# TOPIC 3:

## Sequence data

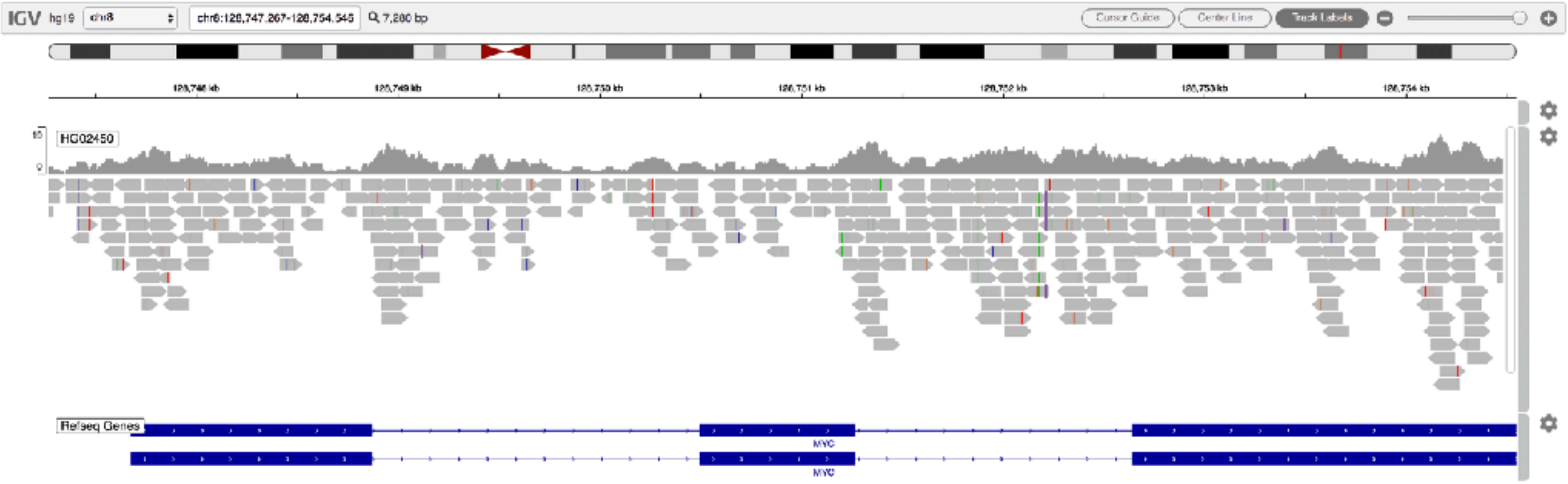
# Outline

1. **Different methods to acquire sequence data**
2. Understand sequence file formats
3. Preparing files for analysis
    - Tutorial looking at sequence data files and quality

# Whole Genome Sequencing

Randomly sheer DNA and sequence all fragments

May use double-stranded nuclease treatment to reduce repetitive elements

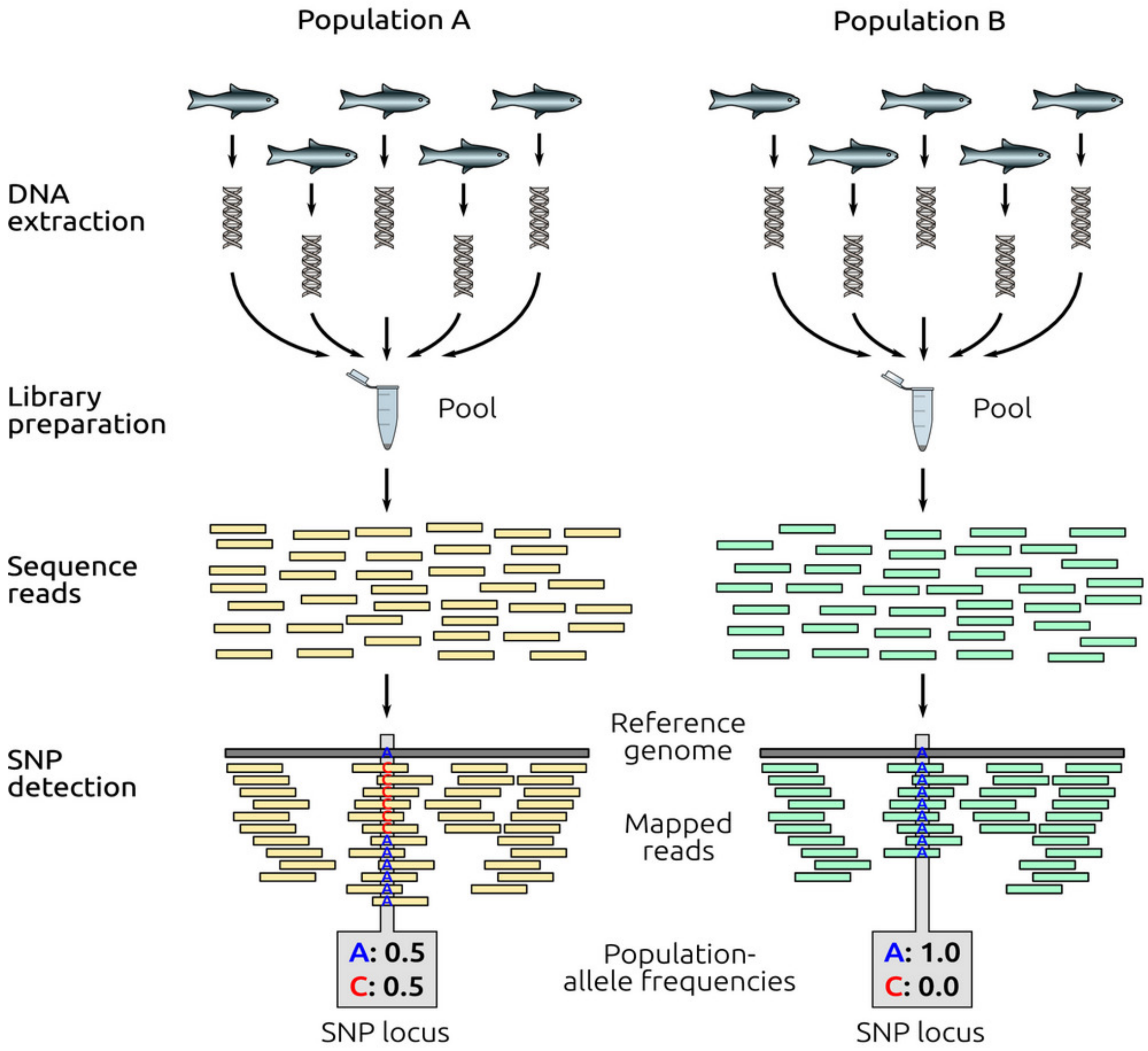


*Screen shot from the Integrated Genomics Viewer*

# Whole Genome Sequencing

| **Pros** | **Cons** |
|---|---|
| All sites possible | Comparatively expensive per sample |
| Simple library prep | Storage and bioinformatics challenging with lots of samples |

# Pool Seq

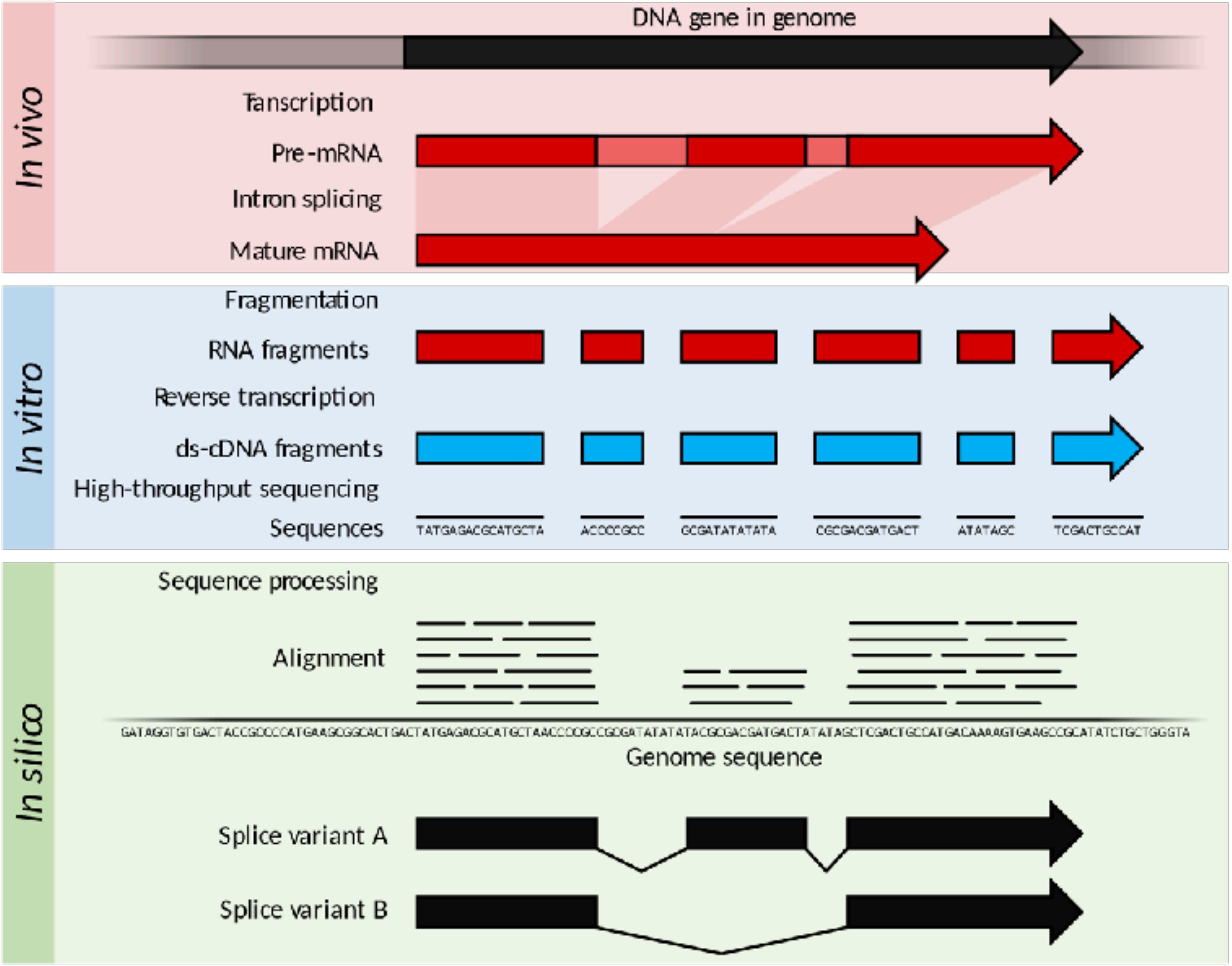


Adapted from Fuentes-Pardo & Ruzzante 2017 Mol. Ecol

# Pool Seq

| **Pros** | **Cons** |
|---|---|
| All sites possible | Limited analysis options |
| Simple library prep | No haplotype information |
| Cheaper than individual WGS | Best in cases where # samples > # reads |

# RNAseq



*From Wikipedia*

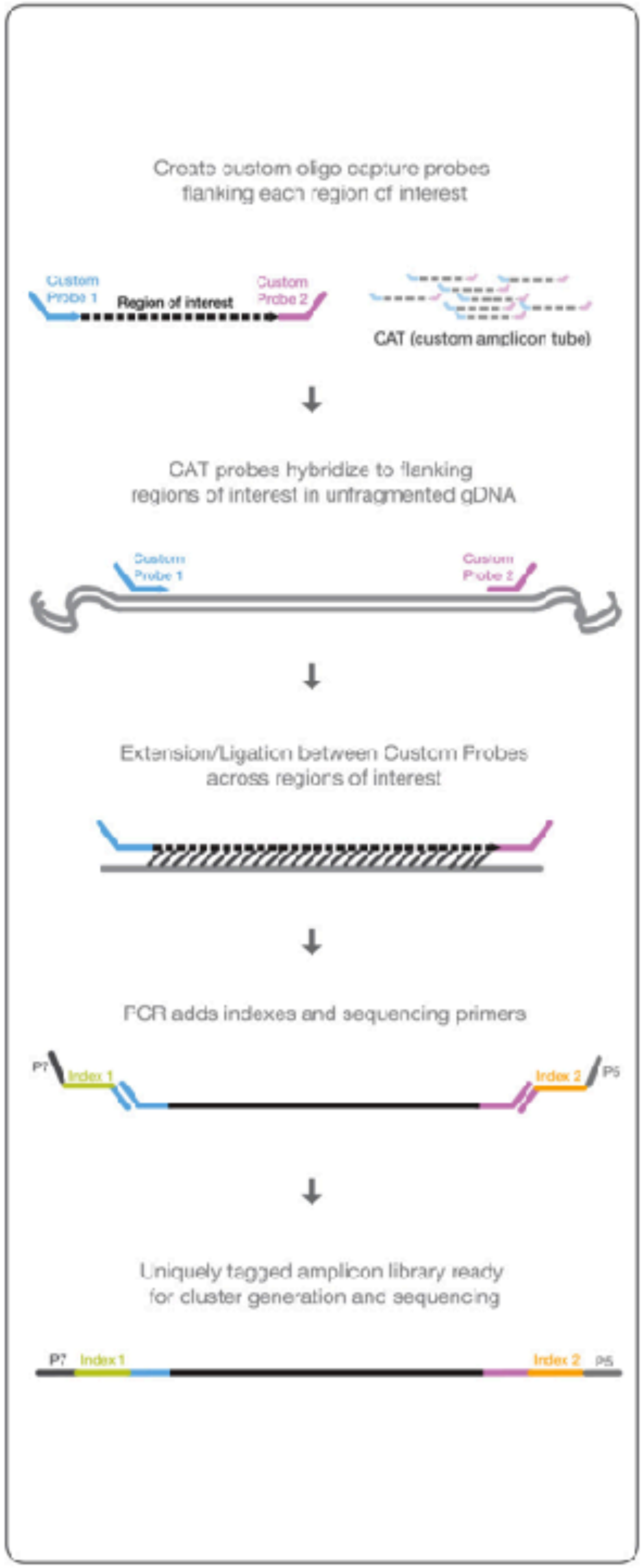
# RNAseq

| Pros | Cons |
|---|---|
| Many sites and only in genes | Expression differences complicate SNP calling |
| Also get expression information | Expensive for pop gen level sampling |
| Relatively easy to assemble | Difficult library prep (or so I’m told!) |

# Amplicon Sequencing

- Use PCR to amplify target DNA. Sequence many barcoded samples in one lane.

- Used to characterise microbiome by sequencing 16s rRNA



Amplicon Generation Workflow

Create custom oligo capture probes flanking each region of interest

CAT probes hybridize to flanking regions of interest in unfragmented gDNA

Extension/Ligation between Custom Probes across regions of interest

PCR adds indexes and sequencing primers

Uniquely tagged amplicon library ready for cluster generation and sequencing

# Amplicon Sequencing

| **Pros** | **Cons** |
|---|---|
| Get incredible depth at single locus | Limited to one or few loci |
| Simple bioinformatics. | Mutations in primer site don't sequence |

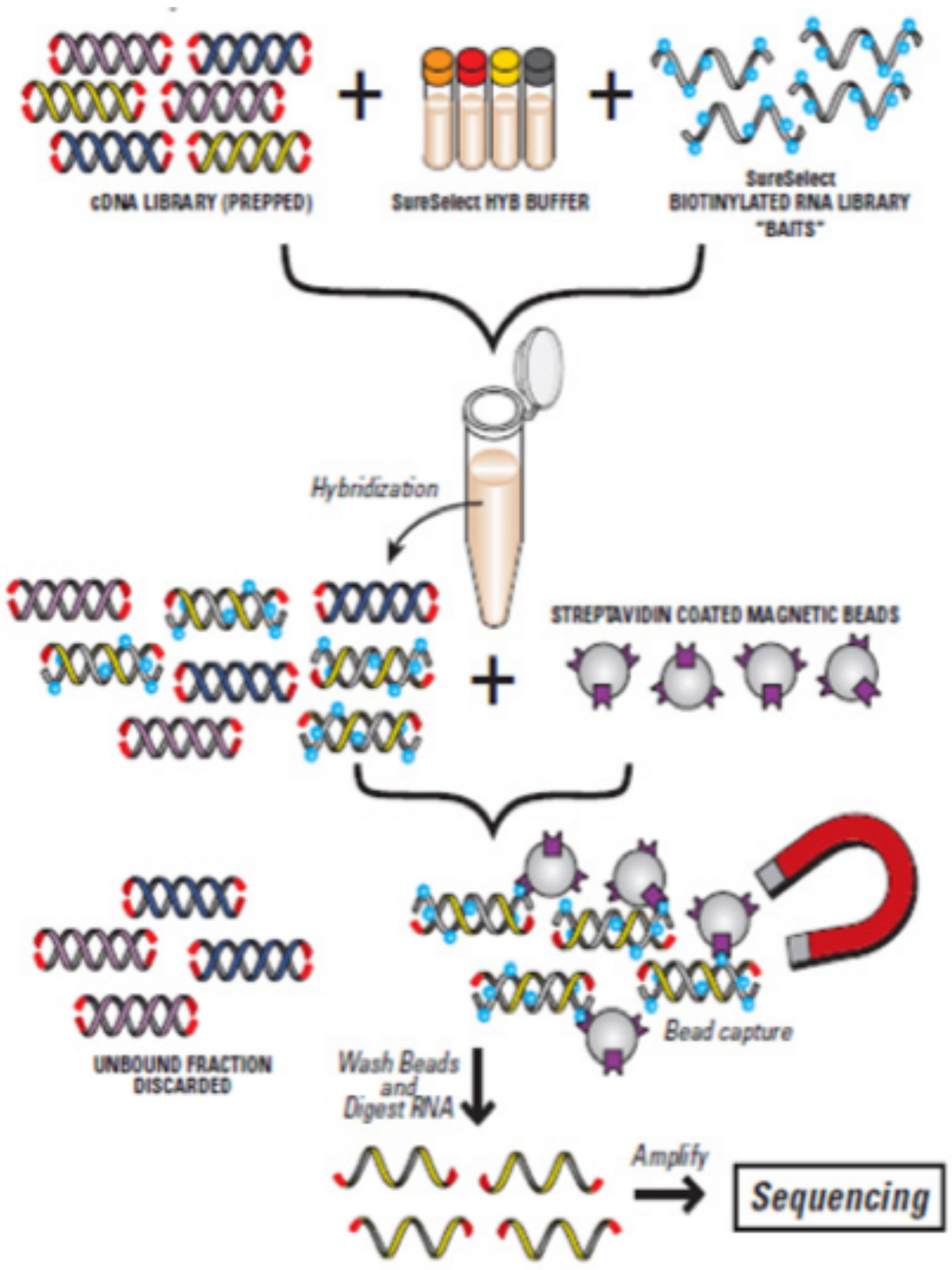
# GT-seq

- Genotyping by Thousands
- Based on Amplicon sequencing
- Multiplex PCR amplify ~200 known SNPs and then sequence pooled PCR products.
- Very cheap ( $1/sample), and bioinformatically simple.
- Useful for genotyping thousands or tens of thousands of samples.
- Complicated initial set-up.

Campbell et al. 2015 *Mol Ecol Resour*.

# Sequence Capture

- Design probe sequences from genome resources, synthesis attached to beads
- Make WGS library, hybridize with probe set. Matching sequence will be captured, all others washed away
- Collect capture sequence, amplify and sequence



# Sequence Capture

| **Pros** | **Cons** |
|---|---|
| Relatively cheap per sample | Requires designing probes |
| Good depth at targeted sites | Long library prep |

# Reduced Representation Sequencing

Instead of sequencing the whole genome, it can be sufficient to sequence just a part of it

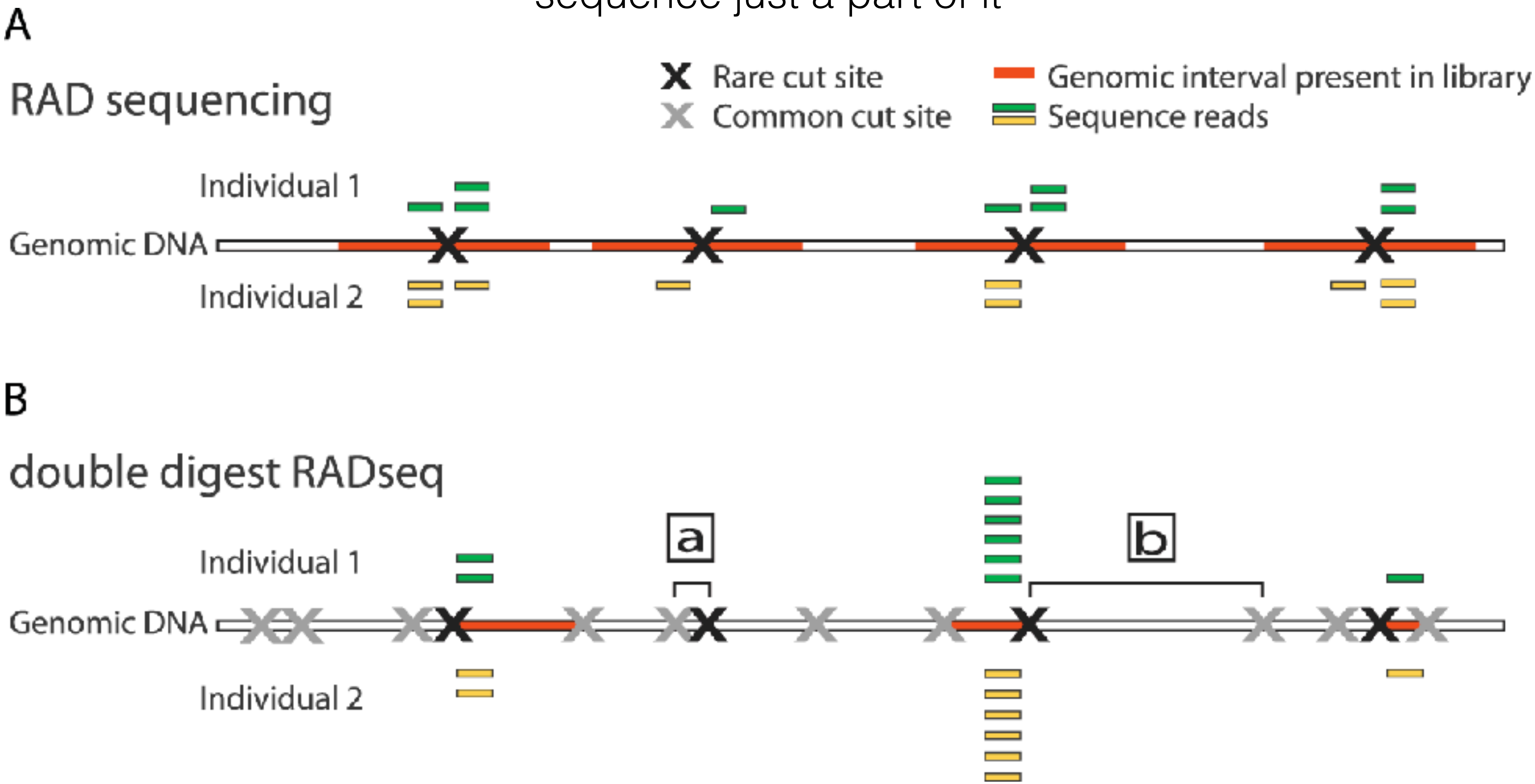
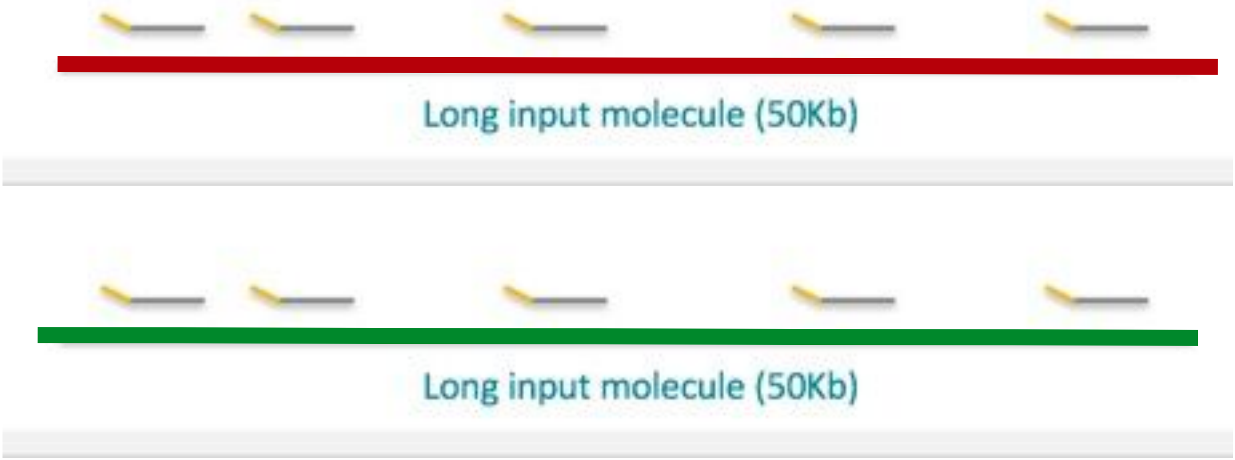


*Figure from Peterson et al PLoS One 2012*

# Reduced Representation Sequencing

| **Pros** | **Cons** |
|---|---|
| Quick library prep for hundreds of samples | Relatively sparse SNPs compared to other methods - limiting analysis options |
| Comparatively cheap per sample cost | Can have problems overlapping different library preps |

# Synthetic long reads



Barcodes read originating from individual DNA molecules

Sequence with Illumina reads

Original molecule can be reconstructed using the barcodes

Potentially very useful for genome assembly and phasing

# Synthetic long reads - i.e. Linked Reads

**E.g. Haplotagging**

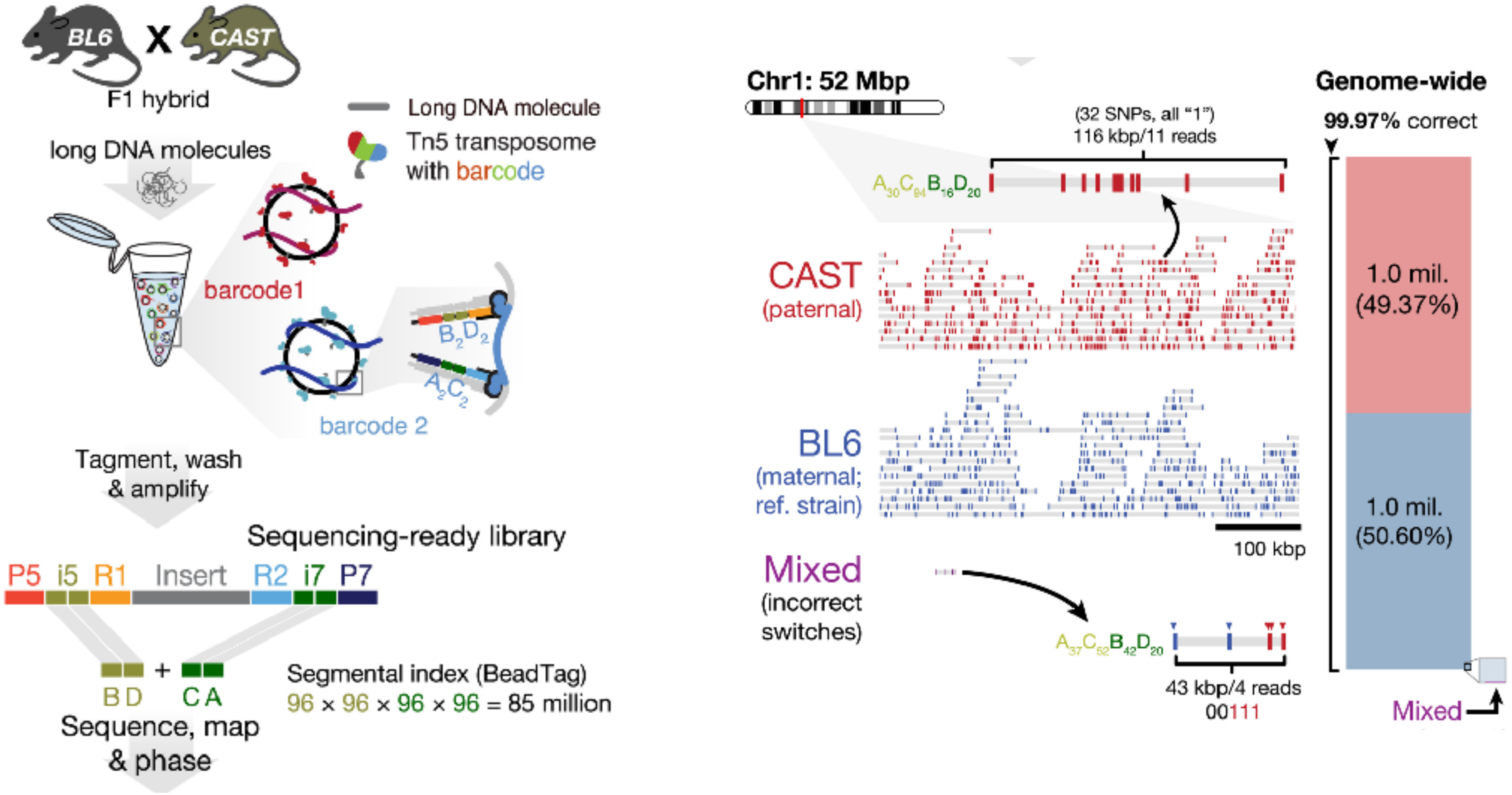
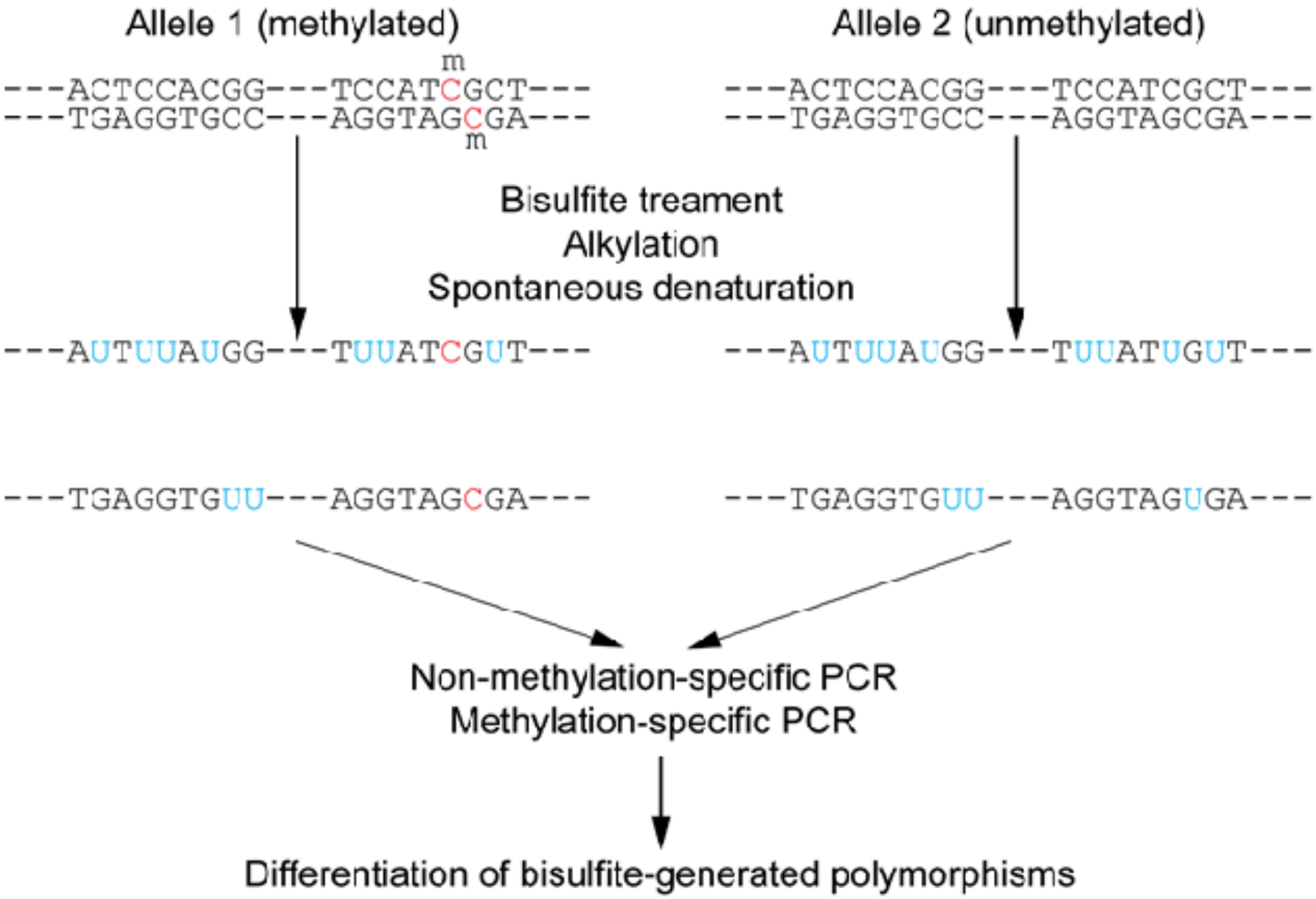


Figure adapted from Meier et al *2021 PNAS*

# Bisulphite Sequencing

Unmethylated cytosines are converted to **U**racil

Methylated **C**pG sites are unchanged and are detected as polymorphisms


Allele 1 (methylated)
Allele 2 (unmethylated)
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
Bisulfite treament
Alkylation
Spontaneous denaturation
---AUTUUAUGG---TUUATCGUT---
---AUTUUAUGG---TUUATUGUT---
---TGAGGTGUU---AGGTAGCGA---
---TGAGGTGUU---AGGTAGUGA---
Non-methylation-specific PCR
Methylation-specific PCR
Differentiation of bisulfite-generated polymorphisms

# How to choose?

For example:

If you wanted to estimate demographic history from the distribution of allele frequencies, a reduced representation method might suffice to obtain an estimate of the site frequency spectrum

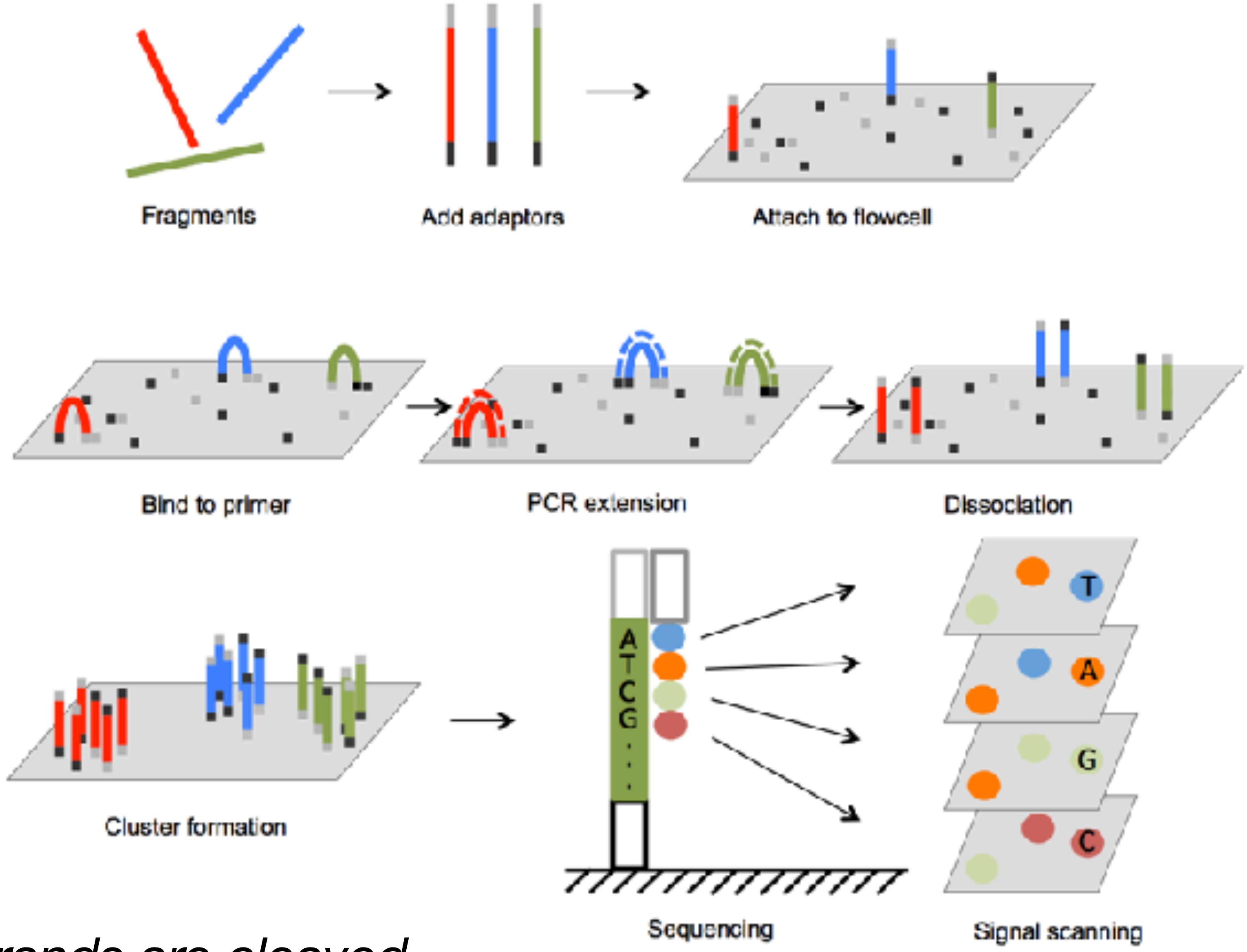
Or, if you want to perform a genome scan, looking at how haplotype frequencies varied among populations, you'd probably need deeper, whole genome information - it all depends on the questions you are tackling

# Outline

1. Different methods to acquire sequence data
2. **Understand sequence file formats**
3. Preparing files for analysis
    - Tutorial looking at sequence data files and quality

# Part 2: Sequence file formats

## Illumina sequencing

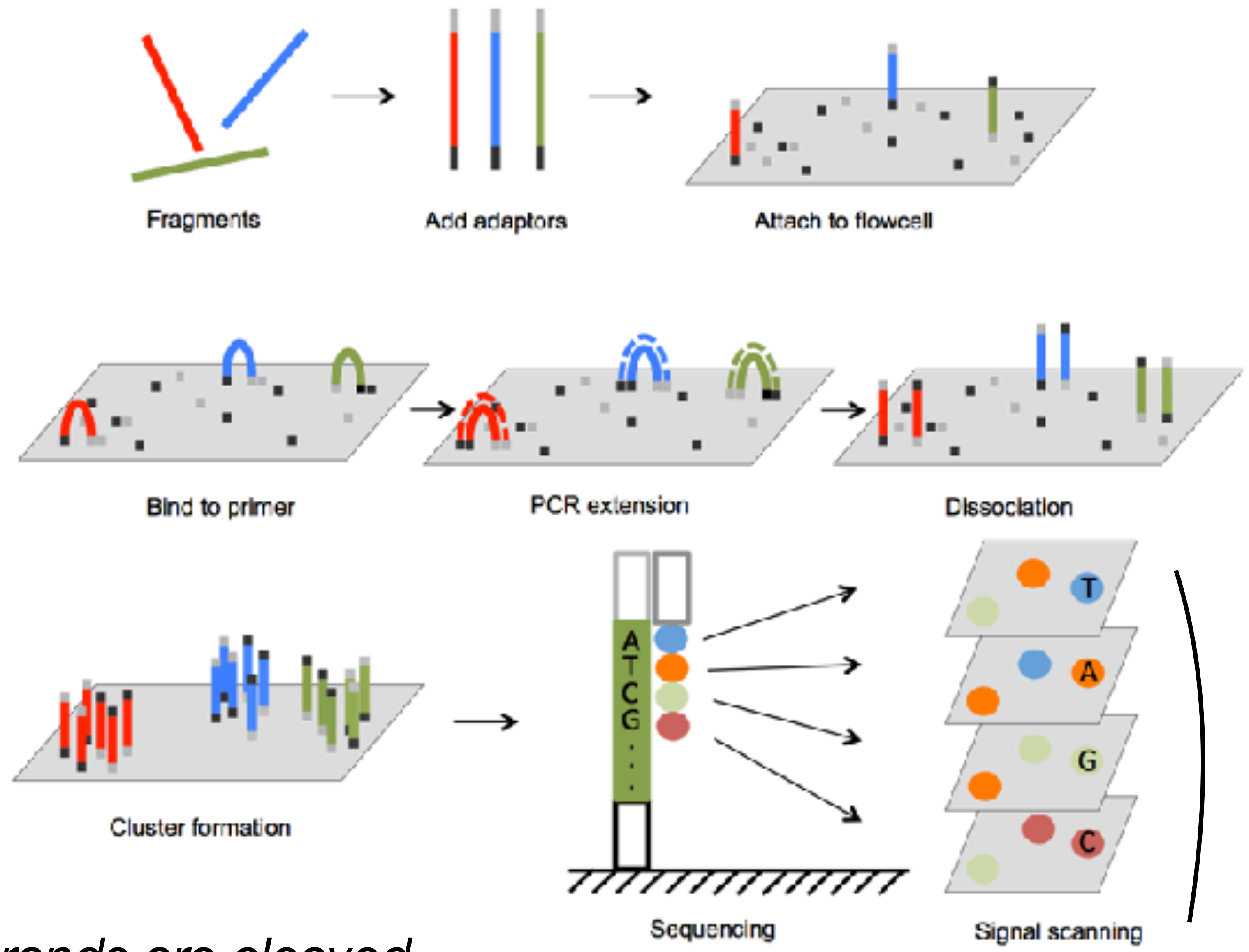


*Reverse strands are cleaved after cluster formation*

*4 cycles are shown, but modern Illumina machines are capable of 600 cycles in one run*

# Part 2: Sequence file formats

## Illumina sequencing



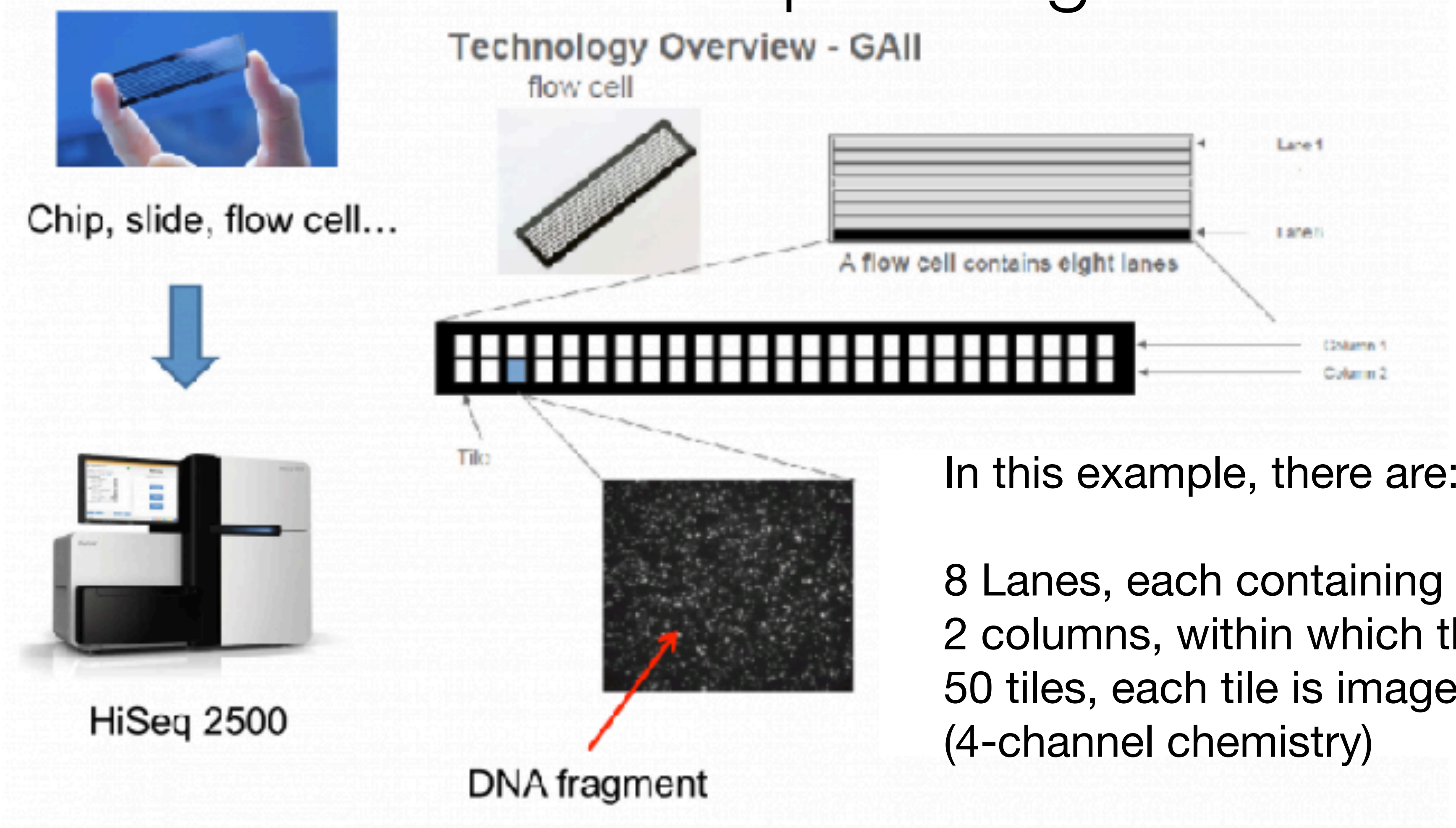
*This process has generated 4 Images*

*Reverse strands are cleaved after cluster formation*

*4 cycles are shown, but modern Illumina machines are capable of 600 cycles in one run*

# Part 2: Sequence file formats

## Illumina sequencing

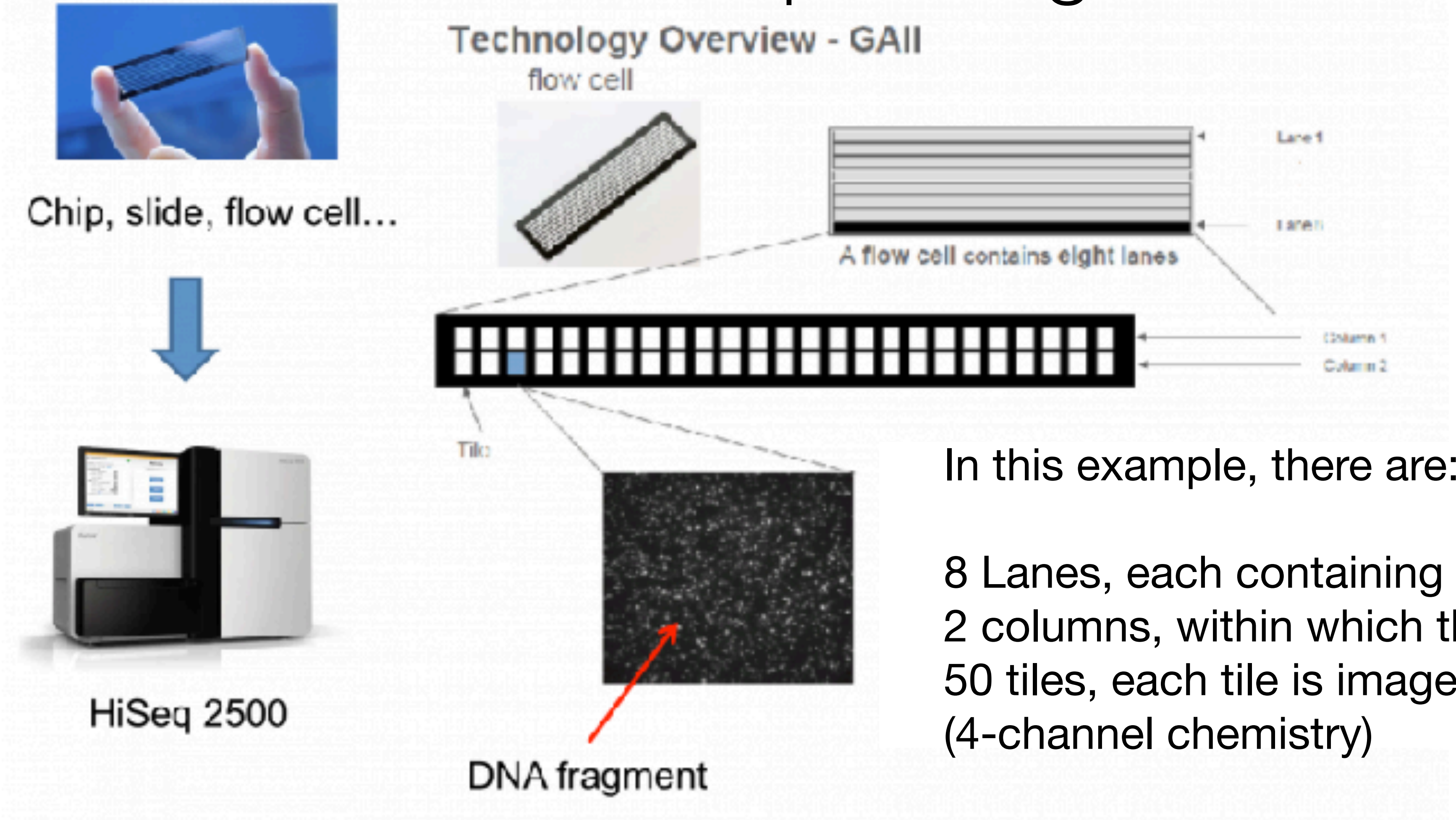


In this example, there are:

8 Lanes, each containing
2 columns, within which there are
50 tiles, each tile is imaged 4 times/cycle
(4-channel chemistry)

# Part 2: Sequence file formats

## Illumina sequencing


Technology Overview - GAII
flow cell
Chip, slide, flow cell...
Lane 1
Lane n
A flow cell contains eight lanes
Column 1
Column 2
Tile
HiSeq 2500
DNA fragment

In this example, there are:

8 Lanes, each containing
2 columns, within which there are
50 tiles, each tile is imaged 4 times/cycle
(4-channel chemistry)

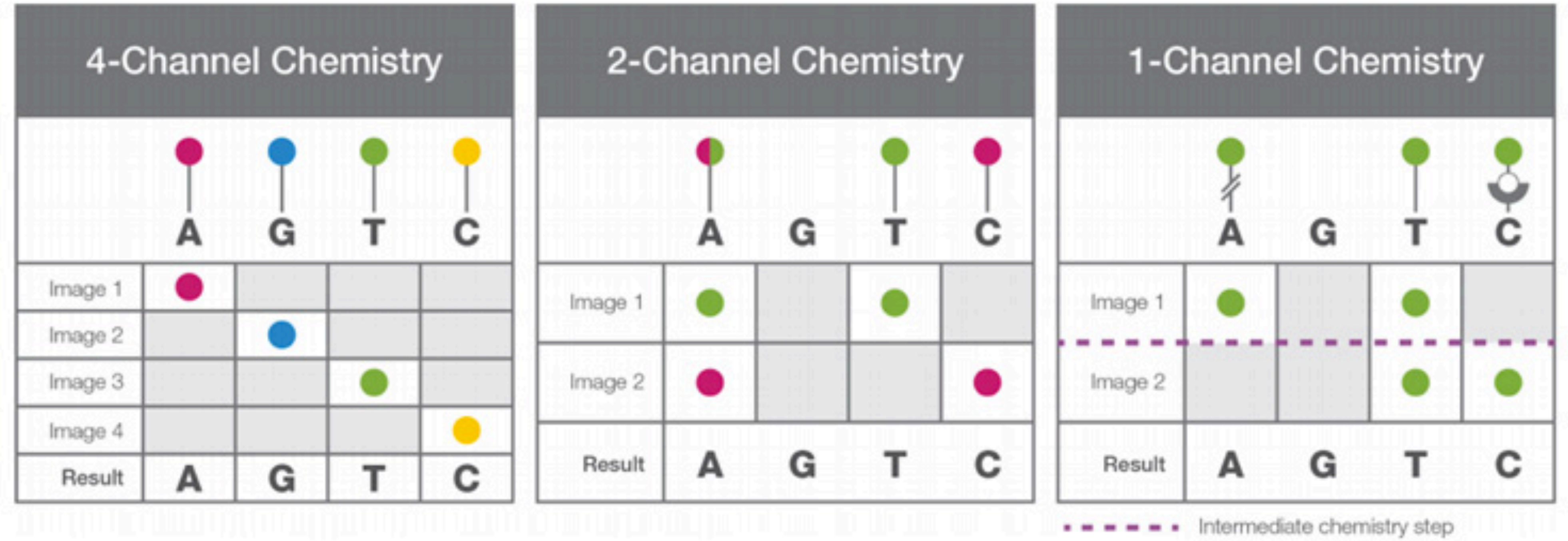
So there are approximately 8x2x50x4 = 3,000 images generated per cycle

Each image is about 3Mb in size

For an Illumina run using 300 cycles, that would be 3000x3x300 = 2,700,000 Mb of data (~2.7 Tb)

# Part 2: Sequence file formats

## Illumina sequencing

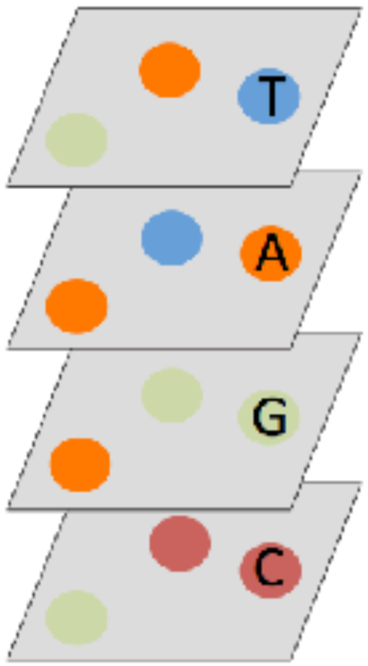


The number of channels refers to the numbers of colours the images detect

4-channel was Illumina's standard chemistry, but now 2-channel is more common

# Part 2: Sequence file formats

## Illumina sequencing



Using the stack of images from an Illumina machine you do the following:

1. Evaluates the light signal from every cluster to calculate the Quality Predictor Value (QPV), measuring things like:

The signal-to-noise ratio
Light Intensity

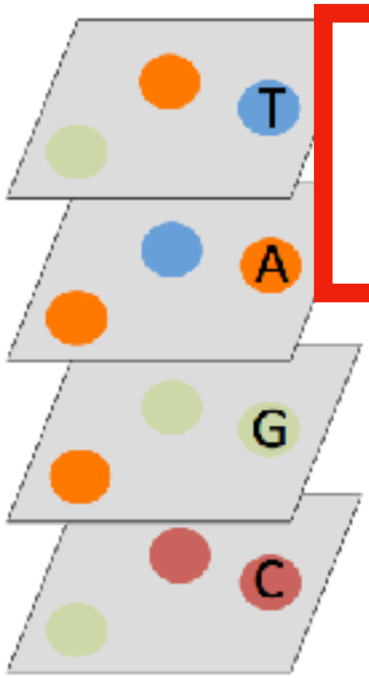
2. QPVs are converted into Phred quality scores (Q-scores) using a calibration curve built using previously sequenced samples

3. Convert the base call and the quality score into a FASTQ file

# Part 2: Sequence file formats

## Illumina sequencing

Using the stack of images from an Illumina machine you do the following:



1. Evaluates the light signal from every cluster to

suring

noise ratio

Light Intensity

**You'll probably never do these steps yourself, but it's good to know where the data come from!**

2. QPVs are converted into Phred quality scores (Q-scores) using a calibration curve built using previously sequenced samples

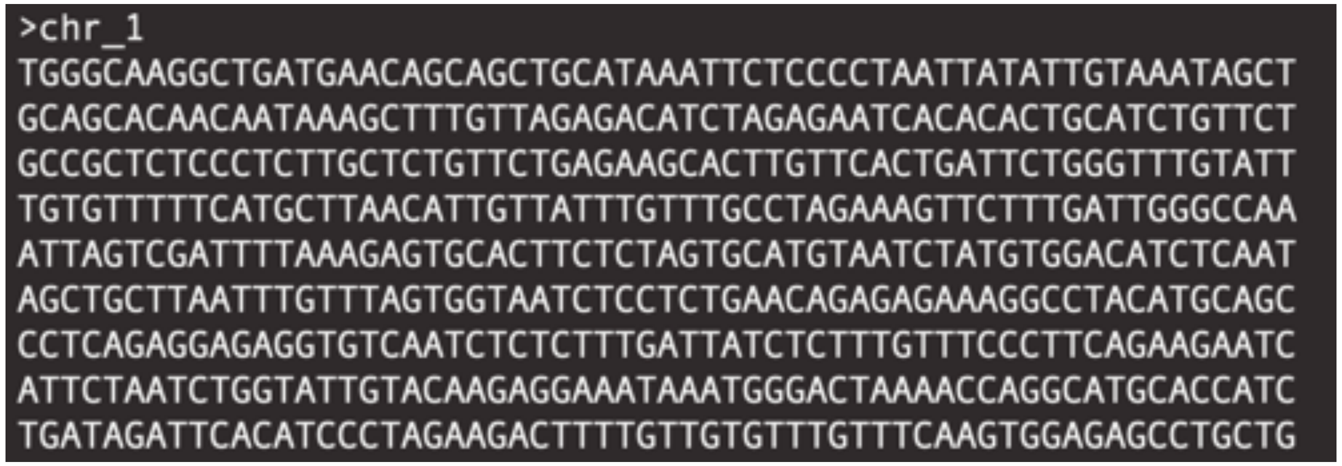
3. Convert the base call and the quality score into a FASTQ file

# Part 2: Sequence file formats

*Remember this from yesterday?*

What a FASTA file looks like:

Sequence name



```
>chr_1
TGGGCAAGGCTGATGAACAGCAGCTGCATAAATTCTCCCCTAATTATATTGTAAATAGCT
GCAGCACAACAATAAAGCTTTGTTAGAGACATCTAGAGAATCACACACTGCATCTGTTCT
GCCGCTCTCCCTCTTGCTCTGTTCTGAGAAGCACTTGTTCACTGATTCTGGGTTTGTATT
TGTGTTTTTCATGCTTAACATTGTTATTTGTTTGCCTAGAAAGTTCTTTGATTGGGCCAA
ATTAGTCGATTTTAAAGAGTGCACTTCTCTAGTGCATGTAATCTATGTGGACATCTCAAT
AGCTGCTTAATTTGTTTAGTGGTAATCTCCTCTGAACAGAGAGAAAGGCCTACATGCAGC
CCTCAGAGGAGAGGTGTCAATCTCTCTTTGATTATCTCTTTGTTTCCCTTCAGAAGAATC
ATTCTAATCTGGTATTGTACAAGAGGAAATAAATGGGACTAAAACCAGGCATGCACCATC
TGATAGATTCACATCCCTAGAAGACTTTTGTTGTGTTTGTTTCAAGTGGAGAGCCTGCTG
```

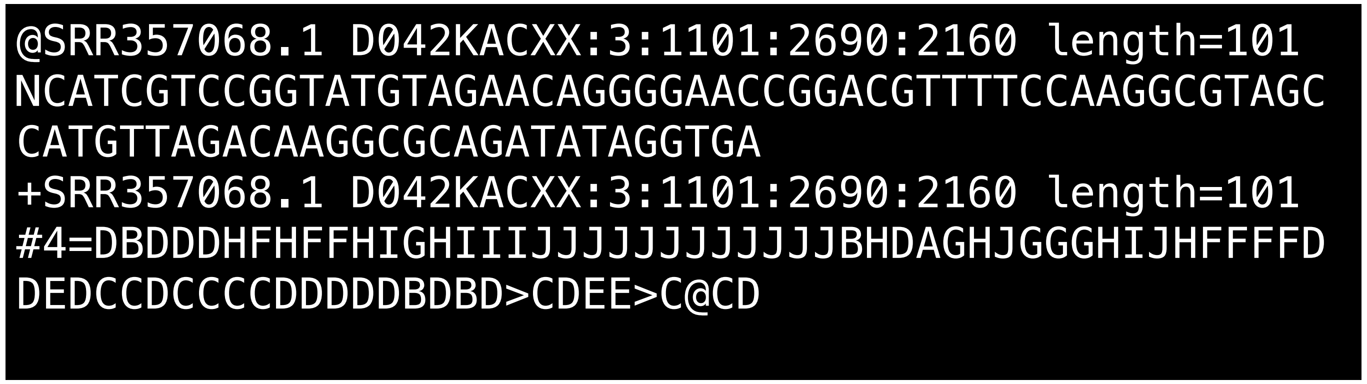
Nucleotide sequence

***FASTAs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:

4 Lines instead of 2



```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

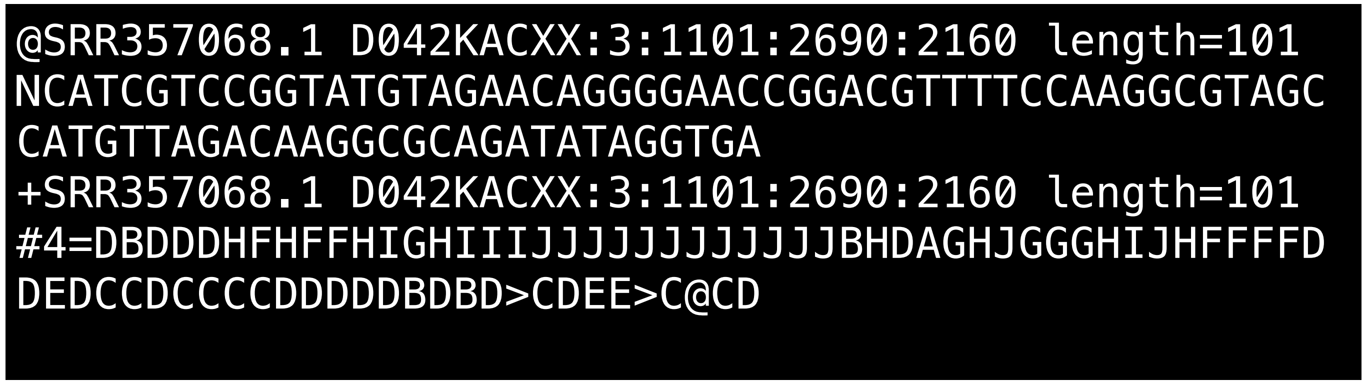
***FASTQs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:

1. Sequence ID
(begins with “@“ not “>”)

Typically contains information on the origin of the read - like which lane and tile it came from, where in the tile the cluster was located



```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

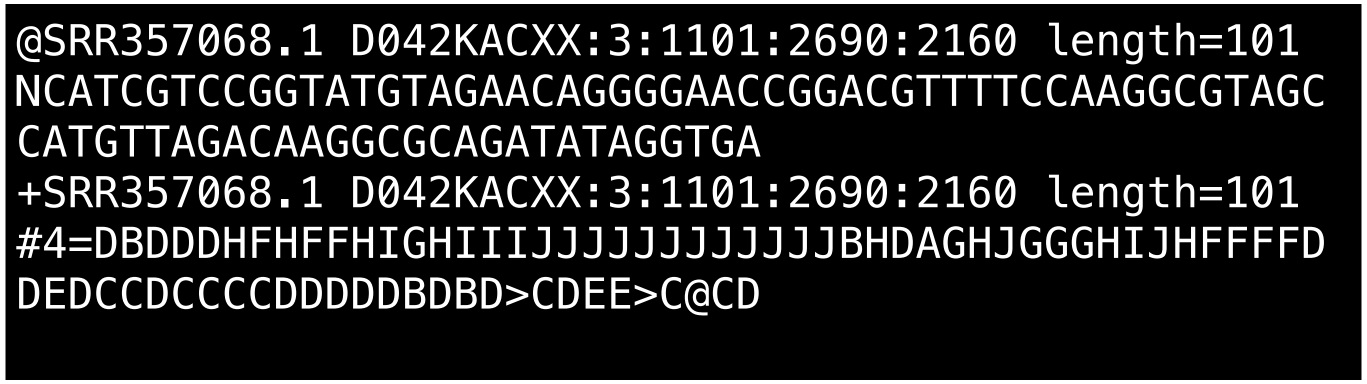
***FASTQs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:

1. Sequence
ID
(begins with “@“ not “>”)

2. Nucleotide
sequence

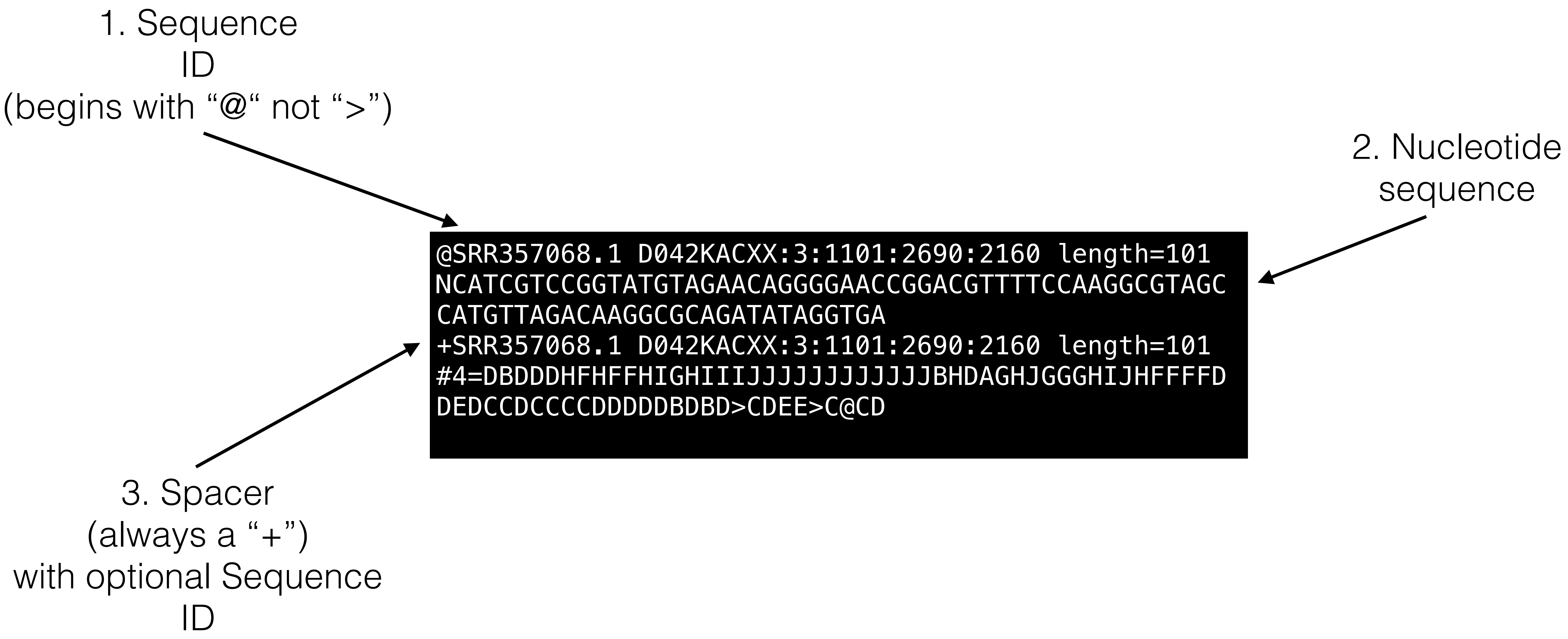


```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

***FASTQs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:



1. Sequence ID (begins with “@“ not “>”)

2. Nucleotide sequence

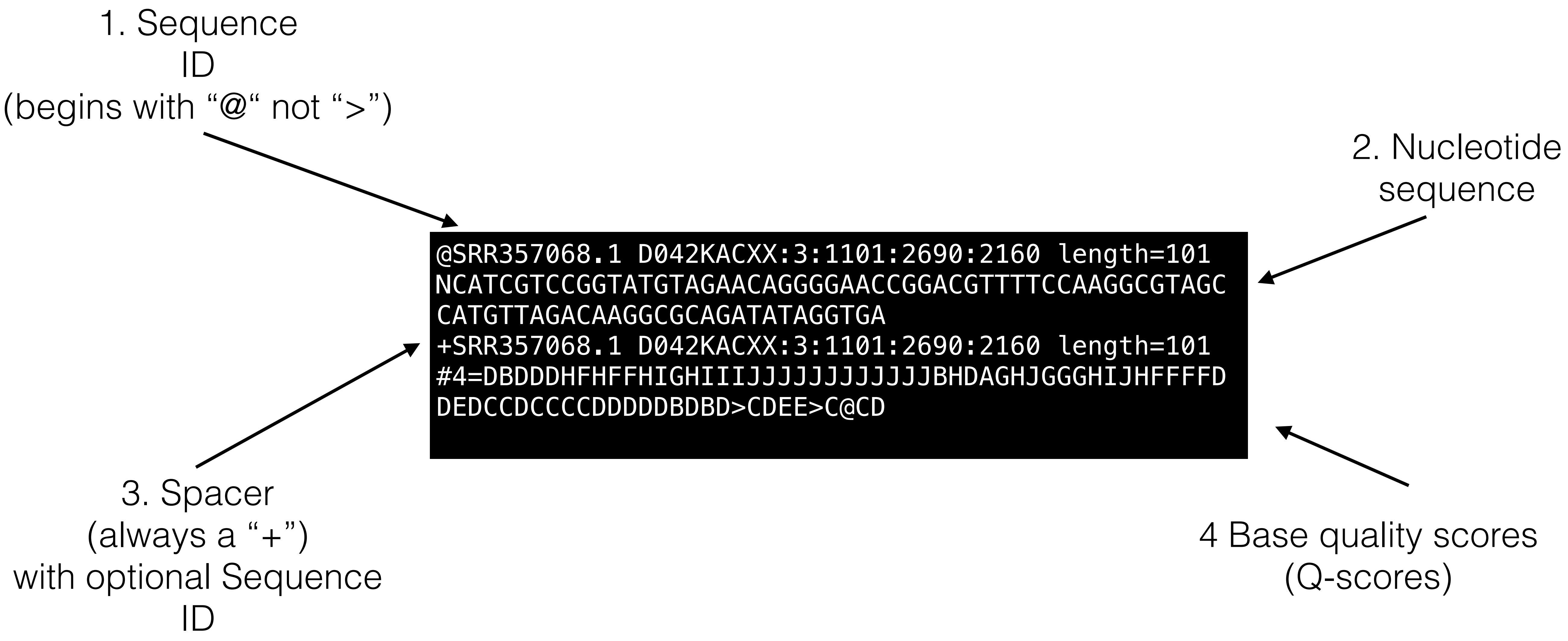
3. Spacer (always a “+”) with optional Sequence ID

```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

***FASTQs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:



1. Sequence ID (begins with “@“ not “>”)

2. Nucleotide sequence

3. Spacer (always a “+”) with optional Sequence ID

4 Base quality scores (Q-scores)

```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

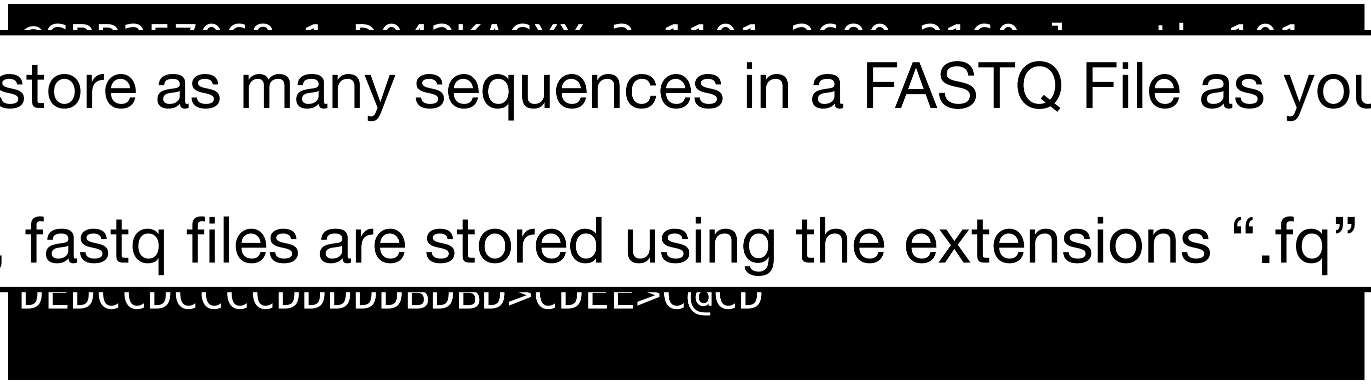
***FASTQs are plain text files***

# Part 2: Sequence file formats

## Anatomy of a FASTQ file:

1. Sequence ID (begins with "@" not ">")

2. Nucleotide sequence



```
@SRR357068.1 D042KACXX:2:1101:2698:2169 length=101
DEDCCDCCCCDDDDDDBDBD>CDEE>C@CD
```

You can store as many sequences in a FASTQ File as you like

By convention, fastq files are stored using the extensions ".fq" or ".fastq"

3. Spacer (always a "+") with optional Sequence ID

4 Base quality scores (Q-scores)

***FASTQs are plain text files***

# Part 2: Sequence file formats

## Base quality scores (Q-scores)

$$Q_{Sanger} = -10log_{10}(p)$$

Where *p* is the probability that a base call is incorrect

$$Q_{Solexa} = -10log_{10}(\frac{p}{1-p})$$

Remember, those probabilities are calculated using the QPVs in Illumina sequencing

# Part 2: Sequence file formats

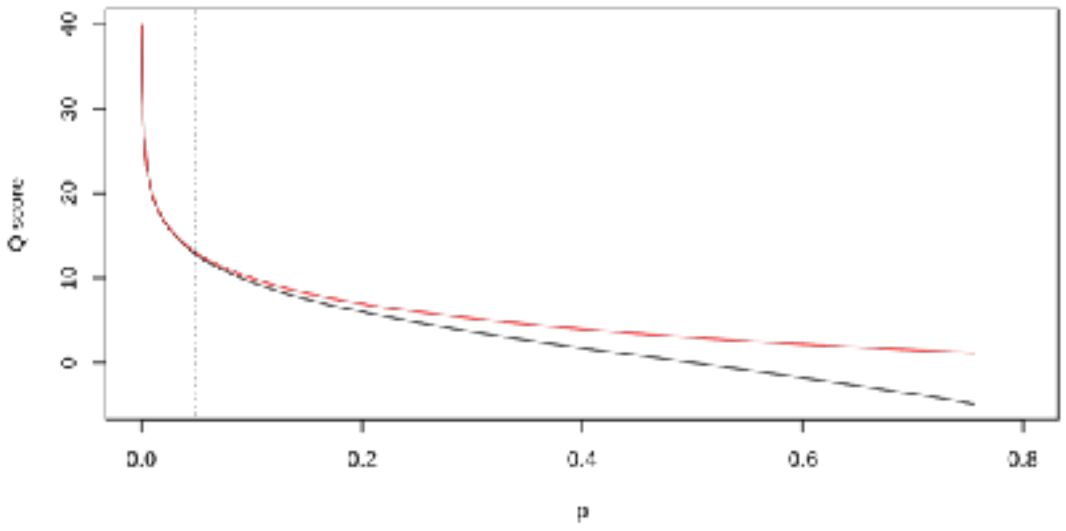
## Base quality scores (Q-scores)

$$Q_{Sanger} = -10log_{10}(p)$$

Where *p* is the probability that a base call is incorrect

$$Q_{Solexa} = -10log_{10}(\frac{p}{1-p})$$

Remember, those probabilities are calculated using the QPVs in Illumina sequencing



***Red line is Sanger***
***Black line is Solexa***

*Asympotically identical when p is small*

# Part 2: Sequence file formats

## Base quality scores (Q-scores)

$$Q_{Sanger} = -10log_{10}(p)$$

What’s the probability that the base is incorrect if Q=30?

# Part 2: Sequence file formats

## Base quality scores (Q-scores)

$$Q_{Sanger} = -10log_{10}(p)$$

What’s the probability that the base is incorrect if Q=30?

p[ Q30 ] = 0.001

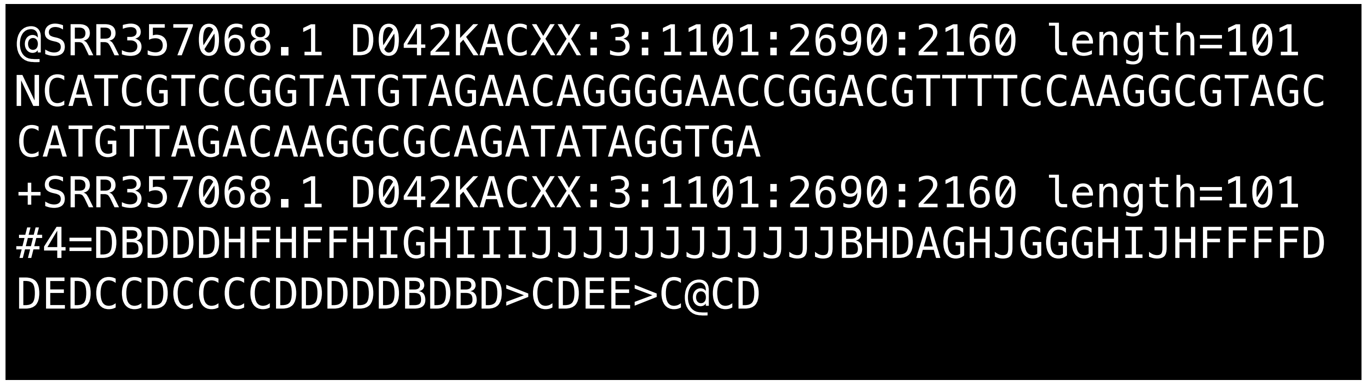
p[ Q20 ] = 0.01

p[ Q10 ] = 0.1

# Part 2: Sequence file formats

## Base quality scores (Q-scores)

You probably noticed that the Q-scores in the FastQ files are not numeric



```
@SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGC
CATGTTAGACAAGGCGCAGATATAGGTGA
+SRR357068.1 D042KACXX:3:1101:2690:2160 length=101
#4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFD
DEDCCDCCCCDDDDDBDBD>CDEE>C@CD
```

Under Illumina sequencing,
ASCII encoding is used to refer to Q scores from 0 to 62

Slightly different encoding strategies are used by the different technologies

# Part 2: Sequence file formats

PacBio Sequencing

Nanopore Sequencing

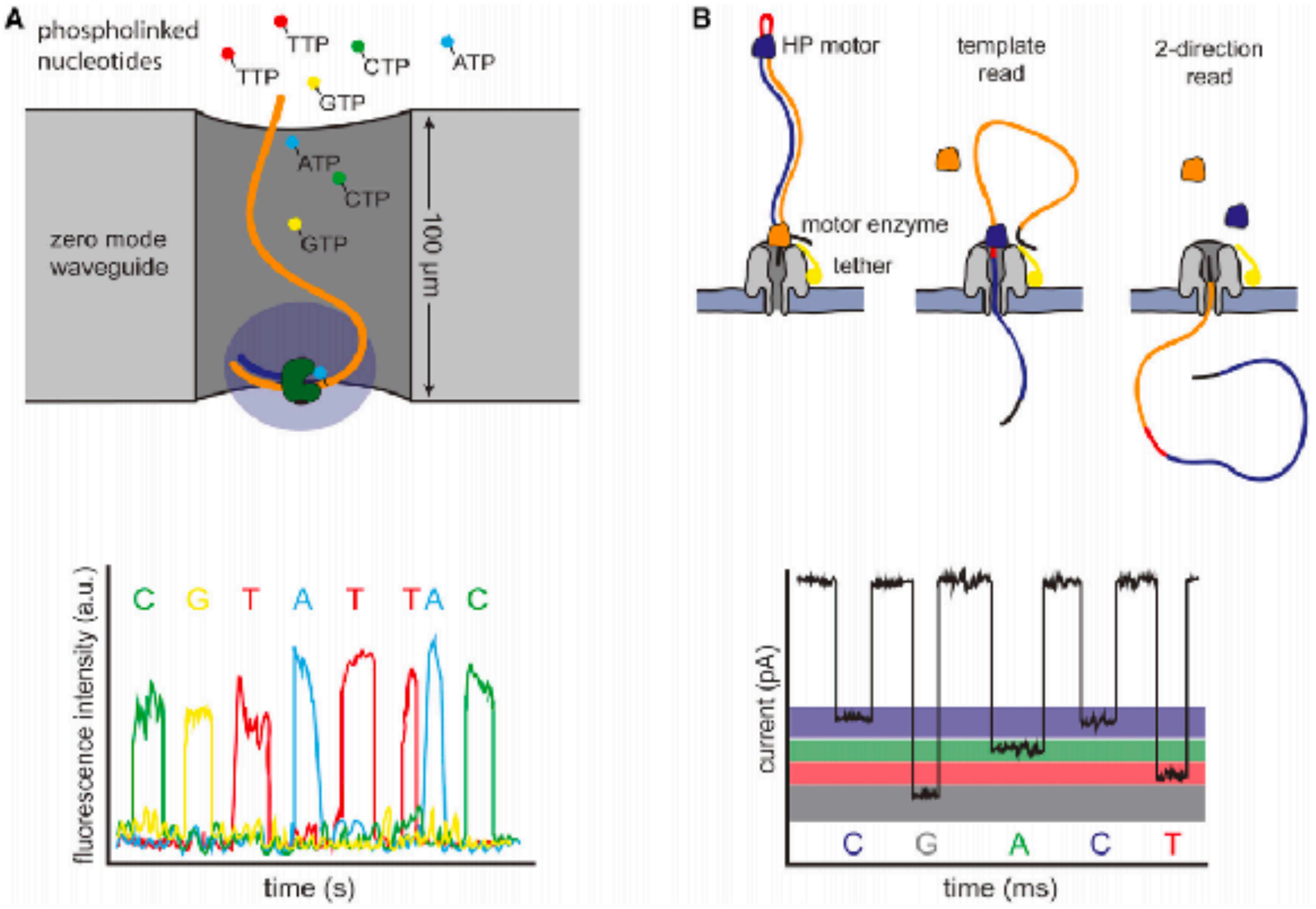


**Figure 3. Single Molecule Sequencing Platforms**

(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides a present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting i detectable fluorescent signal that is captured in a video.

(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme as well as a teth whereas the second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the p are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

Excerpted from Reuter et al 2015 - Molecular Cell

# Outline

1. Different methods to acquire sequence data
2. Understand sequence file formats
3. **Preparing files for analysis**
    - Tutorial looking at sequence data files and quality

# Part 3: Preparing files for analysis

*What do you do when you get your data?*

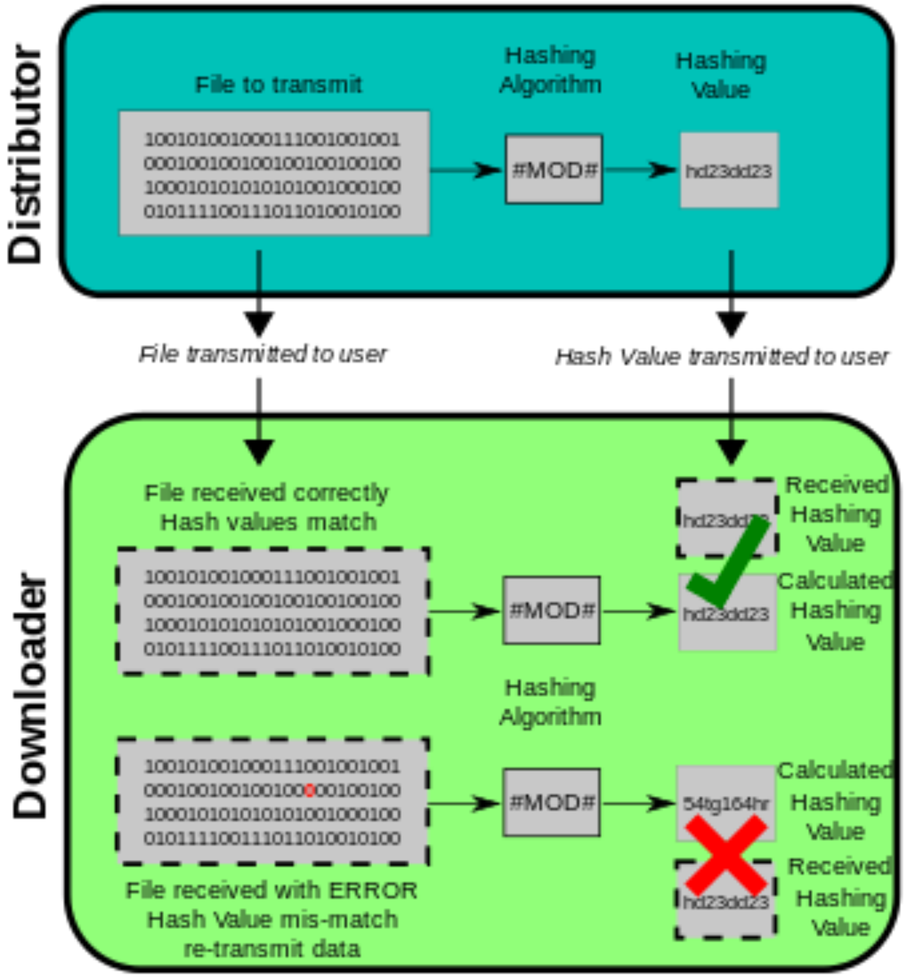
# Part 3: Preparing files for analysis

1) Check files for completeness, use checksums if file corruption is suspected

Downloading large data files takes a long time

There is a possibility of data corruption when files are downloaded

There are command line tools for verifying data complete-ness



*MD5 and SHA-1 are the most common checksum methods*

*There is a short demonstration using SHA-1 sums in the tutorial*

# Part 3: Preparing files for analysis

1) Check files for completeness, use checksums if file corruption is suspected

2) Inspect quality statistics

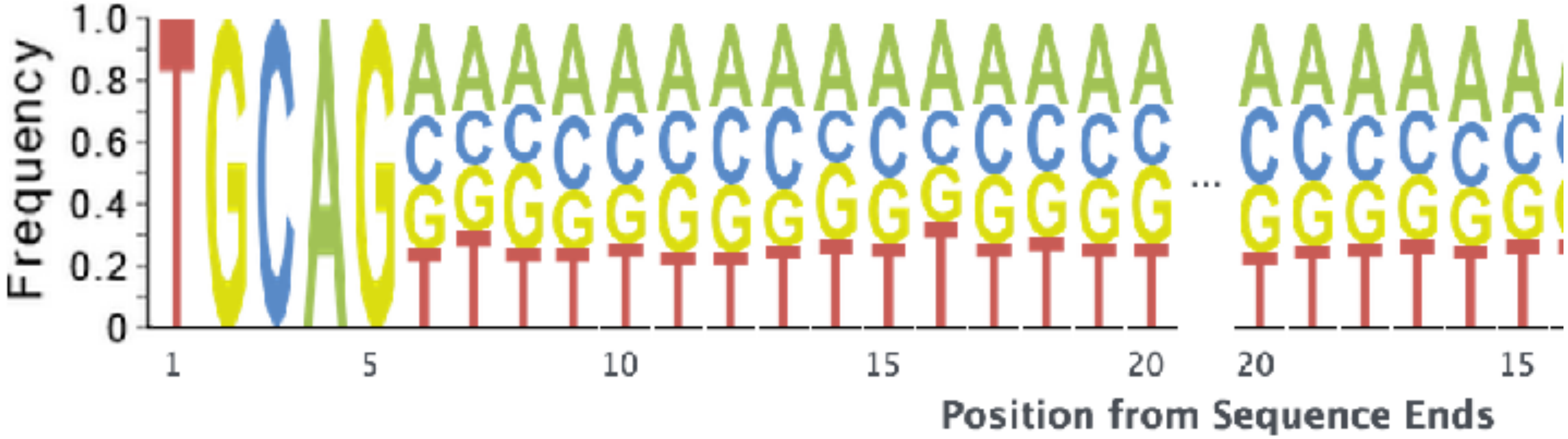
There are many possible statistics to query:

- Number and length of sequences
- Base qualities**
- Poly A/T tails
- Presence of tag sequences (things that you added during library prep.)
- Sequence complexity (e.g. identify repetitive data ATATATATATATATATATATATA)

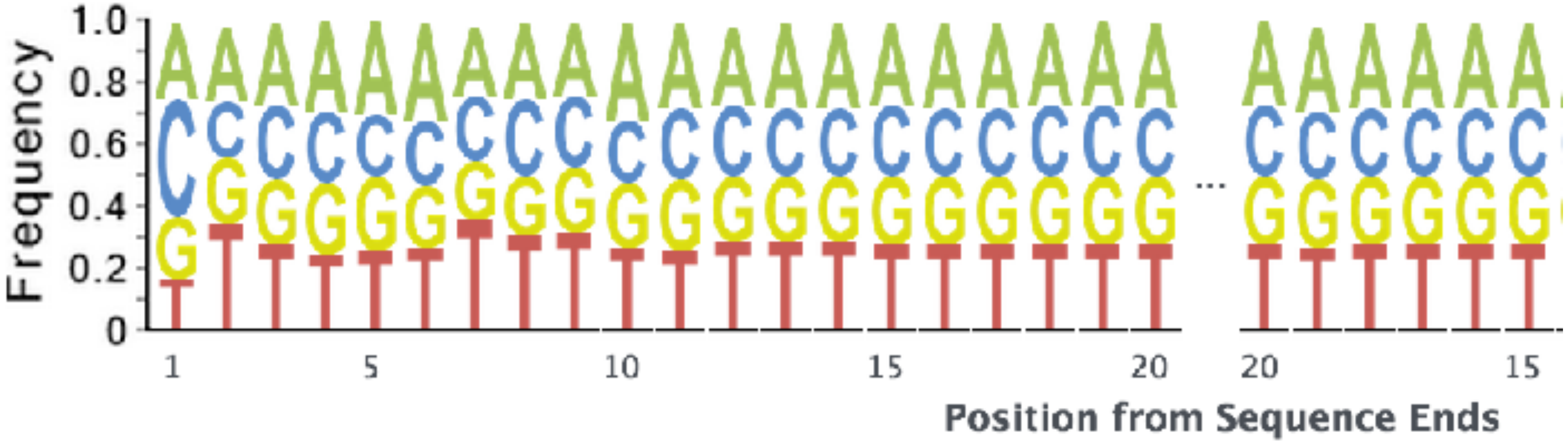
***There are standard tools for examining these, such as prinseq and fastqc***

# Part 3: Preparing files for analysis

Distribution of base frequencies in GBS reads  - with enzyme cut site

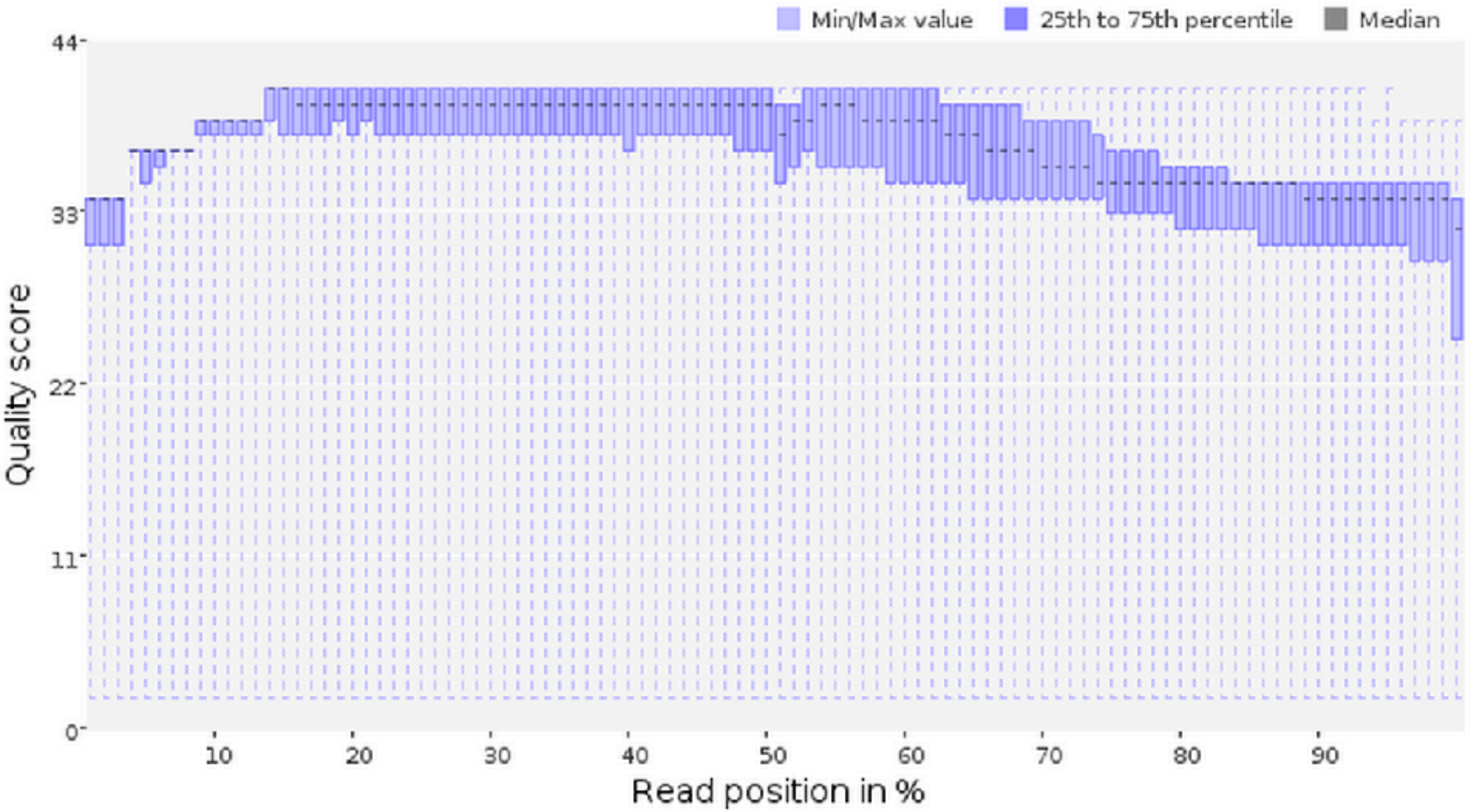


Distribution in RNAseq data - with no adapters/tags used



# Part 3: Preparing files for analysis

A typical quality score distribution for Illumina reads



# Part 3: Preparing files for analysis

1) Check files for completeness, use checksums if file corruption is suspected

2) Inspect quality statistics

3) Possible steps to clean files

– De-multiplex

– Trim adapters

– Filter/trim low quality base calls

– Remove duplicate sequences

– Remove contaminant sequences

– Remove sequences that are mainly adapter

Often done by the sequencing centre

Important for genotyping and RNAseq

Important for reference assembly

Many programs to implement these steps!

# Part 3: Preparing files for analysis

## Quality trimming

Choice of quality score to filter to depends upon the application:

- Too low a quality score cutoff:
  1. increase run times and RAM usage
  2. Bad results (e.g. false SNP calls)
- Too high a quality score cutoff:
  1. Faster run times
  2. Potentially lose useful data (e.g. more fragmented assemblies or missing SNPs)

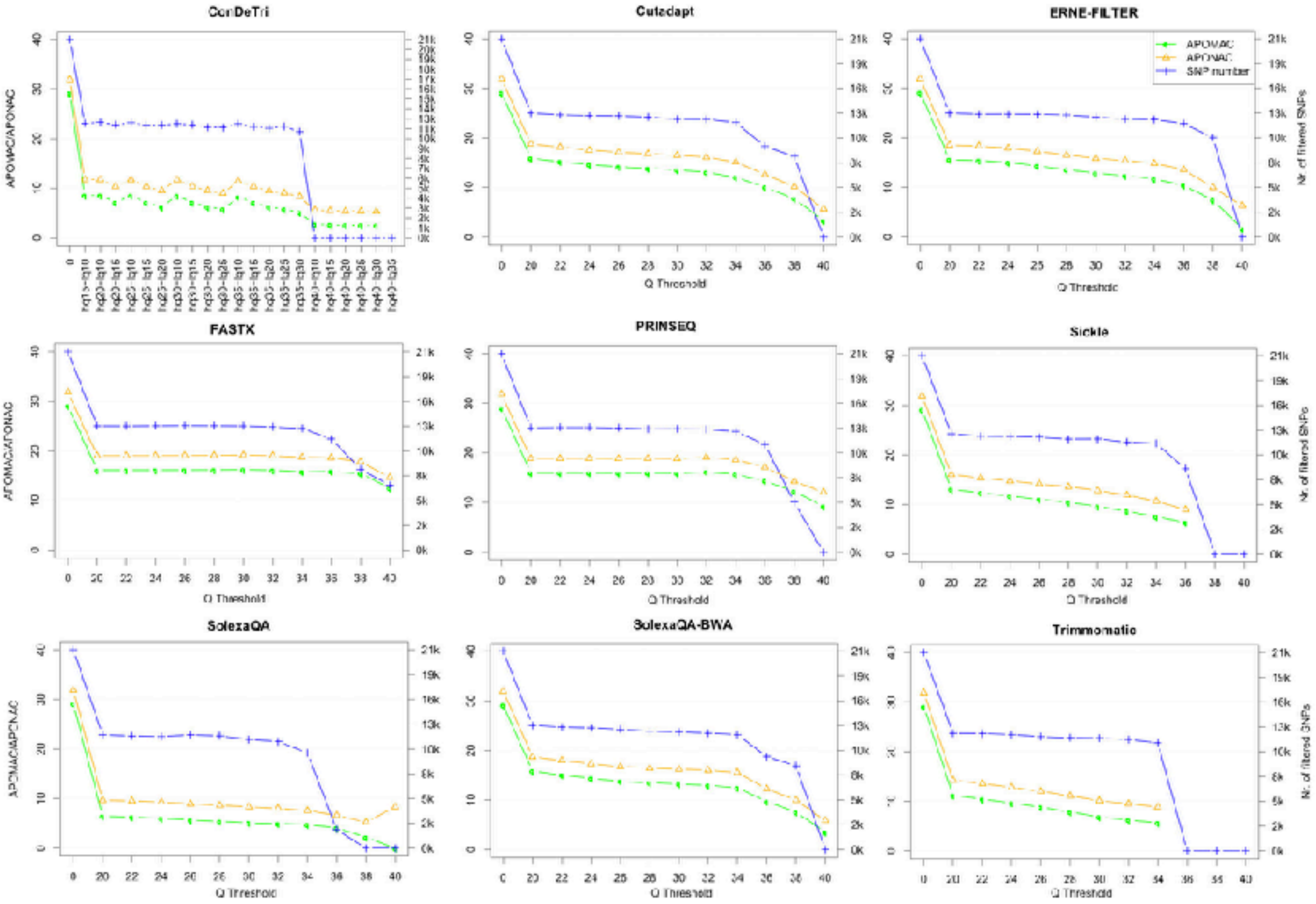
Q20 is a rule of thumb, but it depends on what you’re doing

# Part 3: Preparing files for analysis

## Quality trimming

*Del Fabbro et al 2013*

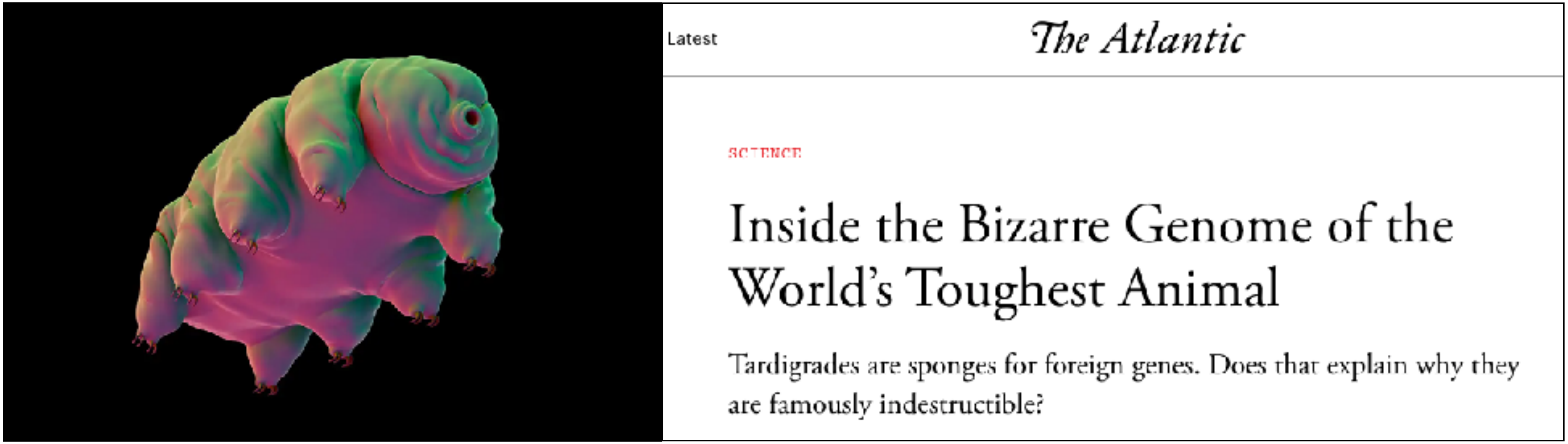
**Number of variants detected**



# Part 3: Preparing files for analysis

## Contamination

**Contamination in your samples can lead to big errors downstream**



Latest

The Atlantic

SCIENCE

Inside the Bizarre Genome of the World's Toughest Animal

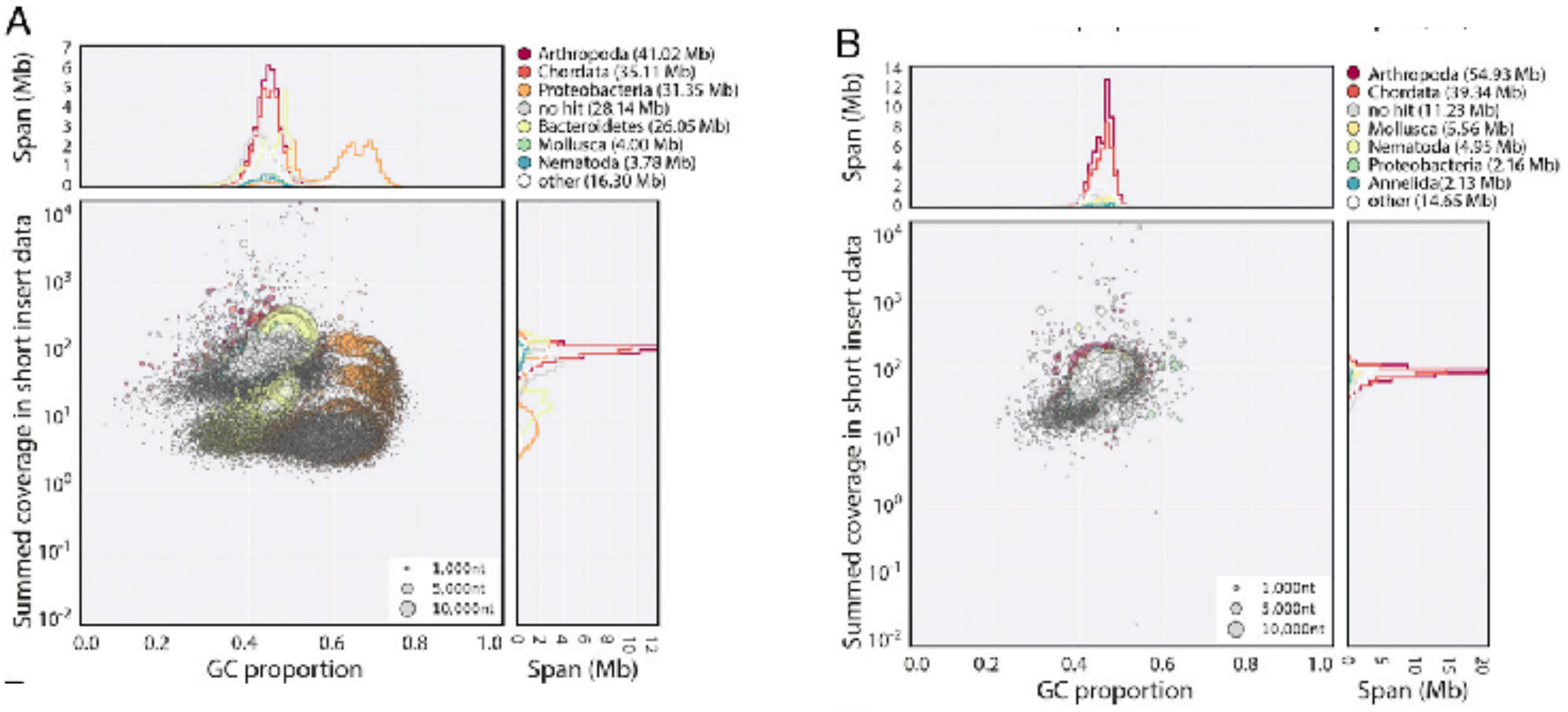
Tardigrades are sponges for foreign genes. Does that explain why they are famously indestructible?

*Koutsovoulosa et al 2016 PNAS*

# Part 3: Preparing files for analysis

## Contamination

**Contamination in your samples can lead to big errors downstream**

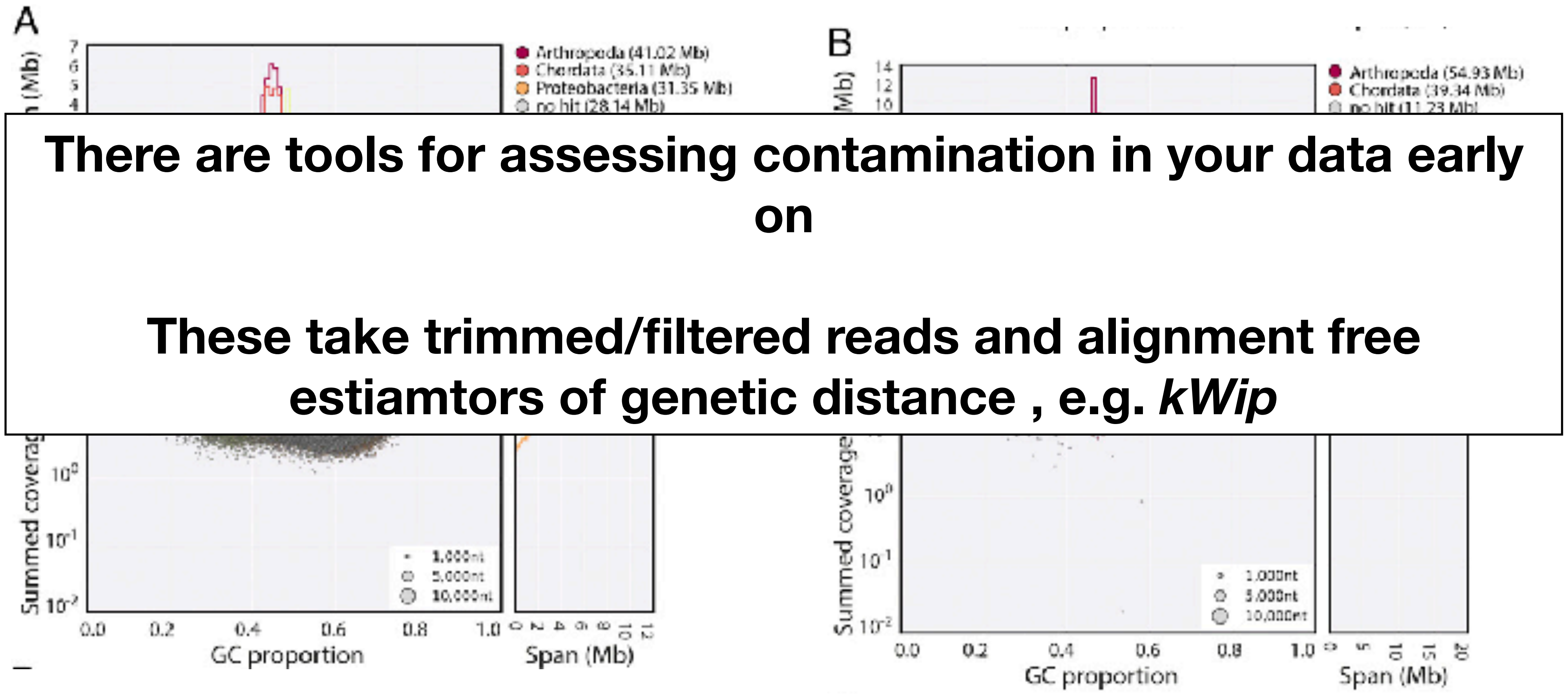


*Koutsovoulosa et al  2016 PNAS*

# Part 3: Preparing files for analysis

## Contamination

**Contamination in your samples can lead to big errors downstream**



**There are tools for assessing contamination in your data early on**

**These take trimmed/filtered reads and alignment free estiamtors of genetic distance , e.g. *kWip***

*Koutsovoulosa et al  2016 PNAS*

# Outline

1. Different methods to acquire sequence data
2. Understand sequence file formats
3. Preparing files for analysis
    - **Tutorial looking at sequence data files and quality**

## Tutorial:
## Work through the tutorial associated with this session