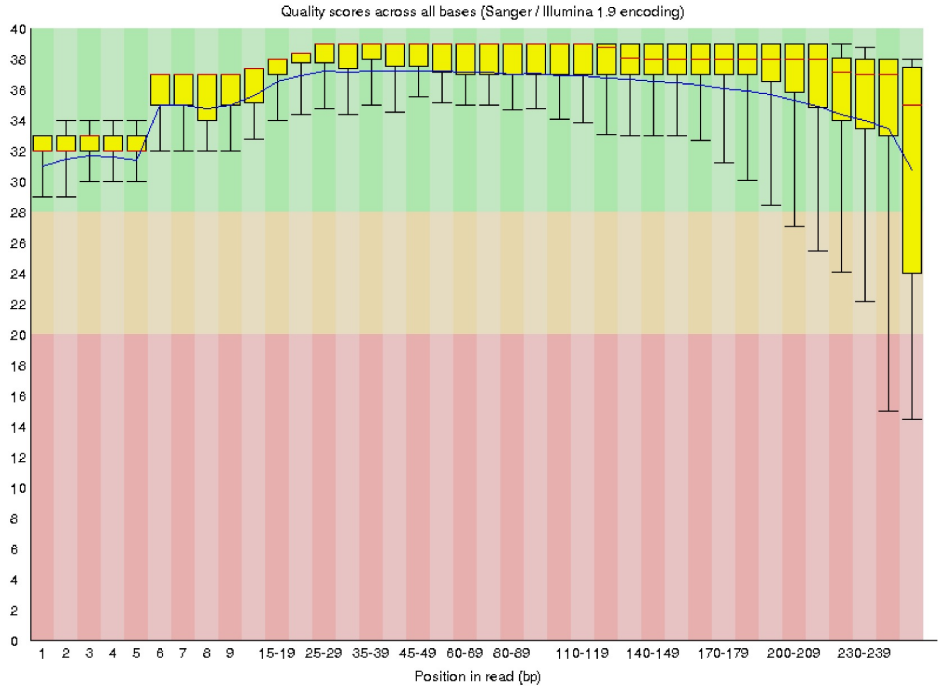
Source: Wikipedia – FASTQ_format

FASTQC - Quality plot of Illumina reads

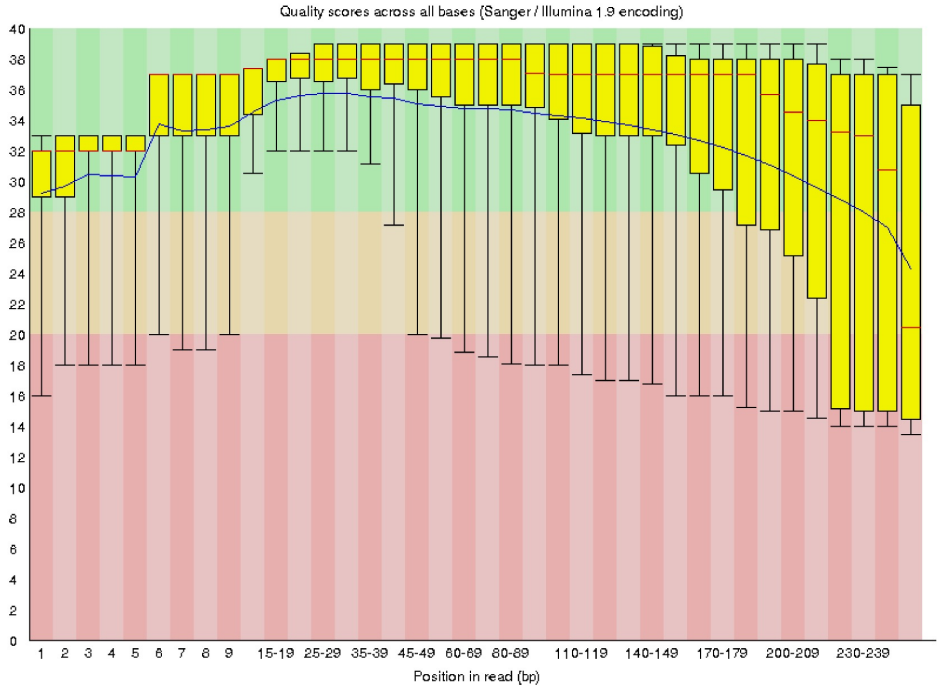
Per base sequence quality



Quality plots of Illumina MiSeq reads



Forward reads



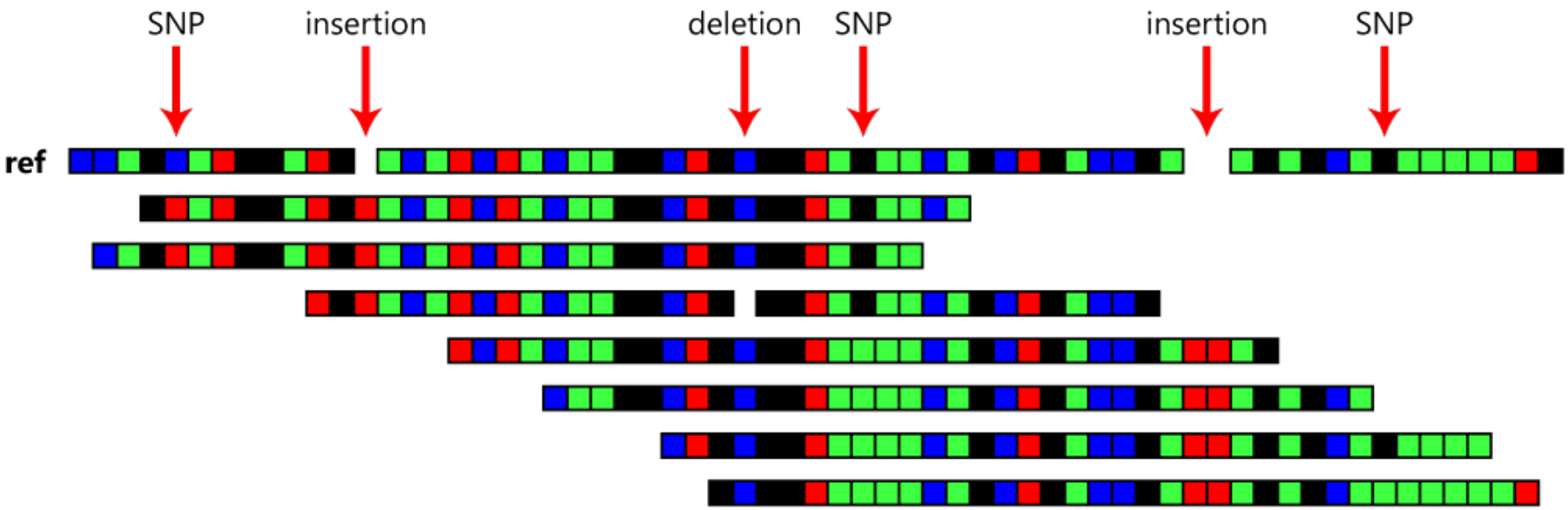
Reverse reads

Common HTS applications

***De novo* genome sequencing**	Determining the complete genome sequence of an organism for the first time
Whole genome re-sequencing	Finding polymorphisms (SNPs) and discover mutations in an individual
Exome sequencing	Sequencing only protein-coding regions of a genome from an individual to identify mutations or polymorphisms (SNPs)
Transcriptomics (RNA-seq)	Sequencing of expressed RNA (after reverse transcription to cDNA), (small RNA, mRNA or total RNA) to determine level
Chromatin immunoprecipitation-sequencing (ChIP-Seq) (ChIP-exo)	Mapping of genome-wide protein-DNA interactions
Methylation sequencing (Methyl-Seq)	Determining methylation patterns in the genome (epigenomics) (often on bisulfite-treated DNA)
Metagenomics	Sequencing genomic DNA of multiple species (microorganisms) simultaneously from a certain environment
Metatranscriptomics	Sequencing RNA from multiple species (microorganisms) simultaneously
Amplicon sequencing	Sequencing of genomic regions selected and amplified by PCR, often from multiple species simultaneously

Resequencing

- Sequencing DNA from a new individual when we already have a reference genome sequence
- Map reads to reference genome instead of assembly



Variation detection by resequencing

- Natural variation discovery
- Mutation detection

- Single Nucleotide Polymorphisms (SNPs)
- Small insertions & deletions (Indels)
- Copy Number Variation (CNV)
- Large inversions, translocations etc

- Requires high coverage, that is, the average number of times each base is sequenced (typically 40X, but may require 100X)

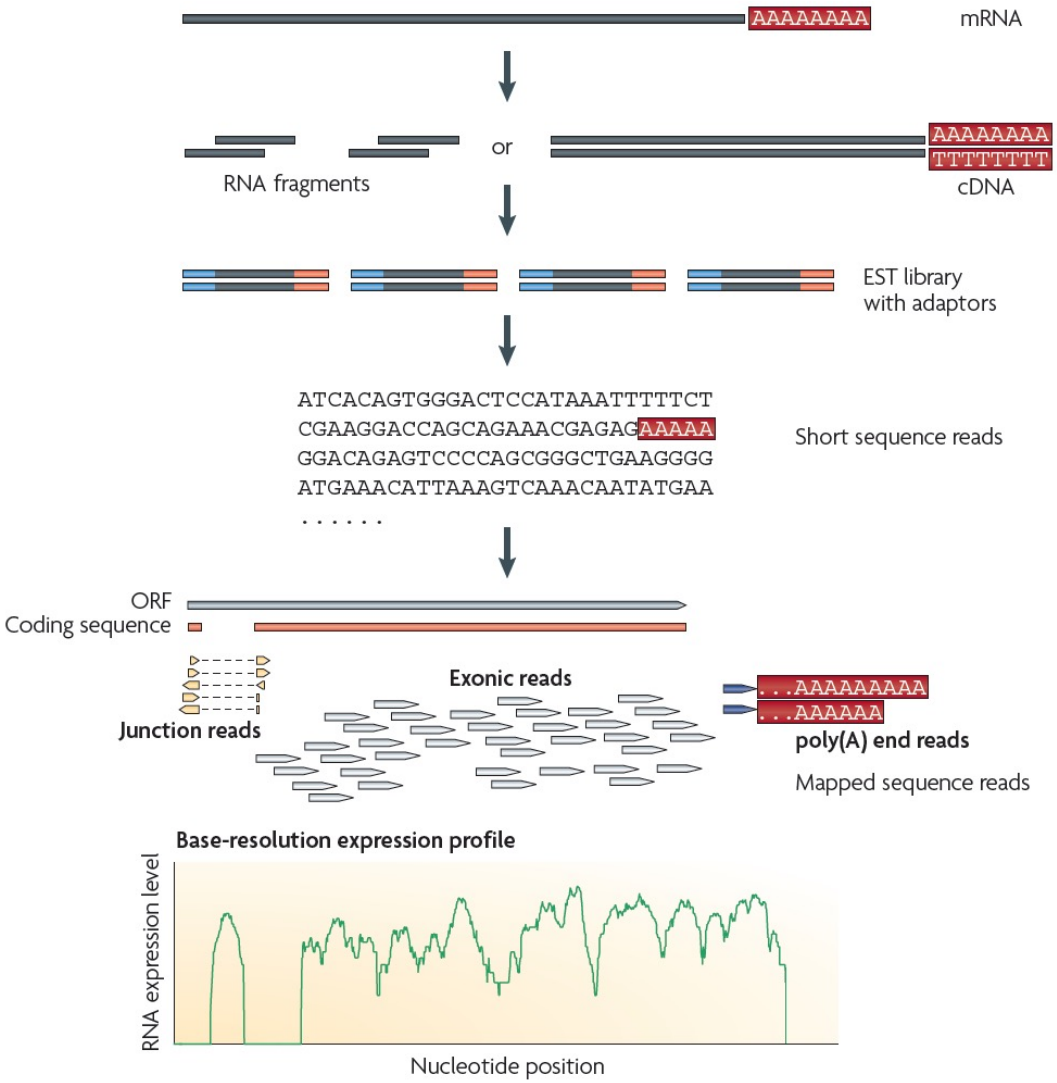
Gene expression (RNA-Seq)

- Gene expression analysis
- “transcriptomics”
- Replaces microarrays

- mRNAs
- Small RNAs (miRNA, piRNA…)

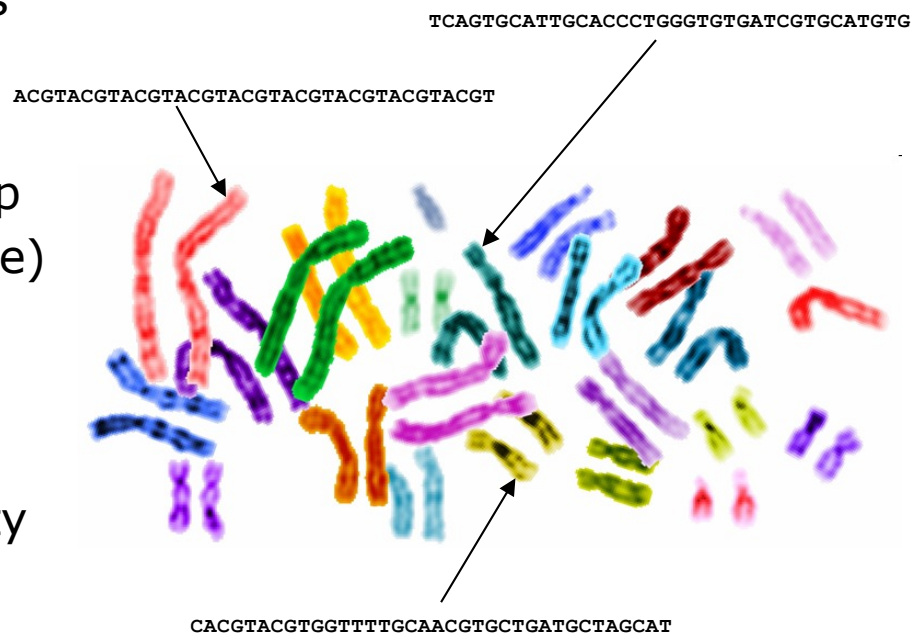
- Splice variants

- Counts the number of reads for each RNA



Mapping reads to a reference genome

Goal: Identify positions in the genome that are most similar to the sequence reads



Input data:

- 10-10000 million reads, each 30-300bp
- Sequencing errors (typ. ~1% error rate)

Reference genome:

- E.g. human genome, 3 Gbp
- Some genome variation, heterozygocity

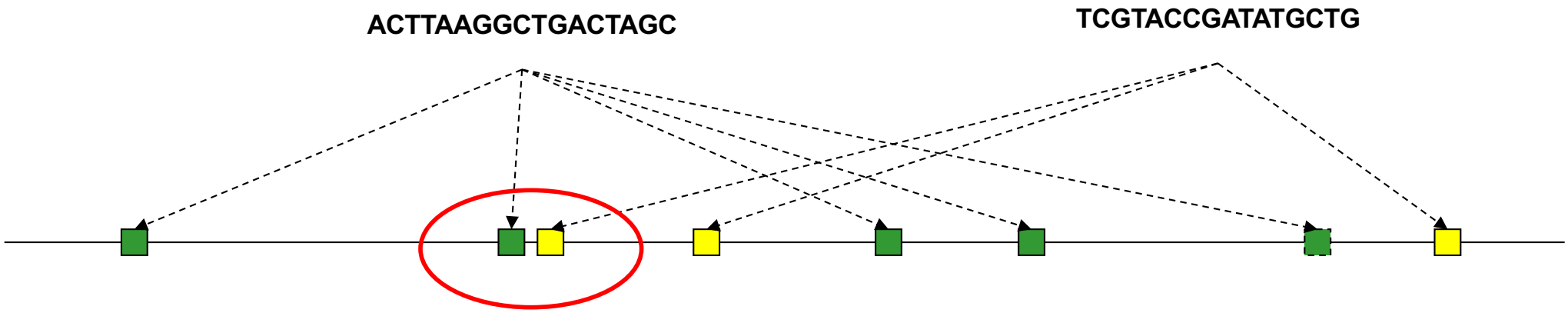
Output:

- 0, 1, or more potential genomic locations for each read
- Mapping quality assignment

Requirements:

- Sensitivity, specificity, speed, compactness

Multiple mapping



- Problem:
  - Short reads (e.g. 100bp) may not map uniquely due to long repeats (>100bp) in the genome
- Solutions:
  - Get longer reads
  - Get paired reads (pairs of reads with fixed distance)

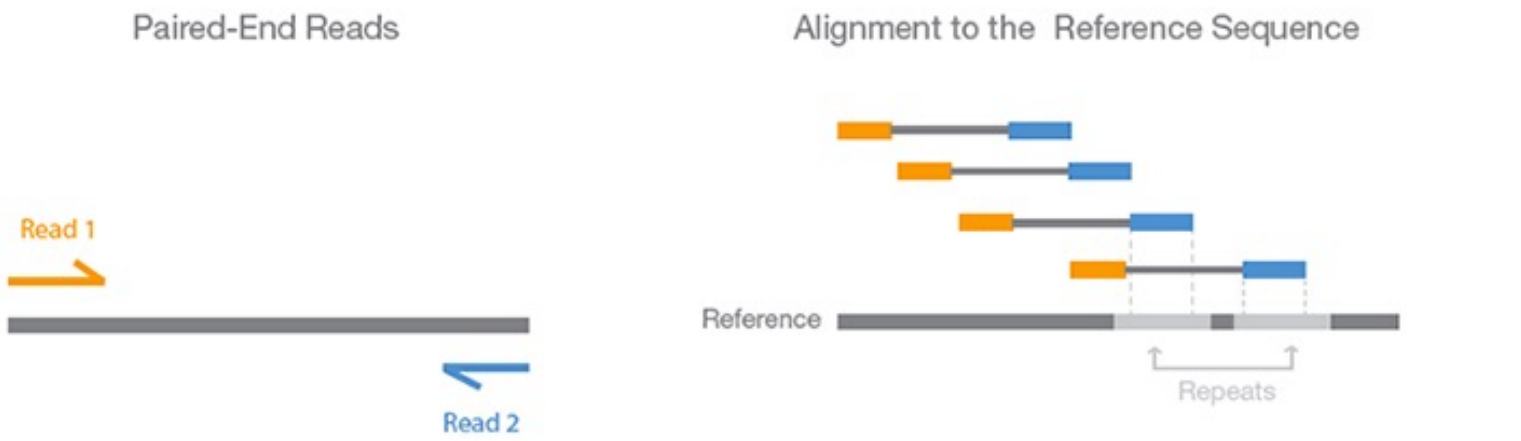
Paired-end / mate-pair sequencing

- Paired-end reads or mate-pair reads are pairs of reads known to come from two close regions in the genome.
- They are located with an approximate fixed distance from each other.
- Typically paired ends are a ~100-500bp apart, while mate pairs are ~2-10kb apart
- Allows short reads to have a larger ”effective” size
- Performed by sequencing fragments from both ends
- Often used with Illumina reads
  - Typically 2 x 100 bp separated by 300bp
  - Allpaths-LG requires ~100 bp from fragments ~ 180bp
- May also overlap (e.g. 2x250bp from 400bp fragments)

Paired-end / mate-pair sequencing

- Both ends of fragments of fixed length (a few kbp) may be sequenced
- Gives information on genomic distance between pairs of reads
- May be used to overcome some problems with short reads

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Source: Illumina

Mapping tools

- Methods based on hashing or indexing of seed sequences (k-mers)
  - Continuous seeds
    - RMAP, Novoalign, Mosaik
  - Spaced seeds
    - Eland, SOAP, Maq, SeqMap, BFAST, ZOOM, SHRiMP, RazerS

- Methods based on Burrows-Wheeler Transform (BWT)
  - **Bowtie**(2), **BWA**(-MEM), SOAP2, BWT

- Methods based on *minimizers*
  - **minimap2**

Mapping methods

- Approach:
  - Initial rapid filtering using some form of indexing
  - Further examination of a small amount of interesting positions
- Index based on
  - Reference genome
  - Reads
- Heuristics:
  - Only a small part of the read is often used initially (the most reliable part of the sequence from Illumina sequencers)
  - Only 2 mismatches (and 1 gap) may be allowed in this region
- Quality of reads:
  - Some nucleotides in the reads are more certain than other
  - The probability of each called base being erroneous is indicated
  - Some programs take the quality values of the read into account
  - Allows mismatches at low quality positions

Indexing

- Indexing uses a table with all possible words of w symbols and their positions in the genome
- All positions in the genome where each word is found can be looked up directly
- Given alphabet size S, genome size N and word length w, it requires $(S^w + N) * \log_2 N$ bits
- Typical index width is 11-19 nucleotides
- Human genome, w=14:
  $(4^{14} + 3*10^9) * 4$ bytes
  = 1GB + 12GB = 13GB
- Example with string “ANANAS” and index width 2

Index key	Positions
AA	
AN	1,3
AS	5
NA	2,4
NN	
NS	
SA	
SN	
SS	

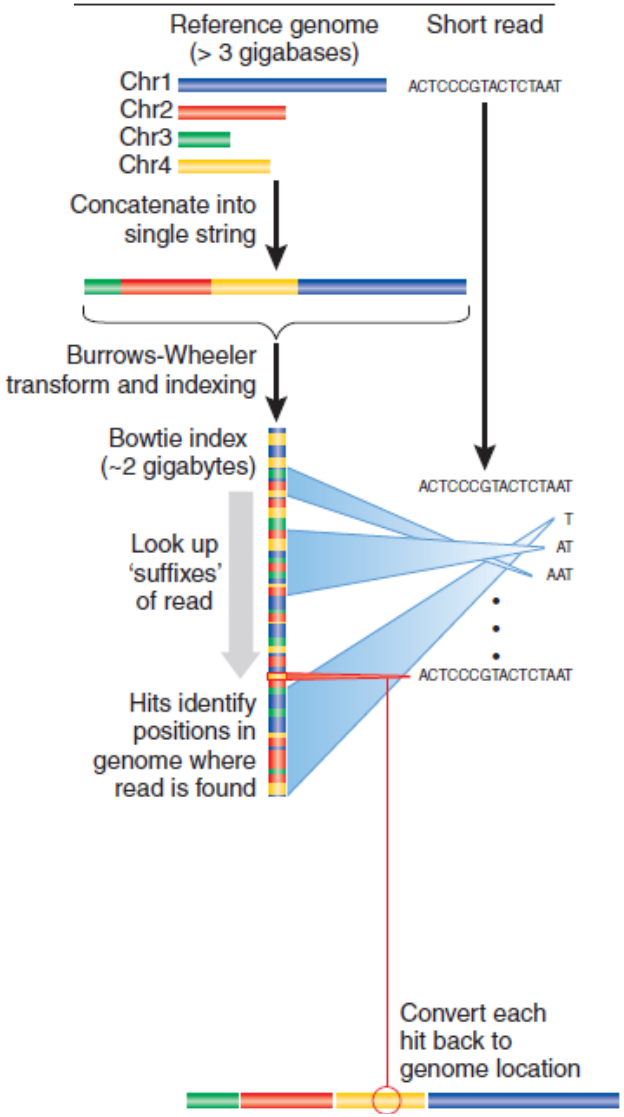
Single seed mapping with *k* mismatches

- Divide the read into k+1 seeds
- Look up each seed in the index
- At least one seed must match
- Example below with k=2 and 36bp long reads
- Guaranteed to find all hits with up to k mismatches

Burrows-Wheeler Transform (BWT)

- A kind of sequence index that allows a program to very rapidly find all perfect matches to given string
- A simple and elegant reversible transformation of a string
- Used in combination with a suffix array and a few other arrays, known together as the FM-index
- Can be stored very compactly. The human genome can be stored in about 1.5GB.
- May be well compressed by storing only parts of the arrays, e.g. only 4 bit / nucleotide
- Takes some time to generate index from the human genome (but can be used over and over), but searching is extremely fast

BWT alignment: Bowtie / BWA



- The reference genome is indexed using BWT
- For each read
  - Initially, only a part of the read is used
  - Exact matches:
    - BWT *backward search*
  - Inexact matches:
    - Introduce substitutions or gaps at all possible positions in the read and then look for exact matches
    - To save time, the search tree may be pruned
- Multiple matches are identified
- Report matches to each read

Thanks!