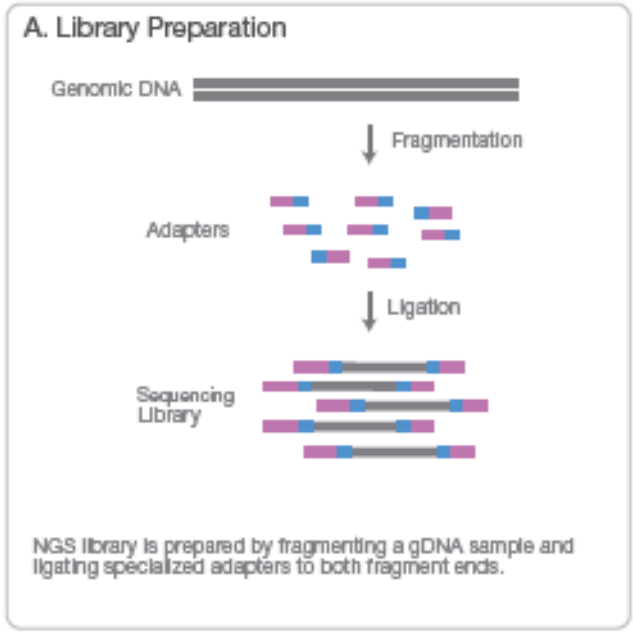
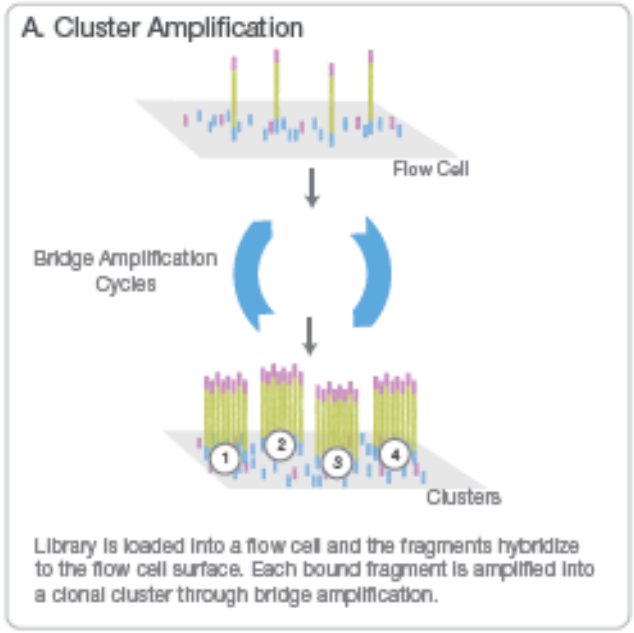
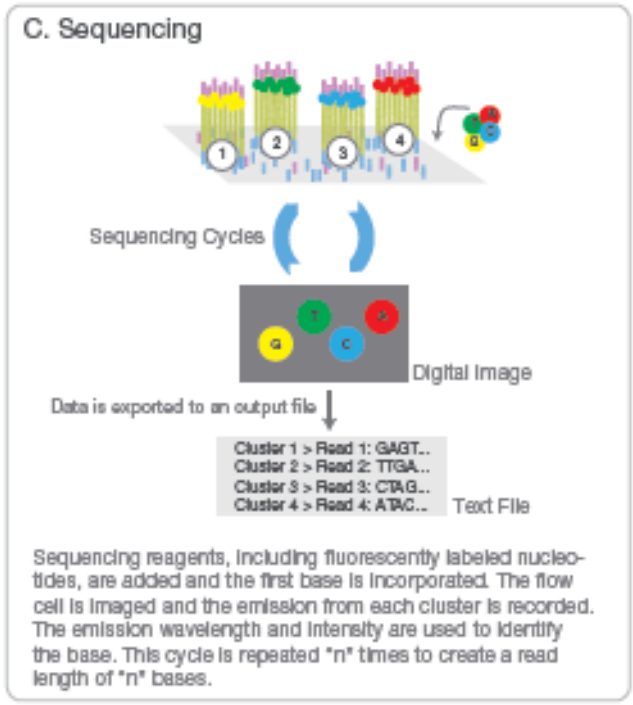
Practical aspects of Illumina sequencing - summary

The steps of Illumina sequencing

1. Fragment genomic DNA, e.g. with a sonicator.
2. Ligate adapters to both ends of the fragments.
3. PCR amplify the fragments with adapters
4. Spread DNA molecules across flowcells. Goal is to get exactly **one DNA molecule** per flowcell lawn of primers. This depends purely on probability, based on the concentration of DNA.
5. Use bridge PCR to amplify the single molecule on each lawn so that you can get a strong enough signal to detect. Usually this requires several hundred or low thousands of molecules.
6. Sequence by synthesis of complementary strand: reversible terminator chemistry.

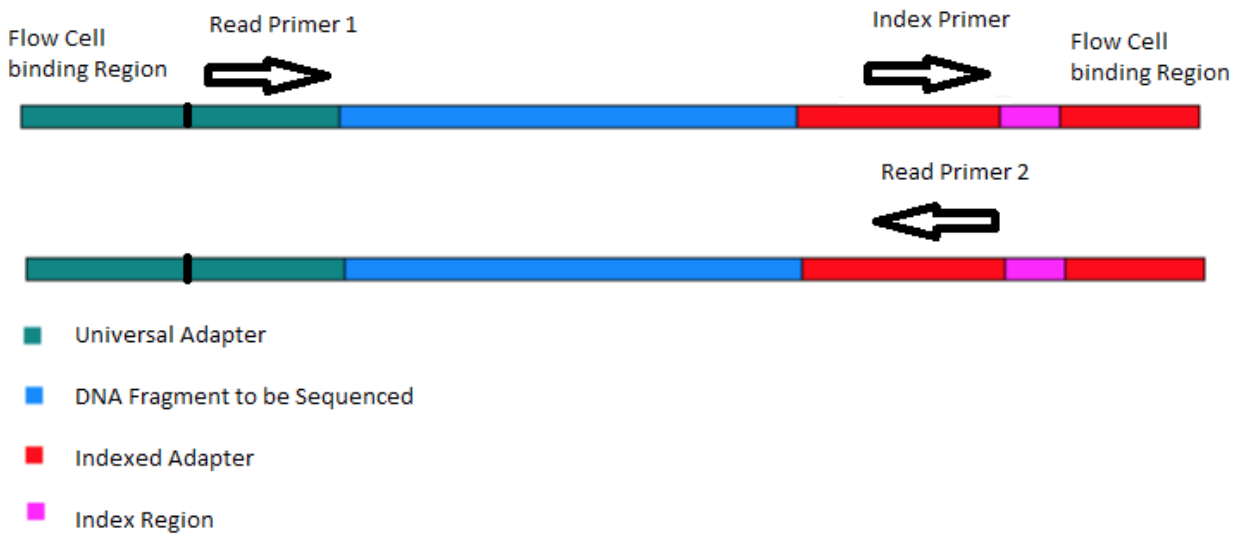






Sources of errors: adapters

- In step 2, adapters are ligated to the end of the fragments



Sequencing random fragments of DNA is possible via the addition of short nucleotide sequences which allow any DNA fragment to:

- Bind to a flow cell for next generation sequencing
- Allow for PCR enrichment of adapter ligated DNA fragments only
- Allow for indexing or 'barcoding' of samples so multiple DNA libraries can be mixed together into 1 sequencing lane (known as multiplexing)

From:
http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

Fragment add-ons

Adapters

Primers

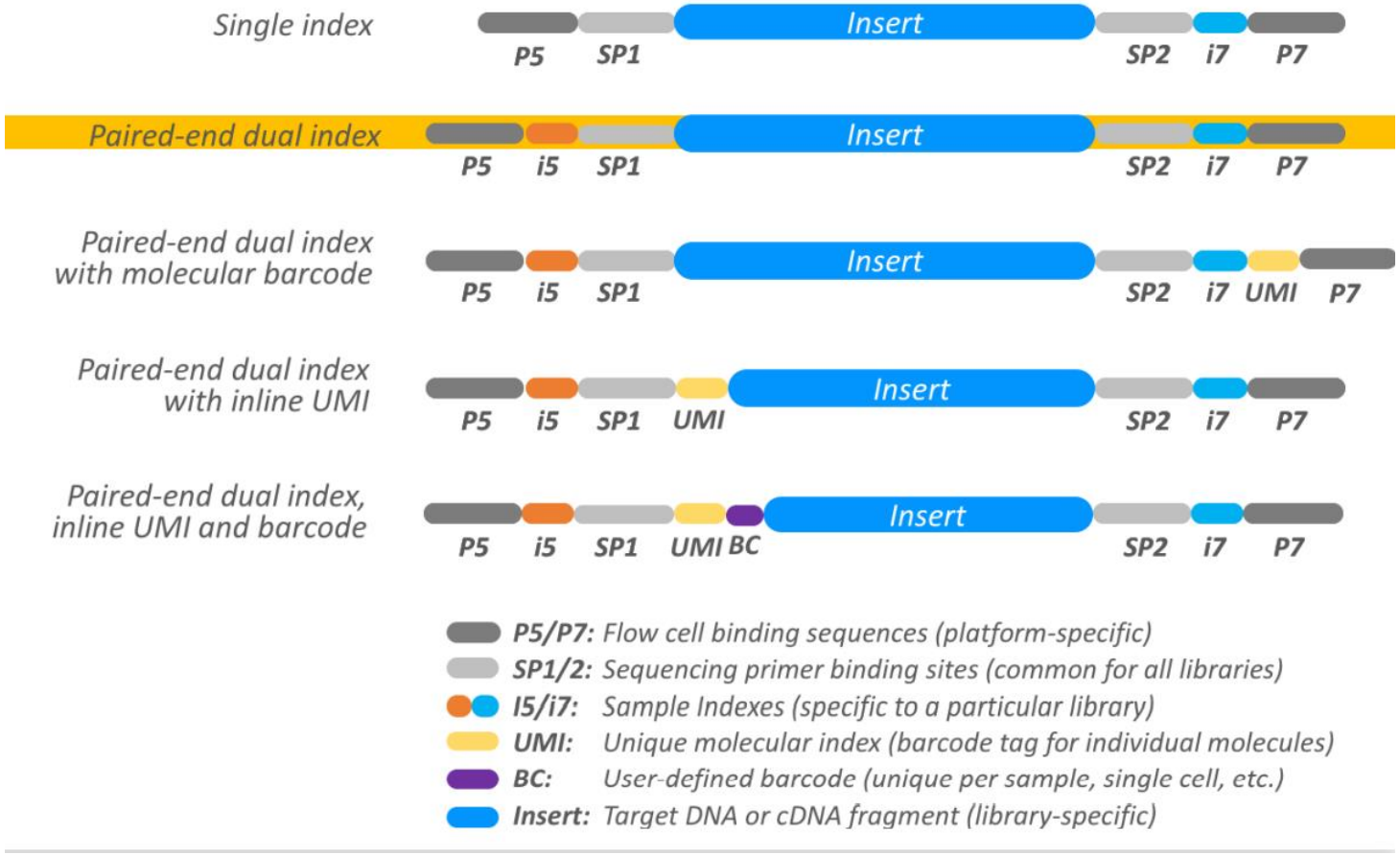
Tags

Barcodes

UMIs

Spacers

Linkers



Fragment add-ons

Have to be present:

P5/P7 – adapters for flow-cell binding

SP1/SP2 – binding point for sequencing primer

Common add-ons:

i5/i7 – sample index – to distinguish sequencing libraries

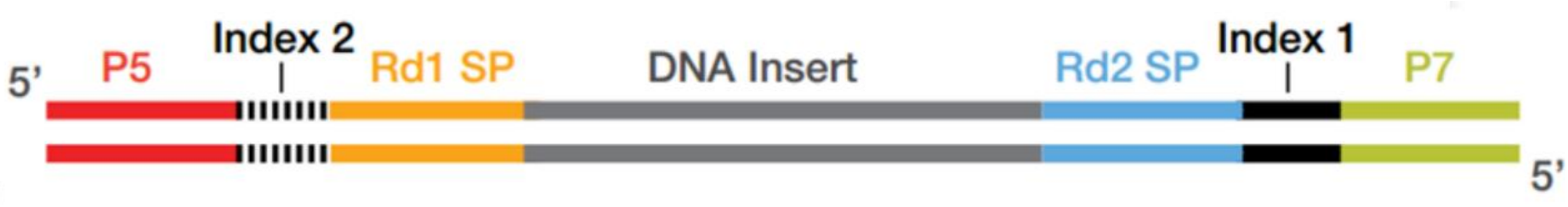
Spacers - for sequence elongation

Linkers – for better binding of oligonucleotides

Optional:

Barcode – unique sequence

UMI – Unique Molecular Identificator – for identification of PCR duplicates



[Indexed Sequencing Overview Guide (15057455) (ox.ac.uk)]

Removal of adapters from library

- Necessary step!
- Removal of unligated adapters and adapter dimers (two adapters ligated to each other) is essential to improve data throughput and quality
- Redundant adapters often compete with library fragments for binding to a flow cell, reducing data output.
- Adapter dimers can also clonally amplify and generate sequencing “noise” that must be filtered out during data analysis.
- An excess of unligated adapters makes libraries more prone to index skipping during sequencing

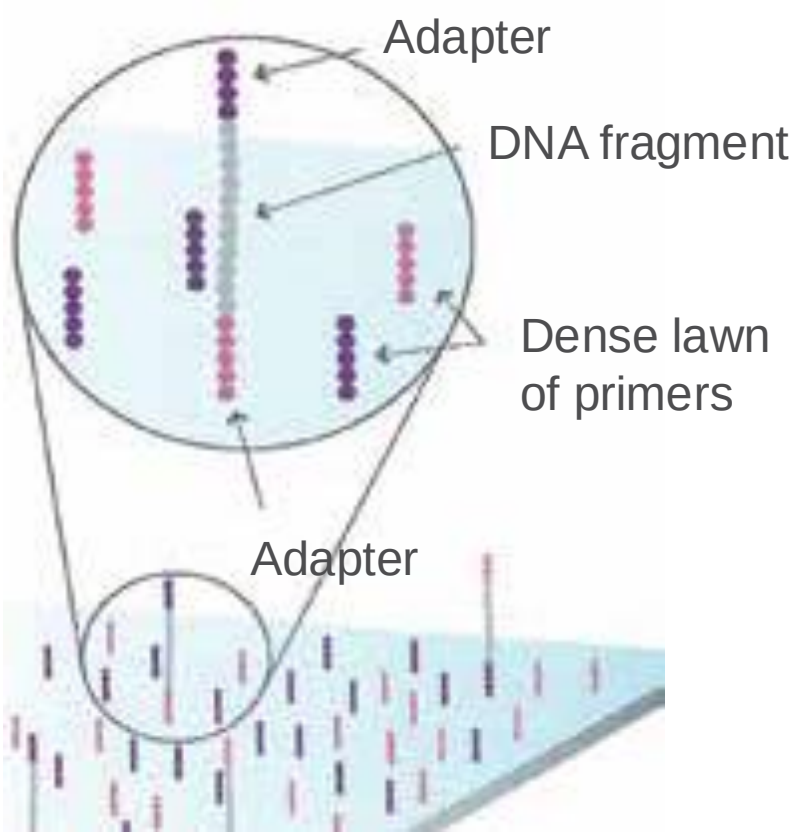
Sources of errors: PCR duplicates

- In step 3 we are *intentionally* creating multiple copies of each original genomic DNA molecule so that we have enough of them.

- PCR duplicates occur when **two copies of the same original molecule get onto different primer lawns in a flowcell**.

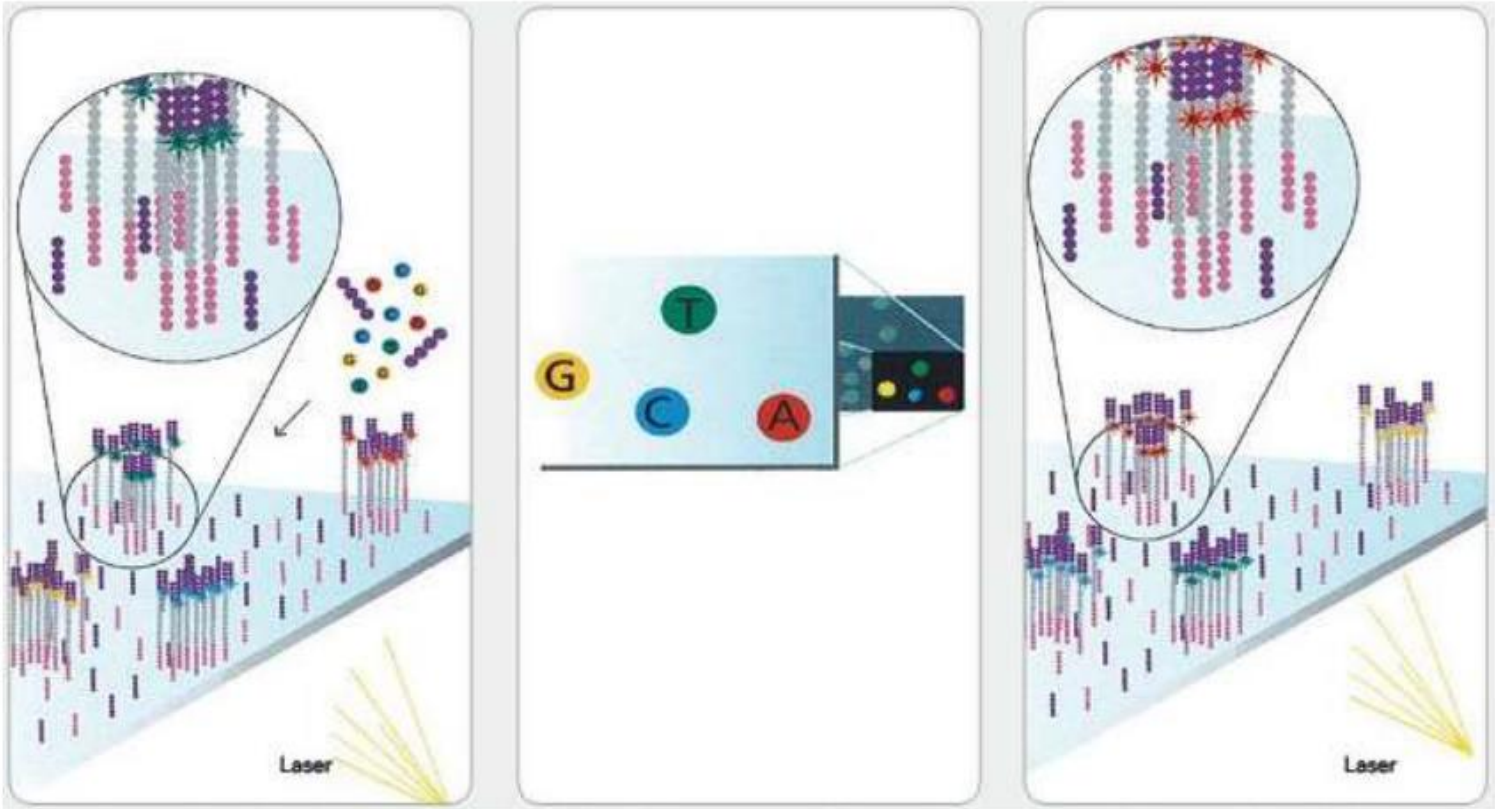
- In consequence we read the very same sequence twice!

Higher rates of PCR duplicates e.g. 30% arise when you have too little starting material such that greater amplification of the library is needed in step 3, or when you have too great a variance in fragment size, such that smaller fragments, which are easier to PCR amplify, end up over-represented.



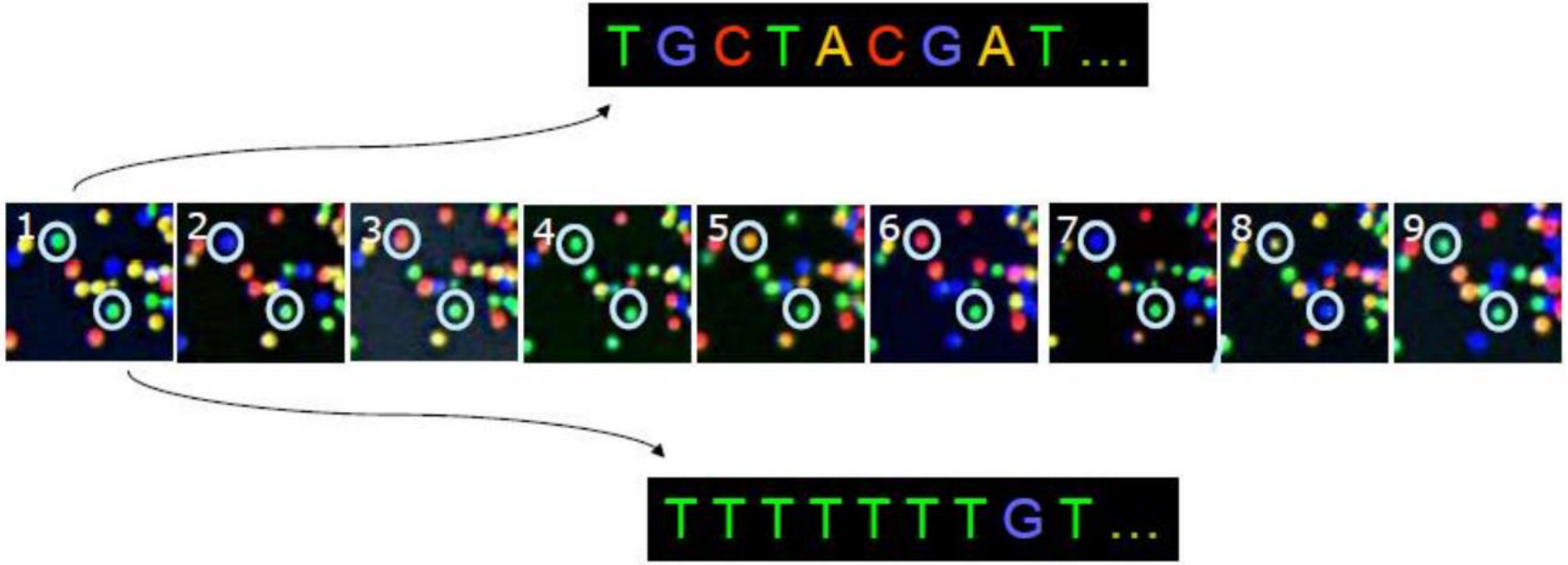
Find beautiful explanation of probabilities and much more at: https://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Clusters of identical sequences are created

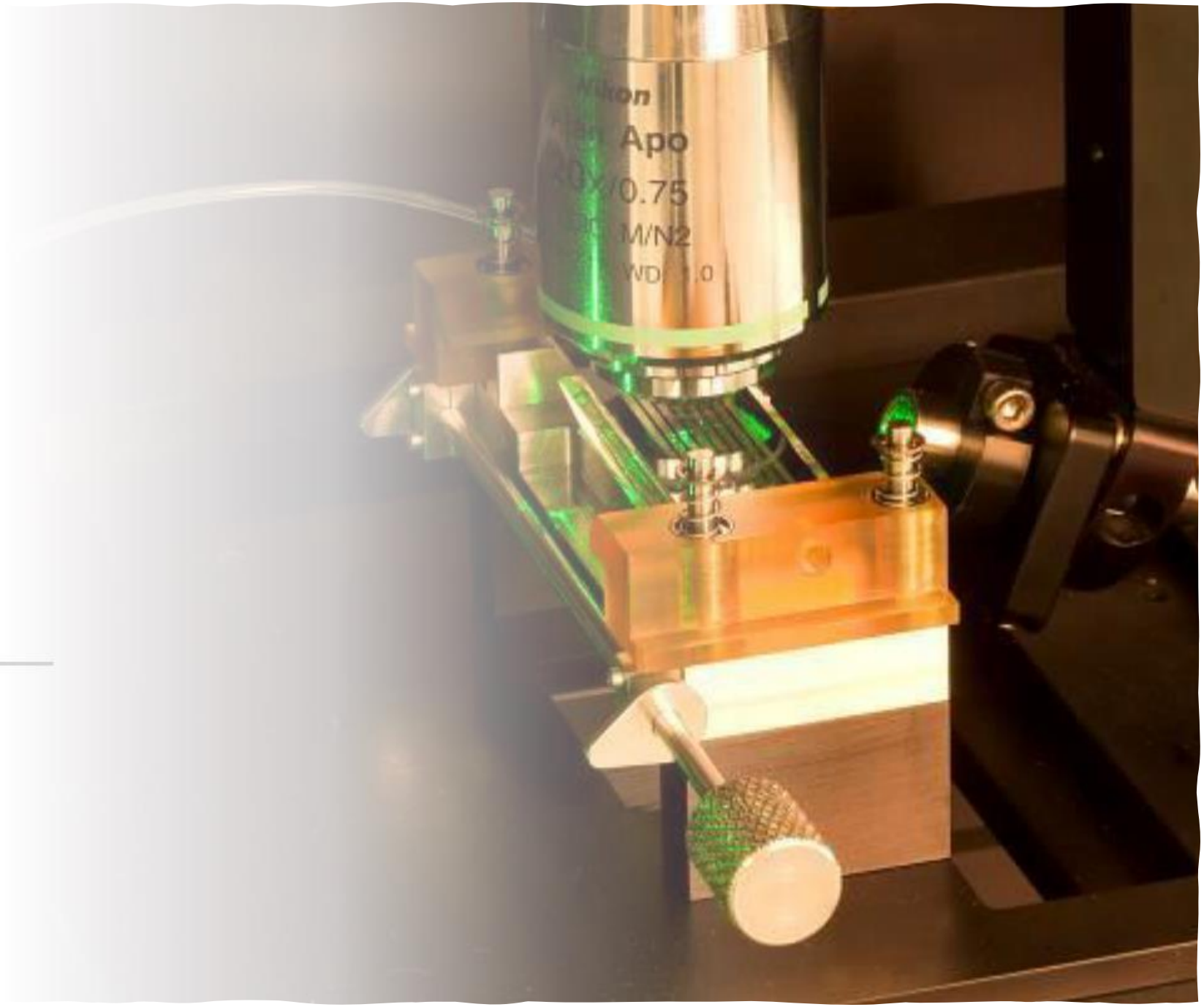


Step 0 of analysis

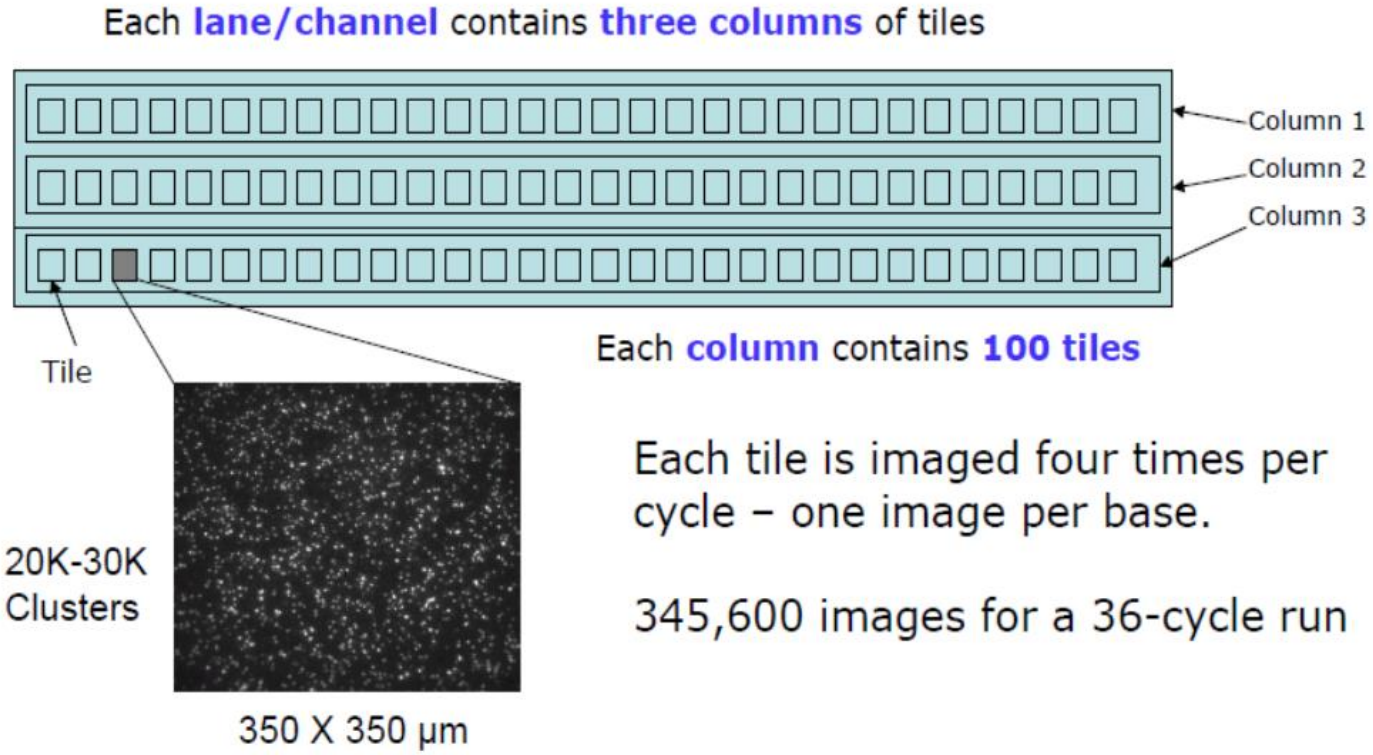
- The identity of each base in the cluster is read from the sequence images
- One cycle -> four images!



Flow-cell imaging

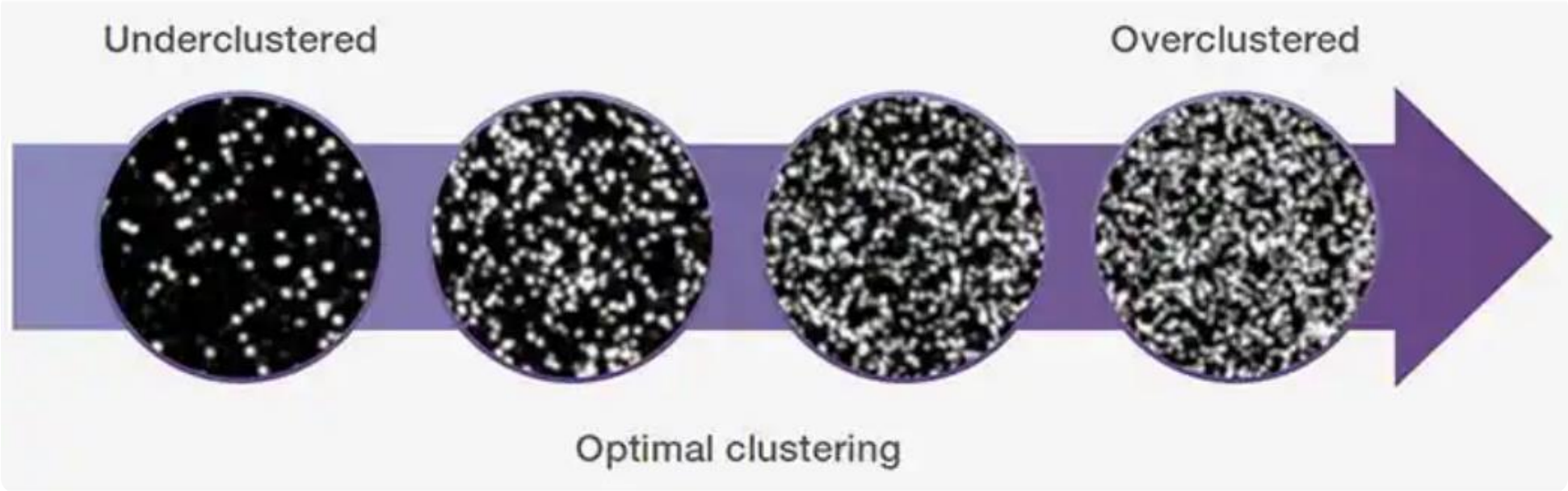


How it works



Each tile is imaged four times per cycle – one image per base.

345,600 images for a 36-cycle run



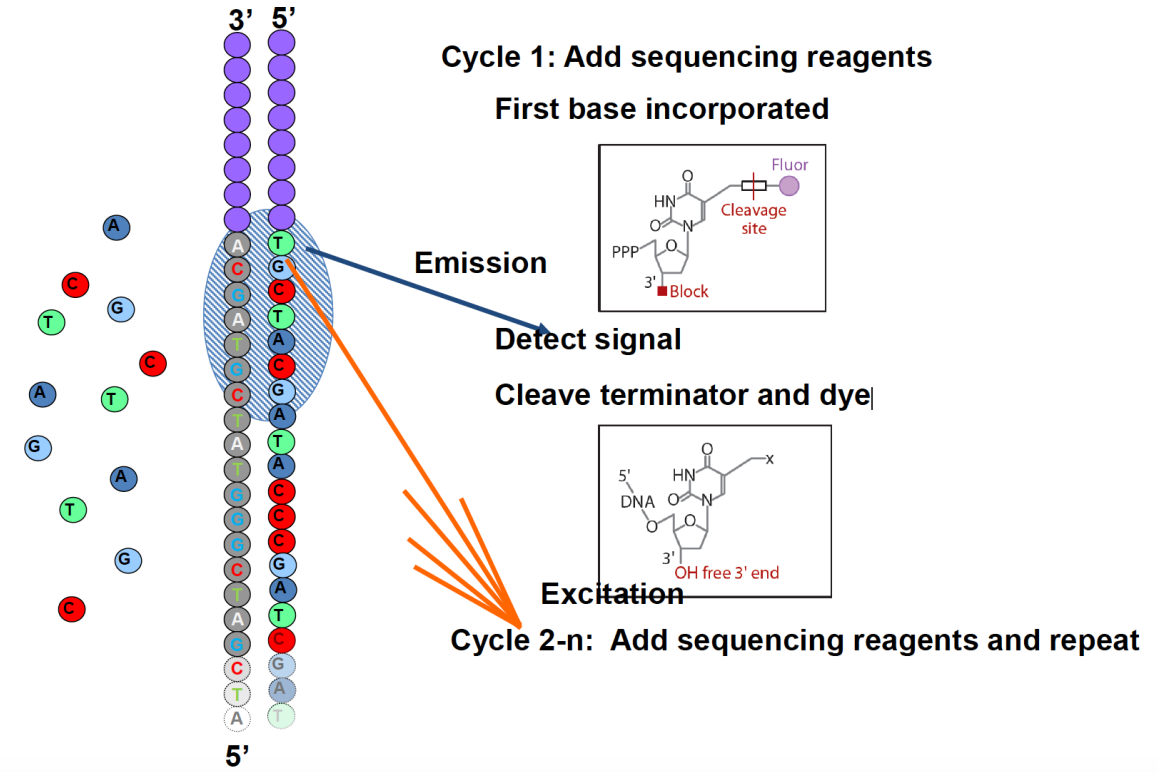
Error source: Library concentration

- The concentrations of prepared NGS libraries can vary widely due to differences in the quantity and quality of input nucleic acid, as well as in the target enrichment method that may be used.
- **underclustering** due to a **low** library concentrations can result in reduced reads against capacity
- too many clusters can result in a low-quality score and problematic subsequent analysis - clusters are poorly distinguished by the image analysis program!

Sources of errors: sequencing by synthesis – the fluorescence

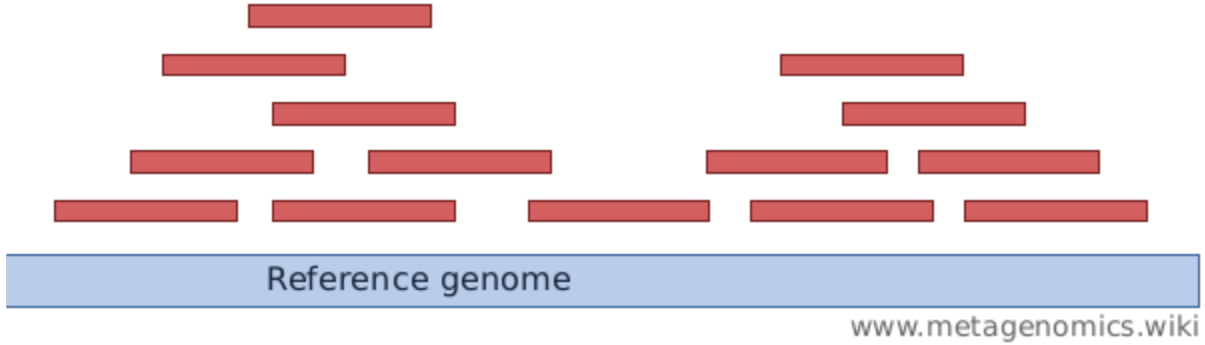
- In step 5 we amplify the signal and detect the fluorescence of each base
- The assumption is that in a cycle, every molecule on the flowcell is extended by one base

- The reality:
  - Some molecules are not extended or their base has no fluorescent dye
  - The previous fluorescent dye is not cleaved – the signal from the cluster after a few cycles is a mix of signals from previous bases



Sequencing **coverage**

Coverage in DNA sequencing **is the number of unique reads that include a given nucleotide** in the reconstructed sequence.



Depth of coverage
(coverage depth / mapping depth)

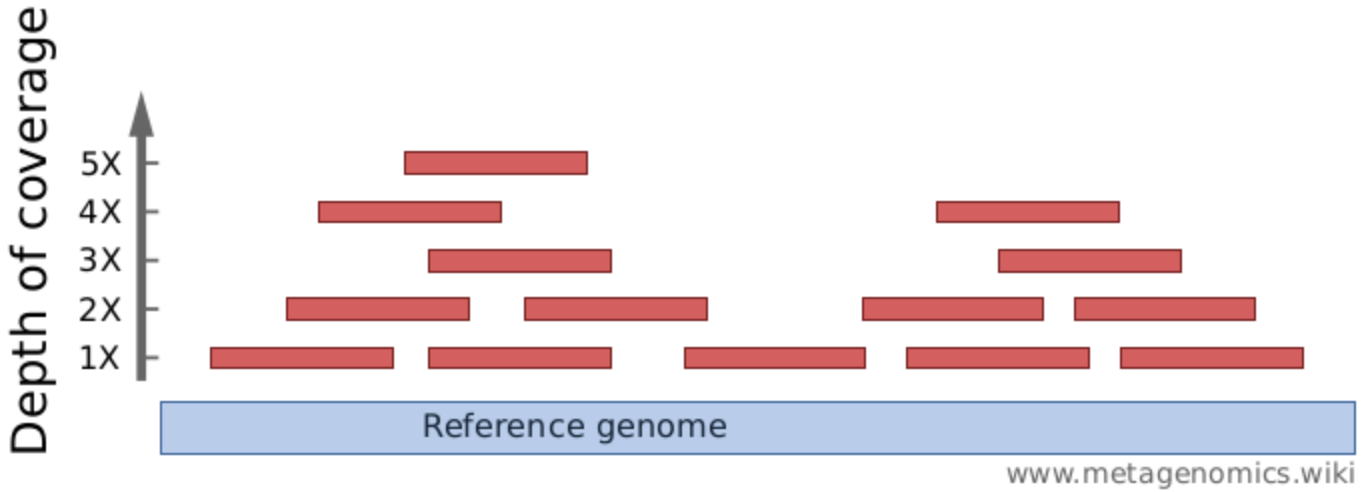
How strongly is the genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced (in other words, how many reads cover it).

Average coverage of the genome (Av)

$$Av = (NxL)/G$$

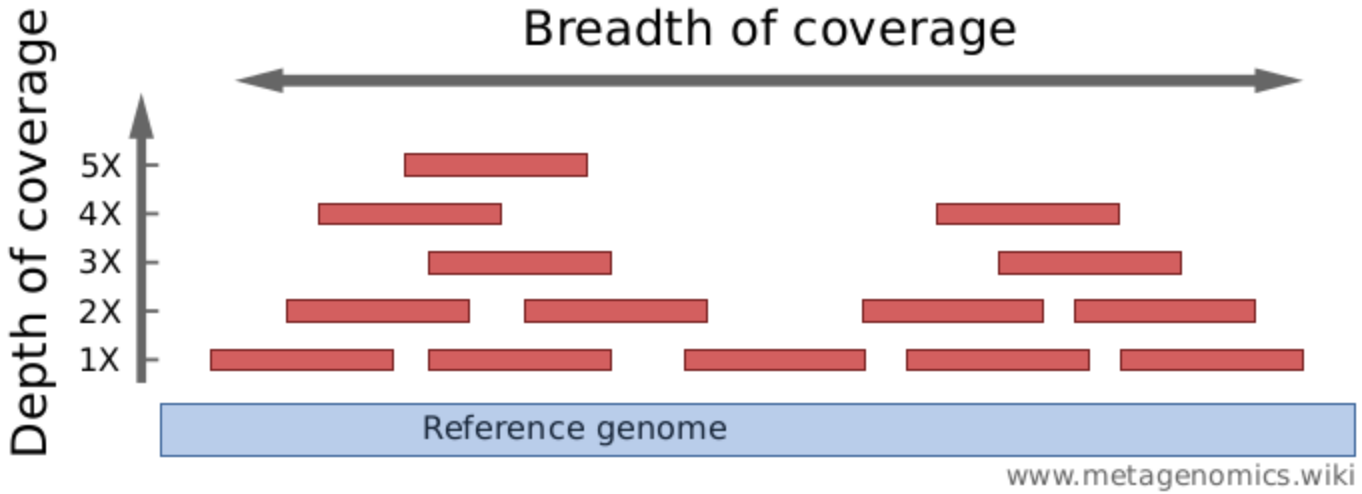
G - length of the original genome
N - number of reads
L - average read length



The coverage depth of a genome is calculated as **the number of bases of all short reads that match a genome divided by the length of this genome**. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).

Breadth of coverage
(covered length)

What proportion of the genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?



Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth.
For example: "90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth."

Coverage recommendations

Coverage is determined based on:

Read lengths

Genome size

Application

Recommendations in the literature

Gene expression levels

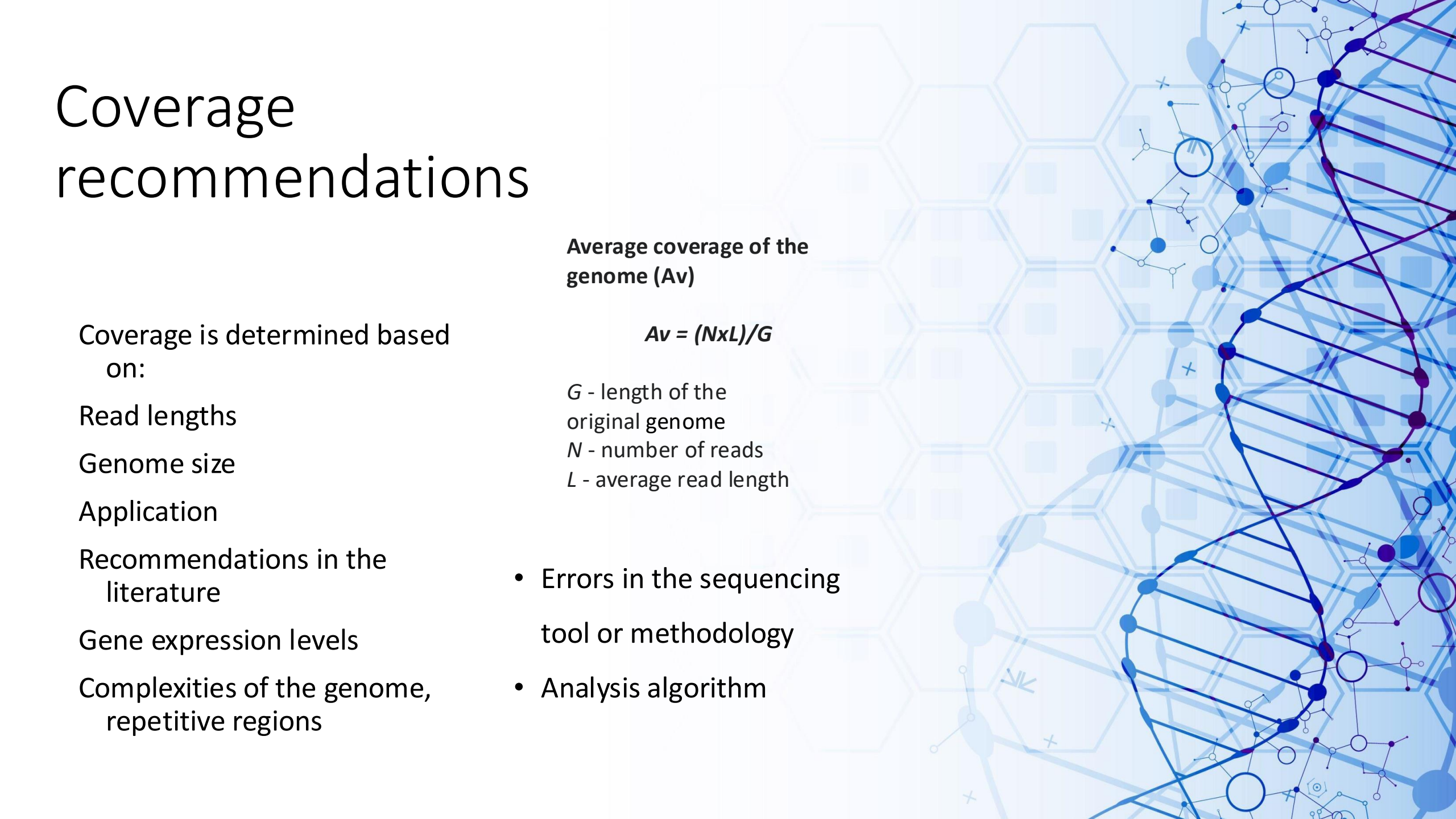
Complexities of the genome, repetitive regions

Average coverage of the genome (Av)

$$Av = (NxL)/G$$

G - length of the original genome
N - number of reads
L - average read length

- Errors in the sequencing tool or methodology
- Analysis algorithm



Coverage recommendations / DNA

Application Type	Coverage
DNA-Seq (Re-Sequencing)	30 - 80X
DNA-Seq (De novo assembly)	100X
SNP Analysis / Rearrangement Detection	10 - 30X
Exome	100 - 200X
ChIP-Seq	10 - 40X

Average coverage of the genome (Av)

$$Av = (NxL)/G$$

G - length of the original genome
N - number of reads
L - average read length

[Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)]

Coverage recommendations / RNA

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

Different transcripts are expressed at different levels => more reads will be captured from highly expressed genes
Transcriptome complexity, alternative expression, 3' associated bias, and distribution of expression levels make coverage estimation difficult.

ATTENTION WHEN CALCULATING! We need to count mapped reads, not total reads.

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Coverage recommendations / application

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev at al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Coverage recommendations / application

DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012
	Hi-C	100M	Belton, J.M et al., 2012
	4C (Circularized Chromosome Confirmation Capture)	1-5M	van de Weken, H.J.G. et al., 2012
	5C (Chromosome Carbon Capture Carbon Copy)	15-25M	Sanyal A. et al., 2012
	ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)	15-20M	Zhang, J. et al., 2012
	FAIRE-Seq	25-55M	ENCODE 2011 Genome; Landt et al., 2012
	DNAse 1-Seq	25-55M	Landt et al., 2012
DNA Methylation Sequencing	CAP-Seq	>20M	Long, H.K. et al., 2013
	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	RRBS (Reduced Representation Bisulfite Sequencing)	10X	ENCODE 2011 Genome
	Bisulfite-Seq	5-15X; 30X	Ziller, M.J et al., 2015; Epigenomics Road Map

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

Coverage recommendations / application

RNA-Target-Based Sequencing	CLIP-Seq	10-40M	Cho J. et al., 2012; Eom T. et al., 2013; Sugimoto Y. et al., 2012
	iCLIP	5-15M	Sugimoto Y. et al., 2012; Rogelj B. et al., 2012
	PAR-CLIP	5-15M	Rogelj B. et al., 2012
	RIP-Seq	5-20M	Lu Z. et al., 2014
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M	Metpally RPR et al., 2013; Campbell et al., 2015
	Discovery	~5-8M	Metpally RPR et al., 2013; Campbell et al., 2015

Coverage and Read Depth Recommendations for Next-Generation Sequencing Applications (genohub.com)

How many samples per run?

It depends on the platform used and its maximum and required number of reads per sample (in millions)

Numbers of Single Reads by Instrument Manufacturer

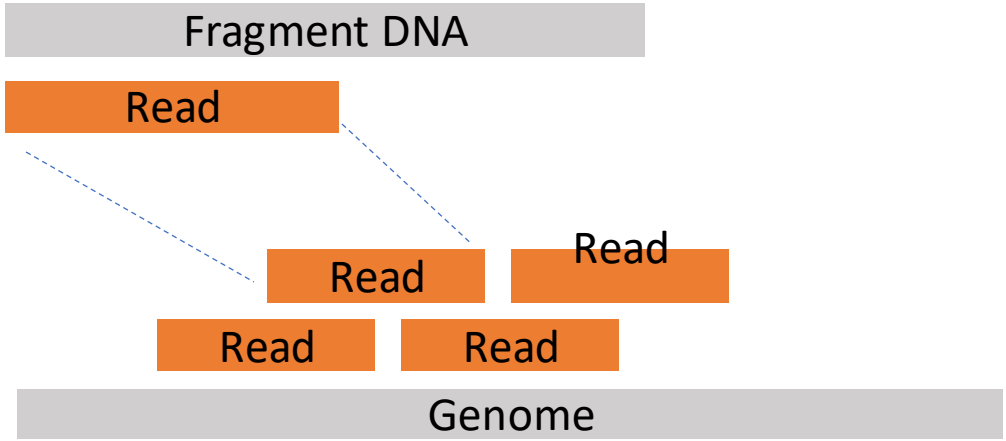
Platform	Instrument	Unit	Reads / Unit	Reference
Illumina	HiSeq X Ten	Lane	375,000,000	1
Illumina	HiSeq 3000/4000	Lane	312,500,000	1
Illumina	HiSeq NextSeq 500 High-Output	Run	400,000,000	2
Illumina	HiSeq NextSeq 500 Mid-Output	Run	130,000,000	2
Illumina	HiSeq High-Output v4	Lane	250,000,000	3
Illumina	HiSeq High-Output v3	Lane	186,048,000	3
Illumina	HiSeq Rapid Run	Lane	150,696,000	3
Illumina	HiScanSQ	Lane	93,024,000	3
Illumina	GAIIx	Lane	42,075,000	3
Illumina	MiSeq v3	Lane	25,000,000	4
Illumina	MiSeq v2	Lane	16,000,000	3
Illumina	MiSeq	Lane	5,000,000	3
Illumina	MiSeq v2 Micro	Lane	4,000,000	5
Illumina	MiSeq v2 Nano	Lane	1,000,000	5

Designing Next-Generation Sequencing Runs (genohub.com)

Single or paired- end?

Single-end sequencing

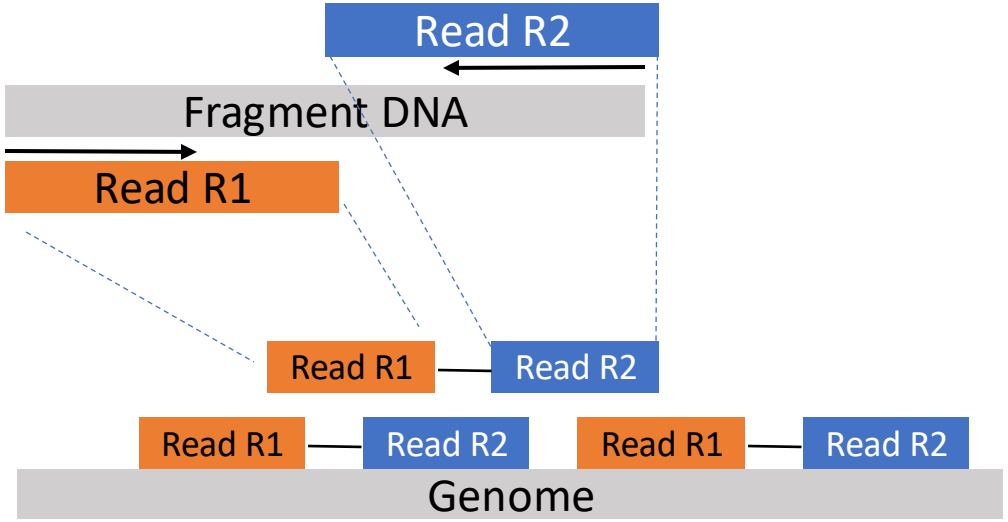
- Pros: fast, cheap
- Cons: limited use
- Useage: usually sufficient for studies looking to detect counts rather than structural changes, such as RNA-Seq or ChIP-Seq



Single or paired- end?

Paired-end sequencing

- Pros:
  - greater accuracy, double the number of reads per sample in one run (higher capacity) for less than the cost of two sequencing runs
- Cons: slower, more expensive (relatively)
- Usage:
  - de novo genome assembly
  - Analysis of structural changes (deletions, insertions, inversions) and SNPs
  - A study of splicing variants
  - Epigenetic modifications (methylation)

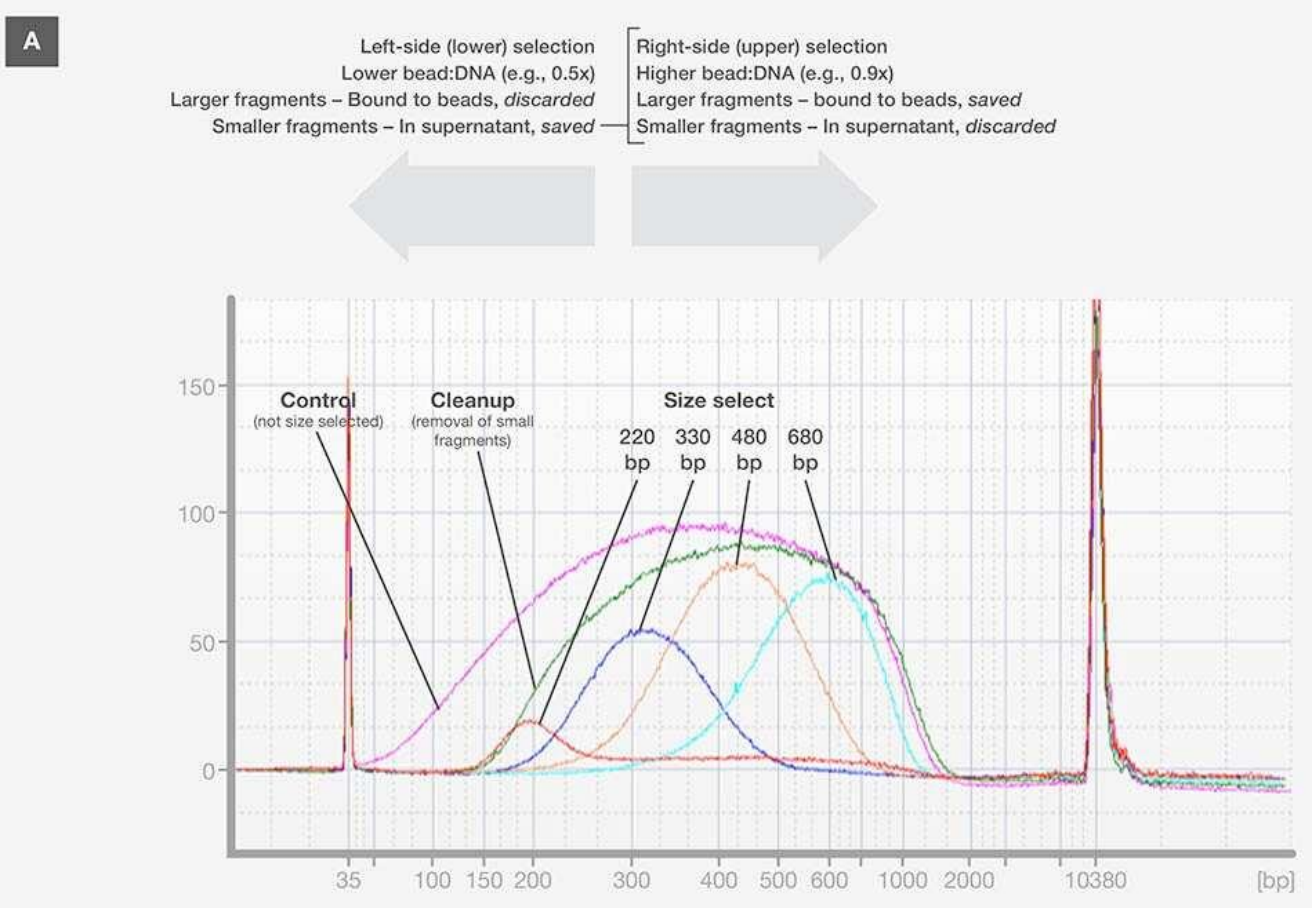


Read length

- Longer read lengths provide more precise information about the relative positions of the bases in the genome, they are more expensive than shorter ones.
- 50-75 cycles are typically sufficient for simple mapping of reads to a reference genome and quantifying experiments e.g. gene expression (RNA-Seq)
- Read lengths greater than or equal to 100 are typically chosen for genome or transcriptome studies that require greater precision
- **The exact read length depends on the length of the inserts!!!**

Read length and fragments

- The length of the fragments should roughly correspond to the length of the read (in the case of paired-end reads their sum)
- Uniformity of fragment sizes is essential because read lengths are limited
- **Significantly longer DNA inserts** => some parts of the inserts remain unsequenced.
- **Shorter than recommended** => suboptimal use of sequencing reagents and resources.
- The combination of short and long inserts => reduces sequencing efficiency and presents problems in data analysis.



[Preparation of DNA Sequencing Libraries for Illumina Systems—6 Key Steps in the Workflow | Thermo Fisher Scientific - CZ]

Read length and fragments!

Read length is limited by the sequencing platform and reagent kit

Reagent Type		Reagent Kit Size	Maximum Number of Cycles	Additional Cycles Needed for Dual Index?
iSeq™ 100	i1 (v1 or v2)	300	322	No
MiniSeq™	Rapid Kit	100	128	Yes - 7 cycles
	High Output or Mid Output	75	92	No
		150	168	
		300	318	
MiSeq™	v2 (including Micro and Nano kits)	50	79	Yes - 7 cycles
		300	329	
		500	529	
	v3	150	179	
		600	629	
NextSeq™ 500/550	High Output or Mid Output	75	92	No
		150	168	
		300	318	
HiSeq™ 1000/1500/2000/2500	Rapid SBS v2	50	79	7 cycles required for paired-end flow cells
		200	229	
		500	529	
	TruSeq SBS v3	50	58	
		200	209	
	HiSeq SBS v4	50	79	
		250	279	

[How many cycles of SBS chemistry are in my kit? (illumina.com)]

Sequencing Platform	SBS Kit Version	Maximum Read Length
iSeq 100	v1	2 x 151bp
	v2	2 x 151bp
MiniSeq	MO*	2 x 151bp
	HO*	2 x 151bp
MiSeq	v2	2 x 251bp
	v3	2 x 301bp
NextSeq 500/550	MO*	2 x 151bp
	HO*	2 x 151bp
NextSeq 1000/2000	P1, P2, P3	2 x 151bp
HiSeq 1000/1500/2000/2500	HO* v3	2 x 101bp
	HO* v4	2 x 126bp
	RR** v4	2 x 251bp
HiSeq 3000/4000	N/A	2 x 151bp

[Maximum read length for Illumina sequencing platforms]

More resources

- Practical tips for lab library preparation: Preparation of DNA Sequencing Libraries for Illumina Systems—6 Key Steps in the Workflow | Thermo Fisher Scientific - CZ

- Practical tips for sequencing run setup: Designing Next-Generation Sequencing Runs (genohub.com)

- Indexed sequencing Illumina guide: Indexed Sequencing Overview Guide (15057455) (ox.ac.uk)
- Sequencing depth and coverage: key considerations in genomic analyses | Nature Reviews Genetics