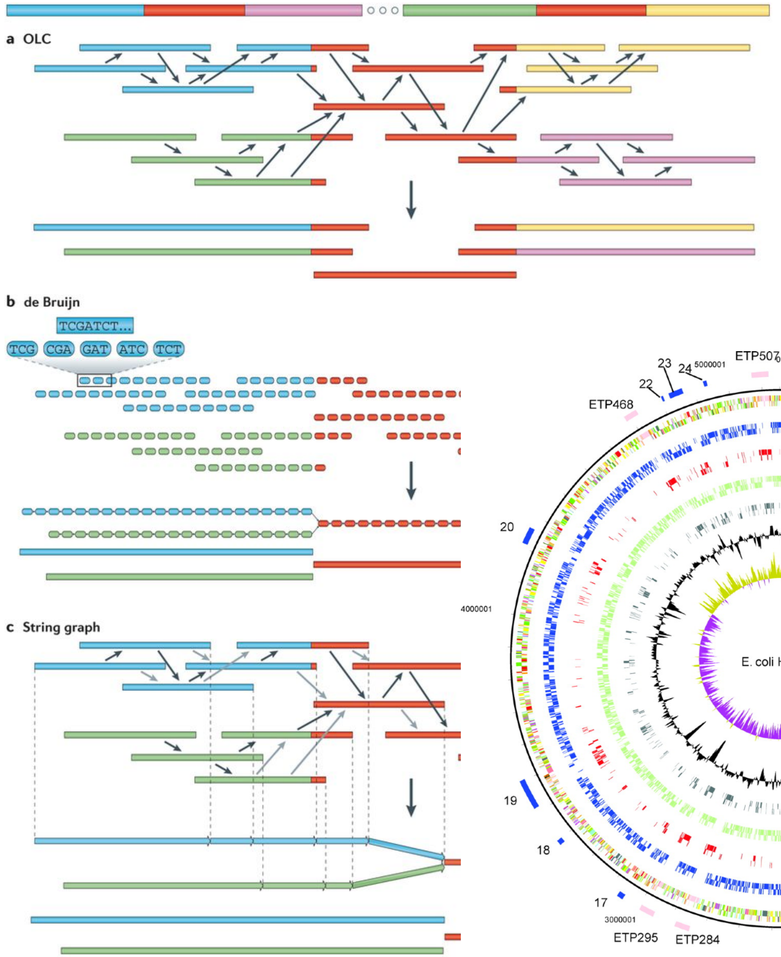
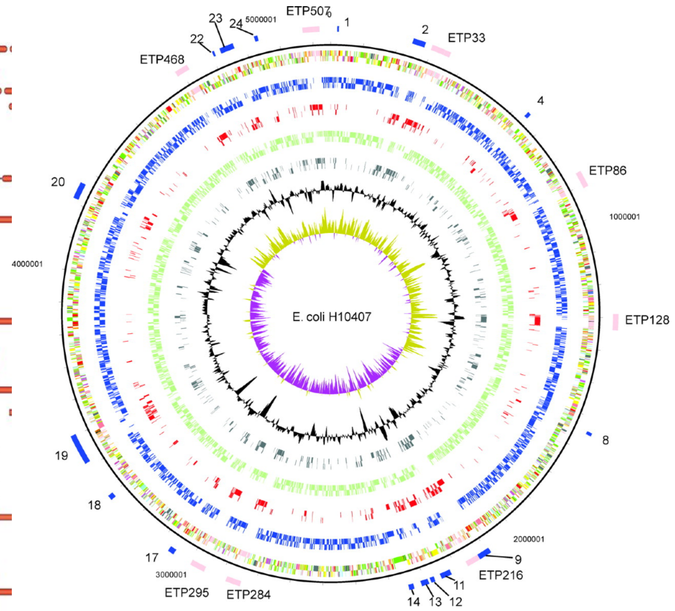
Sequencing Technologies



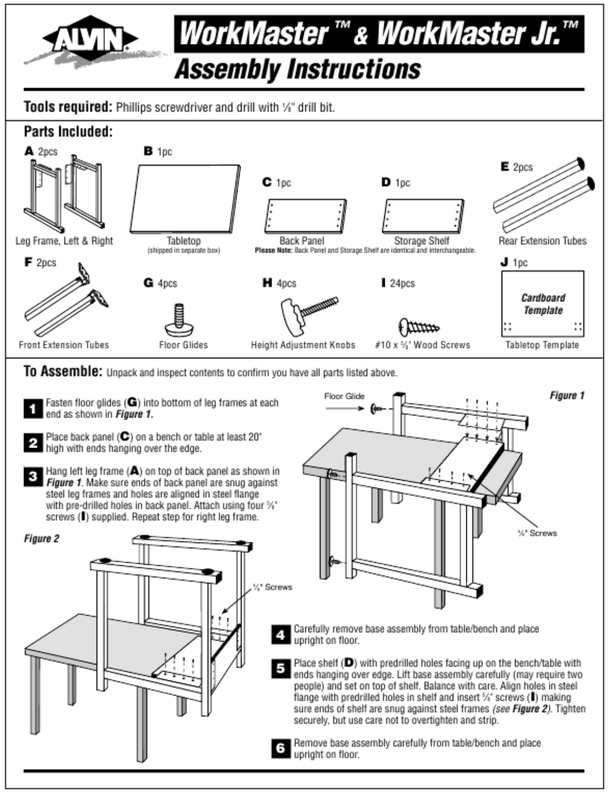


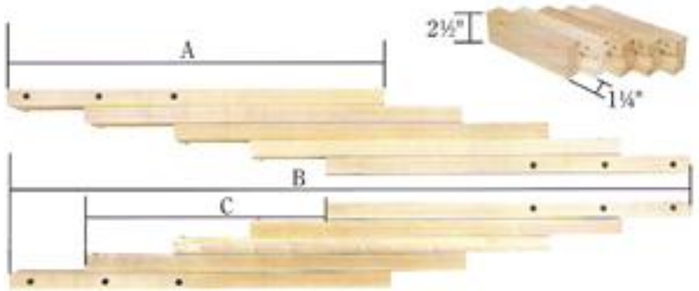
HTS Workshop Genomics & Transcriptomics
KAUST 2019

Robert Lehmann
Octavio Salazar

Why do we want to know the sequence?

We want to know what makes an organism, be.





We can think of a genome as an instruction manual

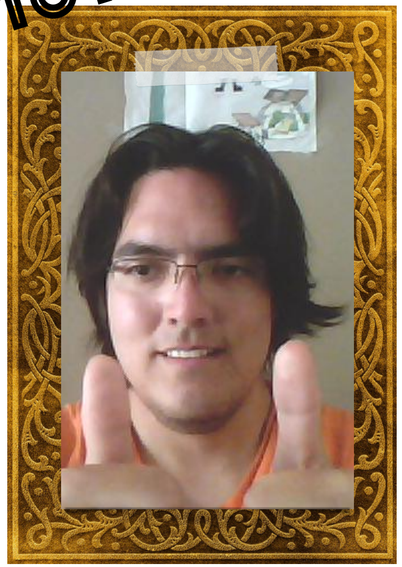
Why do we want to know the sequence?

We want to know what makes species different from each other

HOW TO MAKE A

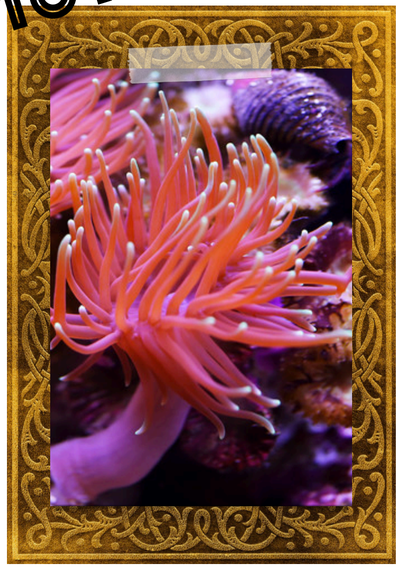
Monkey

HOW TO MAKE A



Uglier monkey

HOW TO MAKE A



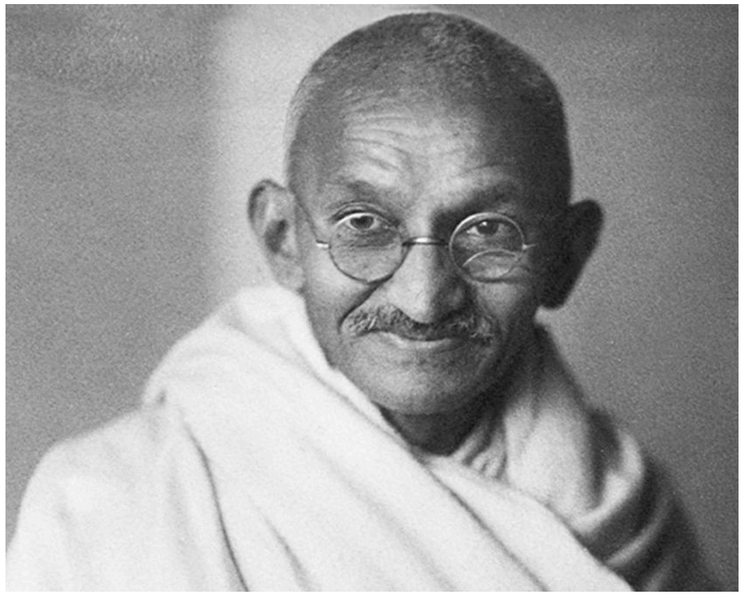
Anemone

HOW TO MAKE A

Tomato

Why do we want to know the sequence?

Not only between species but sometimes we <u>need</u> to know why individuals are different, ahem, special...

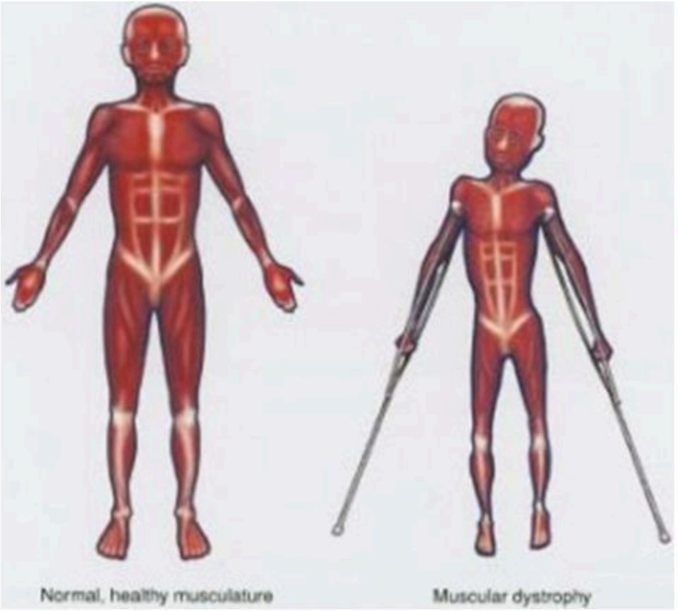


Individuals within a species can have great variation!

Study genetics of disease
e.g. Duchenne Muscular Dystrophy

Why do we want to know the sequence?

What is Duchenne Muscular Dystrophy?

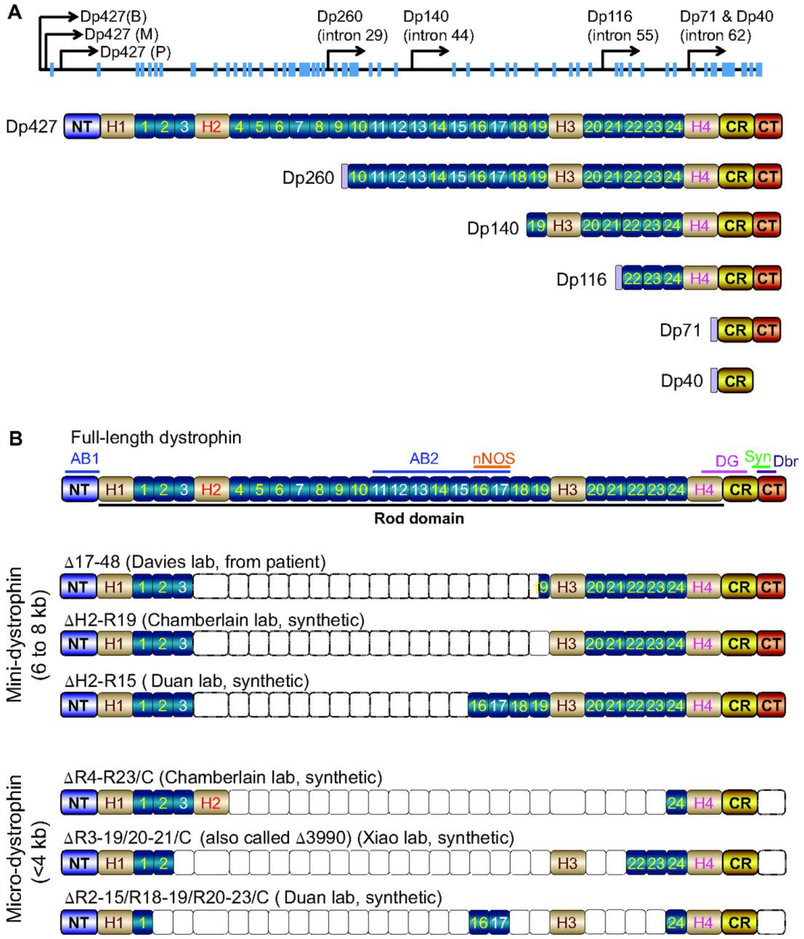


- Duchenne Muscular Dystrophy (DMD) is a type of muscular dystrophy is a recessive genetic disorder causing the fast deterioration of muscles.

Still, no excuse not to be a good bioinformatician!

Why do we want to know the sequence?

Study genetics of disease
e.g. Duchenne Muscular Dystrophy

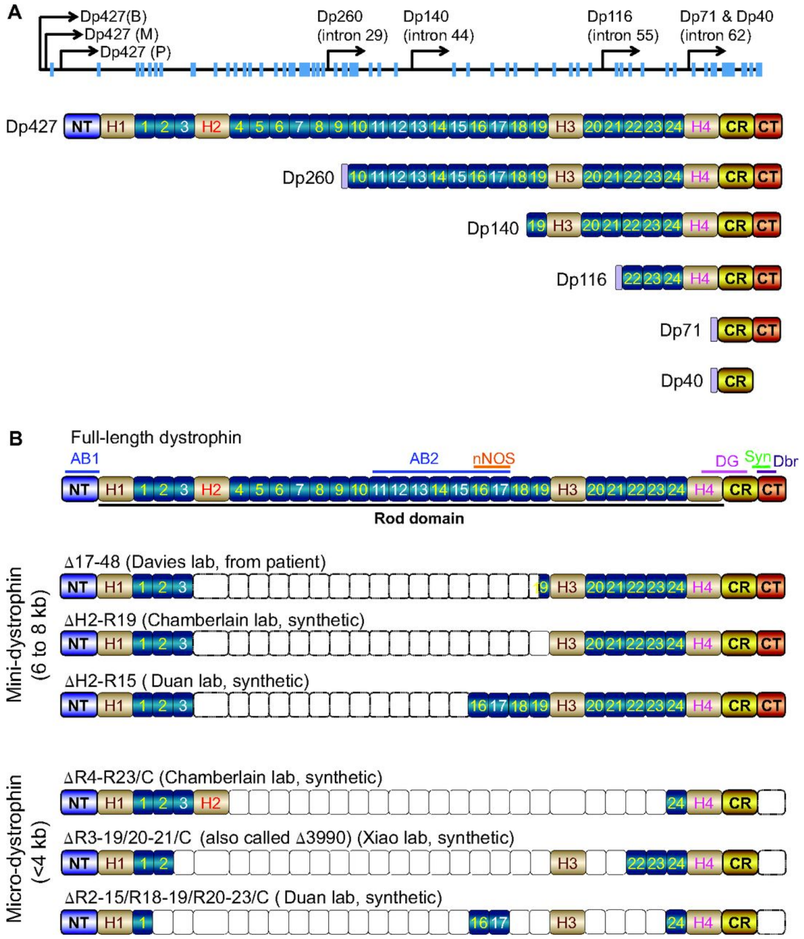


Blame the mother!…

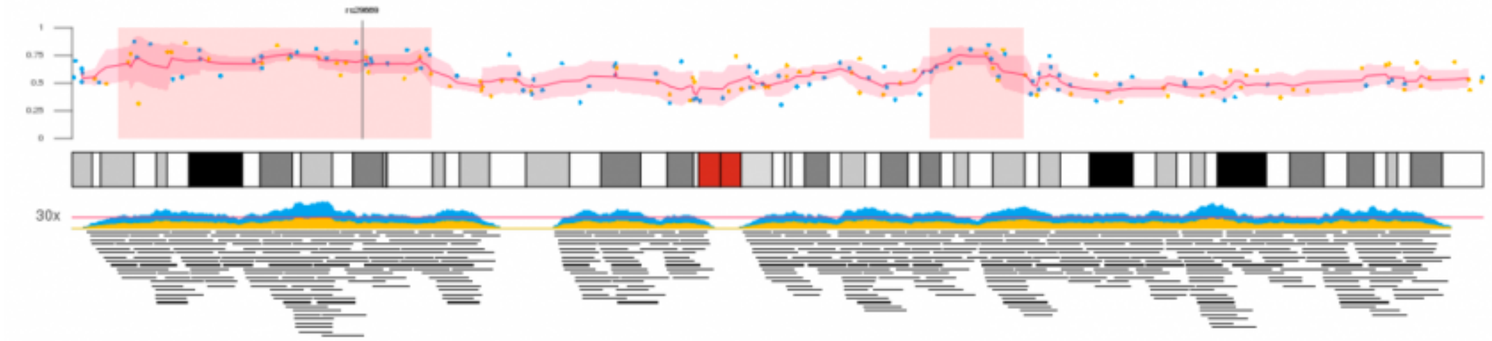
Deletion in a gene in the X chromosome leads to DMD

Why do we want to know the sequence?

Study genetics of disease e.g. Duchenne Muscular Dystrophy



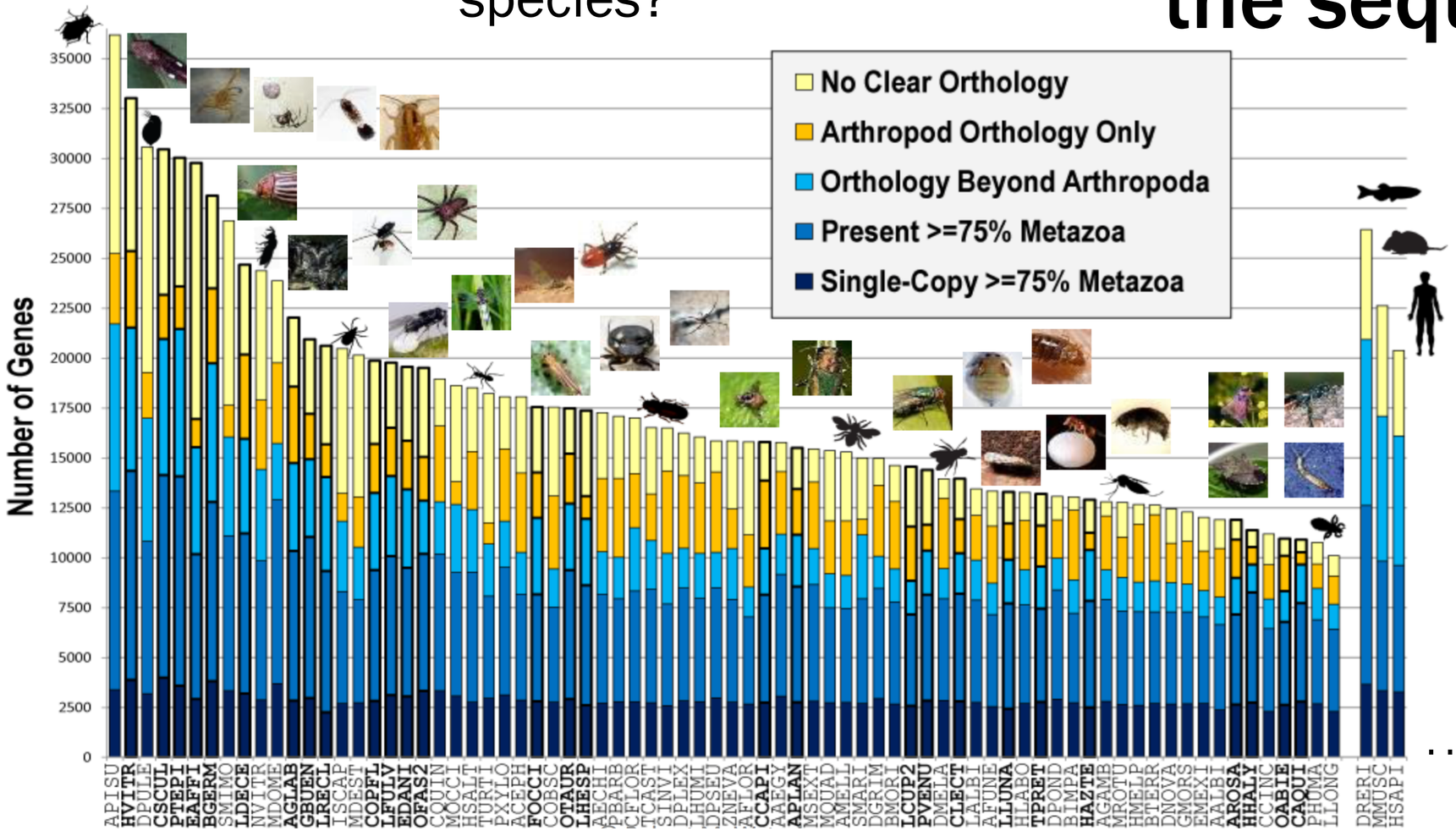
Transcriptomic studies can help us identify mutations such as this

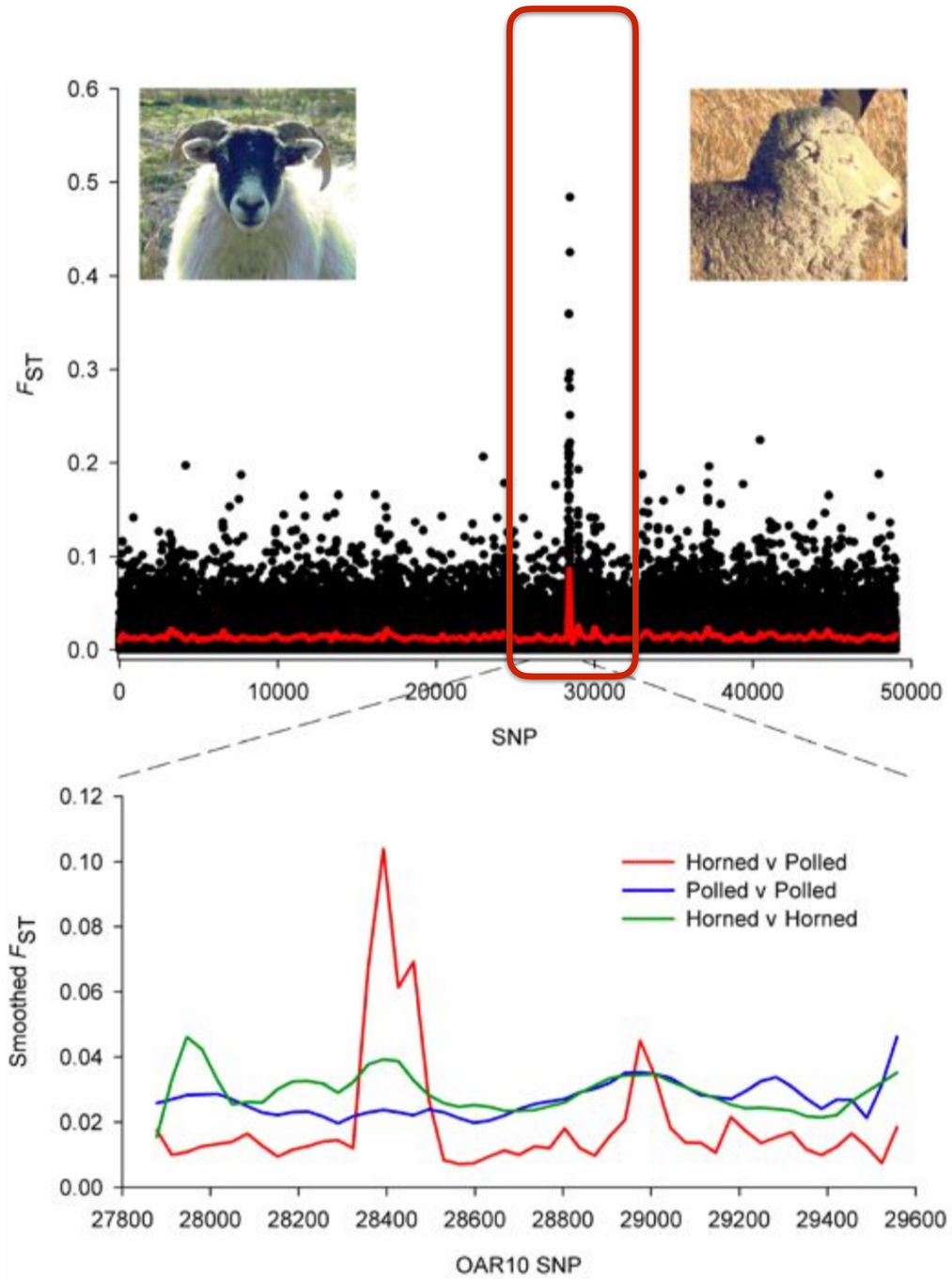


Study evolution

Which genes are shared, which specific to different species?

Why do we want to know the sequence?



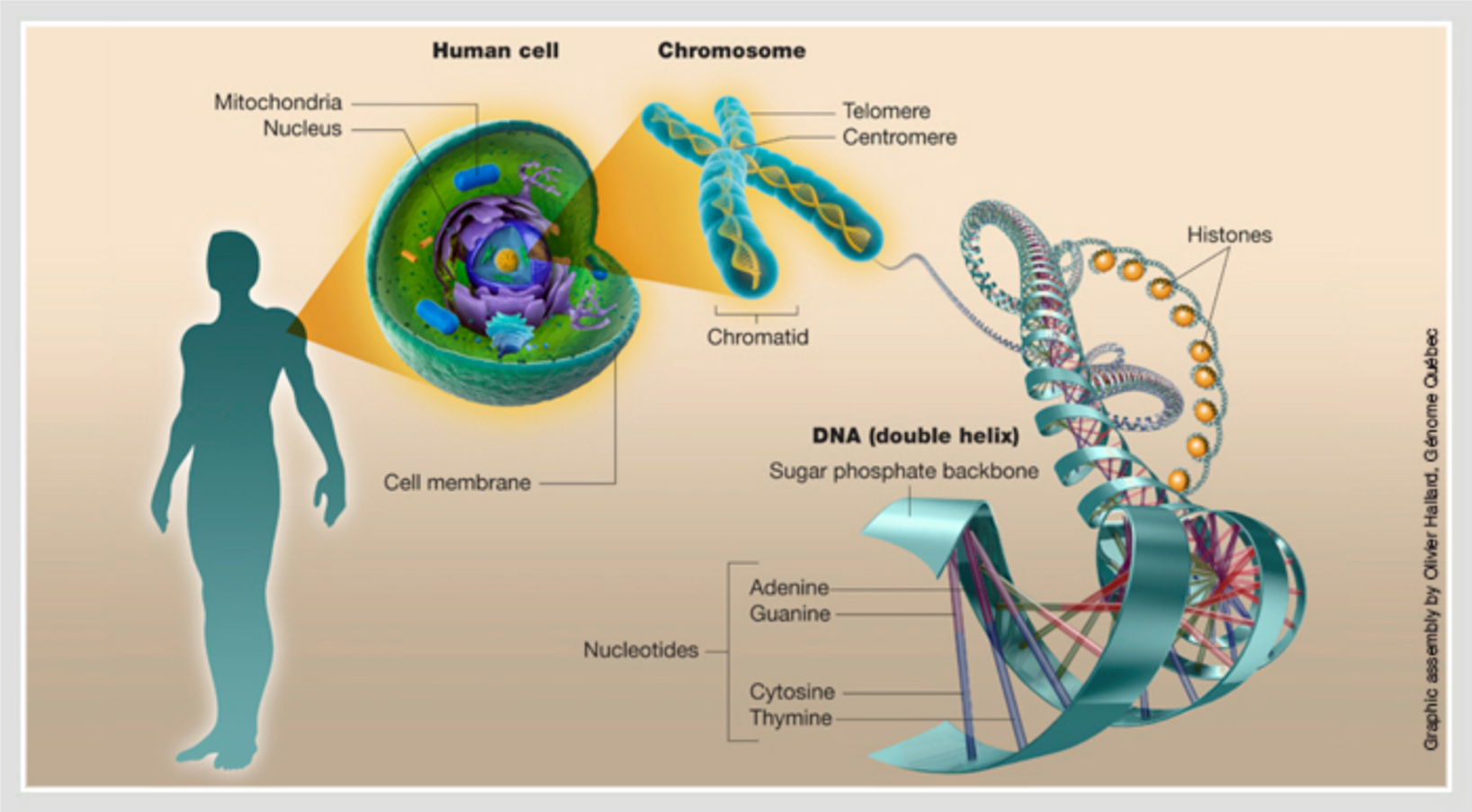


Why do we want to know the sequence?

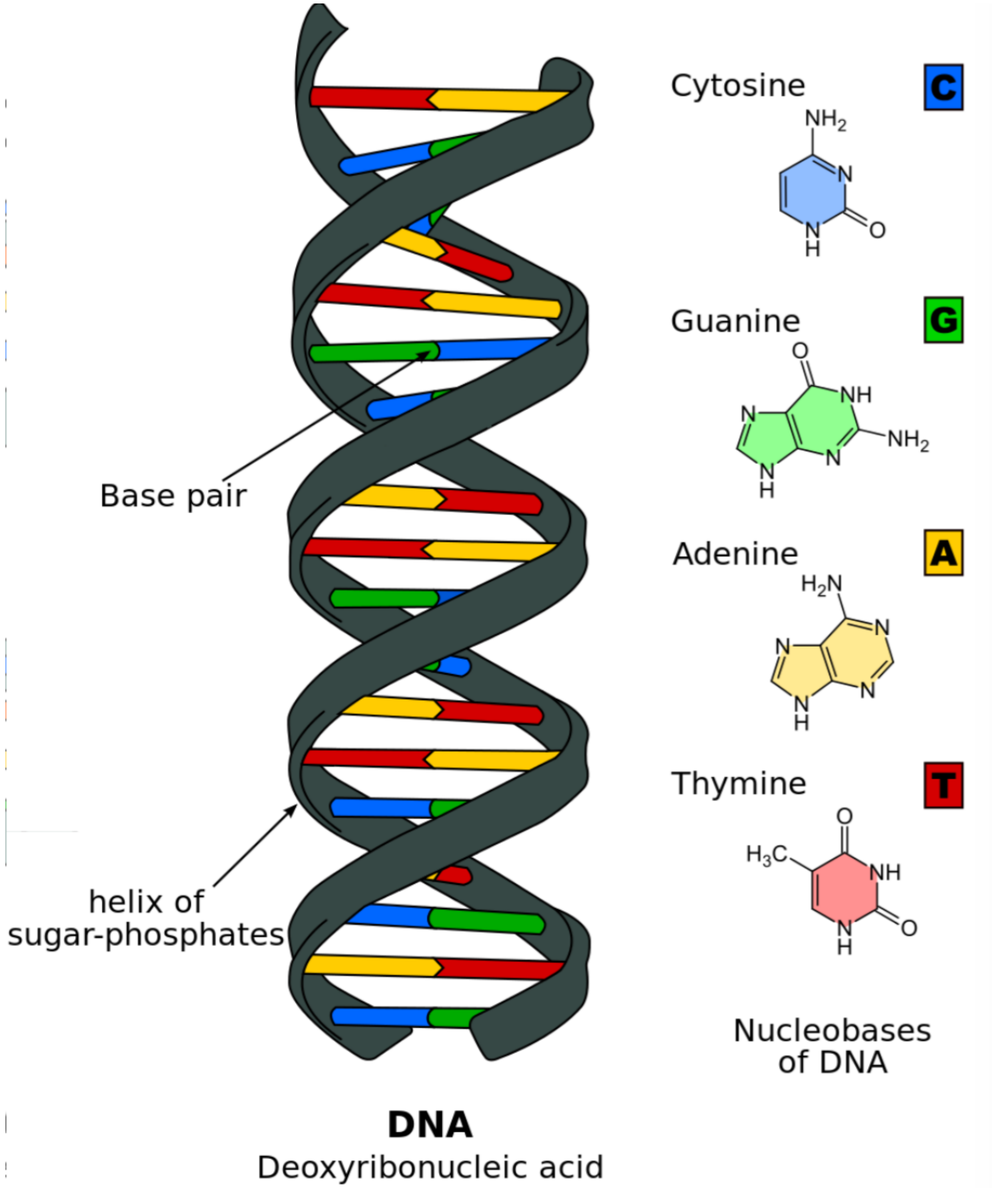
Or study even more important questions!

Population genetics - which genes give sheep horns?

Where is the genetic information stored?

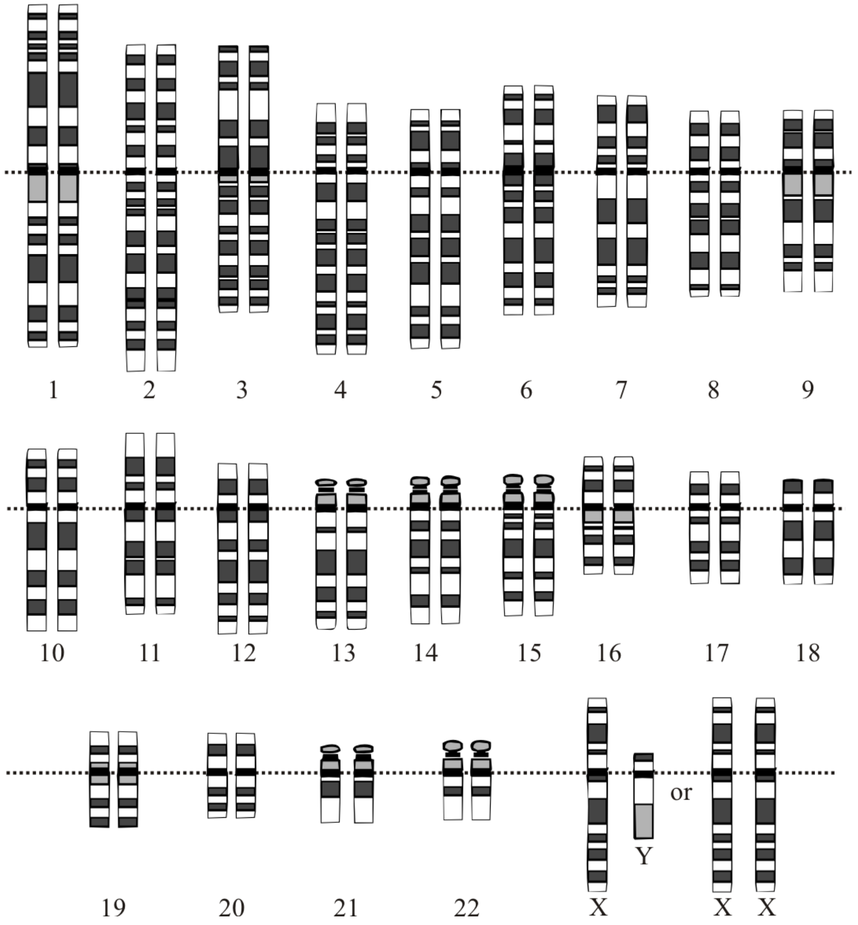


The human Karyotype



...ACGTGATTGACAGGCAGTT...

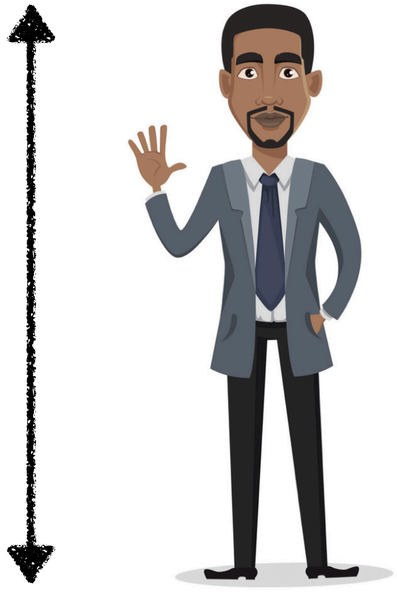
The human Karyotype

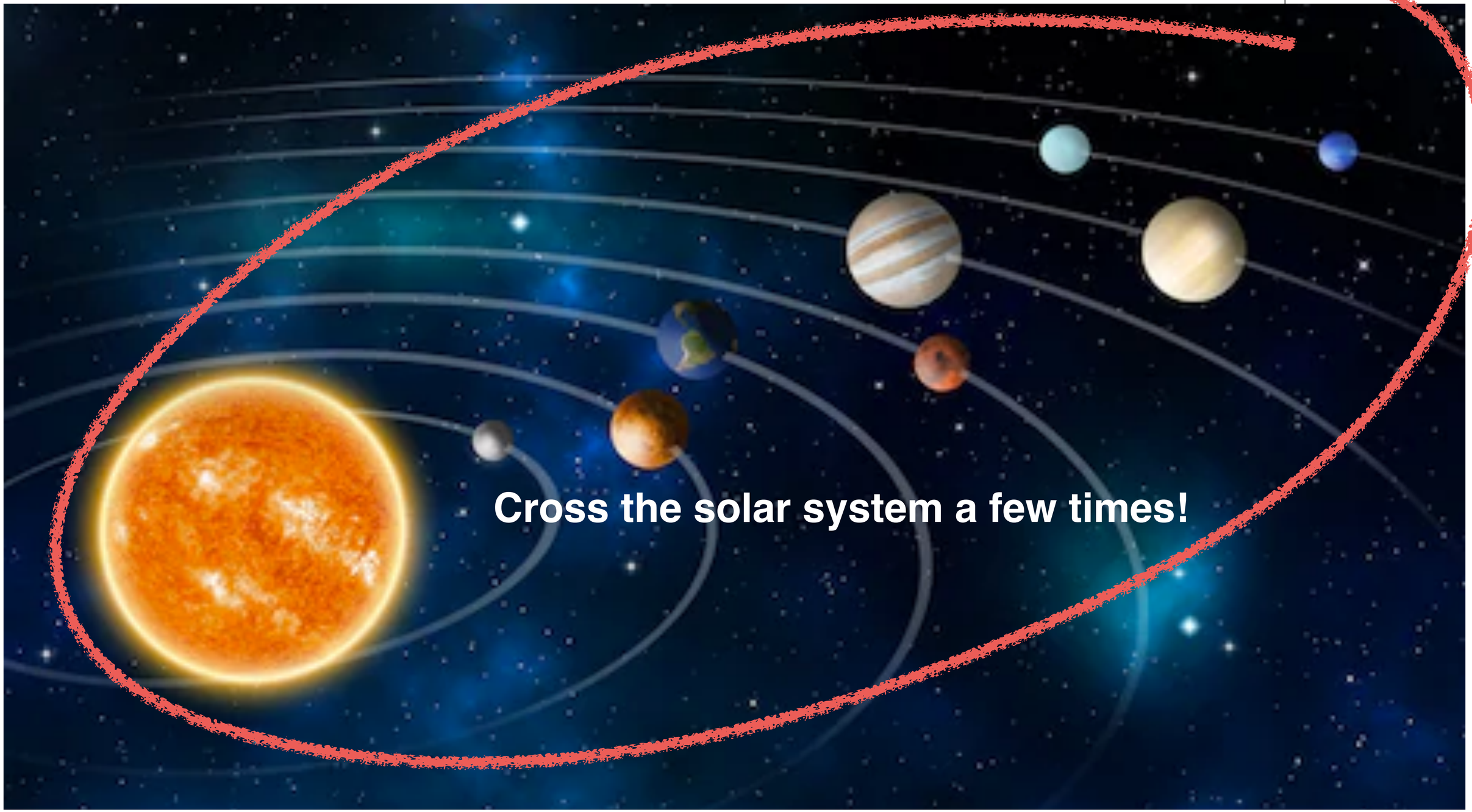


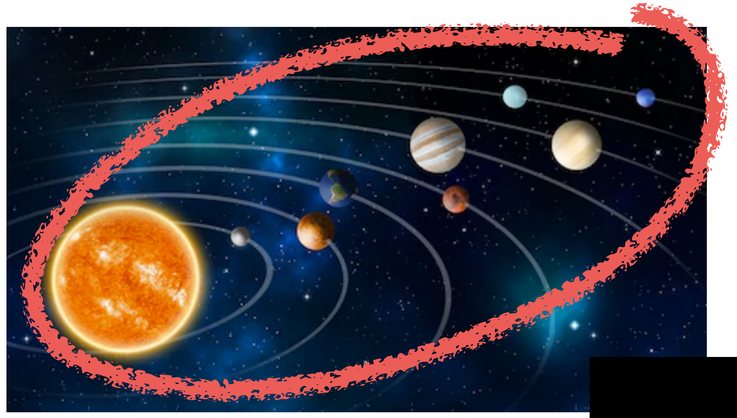
- 22 autosomes 1-22
- 2 allosomes X /Y
- 3.23 billion bp (haploid)
- 6.46 billion bp (diploid)

To put this into perspective

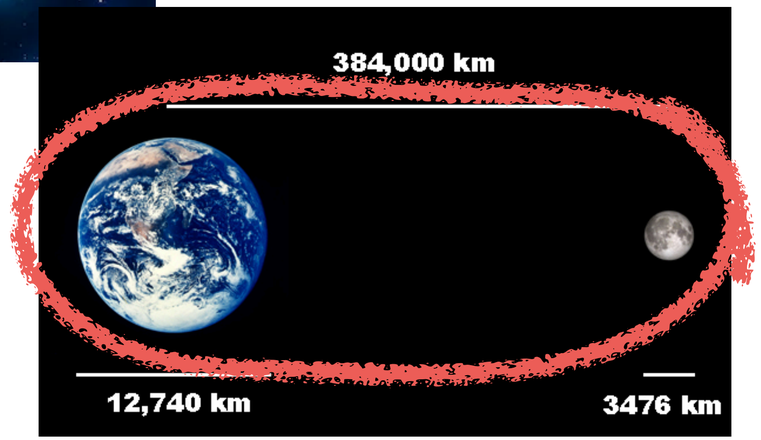
The DNA in one of your cells if stretched = 1.8 m

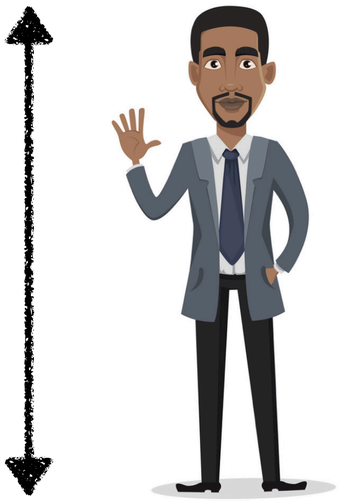


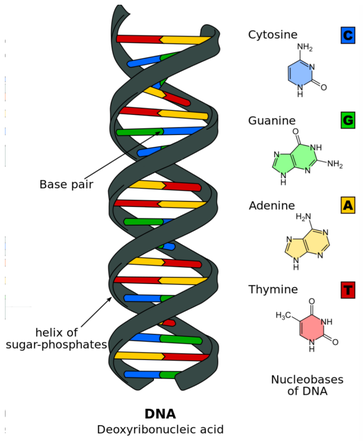

Cross the solar system a few times!



Now the you know that you have the universe within you
Let’s go back a little…

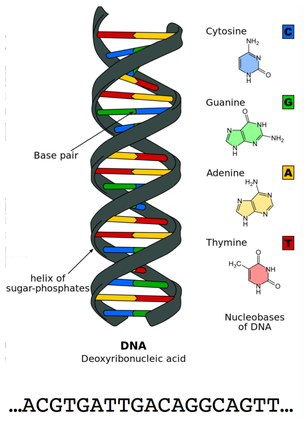






…ACGTGATTGACAGGCAGTT…

Sequence a Genome -What does that mean?



DNA

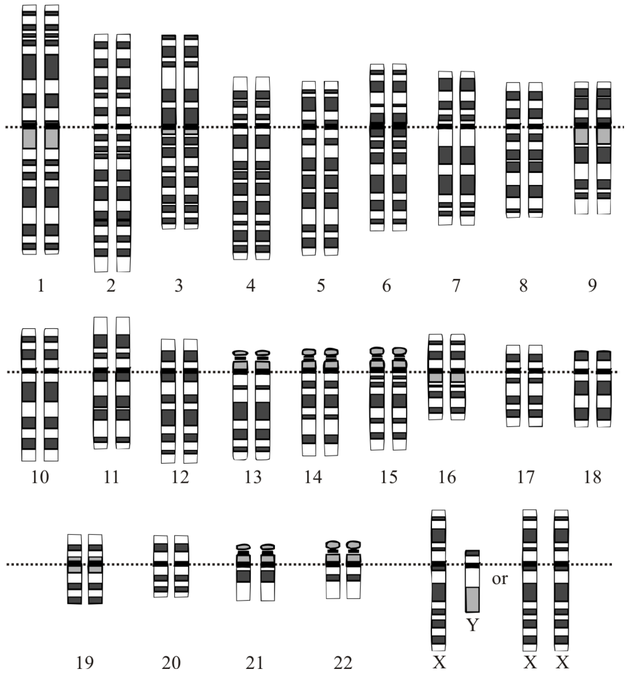
↓

Sequence fragments

↓

Assembler

→



Genome Assembly Pipeline

contigs

Assembly

Scaffolding

Polishing

Assembly

scaffolds / chromosomes

Scaffolding:
chromatin contact map

Gap Filling

Fancy as

Repeat Seq. library

Repeat Characterization

Gene Sequences

Structural Annotation

Gene2function map

Functional Annotation

Analysis

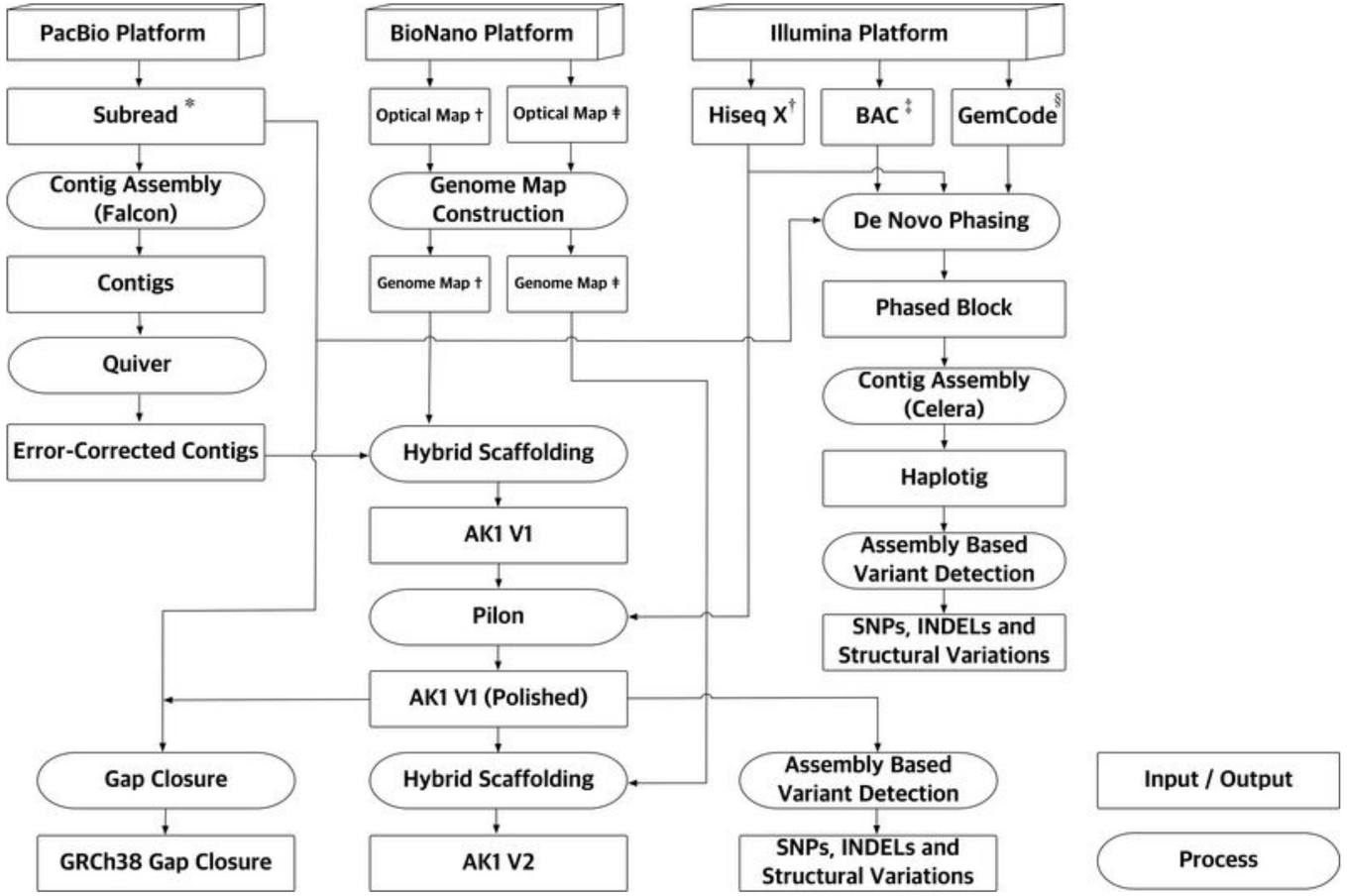
Lol, no

De novo assembly and phasing of a Korean human genome.

Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, Kuk J, Park GH, Kim J, Ryu H, Kim J, Roh M, Baek J, Hunkapiller MW, Korlach J, Shin JY, Kim C

Nature. 2016 Oct 13;538(7624):243-247.

Can involve many many many many steps...



Can involve many many many many steps...

But don’t worry! New technologies have made things easier! Sort of…

Illumina

Oxford Nanopore

PacBio

A brief history lesson…

Making sequence fragments

Making sequence fragments - Sequencing by termination

A brief history lesson…

TGCACTCTAGGACAAGAT

Back in the day sequencing a gene was enough to get a Ph.D.
I feel sorry for you….

AAAA
CCCC
GGGG
TTT
GGG
AA
GG
A
T
C
C
T
G

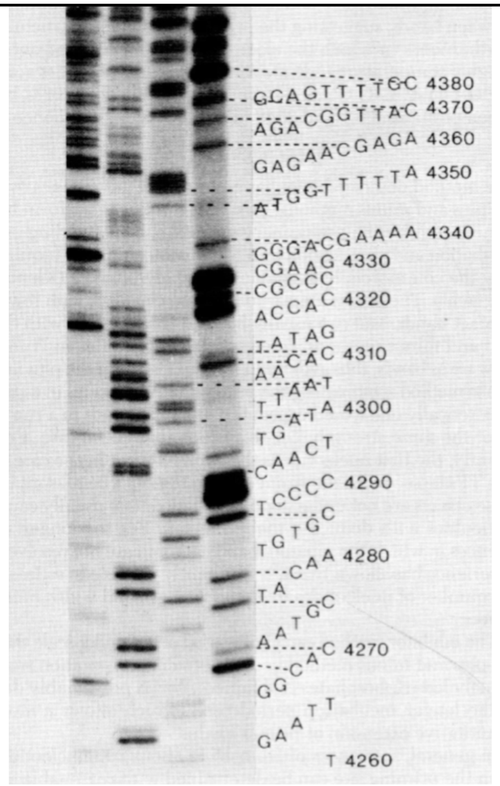
CCC
GGG
TTT
GGG
AAA
GG
AA
T
C
C
T
G
T
T
C
T
A

DNA Pol

AAAAAA
CCCCCC

TT
C
T

DNA Pol



Making sequence fragments

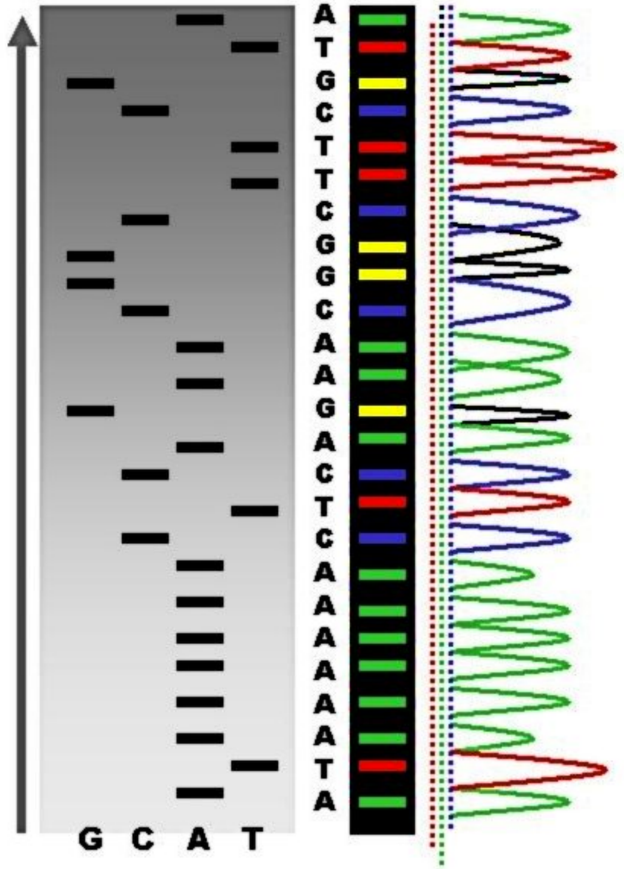
No more radioactivity!

TGCACTCTAGGACAAGAT

AT AT
CG CG

DNA Pol

AAAAAAAAA
CCCCCCCC
GGGGGGG
TTTTTT
GGGGG
AAAA
GGG
AA
T

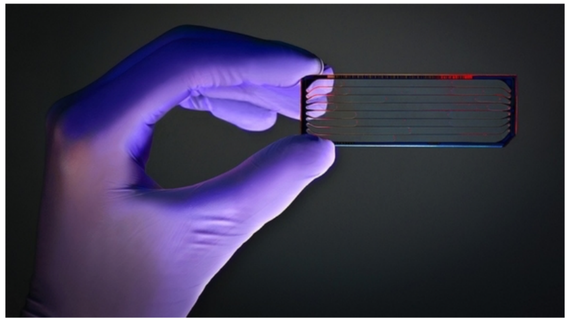


NGS - Sequencing by Synthesis

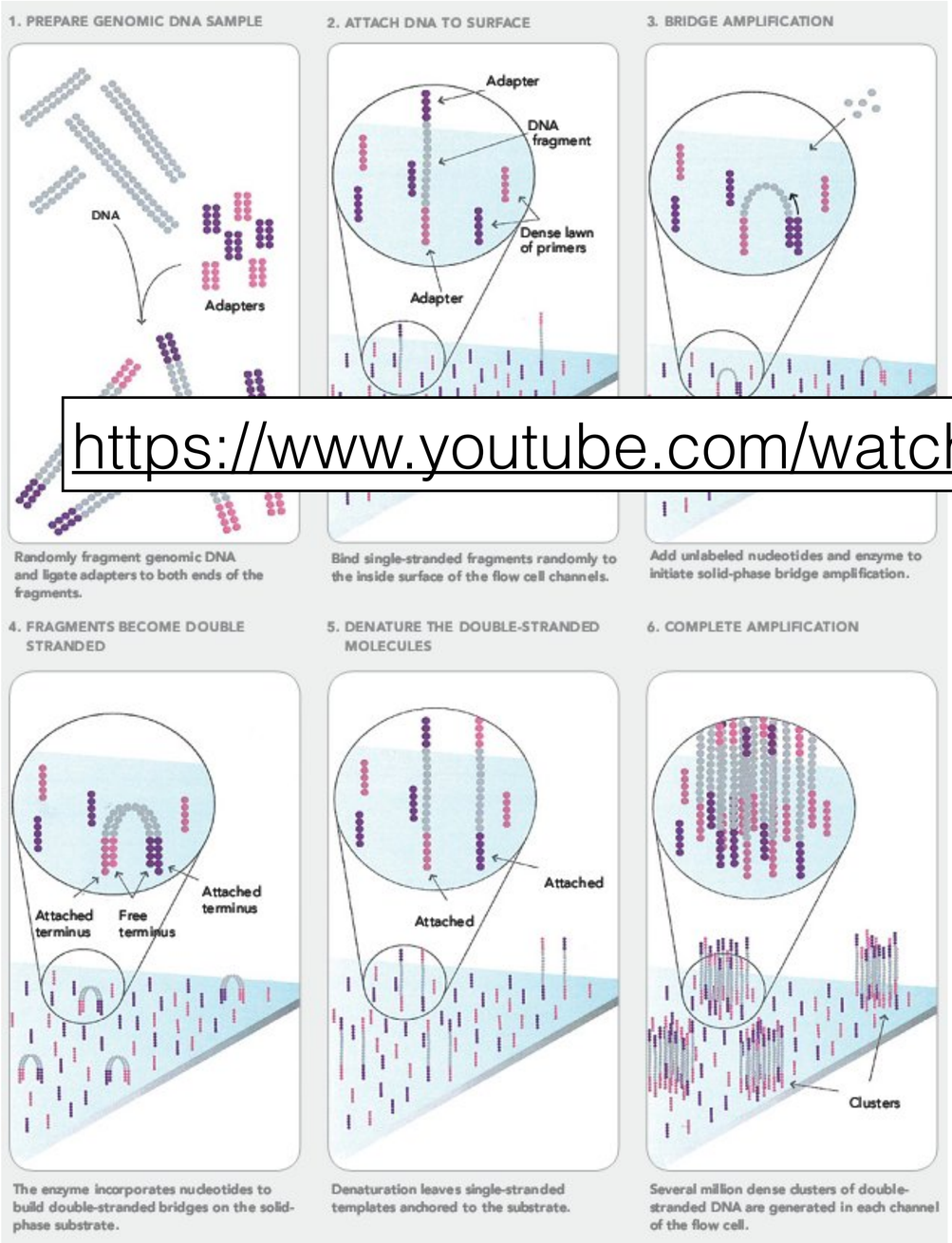
Improvement over termination method:

Watch DNA polymerase during polymerization

Illumina

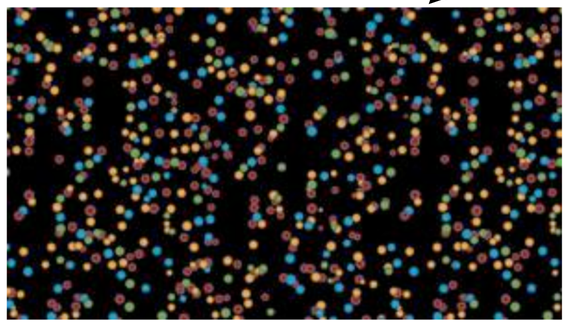


Amplification of fragments

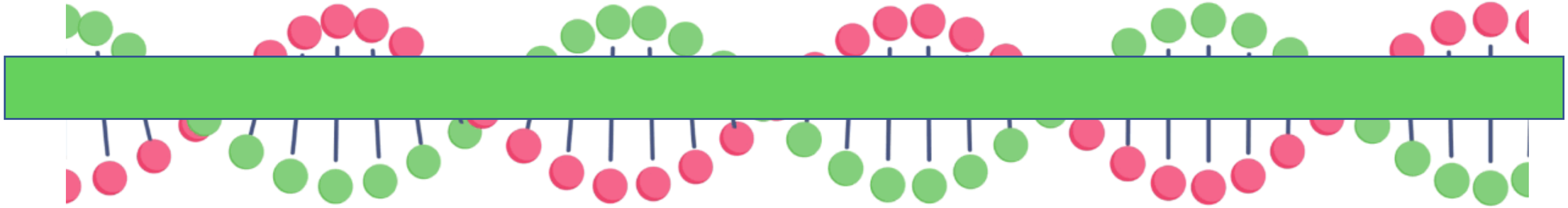


https://www.youtube.com/watch?v=fCd6B5HRaZ8

- Fluorescence signal of single base added to template too weak
- original template amplified into identical sequence “clusters”
- measure light emission per cluster

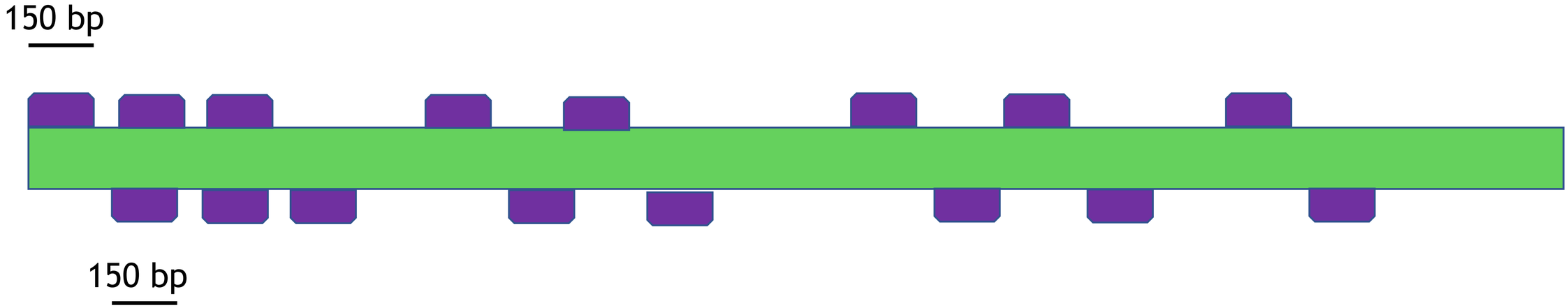


Illumina short read sequencing



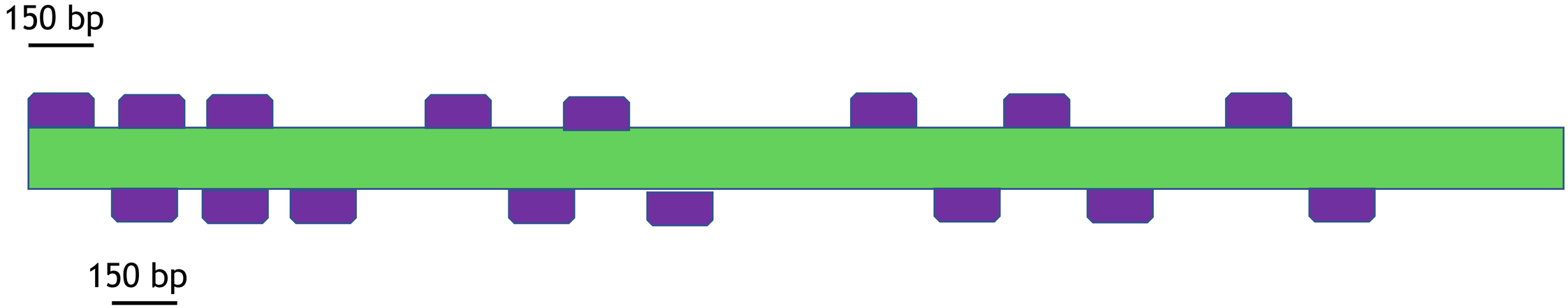
Illumina short read sequencing

Typically



Illumina short read sequencing

Typically



Illumina short read sequencing

Try to stitch together the small pieces of DNA and recreate the original molecule

Works well with small genomes

Sanger vs NGS Illumina sequencing throughput

Sanger throughput: 700bp x 96 = 67 kilobases

Mini

Illumina throughput: 150bp x 2 x (1300- 1500 million) = 400 Gb (~130x human genome) *

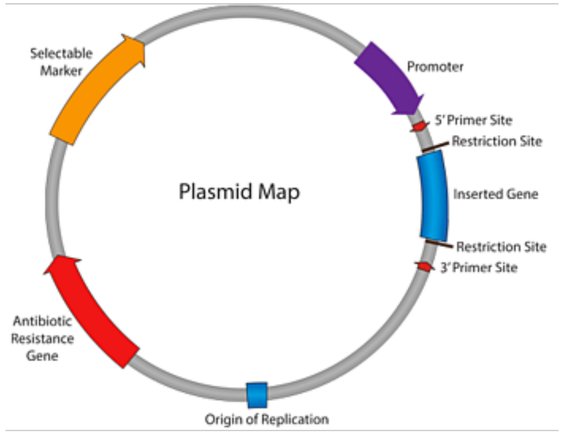
Extra Large

Sanger vs NGS Illumina sequencing throughput

Don’t worry

Everybody is special!

Sanger preferred for plasmid or PCR sequencing due to its longer reads



Illumina is preferred for genome sequencing of small genomes and SNP calling

Illumina short read sequencing

However….

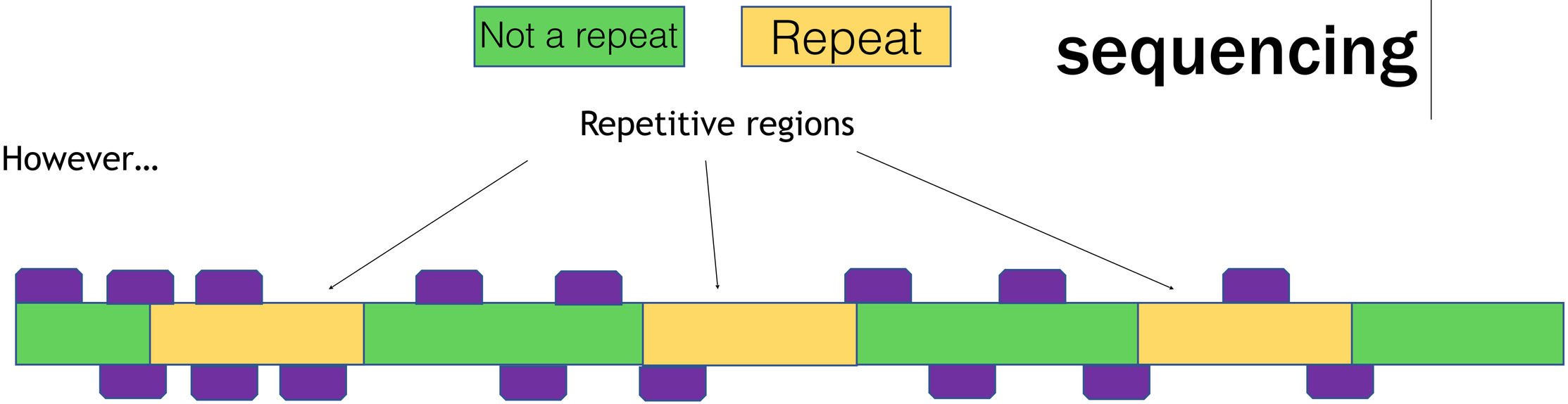
Try to stitch together the small pieces of DNA and recreate the original molecule

Works well with small genomes

Illumina short read sequencing

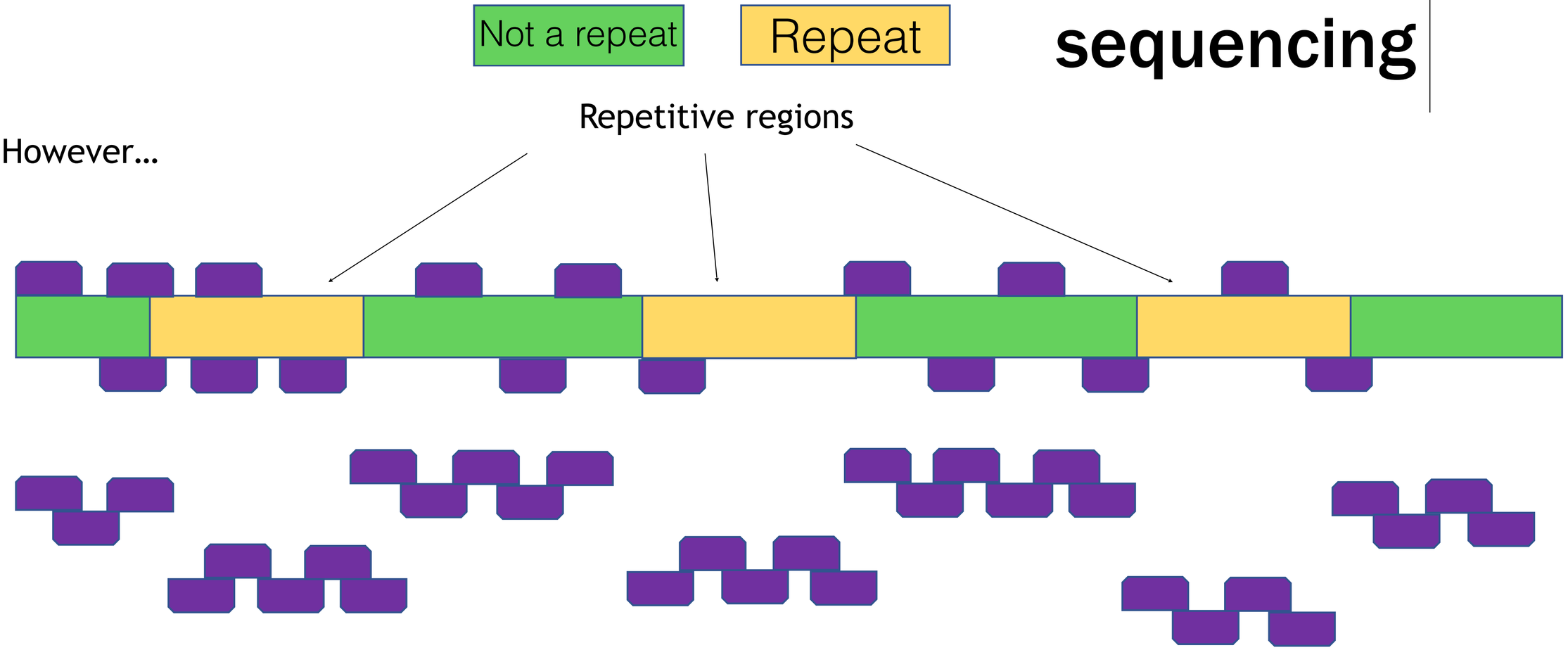
Not the best for repetitive regions!

However…



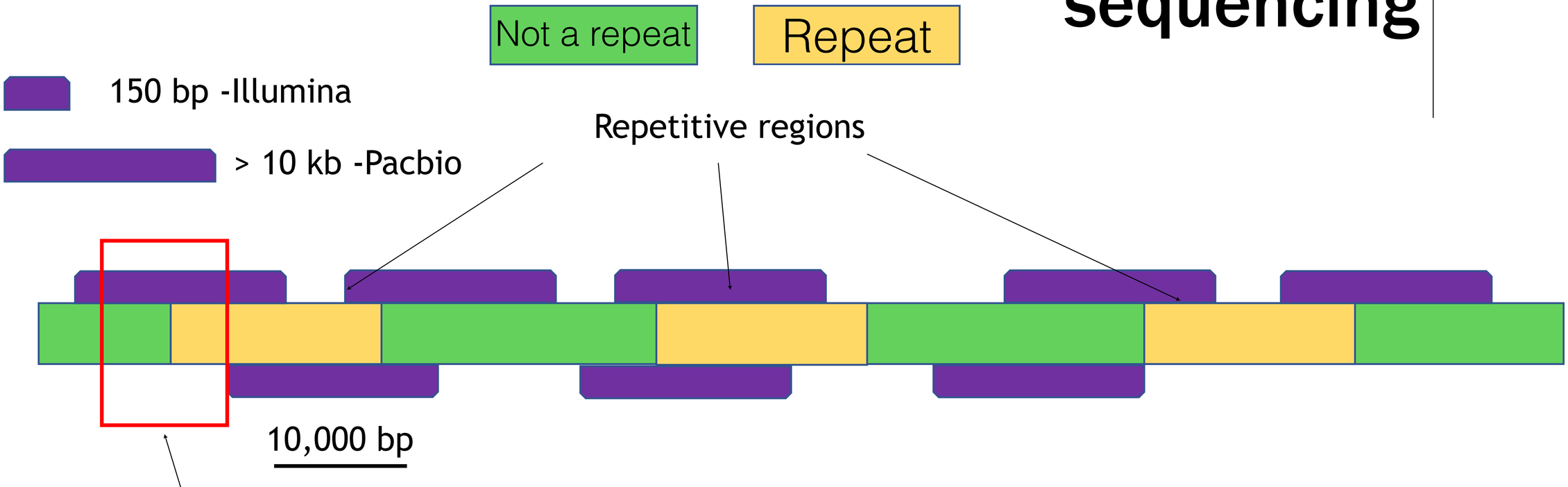
Illumina short read sequencing

However…



PacBio long read sequencing

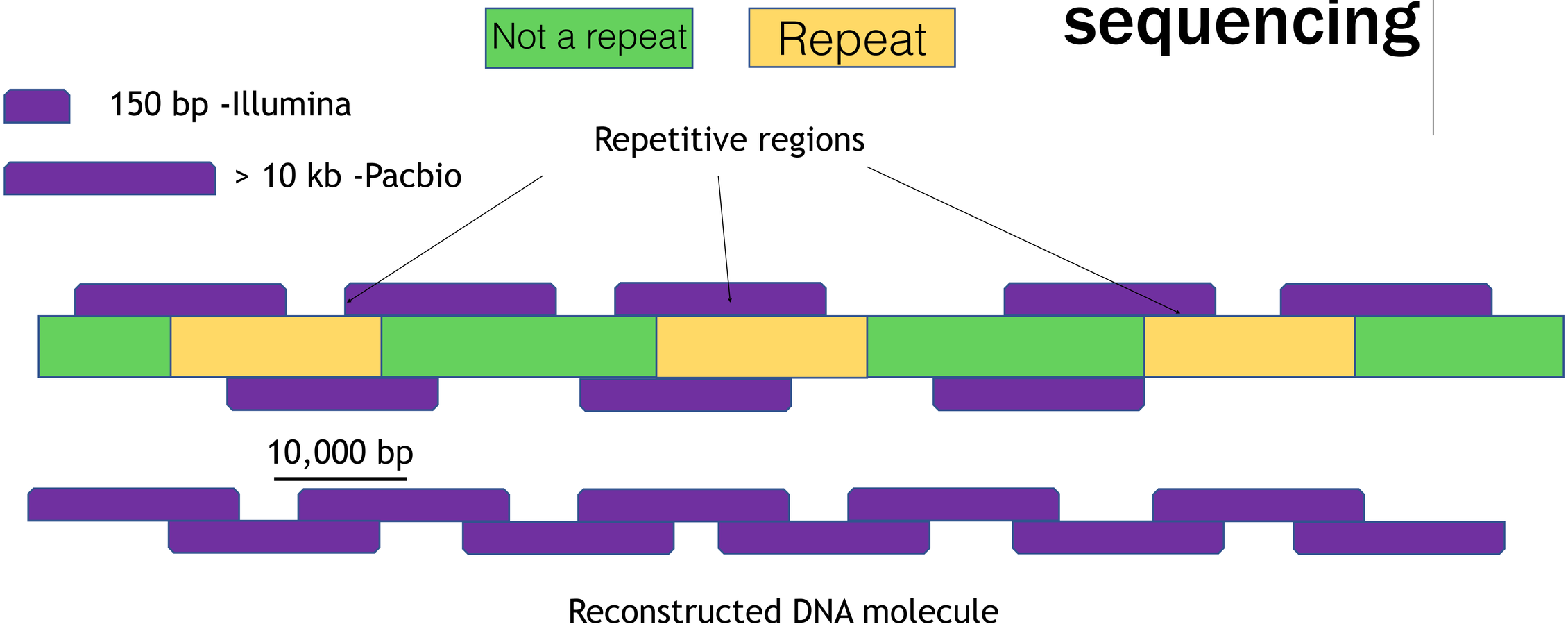
PacBio SMRT sequencing



Can more easily span through repetitive and non-repetitive regions

PacBio long read sequencing

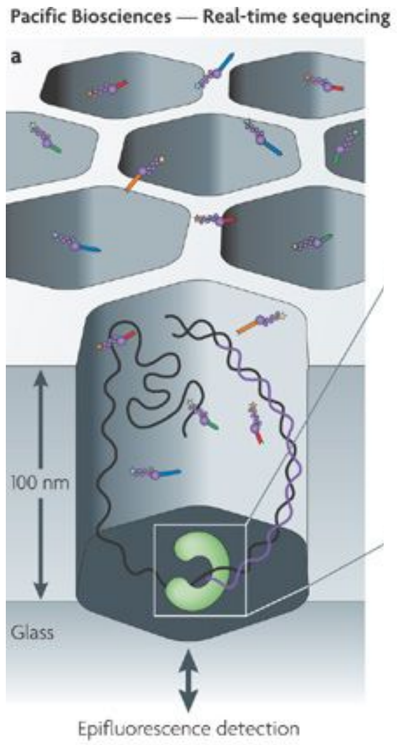
Can help in highly repetitive regions!



However, long reads have higher error rated than short reads and still can’t cover large repetitive regions

Single molecule real-time sequencing

ZMW



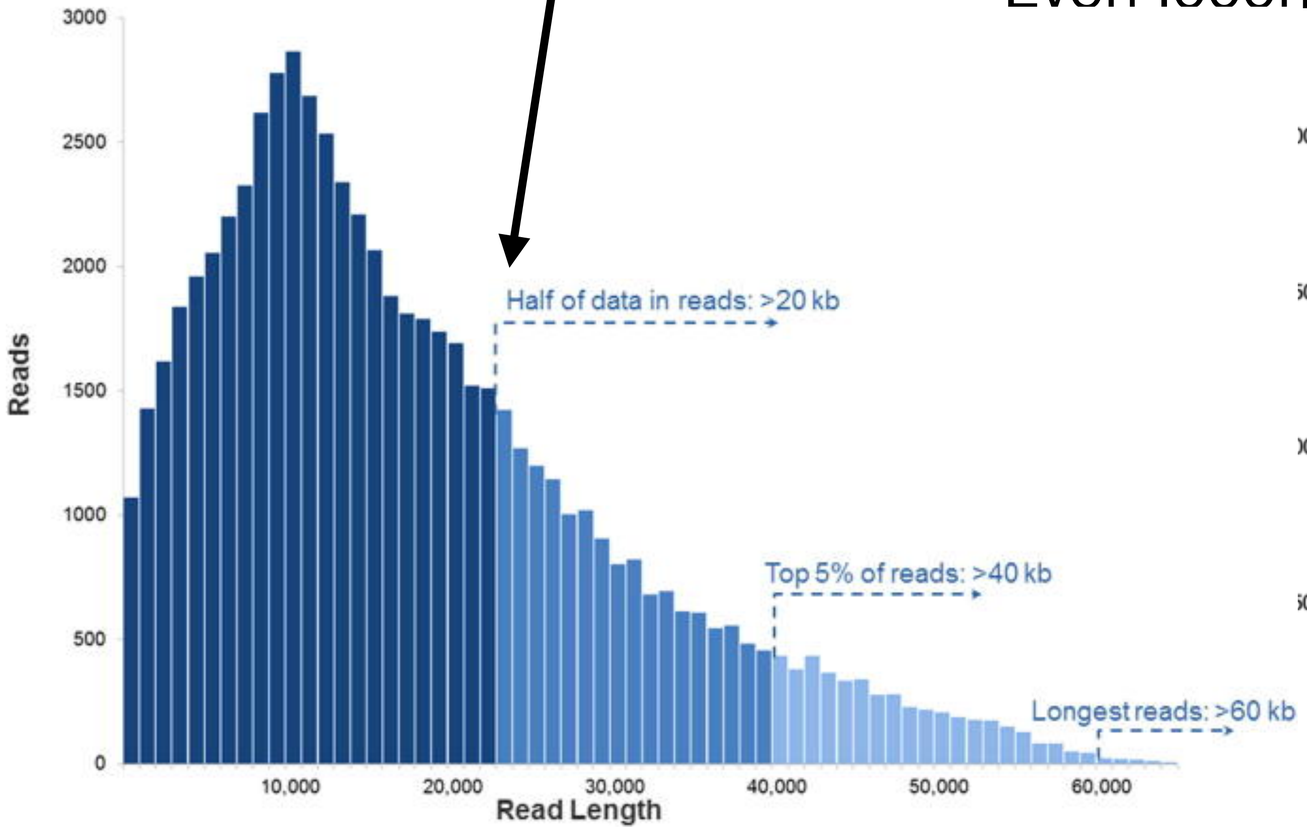
https://www.youtube.com/watch?v=v8p4ph2MAvI

Sequencing technologies - the next generation. Metzker ML, Nat Rev Genet. 2010 Jan;11(1):31-46
PacBio Sequencing and Its Applications. Rhoads A, Kin Fai A, Genomics Proteomics Bioinformatics. 2015 Oct; 13(5): 278–289.

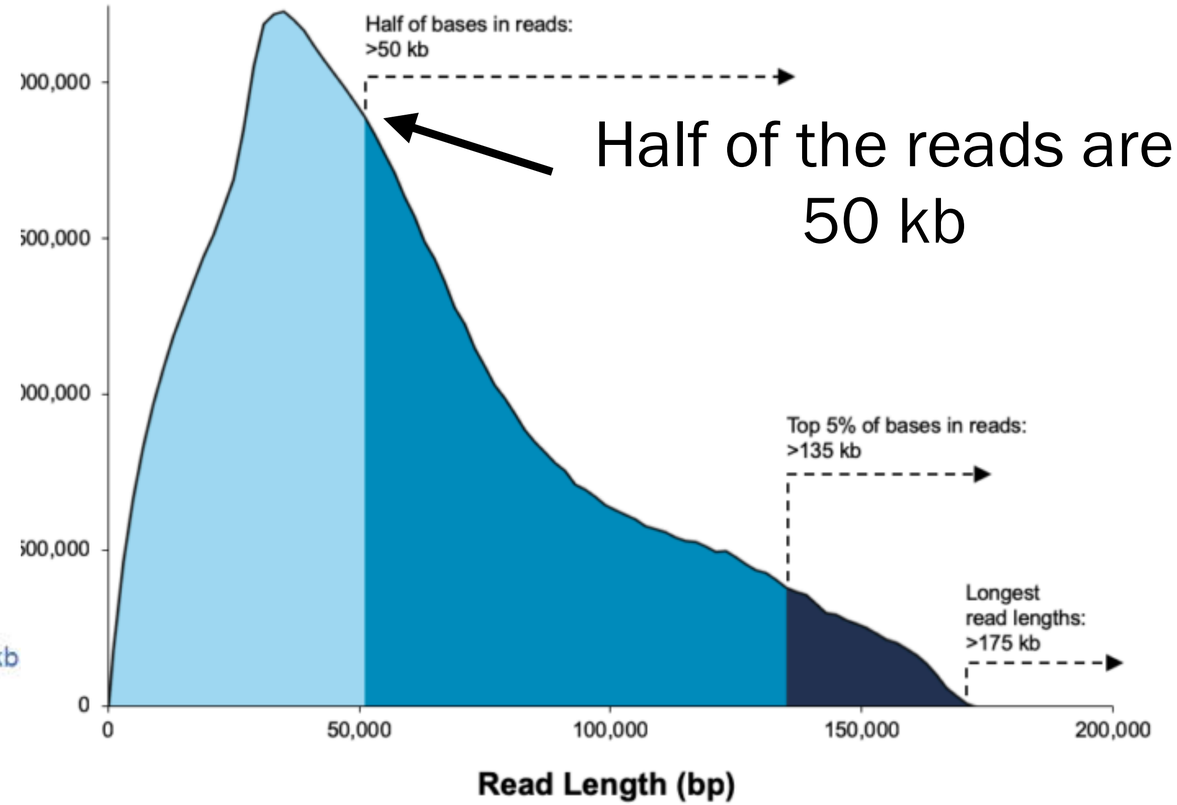
Single molecule real-time sequencing

Long reads

Half of the reads are 20 kb



Even looonger!



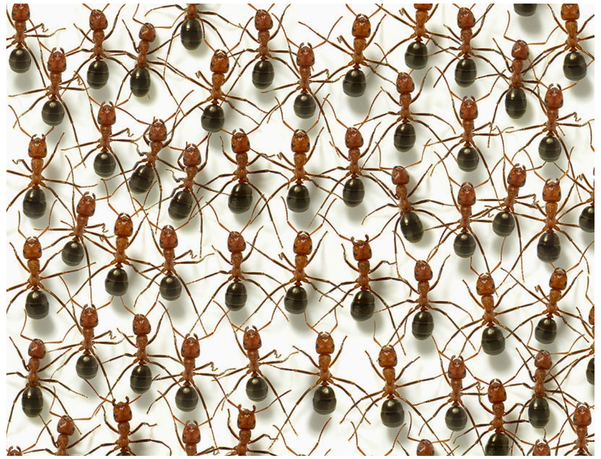
Half of the reads are 50 kb

PacBio Sequencing and Its Applications. Rhoads A, Kin Fai A, Genomics Proteomics Bioinformatics. 2015 Oct; 13(5): 278–289.

NGS Illumina vs SMRT PacBio sequencing throughput

Which one is better??

Illumina throughput: 150bp x 2 x (1300- 1500 million) = 400 Gb (~130x human genome)

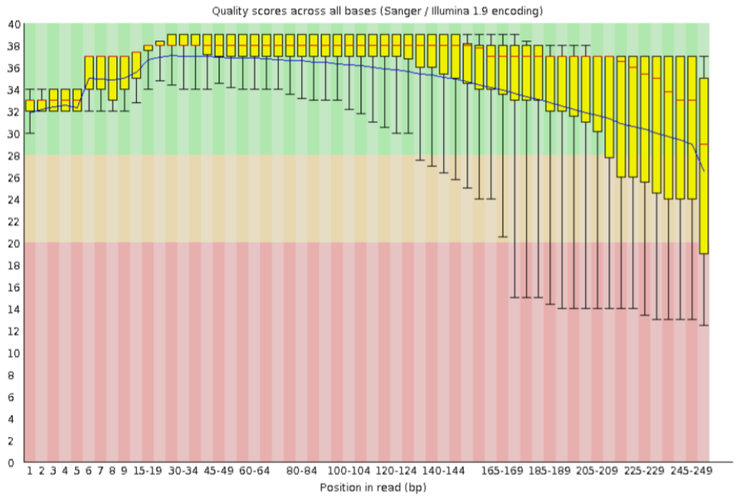


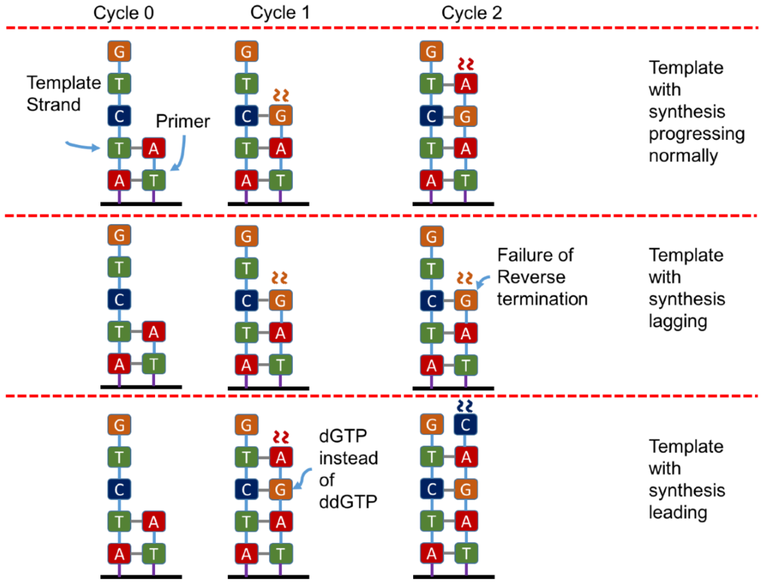
SMRT sequencing PacBio: Half reads up to 50kb = 160 Gb per cell (~53x human genome)

Error rates?

Illumina

Sequencing by synthesis





Phasing - sequences in cluster loose synchronization due to failures in synthesis -> gets worse with read length

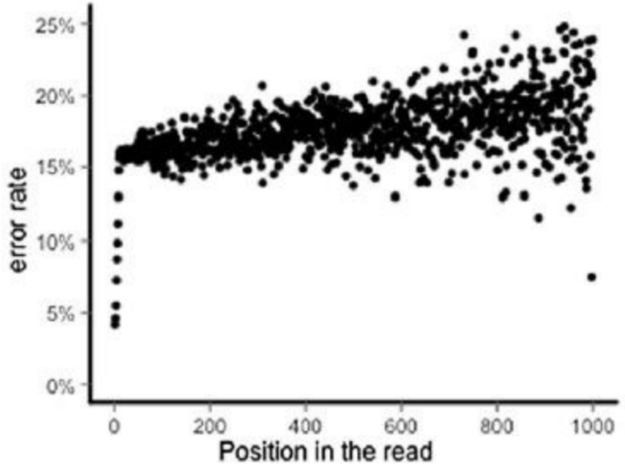
Error rate: ~ 0.1 %

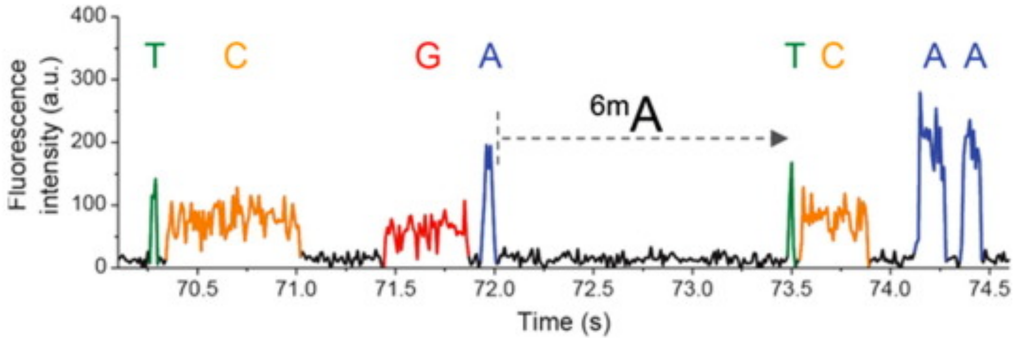
Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Nucleic Acids Research, 2018, (46), 5, 2159–2168

Error rates?

Single molecule real-time sequencing

PacBio





Error rate: ~13 - 15 %

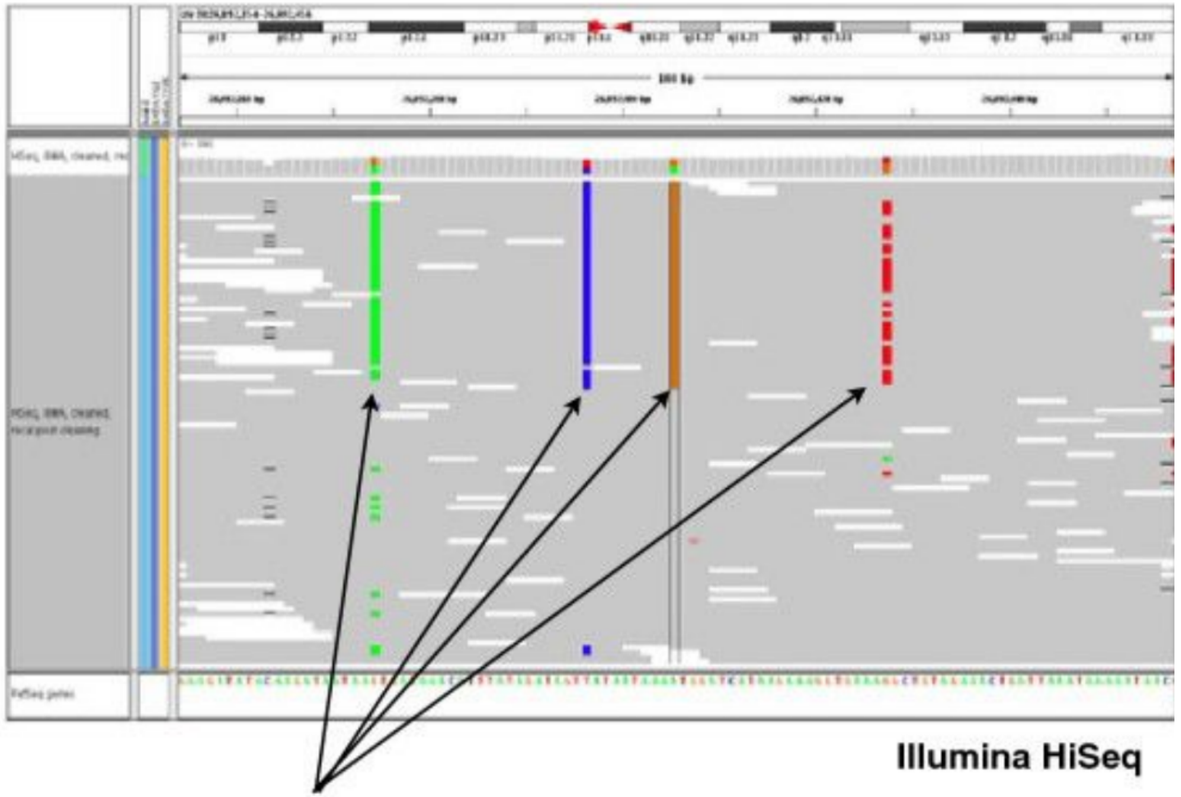
Pacific biosciences sequencing technology for genotyping and variation discovery in human data. Carneiro OM, Russ C, Ross MG, Gabriel SB, Nusbaum C, and DePristo MA. BMC Genomics, 2012, 13:375

PacBio Sequencing and Its Applications. Rhoads A, Kin Fai A, Genomics Proteomics Bioinformatics. 2015 Oct; 13(5): 278–289.

Where are the errors?

Illumina

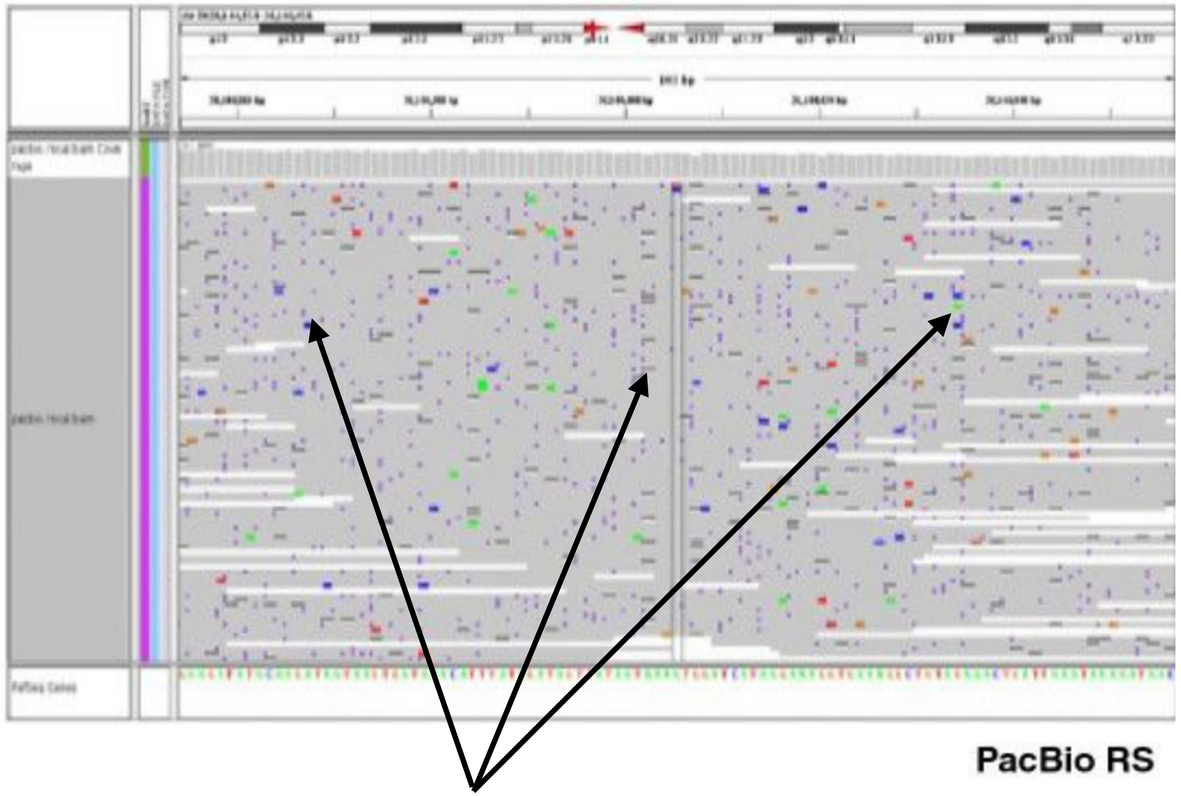
Bias towards low GC regions



SYSTEMATIC ERROR

PacBio

Insertion/deletions



Errors are random

Pacific biosciences sequencing technology for genotyping and variation discovery in human data. Carneiro OM, Russ C, Ross MG, Gabriel SB, Nusbaum C, and DePristo MA. BMC Genomics, 2012, 13:375

Which one is better?

It depends on the application

Illumina:

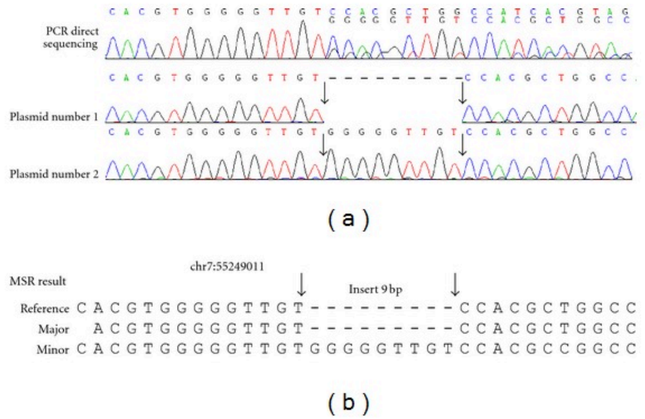
- Good for sequencing small genomes
- SNP analysis
- Correcting long reads

ACTGGTAT...GATCACCAG
ACTCGTAT...GATCACCAG
ACTCGTAT...GATCTCCAG

	SNP I	SNP II
Allele I	G	A
Allele II	C	T

PacBio:

- Good for sequencing larger genomes
- Structural variation analysis (indels)



Sequencing Data - FASTA

FASTA format
or
FAST-All format

PacBio output

Indicates the start of a new entry

Sequence name or header

Sequence

```
>Scaffold_1
CAAAGATTTGATAACCAACATCGCAAAGCTTGACAAGTTTGAGGGCGTGGAGTTCCGTCGGTGGCAGAAGAAGGTGCATTTTCTGTTGACATCGTTAGGG
GTTGTGTACATTTTGACCACNNNGATGCCGGAGGAGTACGAGGAGGGTGAGCAGCTCGAAACTACGAGGAGACGCCTCAAATTTGAGAACGACAAATTAT
ATTTGTCGTGGACACATTCTCAATGGTATGTCCGATGCTTTATTTGATATTTATCAATATTATGACTCCGCATAATCTCTTTGACATGATCTTGAATCTA
AGTATATGGTAGAAGATGTTTCTAGTAAGAAATTCTTGGTTAATAATTTTGATACATATCGTAAGGTTGATTCAAGACCTATCATGGAACAATTTCATGA
ACTTGAAAGGATGTTAGGAAAATTAAGGCAACATAATATTAATATGAATGAATTGTTTGTAGTCTCCTCTATTATTGACAAGTTACCACCTAGTTGGAAA
GATGTAAGGAACCAATTGAAACATAAAAAGGATGATATGAATGTTAAACAACTTGGAGCCCATCTTCGGATTGAAGAAGGGATCCGAGTGCAAGATGGAC
AAAAGGTTGAAAAACCTATGACTTCCACCGTGAATATTGTGGAAGGGAAGTTCAAGGGTGGTAATAACTCGAAAAAGCGTAAGTTCAATGGCAAAAAACA
AAGATTTCGGTAAGAAATCAAAGGGTGACAATGATTCCCAGGAGACATCCACTCGTGGATGTTGGGAATGCGGGAAACCGGGTCACTTCAAGAAGGATTA
CTCTGTGTGGAAGCGAAGGATGAGTAAGAAATCTCATGGAAGCGGNNATGACGGTGGGTCATCCCAAGACAAGCAAGGTATGACTTGTACACAAGTTTCA
CTTAATCGATTTTATTATTATGTGTGTATGATTTCTGAAATCAATTGTGTGCAGGATGATTGCGGGTGGTATATCGATTCAGGCACATCAAAGCATGTTT
GCAAAGATGTAATCTATTCAAAACCTTCGAGGAGGTTGGAGATGGAGAGCGATTACACATGGGGAACGATTCTACAACTAGTATTCGTGGCAAAGGCCAA
GTAAAATTGTTGTTCAGTTCTGGGAAAACAATTATCCTTCGGGATGTATTGTATTCTCCTGAACTTAGGAAAAATCTTAAATCTGGTGCGTTATTGAATC
GATTAGGGTTTAAACTTGTTTTTGAAGCAGACAAATTTGTATTATCTAAAGGTGGTGTGTTTGCAGGACTTGGTTACTTATGTAATGGCTTGTTTAGACT
AAGTGTAGATAATTCTACACAAGACAATGTGTTTGTTGTTGATATTGATTCTAGCTTATGGCATTGTAGGCTAGGTCATGTTAATTATAAAAAGTTGTGT
```

Sequencing Data - FASTQ

FASTQ format

Illumina output

Instrument ID

Flow cell ID

Indicates a new read

Run #

Coordinates

Read directionality

Sample number

Basically the read name

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

The sequence

Quality score in ASCII

Prior quality score

Base Quality

Sequence

Relationship Between Sequencing Quality Score and Base Call Accuracy

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

$Q = -10\log10(e)$

We have our sequencing data
Now what?

Quality Control

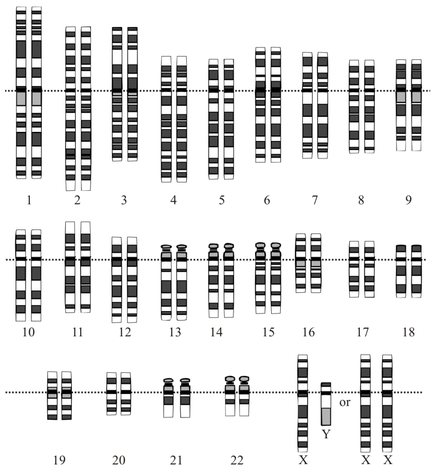
How to test if assumptions hold?

Assumptions

- DNA of the target species
- good quality reads
- even coverage (same number of reads from all genome regions)
- no technical artifacts
- no contaminating DNA from other species

→ Assembler →

Perfect Reconstruction*



(*minus repetitive regions, sequencing errors)

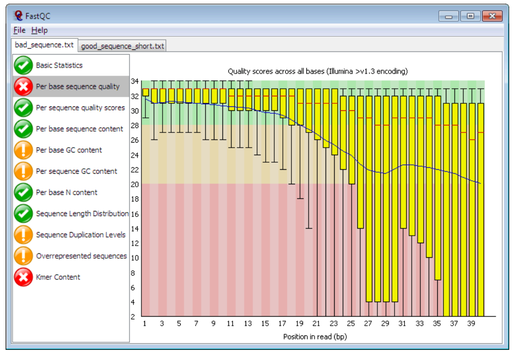
Quality Control

- Read quality:
  - **FastQC** (qcfail.com knowledge base)

- Remove sequencing adapters:
  - **trimmomatic, cutadapt, skewer**

- Duplicates:
  - **sga filter, BBnorm**

- Contamination and Bias detection:
  - **KAT, preqc**



KMER
ANALYSIS
TOOLKIT

FASTQC

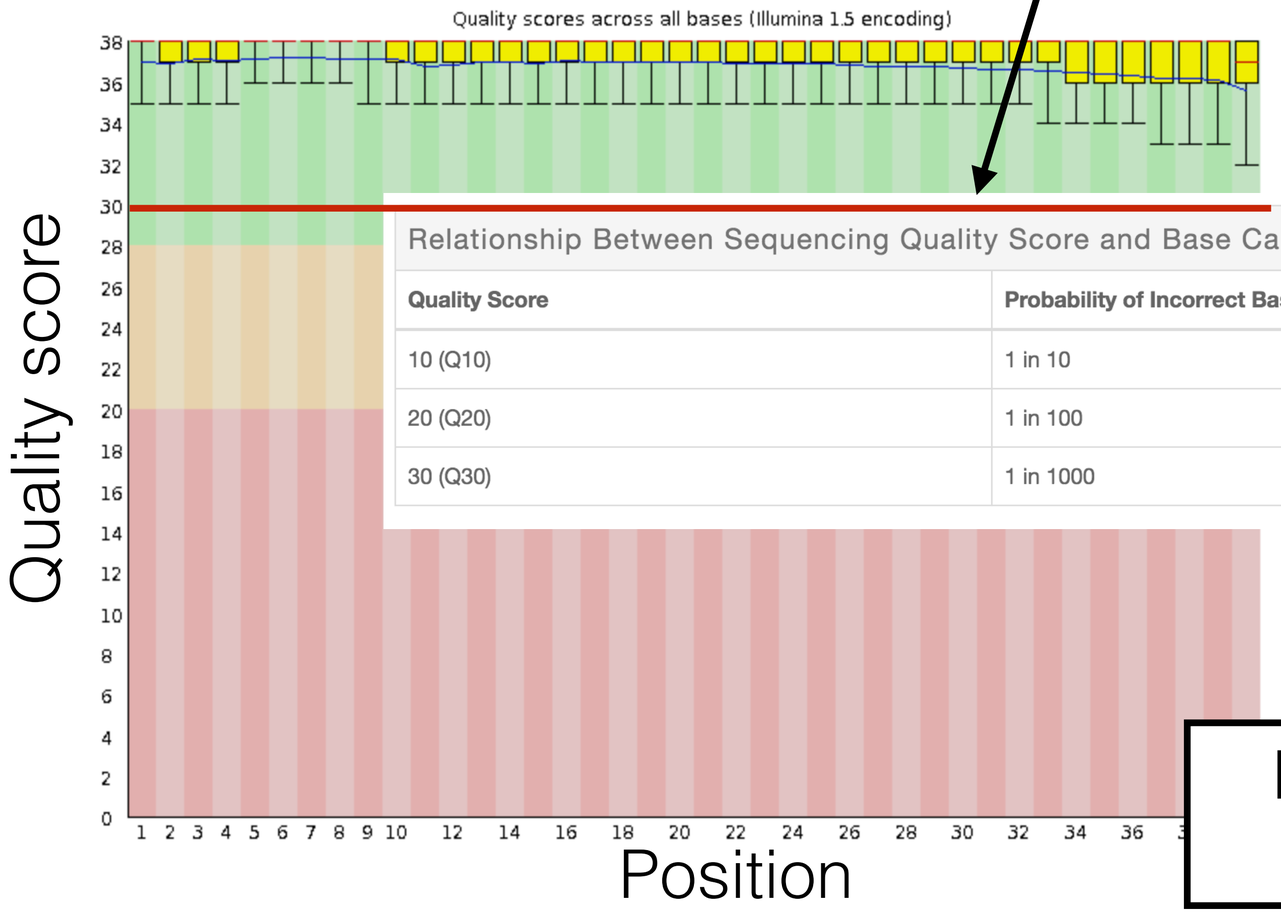
Quality Control

More than 99.9 % accuracy

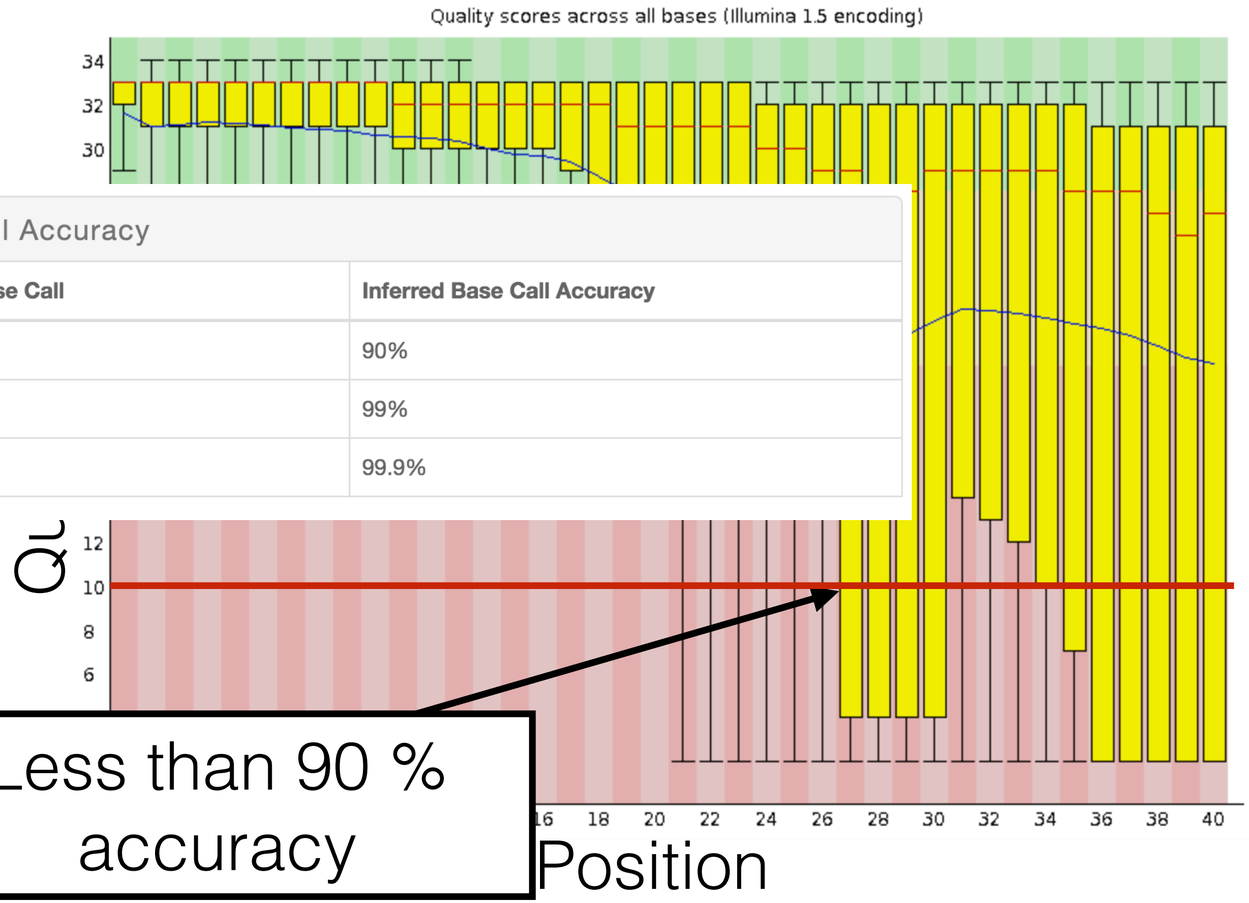
Good

Bad

Per base sequence quality



Per base sequence quality



Relationship Between Sequencing Quality Score and Base Call Accuracy

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Less than 90 % accuracy

Quality Control

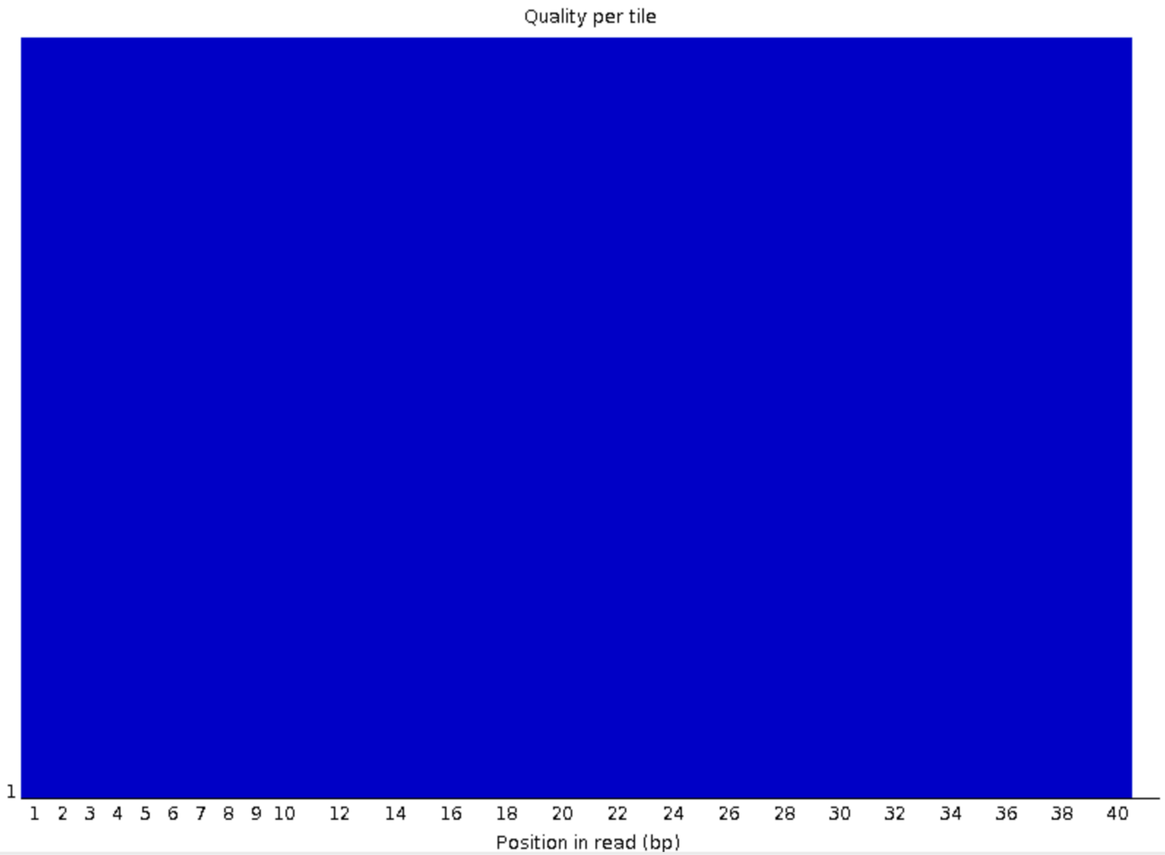
FASTQC

Good

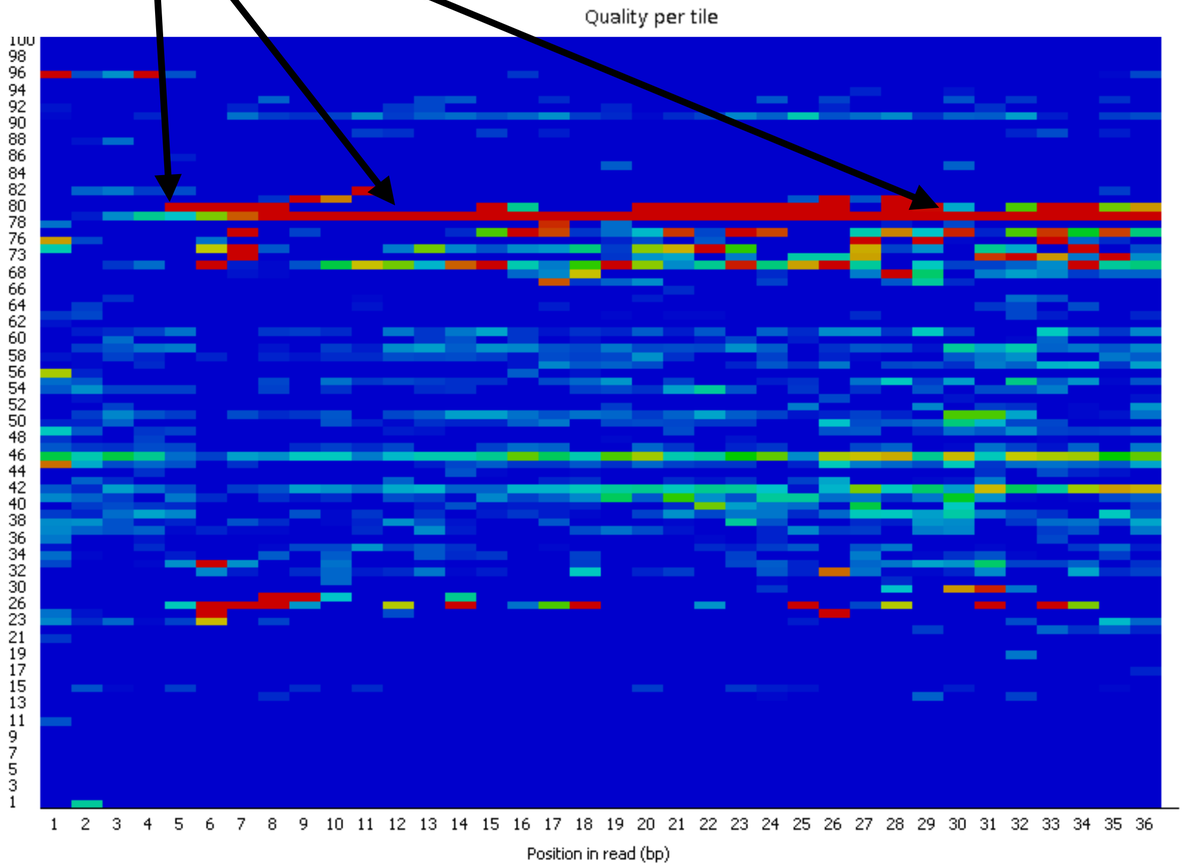
Bad

Something wrong with this tile

Per tile sequence quality



Per tile sequence quality



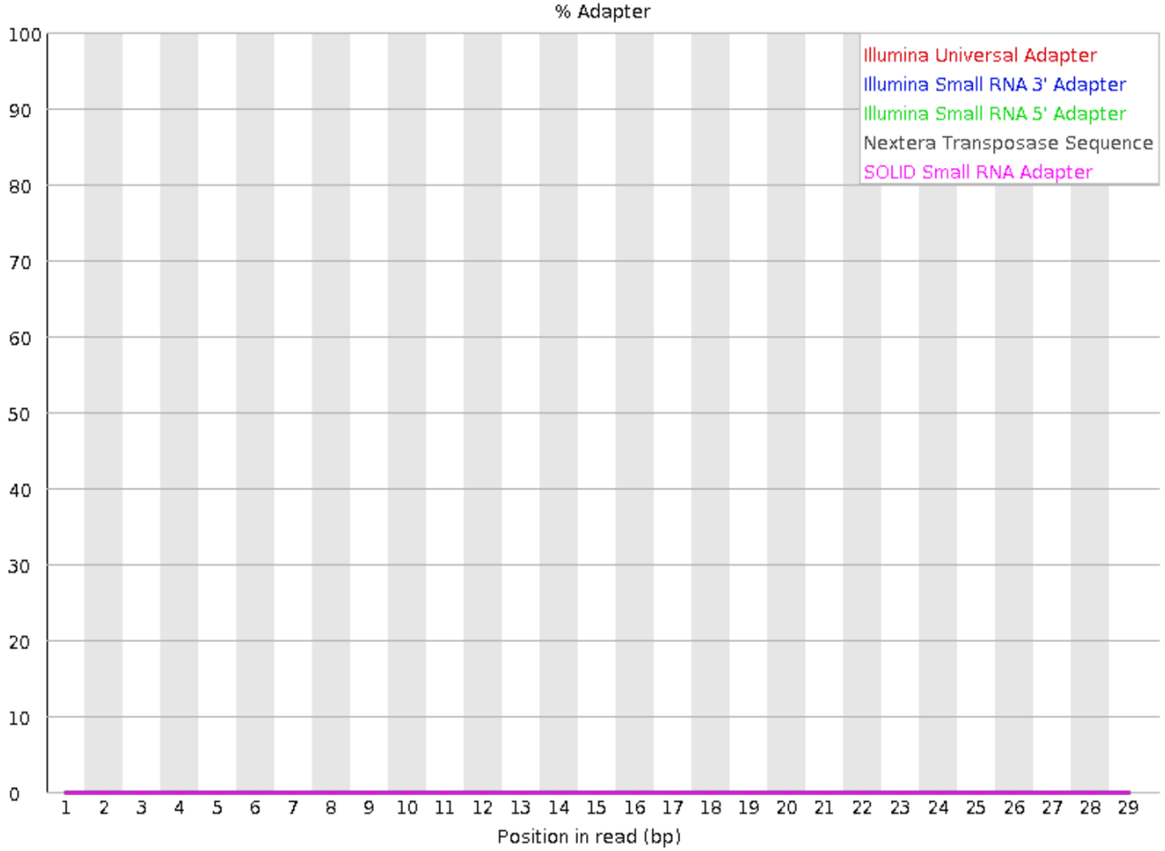
Quality Control

FASTQC

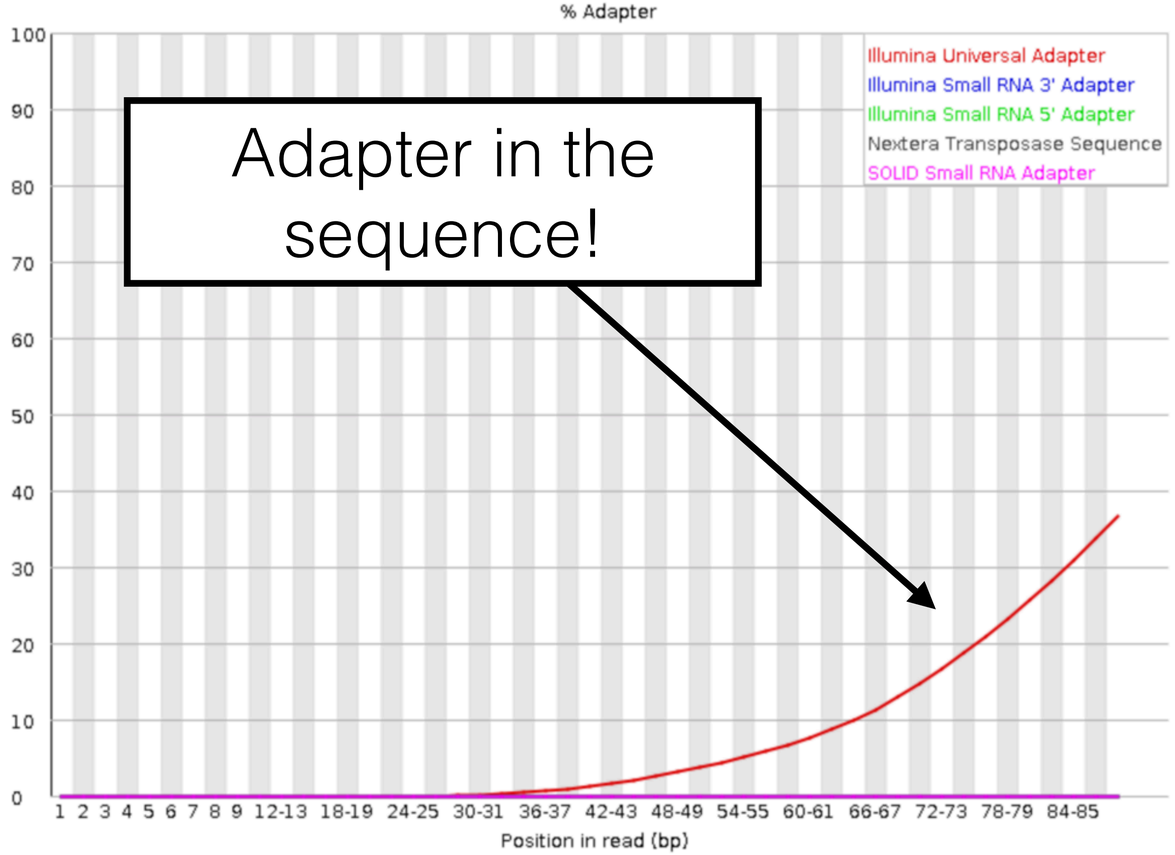
Good

Bad

Adapter Content



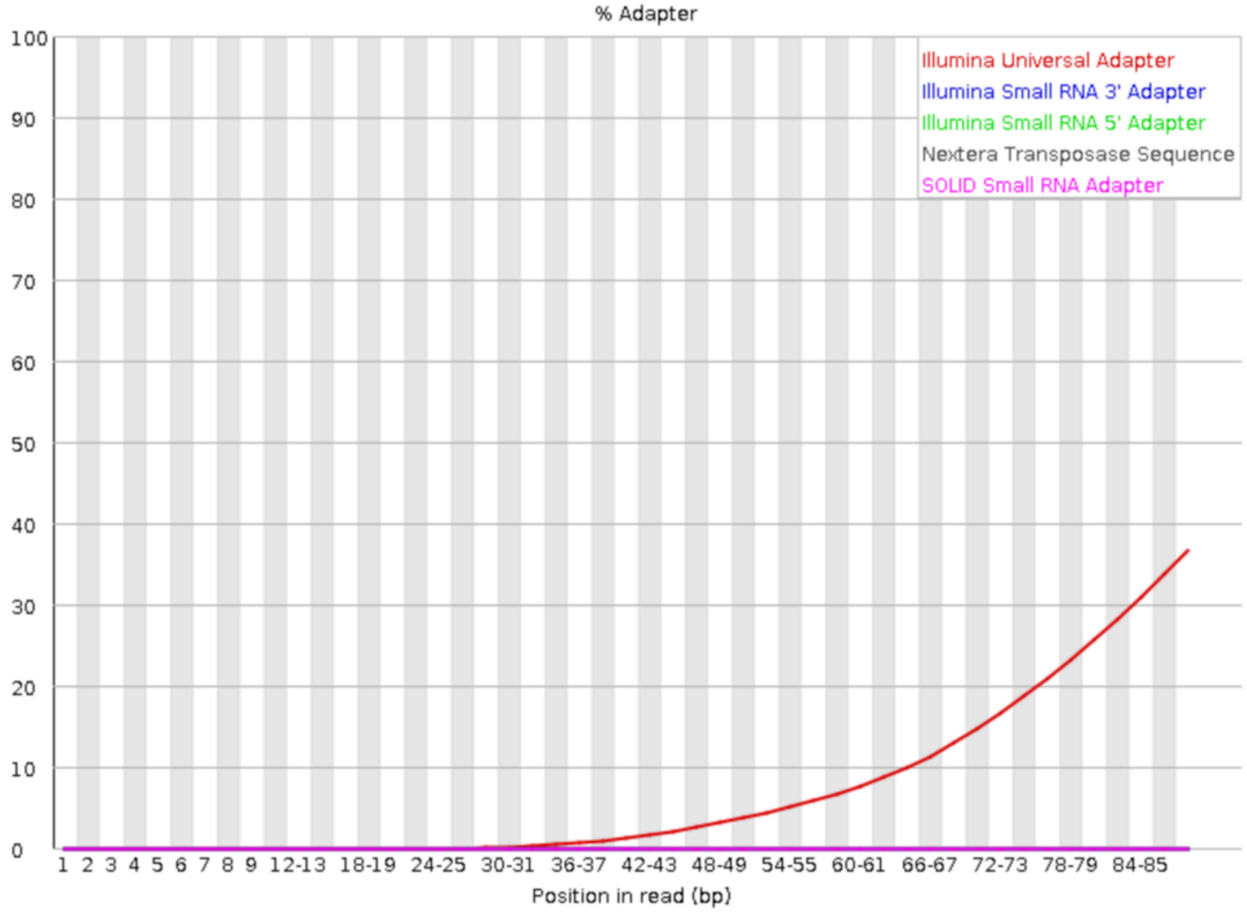
Adapter Content



Adapters trimming

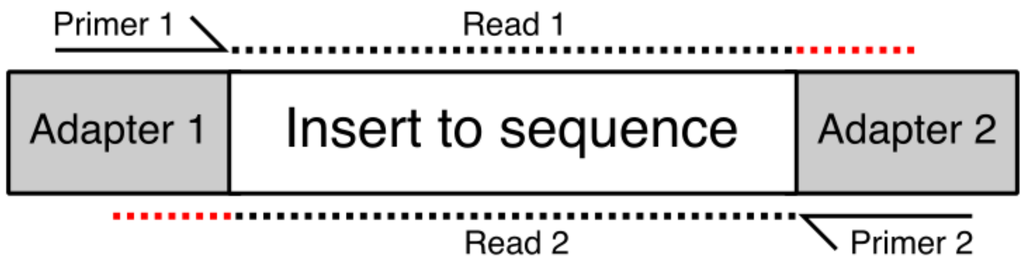
Adapter Content

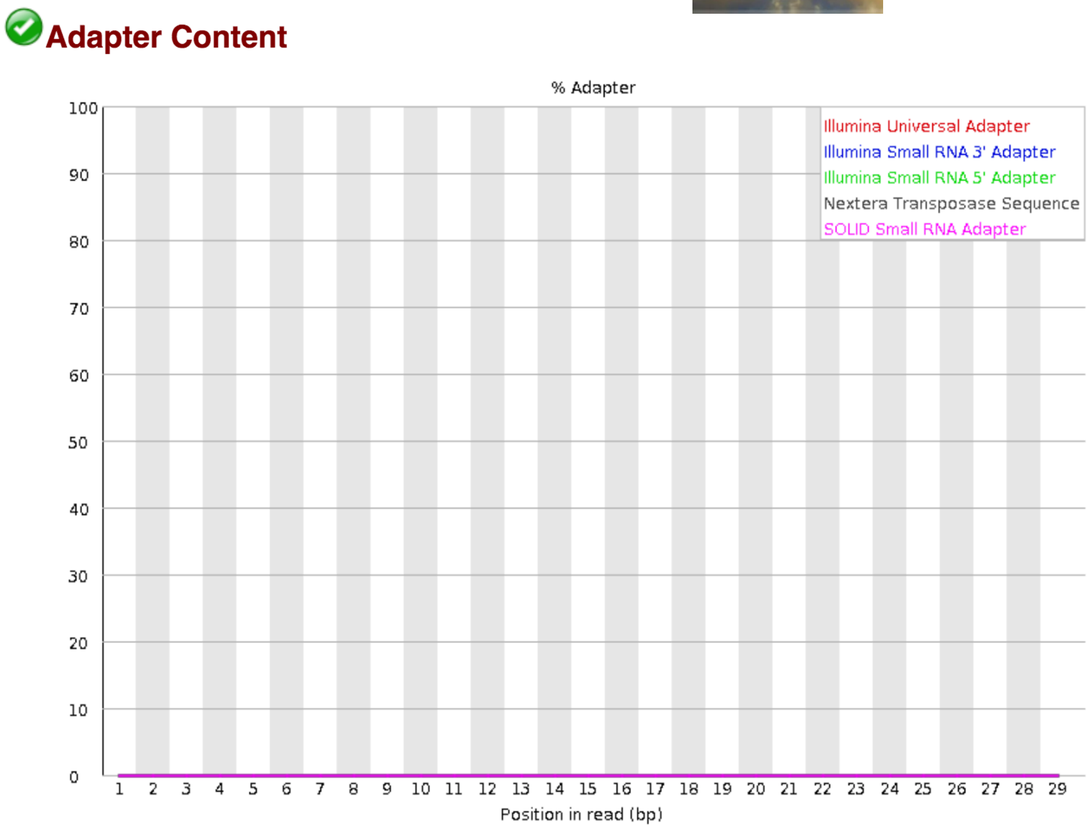
Bad



How do we get rid of adapters?

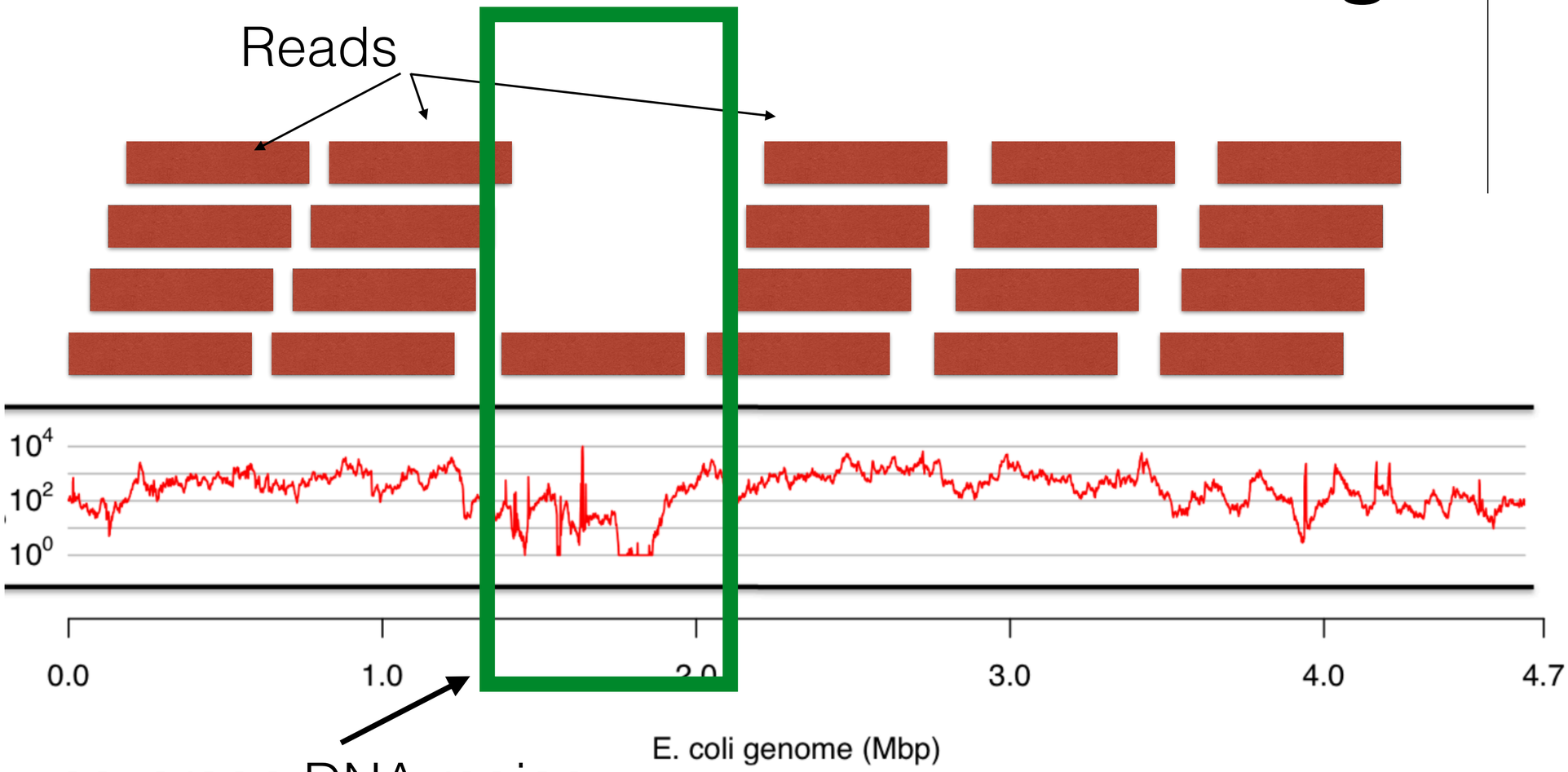
Adapters trimming





- Sequencer / Base caller knows which adapter is used
- Why isn't that done before making the fastq file?
- it is!
- so why do it again?
  —> e.g. Read-through

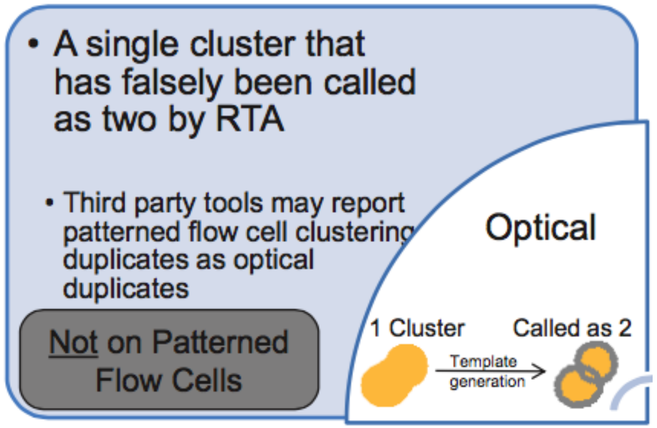
Even coverage?



Low coverage DNA region

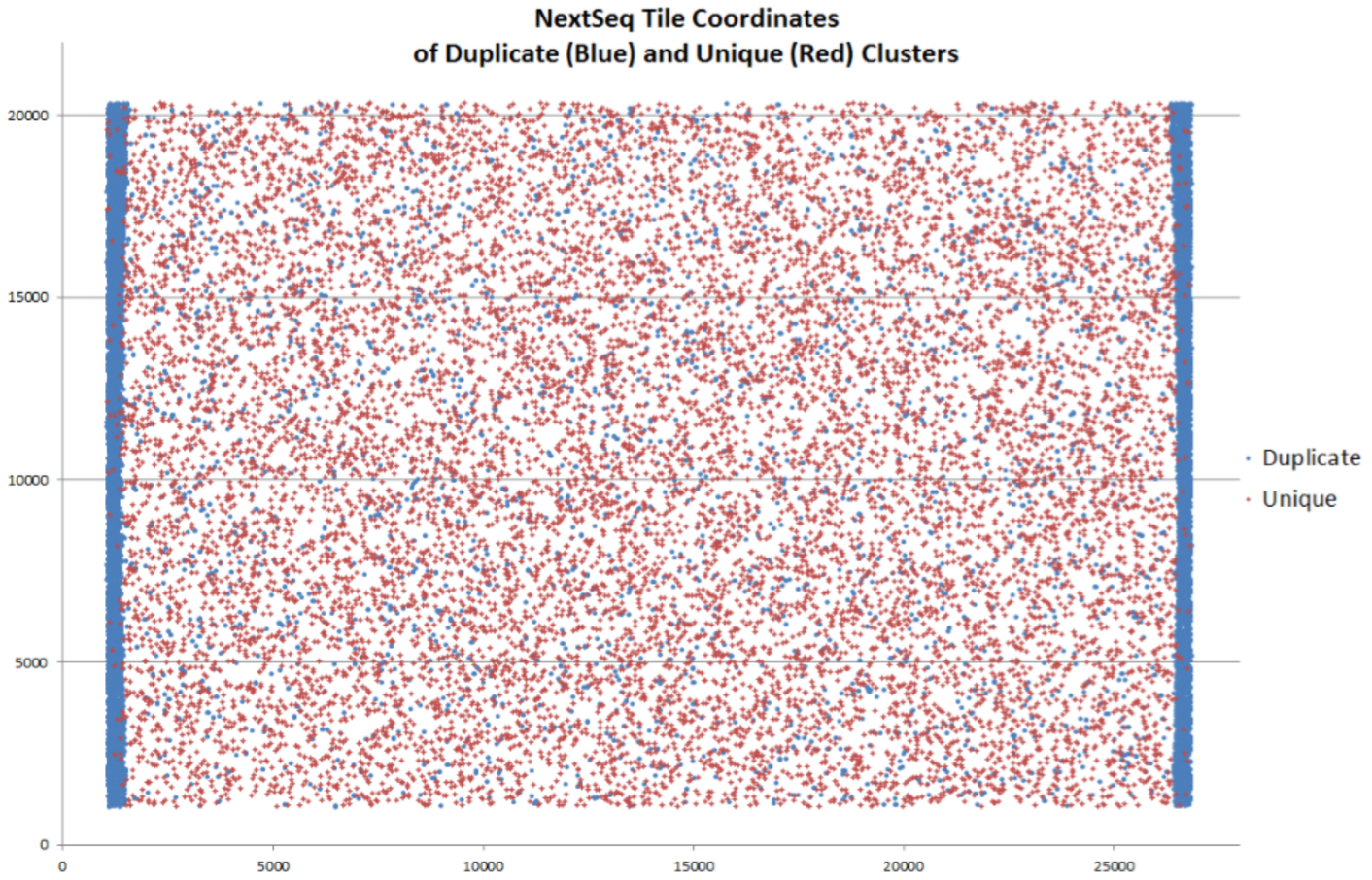
Why?

Coverage biases



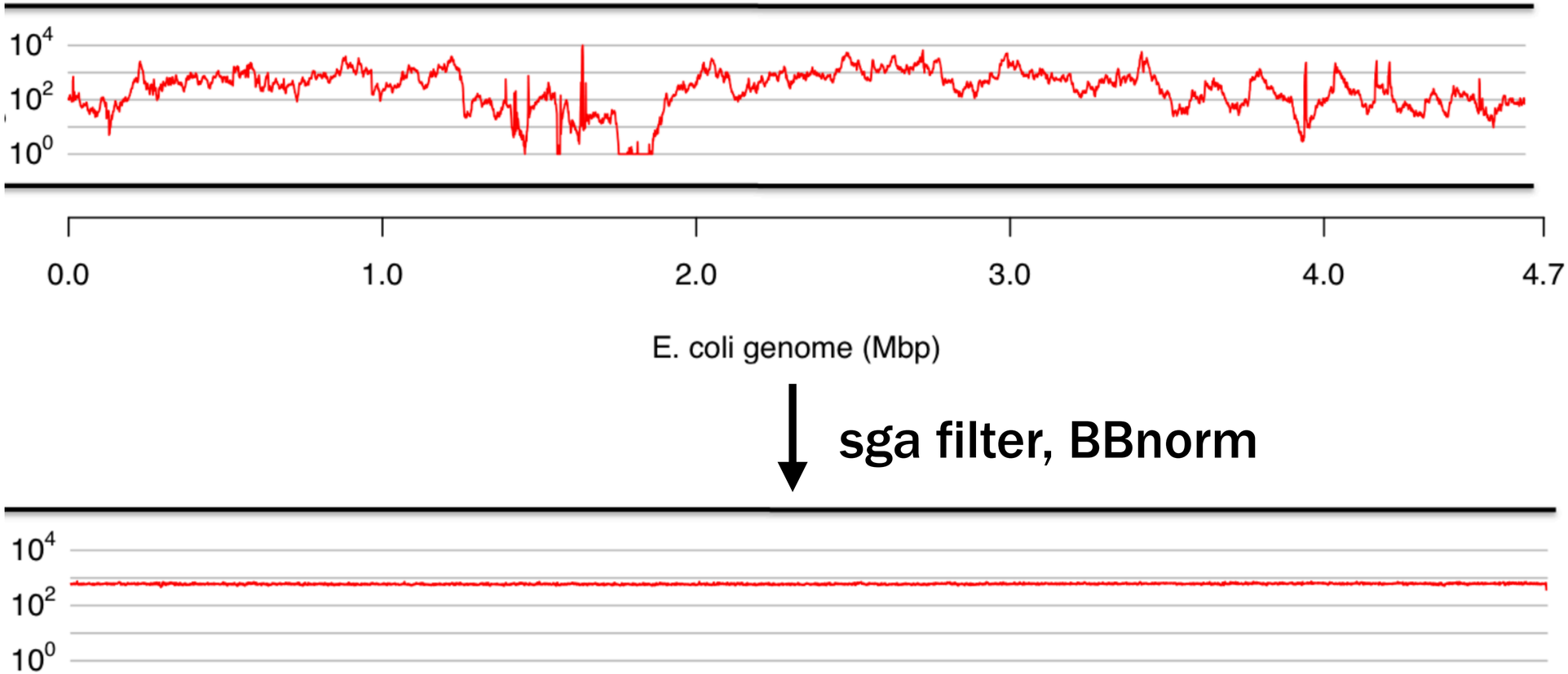
https://www.biostars.org/p/229842/#230390

Coverage biases



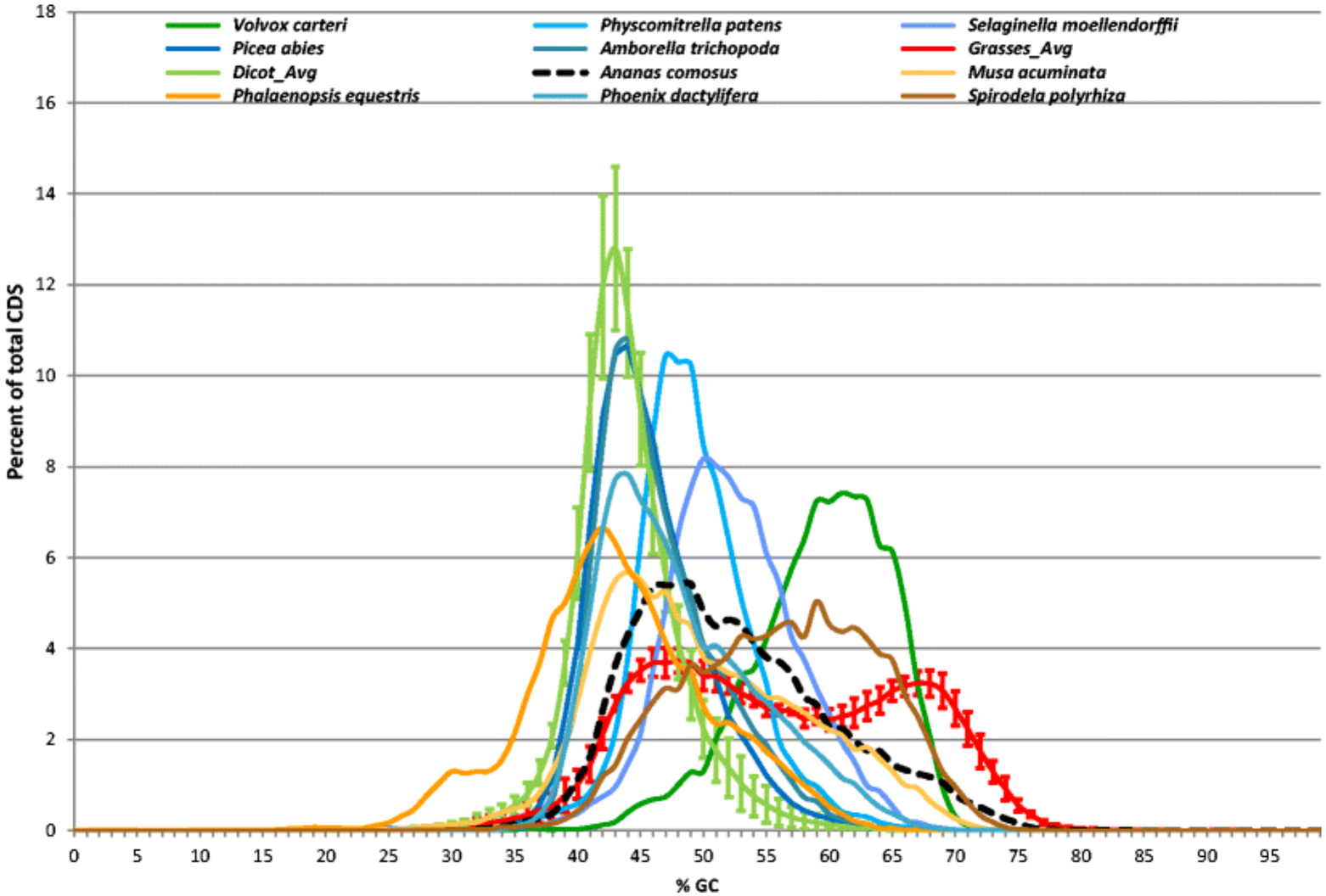
https://www.biostars.org/p/229842/#230390

Even coverage?



Detecting contamination

GC content = proportion of Guanine and Cytosine in a genome



Different species have different GC content

K-mers

Several assemblers use k-mers to generate contigs
Will be discussed in another lecture!

"*k*-mer" is a substring of length *k*

S: GGCGATTCATCG

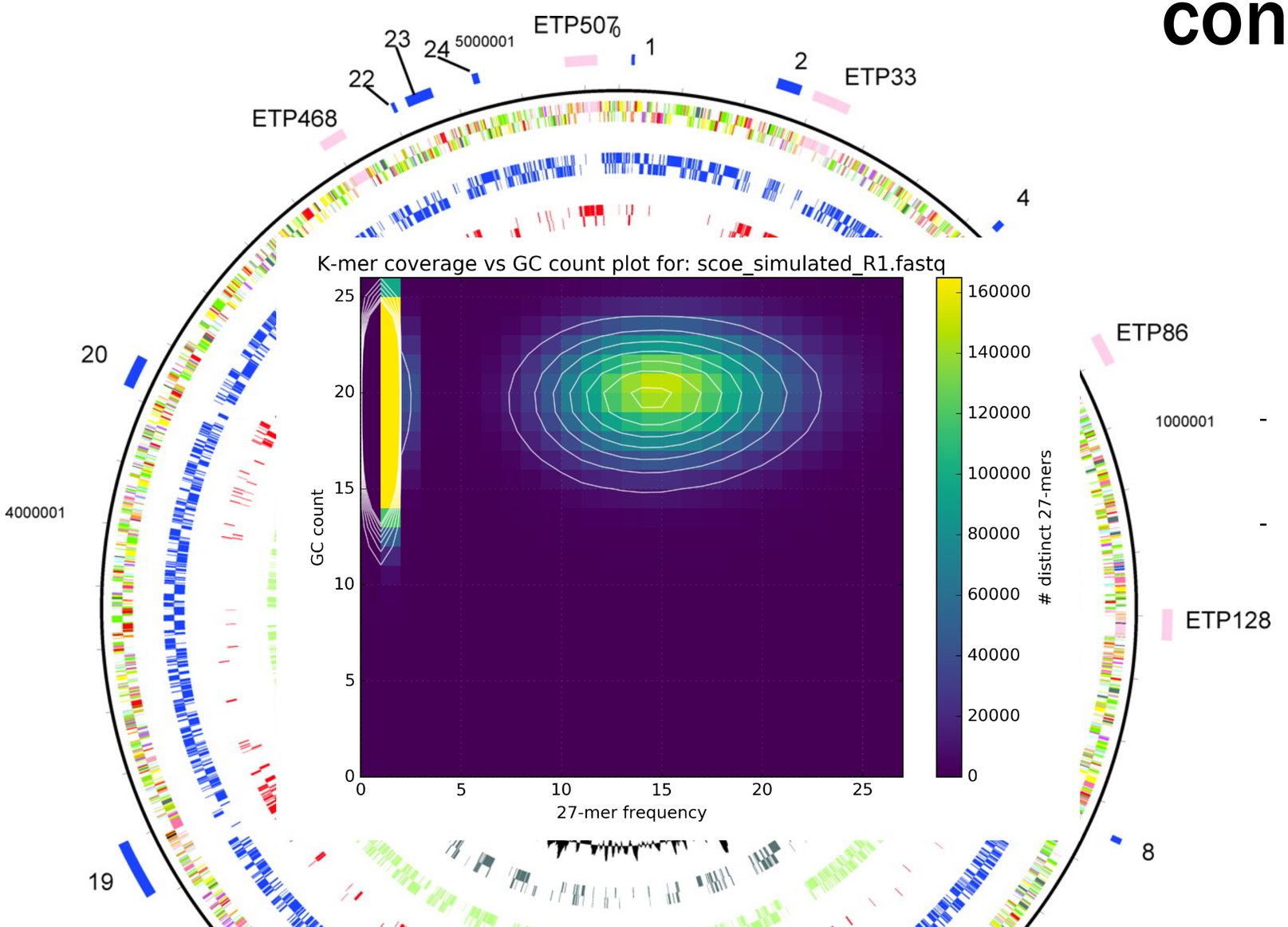
All 3-mers of *S*:

Detecting contamination

In theory, K-mers will be present only once in a genome

Just as PCR primers!

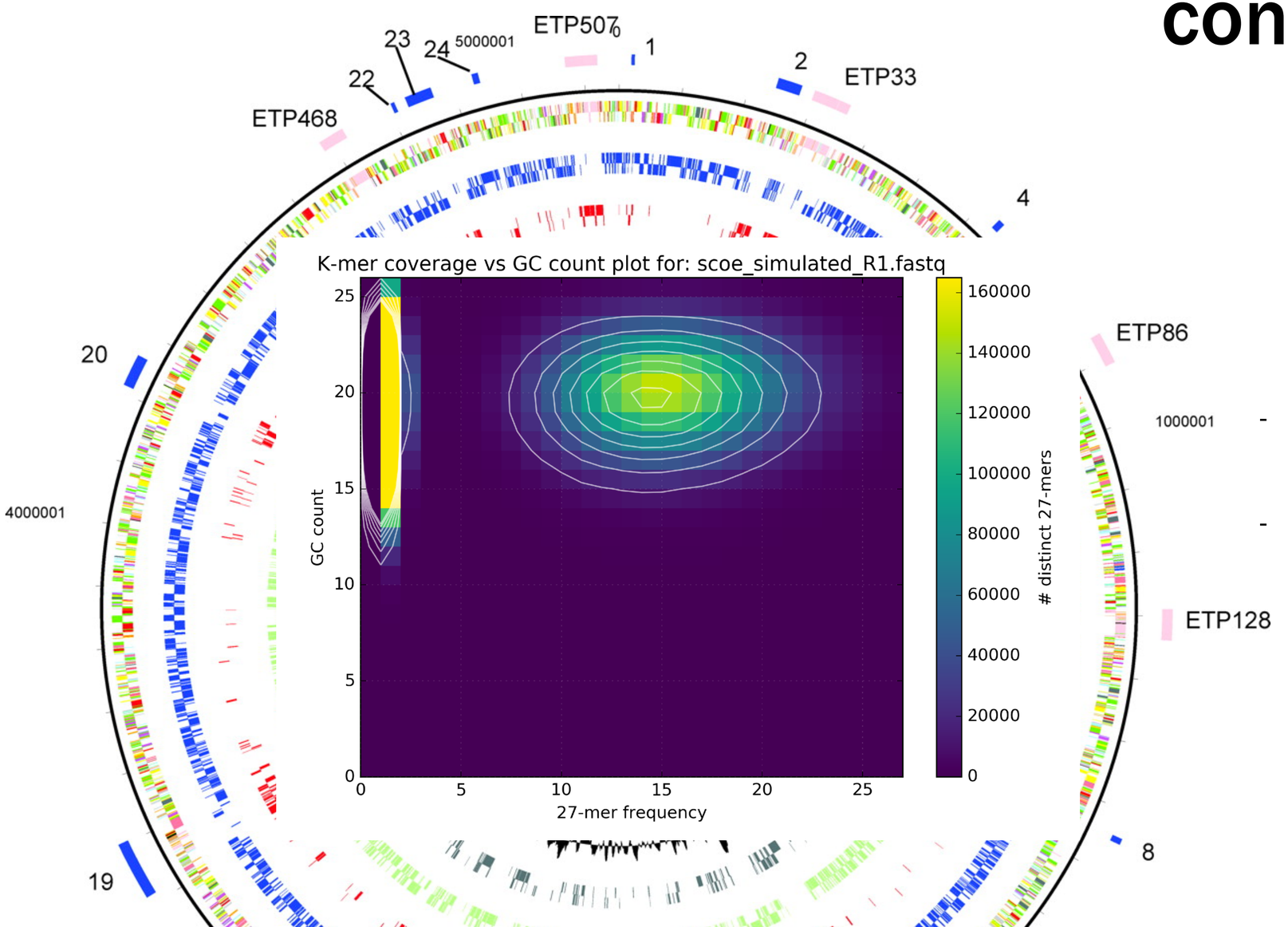
Detecting contamination



Ideal result:

- k-mer frequency centered around coverage
- GC centered around genome-wide average

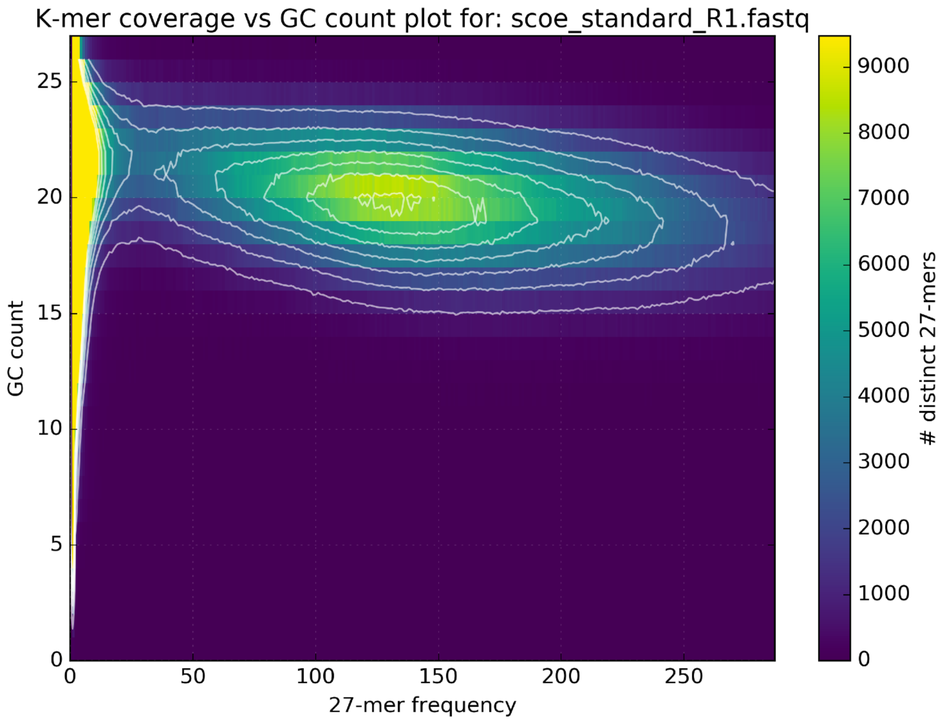
Detecting contamination



Ideal result:

- k-mer frequency centered around coverage
- GC centered around genome-wide average

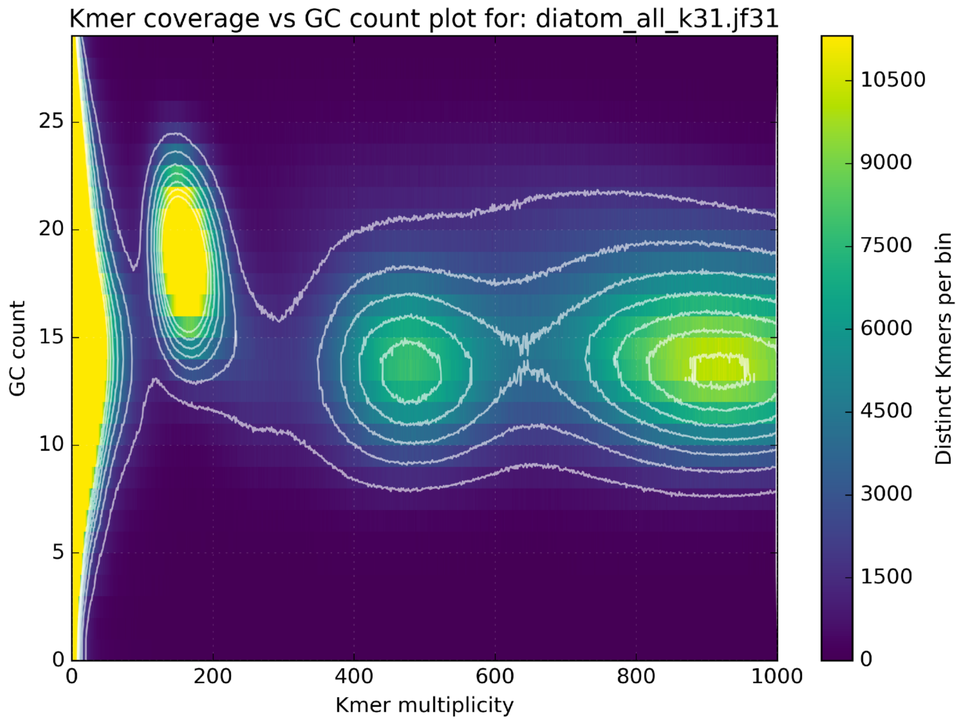
Detecting contamination



Realistic result:

- k-mer frequency distribution stretched out due to sequencing errors
- GC centered properly

Detecting contamination



Contamination:

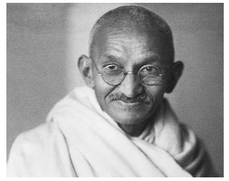
- clearly separated k-mer frequency peaks
- one peak with lower coverage and different average GC

Conclusion

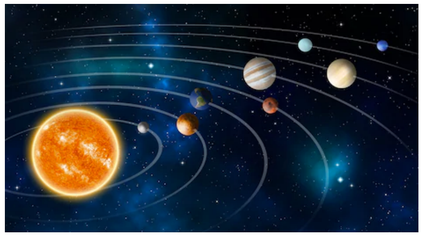
- A genome can be considered an instruction manual with the information required to build an organism

- Genomics can help us understand differences between species

- Genomics can also help us understand differences between Individuals from the same species



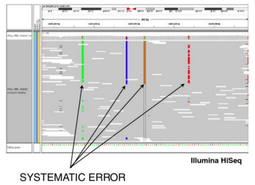
- You have enough DNA in your body to cross the solar system!



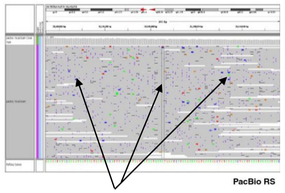
Conclusion

- Different sequencing technologies have their own strengths and weaknesses - it is important to know your aim of the study!

- Illumina sequencing provides very accurate reads, albeit in short reads and shows systematic errors

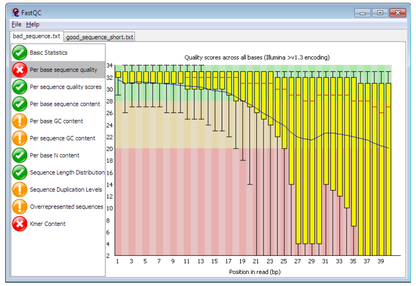


- SMRT sequencing of PacBio offers longer reads but at a lower accuracy



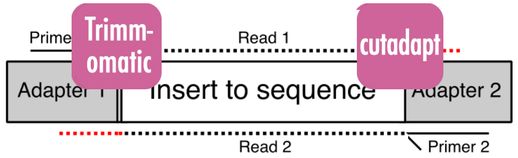
Conclusion

- Quality control is crucial for a good genome assembly

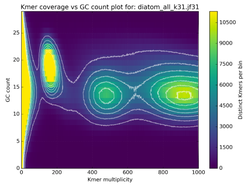


- FASTQC is a good tool to see the quality of your reads

- Adapter trimming can be done with tools such as trimmomatic and cutadapt



- DNA contamination can be detected with GC and K-mer content with KAT



Questions?