Metagenomics:

Quality Control, Decontamination, and Metagenome Assembly

STAMPS – Day 4

July 22, 2024

Michael Nute & Todd Treangen

Agenda

- Quality Control on Sequencing Output
  - What kind of things can go wrong?
  - Identifying problems with FastQC
- Adapter Readthrough
  - What it is
  - Why it’s bad
  - How to fix
- Contamination
  - What if there are no negative controls?
- Metagenome Assembly
  - De Novo (MEGAHIT & metaSPAdes)
  - Reference Guided (Metacompass)
- Tutorial

Quality Control

High-throughput sequencing: what could possibly go wrong?

Quality Control: Why

- High-throughput DNA sequencing (HTS)
  - Highly complex process
  - Depends on much going right at the molecular level.
  - Many distinct failure modes, often occurring during prep.

- Left: quality scores degrade with the position in the read.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

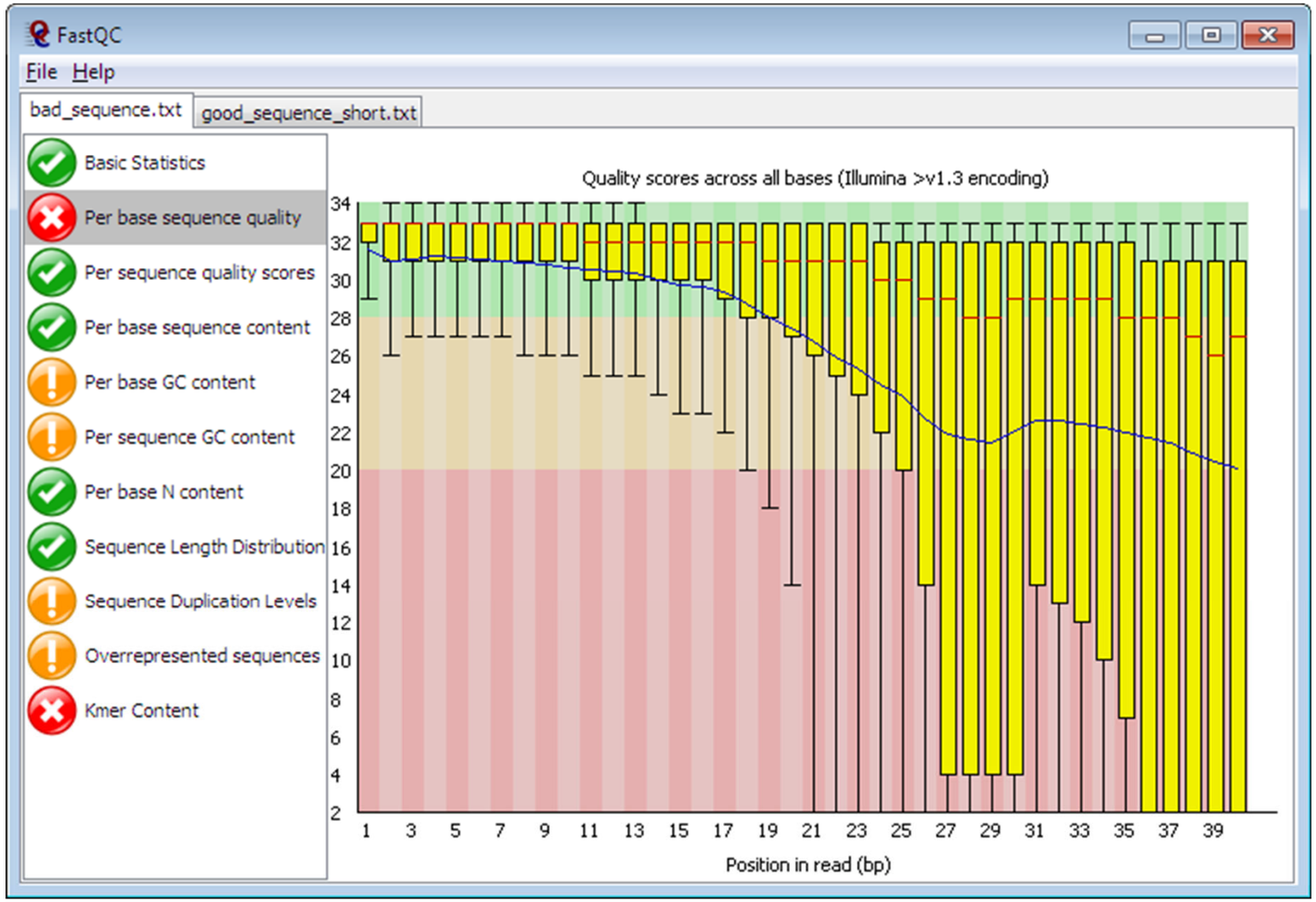


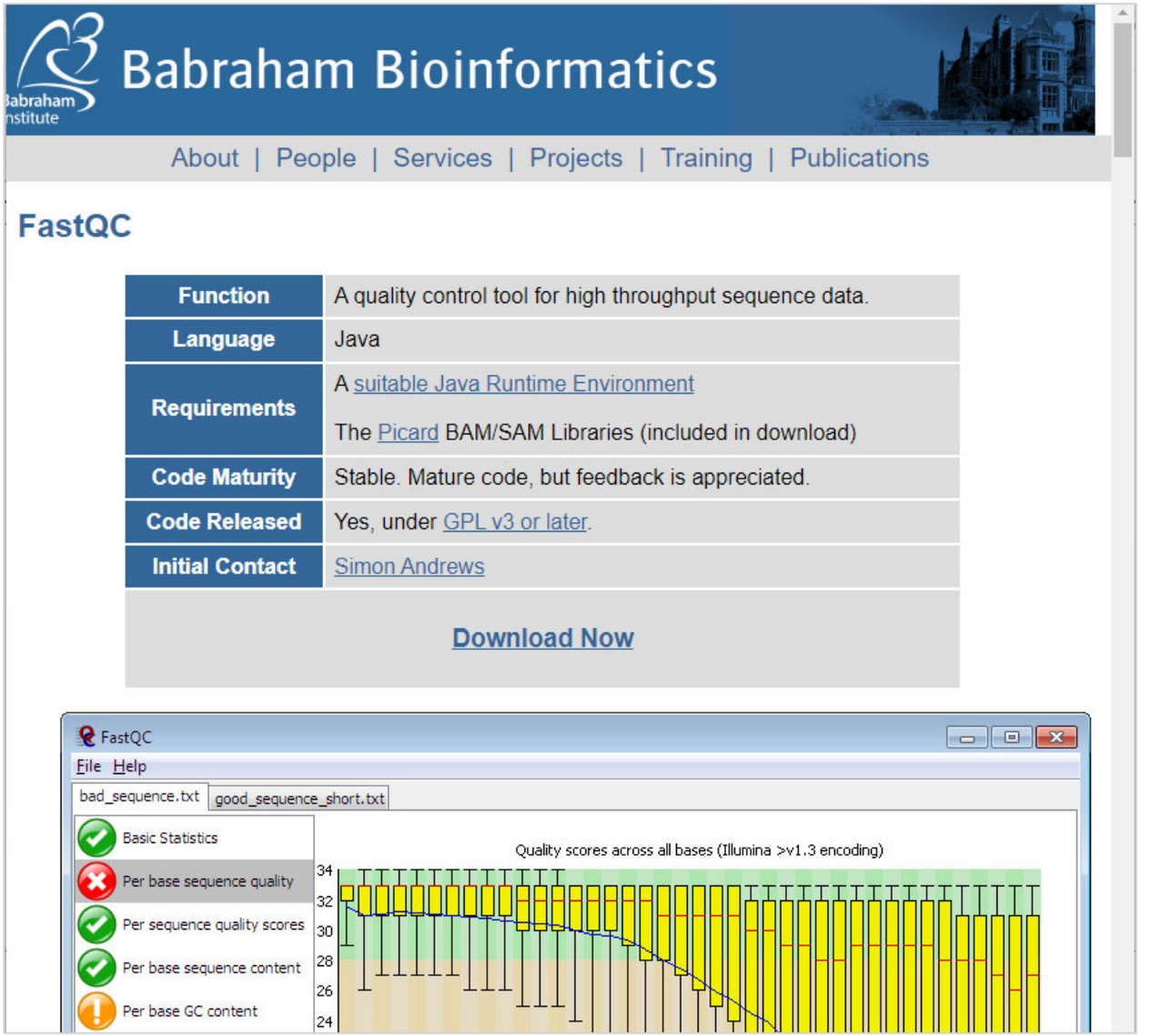
Image credit: FastQC documentation

FastQC: Our Surveillance Monitor

- Command line tool run on FastQ files
  - Produces graphical & text output
  - Only checks for issues, does *not* modify the reads.
  - Simple to run:
    - Few true parameters at command line
    - (mostly output formatting options)

- Example command:

```
fastqc -o <output_folder> <input_fastq.gz>
```



Above: FastQC Project Website (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

FastQC Reports: Per Sequence GC Content

From the FastQC Documentation:

*Warnings in this module **usually indicate a problem with the library**.*

- ***Sharp peaks** on an otherwise smooth distribution are normally the result of a **specific contaminant** (adapter dimers for example), which may well be picked up by the overrepresented sequences module.*
- ***Broader peaks** may represent **contamination with a different species.***

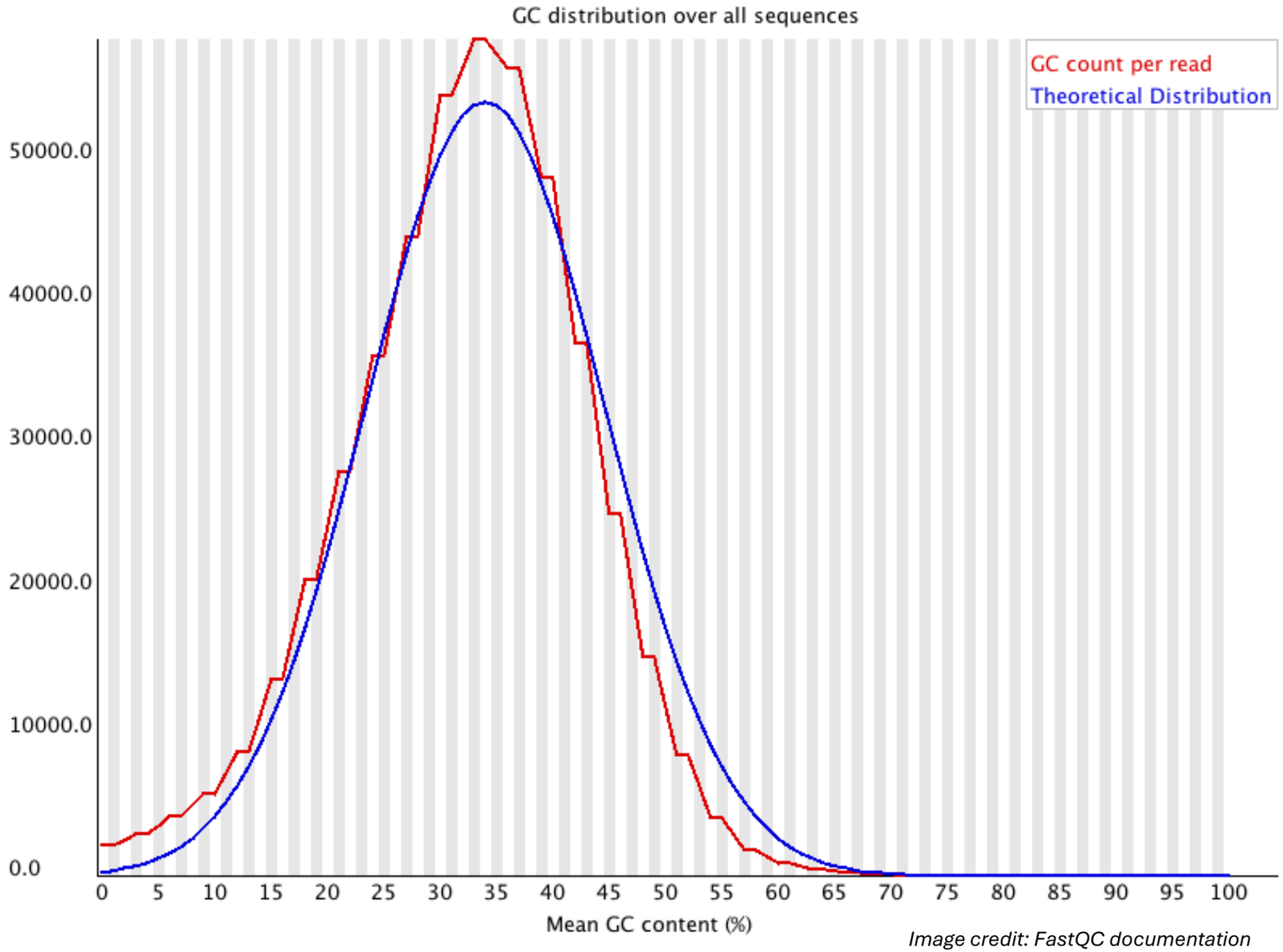
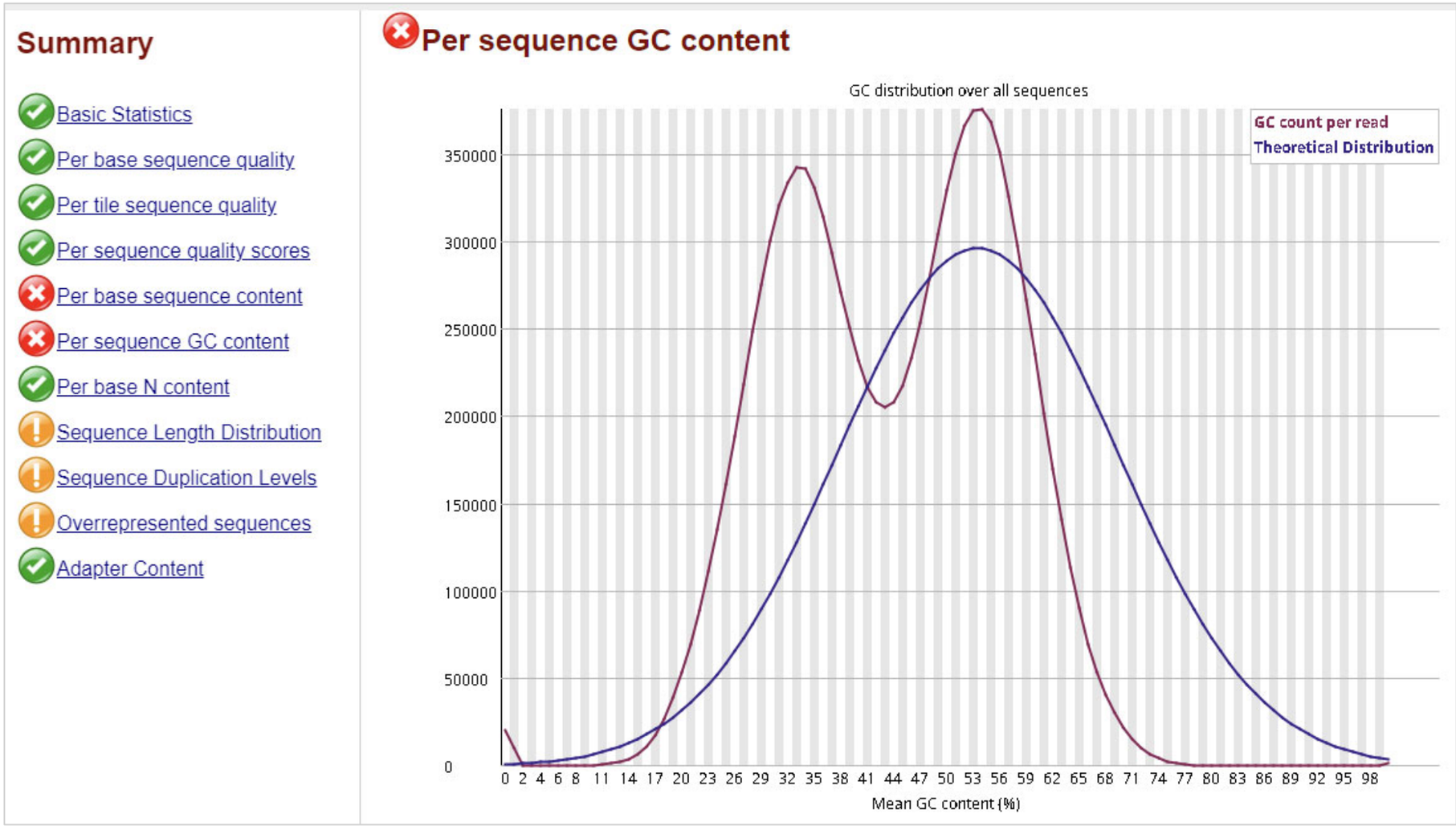


Image credit: FastQC documentation

FastQC Reports: Per Sequence GC Content

Actual report from a metagenomic dataset from a human gut microbiome sample[1].

- What's going on here?
- Why did this fail, and is it a problem?



[1]*BioProject: PRJEB15257, SRA RunID: ERR1600426 (forward reads)*

FastQC Reports: Per Base Sequence Content

From the FastQC Documentation:

It's worth noting that some types of library will always produce biased sequence composition, normally at the start of the read. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) and those which were fragmented using transposases inherit an intrinsic bias in the positions at which reads start...

Common reasons for warnings

1. Overrepresented sequences…
2. Biased fragmentation: Any library which is generated based on the ligation of random hexamers or through tagmentation should theoretically have good diversity through the sequence, but experience has shown that these libraries always have a selection bias in around the first 12bp of each run…
3. Biased composition libraries: Some libraries are inherently biased in their sequence composition. The most obvious example would be a library which has been treated with sodium bisulphite which will then have converted most of the cytosines to thymines…
4. Aggressive adapter trimming…

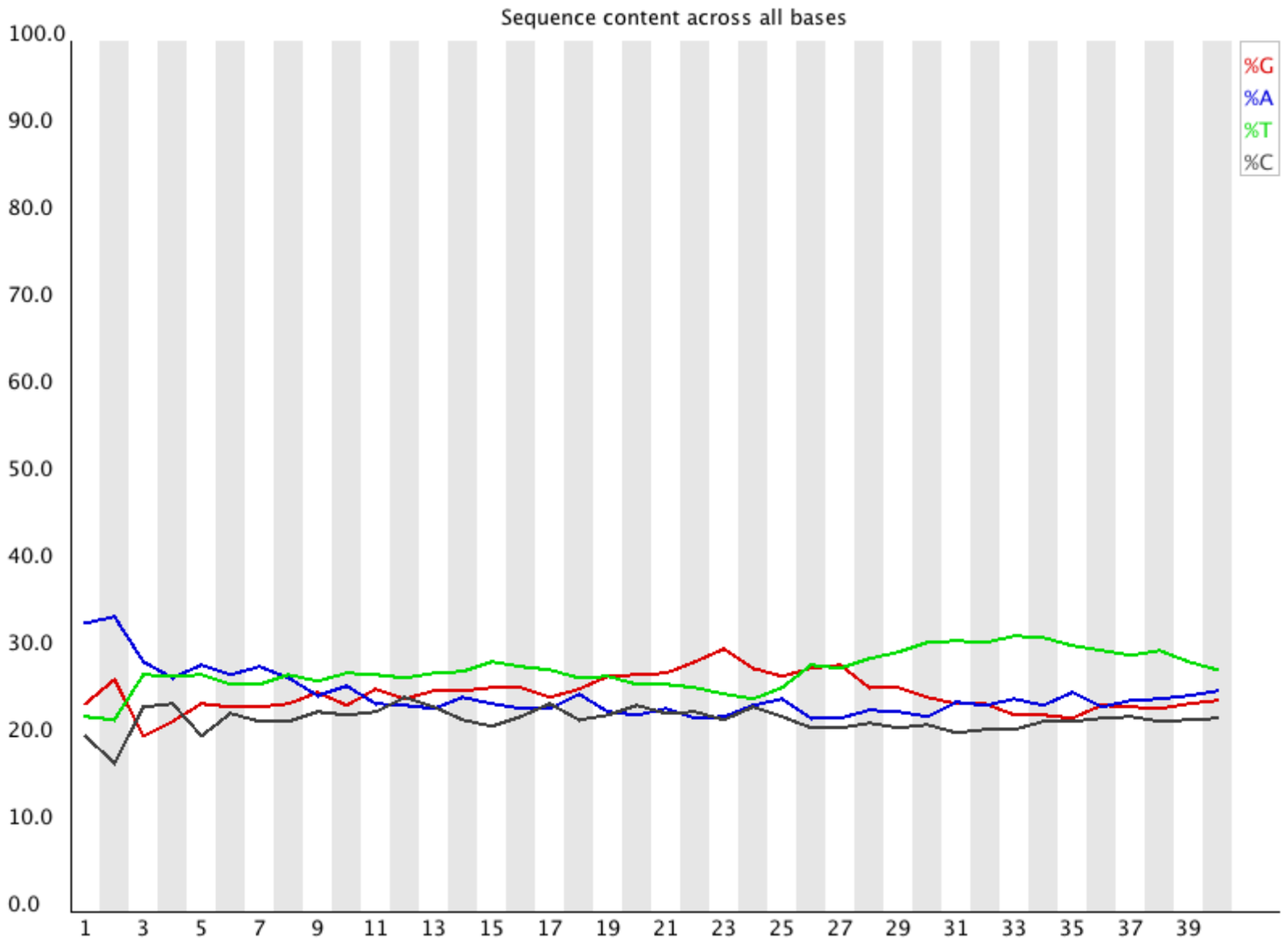


Image credit: FastQC documentation

FastQC Reports: Per Base Sequence Content

Actual report from a metagenomic dataset from a human gut microbiome sample[1].

- What's going on here?
- Why did this fail, and is it a problem?

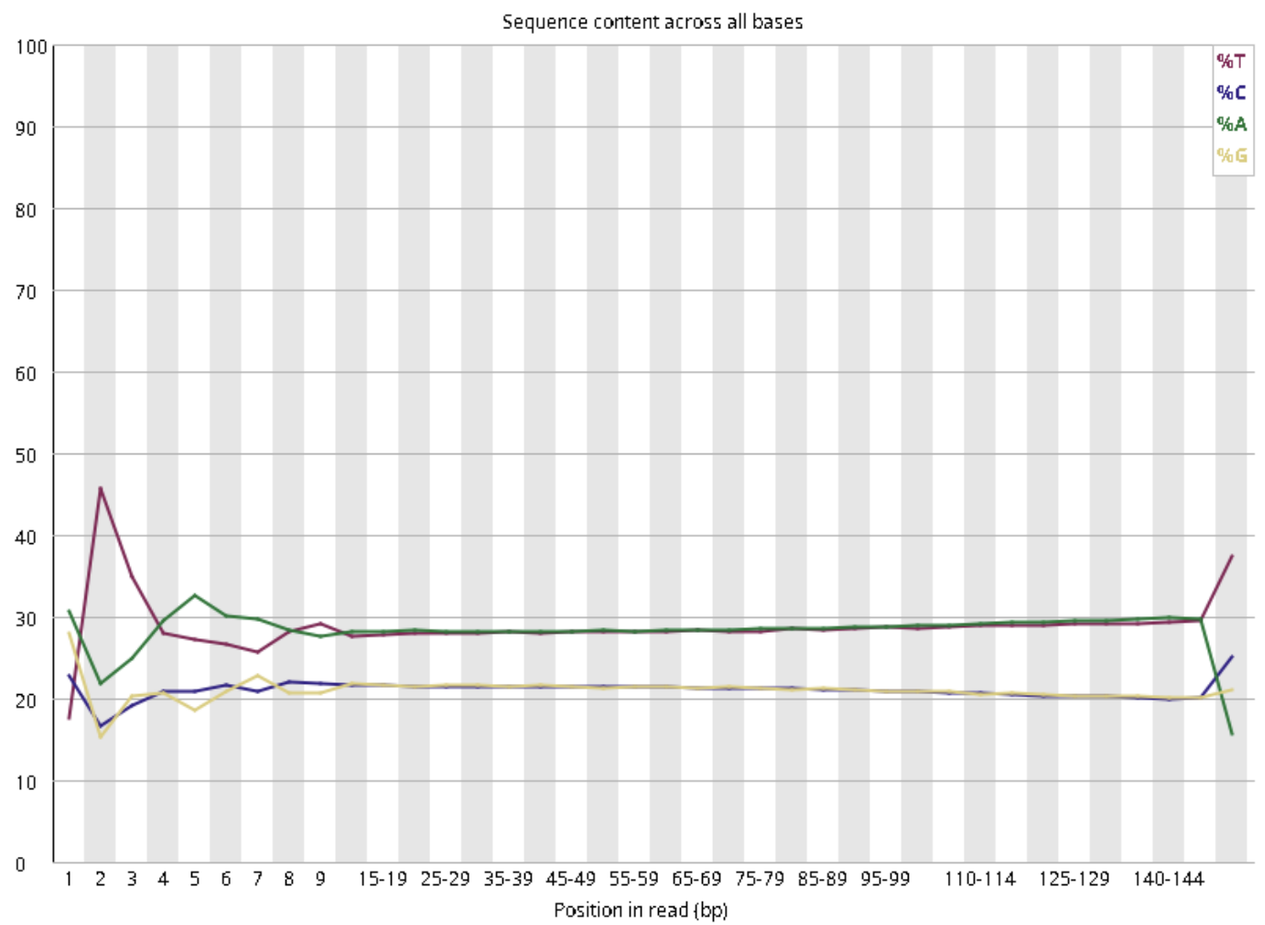
tQC Report

Thu 4 Jan 2024
ERR1600426_1.fastq.gz

ry

tatistics
e sequence quality
sequence quality
uence quality scores
e sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content

Per base sequence content



[1]*BioProject: PRJEB15257, SRA RunID: ERR1600426 (forward reads)*

FastQC Reports: Overrepresented Sequences

Actual report from a metagenomic dataset from a human gut microbiome sample[2].

- Often overrepresented sequences will include some type of adapter, although the adapter content report itself may not report a failure…
- This sample has an issue with adapter readthrough that can be corrected with adapter trimming

FastQC Report

Thu 4 Jan 2024
ERR3726305_1.fastq.gz

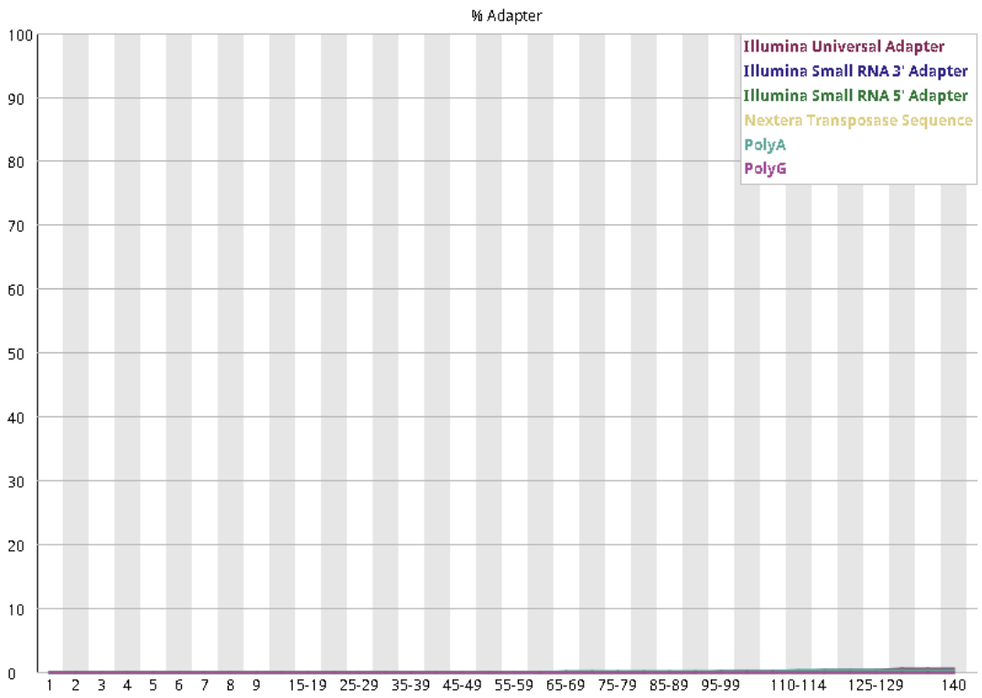
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCTCGTAT	2464829	14.554327100873918	TruSeq Adapter, Index 5 (97% over 37bp)
NATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCTCGTAT	50389	0.29753706577045946	TruSeq Adapter, Index 13 (97% over 35bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCCCGTAT	21420	0.12648085790159044	TruSeq Adapter, Index 5 (97% over 37bp)

Adapter Content



FastQC (version 0.12.1)

[2]*BioProject: PRJEB23147, SRA RunID: ERR3726305 (forward reads)*

FastQC Reports: Per Tile Sequence Quality

Actual report from a metagenomic dataset from a human gut microbiome sample[3].

QC Report

Thu 4 Jan 2024
ERR10149217_1.fastq.gz

Basic Statistics

From the FastQC Documentation:

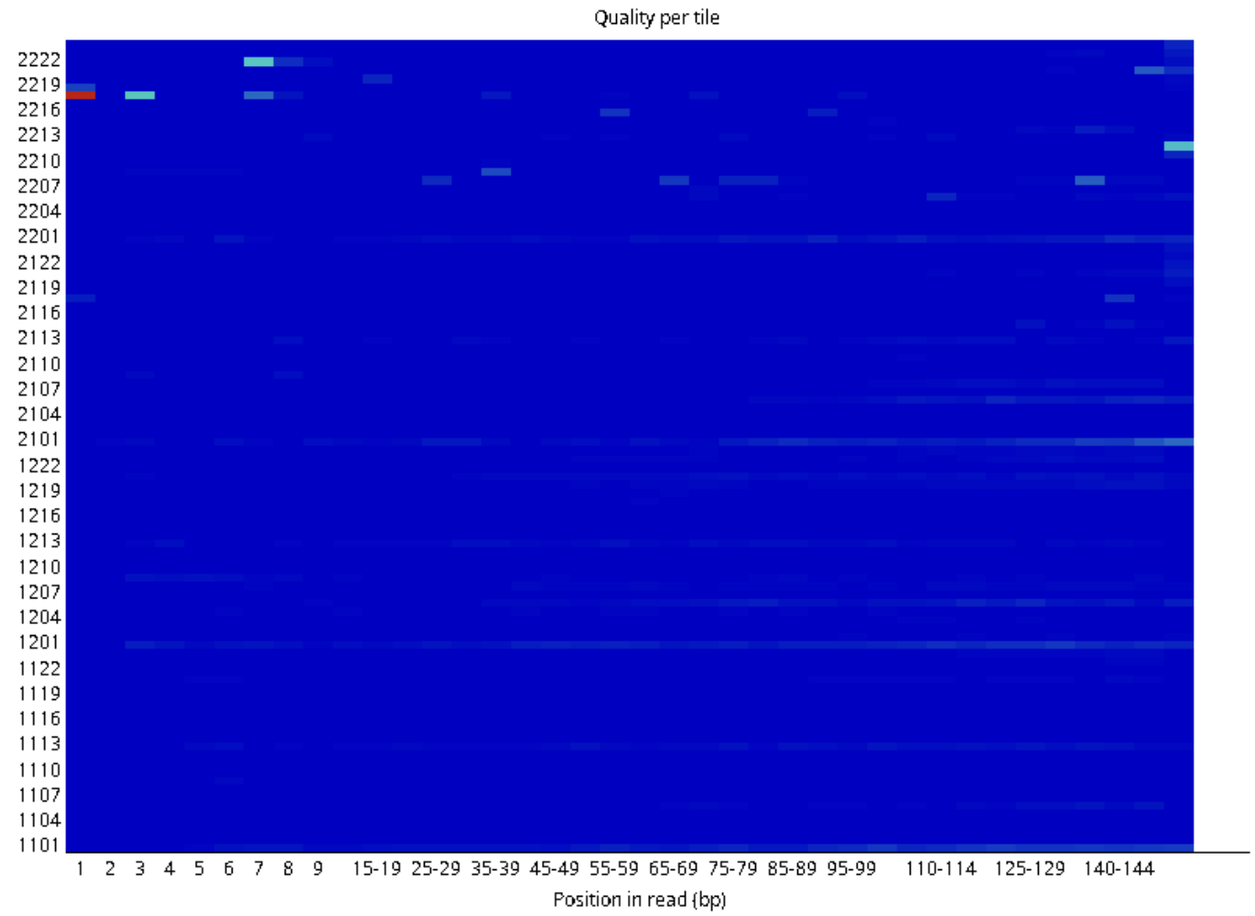
Reasons for seeing warnings or errors on this plot could be transient problems such as bubbles going through the flowcell, or they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane.

Warning: mean Phred score more than 2 less than the mean for that base across all tiles.

Failure: mean Phred score more than 5 less than the mean for that base across all tiles.

Common reasons for warnings
Whilst warnings in this module can be triggered by individual specific events we have also observed that greater variation in the phred scores attributed to tiles can also appear when a *flowcell is generally overloaded*. In this case events appear *all over the flowcell rather than being confined to a specific area or range of cycles*. We would generally ignore errors which mildly affected a small number of tiles for only 1 or 2 cycles, but would pursue larger effects which showed high deviation in scores, or which persisted for several cycles.

Per tile sequence quality



[3]*BioProject: PRJEB55713, SRA RunID: ERR10149217 (forward reads)*

FastQC Reports: Per Tile Sequence Quality (another one)

Actual report from a metagenomic dataset from a human gut microbiome sample[4].

Thu 4 Jan 2024
ERR3229898_1.fastq.gz

Basic Statistics

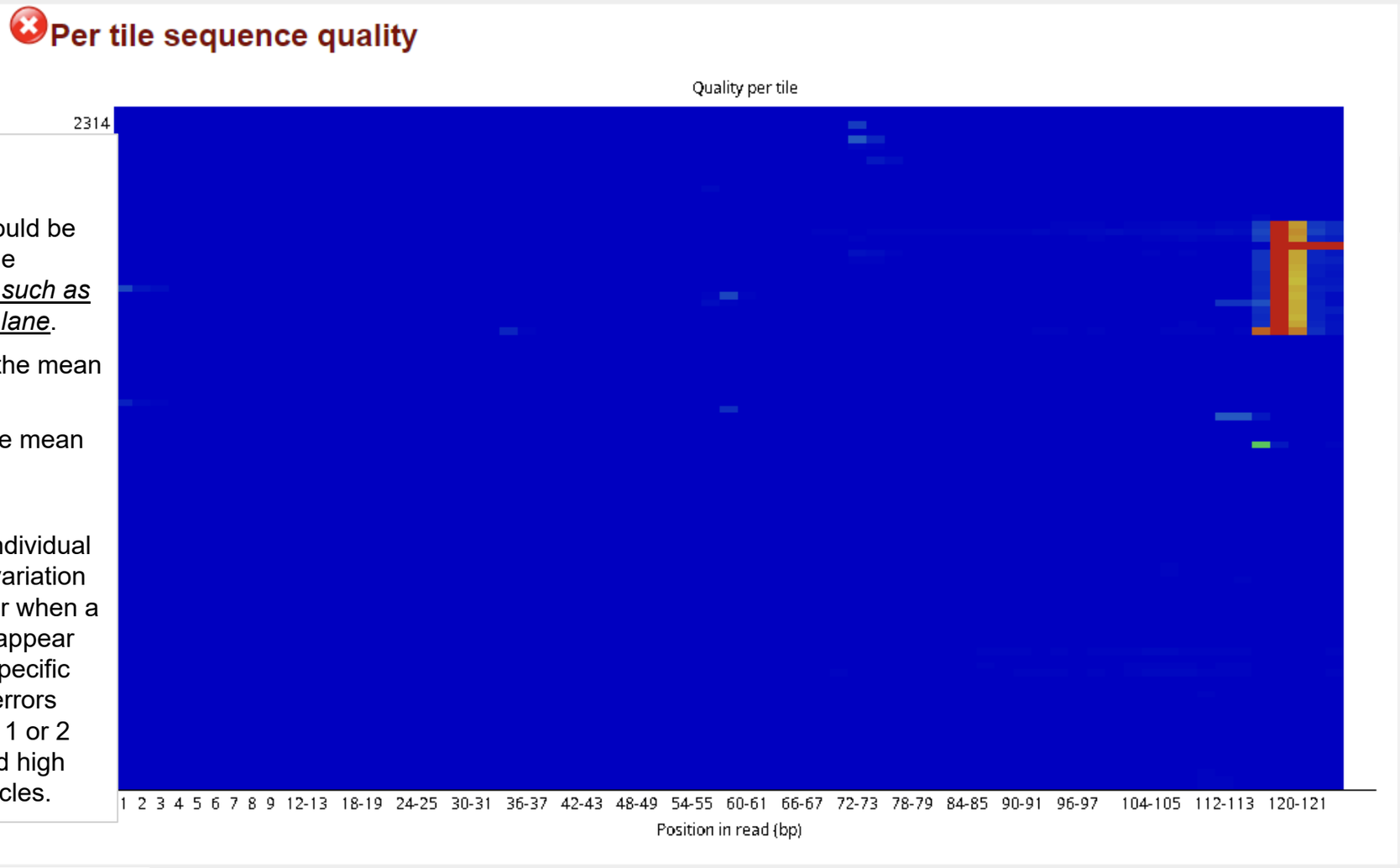
From the FastQC Documentation:

Reasons for seeing warnings or errors on this plot could be transient problems such as bubbles going through the flowcell, or *they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane*.

Warning: mean Phred score more than 2 less than the mean for that base across all tiles.

Failure: mean Phred score more than 5 less than the mean for that base across all tiles.

Common reasons for warnings
Whilst warnings in this module can be triggered by individual specific events we have also observed that greater variation in the phred scores attributed to tiles can also appear when a flowcell is generally overloaded. In this case events appear all over the flowcell rather than being confined to a specific area or range of cycles. We would generally ignore errors which mildly affected a small number of tiles for only 1 or 2 cycles, but would pursue larger effects which showed high deviation in scores, or which persisted for several cycles.



[4]*BioProject: PRJEB23010, SRA RunID: ERR3229898 (forward reads)*

FastQC Reports: Per Base N Content

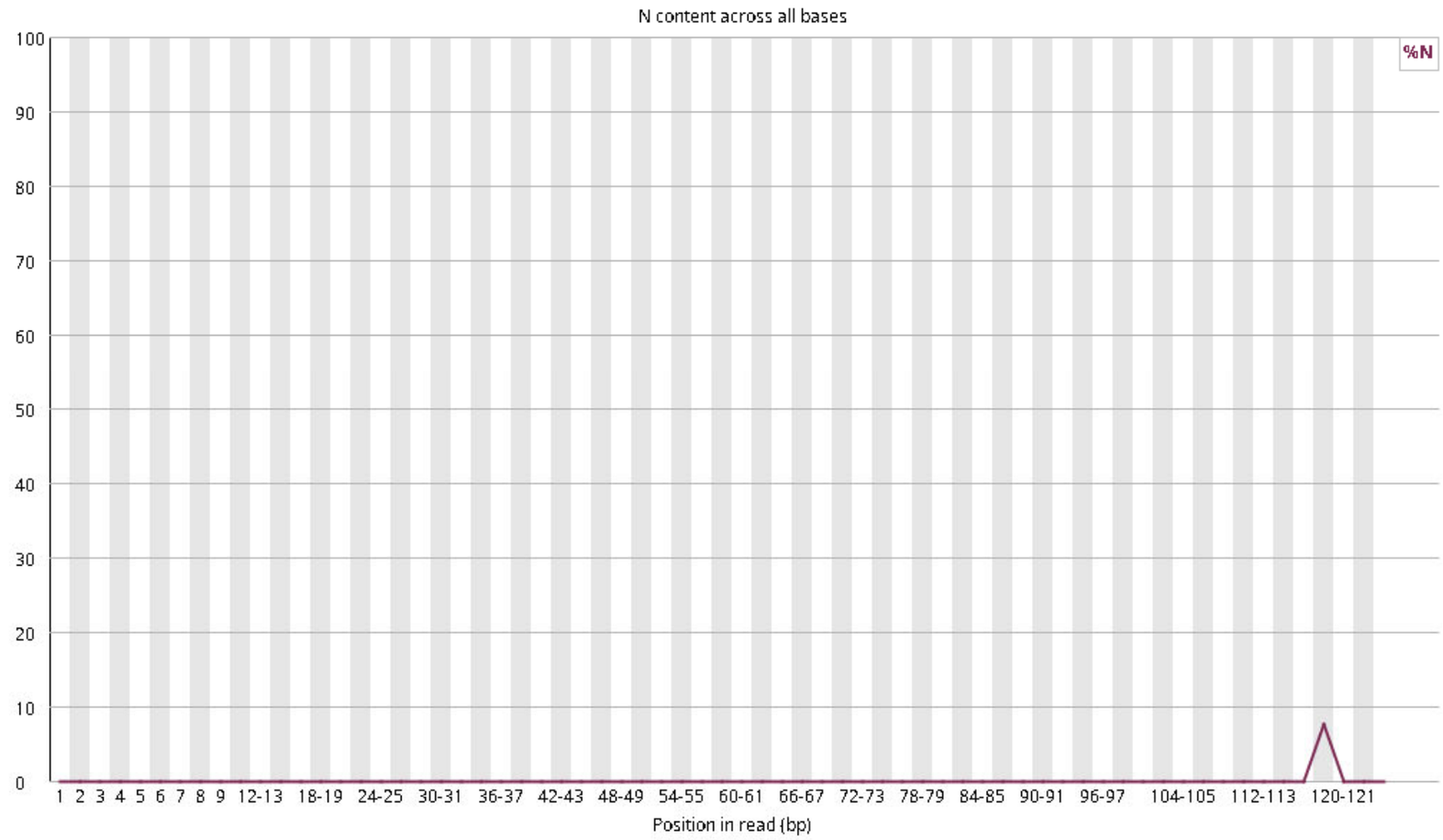
Actual report from a metagenomic dataset from a human gut microbiome sample[4].

Note: all sequences in this sample are 125bp long.

Thu 4 Jan 2024
ERR3229898_1.fastq.gz

- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base N content



[4]*BioProject: PRJEB23010, SRA RunID: ERR3229898 (forward reads)*

Adapter Trimming & Other QC Remediations

Trimming and filtering out low-quality reads

What is adapter readthrough?

From Illumina Website:

...the sequencing primer anneals to the adapter, immediately upstream of the DNA insert (in gray). ***Because the sequencing starts at the first base of the DNA insert*** *in Reads 1 and 2, the adapter is not sequenced at the start of the read. However, if the sequencing extends beyond the length of the DNA insert, and into the adapter on the opposite end of the library fragment, that adapter sequence will be found on the 3' end of the read. Therefore, reads require adapter trimming only on their 3' ends.*

NOTE: Metagenome assembly in particular is HIGHLY sensitive to adapter readthrough if still present in reads.

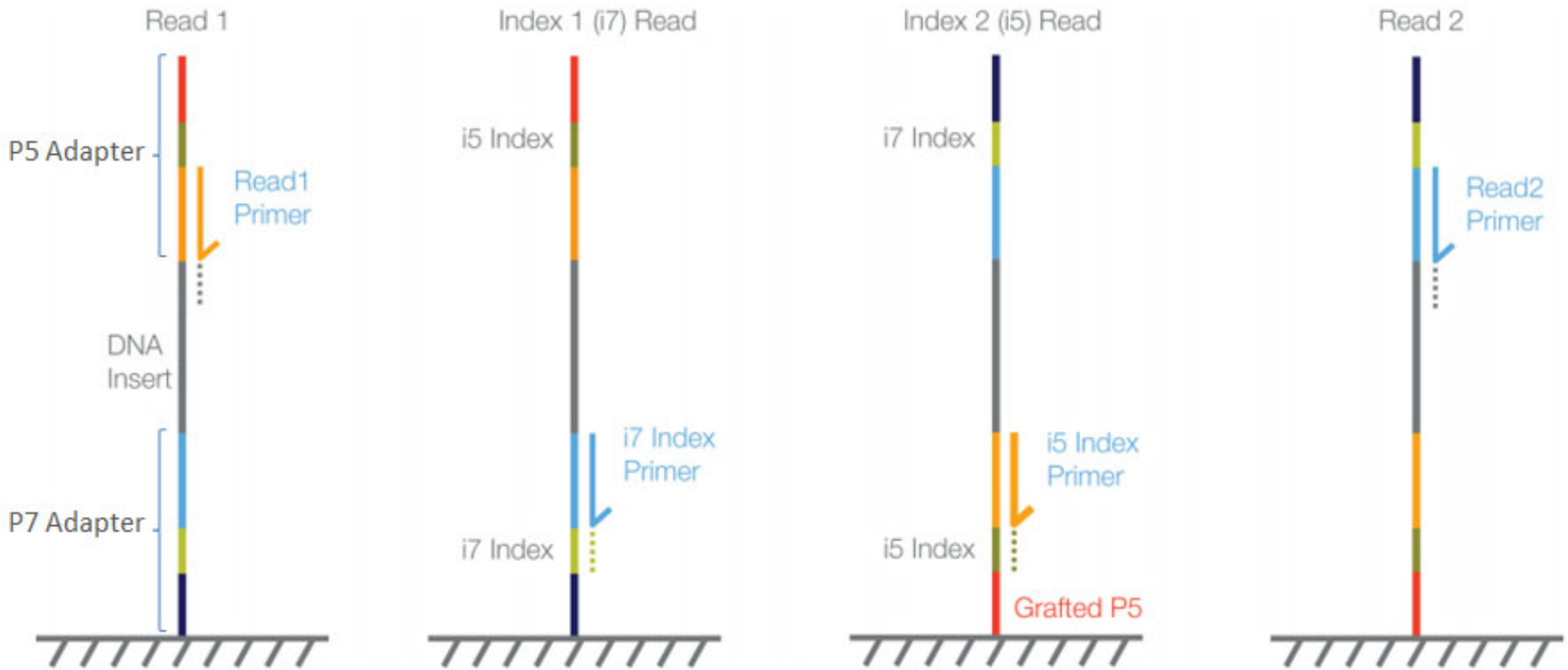


Image credit: Illumina website (link)

Adapter Trimming & Quality Filtering Methods

- Trimmomatic
  - http://www.usadellab.org/cms/?page=trimmomatic
  - Performs several steps:
    - Finds & trims specified Illumina-specific adapter sequences
    - Trims or removes reads based on either length or base-quality specifications
- Cutadapt
  - https://cutadapt.readthedocs.io/en/stable/index.html
  - Adapter-trimming only (but error tolerant)
  - Must provide adapter sequences.
- Bbduk (part of BBtools, from JGI)
  - https://jgi.doe.gov/data-and-tools/software-tools/bbtools/
  - Adapter-trimming, quality/length trimming & filtering, degenerate seq. filtering.
  - Contains own database of adapter sequences (e.g. if adapters not known, as in SRA)

BBTools

- Software suite developed & released by JGI
- Many different tools for manipulating reads and dealing with read QC issues.
- BBDuk:
  - Duk = “Deconamination using k-mers”
  - Trimming/filtering based on quality and/or adapter/contaminant identification.

BBduk for Adapter Trimming: Example

Example command (annotated):

Paths to input & output read file paths.

```
bbduk.sh in1=<fwd.fq.gz> in2=<rev.fq.gz> out1=<out_fwd.fq.gz> out2=<out_rev.fq.gz>
  ref=<bbduk_adapter_db> ktrim=r k=23 mink=11 hdist=1 tpe tbo
```

path to adapter database (packaged with bbduk)

Kmer size to use (lower = more sensitive). Must be no longer than adapter length.

# of mismatches to allow

"r" specifies to trim reads on the right (3' end) only (i.e. for adapters)

Allows shorter kmers to be used at the end of a read, but no shoter than 11.

Recommended additional arguments for adapter trimming:
***tpe**: trims both reads to same length*
***tbo**: also trims based on pair-overlap detection using BBmerge*

Example BBDuk results:

Revisiting the over-represented sequences example from earlier:

Thu 4 Jan 2024
ERR3726305_1.fastq.gz

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCTCGTAT	2464829	14.554327100873918	TruSeq Adapter, Index 5 (97% over 37bp)
NATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCTCGTAT	50389	0.29753706577045946	TruSeq Adapter, Index 13 (97% over 35bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGCTTAATCCCGTAT	21420	0.12648085790159044	TruSeq Adapter, Index 5 (97% over 37bp)

Total: 2,536,638 fwd reads (i.e. 5,073,276 total)

BBDuk summary:

processing time (s)	183.0
input readct	33,870,738
input basect (mb)	5,114.5
ktrimmed readct	5,923,398
ktrimmed basect (mb)	855.2
trimmed by overlap readct	93,486
trimmed by overlap basect (mb)	1.1
total removed readct	5,550,164
total removed basect (mb)	856.4
result readct	28,320,574
result basect (mb)	4,258.1

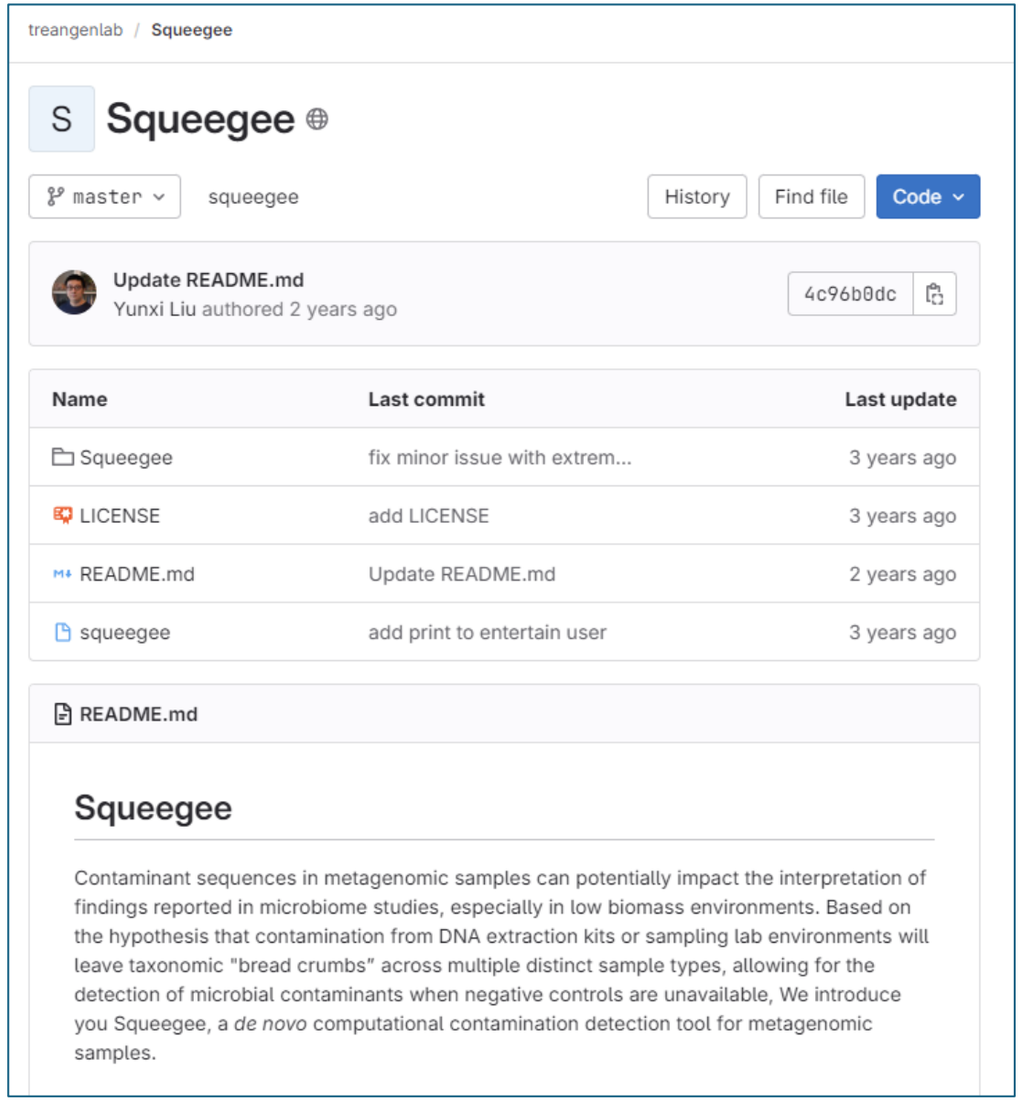
Likely includes all 5m overrepresented sequences found, plus others...

Quick Advertisement: Squeegee

Hey, speaking of quality control…

- Contaminant detection for metagenomic samples where…
  - No negative control is available, *but*…
  - …multiple sample types are.
- https://gitlab.com/treangenlab/squeegee

```
conda install -c bioconda squeegee
```



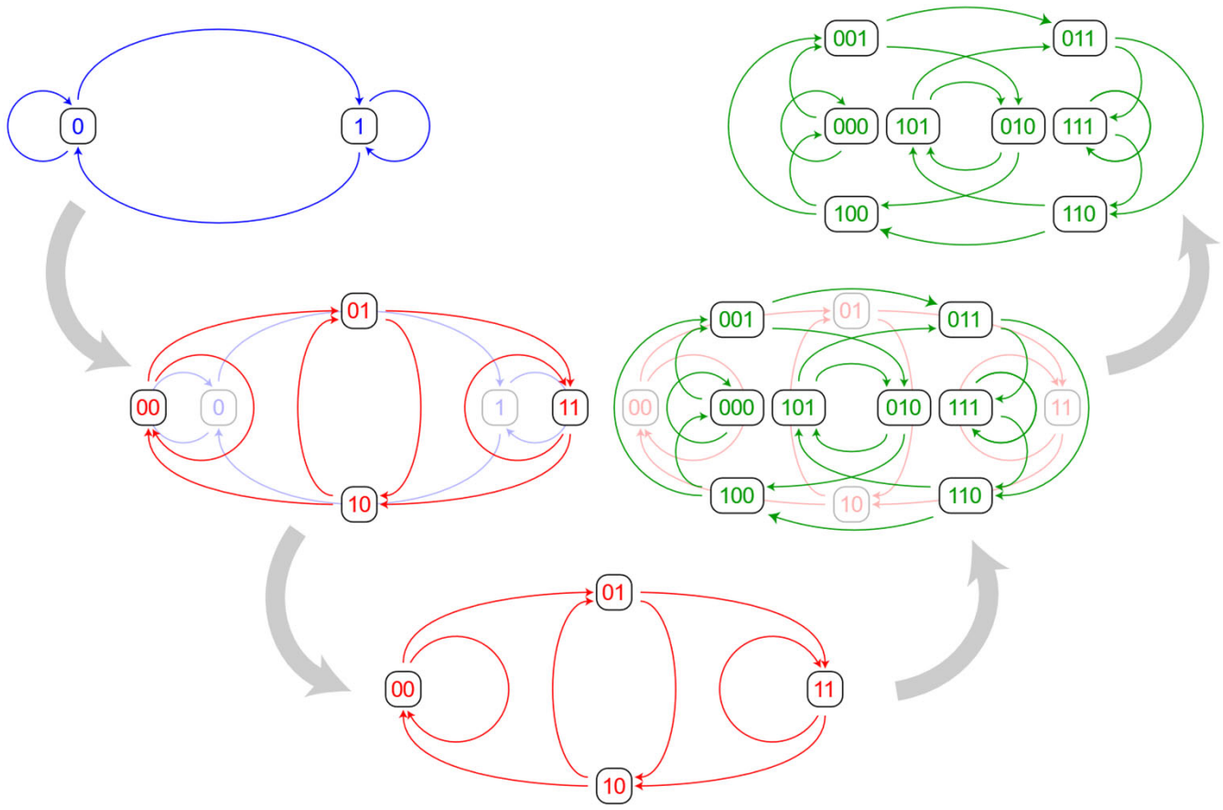
Metagenome Assembly

Concepts & Examples

Basics: De Bruijn Graphs

- Given a *set* of *sequences* of symbols…
  - *(like, for example, a set of DNA sequencing reads)*
- Each sequence is a node in the graph
- Node A is (directionally) connected to Node B if they overlap.
- Traveling this graph in order gives us a contiguous sequence.
  - *But the path is not necessarily unique…*

Bruin Graph



De Bruijn Graph

Image credits: Wikipedia (both)

Basics: De Bruijn Graphs

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph

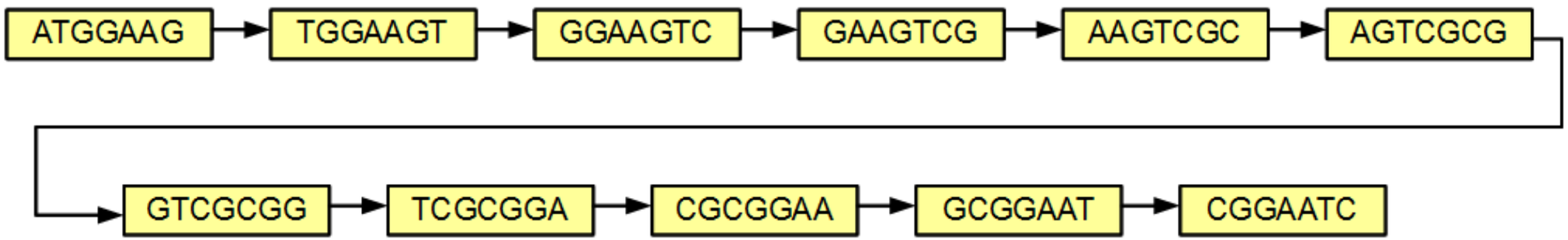
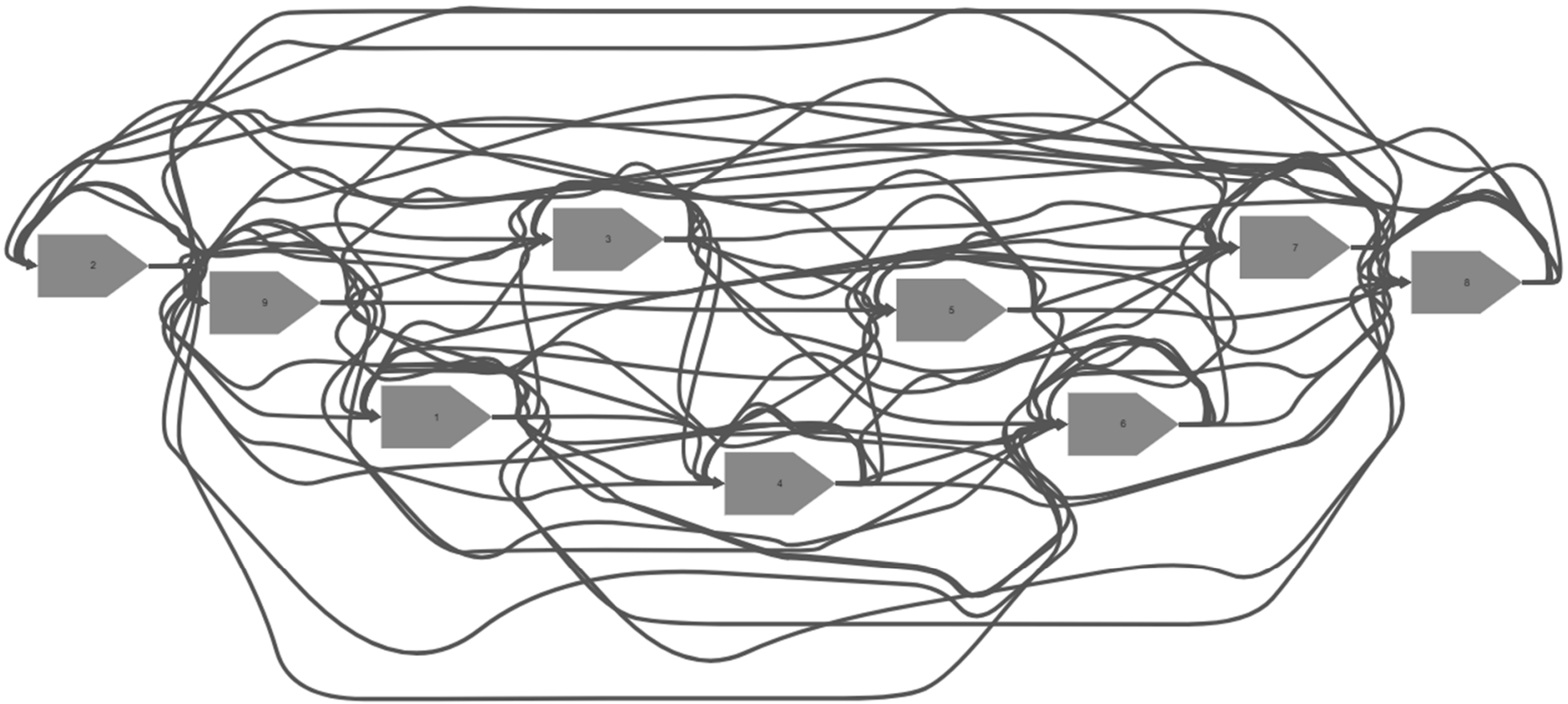


Image credit: https://homolog.us/Tutorials/book4/p2.1.html (via Google Images)

Example 1: AAAAAAAAAAAA (no unique 3-mer)

- 4bp reads:
  - (1) AAAA, (2) AAAA, (3) AAAA, (4) AAAA, (5) AAAA, (6) AAAA, (7) AAAA, (8) AAAA, (9) AAAA
- 3-mers:
  - AAA
- Overlaps:
  - (1)->(2),(1)->(3),(1)->(4),(1)->(5),(1)->(6),(1)->(7),(1)->8,(1)->(9)
  - (2)->(1),(2)->(3),(2)->(4),(2)->(5),(2)->(6),(2)->(7),(2)->8,(2)->(9)
  - ....
  - (9)->(1),(9)->(2),(9)->(3),(9)->(4),(9)->(5),(9)->(6),(9)->(7),(9)->(8)

Example 1: AAAAAAAAAAAA (no unique 3-mer)

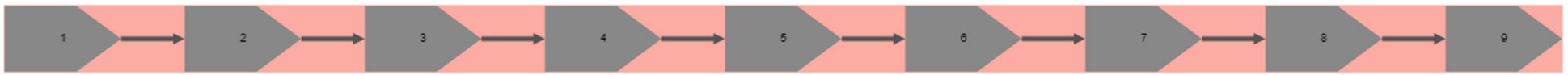


Slide credits: MetagenomeScope

Example 2: AATCCGTTCGGA (no 3-mer repeats)

- 4bp reads:
  - (1) AATC, (2) ATCC, (3) TCCG, (4) CCGT, (5) CGTT, (6) GTTC, (7) TTCG, (8) TCGG, (9) CGGA
- 3-mers (10)
  - (i) AAT, (ii) ATC, (iii) TCC, (iv) CCG, (v) CGT, (vi) GTT, (vii) TTC, (viii) TCG, (ix) CGG, (x) GGA
- Overlaps
  - (1)->(2)
  - (2)->(3)
  - (3)->(4)
  - (4)->(5)
  - (5)->(6)
  - (6)->(7)
  - (7)->(8)
  - (8)->(9)

Example 2: AATCCGTTCGGA (no 3-mer repeats)



Slide credits: MetagenomeScope

Example 2: AATCCGTTCGGA (no 3-mer repeats)

```
AATCCGTTCGGA
AATC
 ATCC
  TCCG
   CCGT
    CGTT
     GTTC
      TTCG
       TCGG
        CGGA
```

Example 3: AATCCGTTCGGA (sequencing error)

- 4bp reads:
  - (1) AAT**G**, (2) ATCC, (3) TCCG, (4) CCGA, (5) CGTT, (6) GTTC, (7) TTCG, (8) TCGG, (9) CGGA
- 3-mers (10)
  - (i) AAT, (ii) ATC, (iii) TCC, (iv) CCG, (v) ATG, (vi) CGT, (vii) GTT, (viii) TTC, (xi) TCG, (x) CGG, (xi) GGA
- Overlaps
  - ~~(1)->(2)~~
  - (2)->(3)
  - (3)->(4)
  - (4)->(5)
  - (5)->(6)
  - (6)->(7)
  - (7)->(8)
  - (8)->(9)

ATCCGTTCGGA

ATCC

TCCG

CCGT

CGTT

GTTC

TTCG

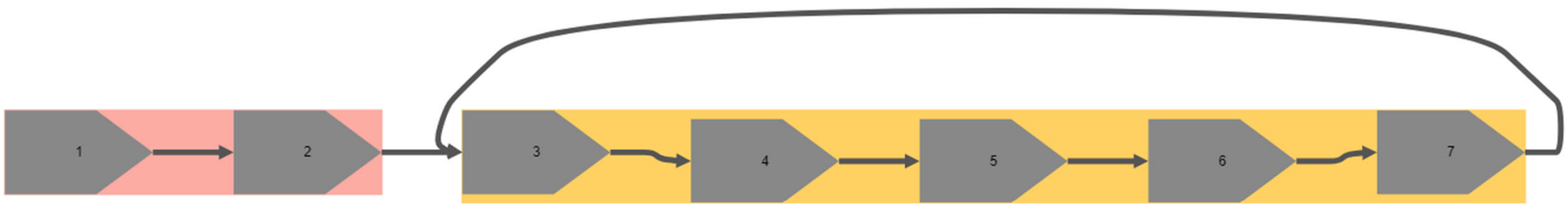
TCGG

CGGA

Example 4: AATCCGTTCGGA (sequencing error)

- 4bp reads:
  - (1) AATC, (2) ATCC, (3) TCCG, (4) CCGT, (5) CGTT, (6) GTTC, (7) TTC**C**, (8) TCGG, (9) CGGA
- 3-mers (10)
  - (i) AAT, (ii) ATC, (iii) TCC (X2), (iv) CCG, (v) CGT, (vi) GTT, (vii) TTC, (viii) TCG, (ix) CGG, (x) GGA
- Overlaps
  - (1)->(2)
  - (2)->(3)
  - (3)->(4)
  - (4)->(5)
  - (5)->(6)
  - (6)->(7)
  - ~~(7)->(8)~~
  - (7)->(3)
  - (8)->(9)

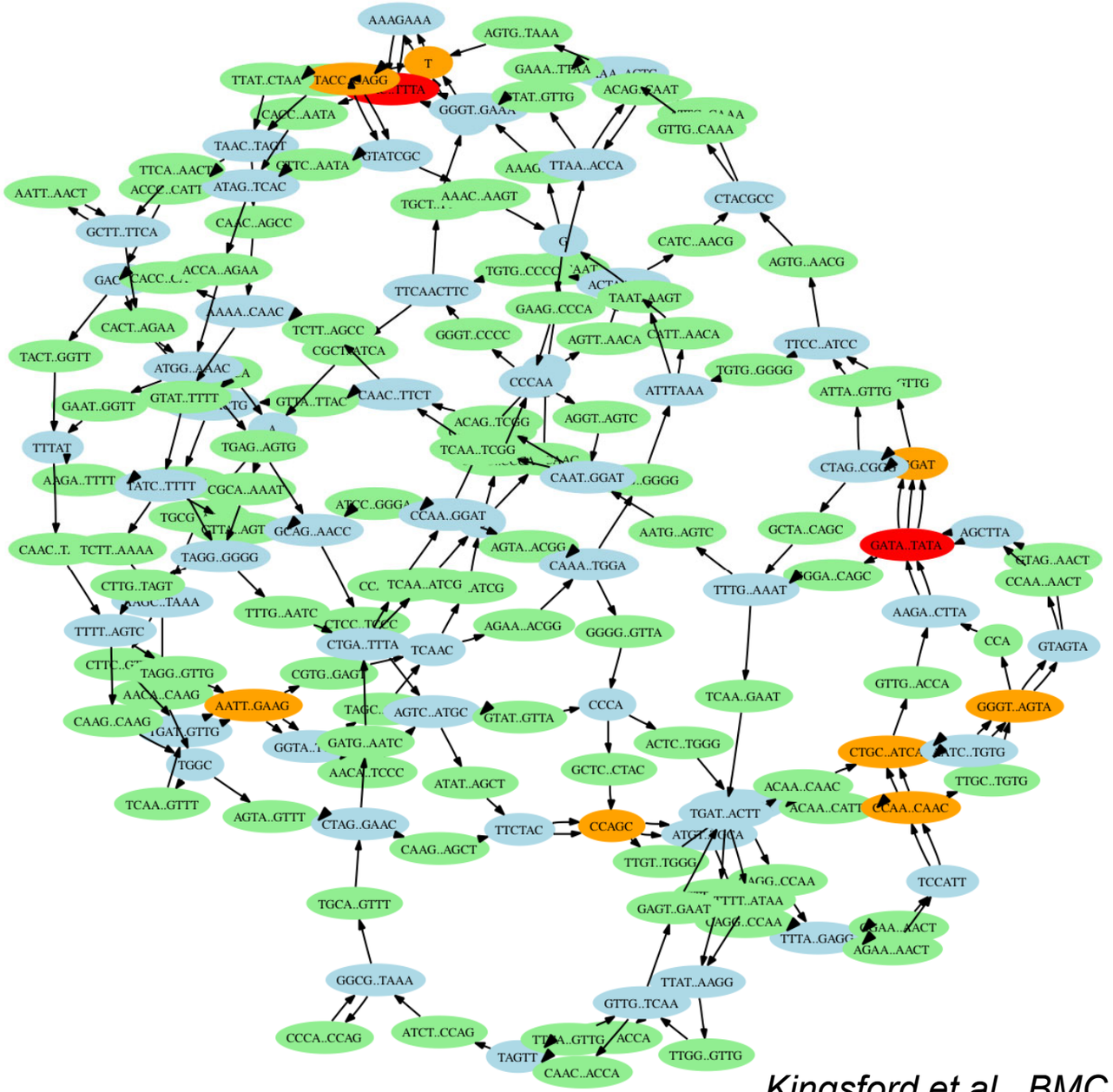
Example 4: AATCCGTTCGGA (sequencing error)



Slide credits: MetagenomeScope

Example 5: AATCCGTTCGGA (coverage gap)

- 4bp reads:
  - (1) AATC, (2) ATCC, (3) TCCG, ~~(4) CCGT, (5) CGTT~~, (6) GTTC, (7) TTCG, (8) TCGG, (9) CGGA
- 3-mers (10)
  - (i) AAT, (ii) ATC, (iii) TCC, (iv) CCG, (v) TTC, (vi) TCG, (vii) CGG, (viii) GGA
- Overlaps
  - (1)->(2)
  - (2)->(3)
  - ~~(3)->(4)~~
  - ~~(4)->(5)~~
  - ~~(5)->(6)~~
  - (6)->(7)
  - (7)->(8)
  - (8)->(9)

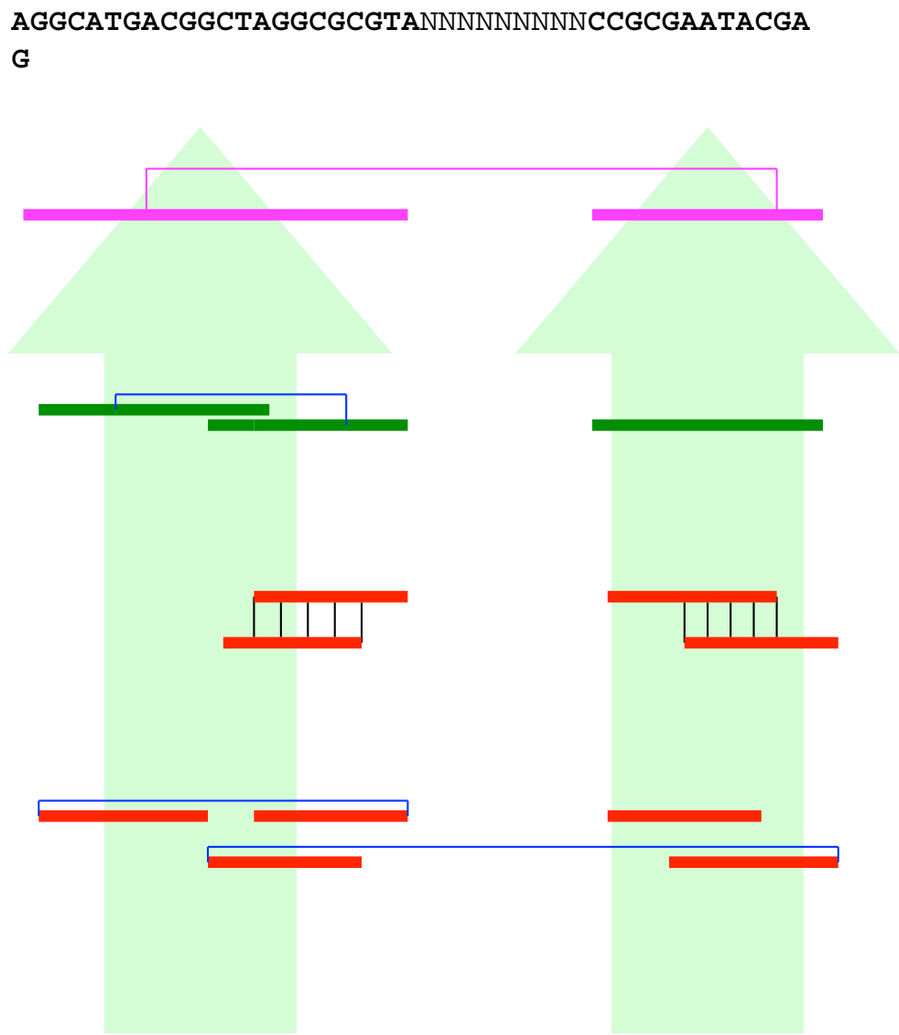


Mycoplasma genitalium, 25 bp reads

Kingsford et al., BMC Bioinformatics 2010

Assembly Glossary

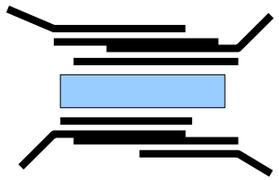
- **Consensus**
  - Multiple alignment of read sequences
- **Scaffold = Contigs + Gaps**
  - group of contigs that can be ordered and oriented with respect to each other (usually with the help of mate--pair data)
- **Con**tigs
  - contiguous segment of DNA reconstructed (unambiguously) from a set of reads
- **Overlaps**
  - Shared sequences between the suffix of one read and the prefix of another
- **Reads**
  - small (50--2000bp) segment of DNA "read" by a sequencing instrument
- **Mate--pair, paired ends**
  - pair of reads whose distance from each other within the genome is approximately known



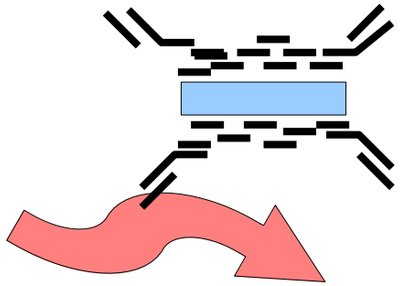
Read length matters…

- Reads (much) longer than repeats – assembly trivial

- Reads roughly equal to repeats – assembly computationally difficult (NP-hard)



- Reads shorter than repeats – assembly undetermined

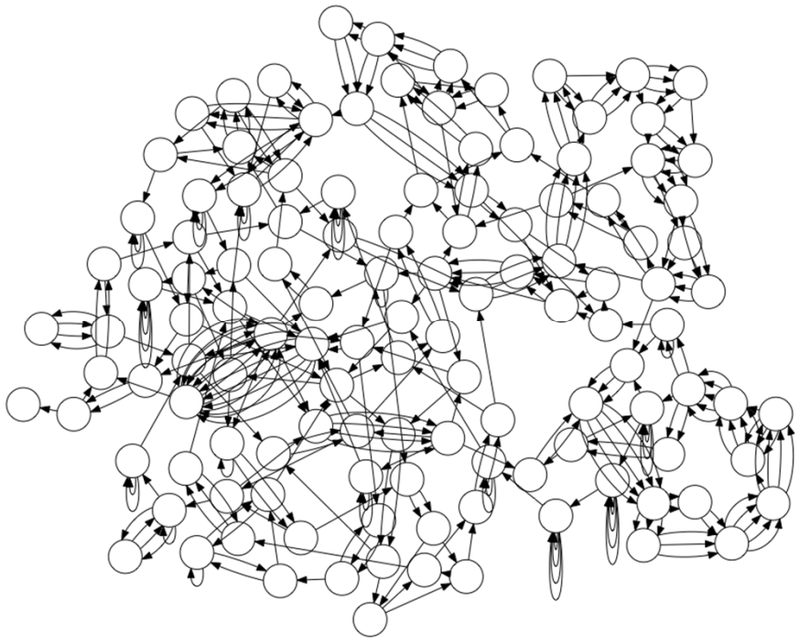


- Number of possible reconstructions exponential in # of repeats

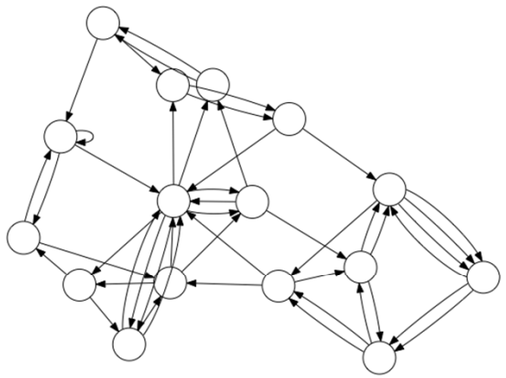
Nagarajan, Pop. J. Comp. Biol. 2009, Kingsford et al., BMC Bioinformatics 2010

Read length matters

k = 50



k = 1,000

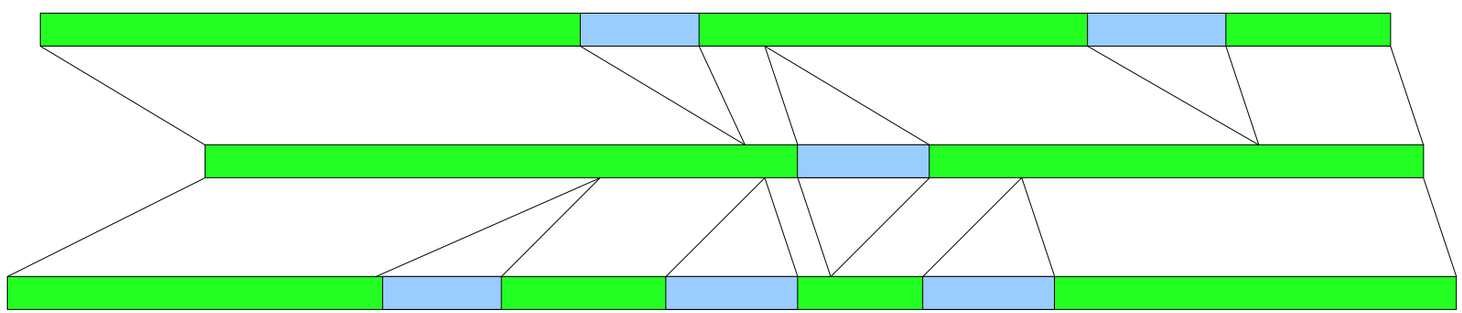


k = 5,000

What are repeats?

Isolate genome

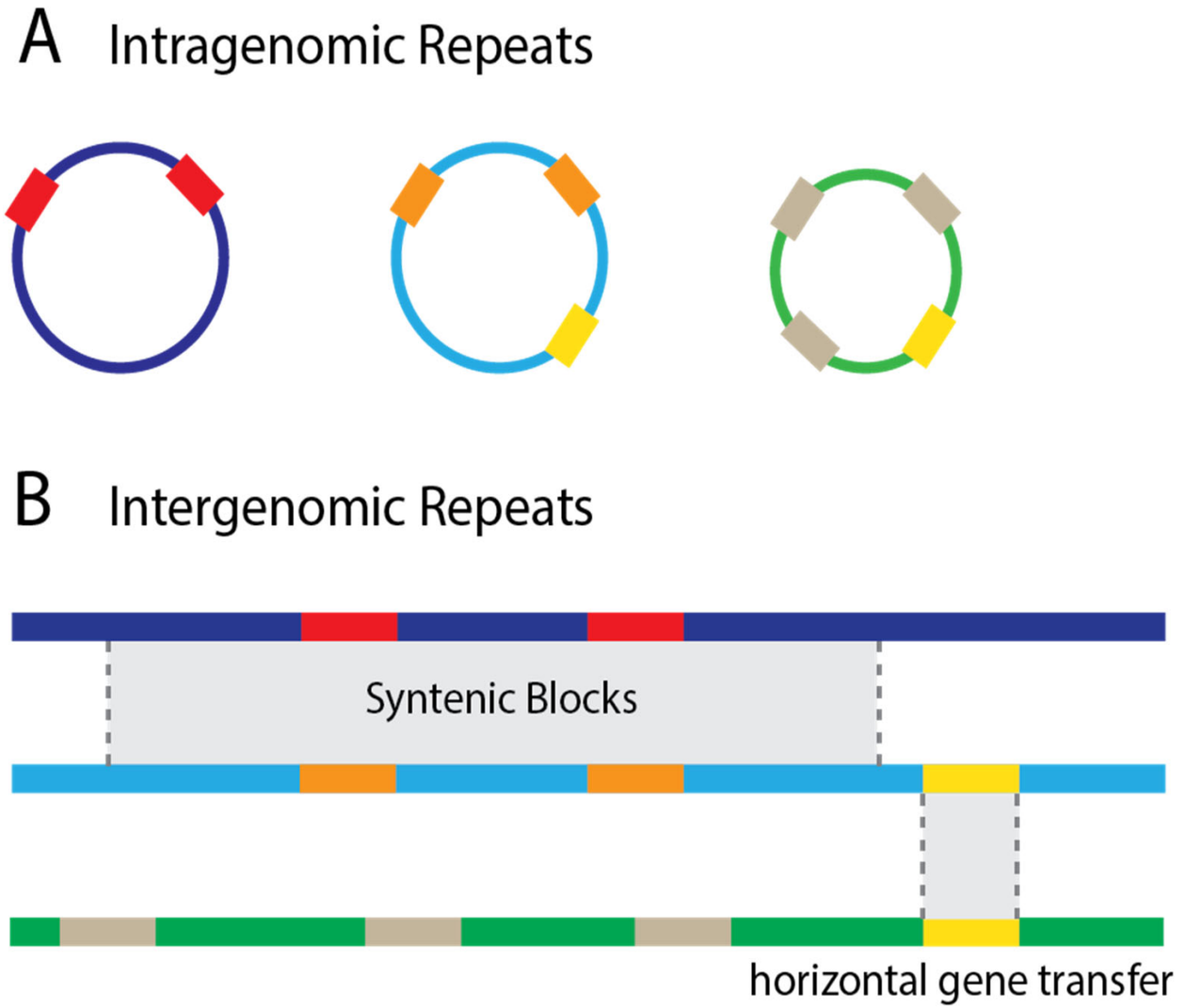
Metagenome



In metagenomes repeats are approximately genome-sized

Haplotype phasing with unknown number of haplotypes

What are repeats?



Challenges in Metagenomic assembly

- Difficult to find repeats
  - coverage vs. over-representation
  - within-genome vs. across-genome repeats

- High genomic variation
  - sequencing experiment has ~10^{15} cells, i.e., each read comes from a different cell
  - phages, transposons, etc. affect only a fraction of the population even in 'homogeneous' strains

Metagenomic questions

- What is the relative abundance of organism X versus organisms Y and Z?

- What proportion of organisms of type X have pathogenicity island P?

- Is pathogenicity island P only found in organism X or also in organisms Y and Z?

  E. coli ETEC, EPEC, EAEC, EHEC, …

- Shiga toxin in *Shigella* or *E. coli,* …

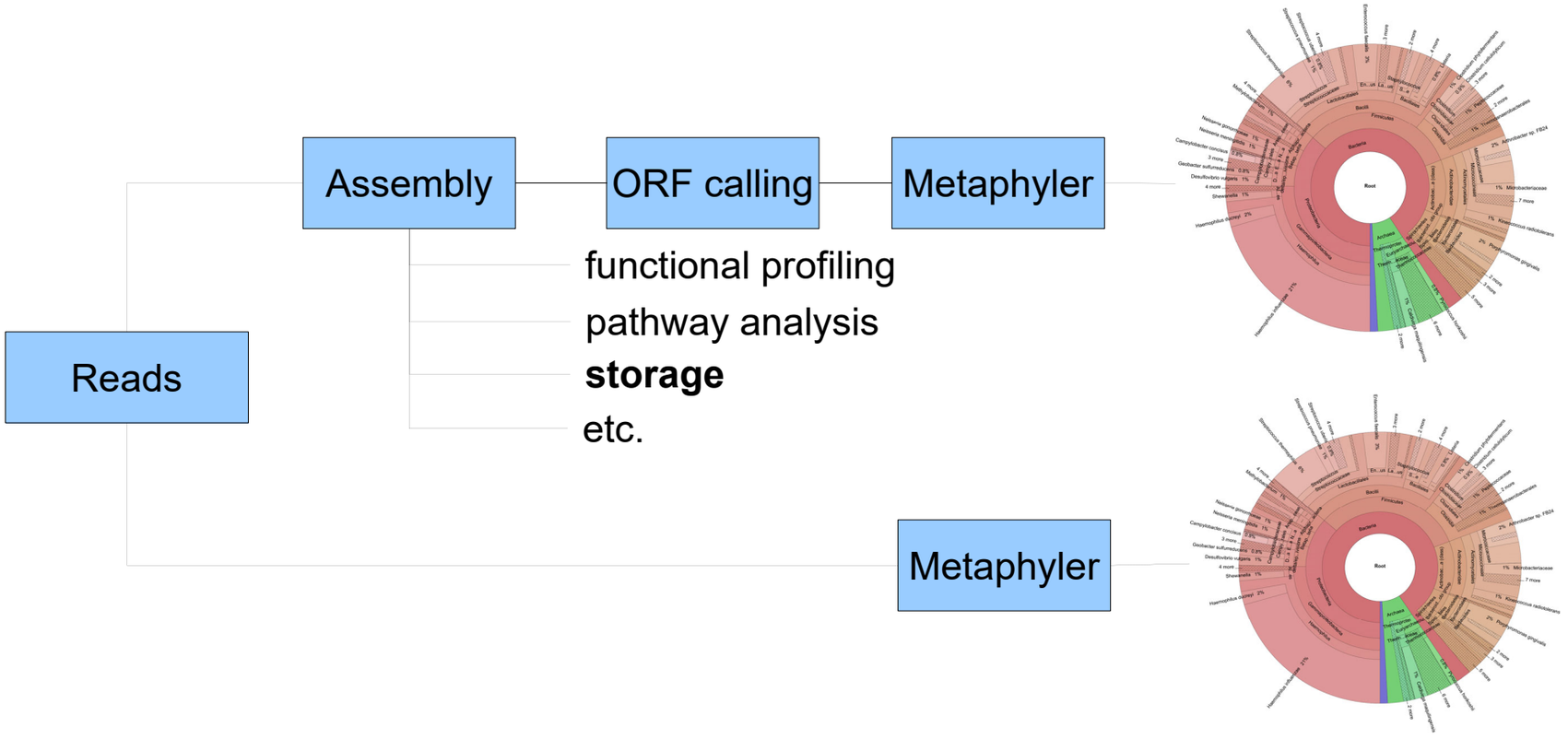
If assembly is impossible, **WHY BOTHER?**

- Long reads (even 10kbp) insufficient as repeats are as long as genomes (100s of kbps to Mbps)
- Errors impossible to avoid in low coverage genomes
- Mate-pairs don't help
- Computationally, assembly is very very hard

Assembly as compression

Stool sample SRS049995

- in: 11.2 Gbp
- out: 174 Mbp + 20 Mbp (unassembled reads)

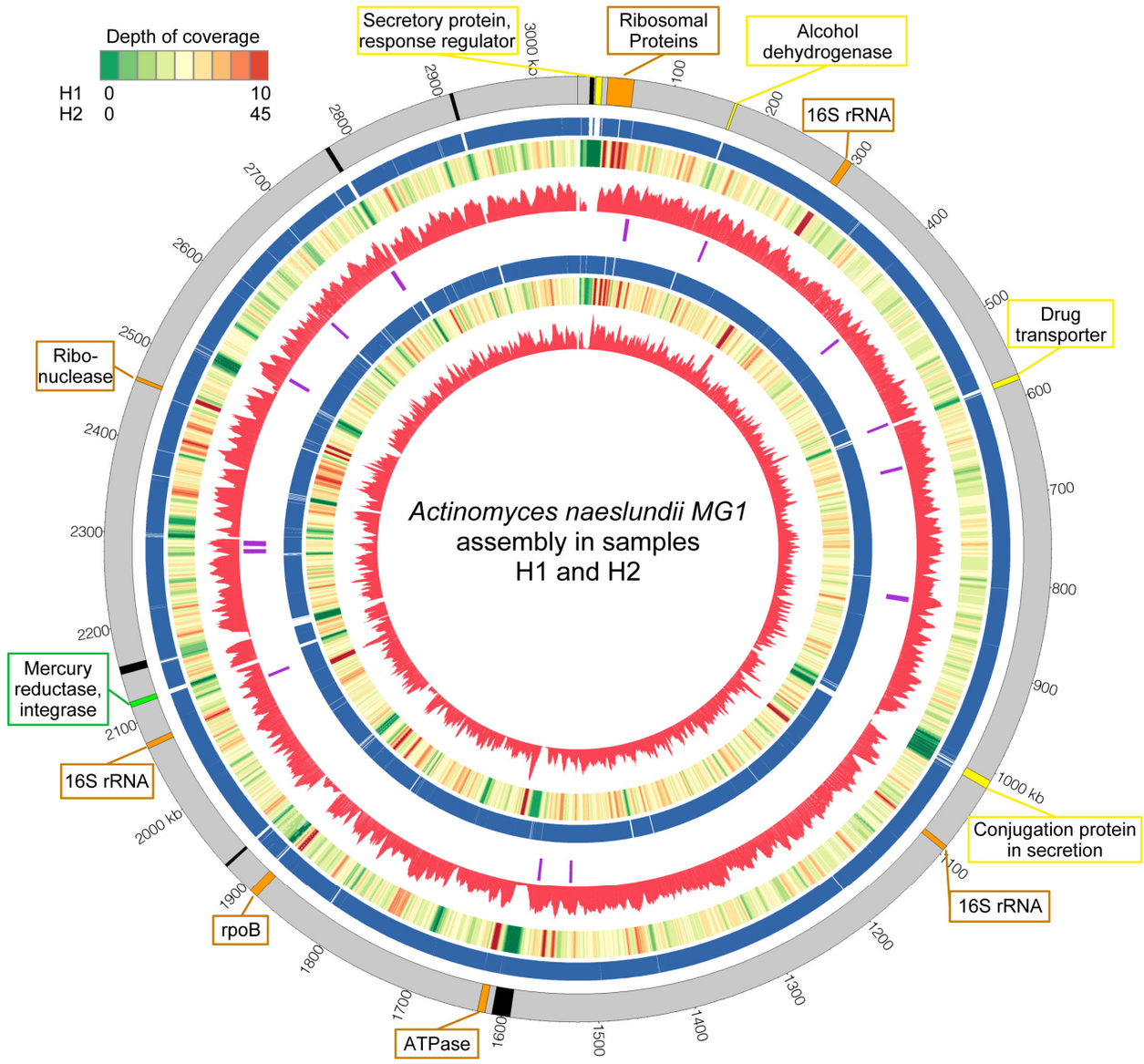


Liu et al. BMC Genomics 2011, Treangen et al. Gen. Biol. 2014

Interesting genomes

- Most microbes are not easily cultured and only known by 16S rDNA signature
  - e.g. RDP grew from ~80,000 (v. 10.4) to 2.1 million (v 10.28)
  - only ~10,000 sequences from type strains
  - only ~150,000 sequences from isolate genomes
  - metagenomic assembly is only way to get the rest
- Clinical studies reveal interesting 16S patterns – what do the genomes do?

Strain structure matters



Metagenome Assembly

Brass Tacks: Software Options & Comparison

Major Software Options

De Novo Methods:

- MEGAHIT (Li, et al. – 2015)
  - https://github.com/voutcn/megahit
  - Fast, Pretty Good
  - Run time on 10-50m paired-end short reads ≈ 1 hour
- metaSPAdes (Nurk, et al. – 2017)
  - https://github.com/ablab/spades
  - Slow, *Slightly* Better
  - Approximately 4x run time vs. MEGAHIT

Reference-Guided Methods:

- MetaCompass (Luan, et al. – 2024)
  - https://github.com/marbl/MetaCompass

MEGAHIT vs. metaSPAdes: SRA Run ID# DRR327028

These outputs are from metaQUAST, which evaluates assemblies.

Easiest comparison is the "without reference" section (in red box).

MEGAHIT

Combined reference | 29 726 918 bp | 9 references | 65 frag

Genome statistics	final.contigs
Genome fraction (%)	56.685
Duplication ratio	1.283
Largest alignment	114 257
Total aligned length	18 744 187
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	942
Misassembled contigs length	7 367 242
Mismatches	
# mismatches per 100 kbp	2617.62
# indels per 100 kbp	84.22
# N's per 100 kbp	0
Statistics without reference	
# contigs	92 479
Largest contig	495 888
Total length	298 680 427
Total length (>= 1000 bp)	263 860 736
Total length (>= 10000 bp)	172 003 526
Total length (>= 50000 bp)	78 037 741

Extended report

metaSPAdes

Combined reference | 59 689 442 bp | 18 references | 341 fragments

Genome statistics	contigs
Genome fraction (%)	63.517
Duplication ratio	1.054
Largest alignment	133 601
Total aligned length	37 274 964
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	1139
Misassembled contigs length	16 017 273
Mismatches	
# mismatches per 100 kbp	1859.94
# indels per 100 kbp	59.26
# N's per 100 kbp	0
Statistics without reference	
# contigs	94 589
Largest contig	752 315
Total length	300 077 086
Total length (>= 1000 bp)	263 798 233
Total length (>= 10000 bp)	172 362 273
Total length (>= 50000 bp)	80 891 785

Extended report

MEGAHIT vs. metaSPAdes: SRA Run ID# ERR1398068

These outputs are from metaQUAST, which evaluates assemblies.

Easiest comparison is the "without reference" section (in red box).

MEGAHIT

Combined reference | 44 002 092 bp | 13 references | 70 fragments

Genome statistics	final.contigs
Genome fraction (%)	43.882
Duplication ratio	1.061
Largest alignment	167 449
Total aligned length	17 240 928
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	474
Misassembled contigs length	4 191 320
Mismatches	
# mismatches per 100 kbp	1776.64
# indels per 100 kbp	46.93
# N's per 100 kbp	0
Statistics without reference	
# contigs	63 786
Largest contig	374 151
Total length	112 012 534
Total length (>= 1000 bp)	85 896 097
Total length (>= 10000 bp)	30 982 957
Total length (>= 50000 bp)	10 174 731

Extended report

metaSPAdes

Combined reference | 68 768 022 bp | 19 references | 117 fragments

Genome statistics	contigs
Genome fraction (%)	49.852
Duplication ratio	1.022
Largest alignment	44 369
Total aligned length	33 561 685
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	460
Misassembled contigs length	2 075 607
Mismatches	
# mismatches per 100 kbp	1215.94
# indels per 100 kbp	27.61
# N's per 100 kbp	0
Statistics without reference	
# contigs	68 611
Largest contig	316 066
Total length	114 681 501
Total length (>= 1000 bp)	85 909 546
Total length (>= 10000 bp)	29 769 961
Total length (>= 50000 bp)	9 439 558

Extended report

MEGAHIT vs. metaSPAdes: SRA Run ID# ERR1398155

These outputs are from metaQUAST, which evaluates assemblies.

Easiest comparison is the "without reference" section (in red box).

MEGAHIT

Combined reference | 19 889 285 bp | 6 references | 14 fragments

Genome statistics	final.contigs
Genome fraction (%)	31.07
Duplication ratio	1.169
Largest alignment	88 423
Total aligned length	6 366 843
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	229
Misassembled contigs length	1 550 069
Mismatches	
# mismatches per 100 kbp	1914.83
# indels per 100 kbp	44.6
# N's per 100 kbp	0
Statistics without reference	
# contigs	89 409
Largest contig	243 923
Total length	150 848 367
Total length (>= 1000 bp)	114 753 870
Total length (>= 10000 bp)	34 931 512
Total length (>= 50000 bp)	7 578 590

metaSPAdes

Combined reference | 35 706 667 bp | 11 references | 86 fragments

Genome statistics	contigs
Genome fraction (%)	39.505
Duplication ratio	1.06
Largest alignment	161 720
Total aligned length	13 935 121
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	271
Misassembled contigs length	2 449 316
Mismatches	
# mismatches per 100 kbp	1852.45
# indels per 100 kbp	47.65
# N's per 100 kbp	0
Statistics without reference	
# contigs	80 848
Largest contig	234 563
Total length	135 785 774
Total length (>= 1000 bp)	103 522 252
Total length (>= 10000 bp)	29 611 925
Total length (>= 50000 bp)	6 902 070

MEGAHIT vs. metaSPAdes: SRA Run ID# SRR27117388

These outputs are from metaQUAST, which evaluates assemblies.

Easiest comparison is the "without reference" section (in red box).

MEGAHIT

Combined reference | 32 364 850 bp | 10 references | 55 fragments

Genome statistics	final.contigs
Genome fraction (%)	60.665
Duplication ratio	1.201
Largest alignment	73 347
Total aligned length	19 479 654
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	933
Misassembled contigs length	4 463 218
Mismatches	
# mismatches per 100 kbp	1998.16
# indels per 100 kbp	53.58
# N's per 100 kbp	0
Statistics without reference	
# contigs	96 118
Largest contig	609 215
Total length	265 146 582
Total length (>= 1000 bp)	228 488 467
Total length (>= 10000 bp)	135 933 934
Total length (>= 50000 bp)	62 609 535

Extended report

metaSPAdes

Combined reference | 37 825 095 bp | 12 references | 84 fragments

Genome statistics	contigs
Genome fraction (%)	43.837
Duplication ratio	1.044
Largest alignment	217 293
Total aligned length	16 120 059
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	355
Misassembled contigs length	5 088 156
Mismatches	
# mismatches per 100 kbp	1476.36
# indels per 100 kbp	43.85
# N's per 100 kbp	0
Statistics without reference	
# contigs	102 821
Largest contig	496 535
Total length	268 922 252
Total length (>= 1000 bp)	229 506 524
Total length (>= 10000 bp)	130 538 509
Total length (>= 50000 bp)	55 869 523

Sneak Preview: Metagenomic Binning

- Once we have these contigs, what do we do with them?

- Questions:
  - Which organism does it belong to?
  - Which contigs are from the same organism?

Questions?

Tutorial Link:
https://github.com/MGNute/stamps_2024_assembly_tutorial/blob/main/qc_assembly.md