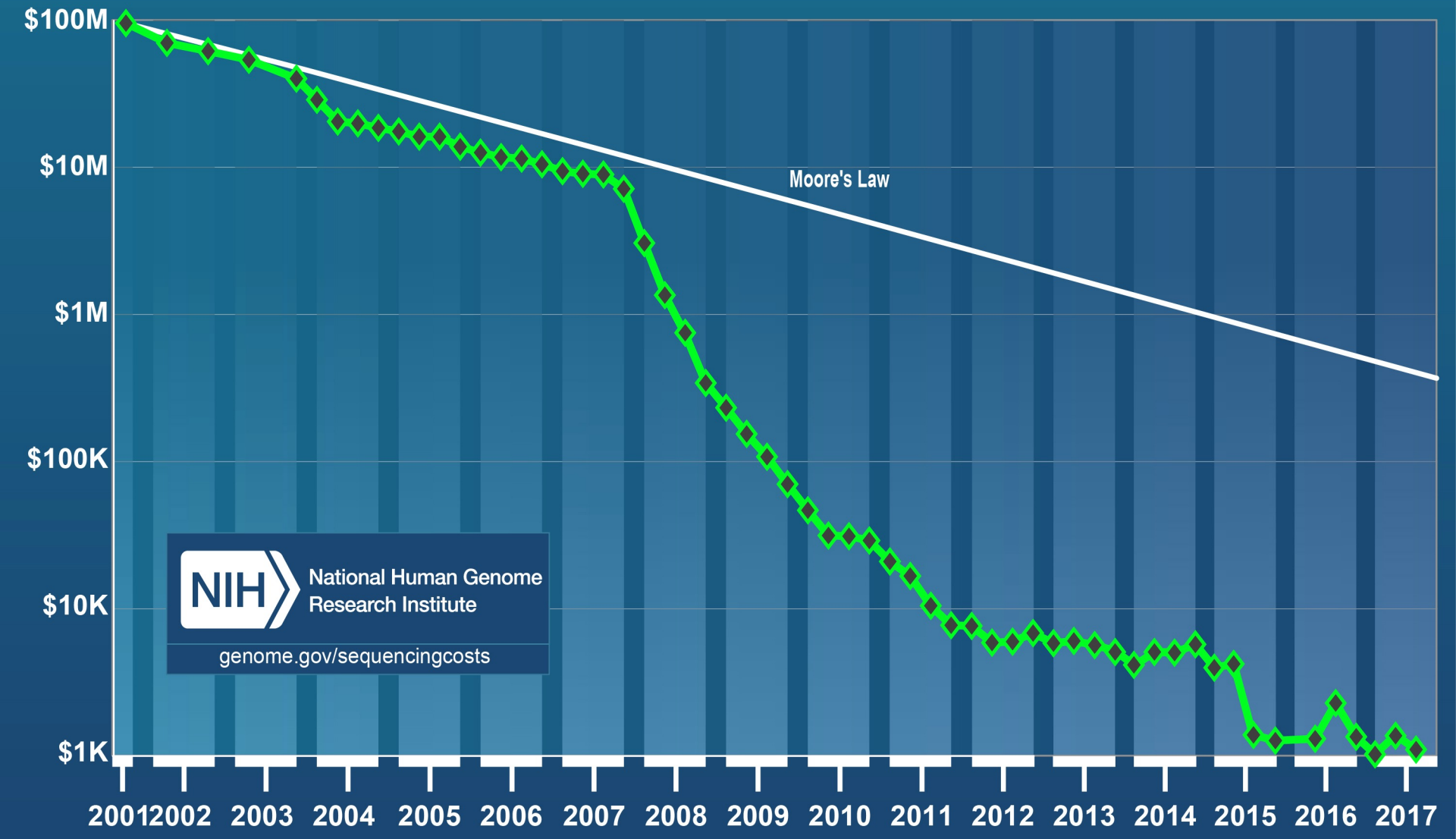
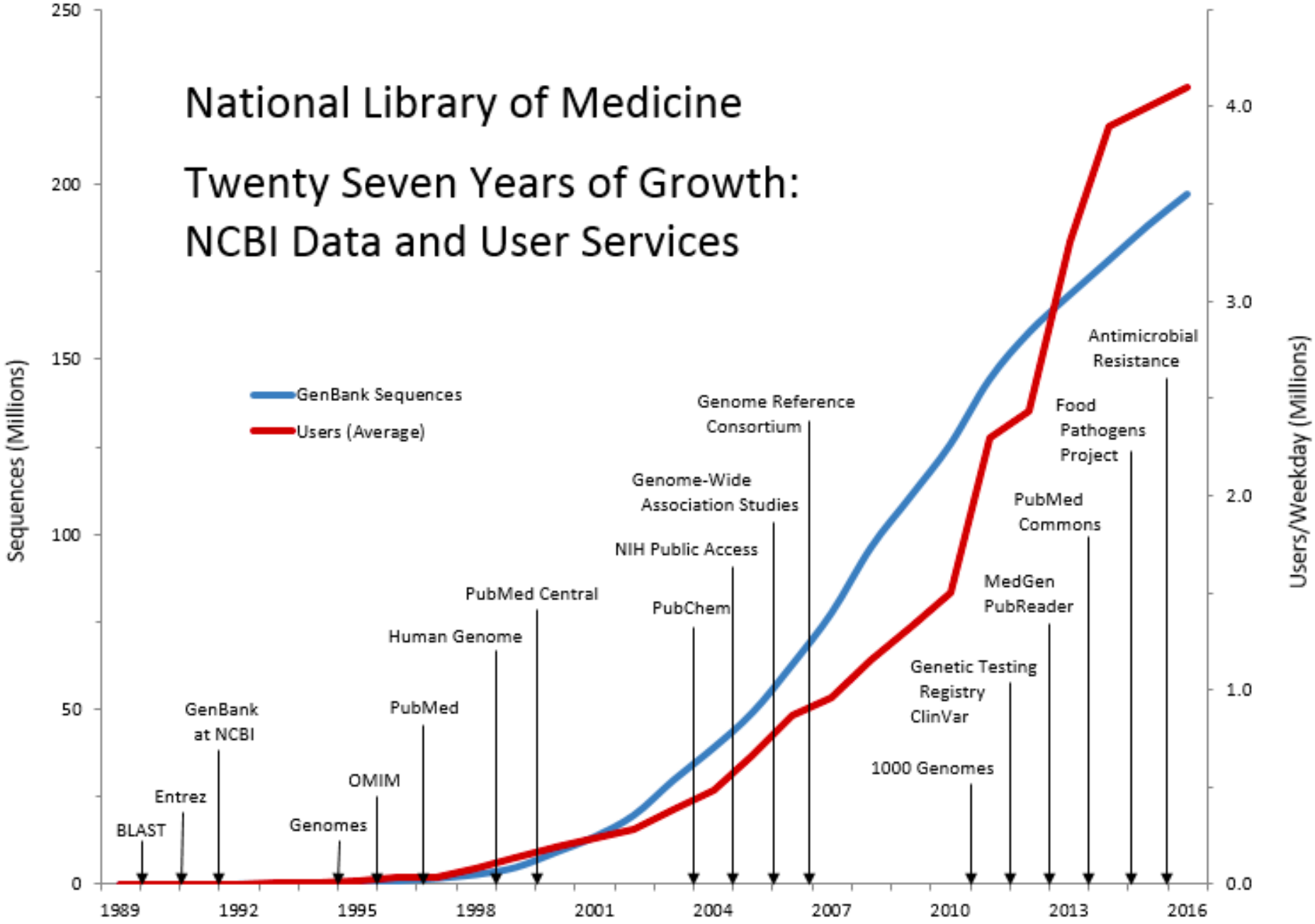
Genome Sequencing

Marc Tollis

Comparative Genomics

Cost per Genome





https://www.nlm.nih.gov/about/2018CJ.html

So You Want to Start a *de novo* Genome Assembly Project

Assuming you have a good reason to sequence and assemble a genome.

1. What is the size of the genome?

2. What will be your sequencing “recipe”?

1. Do you have the computational resources?
   - *i.e.* a machine with 32 processors, 512GB RAM

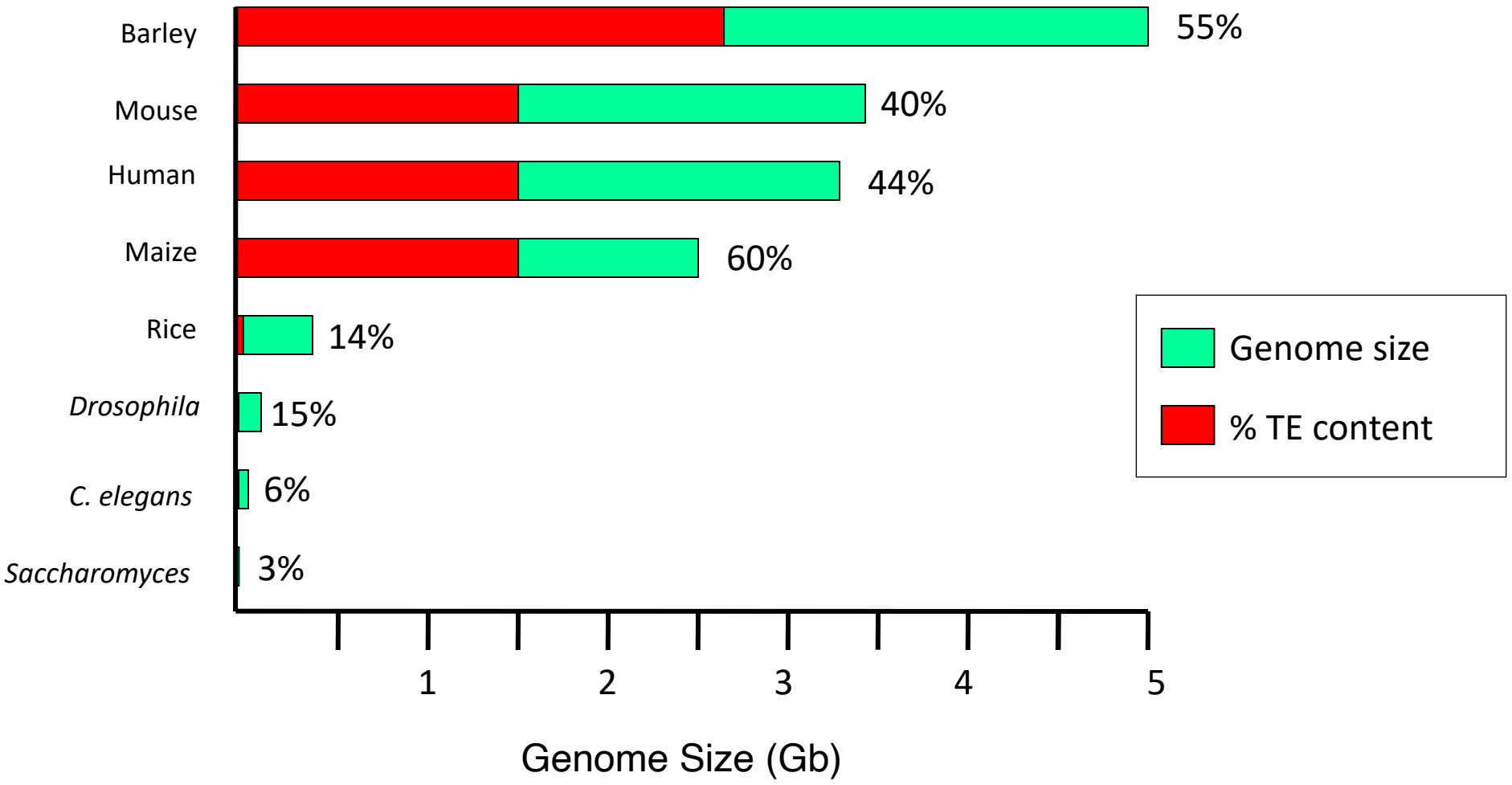
2. Do you have the time? Personnel? Bioinformatics experience?

Animal Genome Sizes

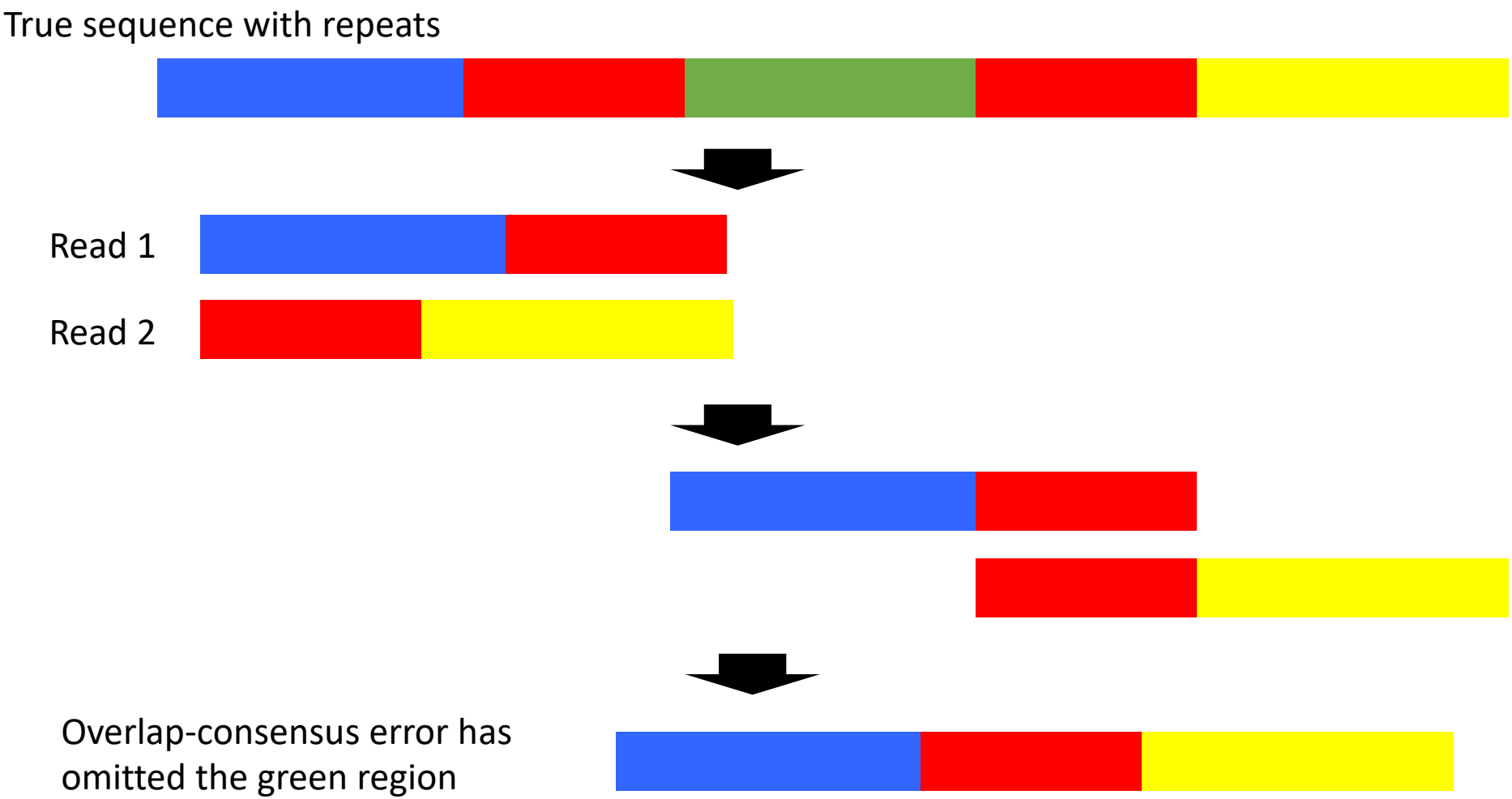
Genome Size in Gigabases

http://genomesize.com/

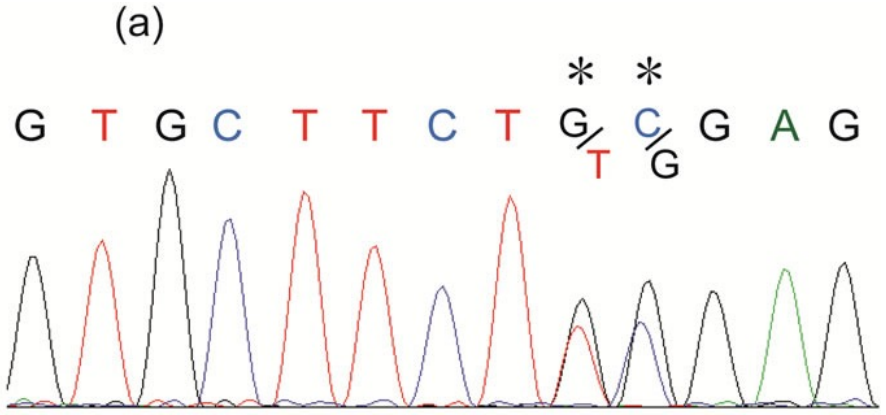
Large genomes have a lot of transposable elements



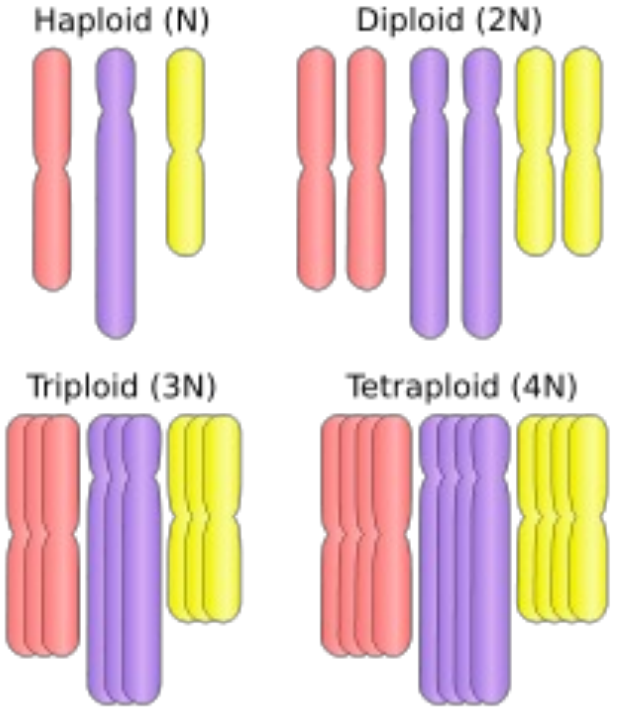
Problem of Repeats



Other Questions



Expected Heterozygosity



Ploidy

Sequencing Technologies

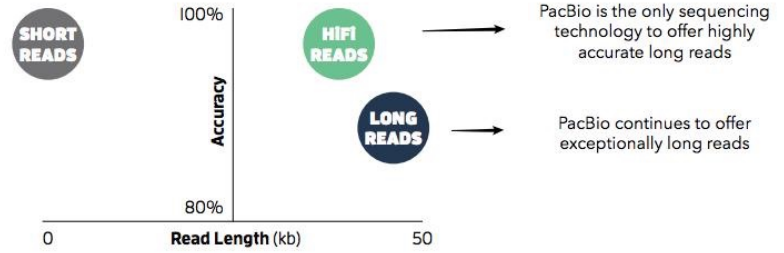
- Sanger method – old workhorse
  - “First generation sequencing”
  - Lower coverage, longer reads, fewer errors
- Next-generation sequencing – now standard
  - “Second generation”, PCR, short reads, more errors
  - 454, Illumina, SOLiD
- Third generation – becoming standard (expensive)
  - Single molecule, long reads, many more errors
  - PacBio, Oxford Nanopore

Basic Overview of Sequencing Technologies

	Company	Platform	Read Length	Time per run	Number of reads per unit	Common error
Short reads	Illumina	NovaSeq 6000	2 X 250bp	≤44 hours	≤40 billion	Substitution
	454 Life Sciences*	GS FLX Titanium XL	≤1000bp	700Mb/day	~150,000	indel
	Applied Biosystems*	SOLiD 5500xl W	25-50bp	5-8 days	~300 million	Substitution
Long reads	Pacific Biosciences	PacBio Sequel	≤60 kb (average 25 kb)	30 hours	≤4 million	indel
	Oxford Nanopore	MinION	≤4 Mb (average 6 kb)	≤72 hours	Tens of millions	indel

*retired

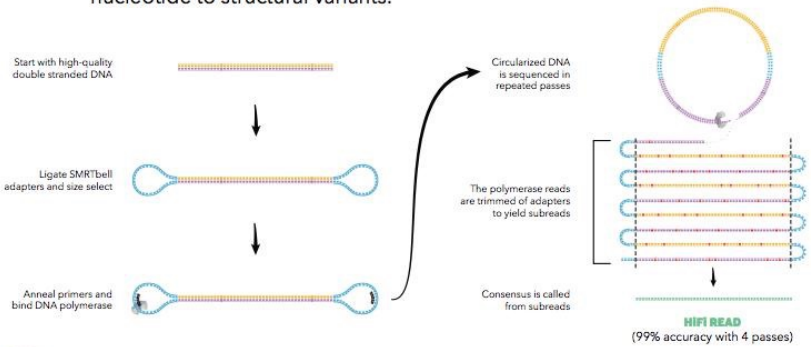
PacBio Sequel System



HiFi READS

Generate Highly Accurate Long Reads

Produce HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution for detection of all variant types from single nucleotide to structural variants.



LONG READS

Optimize Your Run for Even Longer Reads

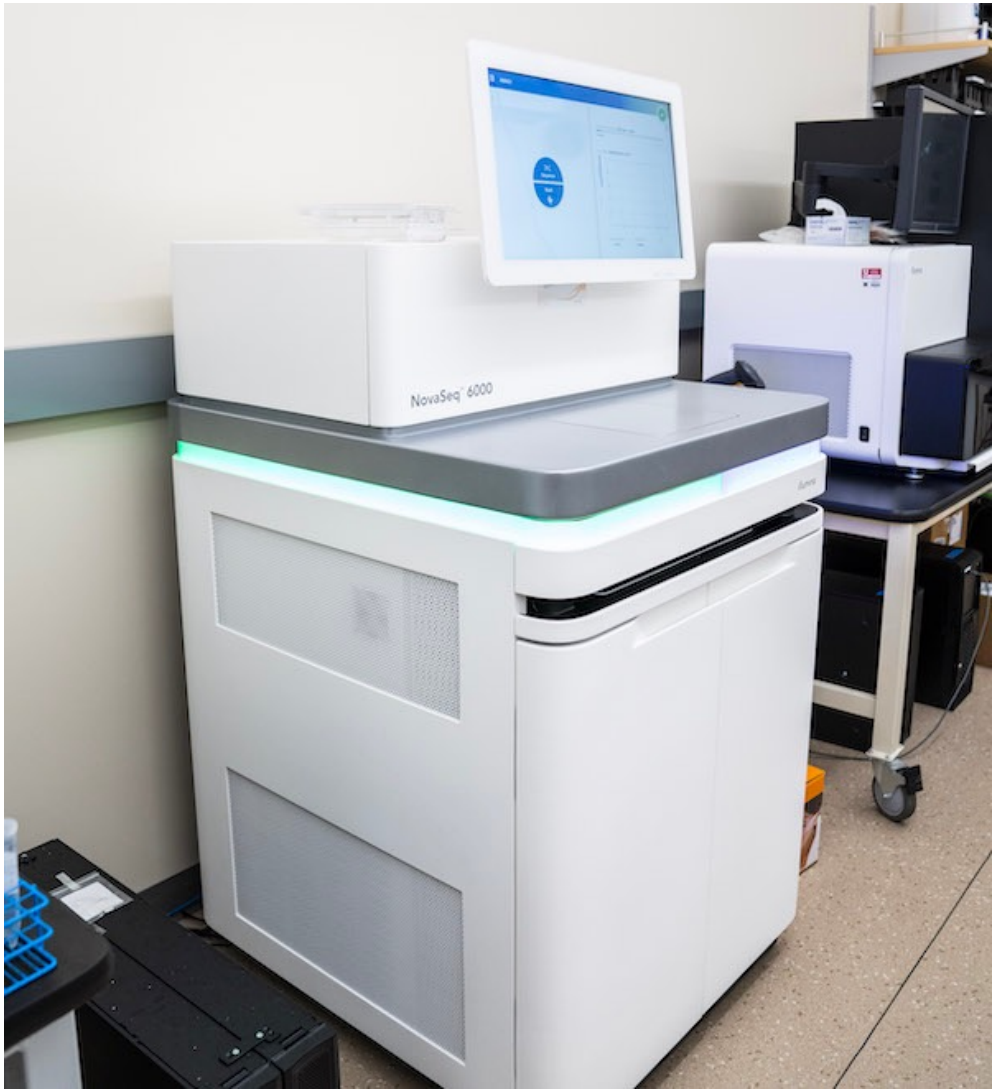
Sequence read lengths in the tens of kilobases using the continuous long read (CLR) sequencing mode to enable high-quality assembly of even the most complex genomes.

Half of Data in Reads
>50 kb

Longest Reads Up To
175 kb

Oxford Nanopore

illumina®



Illumina Paired-end and Mate-pairs

Paired-end (PE) "short insert library" sequencing

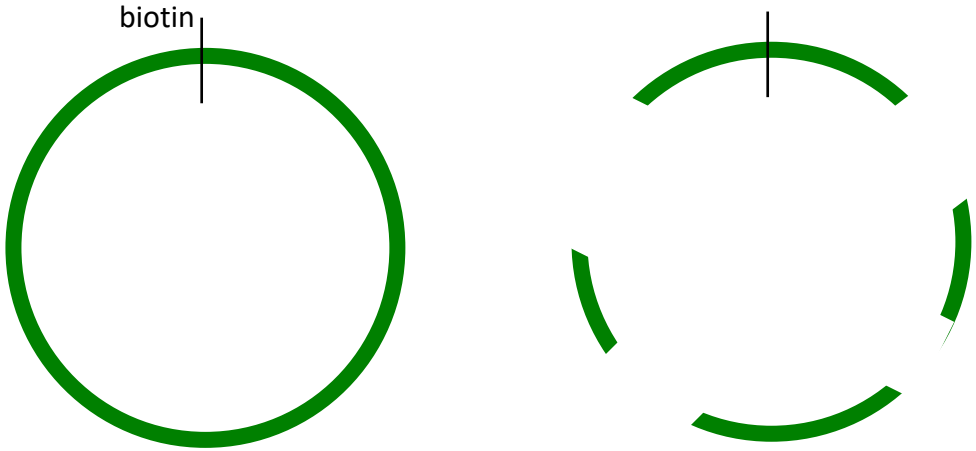
orientation

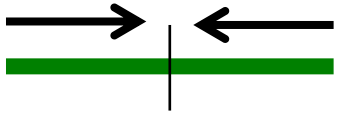
- Genome is fragmented to desired lengths
- Reads one end of the molecule, flips and then reads the other end
- Generates read pairs with a known distance between them

FR

Mate-pair (MP)"jumping library" sequencing

- Circularizes longer molecules (2kb-25kb)
- Biotinylated, fragmented, enriched, and sequenced





RF

Repeats Resolved

- Repeats can be resolved using paired-end information
- If one end of a read is unique, then you can map both reads.

Repeats Resolved

- Repeats can be resolved using paired-end information
- If one end of a read is unique, then you can map both reads.

- However, for longer repeats (*i.e.* LINEs) this will not work.
- Hence Illumina-based genomes tend to be fragmented.

How much to sequence?

$$\text{Coverage} = \frac{(\text{read number} * \text{read length in bp})}{\text{Genome size in bp}}$$

(The number of times a site in the genome is represented by a read)

How much to sequence?

Say, a typical mammal…

$$100X = \frac{(2{,}000{,}000{,}000 * 150)}{3{,}000{,}000}$$

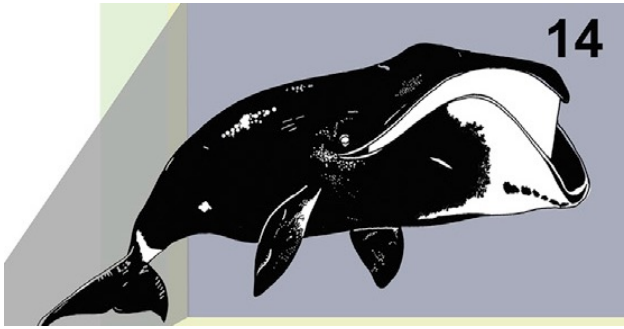
Libraries	Total Data (Gb)	Sequence Coverage
200bp paired-end	149.1	51.2X
500bp paired-end	141.7	48.7X
3kb mate-paired	57.3	19.7X
5kb mate-paired	72.5	24.9X
10kb mate-paired	28.5	9.8X
	449.1	154.3X

Cell Reports

Resource

Insights into the Evolution of Longevity from the Bowhead Whale Genome

Keane et al. (2015), *Cell Reports*



Fastq format

- Each sequence has four lines
  1. sequence name starting with “@”
  2. Nucleotide sequence
  3. Empty line except for “+”
  4. Quality score information.

Phred Quality Scores

Logarithmically linked to error probabilities
Estimates several parameters based on peak shape and resolution at each base
These estimated parameters are then compared to lookup tables from known sequences
Allows the automation of quality control – especially helpful for large numbers of reads
Was invented for the Human Genome Project

Base-calling of automated sequencer traces using phred. II. Error probabilities. Ewing B, Green P (1998). Genome Research.

Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Ewing B, Hillier L, Wendl MC, Green P. (1998). Genome Research.

Phred Quality Score	Probability of Error	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

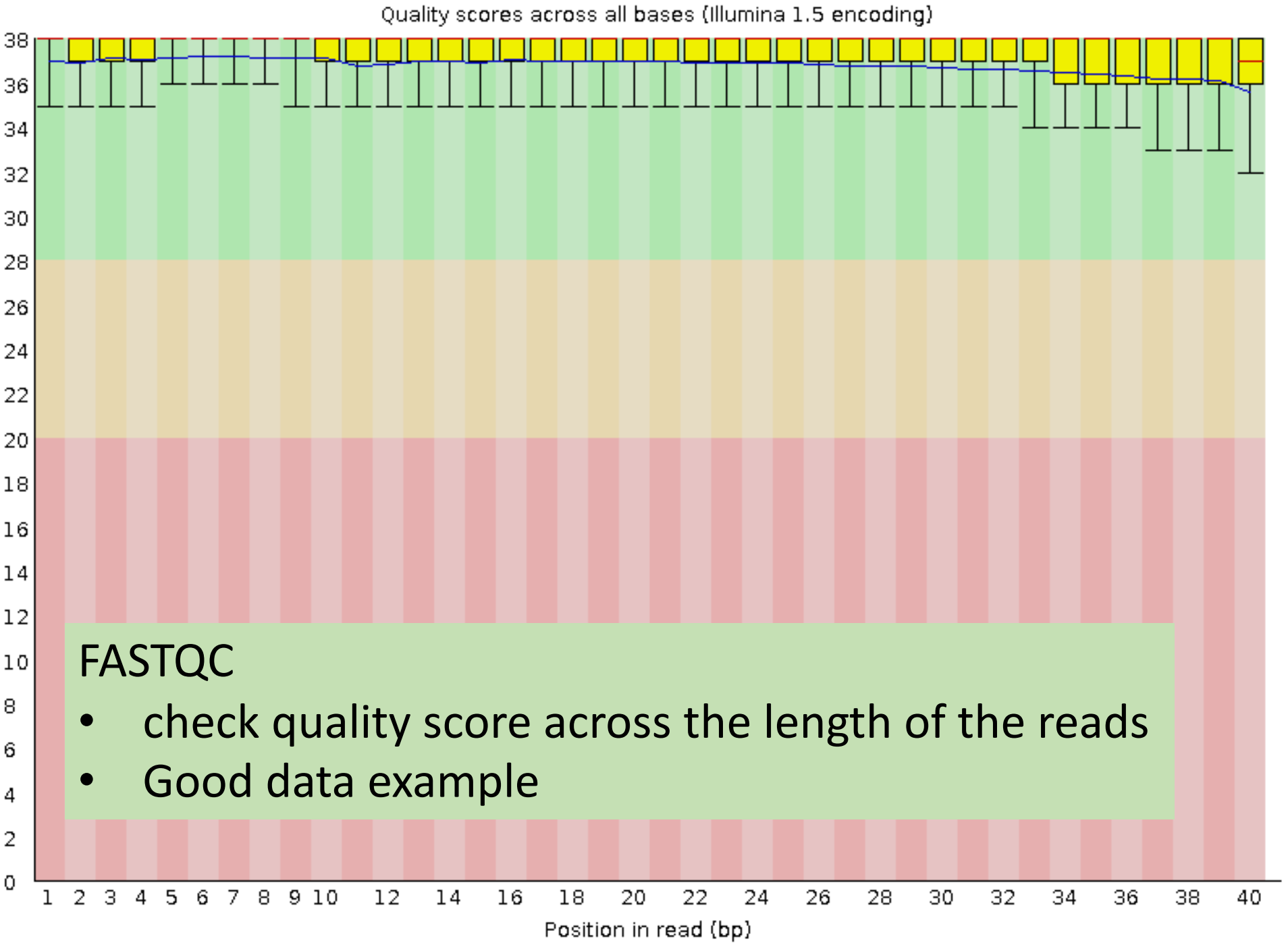
Data Quality Control

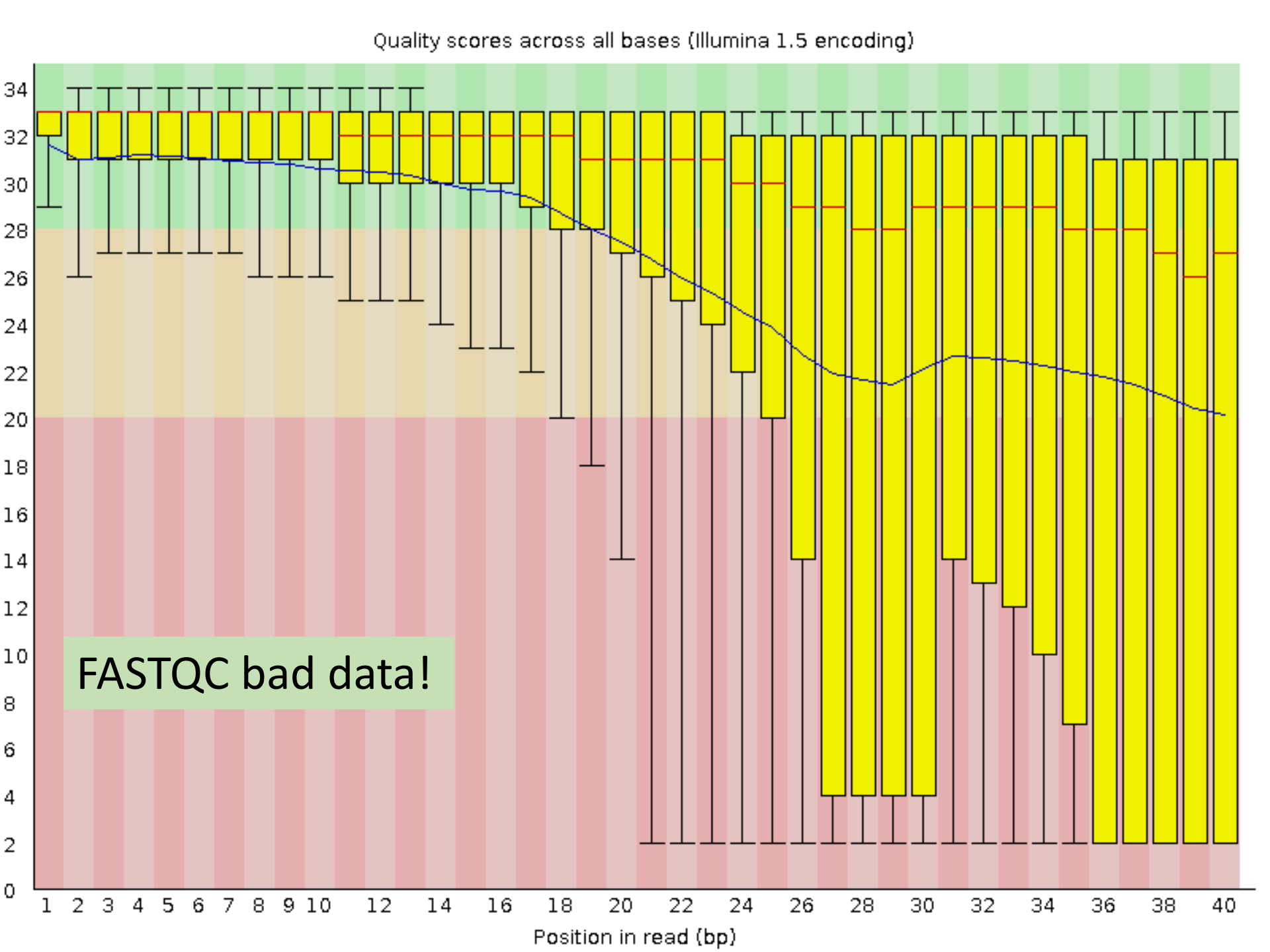
Quality score?

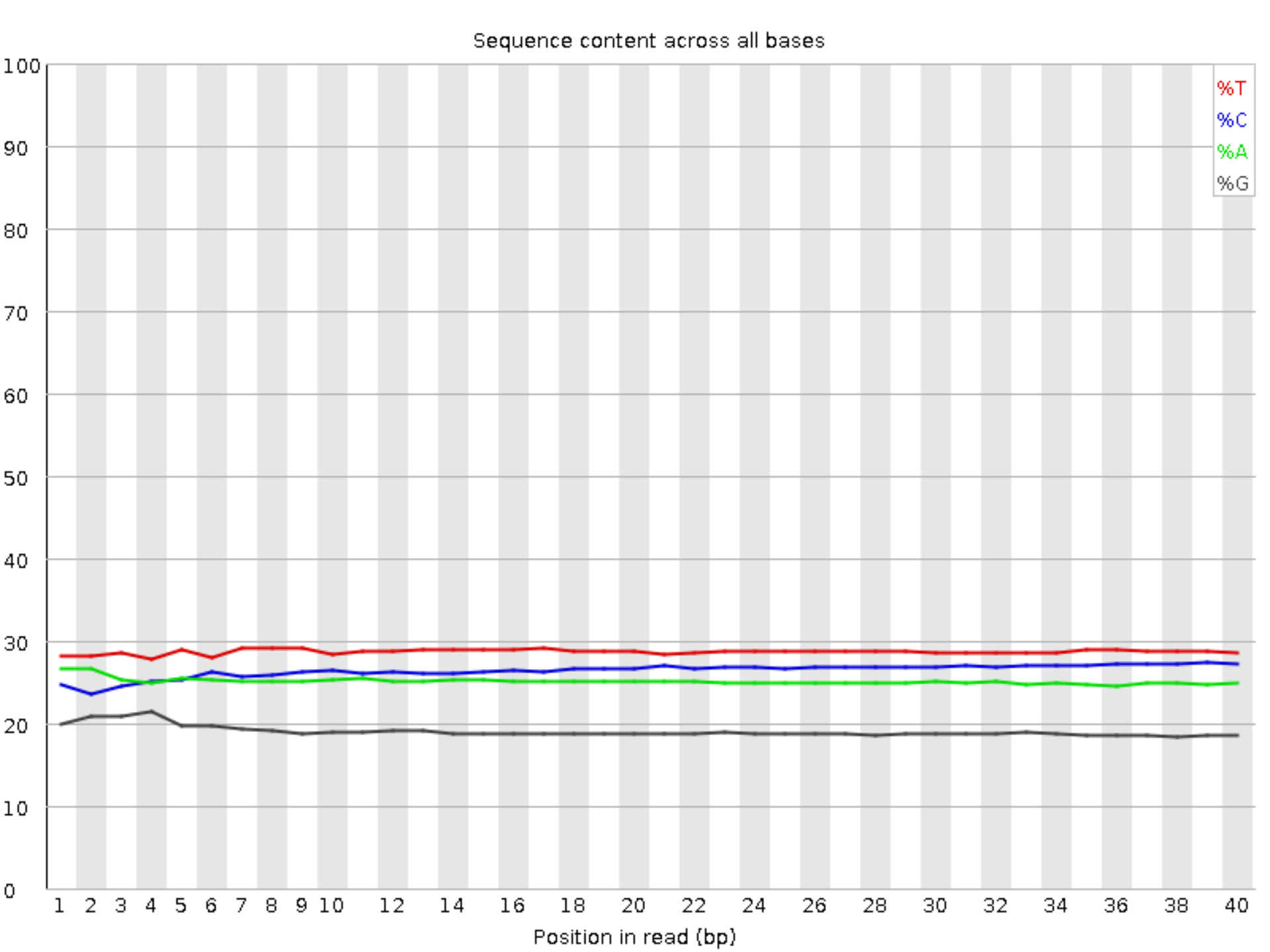
GC content?

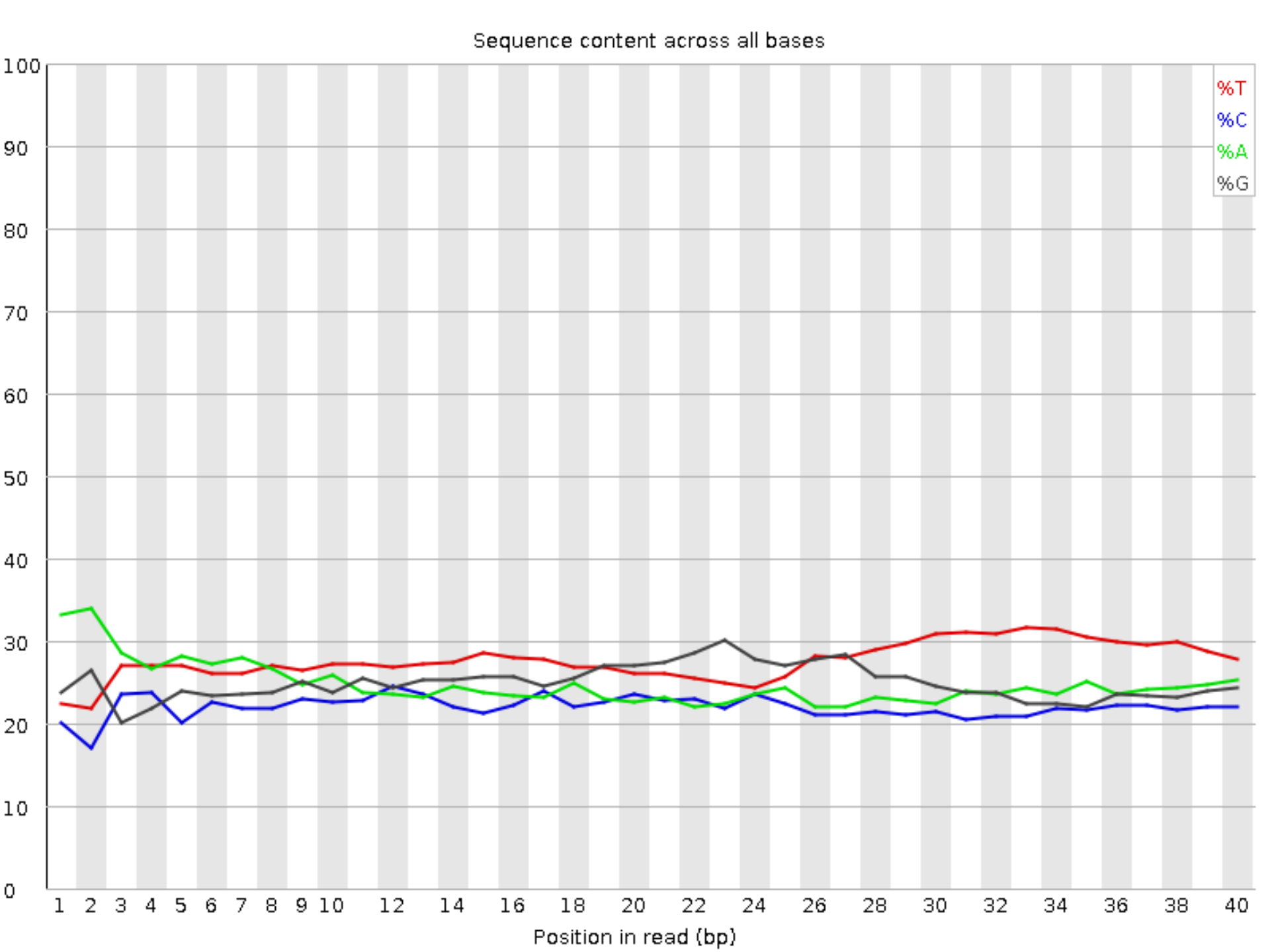
Sequence duplication levels?

Overrepresented sequences, or adaptors?


Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)
FASTQC
• check quality score across the length of the reads
• Good data example


Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)
FASTQC bad data!


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

Trimming

- Trimmomatic (Bolger et al. 2014 *Bioinformatics*)
  - Paired end mode:
    - ILLUMINACLIP – trim adapters using database
    - HEADCROP
    - LEADING:3
    - TRAILING:3
    - SLIDINGWINDOW:4:15
    - MINLEN
  - Will trim entire dataset in pairs (F and R), and output singletons whose mates were eliminated

Three types of biases in NGS data

- Systematic bias
  - Problem with PCR, sequencer or library prep
  - Errors in base-calling
  - GC bias
  - High duplication rates
- Coverage bias
- Batch effects
  - Non-biological differences between experimental groups

Error correction

Substitution errors are most common in Illumina datasets

- In theory, errors should be infrequent (0.0001) and random
- But if you have 1 billion reads, that's 100,000 errors

Can be addressed by laying out all the reads covering a position

Use majority of reads to find (rare) erroneous sites, and correct them

Error correction

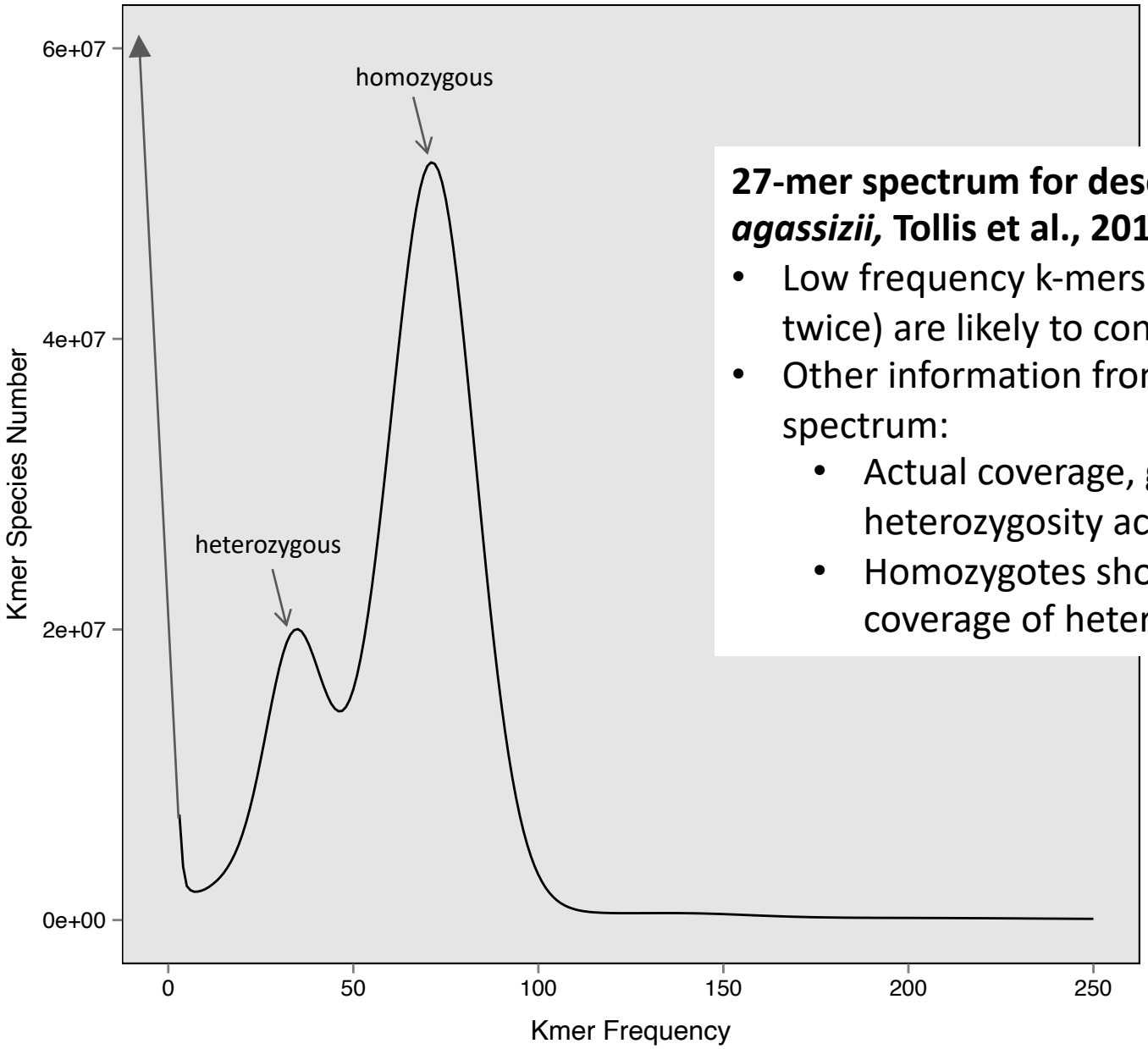
Error correction needs to be done before assembly!

K-spectrum-based correcting:

- Reads are broken up into k-mers
- Distances between k-mers are calculated
- Errors are corrected

Some tools are:

- SOAPdenovo error correction (Li et al. 2010)
- Quake (Kelley et al. 2010)
- Reptile (Yang et al. 2010)



27-mer spectrum for desert tortoise (*Gopherus agassizii,* Tollis et al., 2017 *PLoS One*)

- Low frequency k-mers (those occurring once or twice) are likely to contain errors
- Other information from k-mer frequency spectrum:
  - Actual coverage, given repetitiveness and heterozygosity across genome
  - Homozygotes should in theory be 2X coverage of heterozygotes

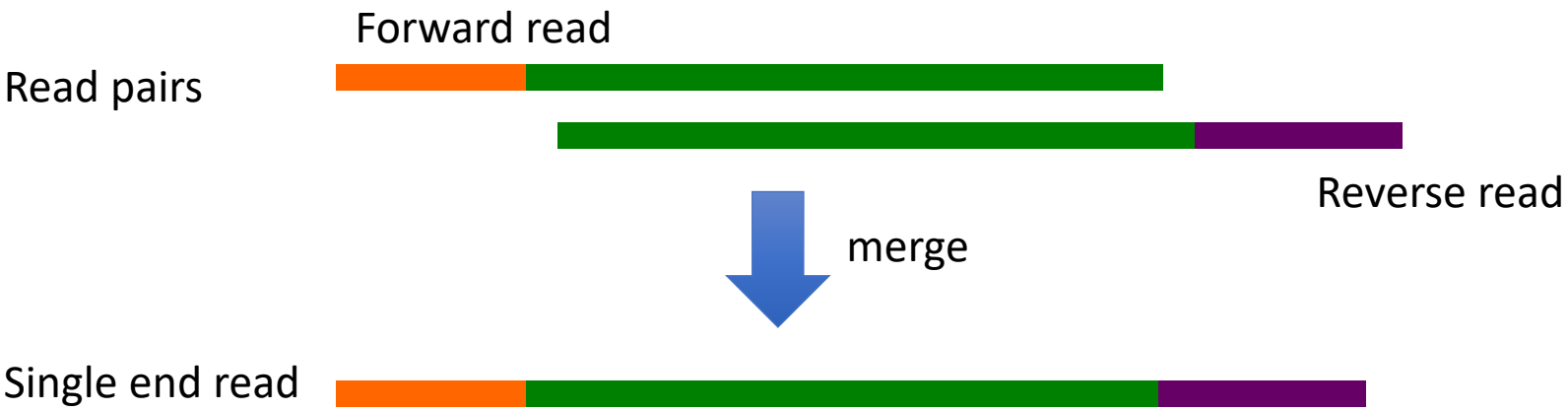
Merging reads

Joins together overlapping reads to create single-end reads

If you have 180/200bp libraries and ~100bp reads, this is recommended.

Speeds up assembly considerably so the assembler does not have to calculate the distance between these reads.

FLASH, Cope, PEAR are some tools that do this.



Choose genome, gather info

↓

DNA Library preparation

↓

Sequencing

↓

Quality check

↓

Trimming

↓

Error correction

↓

Merge overlapping reads

↓

ASSEMBLE!