Introduction to Next Generation Sequencing

Shamith Samarajiwa

CRUK Summer School in Bioinformatics
Cambridge, September 2018

UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

Where to get help!

SEQanswers
the next generation sequencing community

http://seqanswers.com

http://www.biostars.org

Biostar
BIOINFORMATICS EXPLAINED

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

http://www.bioconductor.org/help/mailing-list

Read the posting guide before sending email!

Overview

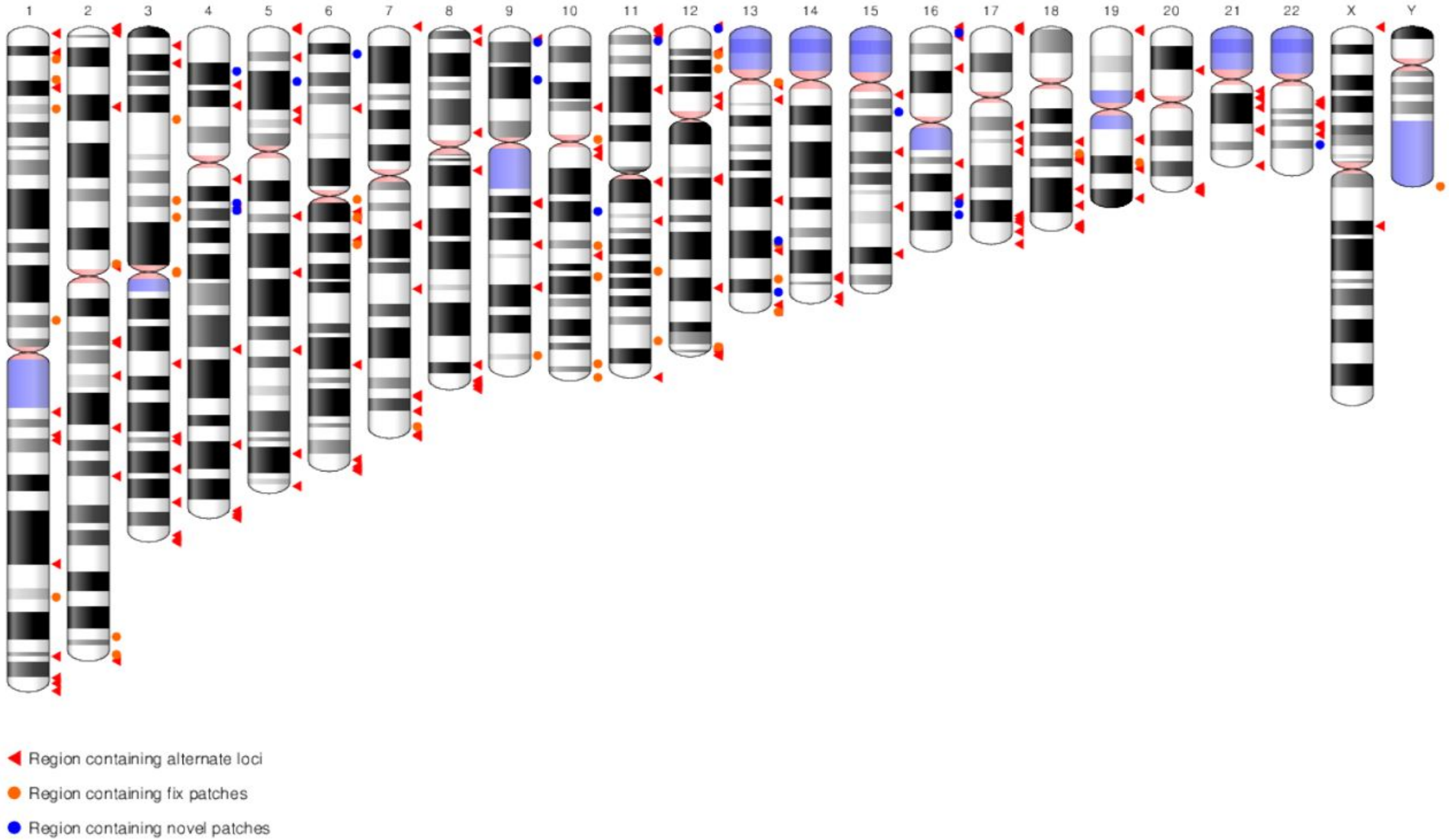
- Understand the difference between reference genome builds
- Introduction to Illumina sequencing
- Short read aligners
  - BWA
  - Bowtie2
  - STAR
  - Other aligners
- Genomic Coverage and Depth
- Mappability
- Use of decoy and sponge databases
- Alignment Quality, SAMStat, Qualimap
- Samtools and Picard,
- Visualization of alignment data (IGV)
- A very brief look at long reads, graph genome aligners and *de novo* genome assembly

Reference Genomes

- A haploid representation of a species genome.
- The human genome is a haploid mosaic derived from 13 volunteer donors from Buffalo, NY. USA.
- In regions where there is known large scale population variation, sets of alternate loci (178 in GRCh38) are assembled alongside the reference locus.
- The current build has around 500 gaps, whereas the first version had ~150,000 gaps.
- Allelic diversity and structural variation present challenges.

Genome Reference Consortium: https://www.ncbi.nlm.nih.gov/grc

GRCh 38

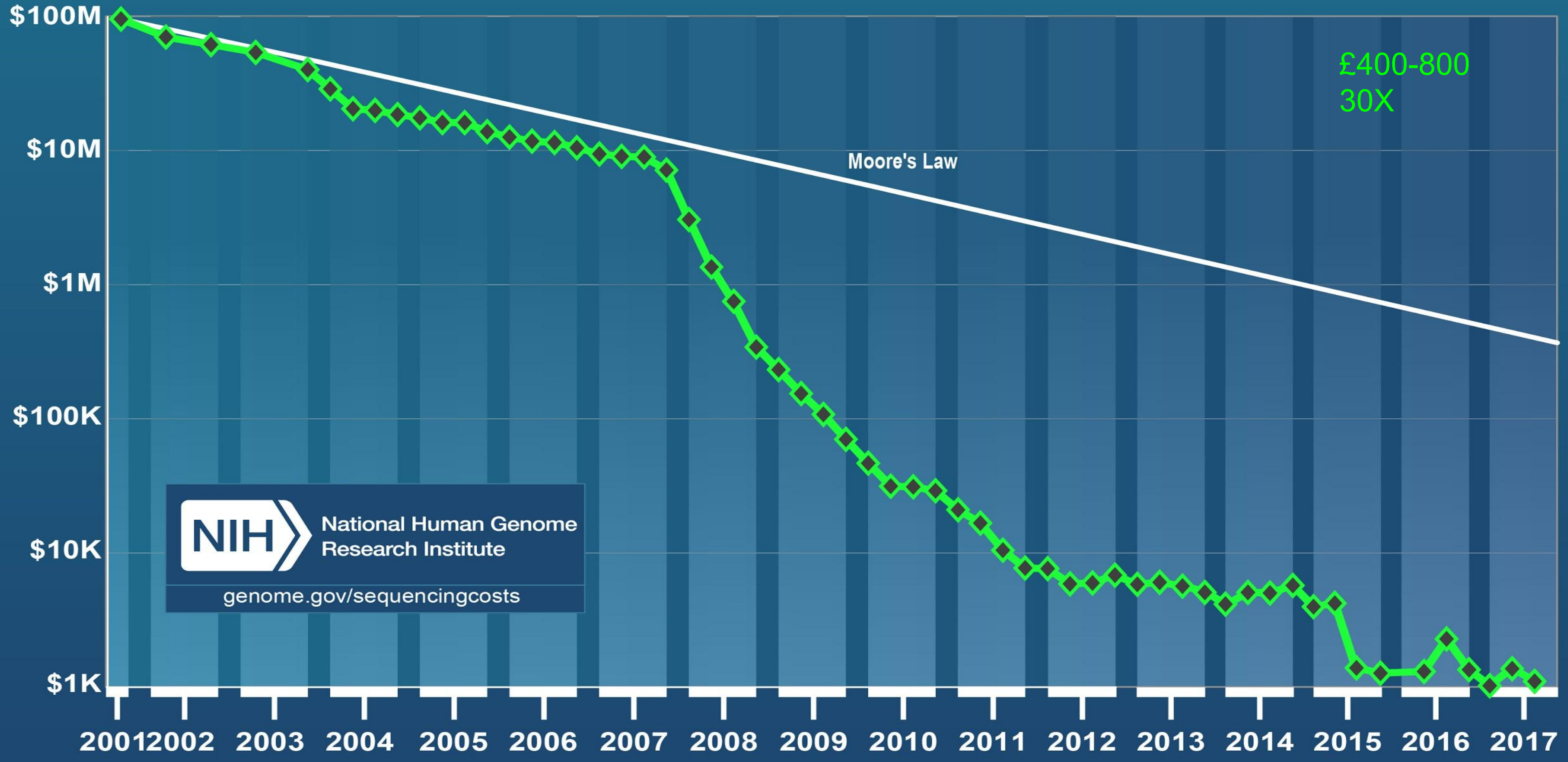


◀ Region containing alternate loci

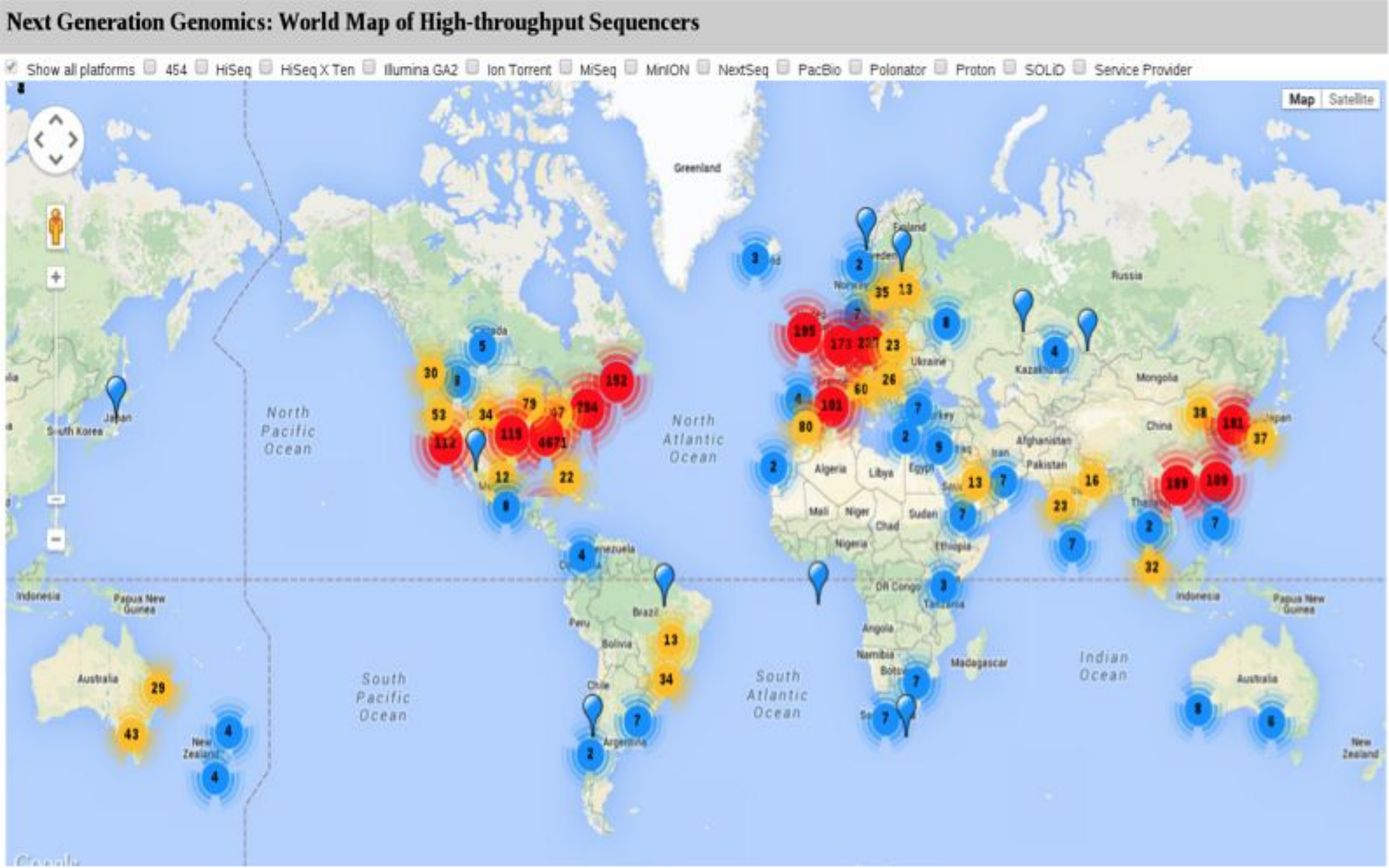
● Region containing fix patches

● Region containing novel patches

Cost per Genome



£400-800
30X



Next Generation Genomics: World Map of High-throughput Sequencers

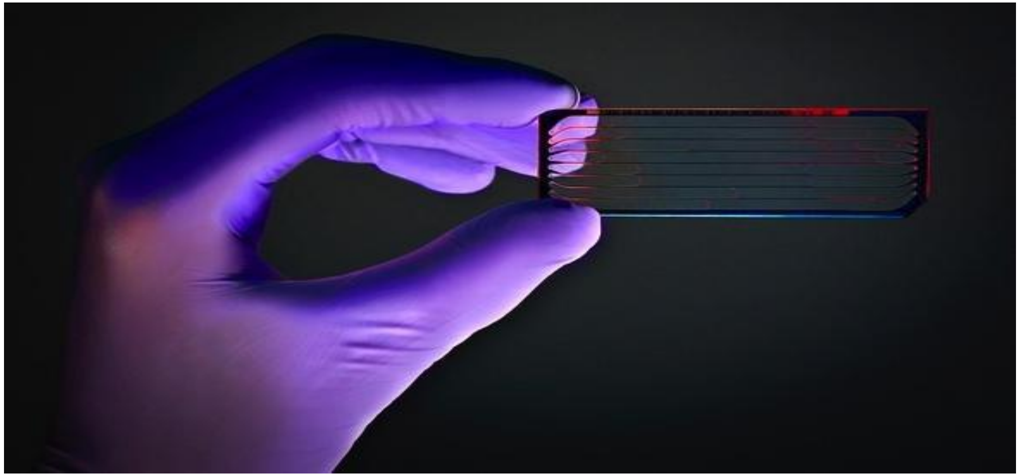
☑ Show all platforms □ 454 □ HiSeq □ HiSeq X Ten □ Illumina GA2 □ Ion Torrent □ MiSeq □ MinION □ NextSeq □ PacBio □ Polonator □ Proton □ SOLiD □ Service Provider

	 NextSeq[*†]	 HiSeq 4000[*]	 NovaSeq 6000[*††]
Output Range	20–120 Gb	125–1500 Gb	134–6000 Gb
Run Time	11–29 hr	< 1–3.5 days	13–44 hr
Reads per Run	130–400 million	2.5–5 billion	Up to 20 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 150 bp
Samples per Run[‡]	2–8	50–100	26–400
Relative Price per Sample[‡]	Higher Cost	Mid Cost	Lower Cost
Relative Instrument Price[‡]	Lower Cost	Mid Cost	Higher Cost

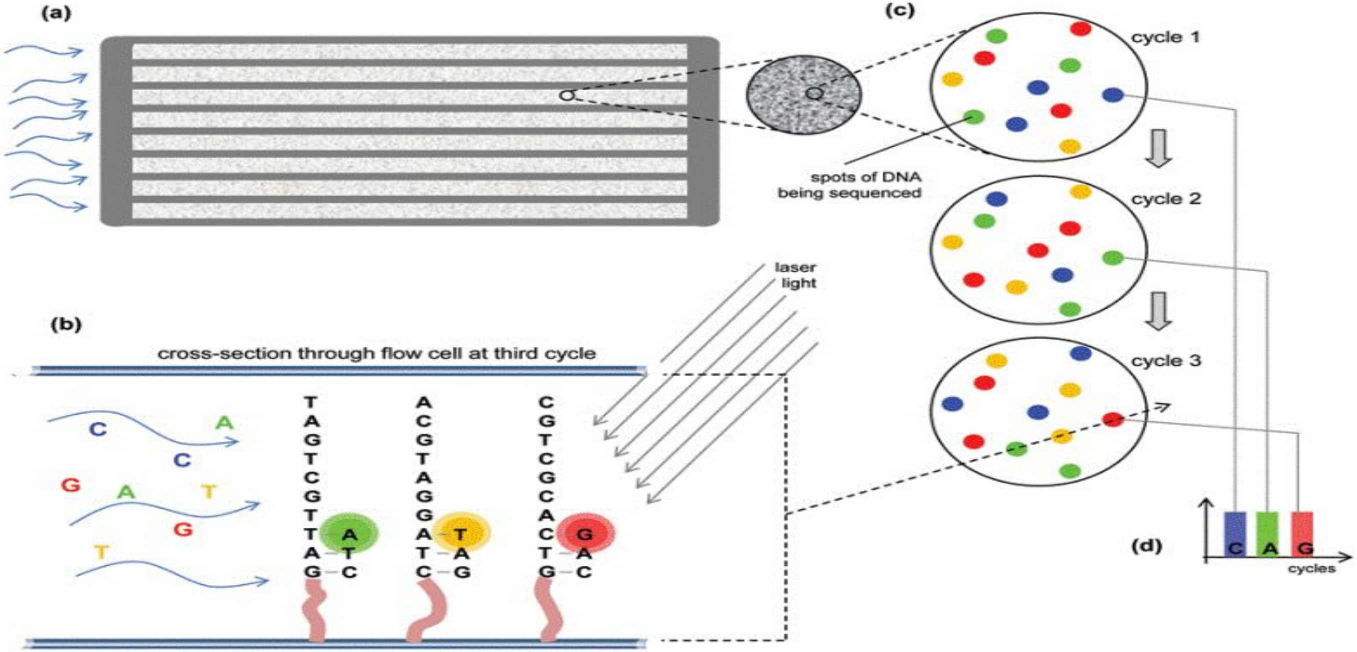
Illumina Genome Analyzer

Illumina sequencing technology

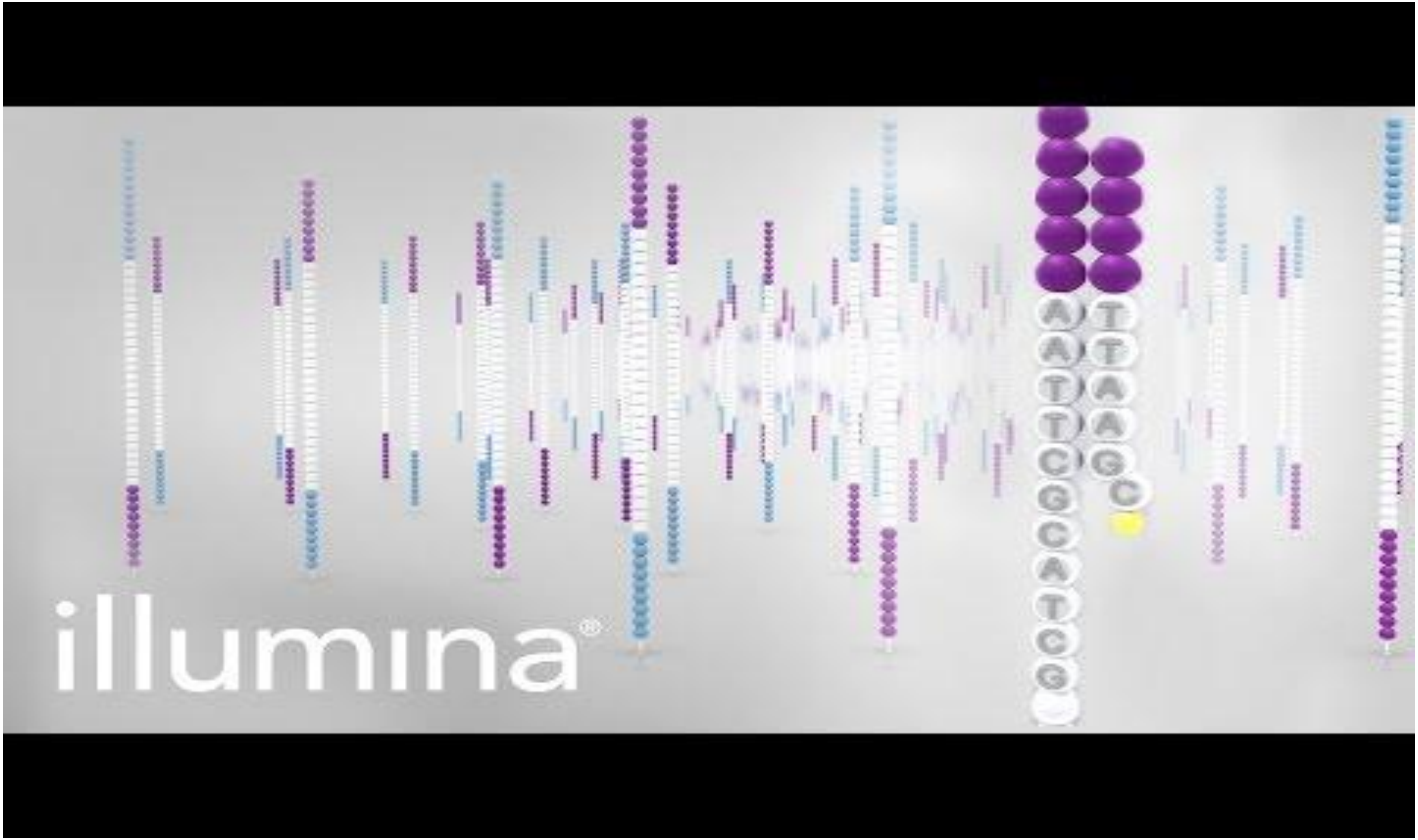
- Illumina sequencing is based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** (1998) at the University of Cambridge.
- Multiple steps in “Sequencing by synthesis” (explained in next slide)
  - Library Preparation
  - Bridge amplification and Cluster generation
  - Sequencing using reversible terminators
  - Image acquisition and Fastq generation
  - *Alignment and data analysis*



Illumina Flow-cell

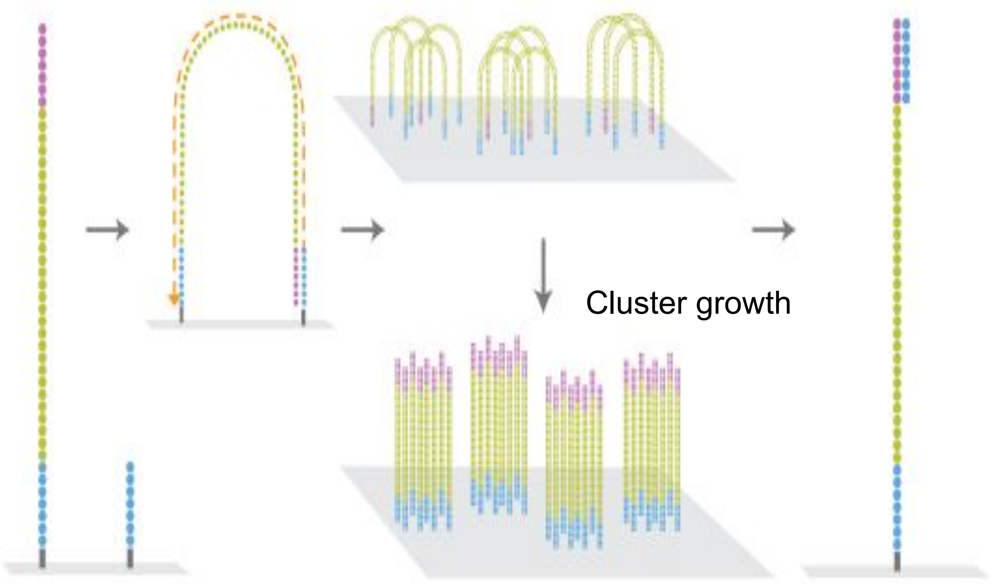


Sequencing by Synthesis technology



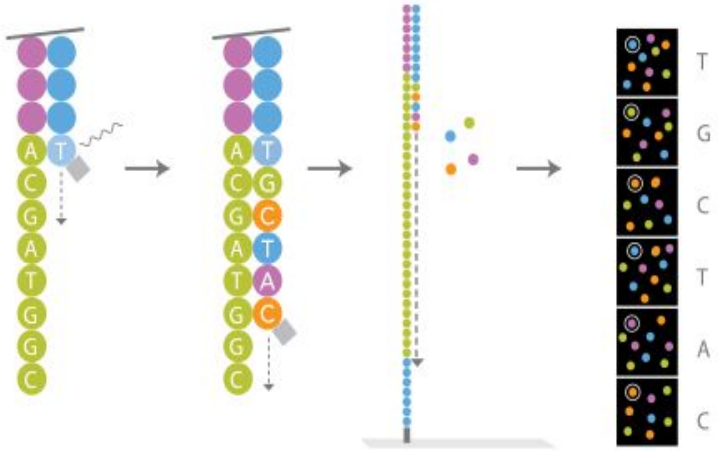
Illumina Sequencing

Bridge Amplification



Cluster growth

Sequencing



Incorporation of fluorescence,
reversibly terminated tagged nt

Multiplexing

- Multiplexing gives the ability to sequence multiple samples at the same time.
- Useful when sequencing small genomes or specific genomic regions.
- Different barcode adaptors are ligated to different samples.
- Reads de-multiplexed after sequencing.

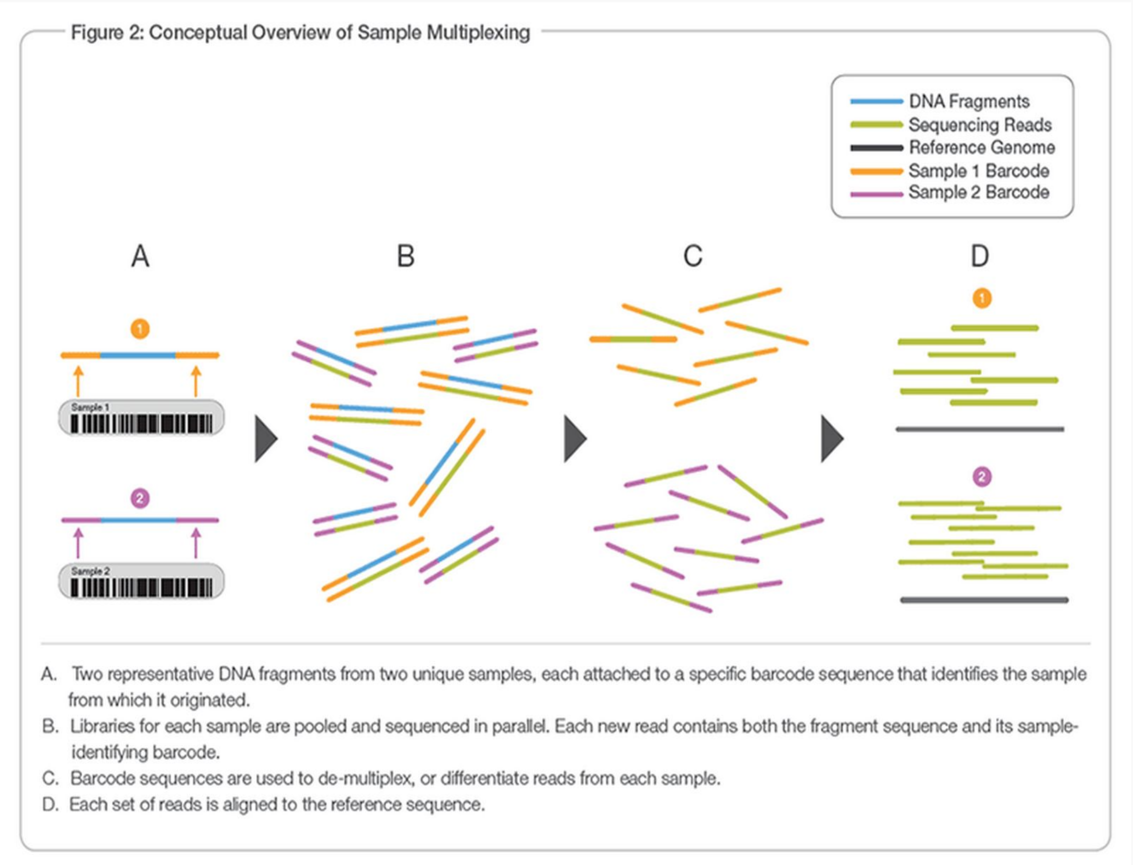


Figure 2: Conceptual Overview of Sample Multiplexing

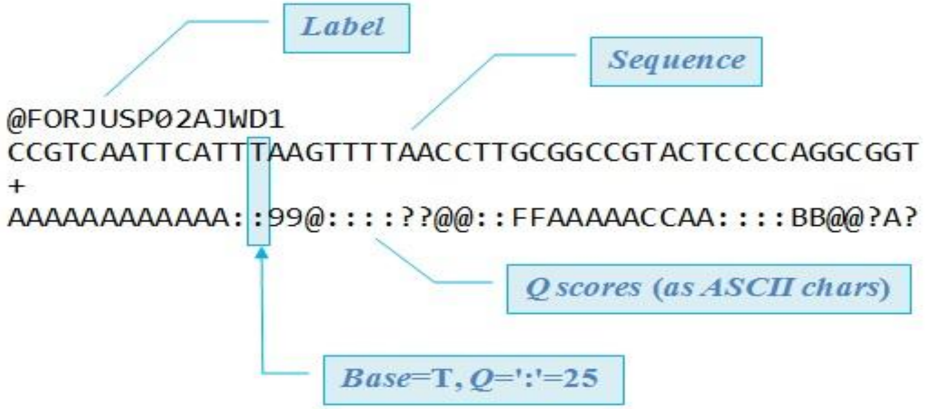
A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.

B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.

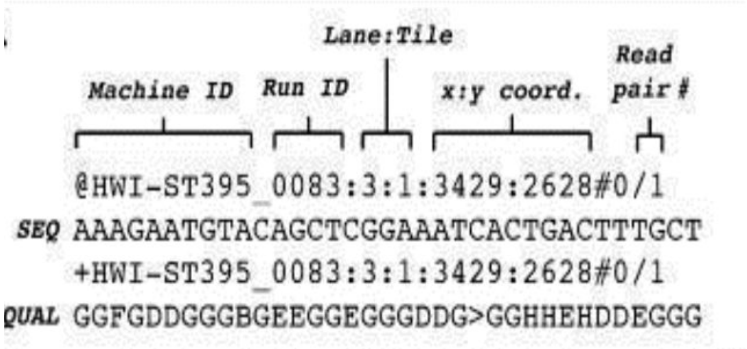
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.

D. Each set of reads is aligned to the reference sequence.

FASTQ format



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```



```
@HWI-ST395_0083:3:1:3429:2628#0/1
AAAGAATGTACAGCTCGGAAATCACTGACTTTGCT
+HWI-ST395_0083:3:1:3429:2628#0/1
GGFGDDGGGBGEEGGEGGGDDG>GGHHEHDDEGGG
```

A FASTQ file normally uses four lines per sequence.

Line-1 begins with a **'@'** character and is followed by a sequence identifier and an optional description.

Line-2 is the raw sequence letters.

Line-3 begins with a **'+'** character and is optionally followed by the same sequence identifier again.

Line-4 encodes the quality scores (ASCII) for the sequence in Line 2.

Historically there are a number of different FASTQ formats. These include the Sanger Format, Illumina/Solexa 1.0, Illumina 1.3, 1.5, 1.8 and 1.9

Cock et al., Nucleic Acids Res. 2010 Apr;38(6):1767-71.