Introduction to metagenomics: key concepts and examples of analyses

Dr. Natalia Zajac

natalia.zajac@fgcz.ethz.ch

29.03.2023

Schedule for the data analysis part

Data QC

- **Theory** – this lecture
- **Hands-on excercises** – Using SUSHI platform developed in Functional Genomics Center Zurich

Amplicon-based metagenomics

- **Theory** – main workflows in data analysis
- **Hands-on excercises** – data analysis by **QIIME2** bioinformatics platform on two datasets

Shotgun metagenomics

ASSEMBLY-BASED METHODS

- **Theory** – main workflows in data analysis
- **Examples** – a guest talk by Dr. Silas Kieser on **metagenome-atlas**
- **Hands-on excercises** – data analysis with metagenome-atlas

Schedule for the data analysis part

Shotgun metagenomics

ASSEMBLY-FREE METHODS

- **Theory** – main workflows in data analysis
- **Hands-on excercises** – data analysis with **Kraken2**

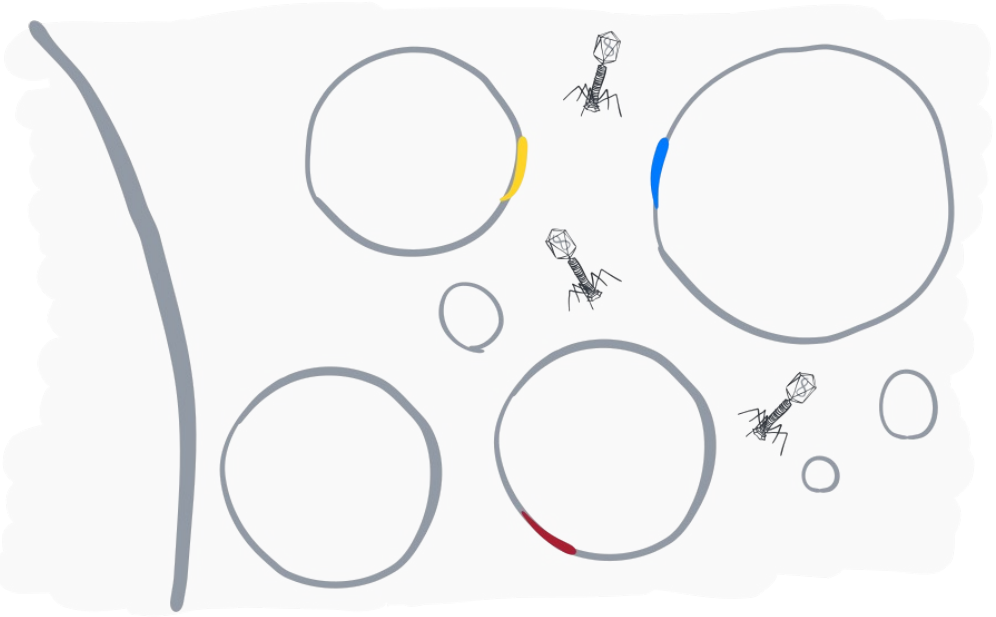
Shotgun metatranscriptomics

- **Theory** – main workflows in data analysis
- **Examples** – guest talk by Dr. Jonas Grossmann on metaproteomics
- **Hands-on excercises** – data analysis with Samsa2

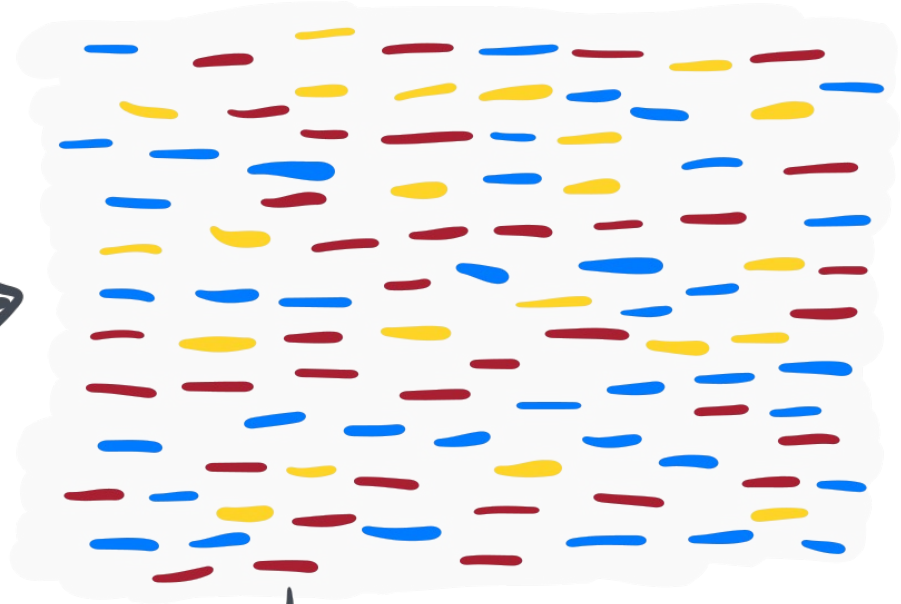
16S vs Shotgun Metagenomic Sequencing

Factors	16S/ITS/18S sequencing	Shotgun Metagenomic Sequencing
Cost	~$50 USD	Starting at ~$150 but price will depend on sequencing depth required
Sample preparation	Similar complexity to shotgun sequencing	Similar complexity to 16S rRNA sequencing
Functional profiling (profile microbial genes)	No (but 'predicted' functional profiling is possible)	Yes (but it only reveals information on functional potential)
Taxonomic resolution: Genus, species, strain?	Bacterial genus (sometimes species); dependent on region(s) targeted	Bacterial species (sometimes strains and single nucleotide variants, if sequencing is deep enough)
Taxonomic coverage	Bacteria and archaea	All taxa, including viruses
Bioinformatics requirements	Beginner to intermediate expertise	Intermediate to advanced expertise
Databases	Established, well-curated	Relatively new, still growing
Sensitivity to host DNA contamination	Low (but PCR success depends on the absence of inhibitors and the presence of a detectable microbiome)	High , varies with sample type (but this can be mitigated by calibrating the sequencing depth)
Bias	Medium to high (retrieved taxonomic composition is dependent on selected primers and targeted variable region)	Lower (while metagenomics is "untargeted", experimental and analytical biases can be introduced at various stages)

Taken from: Microbiome Insights Blog, Ruairi Robertson, PhD, https://blog.microbiomeinsights.com/16s-rrna-sequencing-vs-shotgun-metagenomic-sequencing#rrna-sequencing-16s

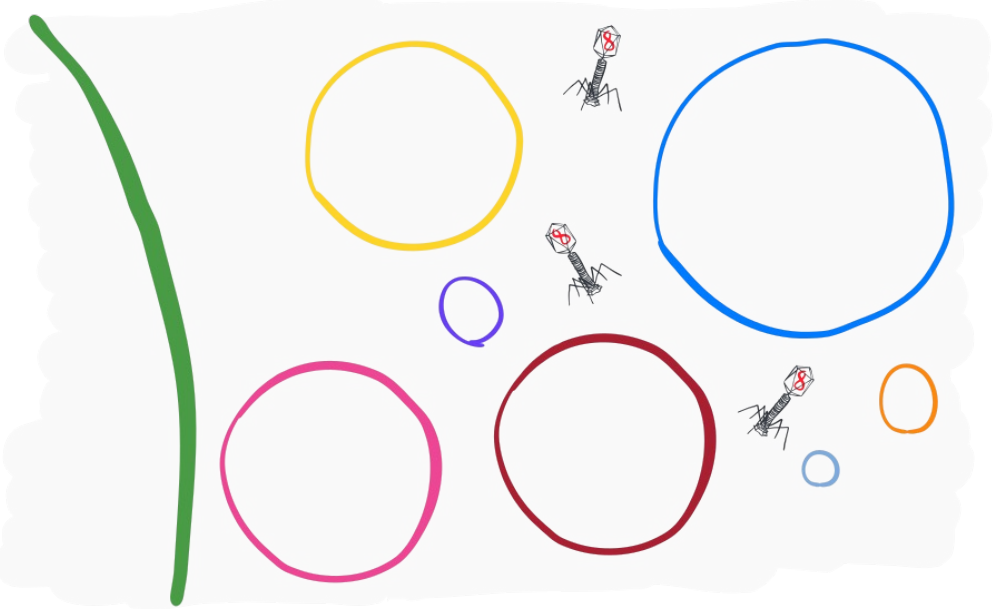


AMPLICON SEQUENCING

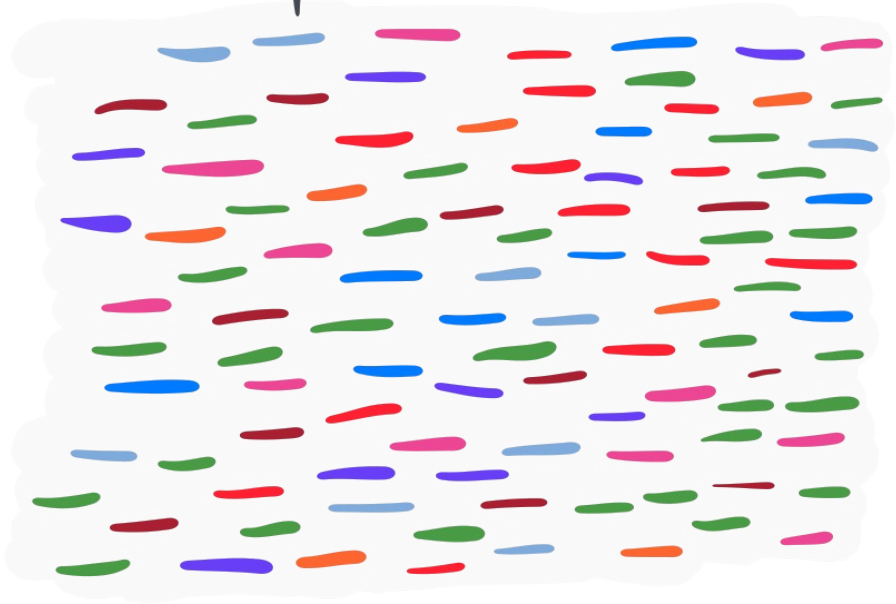


AMPLICONS

METAGENOMIC SHORT READS



SHOTGUN SEQUENCING

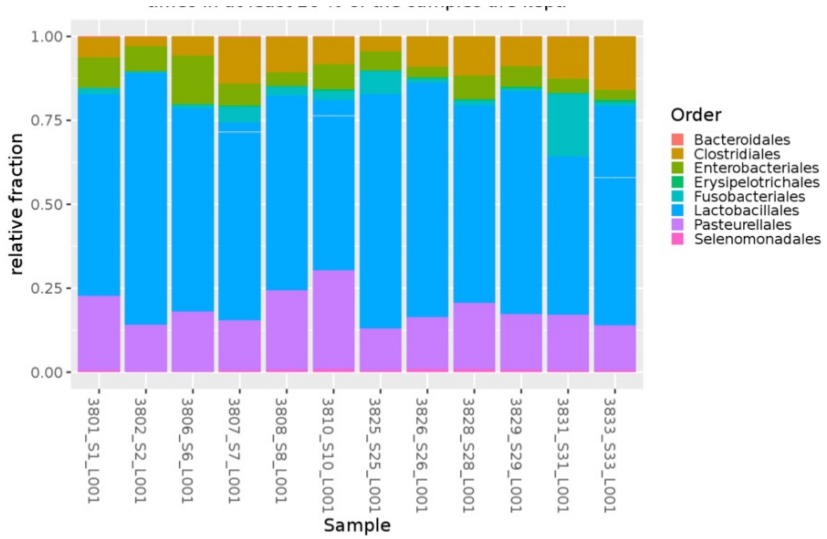


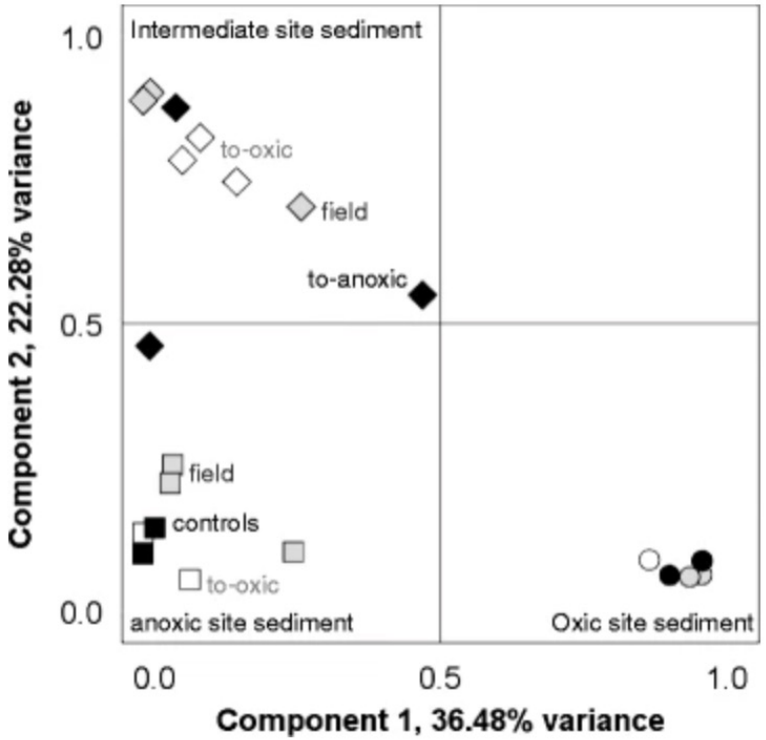
https://merenlab.org/momics/

Amplicon-based metagenomics

Community composition:

- **do samples cluster according to their community composition?**
- **which organisms are present in which abundance?**





Principal component analysis of the OTU communities in the top 1-cm sediment layer. Symbols: *gray circles*, oxic field; *black circles*, oxic control; *white circles*, oxic-to-anoxic; *gray squares*, anoxic field; *black squarse*, anoxic control; *white squares*, anoxic-to-oxic; *gray diamonds*, intermediate field; *white diamonds*, intermediate-to-oxic; and *black diamonds*, intermediate-to-anoxic

Source: Broman et al. 2017 Shifts in coastal sediment oxygenation cause pronounced changes in microbial community composition and associated metabolism

Amplicon-based metagenomics

Shotgun metagenomics

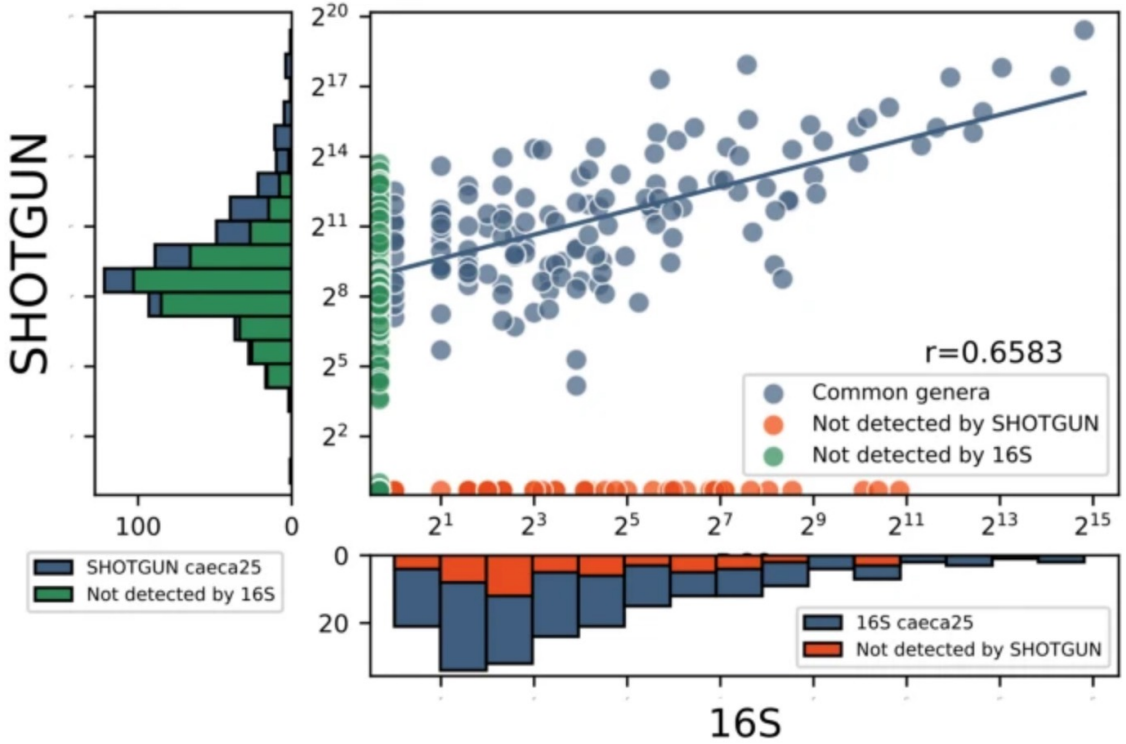
Taxonomic deconvolution, Community diversity

Article | Open Access | Published: 04 February 2021

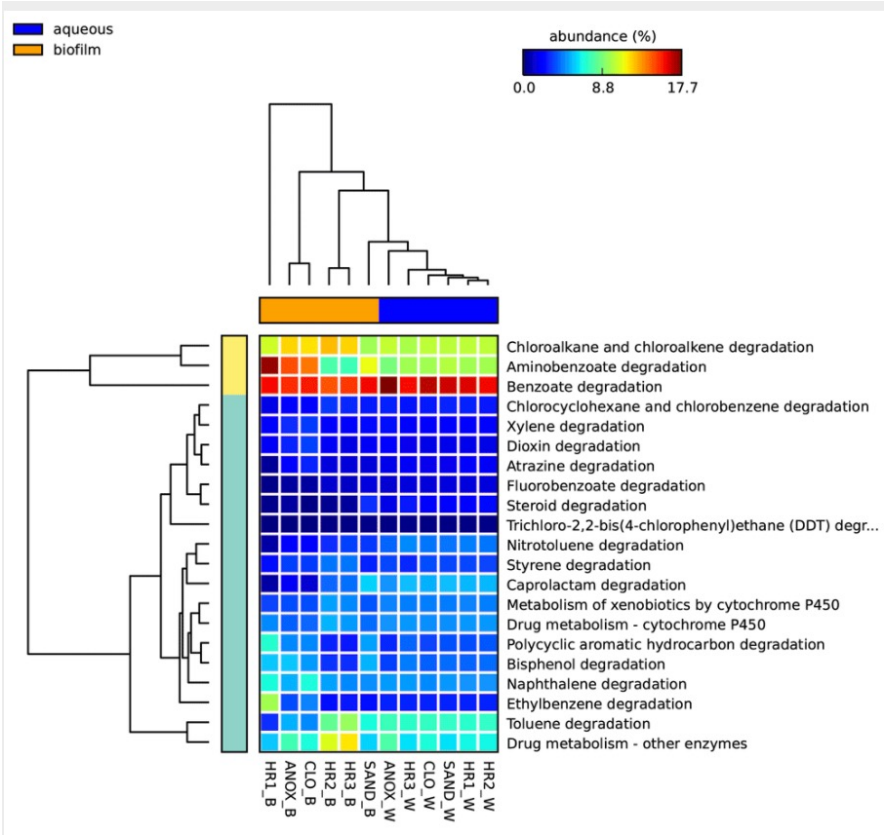
Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota

Francesco Durazzi, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Re
Alessandra De Cesare

Scientific Reports **11**, Article number: 3030 (2021) | Cite this article



Amplicon-based metagenomics



Heat map indicating the relative number of sequence reads associated with the KEGG xenobiotic metabolism categories (on right) identified in each sample location and phase (on bottom)

Source: Balcom et al. 2016
Metagenomic analysis of an ecological wastewater treatment plant's microbial communities and their potential to metabolize pharmaceuticals

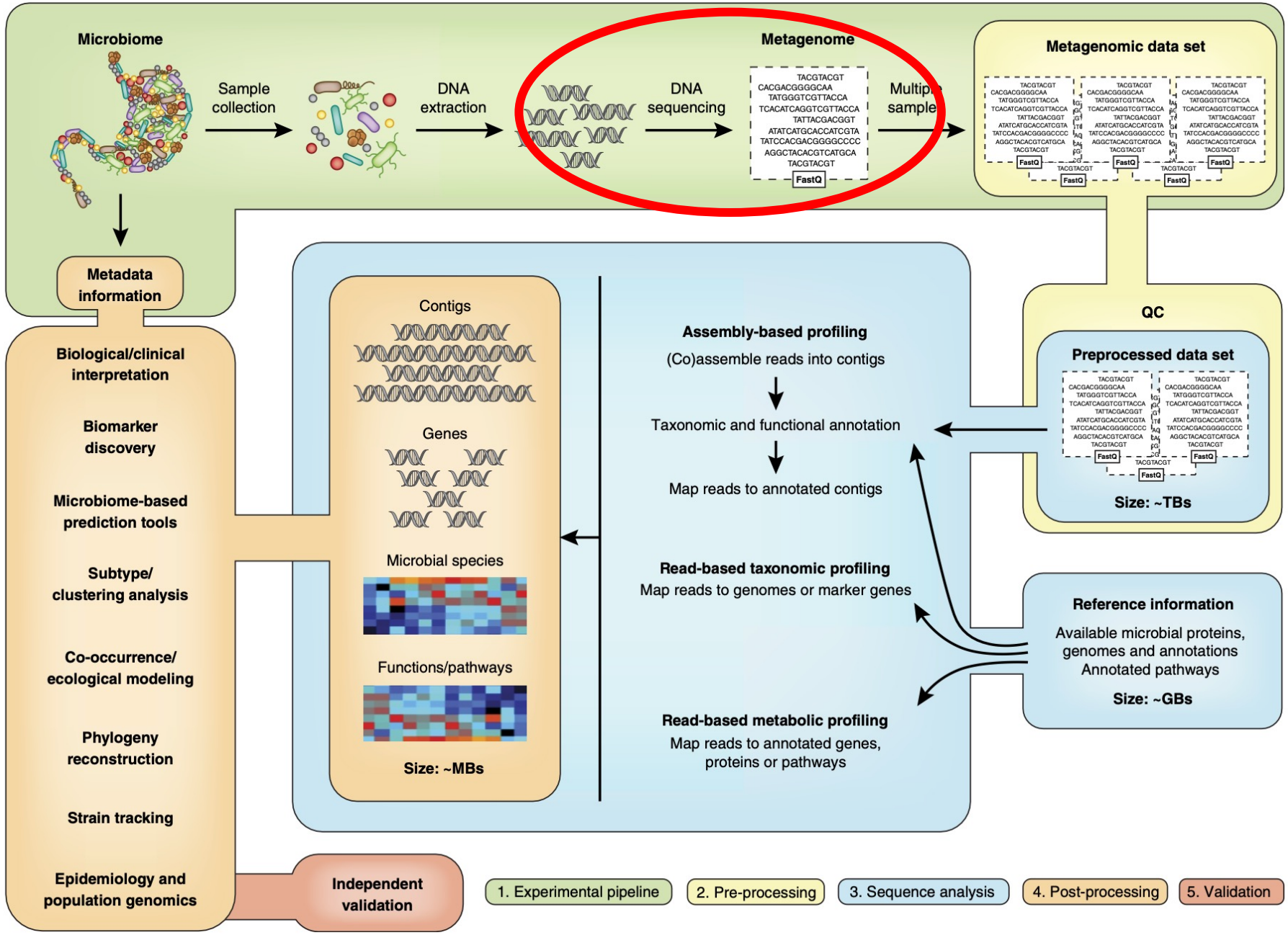
Shotgun metagenomics

Assembly of genomes and taxonomic profiling

Functional annotation:

- **Importance of pathways**: how are the different taxa distributed in a functional category?
- **Comparative:** Do we see enrichment on certain gene ontologies in species present in the sample.

1. Choosing the type of sequencing

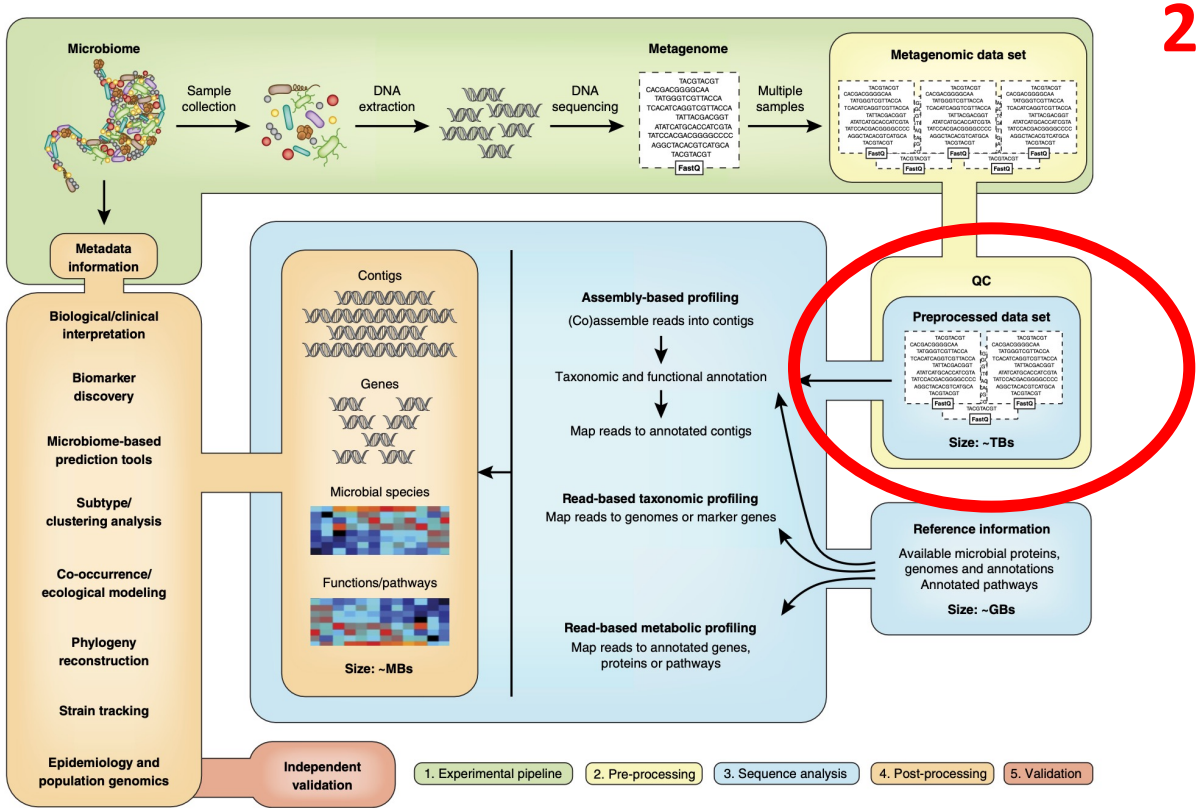


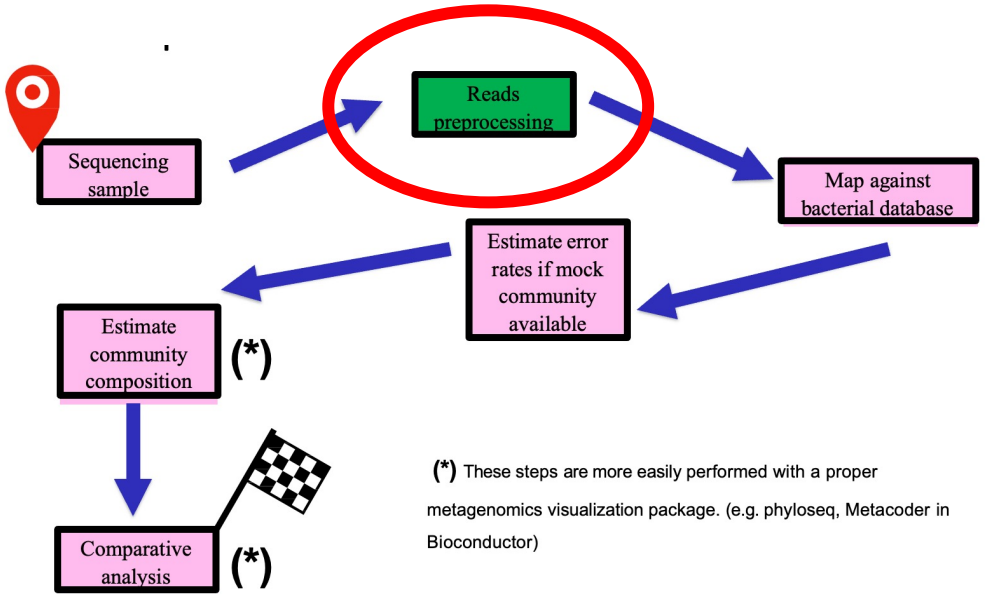
Source: Shotgun metagenomics, from sampling to analysis. Quince et al. 2017. Nature biotechnology

More on the different technologies maybe

	Short read Illumina sequencing	Long reads PacBio or ONT sequencing
Accuracy	Illumina > 99.9%	Pacbio hifi reads >99% ONT 90-100%
Length	2 × 251 bp , 2 x 150 bp	Pacbio: 16S: 1.6 kb HiFi reads generate up to 3.5 million >Q20 Shotgun: 10 kb HiFi reads generate up to 2.4 million >Q20 ONT: 10-20kb
Primary error	Substitution	Pacbio: Insertion ONT: Deletion/Substitution

2. QC and pre-processing



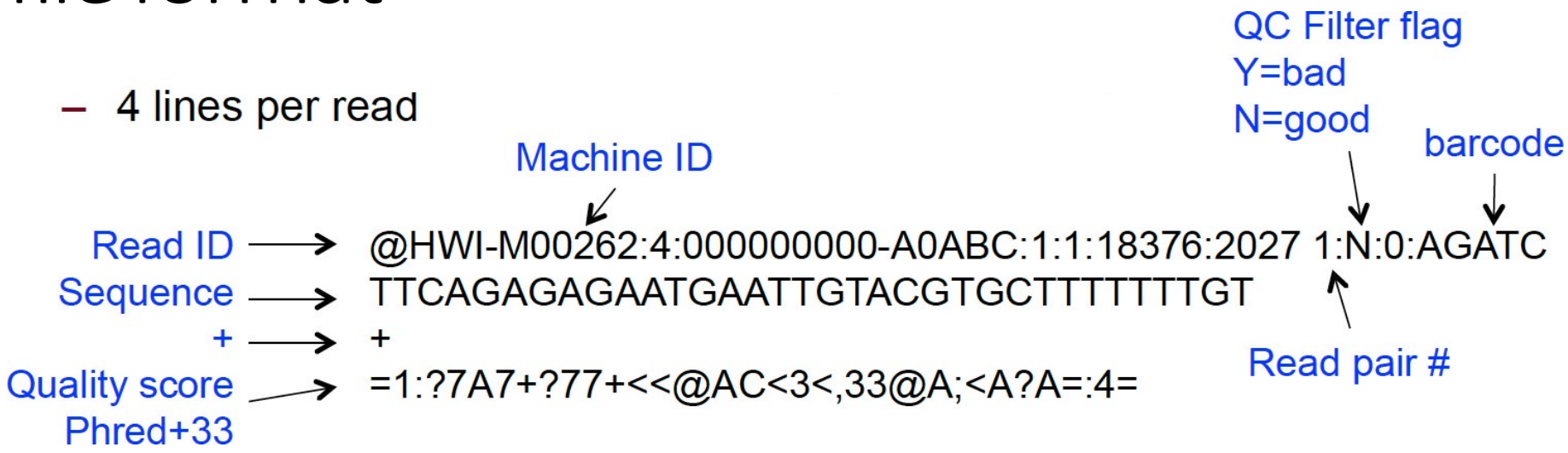


(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

Source: Shotgun metagenomics, from sampling to analysis. Quince et al. 2017. Nature biotechnology

Fastq file format

- 4 lines per read



1. Header line for Read (starts with "@" and the sequence ID)
2. Sequence
3. Header line for Qualities (starts with "+")
4. Quality score

fastq header format (version > 1.8)

Sequence Header +Sequence ID

a b c d e f g h i j k

@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

a. unique instrument name
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile

h. the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
i. Y if the read fails filter (read is bad), N otherwise
j. 0 when no control bits are on
k. index sequence

QC and pre-processing – why and how?

- Raw unprocessed data from a sequencer.
- Will ultimately affect the final taxonomic and functional composition.
- Will save you a lot of time when interpreting the results

1. General Statistics:
   - GC content
   - PCR duplicates
   - Adapter Sequences
   - Quality score
   - Ambiguous bases
   - Sequence length
2. Trimming
3. Removing sequencing artifacts
4. Host decontamination

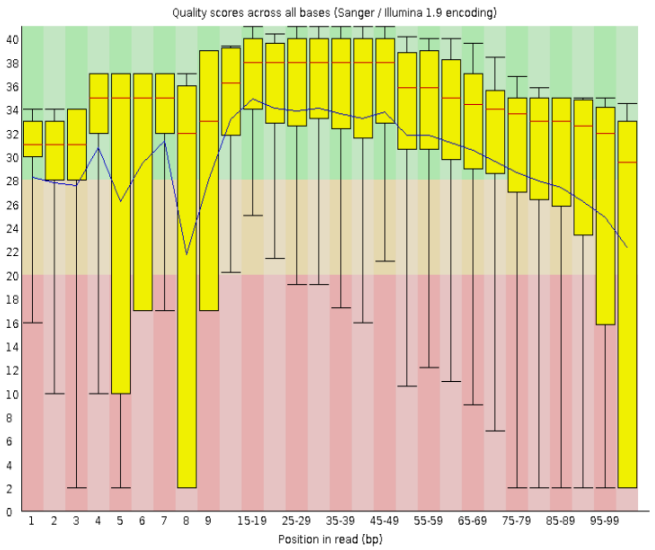
Phred scores

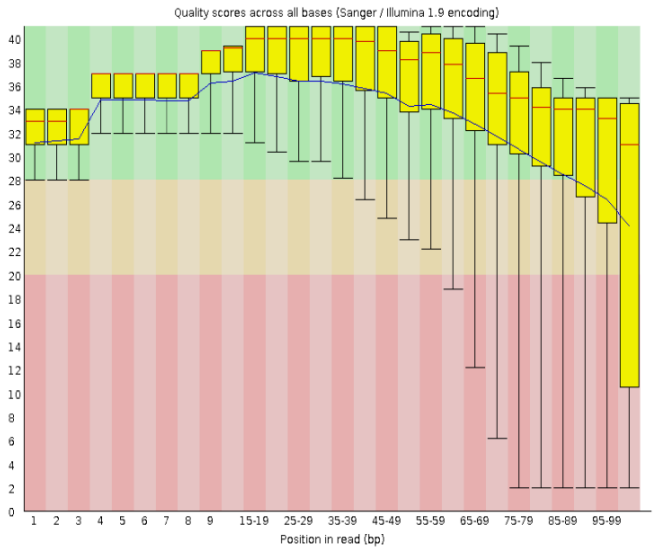
- Measure base calling accuracy
  Accuracy of assigning **bases** (nucleobases) to signal peaks
- P
  error probability of a given base call
- *Q*
  $-10\log_{10}P$
- Assign to each base
- Range from 0-41 for Illumina sequencing

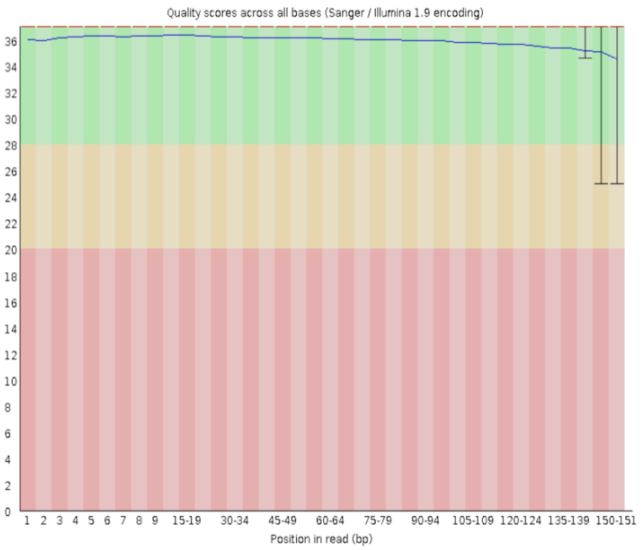
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Ewing B, Green P. 1998. Genome Res. 8(3):186-194.

Phred score distribution per base



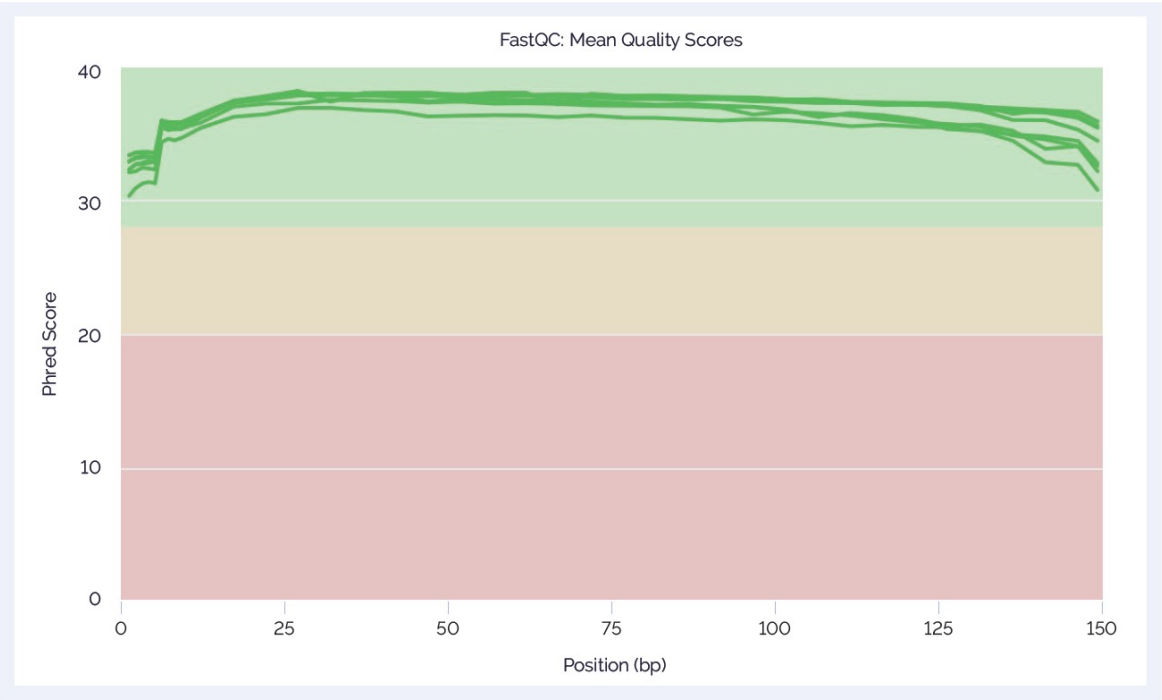


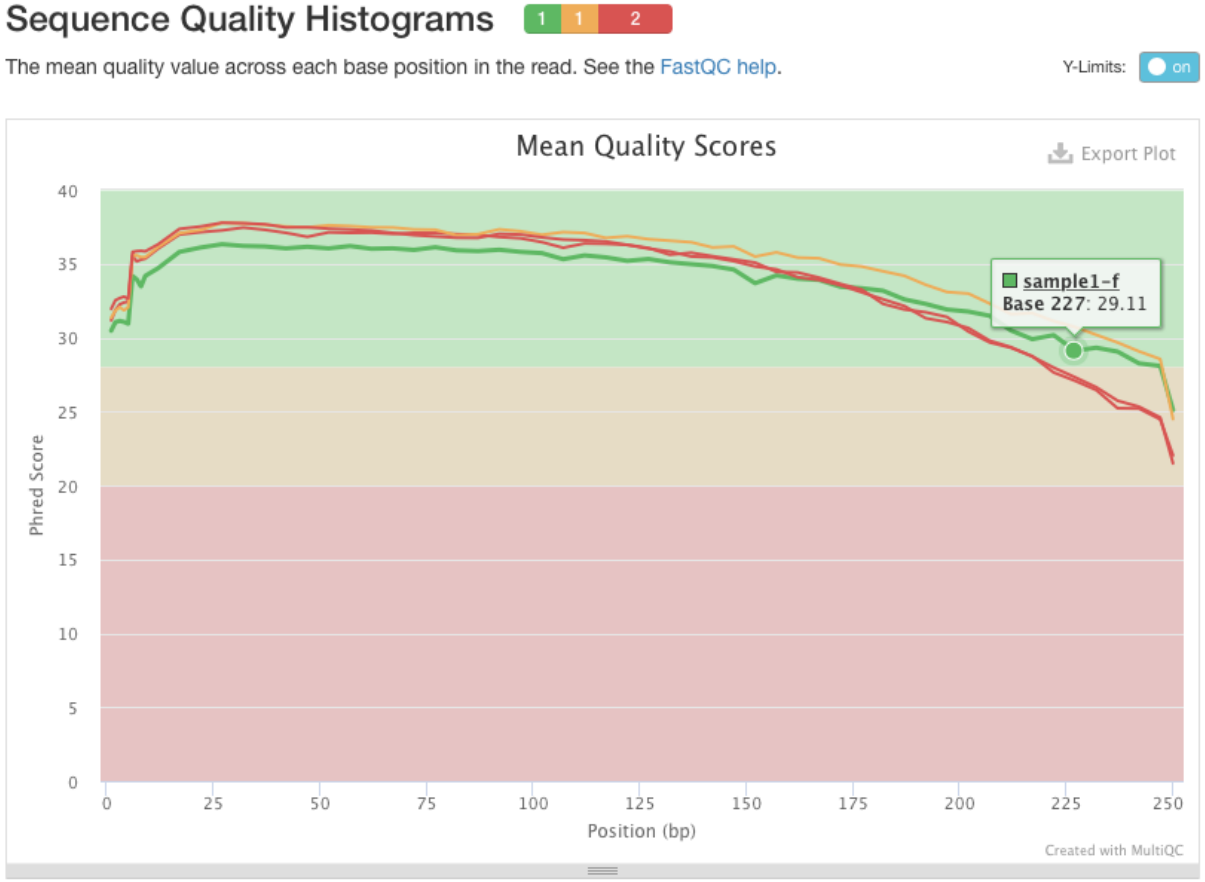


- **Poor scores on both ends**
- **Large variance**

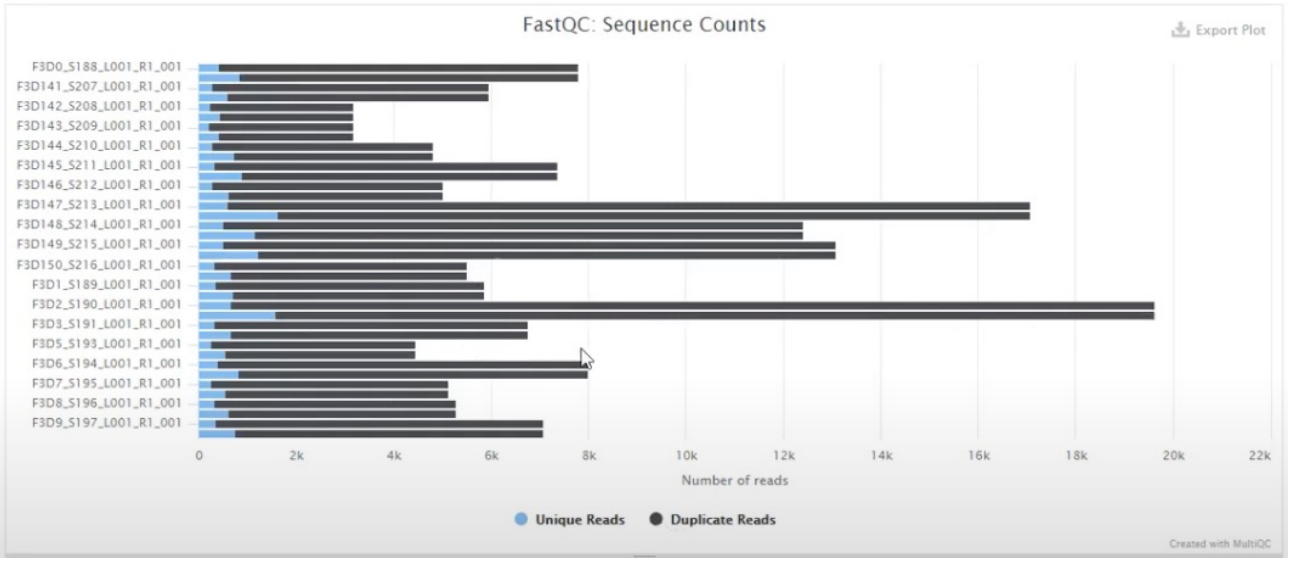
- **Very high and consistent quality along the reads**

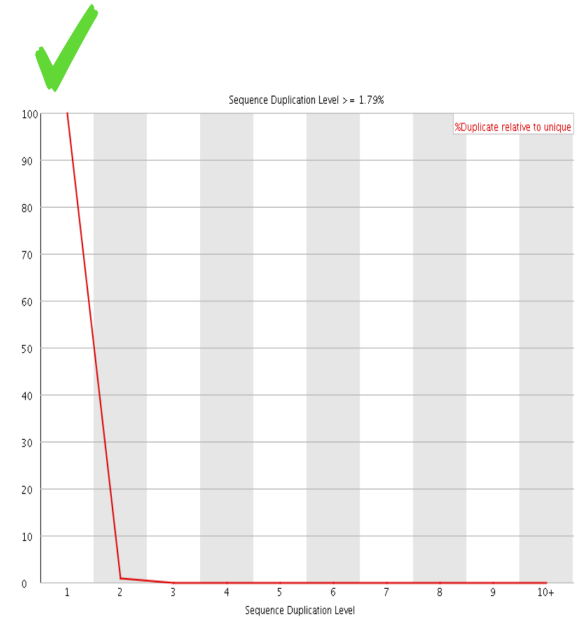
Sequence quality histograms



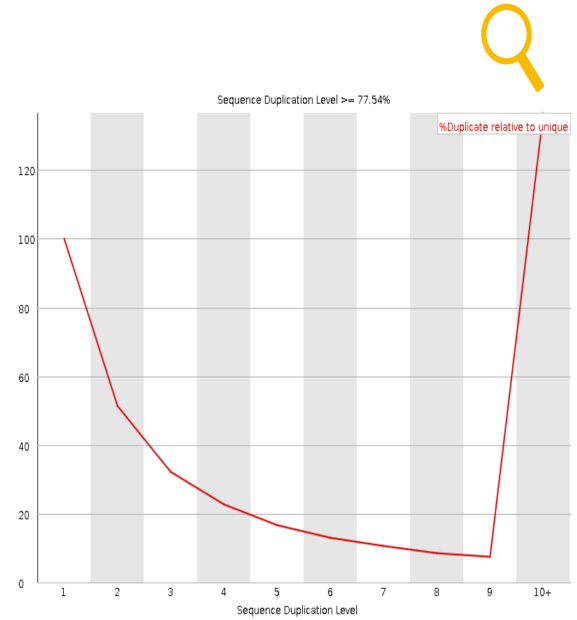


Sequence duplication





- Essentially no duplication



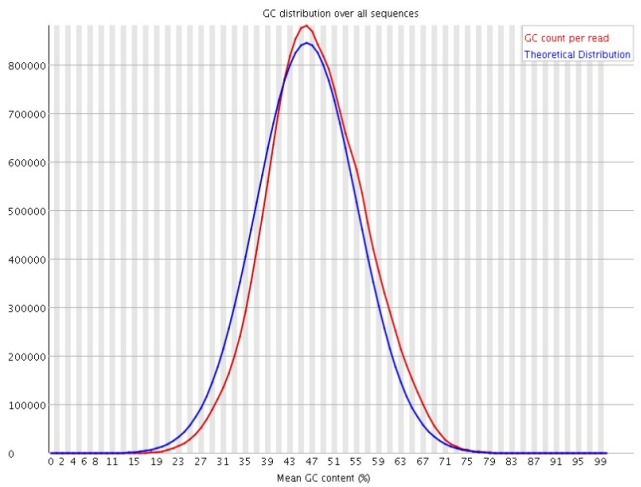
- Enrichment bias
- PCR over amplification, too little input material
- Often seen in RNA-seq
- Of course normal in Amplicons

GC content

Bacterial genomic GC content varies from **less than 15% to more than 75%**

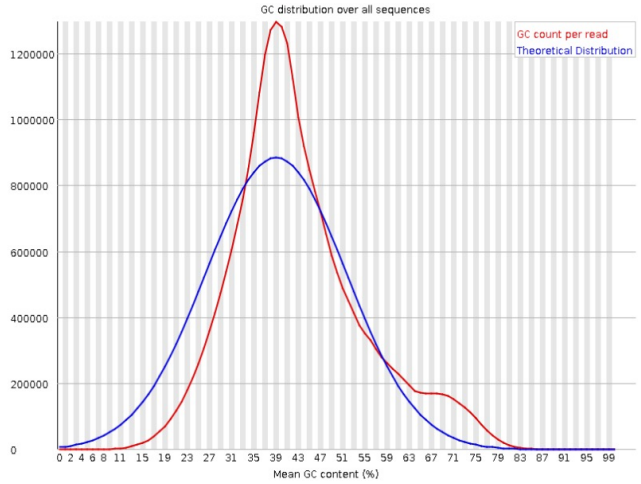
Distribution of average GC content

✗



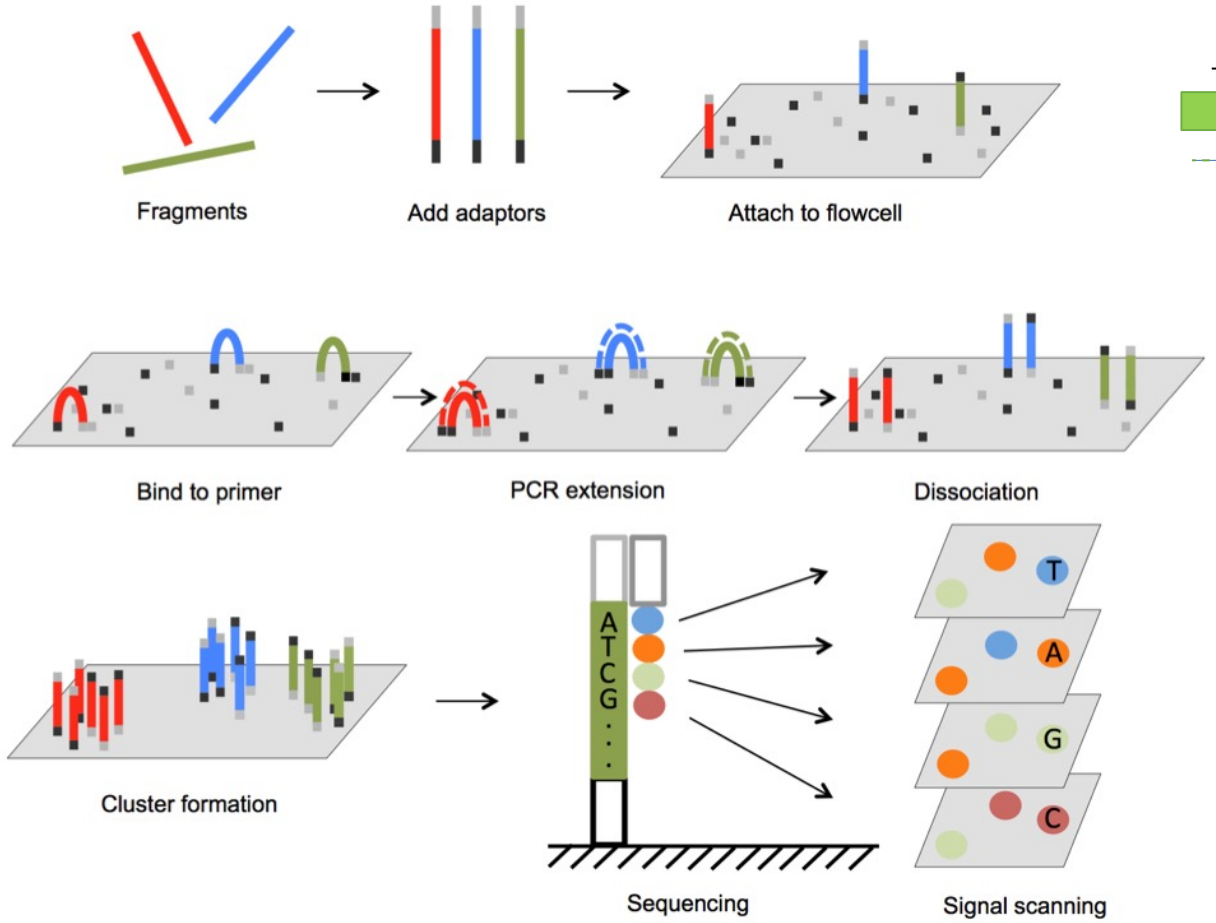
- **Normal distribution**
- **Fit with expected GC (organism dependent)**

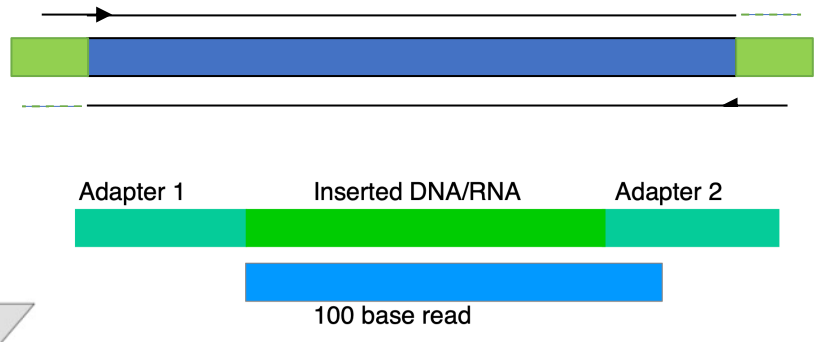
✓



- **Bi-modal distribution**
- **Large tail towards high GC**

Adapter contamination

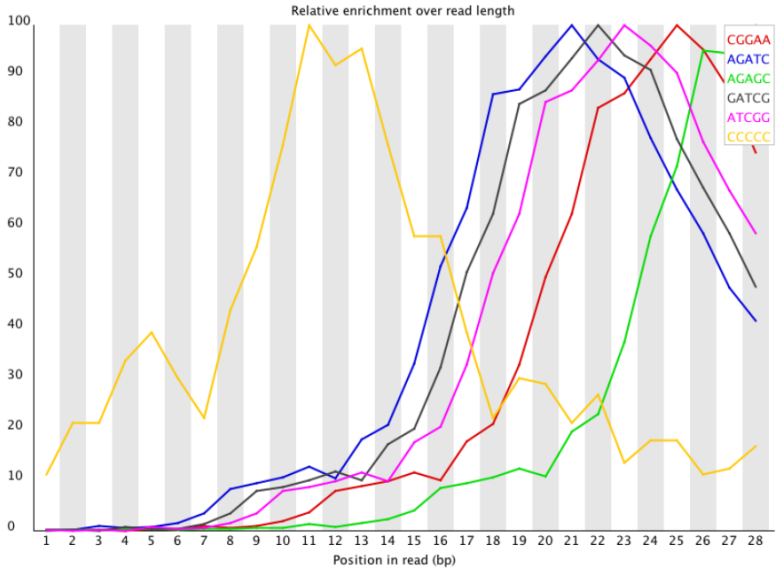




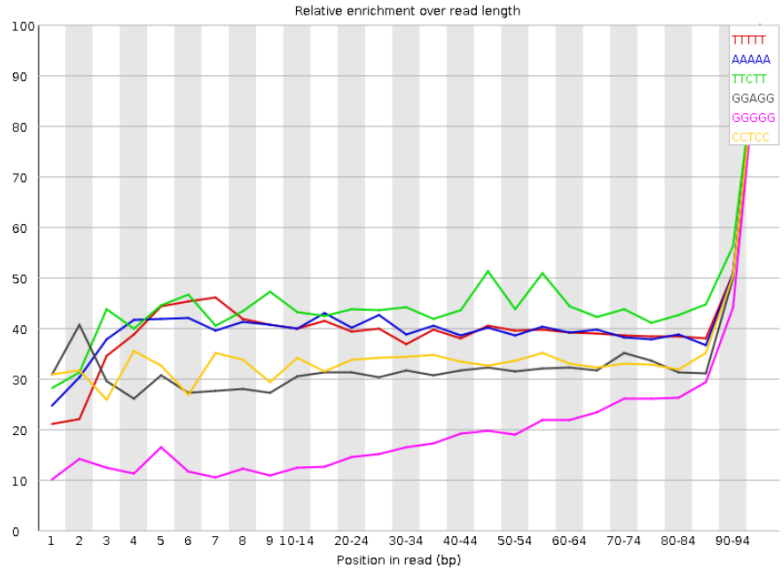
1) If the inserted DNA/RNA fragment is too short, the read will contain part of the adapter

2) Adapter trimming can be challenging if
- if the insert is 90 – 100 bases
- if the read has many sequencing errors

Overrepresented Kmers - Adapter contamination



- **Progressive patterns on 5‘ or 3’ end:**
- **partial adaptor contamination**

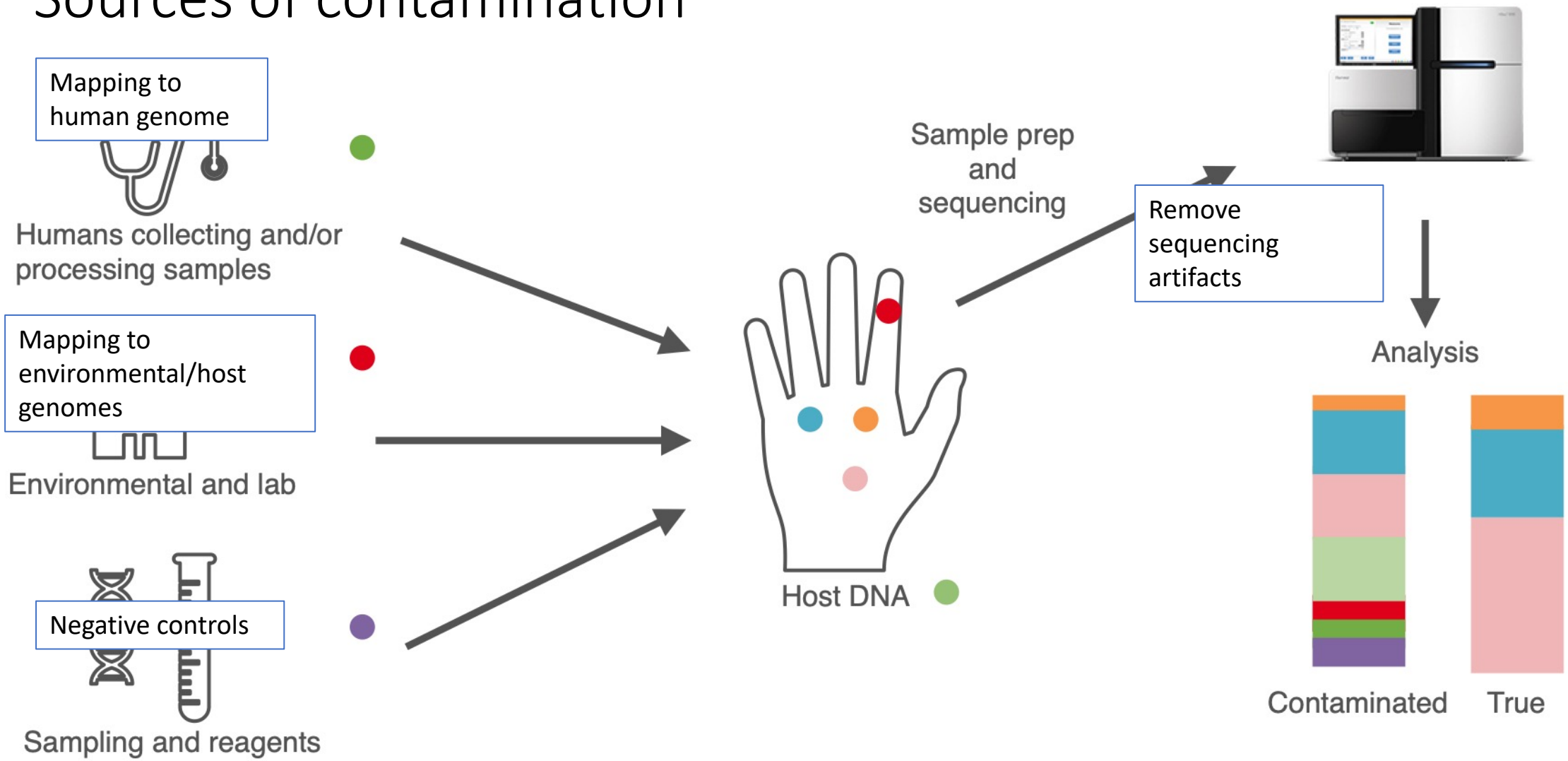


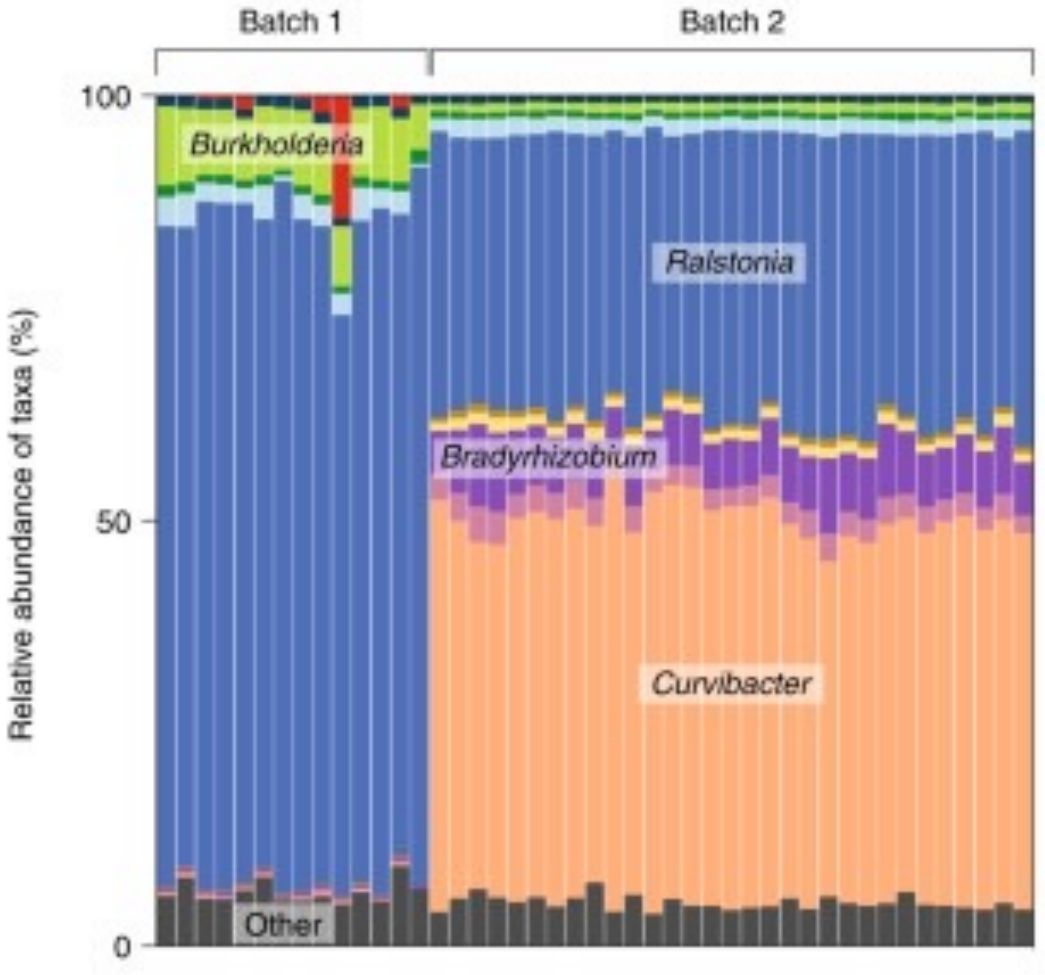
- **Homopolymer runs at 3‘**
- **Poor sequence stretches**

Other possible causes:

- Biologically significant (i.e. highly expressed transcripts)
- Less diverse library (low complexity)

Sources of contamination





de Goffau, M.C., Lager, S., Salter, S.J. *et al.* Recognizing the reagent microbiome. *Nat Microbiol* **3,** 851–853 (2018). https://doi.org/10.1038/s41564-018-0202-y

Prevention and filtering

- Sterile sampling environment
- Negative controls
- What to do with a negative control?
  - Check the general microbial content
  - Compare with the content of the sample
  - Is there anything in the sample/control that doesn't make sense?
- Filtering: e.g. reads mapping to human genome or reads mapping to PhiX, reagent contaminants

Sequencing Artifacts – calibration/controls

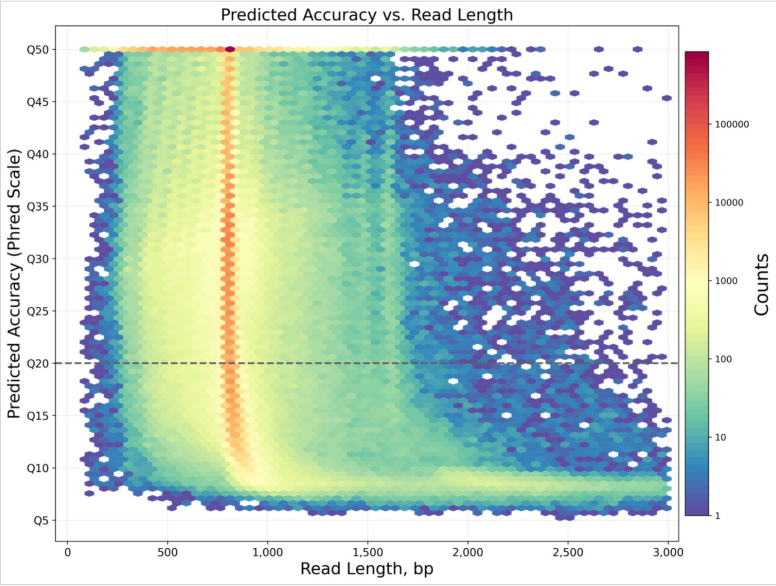
PhiX

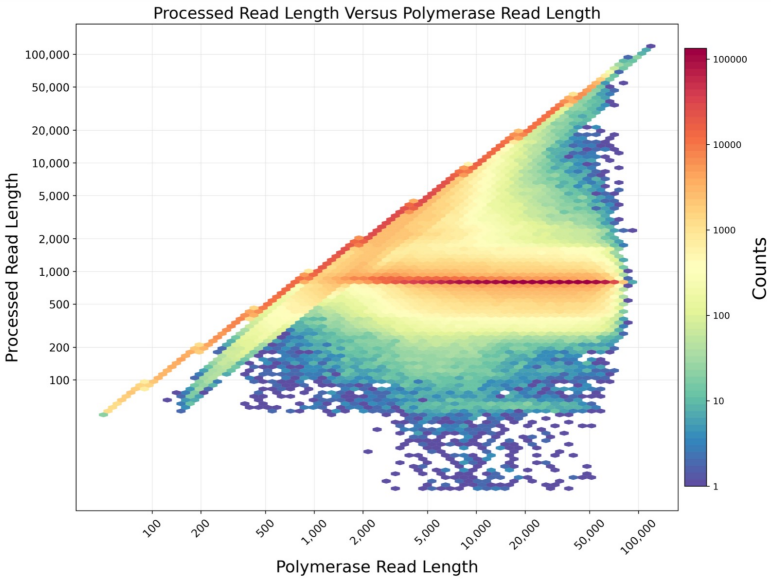
- Bacteriophage with a small genome.
- Used as a control during sequencing.
- Normally used a low concentrations.
- Concentration can be increased for low base diversity samples.
- Can appear as contaminants in assemblies.

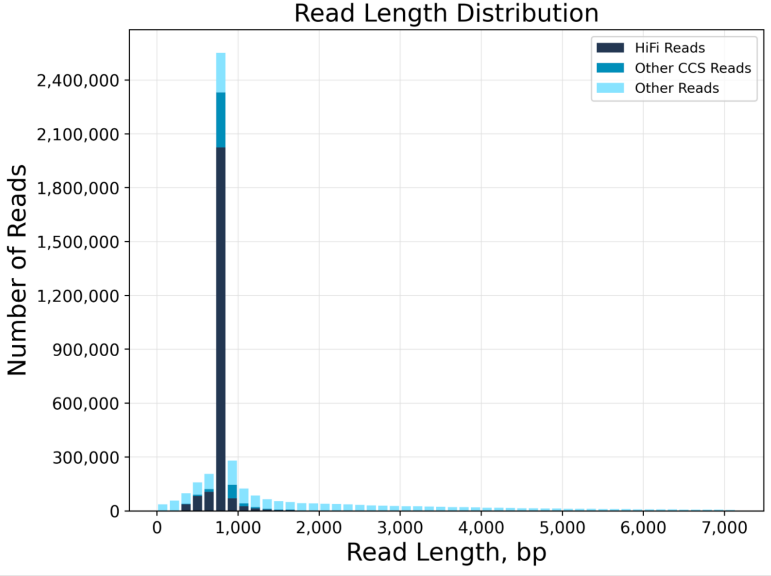
Mukherjee et al. Standards in Genomic Sciences 2015, 10:18
http://www.standardsingenomics.com/content/10/1/18

PacBio Run metrics

	Sa...	Run Sett...				Productivity (%)			Reads				Control
									HiFi Reads				
W...	Name	Movie Time...	Status	Total Bases (Gb)	Unique M...	P0	P1	P2	≥Q20 Reads	Yield	Mean Length	Median QV	Poly RL Mean (b...
A01	p2321...	8	Comple...	81.11	12.54	3.5	60.5	36.1	2379432	1.88 Gb	790	Q42	16111
B01	p2681...	30	Comple...	408.84	81.20	33.7	64.8	1.6	1883124	25.90 Gb	13752	Q33	68721



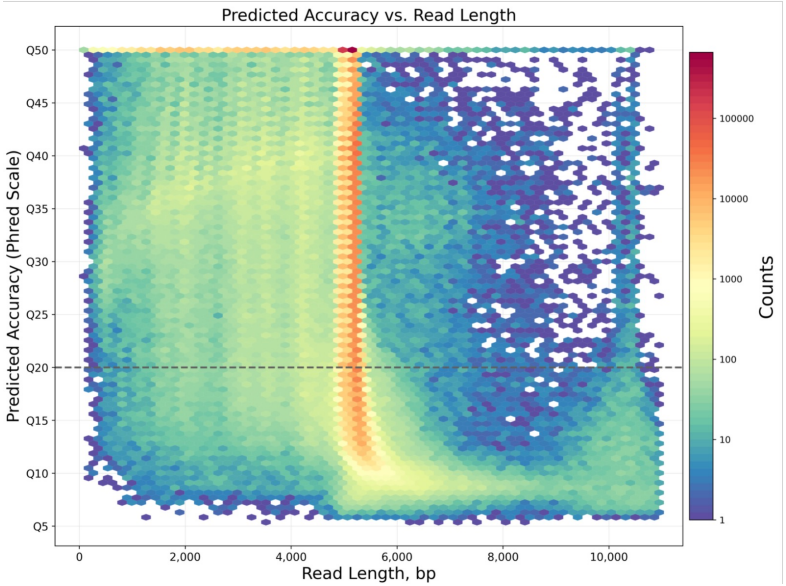


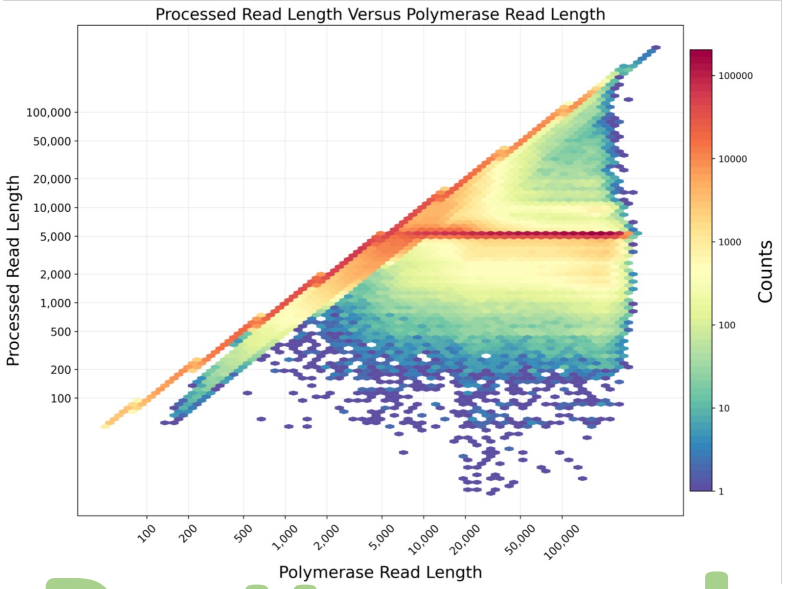


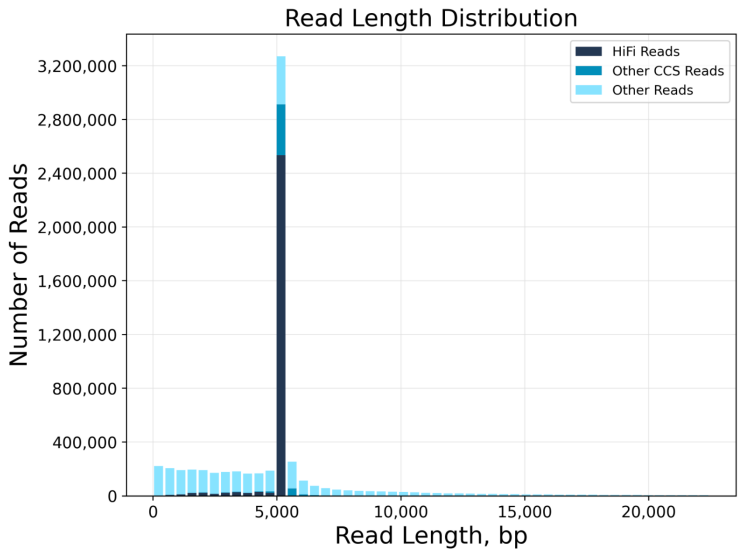
Not so good

PacBio Run Metrics

	S...	Run Setti...				Productivity (%)			Reads				Control
									HiFi Reads				
W...	N...	Movie Time ...	Status	Total Bases (Gb)	Unique ...	P0	P1	P2	≥Q20 Reads	Yield	Mean Length	Median QV	Poly RL Mean (bp)
B01	p31...	30	Comple...	377.45	48.54	13.7	83.1	3.2	2747812	13.7...	5011	Q41	57083
C01	p26...	30	Comple...	500.51	50.37	47.1	51.6	1.3	2362379	19.3...	8192	Q43	51072
D01	p26...	30	Comple...	273.92	29.07	72.5	26.8	0.7	1220602	10.5...	8677	Q43	51440
E01	p24...	30	Comple...	0.30	0.22	99.7	0.2	0.1	340	3.13 ...	9205	Q43	74238



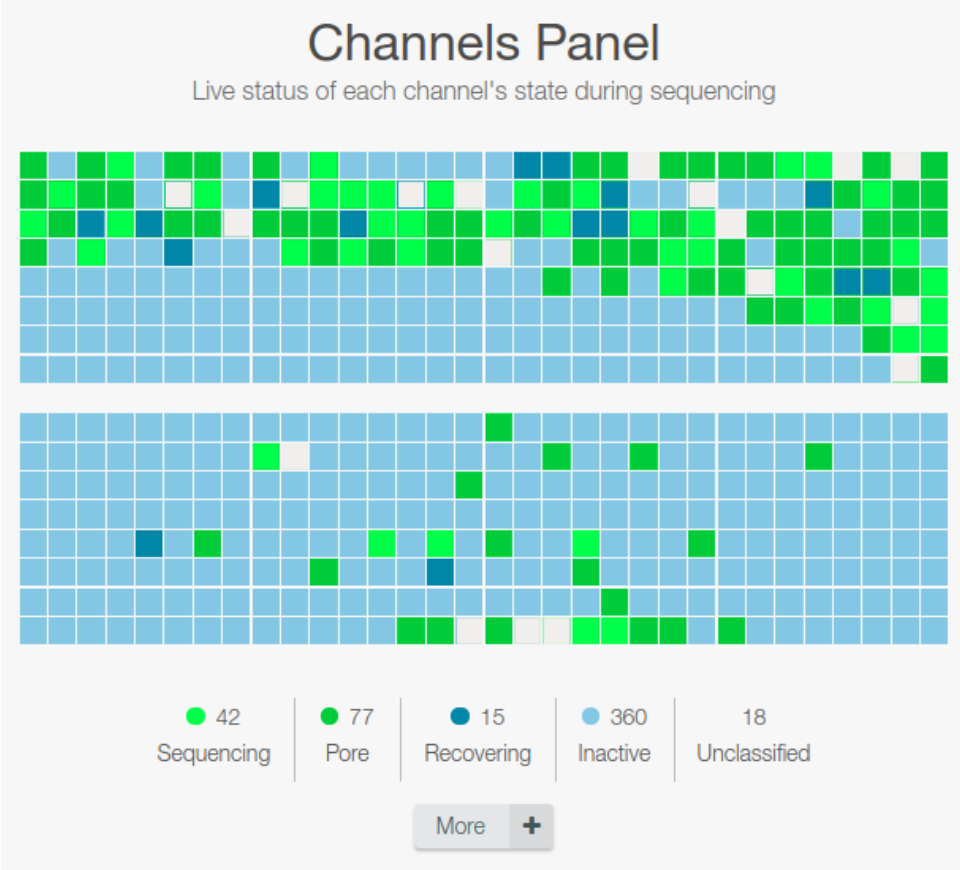




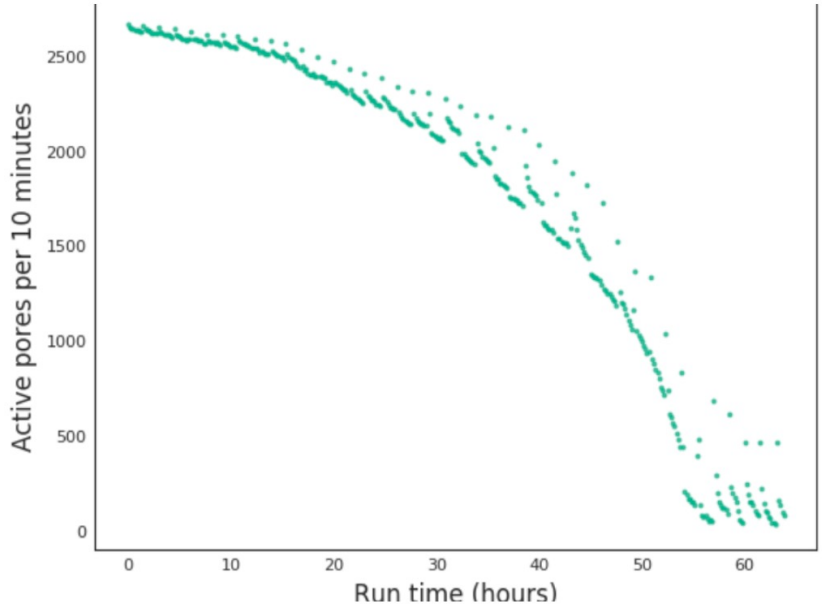
Pretty good

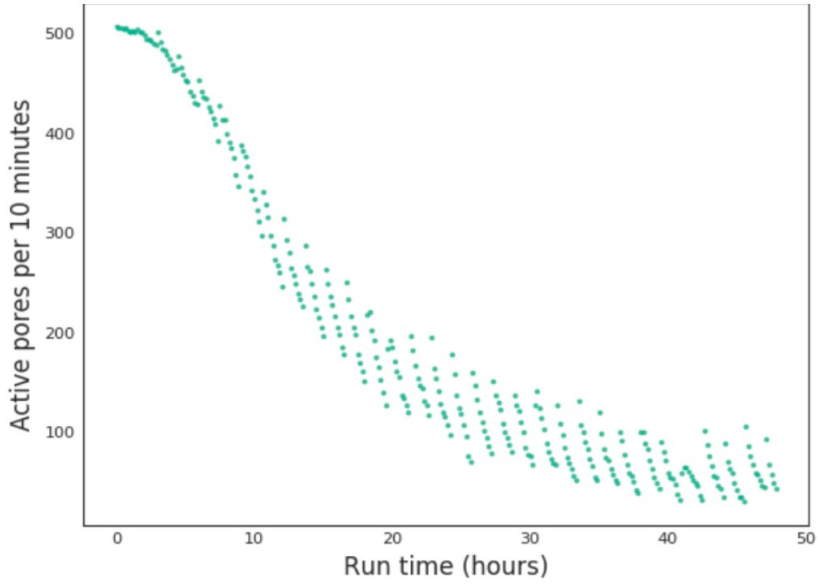
ONT Run metrics

- Too few active channels
  - Flow cell defect
  - loading error etc
  - Lower yield



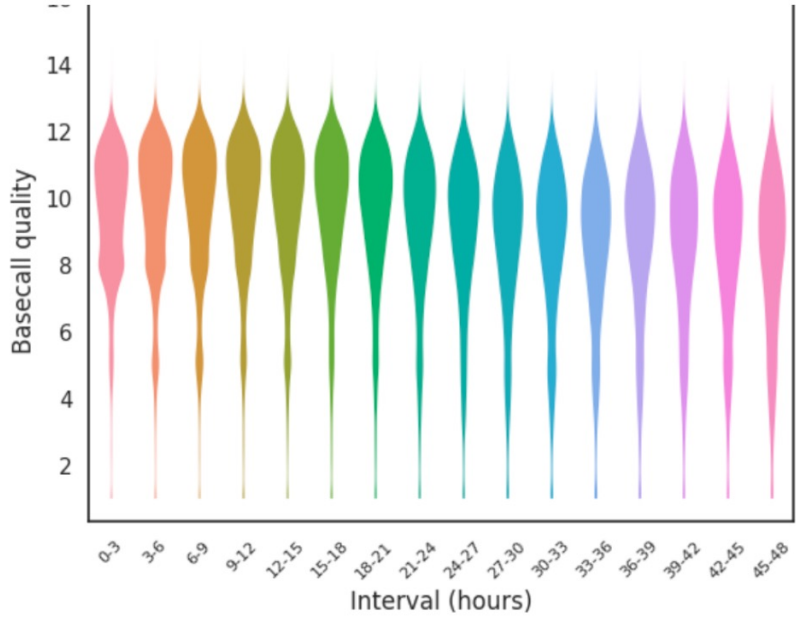
ONT Run metrics

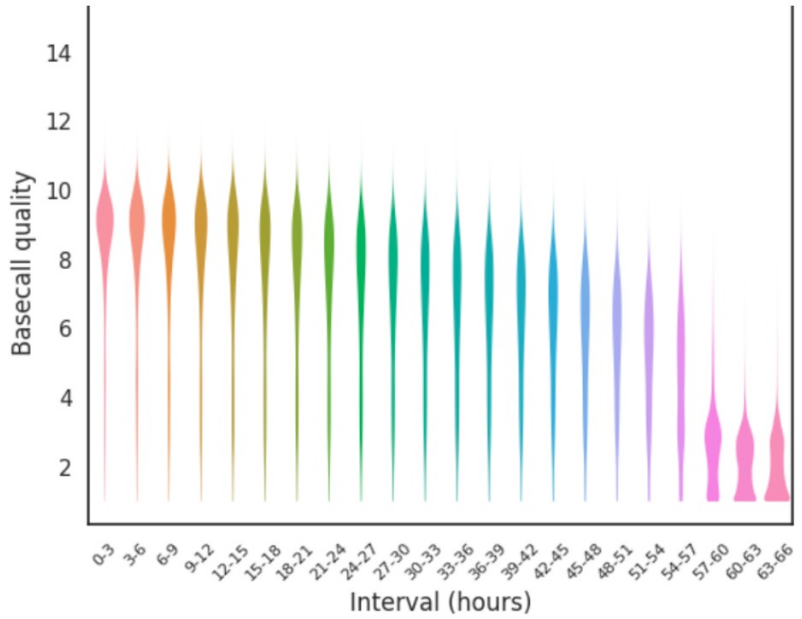




- Decreasing active pores with time is expected
- Rapid decrease: library contamination

ONT Run metrics





- Quality should not decrease over time until about 50 hrs

Tools with which to perform quality control

Do it yourself tools:

- FastQC/MultiQC - quality control, visualisation of metrics from Illumina reads
- Fastqscreen – contamination assessment from Illumina reads
- Kraken – contamination assessment from raw reads
- Blast (command line) + Megan - contamination from raw reads
- Trimmomatic – trimming of adapter sequences
- Prinseq – removal of duplicates
- Picard tools – quality control, filtering options

Tools that do it all for you:

- QIIME2, MG-RAST
- Kneaddata - https://huttenhower.sph.harvard.edu/kneaddata/

Especially for microbiome experiments, quality control with fastqc, trimming of adapters with Trimmomatic, principled in silico separation of bacterial reads from these “contaminant” reads

- Snakemake and nextflow pipelines: e.g.
  - Metagenome atlas (https://metagenome-atlas.readthedocs.io/en/latest/),
  - MAG (https://nf-co.re/mag)

Summary – Data preprocessing

a. Trimming
   - By quality (soft trimming) - fast
   - By length (hard trimming) - fast
   - By adapter – most time consuming

b. Filtering
   - By quality (average, percentage etc within a scanning window)
   - By length (kmer length for mapping and assembly)
   - By sequence content (N, GC)
   - By duplication level
   - Based on similarity to reference genomes or contamination databases

c. Chimera search and removal

d. Pacbio-specific
   - Apply circular consensus to eliminate most of the errors