MUNI | RECETOX

MUNI FI

NGS introduction

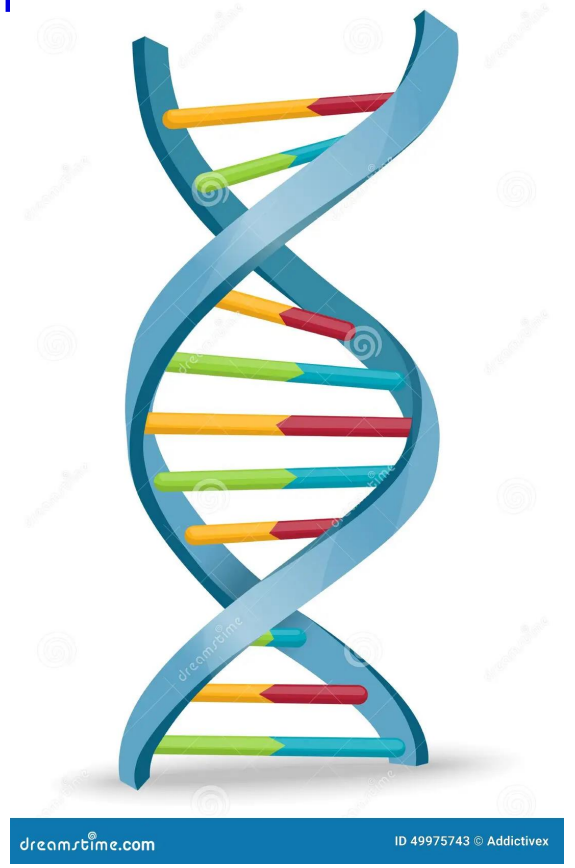
IV110 Projekt z bioinformatiky I
IV114 Projekt z bioinformatiky a systémové biologie
E4014 Projekt z Matematické biologie a biomedicíny - biomedicínská bioinformatika

Mgr. Eva Budinská, Ph.D.
doc. Ing. Matej Lexa, Ph.D.

1 eva.budinska@recetox.muni.cz, matej.lexa@fi.muni.cz

Next-generation sequencing introduction

- Deciphering DNA sequence is essential for all the branches of “biological” research
- It has become widely adopted in numerous laboratories all over the world
- **Next-generation sequencing (NGS)** is a new (almost) technology in the sequencing
- It helps to overcome the limitations of older techniques such as speed, scalability, throughput and resolution



Year 2010


华大基因
BGI


Complete
genomics


illumina


Roche


life
technologies


PACIFIC
BIOSCIENCES


Helicos
BioSciences Corporation


VISIGEN
BIOTECHNOLOGIES, INC


ion torrent


Oxford
NANOPORE
Technologies


RainDance
Technologies

Year 2023

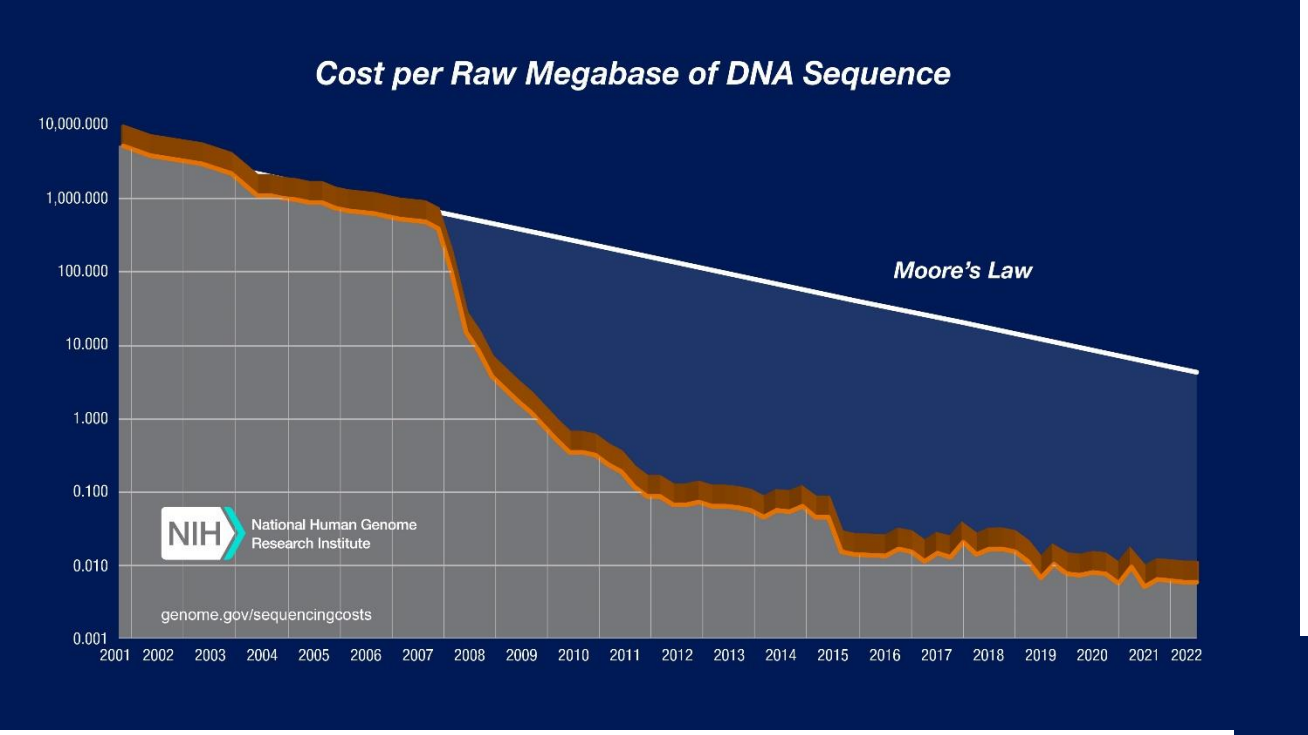

10x GENOMICS
华大基因 BGI
Complete genomics
illumina
Roche
life technologies
PACIFIC BIOSCIENCES
Helicos BioSciences Corporation
VISIGEN BIOTECHNOLOGIES, INC
ion torrent
Oxford NANOPORE Technologies
RainDance Technologies
ThermoFisher SCIENTIFIC

Comparison of NGS

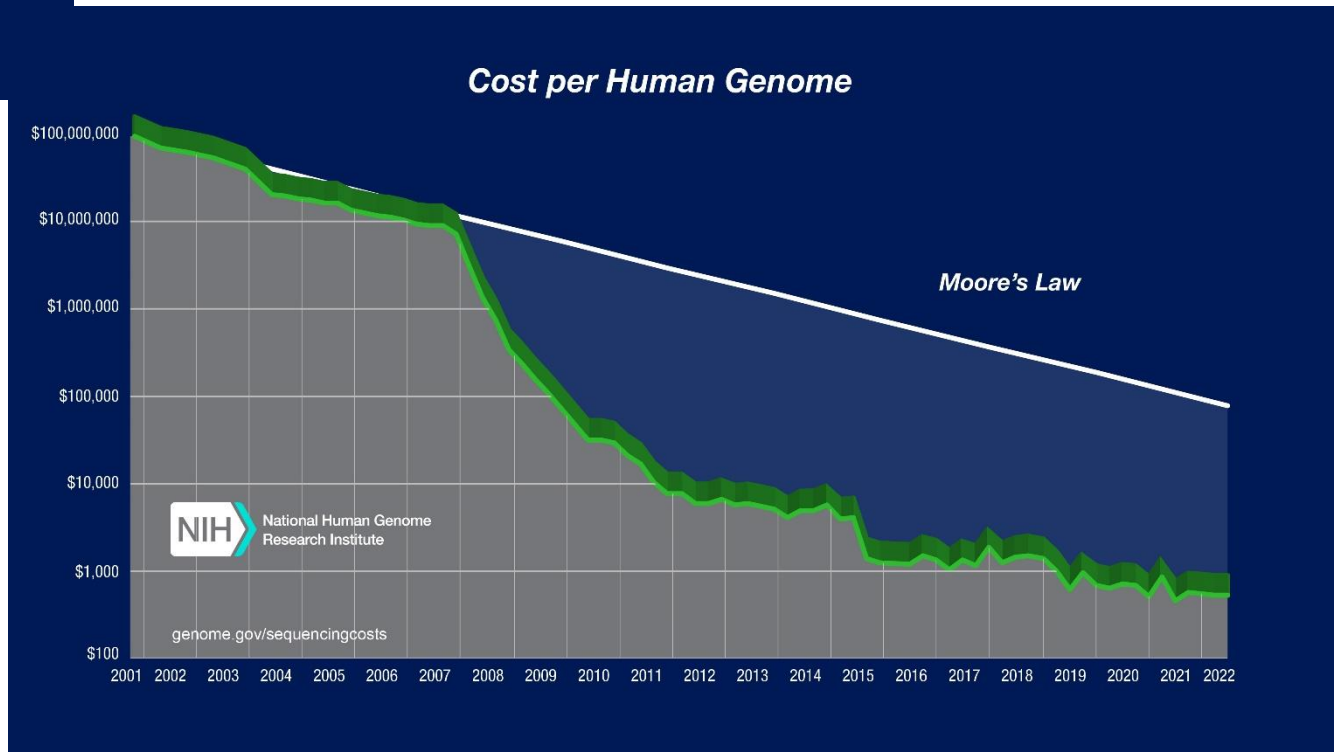
Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	30,000 bp (N50); maximum read length >100,000 bases[66][67][68]	87% raw-read accuracy[69]	500,000 per Sequel SMRT cell, 10–20 gigabases[66][70][71]	30 minutes to 20 hours[66][72]	$0.05–$0.08	Fast. Detects 4mC, 5mC, 6mA.[73]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 600 bp[74]	99.6%[75]	up to 80 million	2 hours	$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[76]	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Combinatorial probe anchor synthesis (cPAS- BGI/MGI)	BGISEQ-50: 35-50bp, MGISEQ 200: 50-200bp, BGISEQ-500, MGISEQ-2000: 50-300bp[77]	99.9% (Phred30)	BGISEQ-50: 160M, MGISEQ 200: 300M, BGISEQ-500: 1300M per flow cell, MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.	1 to 9 days depending on instrument, read length and number of flow cells run at a time.	$0.035- $0.12		
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences.[78]
Nanopore Sequencing	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)	~92–97% single read	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$500–999 per Flow Cell, base cost dependent on expt	Longest individual reads. Accessible user community. Portable (Palm sized).	Lower throughput than other machines, Single read accuracy in 90s.
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2400	Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time

MUNI FI

CETOX



[DNA Sequencing Costs: Data (genome.gov)]

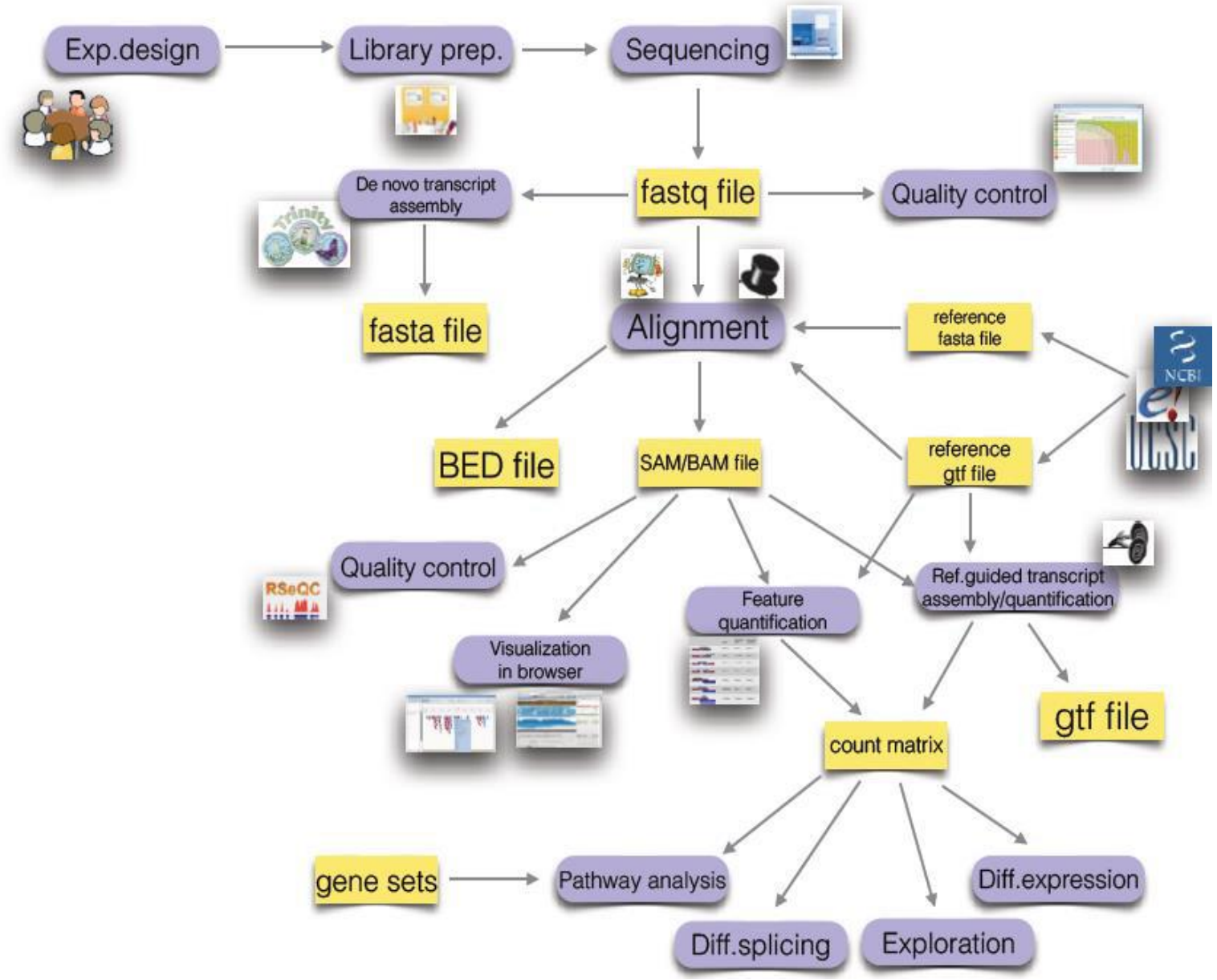


*Seq things

- NGS sequencing has a **wide range of use**
- One of many nice list give you an example of all possible applications
- http://enseqlopedia.com/enseqlopedia/

- Approximately (on this list) ~**200 different** techniques…
- Another (simple) list of NGS based techniques
- https://liorpachter.wordpress.com/seq/

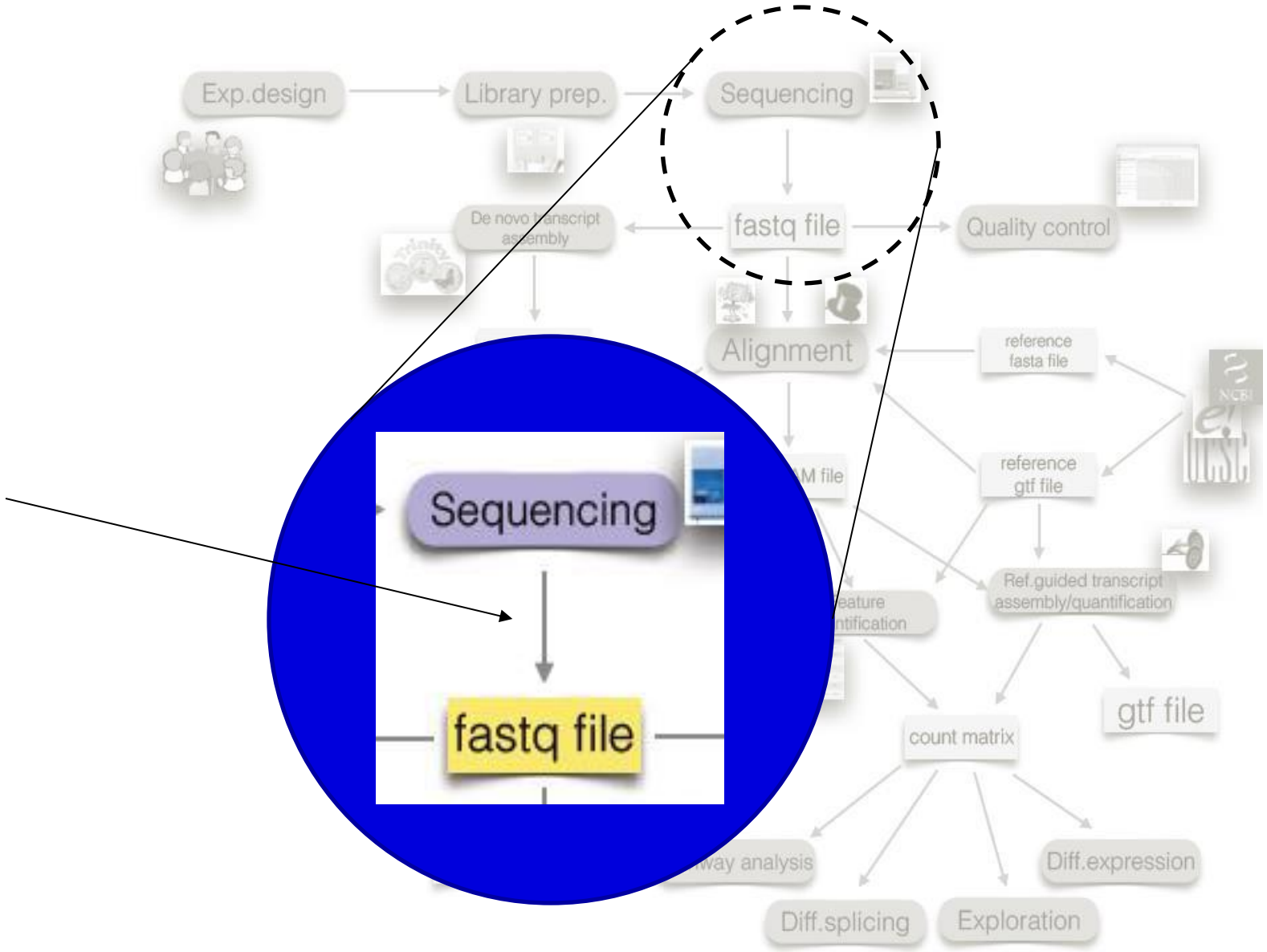
The NGS analysis pipeline



MUNI FI

MUNI | RECETOX

Step 0: base calling (image analysis) + base quality control

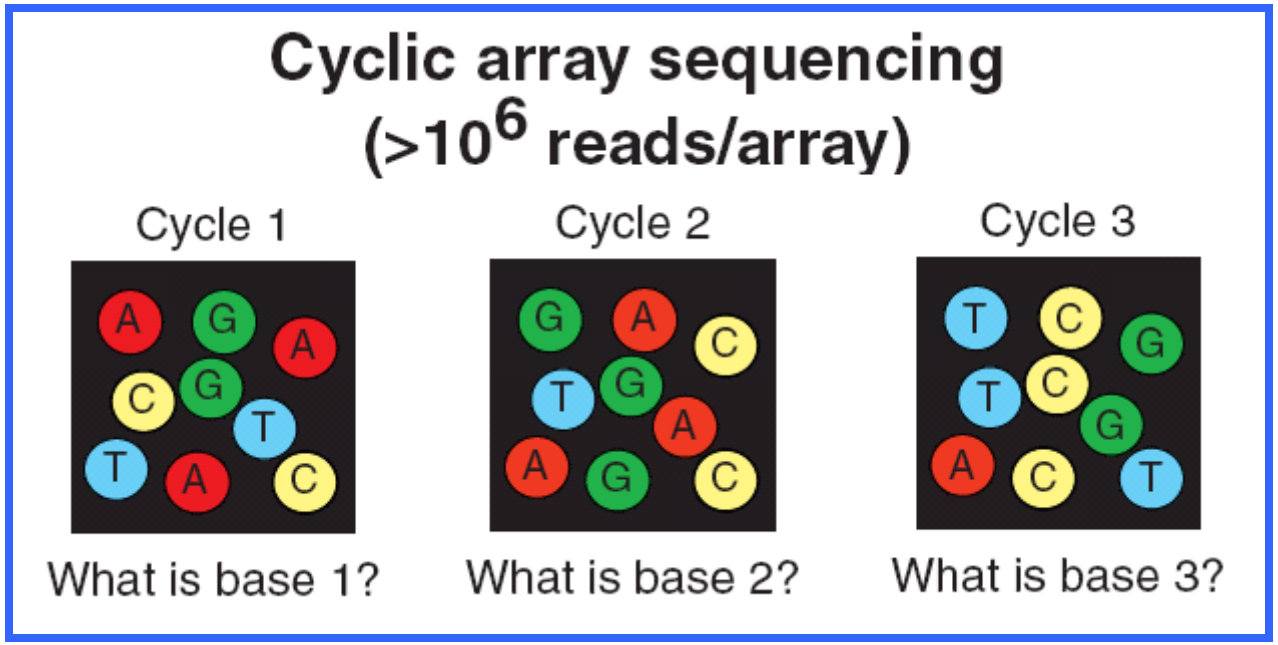


MUNI FI

MUNI | RECETOX

NGS sequencing is a **high-throughput** sequencing

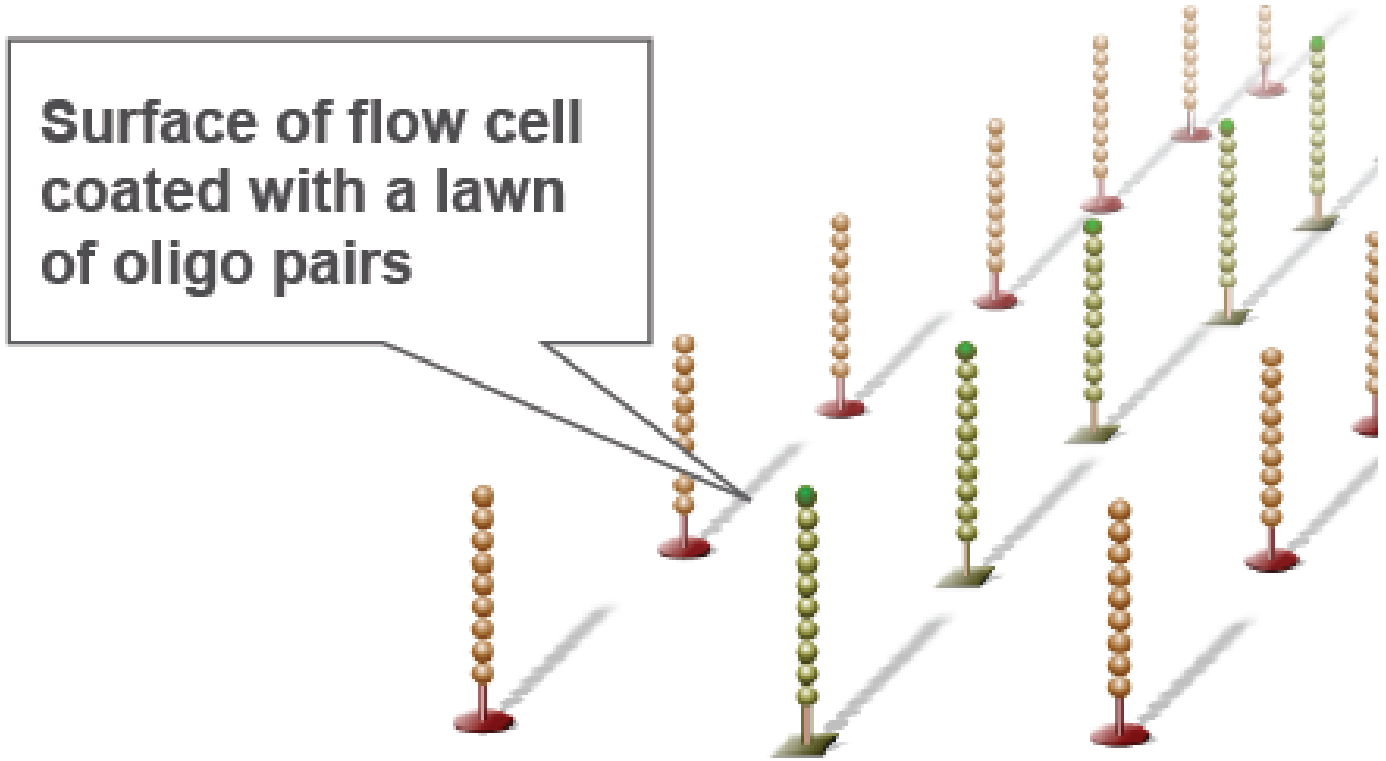
Sanger **Polony**



•Polony sequencing takes place using array of polonies, in which all amplicons of the same DNA fragment are clustered together on the same region of the array. These groups of amplicons were termed polonies, shortcut for polymerase colonies.

DNA Library Preparation

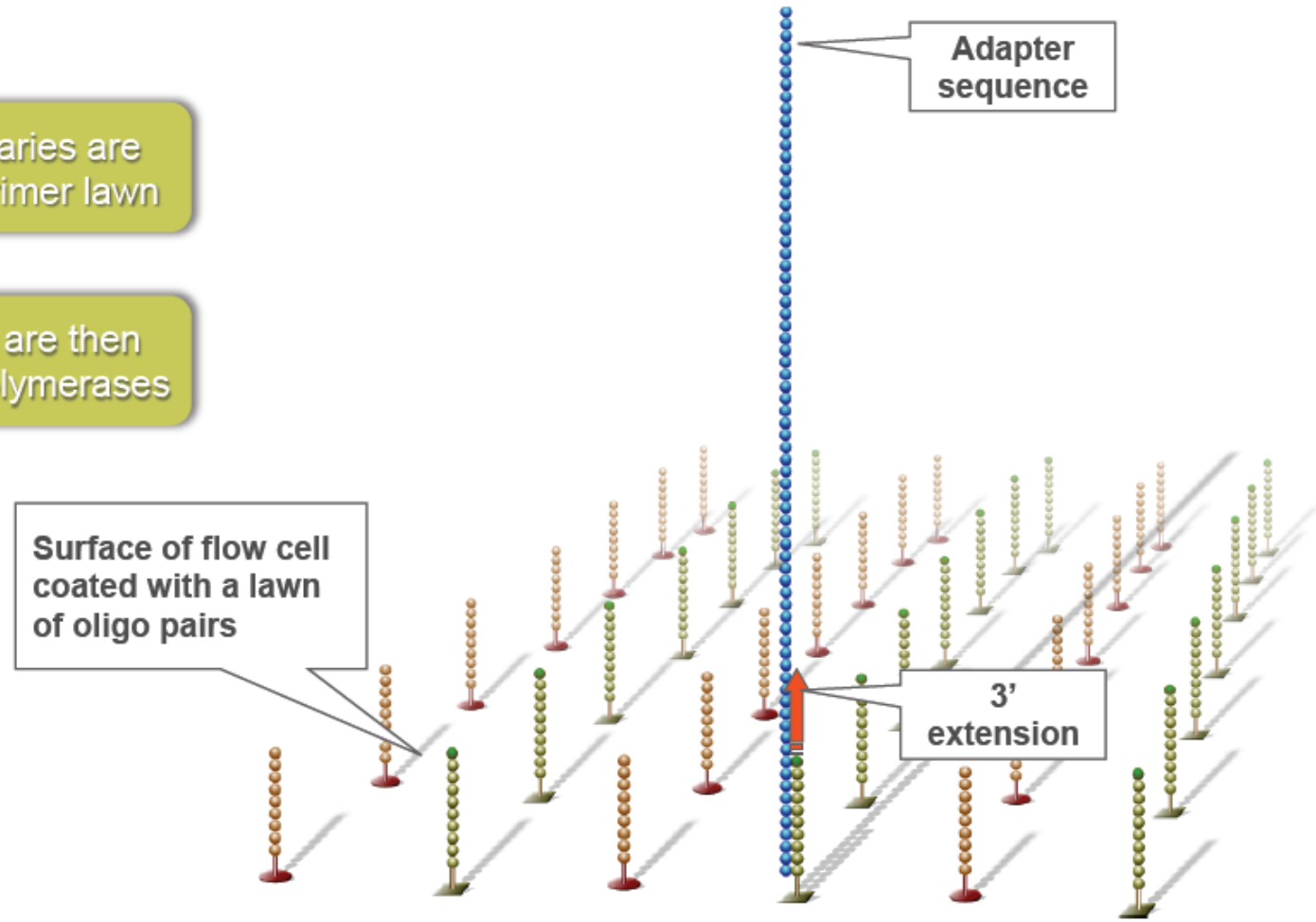
Two PCR primers are attached to the surface of flowcell. One of the primers has a cleavable site



Hybridize Fragment & Extend

Single DNA libraries are hybridized to primer lawn

Bound libraries are then extended by polymerases

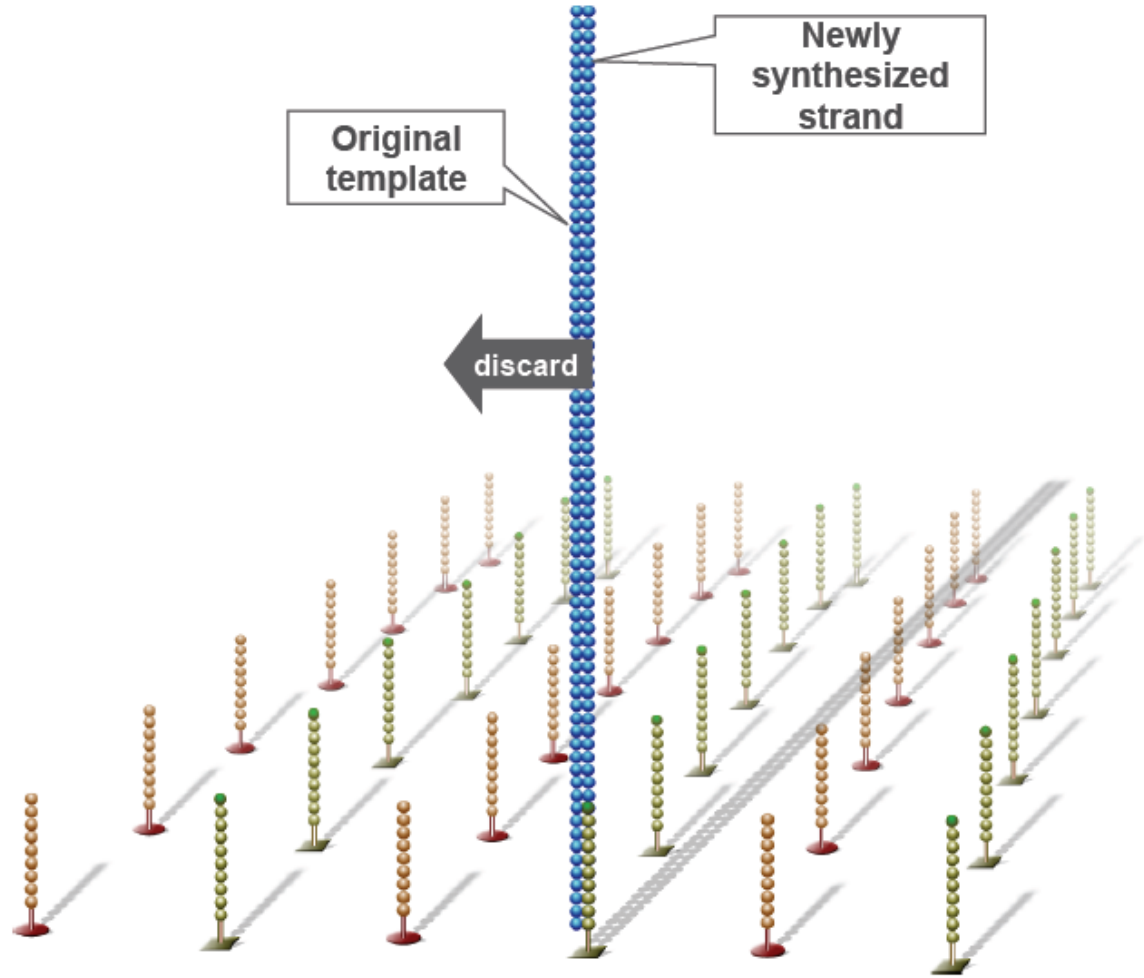


Denature Double-Stranded DNA

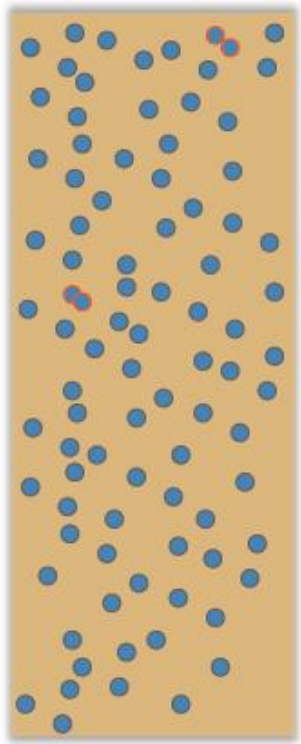
Double-stranded molecule is denatured

Original template washed away

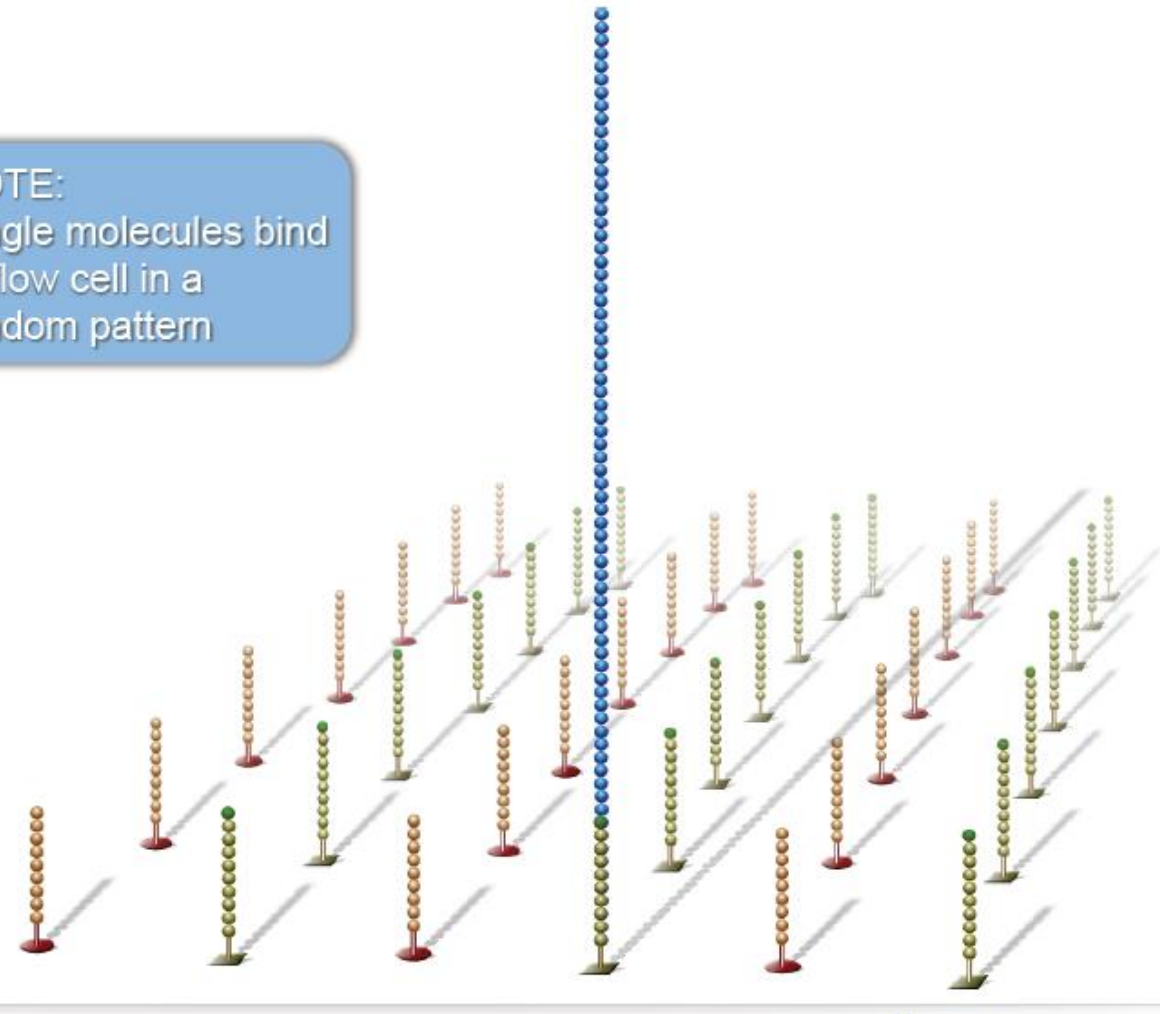
Newly synthesized strand is covalently attached to flow cell surface



Single-Stranded DNA



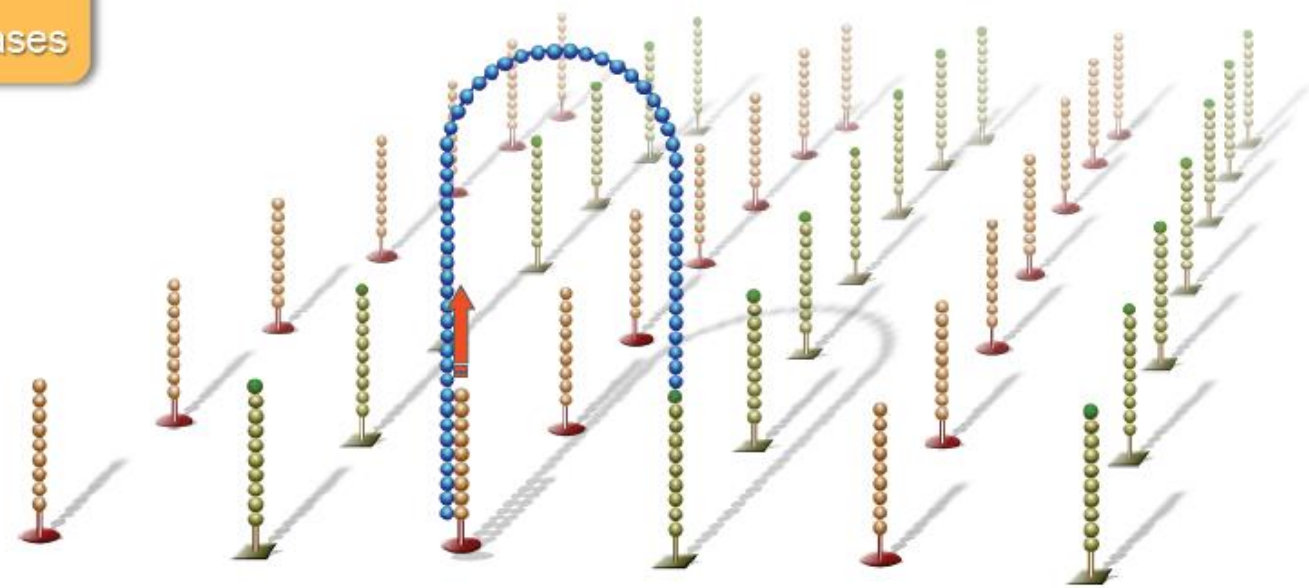
NOTE:
Single molecules bind to flow cell in a random pattern



Bridge Amplification

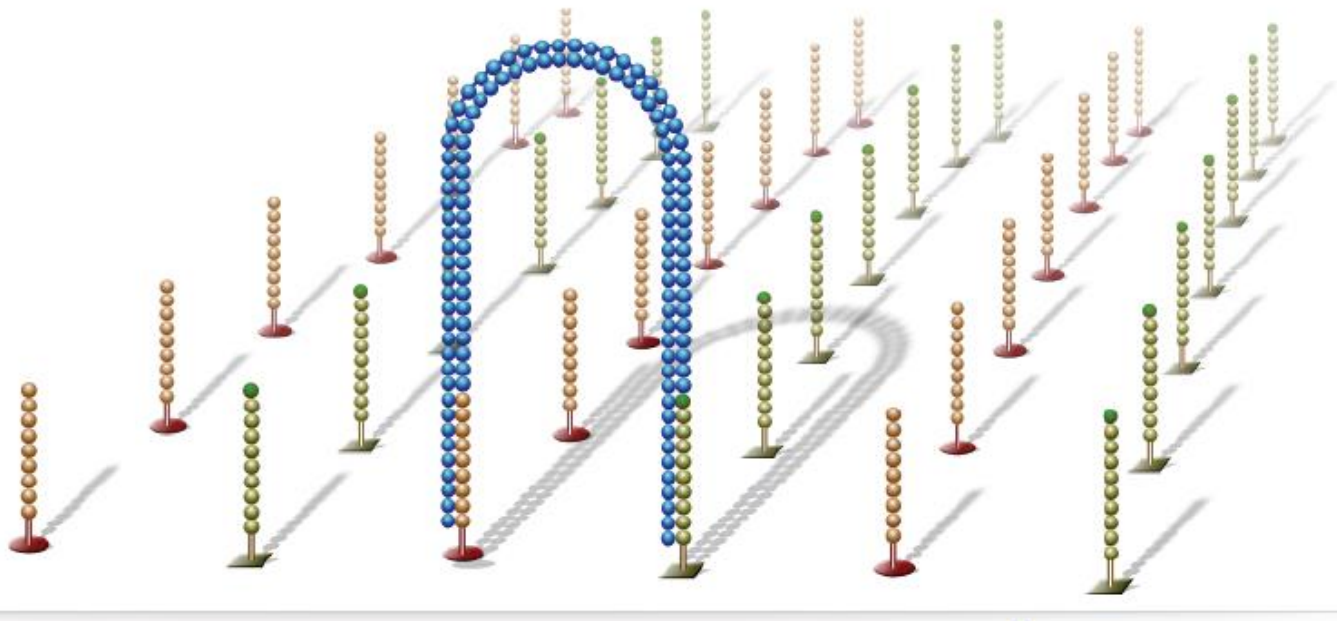
Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



Bridge Amplification

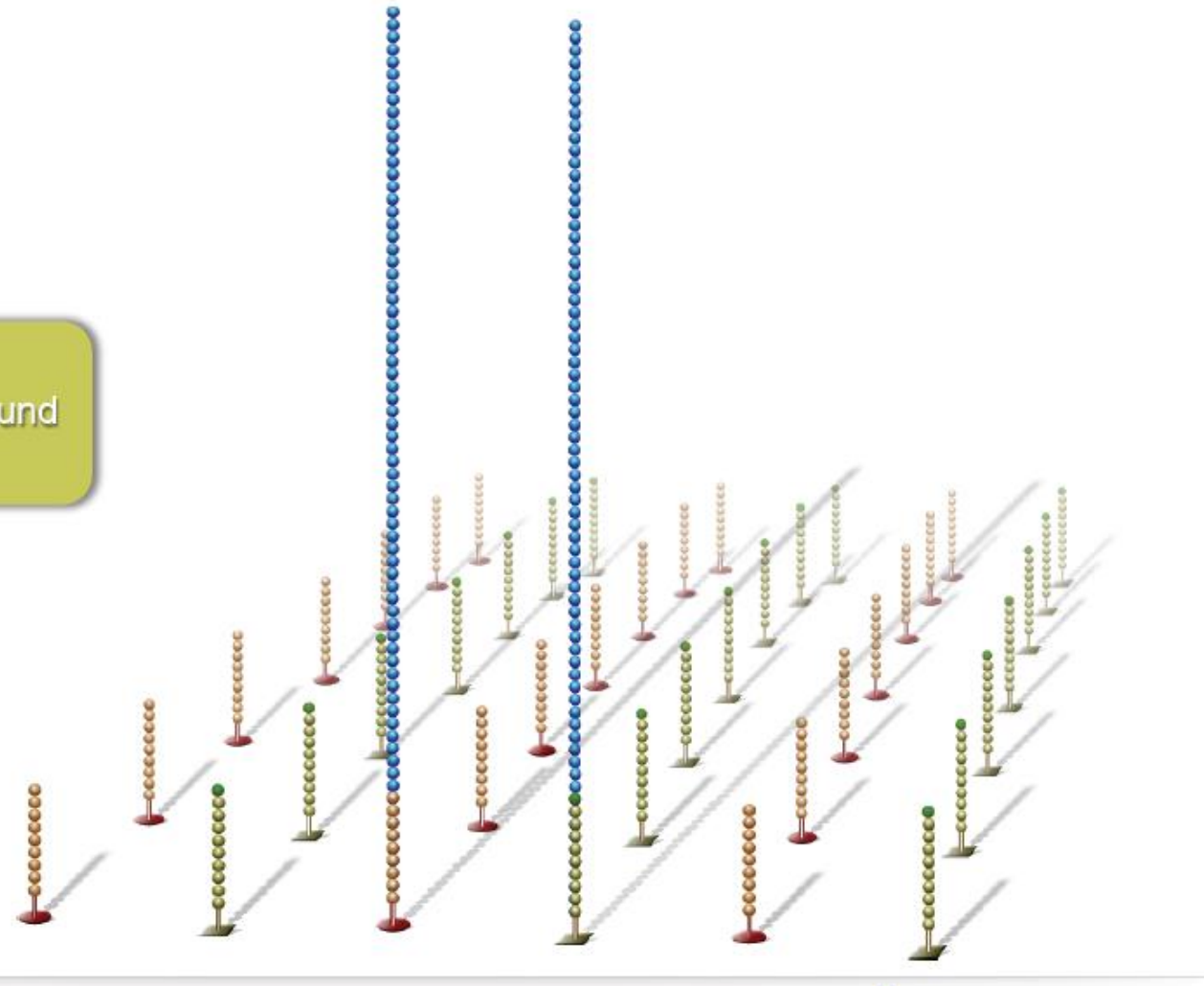
Double-stranded bridge is formed



Denature Double-Stranded Bridge

Double-stranded bridge is denatured

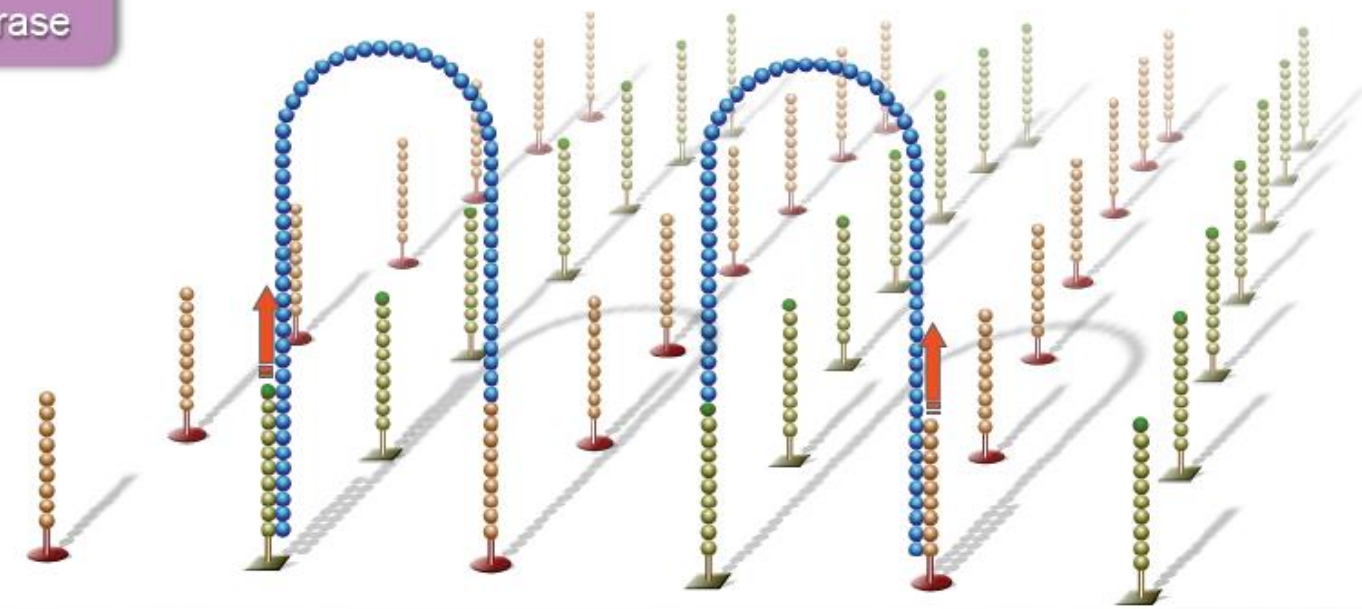
Result:
Two copies of covalently bound single-stranded templates



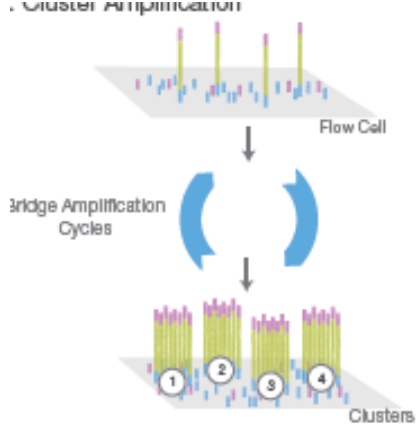
Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

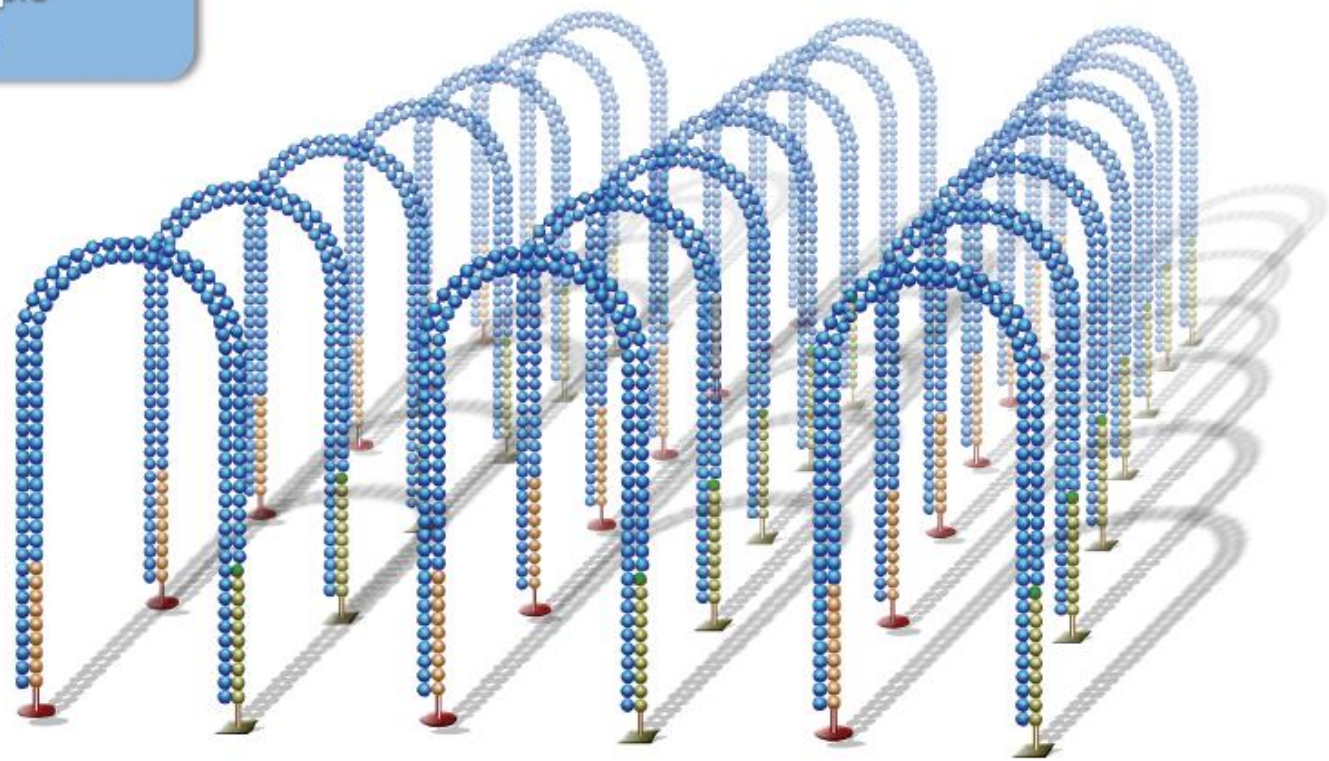
Hybridized primer is extended by polymerase



Bridge Amplification

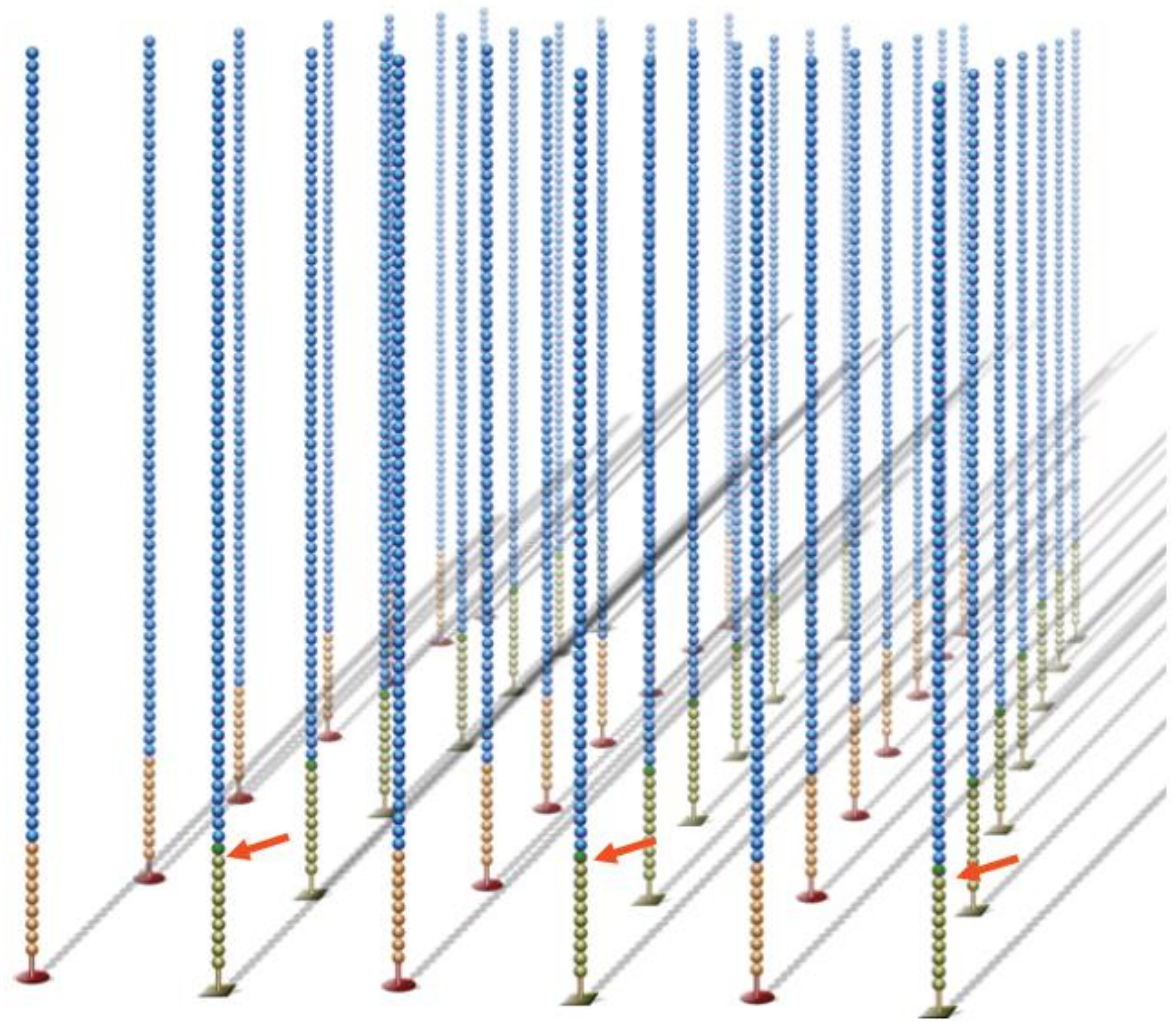


Bridge amplification cycle is repeated until multiple bridges are formed



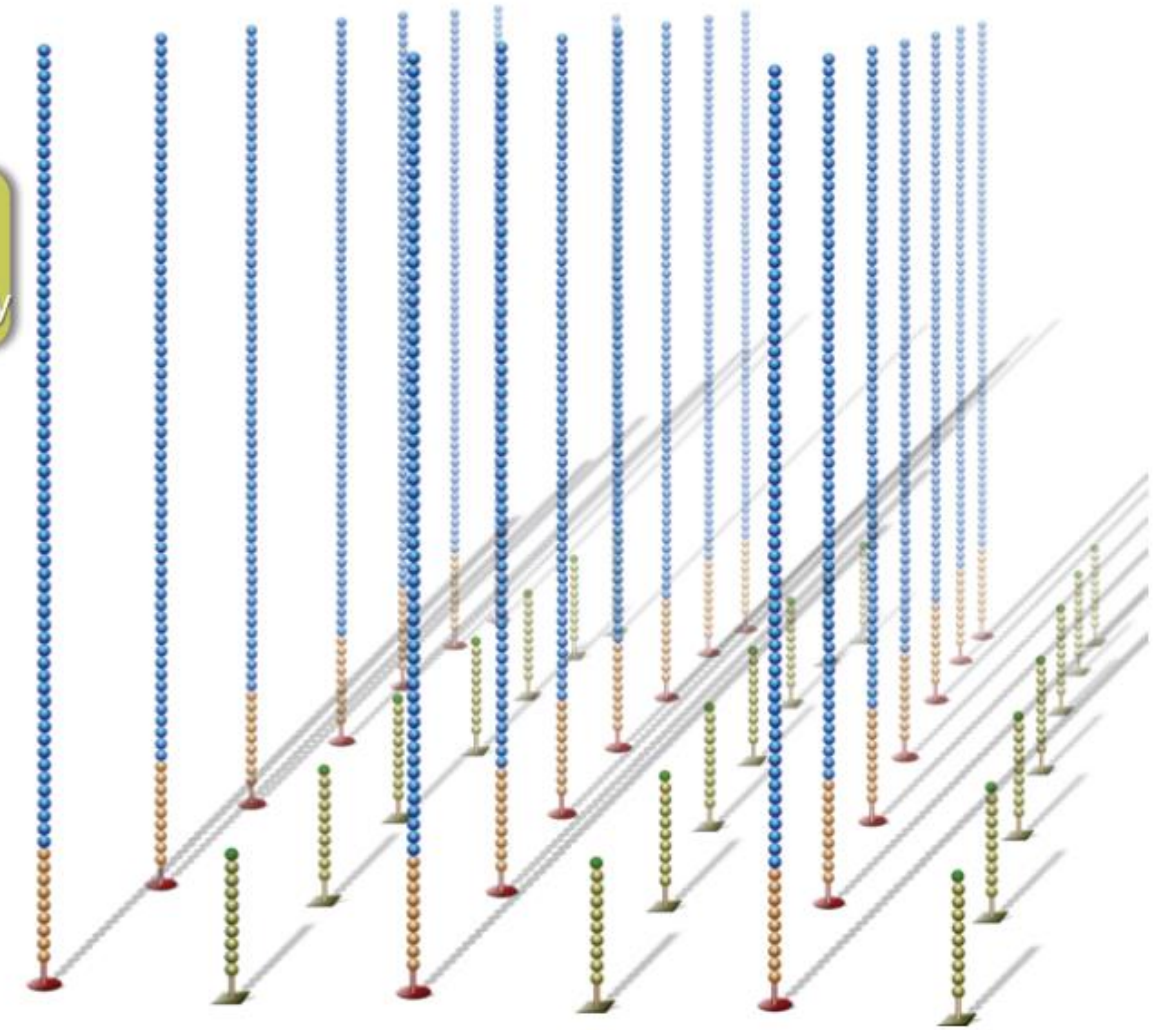
Linearization

dsDNA bridges are denatured



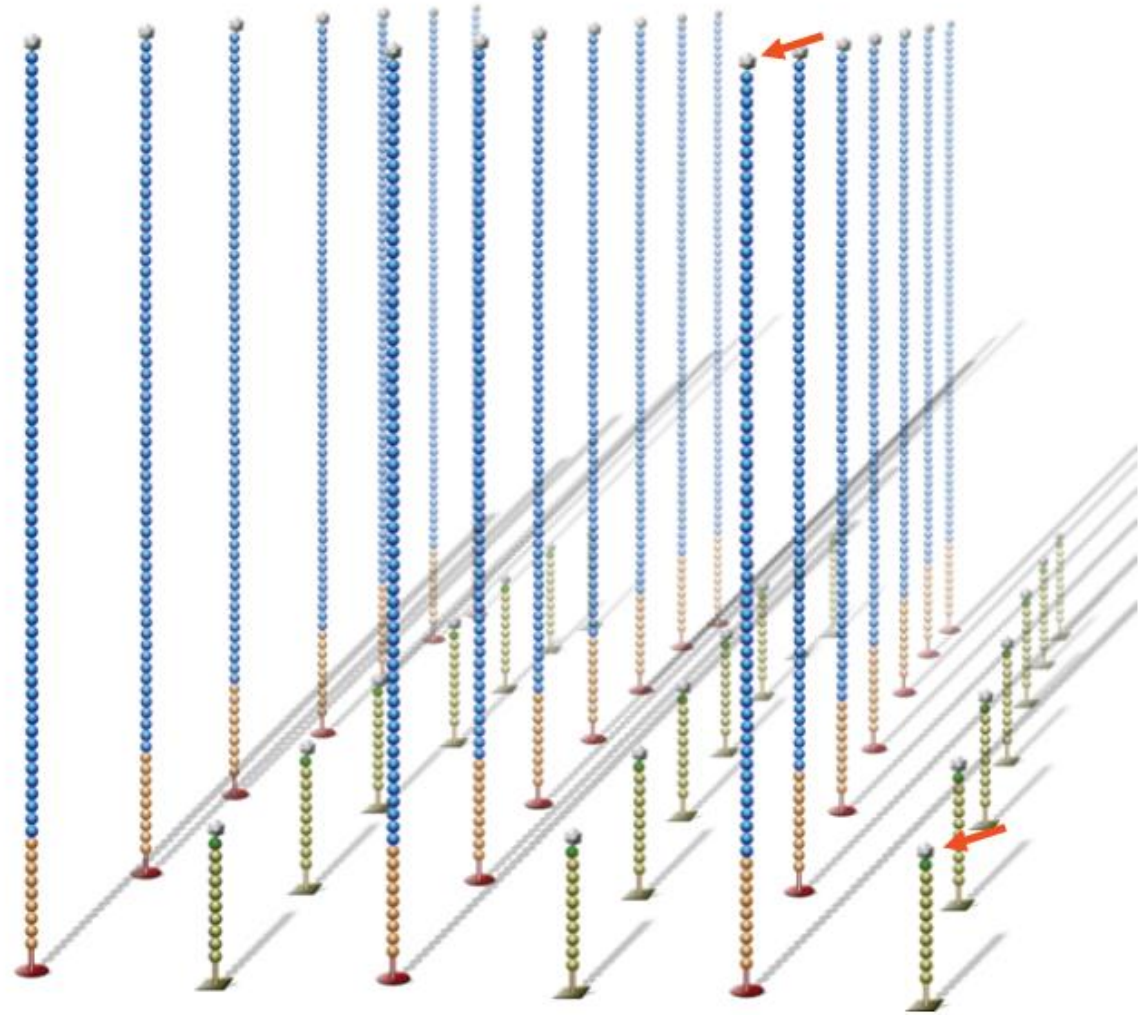
Reverse Strand Cleavage

Reverse strands are cleaved and washed away, leaving a cluster with forward strands only



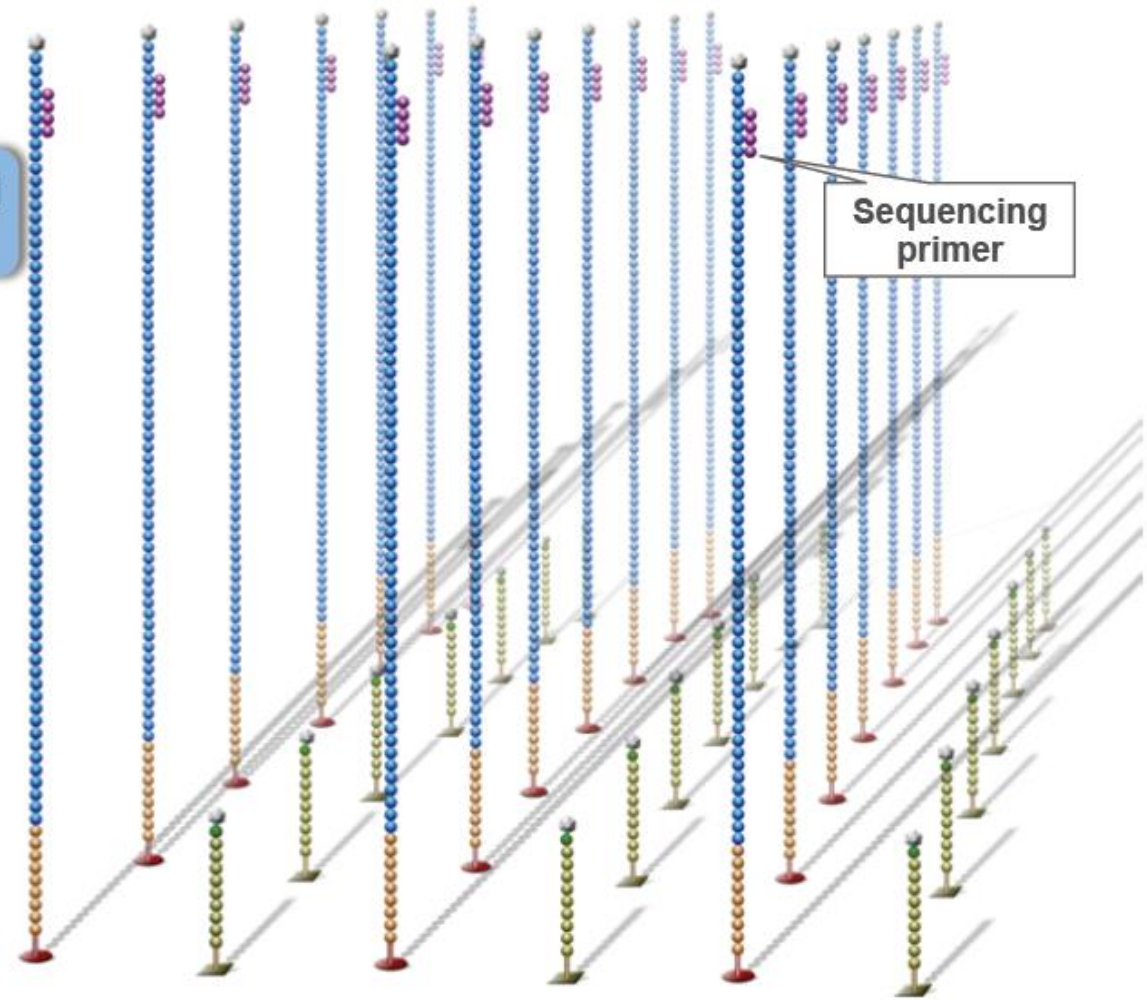
Blocking

Free 3’ ends are blocked to prevent unwanted DNA priming



Read 1 Primer Hybridization

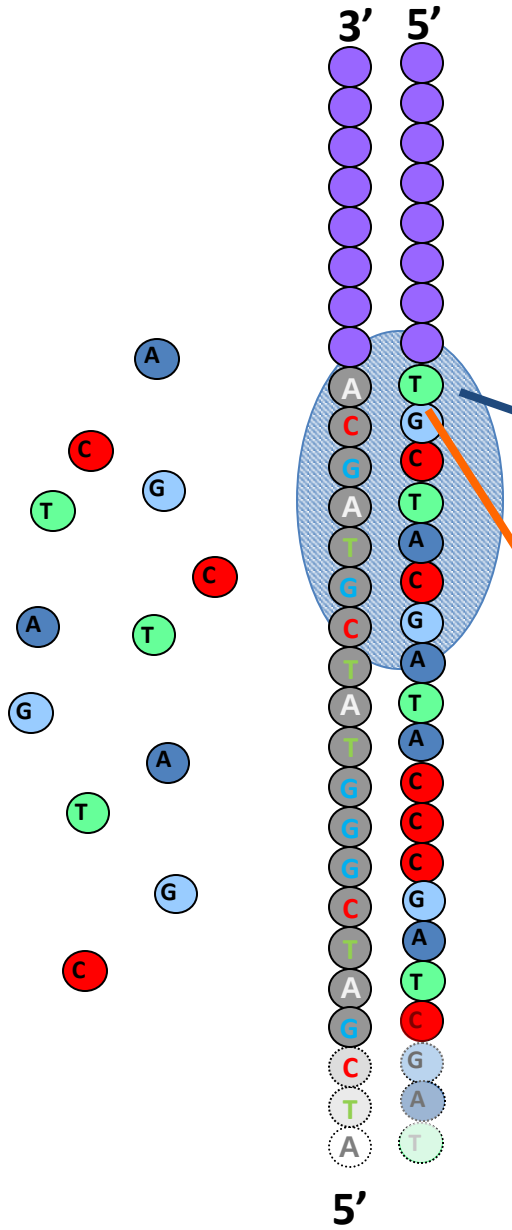
Sequencing primer is hybridized to adapter sequence



MUNI FI

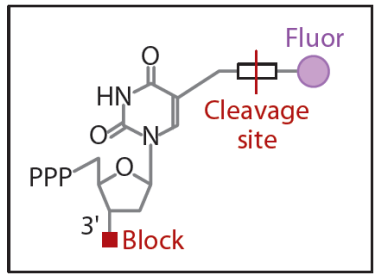
MUNI | RECETOX

Sequencing by synthesis



Cycle 1: Add sequencing reagents

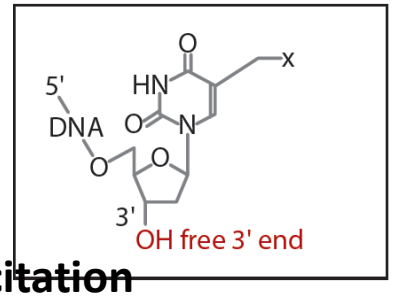
First base incorporated



Emission

Detect signal

Cleave terminator and dye

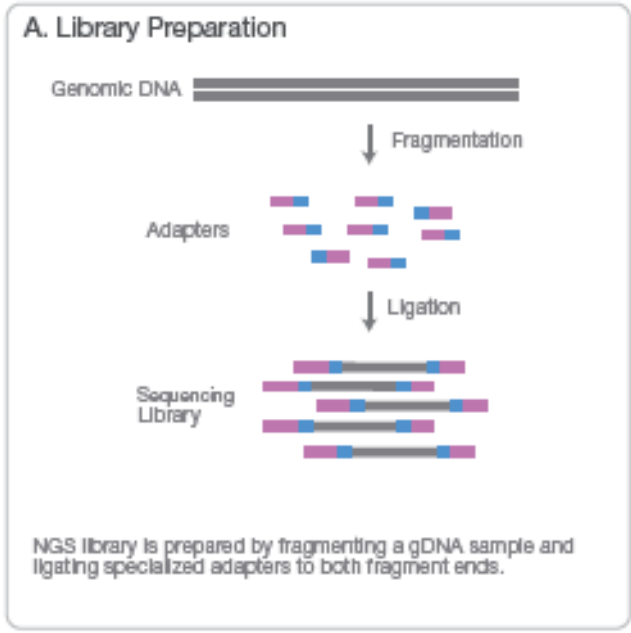


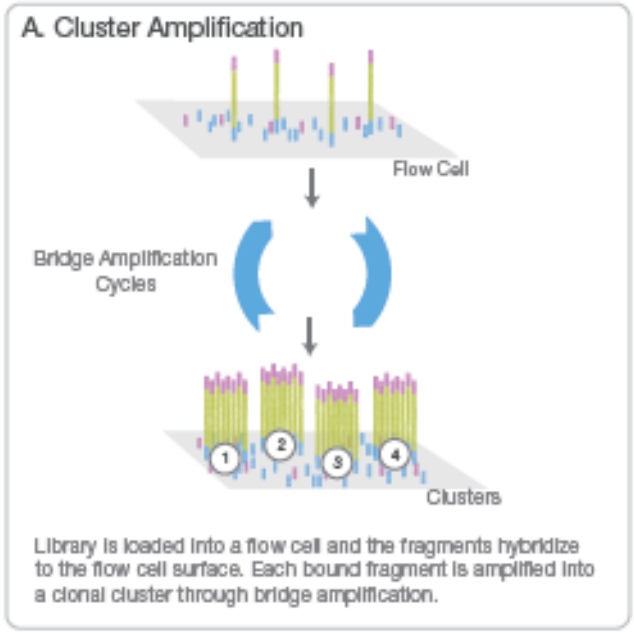
Excitation

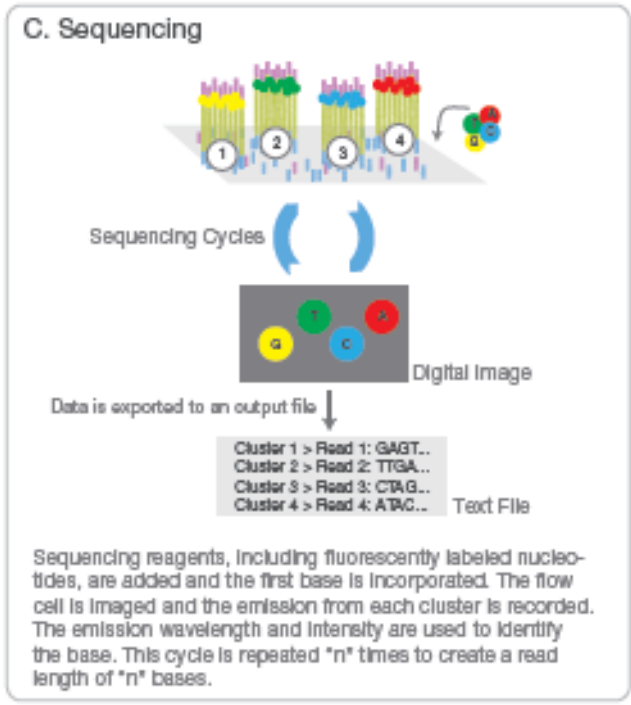
Cycle 2-n: Add sequencing reagents and repeat

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.





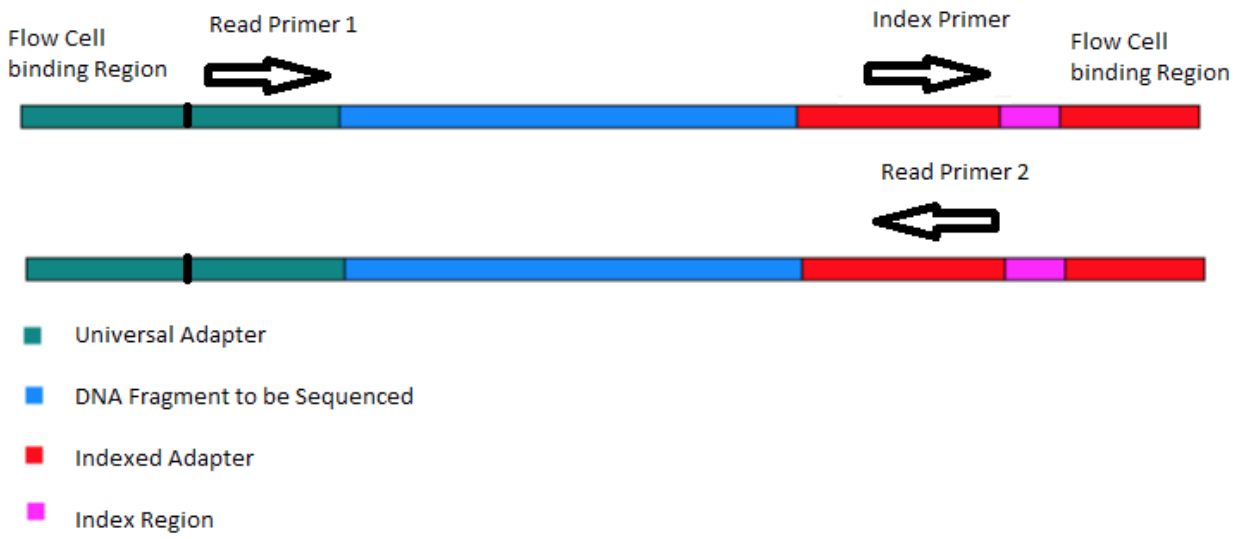


Lets see a video

https://www.youtube.com/watch?v=womKfikWlxM

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

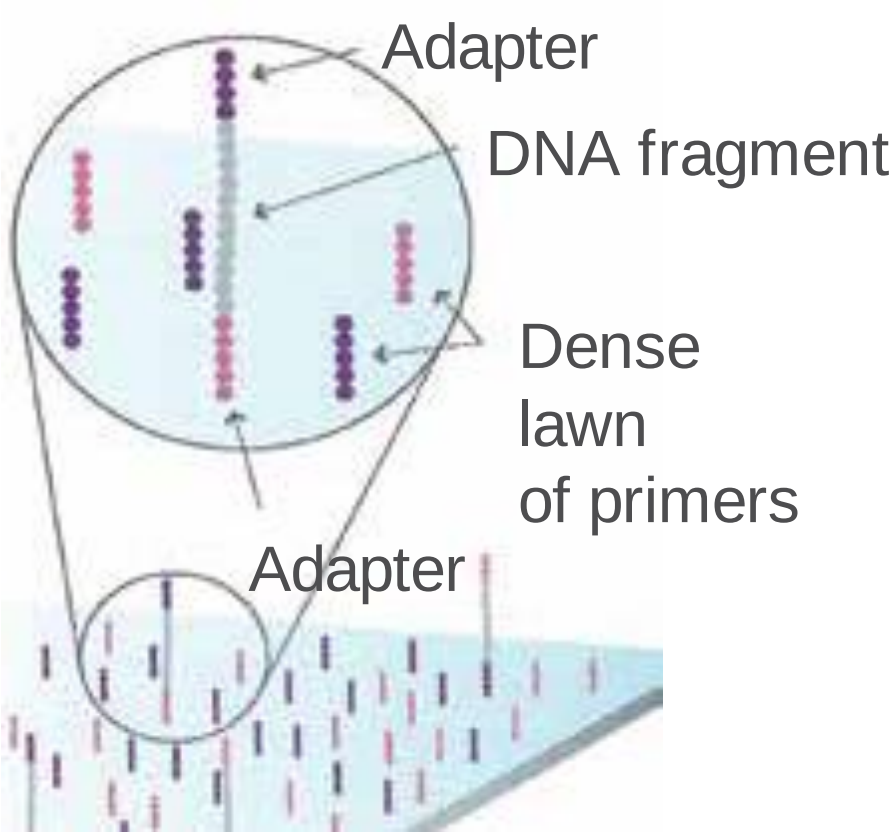
Sources of errors: PCR duplicates

- 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!

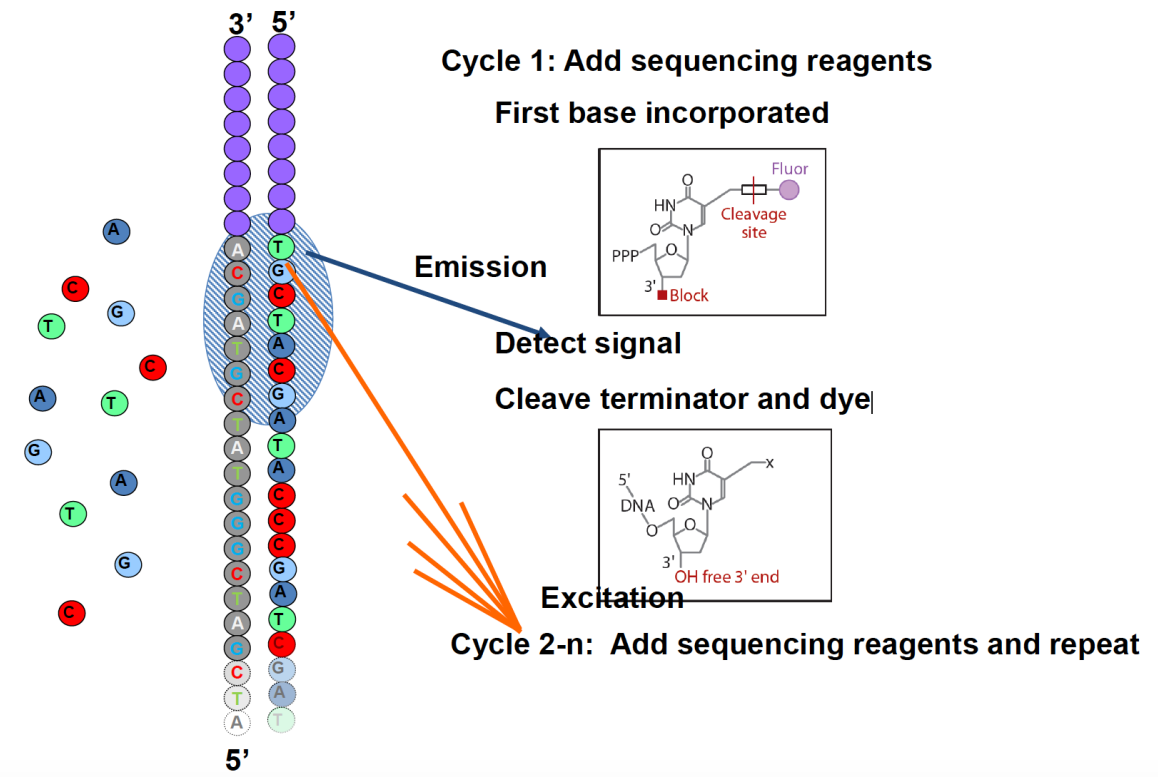
Higher rates of PCR duplicates e.g. 30% arise when you have too little starting material such that greater amplification of the library is needed in step 3, or when you have too great a variance in fragment size, such that smaller fragments, which are easier to PCR amplify, end up over-represented.



Sources of errors: sequencing by synthesis – the fluorescence

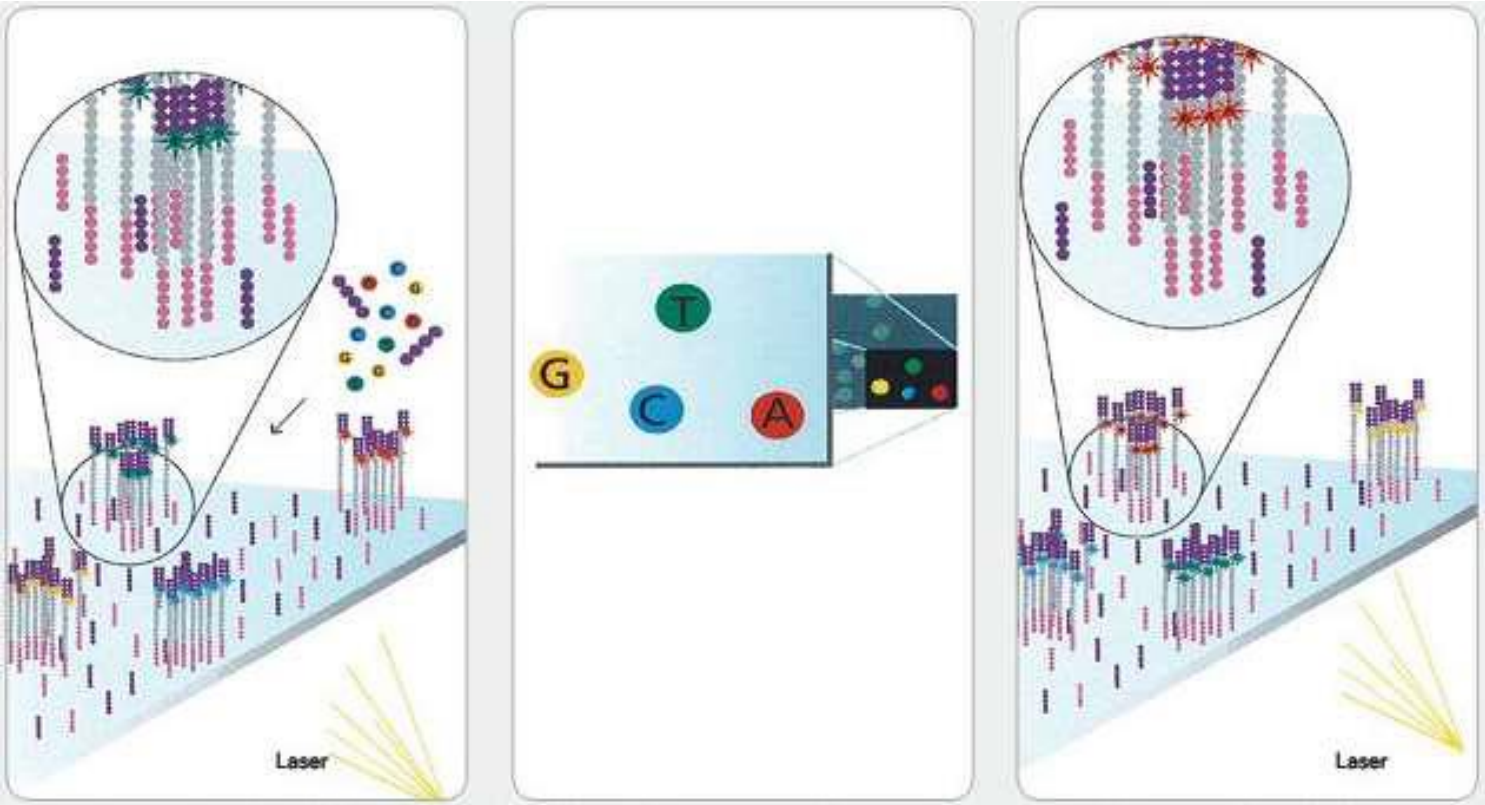
- In step 5 we amplify the signal and detect the fluorescence of each base
- The assumption is that in a cycle, every molecule on the flowcell is extended by one base

- 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



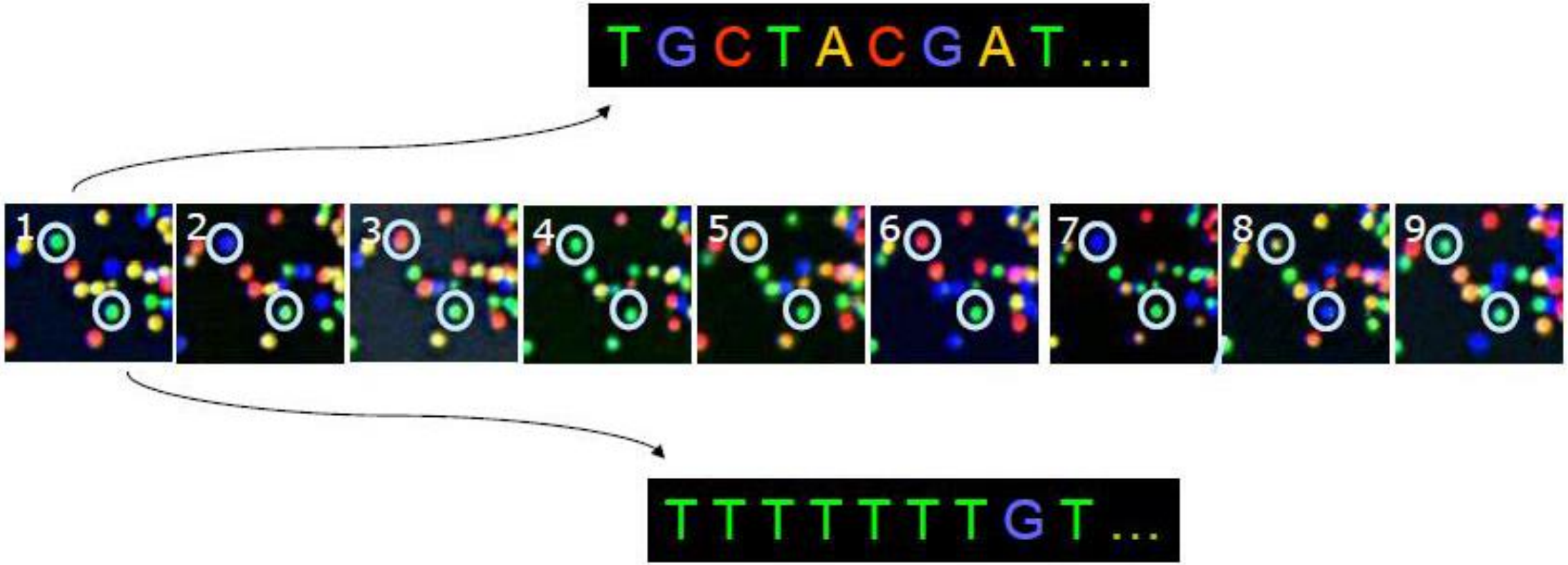
Sequencing by Synthesis - Fluorescently labeled Nucleotides (Illumina)

- During the process, **clusters of same sequences** are created

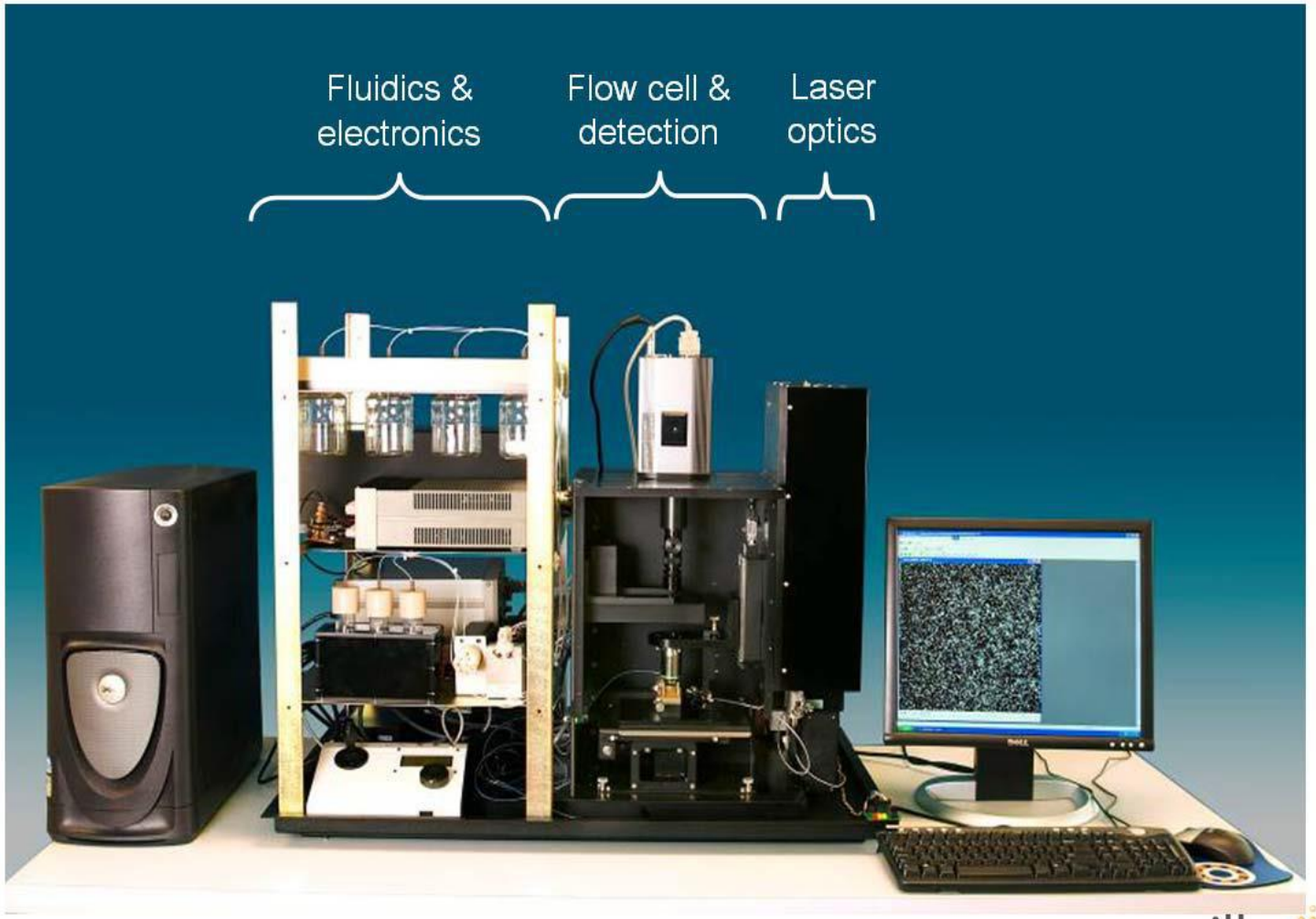


Step 0: base calling (image analysis)

- The identity of each base of a cluster is read off from **sequential images**
- One cycle -> one image

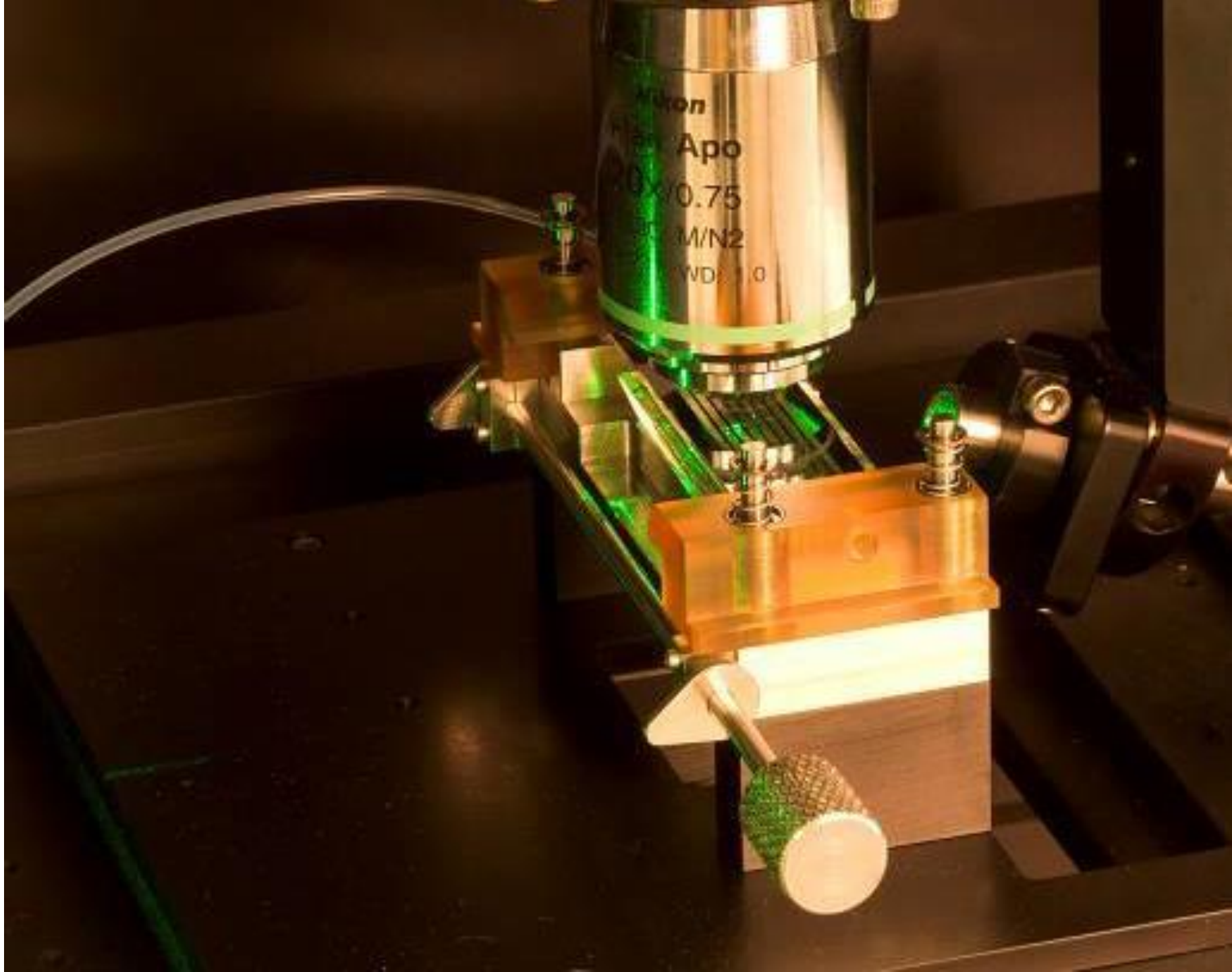


Instrument without Covers

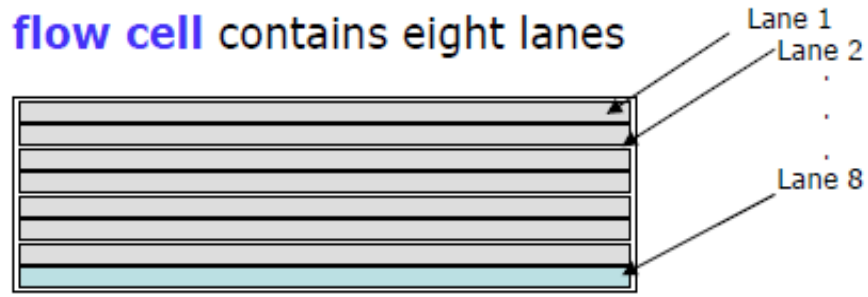


illumina

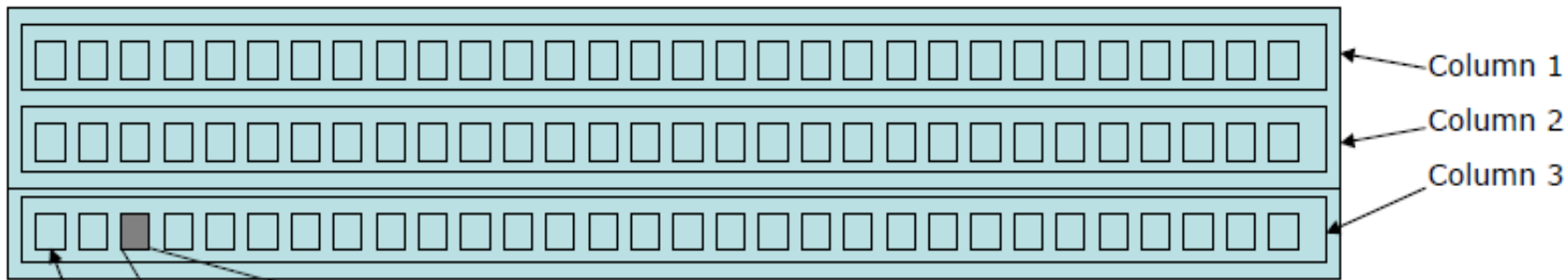
Flow-cell imaging



A **flow cell** contains eight lanes

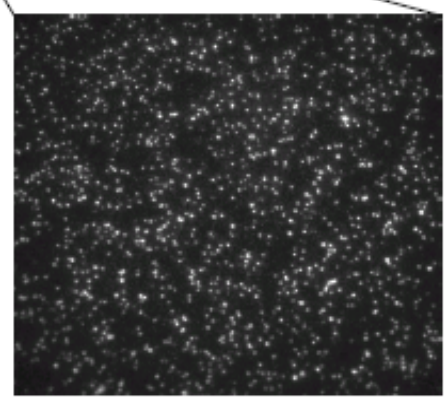


Each **lane/channel** contains **three columns** of tiles



Each **column** contains **100 tiles**

Tile



20K-30K Clusters

350 X 350 µm

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

345,600 images for a 36-cycle run

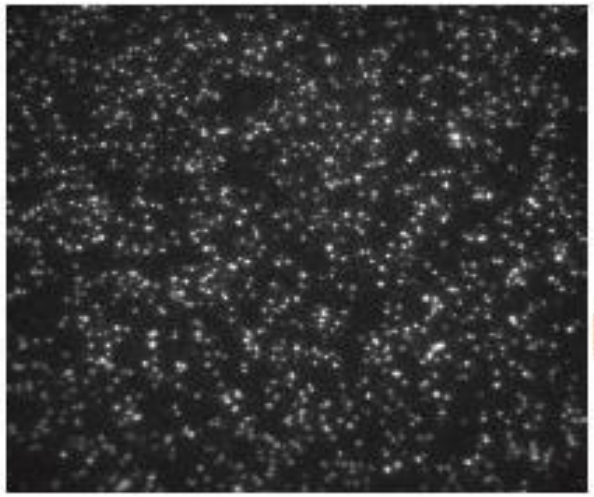
MUNI FI

RECETOX

Getting the sequences from clusters

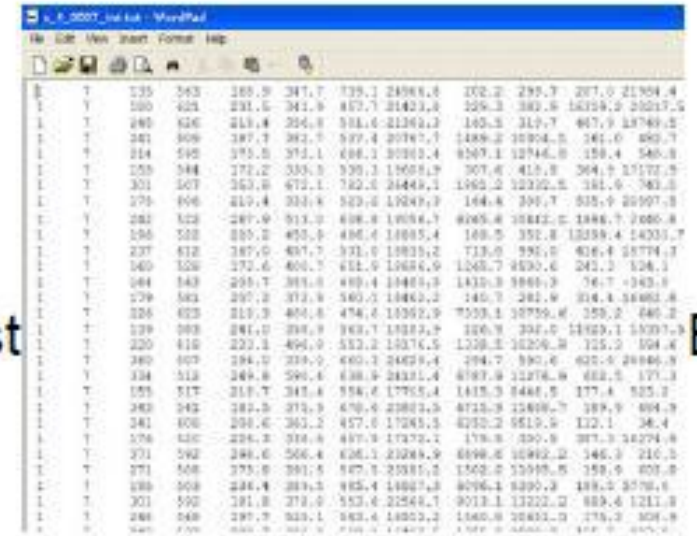
– Illumina pipeline

Firecrest (image analysis)
Locates clusters and calculates intensity and noise



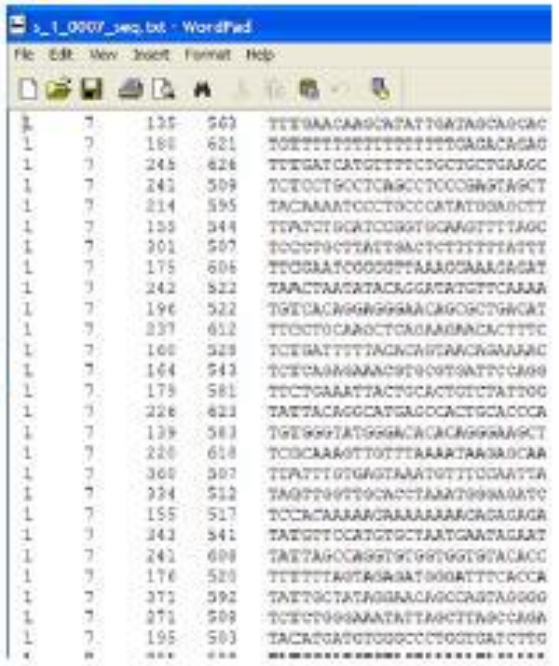
tiff image files
(345,600)

Firecrest



intensity files

Bustard



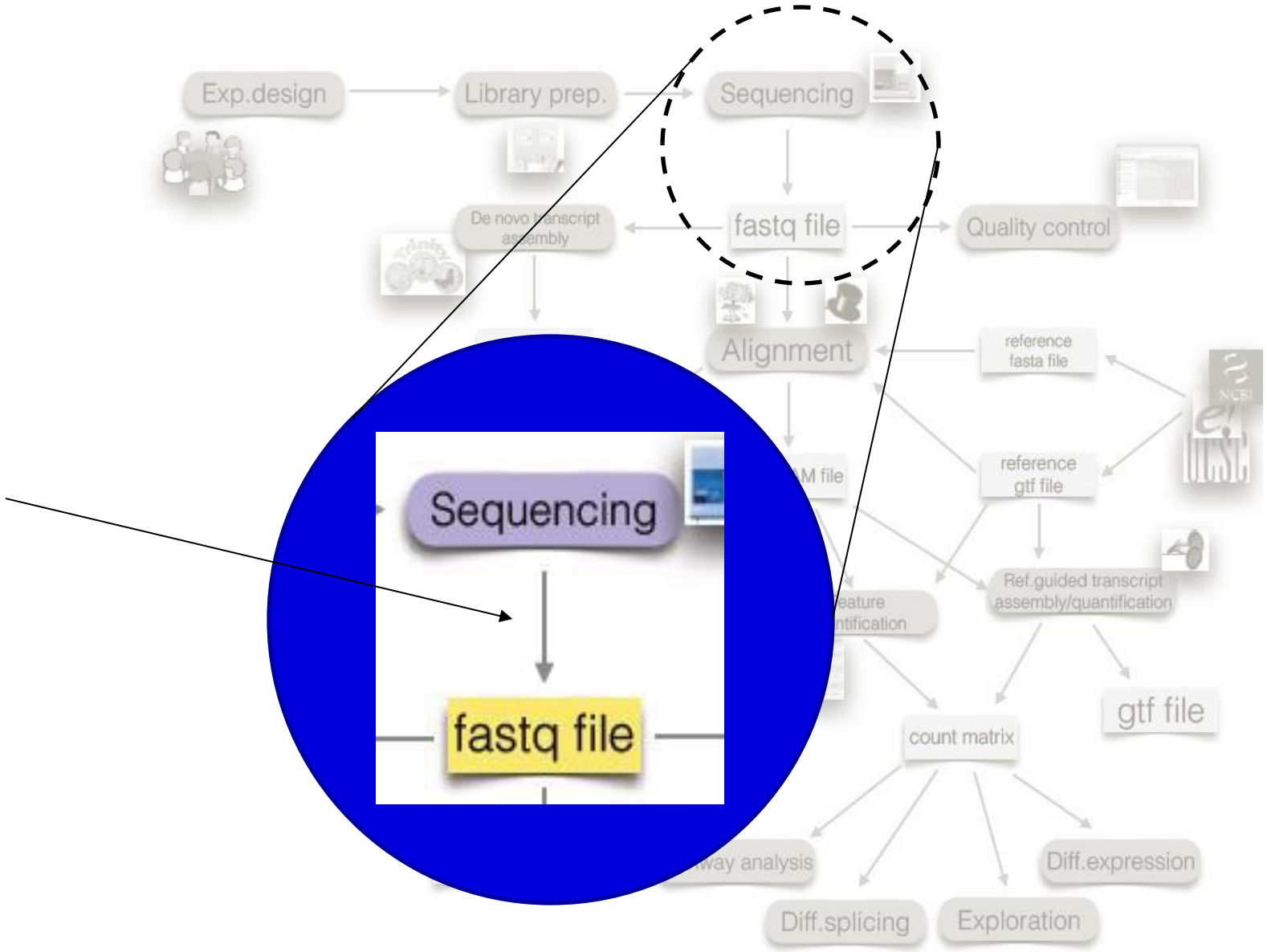
Sequence files

Bustard (base calling)
Deconvolutes signal and corrects for cross-talk, phasing

Image analysis data output

- 100 tiles per lane, 8 lanes per flow cell, 36 cycles
- 4 images (A,G,C,T) per tile per cycle = 115,200 images
- Each tiff image is ~ 7 MB = 806,400 MB of data
- 1.6 TB per 70 nt read, 3.2 TB for 70 nt paired-end read
- Most technologies are erasing intensities as they are sequencing, because of a too high amount of data

Step 0: base calling (image analysis) + base quality control

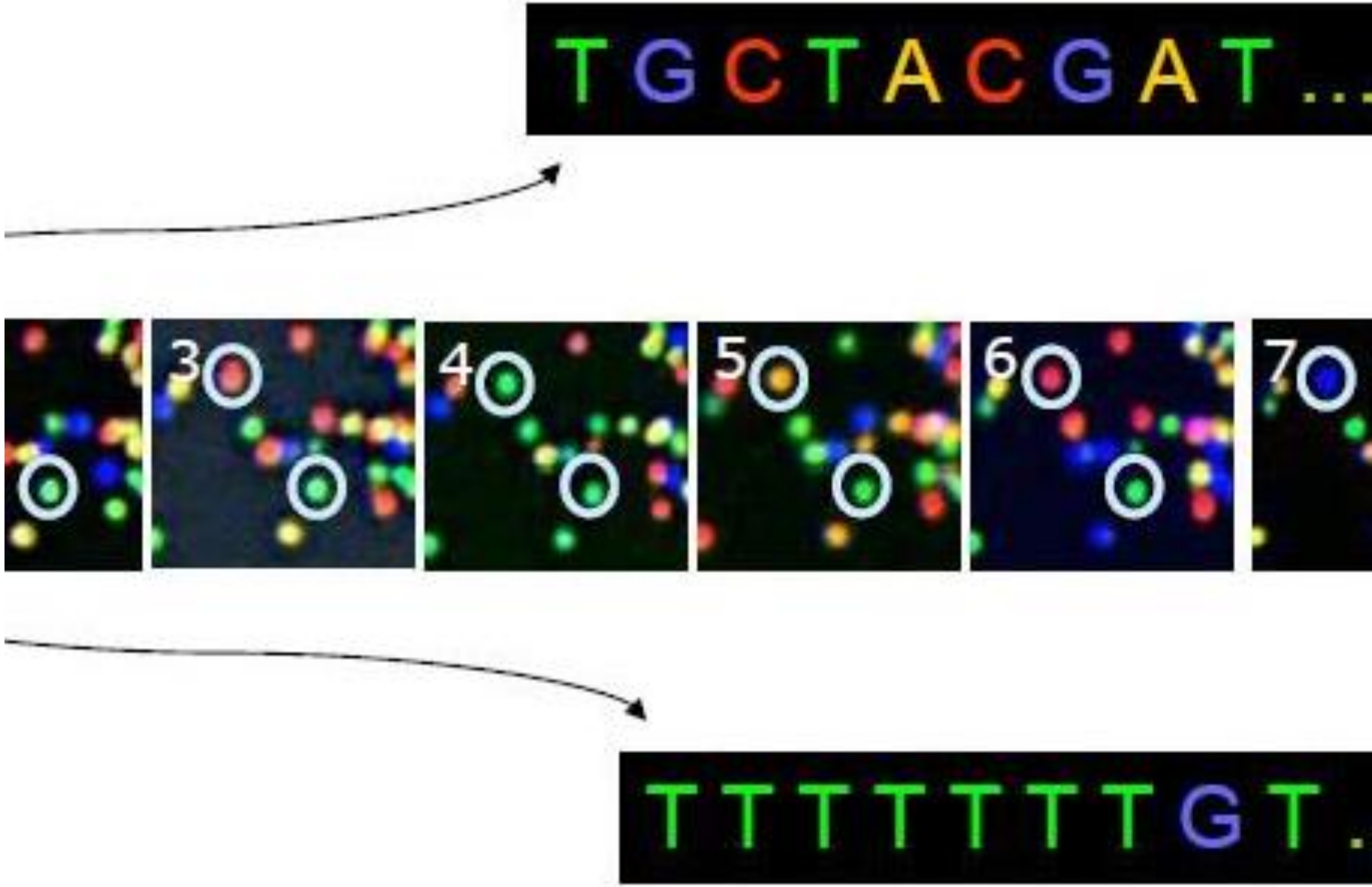


MUNI FI

MUNI | RECETOX

Base call quality control

- Quality control (QC) of each base call is automatically performed by the sequencing platform
- In other words: *For each letter in a read, we estimate the probability of it being erroneous (P).*
- QC per base is specialized for each platform – each platform must solve challenges unique to the underlying sequencing technology



The PHRED score

$$Q_{phred} = -10 \times \log_{10} P(\textbf{error})$$

- The *Phred* quality score is the negative ratio of the error probability to the reference level of $P = 1$ expressed in Decibel (dB).
- The **error estimate** is based on **statistical model** providing measure of **certainty** of each base call in addition to the nucleotide itself
- These statistical models base their error estimate on:
  - Signal intensities from the recorded image
  - Number of the sequencing cycle
  - Distance to other sequence colonies
- *Phred* score is recoded using ASCII in fastq file

Phred score	Probability of incorrect base call	Base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%

Phred score encoding in ASCII

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

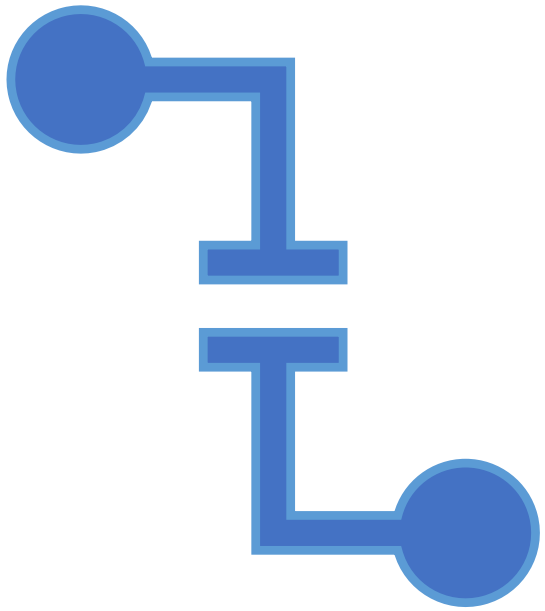
```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  .............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  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........................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)
```

MUNI FI

RECETOX



FASTA and FASTQ formats

- The reads obtained from the sequencer are typically stored in **fasta** (just the sequences) or **fastq** (sequences + QC measure) format files.
- For **paired-end** reads, we usually obtain **two files**.
- **Reads** are *not* generally grouped by strand, only **by the order in which they were sequenced**.

FASTA format

- General format to represent sequences
- **Two lines per sequence** (read)
  - ID line (starting with >)
  - Sequence line
- Typical file extension: `.fa` or `.fasta`

```
>HWI-ST132:633:D17U2ACXX:8:1101:14830:2376 1:N:0:GATCAG
CTCAGACCGCGTTCTCTCCCTCTCACTCCCCAATACGGAGAGAAAAACGA
```

- HWI-ST132 - unique instrument name
- 633 - run ID
- D17U2ACXX - flowcell ID
- 8 - flowcell lane
- 1101 - tile number within lane
- 14830 - x-coordinate of cluster within tile
- 2376 - y-coordinate of cluster within tile
- 1 - member of pair (1 or 2). Older versions: /1 and /2
- Y/N - whether the read failed quality control (Y = bad)
- 0 - none of the control bits are on
- CATGCA - index sequence (barcode)

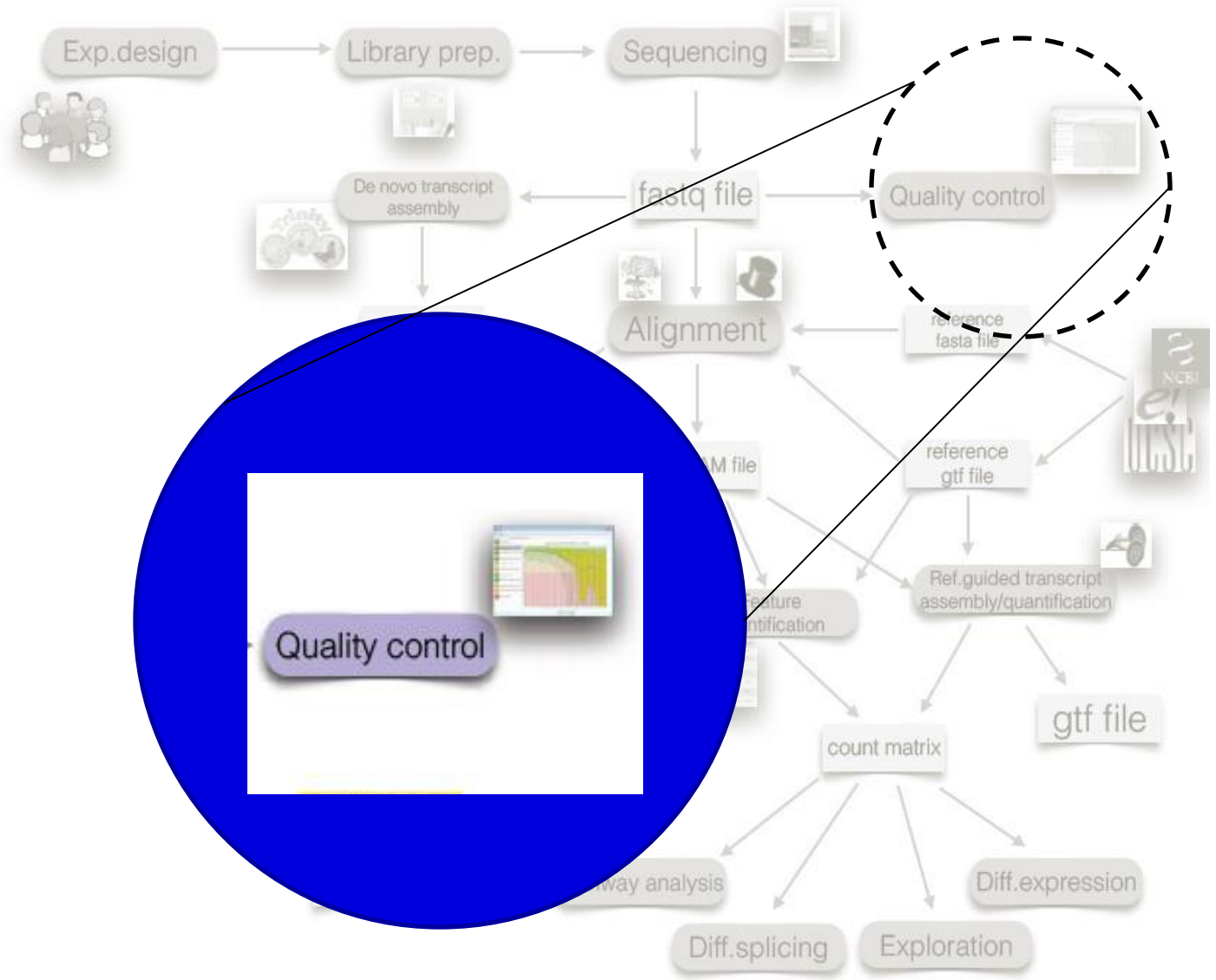
FASTQ format

- Combines sequence and base call quality information.
- Typical file extension: `.fastq`

```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

- Four lines per sequence (read):
  - ID (starting with @)
  - Sequence line
  - Another ID line (starting with +)
  - Base qualities (one for each letter in the sequence)

Step 1: Read quality control and data filtering



MUNI FI

MUNI | RECETOX

Step 1: Read quality control and data filtering

- Uses the output file with information about the quality of base calls (.fastq)
- First step in the pipeline that **deals with actual sequencing data** in base or color space

- Several metrics are evaluated, not all of them use the Phred score information:
  - Distribution of quality scores at each sequence, Sequence composition, Per-sequence and per-read distribution of GC content, Library complexity, Overrepresented sequences
- Initial overview – already in base calling SW
- More quality overview – SW solutions `SolexaQA, FastQC`

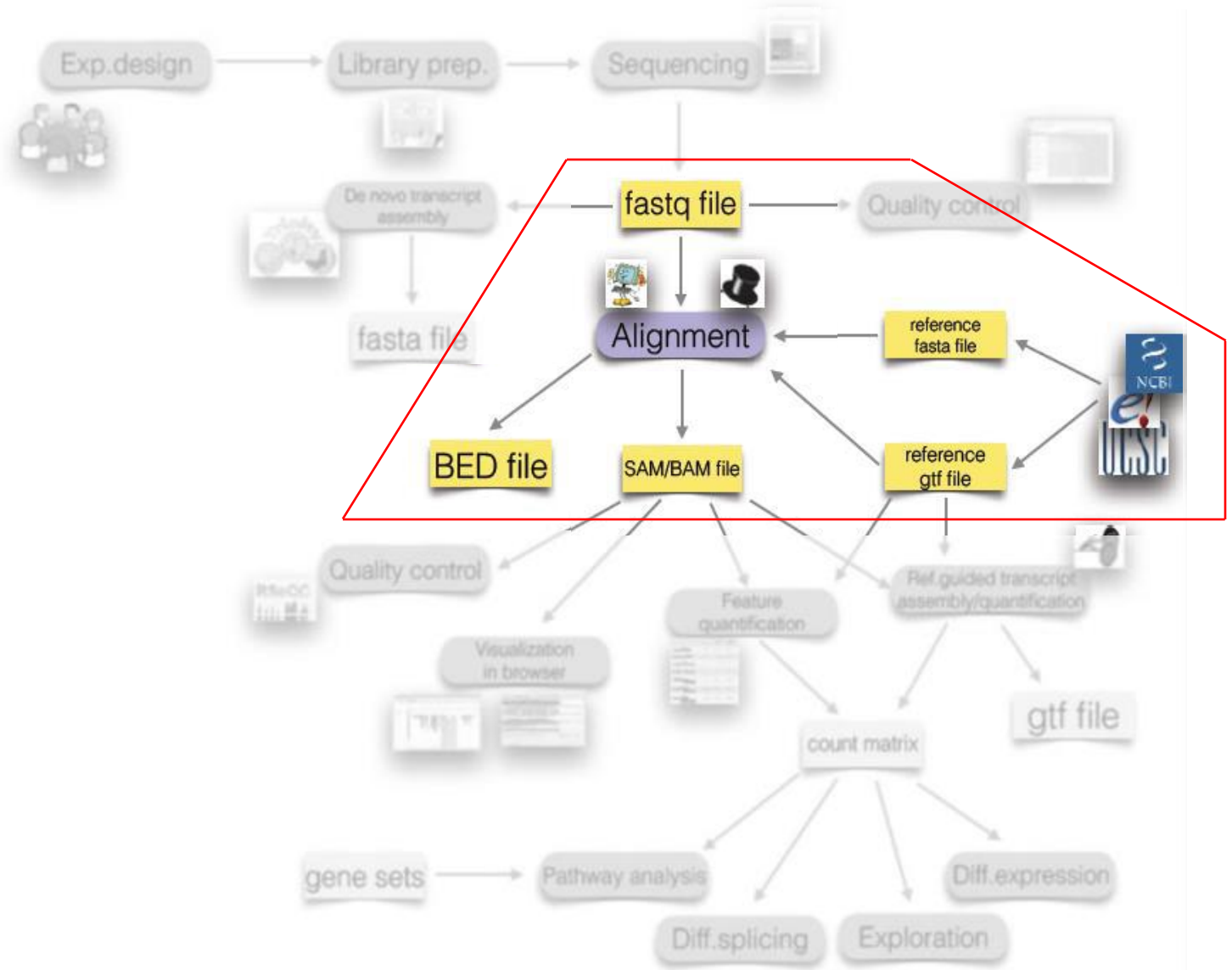
Step 1: Read quality control and data filtering

– Based on the quality measures, we decide to remove low quality bases and reads

- **Trimming** – removes low quality or unwanted bases from reads, thus shortening them. Is applied to increase the number of mappable reads.

- **Filtering –** removes whole reads that do not meet quality standards (e.g. too short etc)

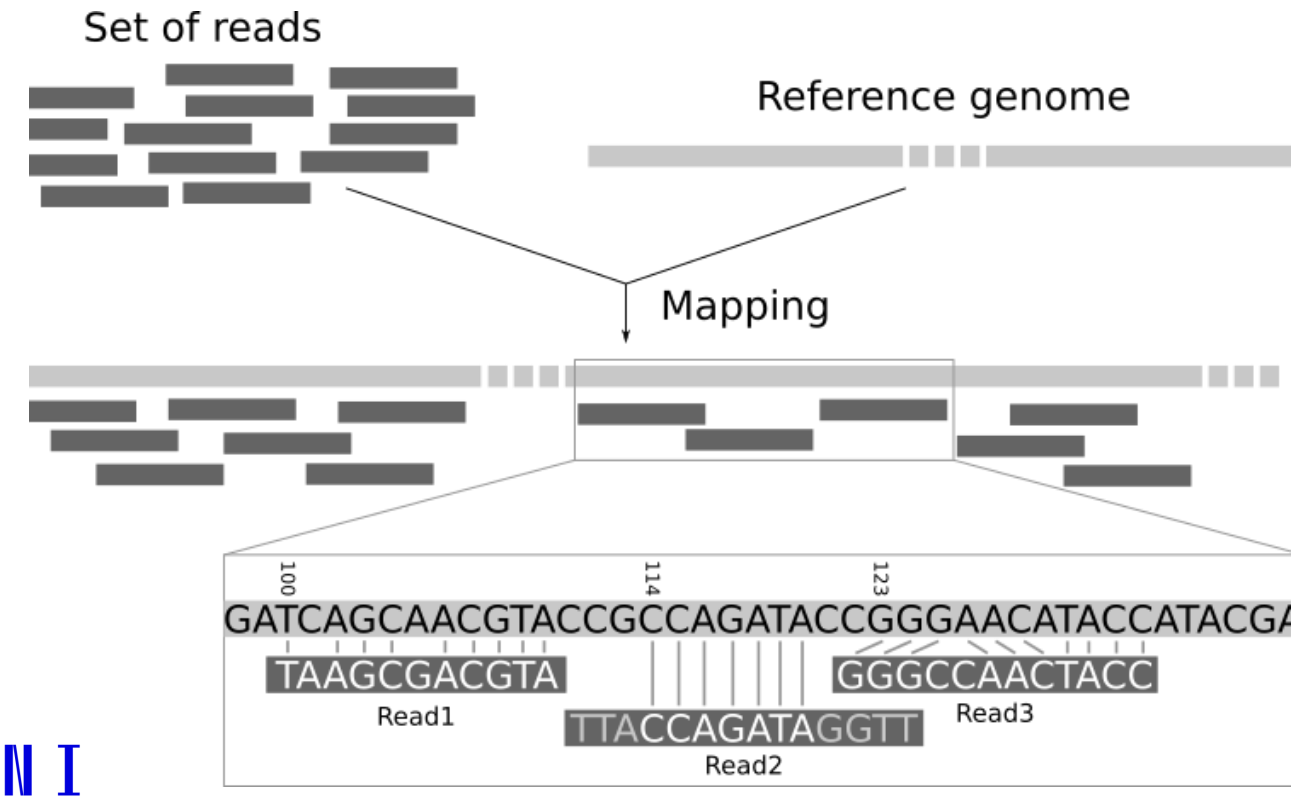
Step 2: Alignment (mapping)



MUNI
FI

MUNI | RECETOX

Step 2: Alignment (mapping)



- To know, where the **short reads** (in our filtered .fastq file) come from (which part of the genome or transcriptome do they represent) they need to be (in most instances) aligned to a **reference sequence**

Reference sequence

- The reference sequence can be a genome, a transcriptome or a collection of specific sequences.
- Typically, the reference sequence(s) is given in a `.fa` or `.fasta` file
- An alternative is the GTF (gene transfer format) - stores gene structure
- BED format (designed for annotation tracks in genomic browsers)

(we will learn about where to get the reference sequences in one of the next lectures)

Step 2: Alignment (mapping)

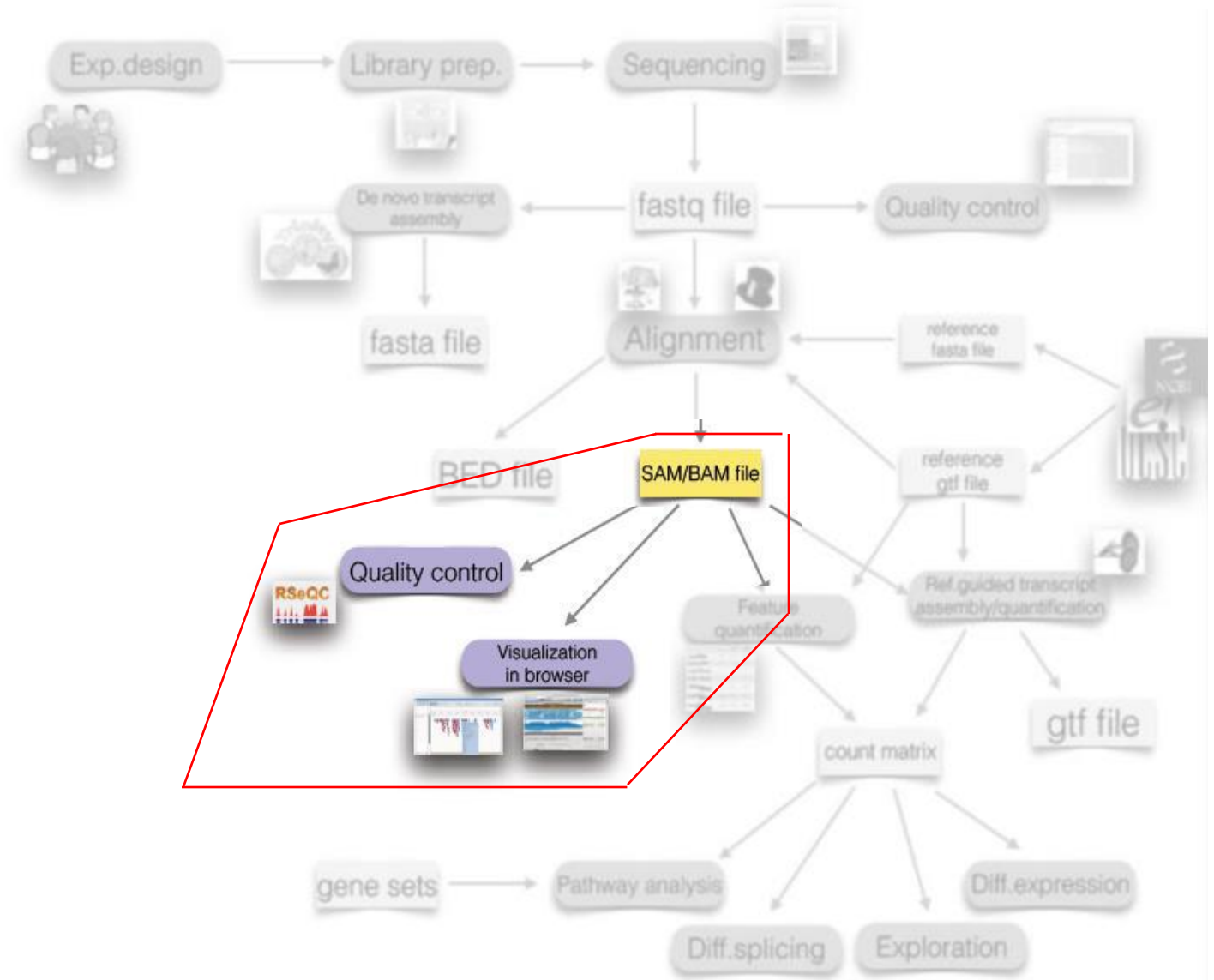
```
GTGCTCGCTGACACAGAAAGTTCGGCA
CTCAGACA
11111111
```

- Intuitively an easy task
- However, trying all the possible options (alignments), is very time consuming!
- Efficient algorithms (**aligners**) exist

- The result of mapping is stored by many algorithms in the **Sequence alignment/map (SAM) format**

- We will talk about mapping a in one of the future lectures

Step 3: Post-alignment QC and visualization



MUNI
FI

MUNI | RECETOX

Step 3: Post-alignment QC and visualization

– Necessary in order to see the **efficiency of the alignment**.

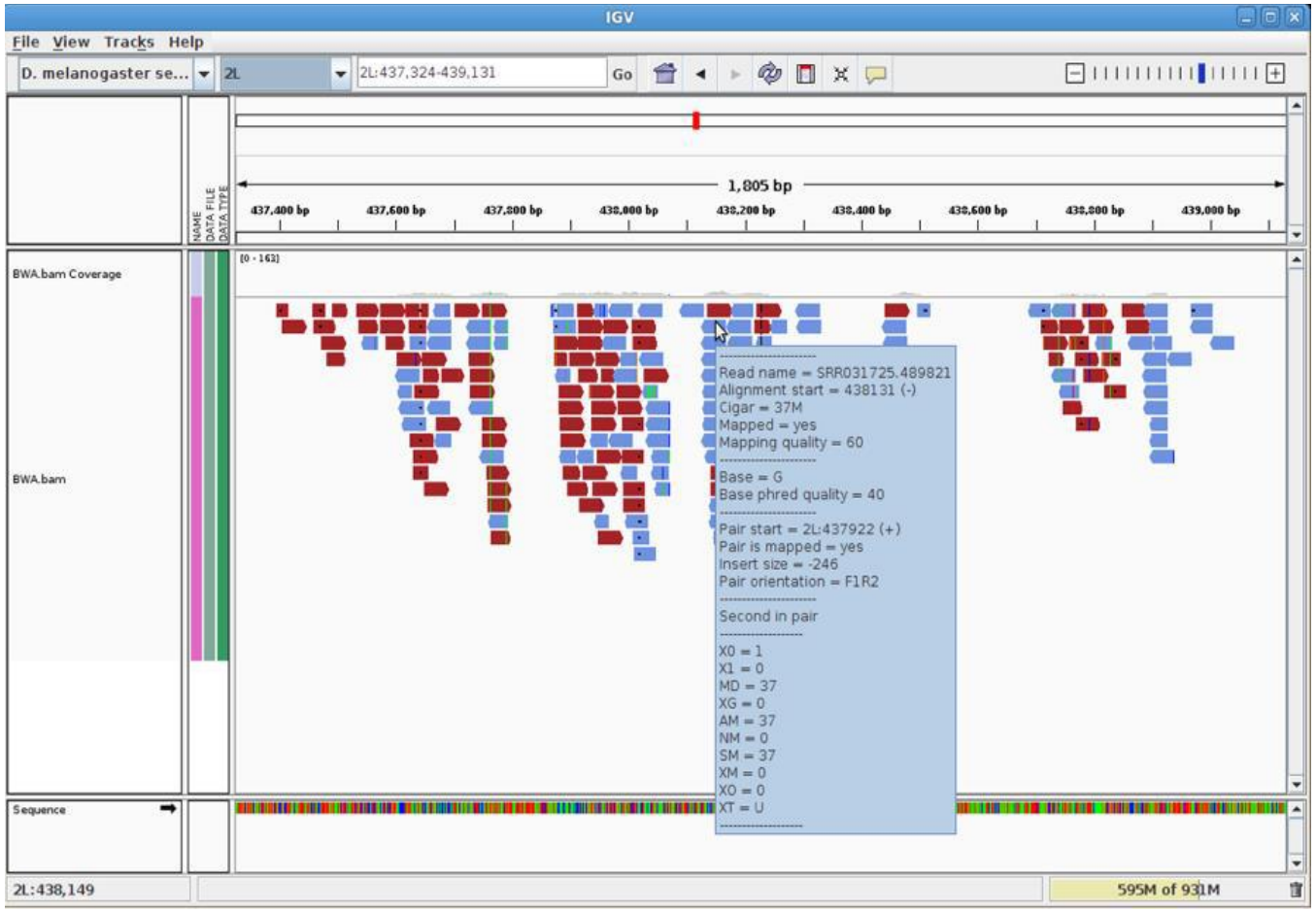
- During the alignment, not all the reads are aligned – but what proportion?
- If they were aligned – are there any errors?
- How well is the reference genome covered?

- Important in determining whether:
  - we can proceed with the analysis or some pre-processing needs to be done
  - we need to possibly redo the alignment
  - or we need to realign those unaligned reads

Step 3: Post-alignment QC and visualization

Allows us to get a detailed look on the **coverage** of a **given** region.

IGV genome browser

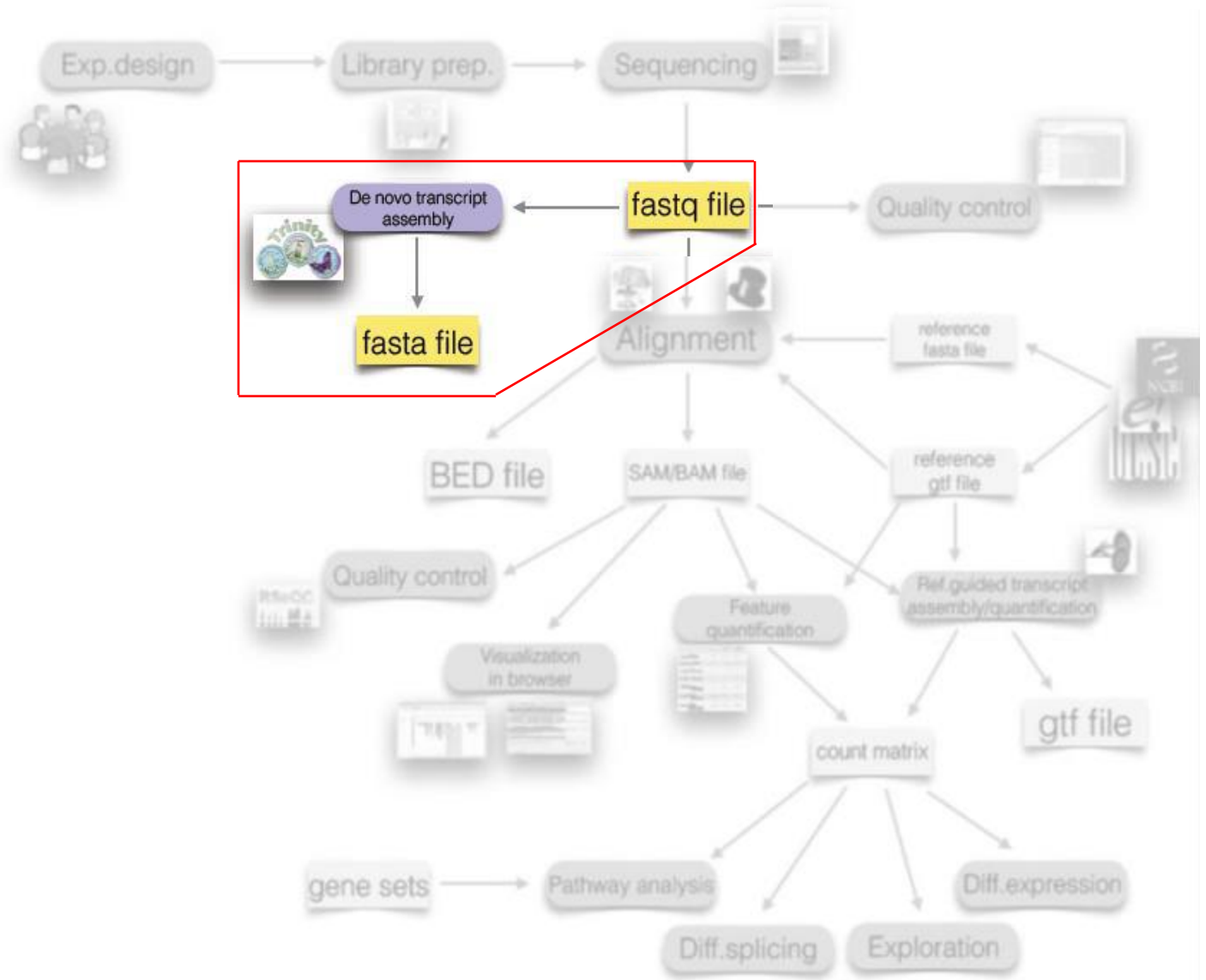


http://software.broadinstitute.org/software/igv/

MUNI FI

MUNI | RECETOX

Alternative step 2: Genome/transcript (de-novo) assembly



MUNI FI

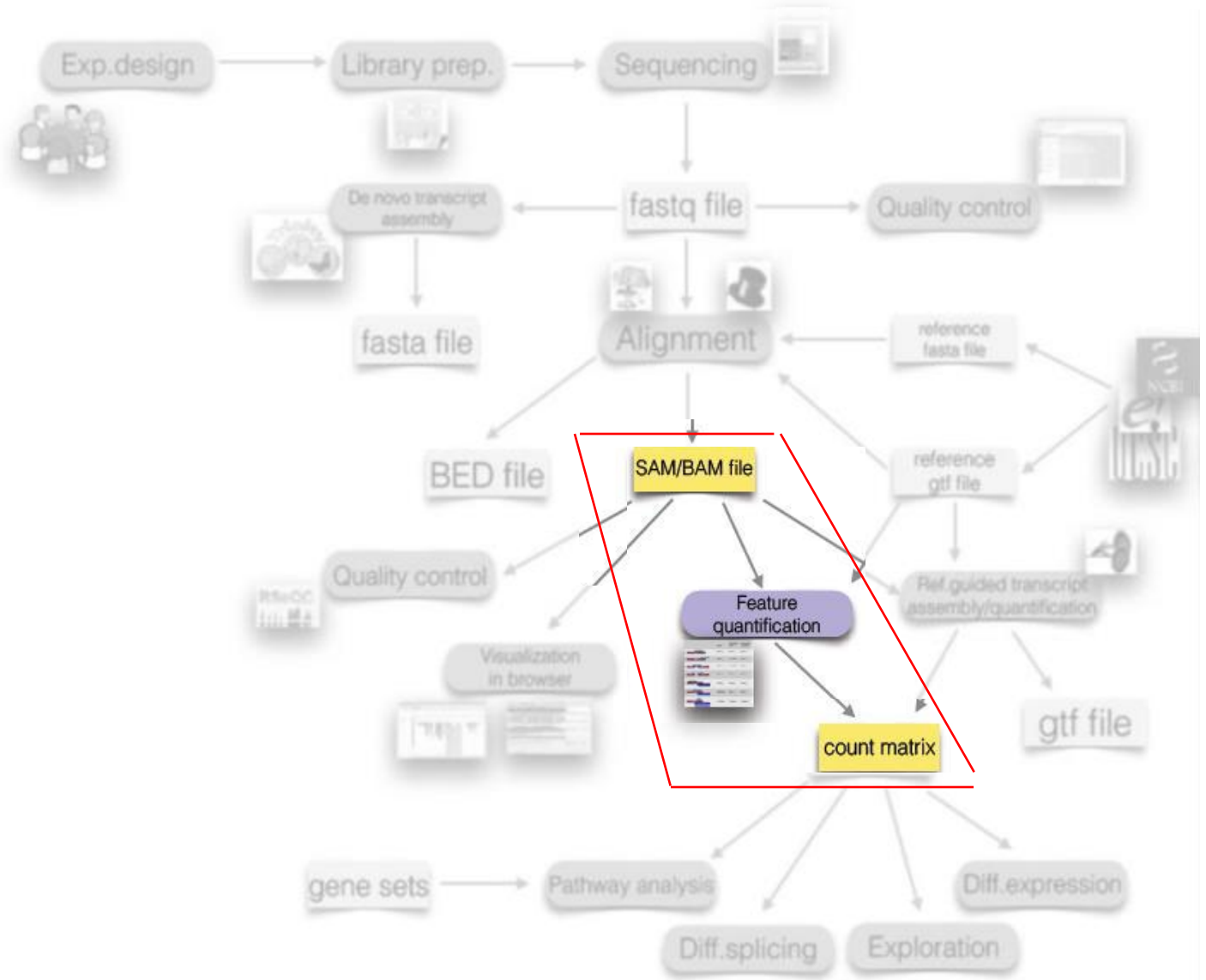
MUNI | RECETOX

Alternative step 2: Genome/transcript (de-novo) assembly

– When the reference sequence does not exist

- Alignment is dependent on the existence of reference sequence.
- However – sometimes this reference does not exist! – ***de novo*** genome assembly – we need to practically create the reference genome.
- The assembly is sometimes preferred in order to identify large structural rearrangements even when reference genome is known. In transcriptomics we can use it to detect **alternative splicing** events

Step 4: Feature detection (quantification)



MUNI
FI

MUNI | RECETOX

Step 4: Feature detection (quantification)

– Creates the final table with read counts for further statistical analyses

- A feature of interest differs based on the experiment:
  - gene, exon, intron… (WGS, WES)
  - transcript, isoform (RNA-seq)
  - variant - SNP, insertion, deletion, CNV - (WGS, WES, targeted sequencing)
  - promotor sequence (ChIP-Seq)

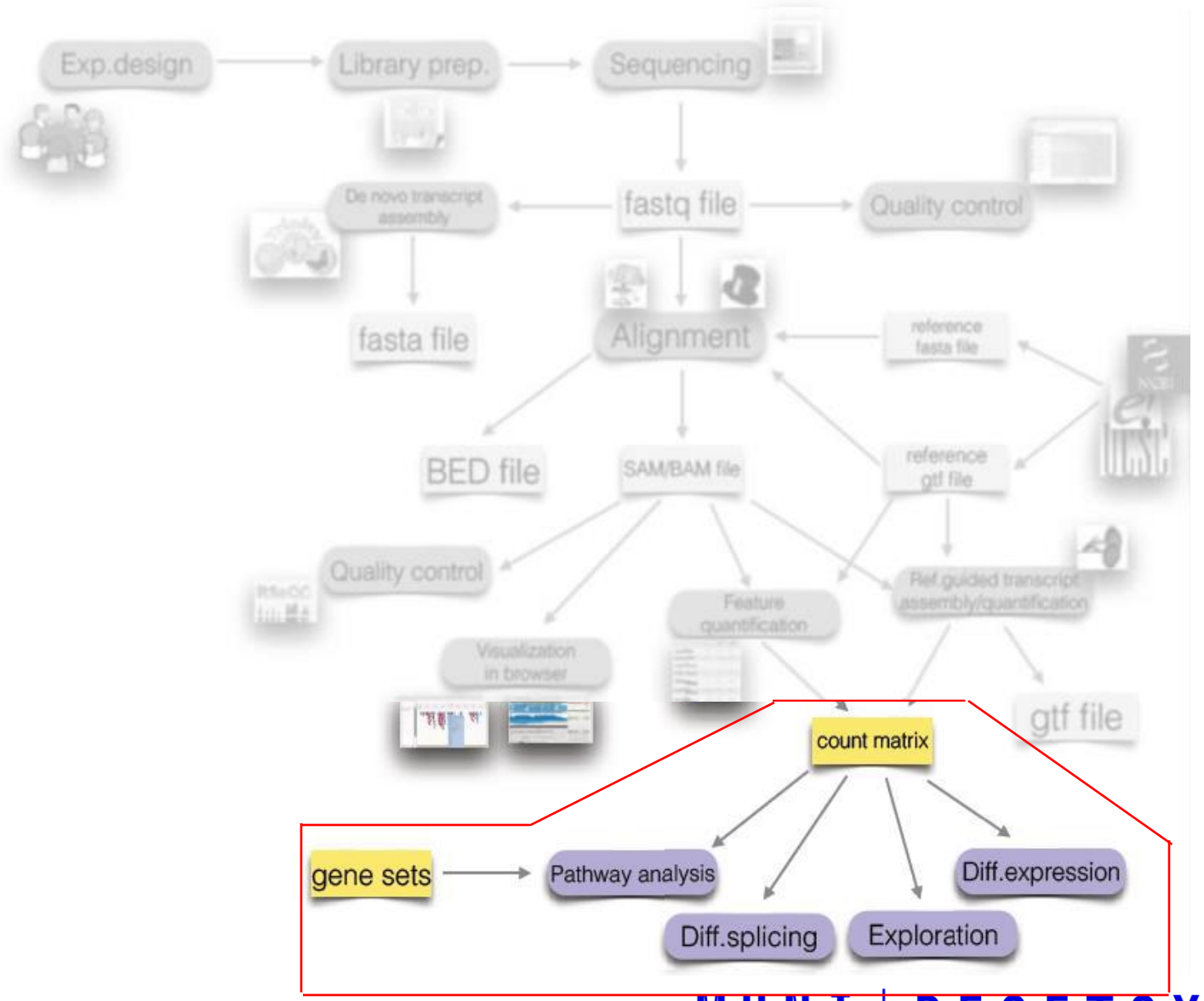
- In **transcriptomics** NGS experiments, the emphasis is on **quantification** of known transcripts (unless the aim is to get new isoforms) – we quantify the abundance of the RNA.
- In **genomic** NGS experiments, the emphasis is more on the **detection** of structural changes (the quantification is the % of alternative alleles found).

Step 4: Feature detection (quantification)

– Creates the final table with read counts for further statistical analyses

- The final output of this step is always a matrix with:
  - **Information** about the feature (ID, name, variant…)
  - **Quantification** of this feature in each of the samples

Step 5: Statistical data analysis



MUNI FI

MUNI | RECETOX

Step 5: Statistical data analysis

– The final matrix is input to four main analysis types:

Group comparison (between groups of samples or groups of features)
•Differential gene expression / splicing
•Differential variants detection

Group discovery (within samples or features)

•Clustering of patients into unknown subtypes based on their sequencing profiles
•Searching for genes with similar expression

Group prediction (usually for samples)

•Finding genes for diagnosis…

Special analyses: pathway analysis, construction of gene networks, analysis of survival, …

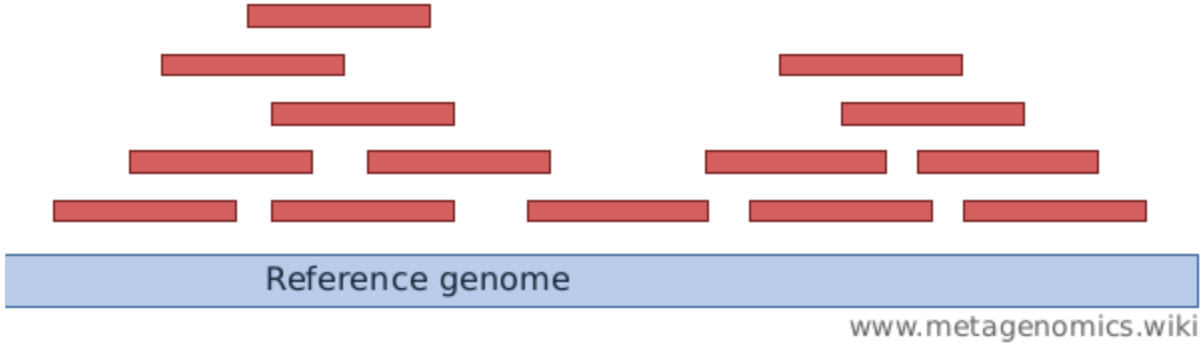
To remember:

- **Bioinformatics** (and especially the sequencing bioinformatics) is a **very new field**
- No good books, no standards, nothing lasts forever, … **almost everything** is old and **outdated**!
- **Bioinformaticians** have to be **always** looking for **new methods**, tools, algorithms, … it's the same when wet-lab people must search for novel methods which for decrease bias, are faster, require less input material, …
- **Garbage in –garbage out**
- If you **do not understand** the whole process you **don't know** what the **results** mean

Some important terms

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)

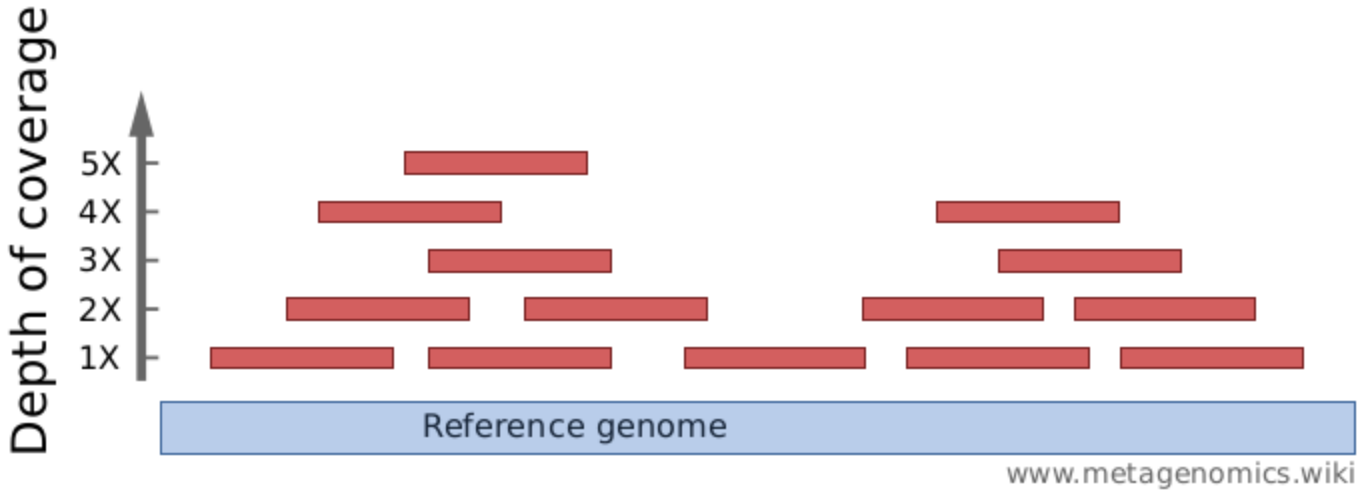
How strongly is the genome "covered" by sequenced fragments (short reads)?

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)

$$Av = (NxL)/G$$

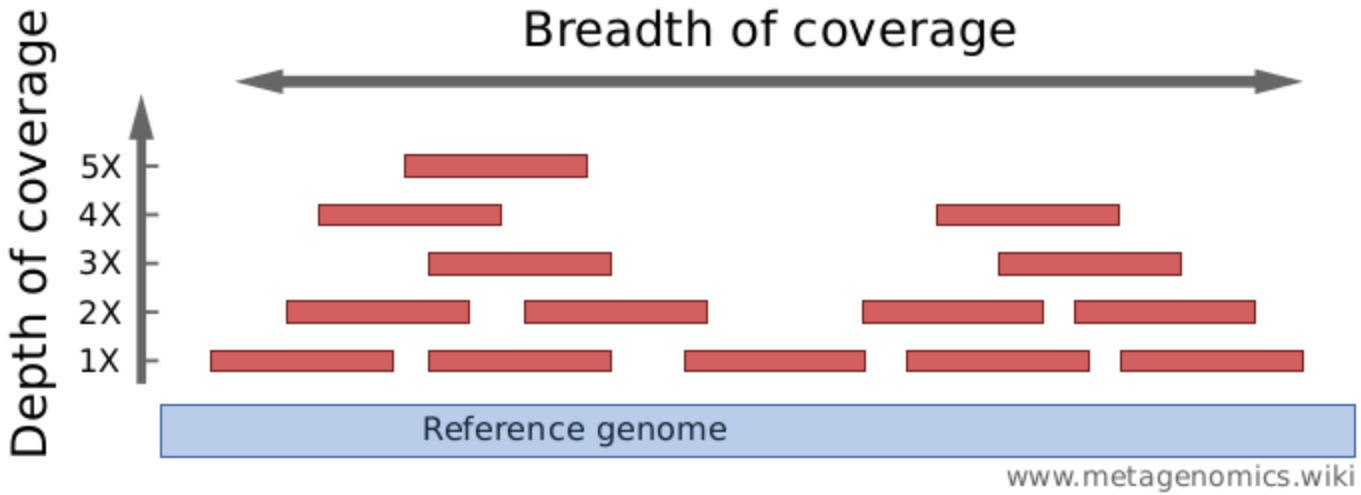
G - length of the original genome
N - number of reads
L - average read length



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?

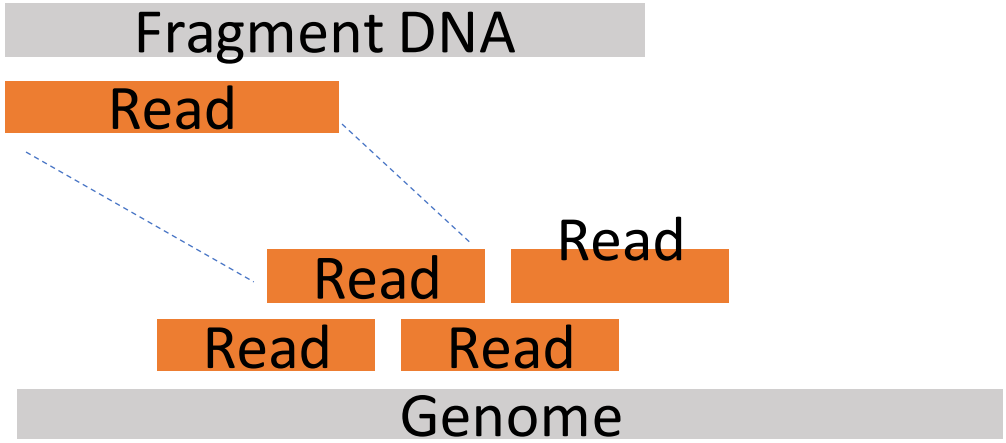


<u>Breadth of coverage</u> is the percentage of bases of a reference genome that are covered with a certain depth. For example: "90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth."

Single or paired- end?

Single-end sequencing

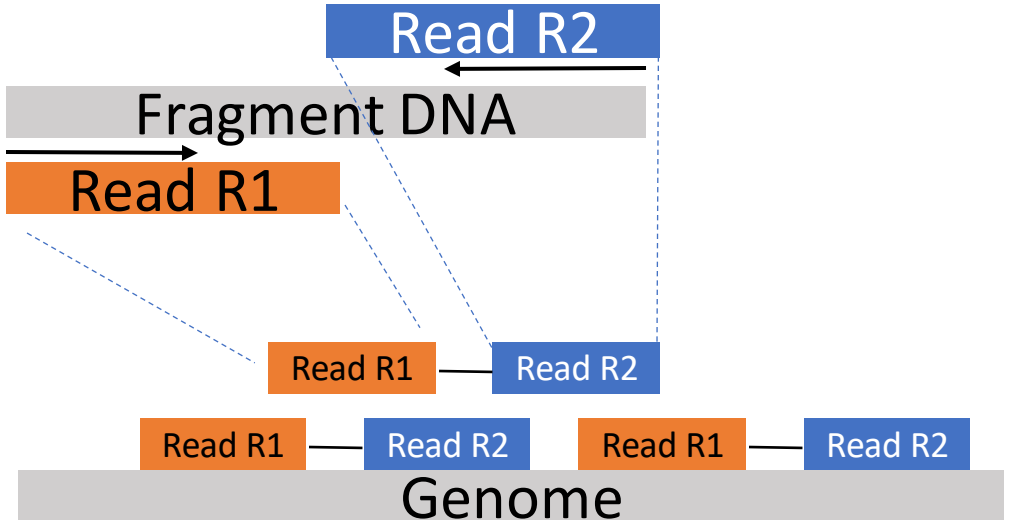
- Pros: fast, cheap
- Cons: limited use
- Usage: usually sufficient for studies looking to detect counts rather than structural changes, such as RNA-Seq or ChIP-Seq



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!!!**