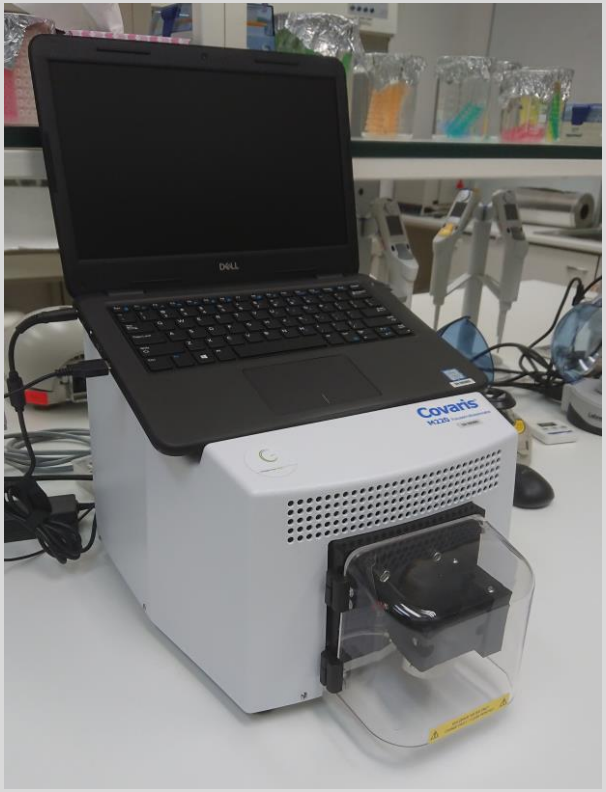
Introduction to NGS data analysis

Tomáš Fér
Department of Botany, Charles University, Prague
tomas.fer@centrum.cz
http://botany.natur.cuni.cz/fer/markers/practicals/NGS.htm
2023

Outline

- library preparation – sonication, NEBNext Ultra II
- sequencing – Illumina
- FASTQ files – quality scores
- quality check – FastQC
- trimming – adaptors & bad quality
- mapping to reference – BWA
- SAM/BAM file structure – flags
- variant calling – samtools, bcftools
- VCF file structure
- de novo assembly – de Bruijn graphs (Velvet)

NGS library prep – sonication



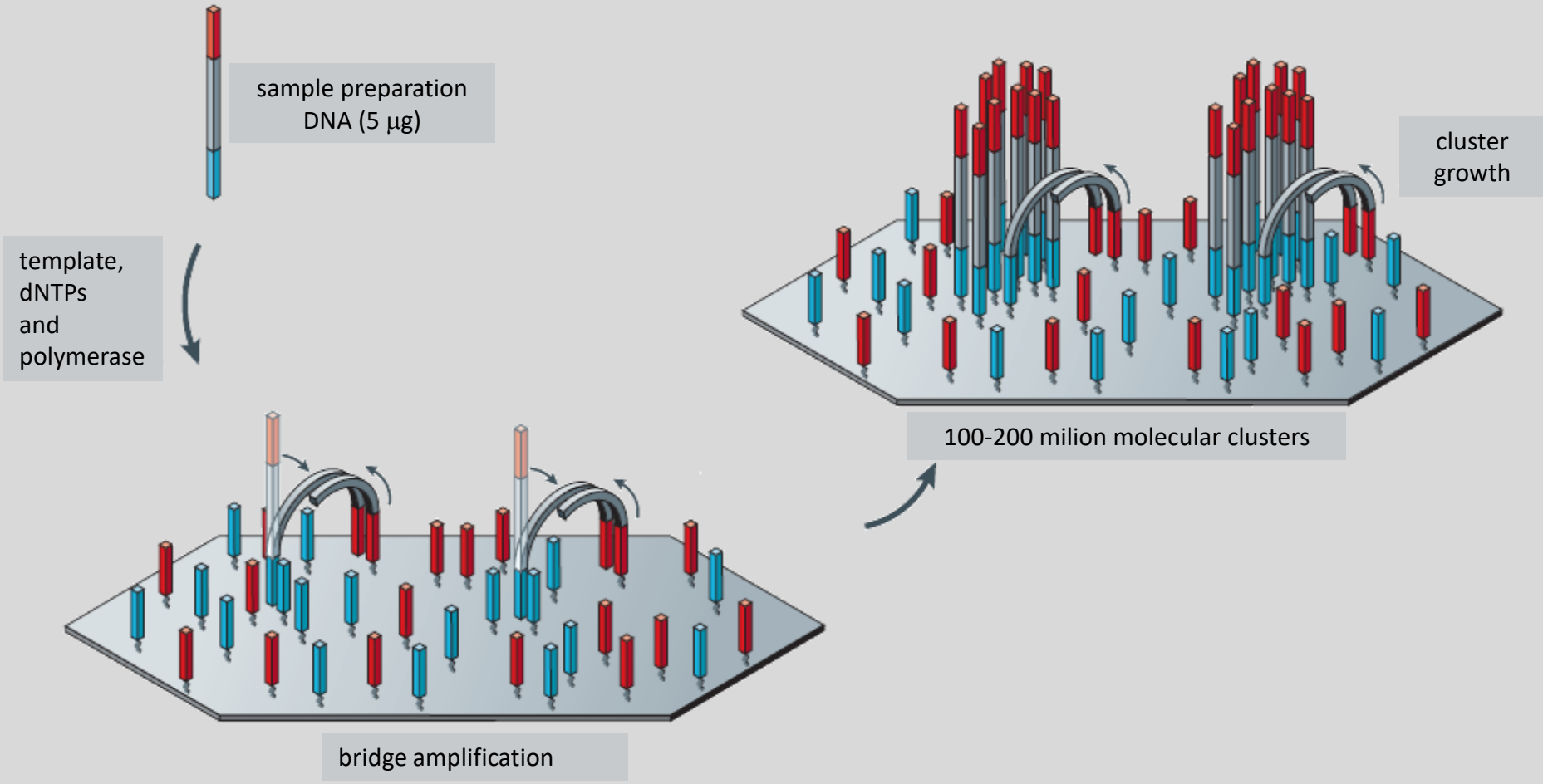
Covaris M220

- DNA fragmentation using ultrasound
- Adaptive Focused Acoustics (AFA) technology

NGS library prep – NEBNext Ultra II

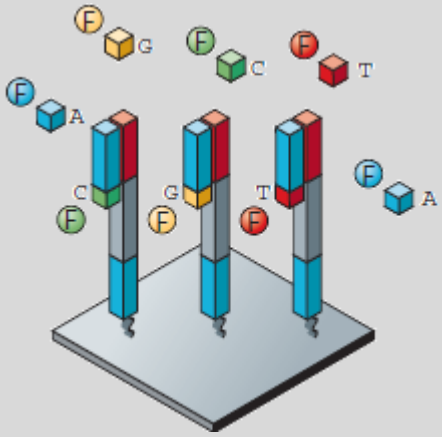
- End repair and A-tailing
- Adaptor ligation
- Sample clean-up
- Size selection
- Final concentration measurement (e.g., Qubit)

Solid-phase amplification (Illumina)

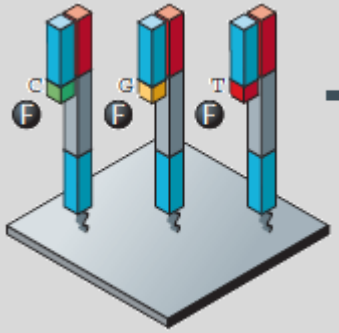


Metzker 2010

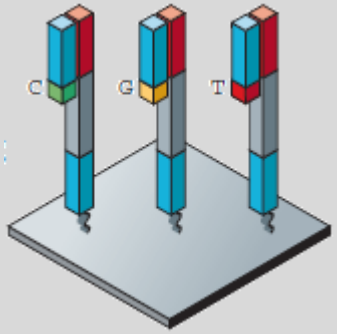
Cyclic reversible termination (Illumina)



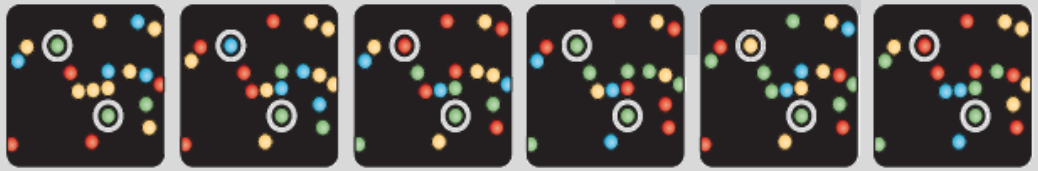
incorporate all four nucleotides, each label with a different dye



wash, four-colour imaging



cleave dye and terminating groups, wash



Top: CATCGT
Bottom: CCCCCC

instruments – MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq

Metzker 2010

FASTQ

- FASTA + quality scores

M01691	the unique instrument name
85	the run id
000000000-ABGJG	the flowcell ID
1	flowcell lane
1101	tile number within the flowcell lane
15345	'x'-coordinate of the cluster within the tile
2139	'y'-coordinate of the cluster within the tile
2	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
19	sample number

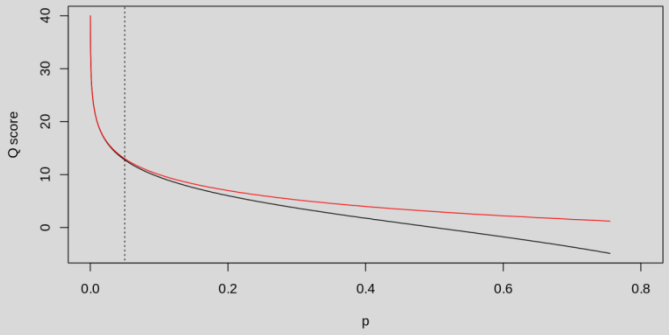
```
@M01691:85:000000000-ABGJG:1:1101:15345:2139 2:N:0:19
GGCAAGCTCTGTACGGTTATGAACCCCCCAAGTCGGACAAGCGGCCAAAGATGACGTGTGACTGCTTGCCTTCATGGTGT
TTCTGCTGCTGTTGCTGCTGCTTTGGCTCGAGAAAATCCAAAGCTAAGAAAGGAGGAGCGAAGGGGGGTTT
+
>11>A11B1@D@FGGE1AEGFDFCGAA?A0AA1BEG?/A00AEEEE/FB0B@D11EFCG>1GHHHHEFH11BBFF2F11?
?BFH2DB1B11>111BGFFHH1DG1BG1<E/C//1BGFHH<G0GF1>@11@GC0?A?//</!-<C?<;@-A
```

line1 starts with '@' followed by sequence identifier
line2 raw sequence letters
line3 '+' character
line4 quality values

Quality scores

- quality between 0 and 41 (for Illumina 1.8 and higher)
  `!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ`

- $Q_{Phred} = -10 \log_{10} p$
- $Q_{solexa\text{-}prior\ to\ v1.3} = -10 \log_{10} (p/(1-p))$



Sanger (Phred)

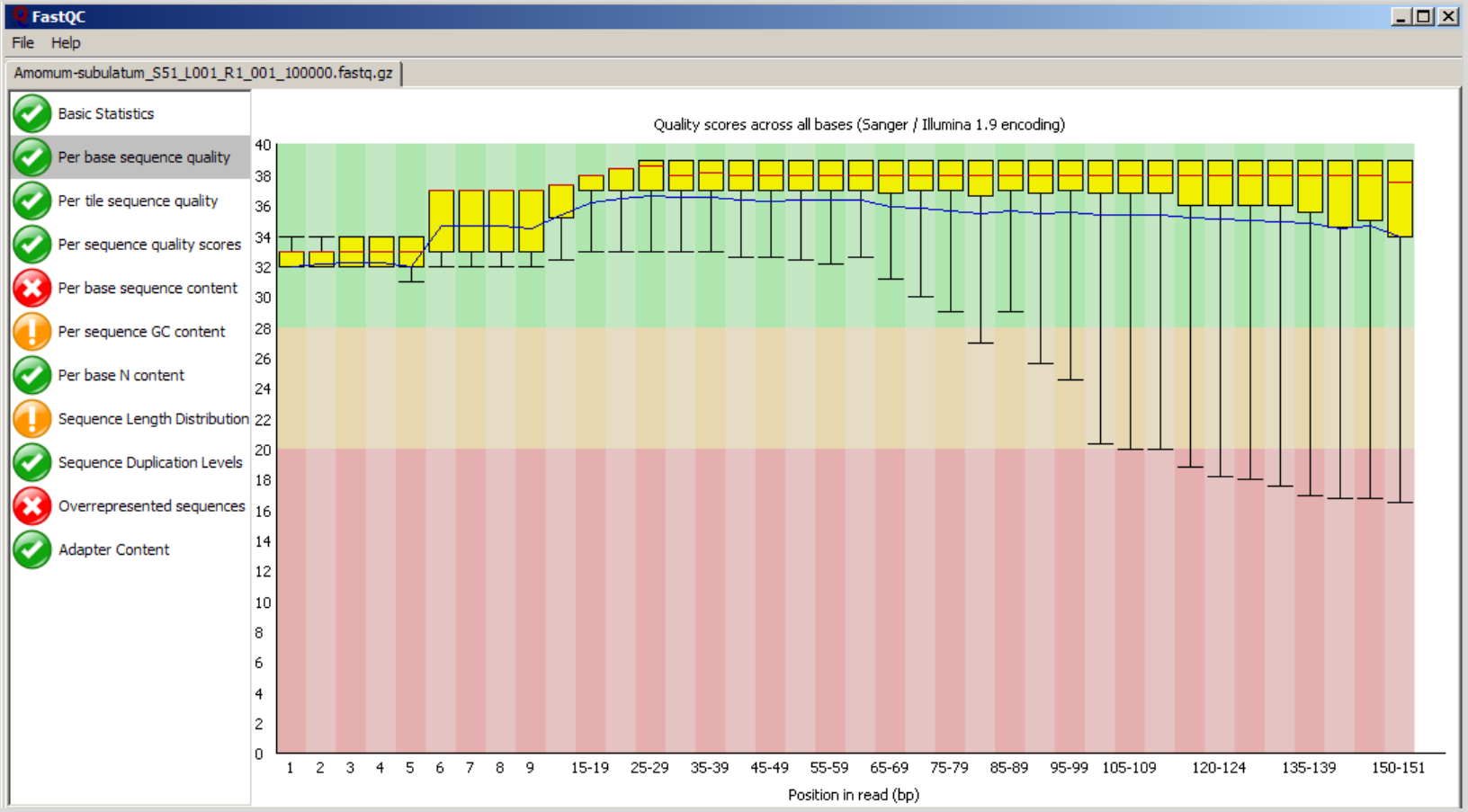
Solexa

Quality score	Probability of incorrect base call
10	1 in 10
20	1 in 100
30	1 in 1,000
40	1 in 10,000
50	1 in 100,000

- https://en.wikipedia.org/wiki/FASTQ_format

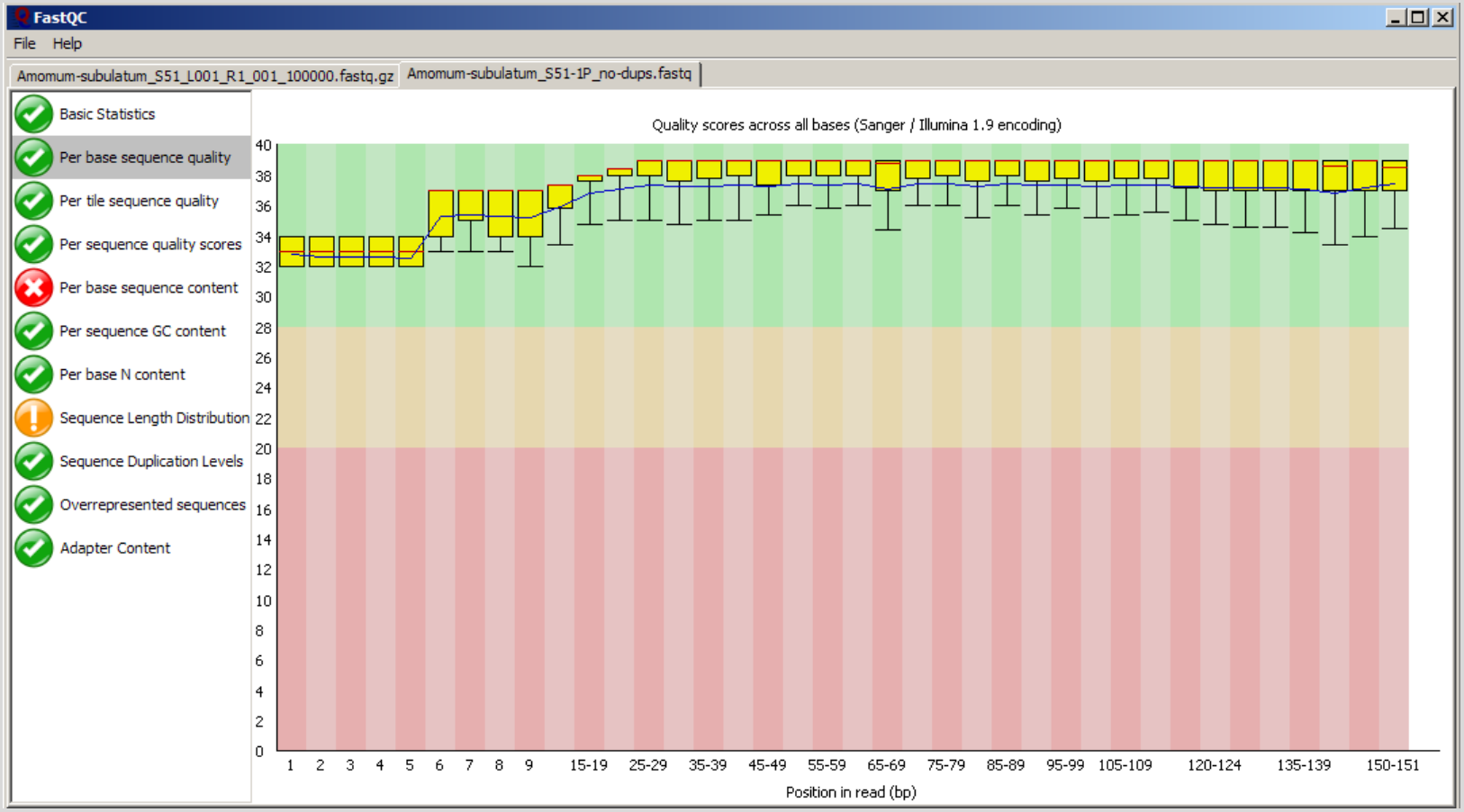
Quality check – FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Trimming

- adaptor contamination
- low quality sequences



- Trimmomatic, cutadapt, Trim Galore, BBMap…

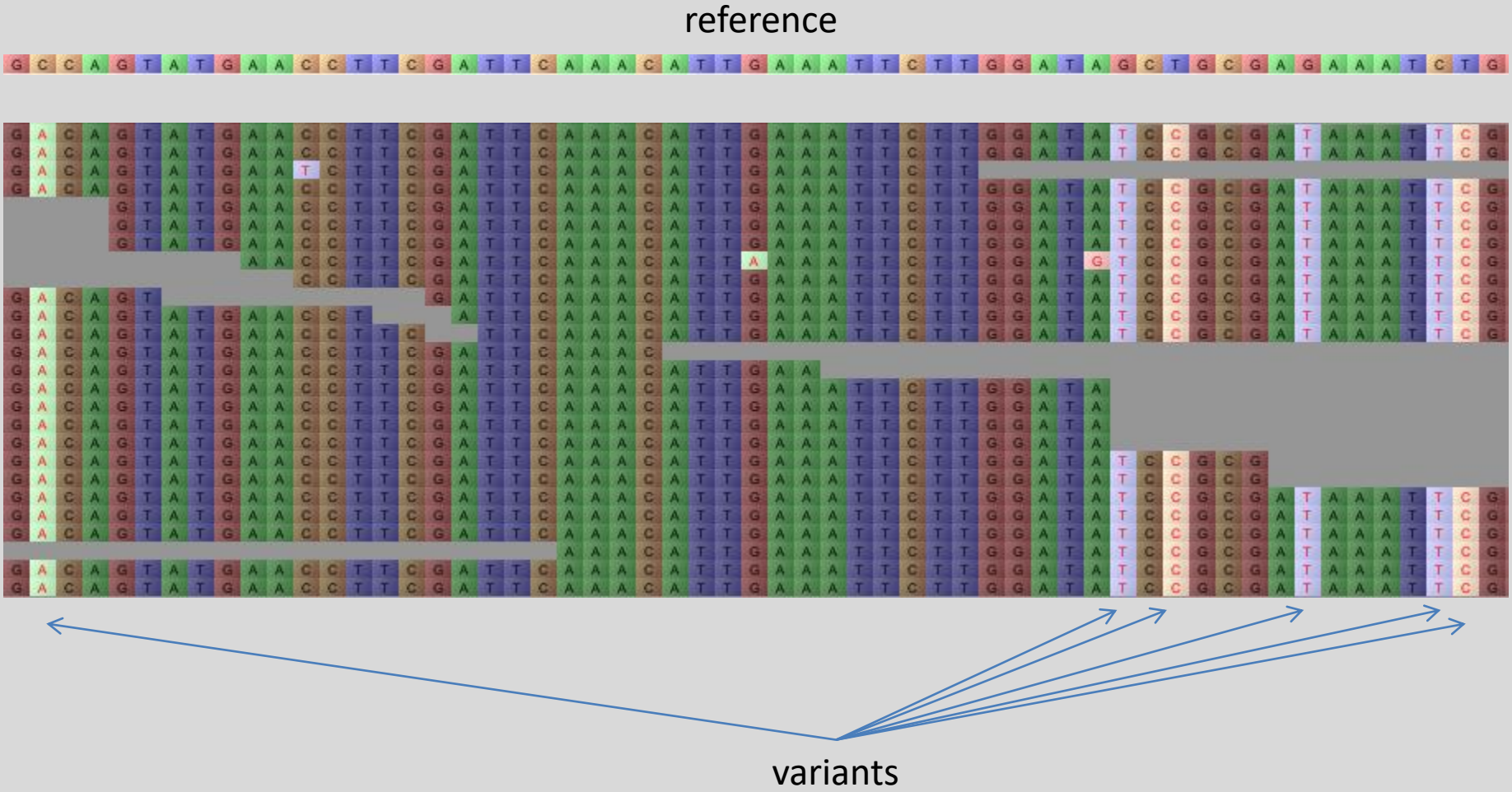
Mapping to the reference

- bowtie2
- BWA (Burrows-Wheeler Aligner)

```
reference     CTGCGTAACTGTCCATGCTGGTTTCATG
read 1        CTGCGTAACTGACC
read 2           CGTAACTGACCATG
read 3                CTGACCATGCTGGTT
read 4                      ATGCTGGTTTCATG
```

Read mapping

- SAM – text file
- BAM – binary file, compressed



SAM/BAM file

```
@SQ	SN:curcuma_HybSeqProbes_test_with400Ns_beginend	LN:77942
@PG	ID:bwa	PN:bwa	VN:0.7.12-r1039	CL:bwa mem curcuma_HybSeqProbes_test_with400Ns_beginend.fas Amomum-subulatum_S51_L001_R1_001_100000.fastq.gz Amomum-subulatum_S51_L001_R2_001_100000.fastq.gz
#QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	TLEN	SEQ	QUAL	OPT
M01691:91:14994	83	curcuma	19770	60	145M4S	=	19601	-314	CCGCT	CGHBF	NM:i:12	MD:Z:1A11A20A2C8T2C5G2T32A23G5C14A8	AS:i:88	XS:i:0
M01691:91:20809	77	*	0	0	*	*	0	0	TGTCA	BBBAB	AS:i:0	XS:i:0
```

@SQ: Reference sequence dictionary

- SN: Reference sequence name
- LN: Reference sequence length

@PG: Program

- ID: Program record identifier
- PN: Program name
- VN: Program version
- CL: Command line

data

- QNAME: query name
- FLAG: bitwise FLAG
- RNAME: reference name
- POS: leftmost position
- MAPQ: mapping quality
- CIGAR: CIGAR string
- MRNM: mate reference seq. name
- MPOS: mate position
- TLEN: inferred insert size
- SEQ: query sequence
- QUAL: query quality
- OPT: optional fields (TAG: VTYPE: VALUE), e.g. NM (edit distance to the reference), MD (String for mismatching positions), AS (alignment score generated by aligner)

Bit position	Hexadecimal	Decimal	Description
1	0x1	1	Read paired
2	0x2	2	Read mapped in proper pair
3	0x4	4	Read unmapped
4	0x8	8	Mate unmapped
5	0x10	16	Read reverse strand
6	0x20	32	Mate reverse strand
7	0x40	64	First in pair
8	0x80	128	Second in pair
9	0x100	256	Not primary alignment
10	0x200	512	Read fails quality checks
11	0x400	1024	Read is PCR or optical duplicate
12	0x800	2048	Supplementary alignment
SUM	**0x53**	**83**	

https://www.samformat.info/sam-format-flag

CIGAR string

Compact **I**diosyncratic **G**apped **A**lignment **R**epresentation

- describes how the read aligns with the reference

Operator	Description
D	**Deletion**; the nucleotide is present in the reference but not in the read
H	**Hard Clipping**; the clipped nucleotides are not present in the read.
I	**Insertion**; the nucleotide is present in the read but not in the reference.
M	**Match**; can be either an alignment match or mismatch. The nucleotide is present in the reference.
N	**Skipped region**; a region of nucleotides is not present in the read
P	**Padding**; padded area in the read and not in the reference
S	**Soft Clipping**; the clipped nucleotides are present in the read
X	**Read Mismatch**; the nucleotide is present in the reference
=	**Read Match**; the nucleotide is present in the reference

```
reference      CTGCGTAA**CTGTCCATGCTGGTTTCATG      CIGAR
read 1         CTGCGTAA                            8M
read 2            aaTAA**CTGACCATG                 2S3M2P9M
read 3               AAGCCTGACCATGCTGGTT           2MI215M
read 4                     tggTGGT**TTTCATG        3H4M2D7M

consensus      CTGCGTAA**CTGACCATGSTGGTTTCATG
```

Creating consensus sequence

```
reference          CTGCGTAACTGTCCATGCTGGTTTCATG
read 1             CTGCGTAACTGACC
read 2                CGTAACTGACCATG
read 3                     CTGACCATGCTGGTT
read 4                           ATGCTGGTTTCATG

read consensus     CTGCGTAACTGACCATGCTGGTTTCATG
```

Variant calling

- variant sites exported only
  1. mpileup – reads the alignments, for each position of the genome constructs a vertical slice across all reads covering the position (“pileup”); genotype likelihoods are calculated – base qualities, mapping qualities, probability of local misalignment (per-base alignment quality; BAQ)
  2. call – most likely genotype under HW evaluated
- VCF (Variant Call Format), BCF (Binary compressed)
- raw SNPs are further filtered
  - % of missing data, read depth, minor allele frequency (MAF), mapping quality
  - VCFtools, GATK, SnpSift, vcfR…
- GATK, SAMtools/BCFtools, FreeBayes, SNVer…

VCF format

VCF header

Body

```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.19-44428cd
##reference=file://curcuma_HybSeqProbes_test_with400Ns_beginend.fas
##contig=<ID=curcuma_HybSeqProbes_test_with400Ns_beginend,length=77942>
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=RPB,Number=1,Type=Float,Description="Read Position Bias">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias (v2) for filtering splice-site artefacts in RNA-seq data. Note: this version may be broken.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Amomum
curcuma	447	.	T	A	107	.	DP=6;VDB=7.293871e-02;AF1=1;AC1=2;DP4=0,0,6,0;MQ=60;FQ=-45	GT:PL:GQ	1/1:140,18,0:33
curcuma	512	.	A	G	165	.	DP=8;VDB=9.040111e-02;AF1=1;AC1=2;DP4=0,0,6,2;MQ=60;FQ=-51	GT:PL:GQ	1/1:198,24,0:45
curcuma	524	.	T	C	173	.	DP=7;VDB=9.724455e-02;AF1=1;AC1=2;DP4=0,0,5,2;MQ=60;FQ=-48	GT:PL:GQ	1/1:206,21,0:39
curcuma	545	.	A	G	188	.	DP=8;VDB=5.242439e-02;AF1=1;AC1=2;DP4=0,0,4,4;MQ=60;FQ=-51	GT:PL:GQ	1/1:221,24,0:45
curcuma	2986	.	G	A	58	.	DP=7;VDB=6.365454e-02;RPB=6.486824e-01;AF1=0.5;AC1=1;DP4=3,1,2,1;MQ=60;FQ=61;PV4=1,1,1,1	GT:PL:GQ	0/1:88,0,109:91
curcuma	2992	.	T	A	55	.	DP=7;VDB=6.621023e-02;RPB=6.486824e-01;AF1=0.5;AC1=1;DP4=3,1,2,1;MQ=60;FQ=57.8;PV4=1,1,1,1	GT:PL:GQ	0/1:85,0,99:88
curcuma	3001	.	G	A	77	.	DP=9;VDB=7.633782e-02;RPB=4.748531e-01;AF1=0.5;AC1=1;DP4=3,2,3,1;MQ=60;FQ=80;PV4=1,1,1,0.49	GT:PL:GQ	0/1:107,0,134:99

mandatory header lines

optional header lines
(metadata about the annotation)

position

reference

alternate

genotype
- 1/1 – homozygote for alternate allele
- 0/1 - heterozygote

Variant columns

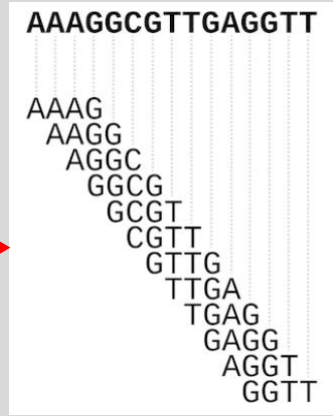
Genotype columns

De novo assembly

- assembling the genome without any reference
- many software – Velvet, SPAdes, DISCOVAR, MaSuRCA
- Velvet – the only assembler working under Windows
  - de Bruijn graph assembler
  - very fast

- *k*-mers – a (DNA) molecule of the length *k*

read to *k*-mer (*k*=4)

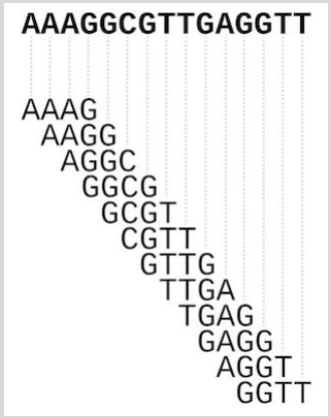


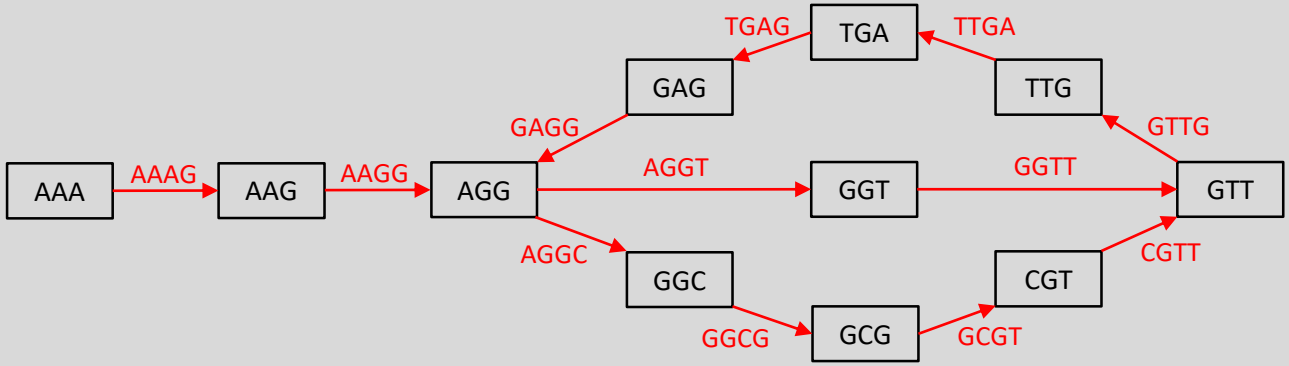
de Bruijn graphs

- network made up of nodes and edges (directed multigraph)
- these comes from the overlaps between *k*-mers

- every possible (*k*-1)-mer is assigned to a node
- edges are all possible *k*-mers
- connect nodes by a directed edge if there is a *k*-mer whose
  - prefix (i.e., all position except the last one) is the former node
  - suffix (i.e., all position except the first one) is the latter node
- Eulerian cycle in the graph (Eulerian walk) – visits each edge exactly ones

k-mer = 4





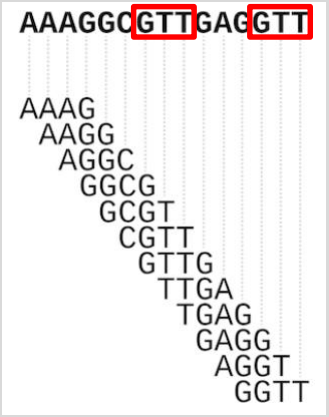
single or multiple Eulerian walk possible? Why?

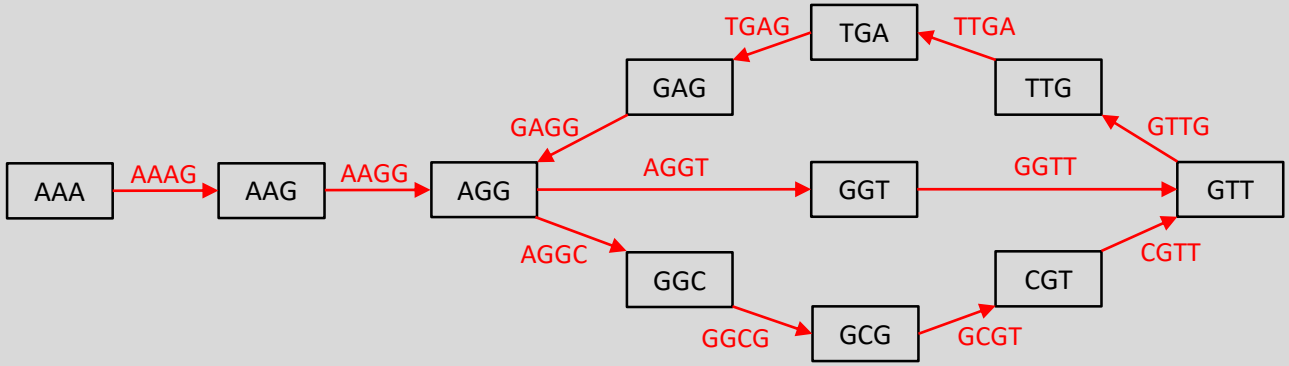
de Bruijn graphs

- network made up of nodes and edges (directed multigraph)
- these comes from the overlaps between *k*-mers

- every possible (*k*-1)-mer is assigned to a node
- edges are all possible *k*-mers
- connect nodes by a directed edge if there is a *k*-mer whose
  - prefix (i.e., all position except the last one) is the former node
  - suffix (i.e., all position except the first one) is the latter node
- Eulerian cycle in the graph (Eulerian walk) – visits each edge exactly ones

k-mer = 4





single or multiple Eulerian walk possible? Why?

de Bruijn graphs

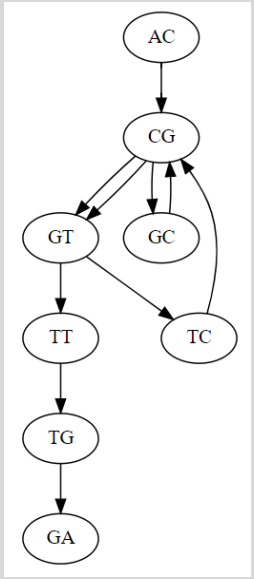
- requirements for straightforward graph
  - all *k*-mers present in the genome sequenced (gaps in sequencing lead to fragmented graphs)
  - all *k*-mers are error-free (error correction possible)
  - each *k*-mer appears at most once in the genome (different coverage requires normalization)
  - genome consists of a single circular chromosome

- play with *k*-mers and graphs using this Jupyter Notebook by B. Langmead https://colab.research.google.com/drive/1pQu9tJZ9RNpk8AaL2ThEYXol3lu7Rw34
- experiment with different *k*-mer settings

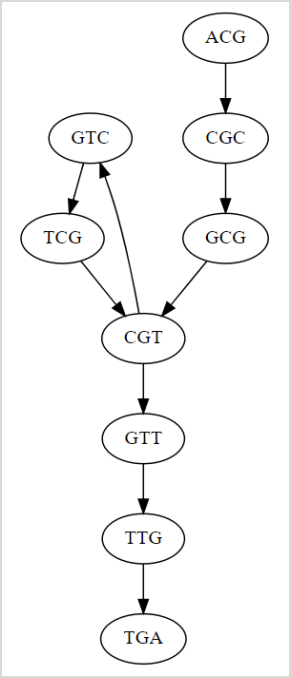
de Bruijn graphs

ACGCGTCGTTGA

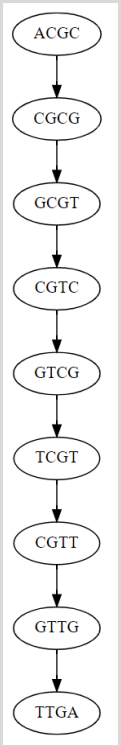
k=3



k=4



k=5



All graphs with Eulerian path

- all nodes (except first and last) are balanced (i.e., # incoming edges = # outcoming edges)
- starting and ending nodes are semibalanced

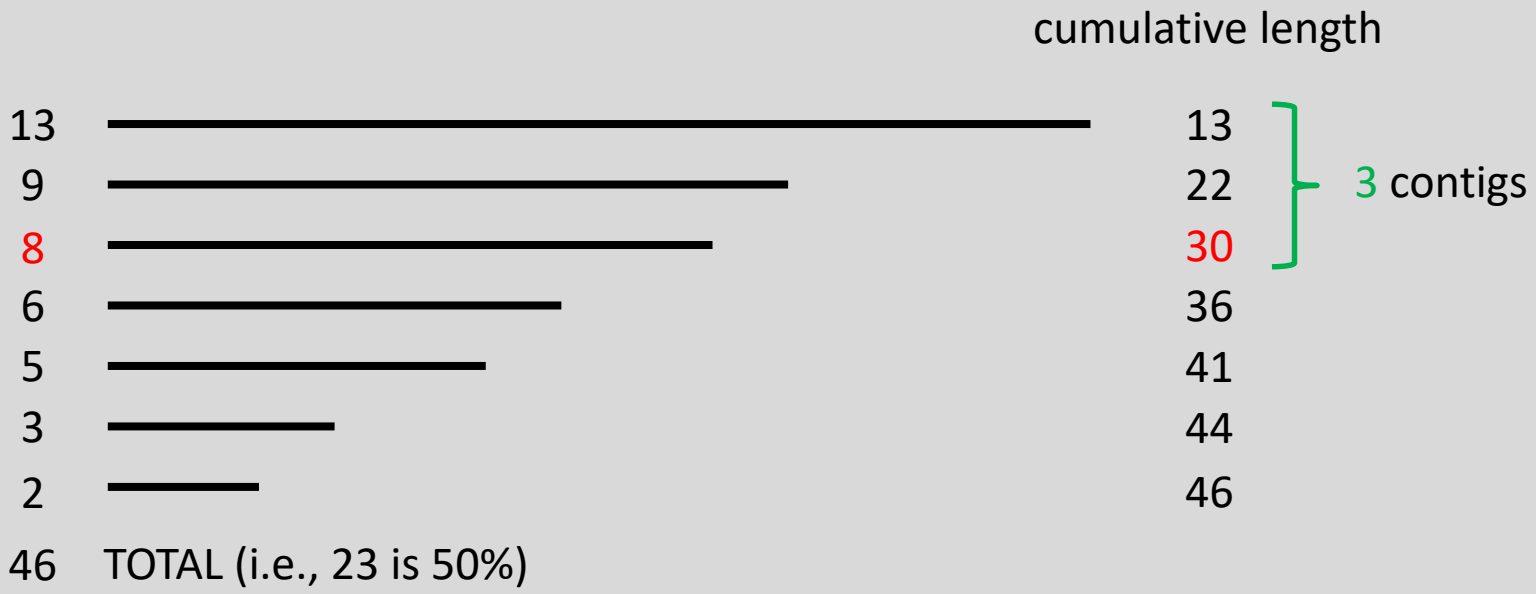
De novo assembly

N50

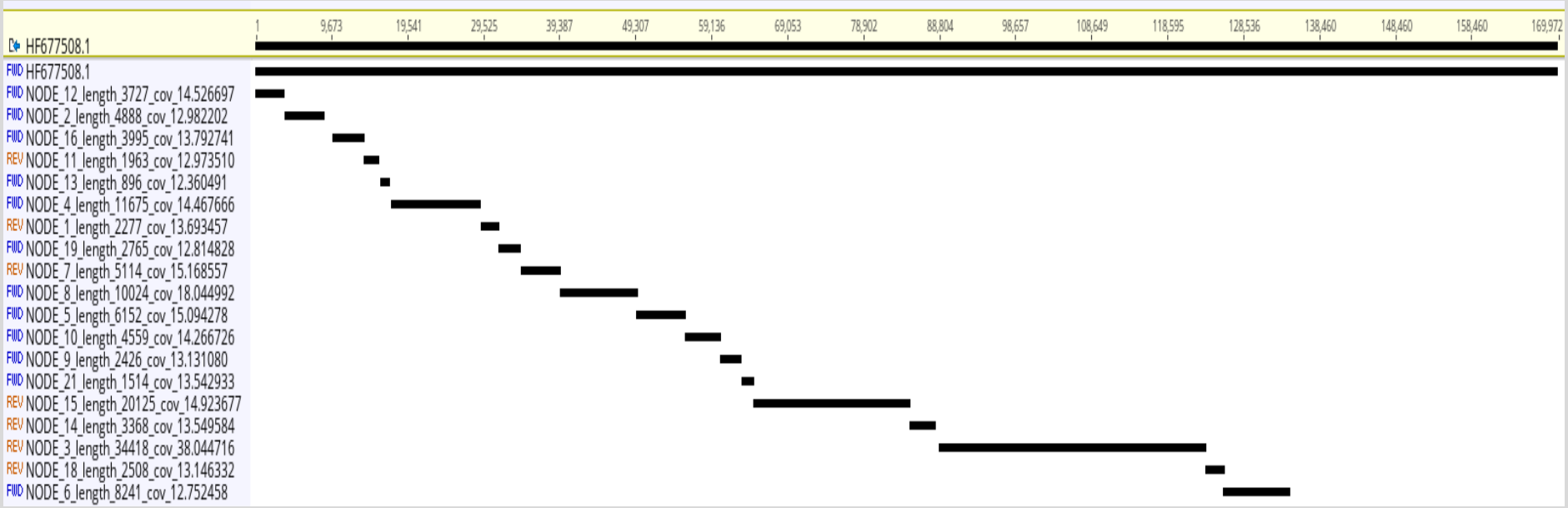
- assembly quality in terms of contiguity
- size of the contig which, along with the larger contigs, contain 50% of the total assembly length

L50

- smallest number of contigs whose length sum makes up 50% of the total assembly length



De novo contigs mapped to reference



- 19 contigs mapped to plastome reference
- nearly complete plastome was assembled
- all IR-derived reads probably assembled a single repeat

Literature

Danecek P. et al. (2021): Twelve years of SAMtools and BCFtools. GigaScience 10(2): giab008

Miller J.R. et al. (2010): Assembly algorithms for next-generation sequencing data. Genomics 95(6): 315–327.

Compeau P.E. et al. (2011): How to apply de Bruijn graphs to genome assembly. Nature Biotechnology 29(11): 987–991.