NHMW Workshop IV

De novo genome assembly

Martin Kapun & Andreas Kroh

Program for today

NHMW Workshop IV

Concepts

- The basics of NGS
- Types of assembly algorithms
- Assembly workflow
- Quality control

Program for today

Concepts

- The basics of NGS
- Types of assembly algorithms
- Assembly workflow
- Quality control

Hands-on

- Raw data QC
- Estimating the genome size
- Assembly with SPAdes and Flye
- Assembly QC

Genome Assembly

A genome assembly is an attempt to accurately represent an entire genome sequence from a large set of short DNA sequences.

.

Source: Keith Bradnam

Genome Assembly

A genome assembly is **an attempt** to accurately represent an entire genome sequence from a large set of short DNA sequences.

.

Source: Keith Bradnam

Genome Assembly

A genome assembly is an attempt to accurately **represent an entire genome sequence** from a large set of short DNA sequences.

.

Source: Keith Bradnam

Genome Assembly

A genome assembly is an attempt to accurately represent an entire genome **sequence from a large set** of short DNA sequences.

.

Source: Keith Bradnam

Genome Assembly

A genome assembly is an attempt to accurately represent an entire genome sequence from a large set of **short** DNA sequences.

.

Source: Keith Bradnam

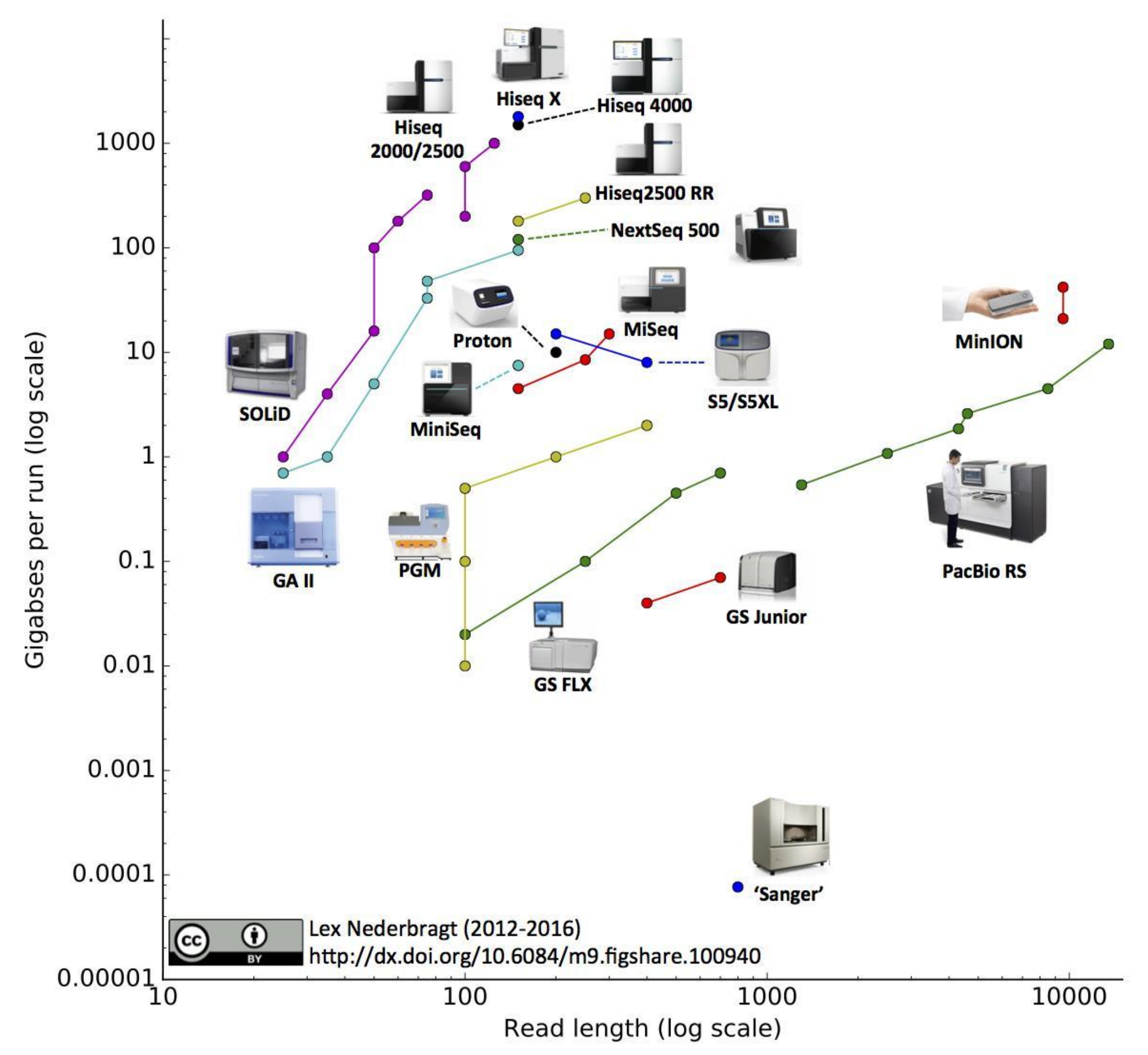
Genome Assembly

A genome assembly is an attempt to accurately represent an entire genome sequence from a large set of short DNA sequences.

It's a bit like trying to do the hardest jigsaw puzzle you can imagine!

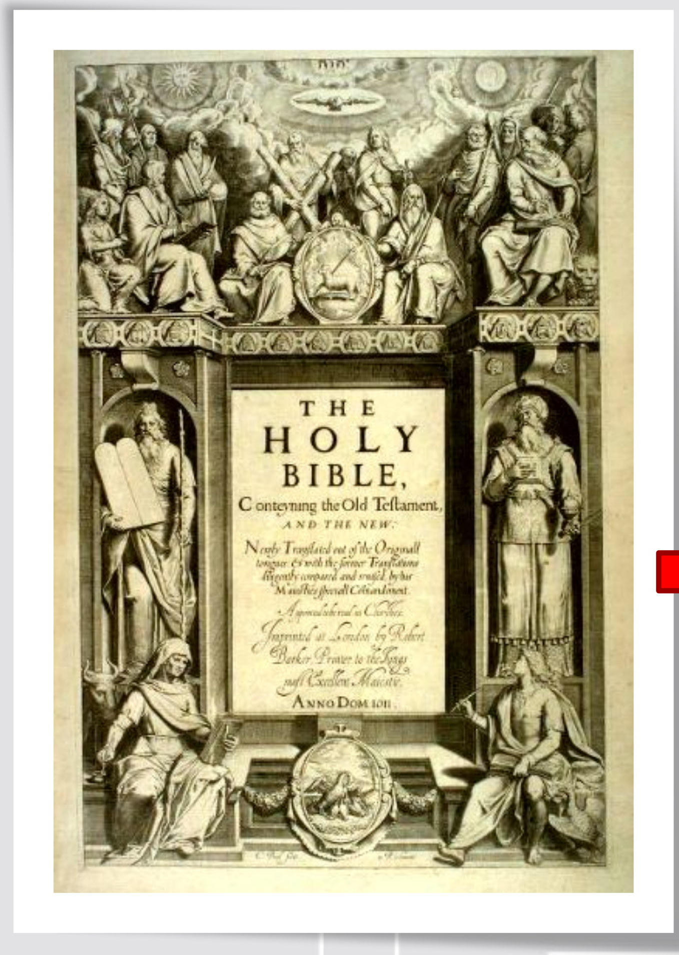
Source: Keith Bradnam

New Sequencing technologies

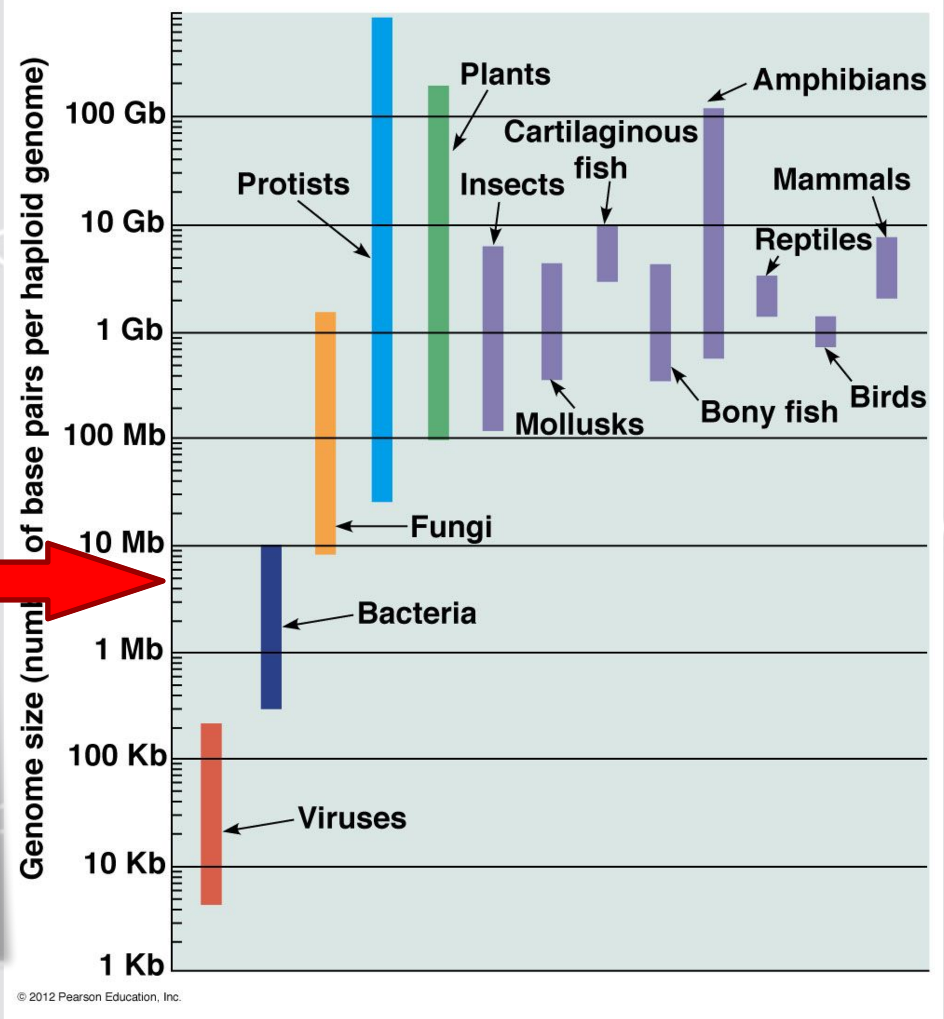


Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

Genome-size variation



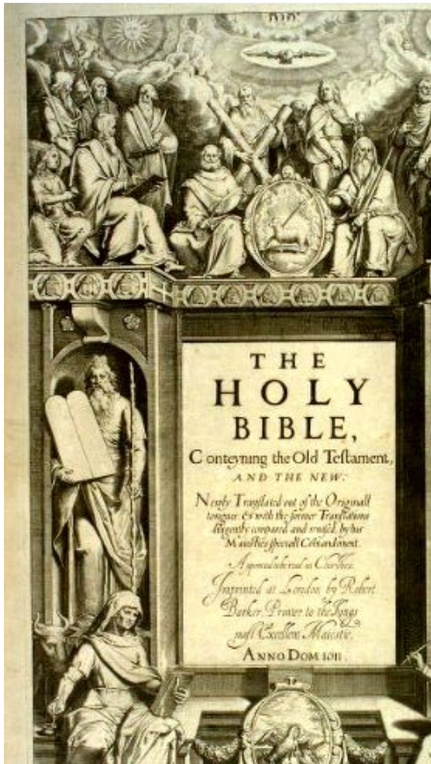
THE
HOLY
BIBLE,
Conteyning the Old Teſtament,
AND THE NEW:
Newly Tranſlated out of the Originall
tongues: & with the former Tranſlations
diligently compared and reviſed, by his
Maieſties ſpeciall Comandement.
Appointed to be read in Churches.
Imprinted at London by Robert
Barker, Printer to the Kings
moſt Excellent Maieſtie.
ANNO DOM. 1611.



© 2012 Pearson Education, Inc.

Next Generation Sequencing

The idea behind it



THE HOLY BIBLE,

Conteyning the Old Teſtament, AND THE NEW:

Newly Tranſlated out of the Originall tongues: & with the former Tranſlations diligently compared and reuiſed, by his Maieſties ſpeciall Cōmandement.

Appointed to be read in Churches.

Imprinted at London by Robert Barker, Printer to the Kings moſt Excellent Maieſtie.

ANNO DOM 1611.

The Holy Bible: *c.* 3.5 million charact

Next Generation Sequencing

The idea behind it

The Holy Bible: *c.* 3.5 million characters (excl. spaces) x 500,000 (!)

Next Generation Sequencing

The idea behind it

Next Generation Sequencing

The idea behind it

GAATTCTTCAGGTAGCTTCCTAGGGTTTCCAAGGCAATACAA

AGGTAGCTTCCTAGGGTTTCCAAGGCAATACAAGAAGAATTTT

TTCTTCAGGTAGCTTCCTAGGGTTTCCAAGG

Problem:

- numerous fragments of the same genomic region exist

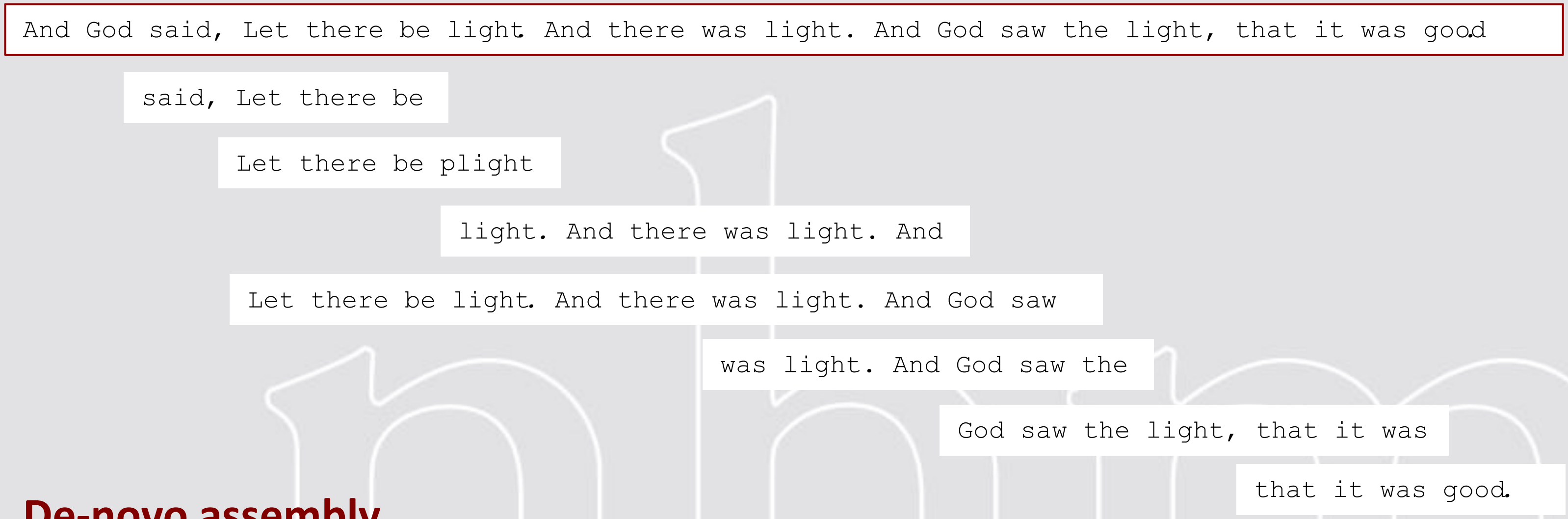
Benefit:

- numerous fragments of the same genomic region exist

What to do with these data?

The idea behind it

Reference guided assembly (read mapping/resequencing)



And God said, Let there be light. And there was light. And God saw the light, that it was good.

said, Let there be

Let there be plight

light. And there was light. And

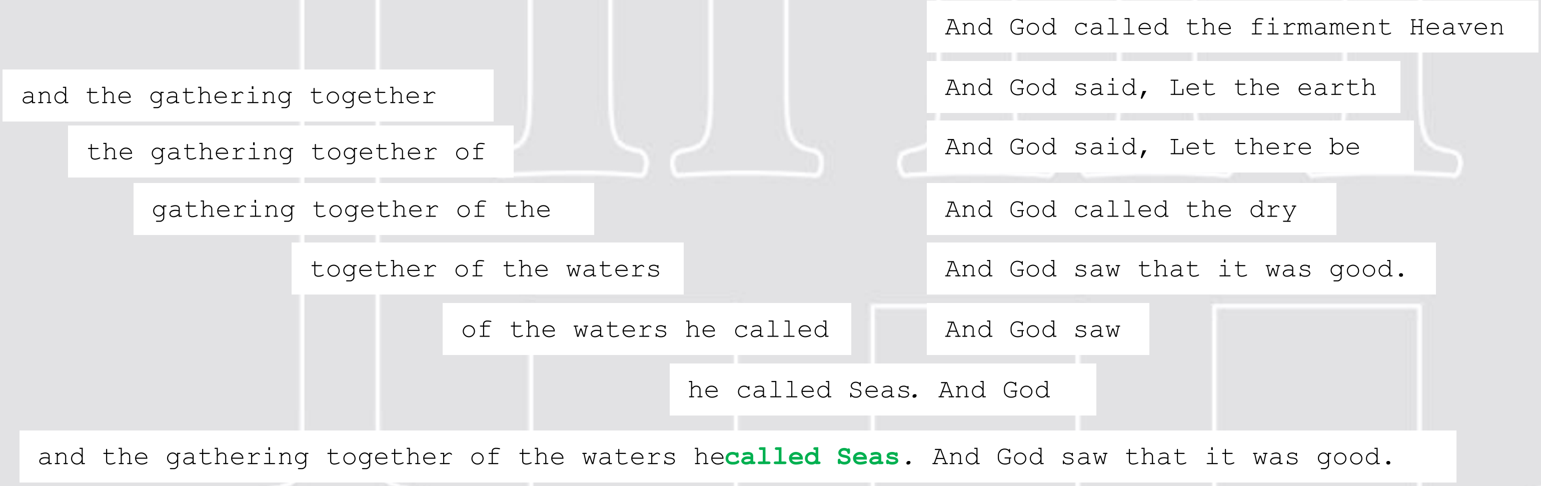
Let there be light. And there was light. And God saw

was light. And God saw the

God saw the light, that it was

that it was good.

De-novo assembly



and the gathering together

the gathering together of

gathering together of the

together of the waters

of the waters he called

he called Seas. And God

And God called the firmament Heaven

And God said, Let the earth

And God said, Let there be

And God called the dry

And God saw that it was good.

And God saw

and the gathering together of the waters he**called Seas**. And God saw that it was good.

Read types

Single-End vs. Paired-End

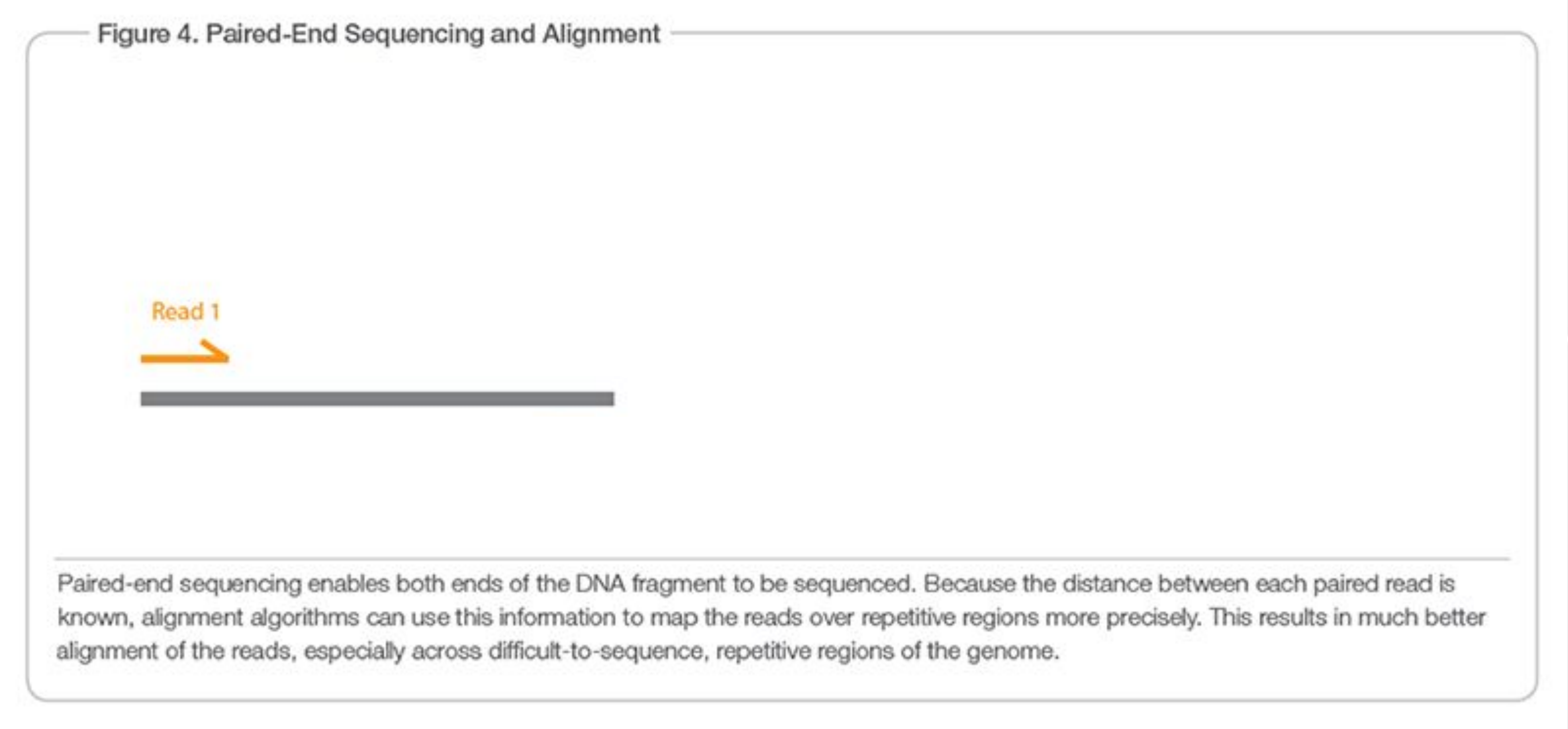
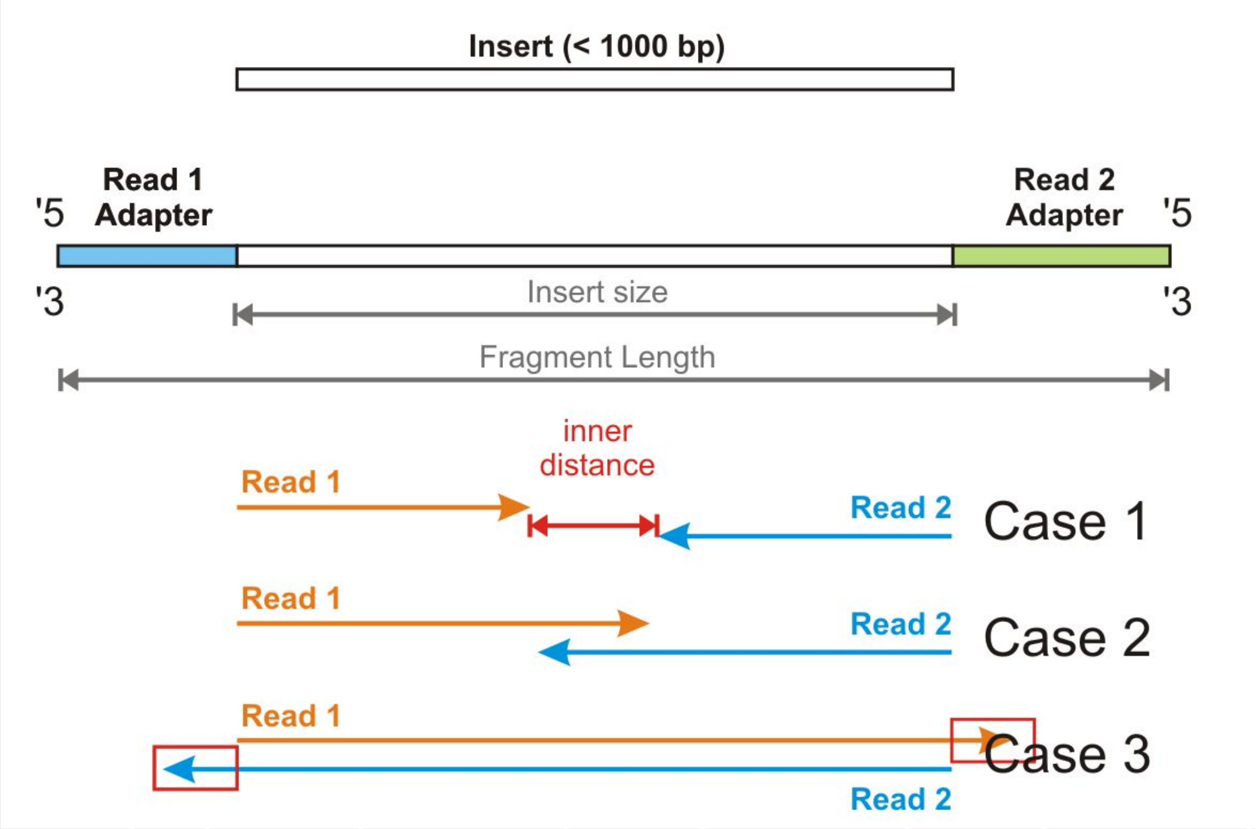


Figure 4. Paired-End Sequencing and Alignment

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Read types

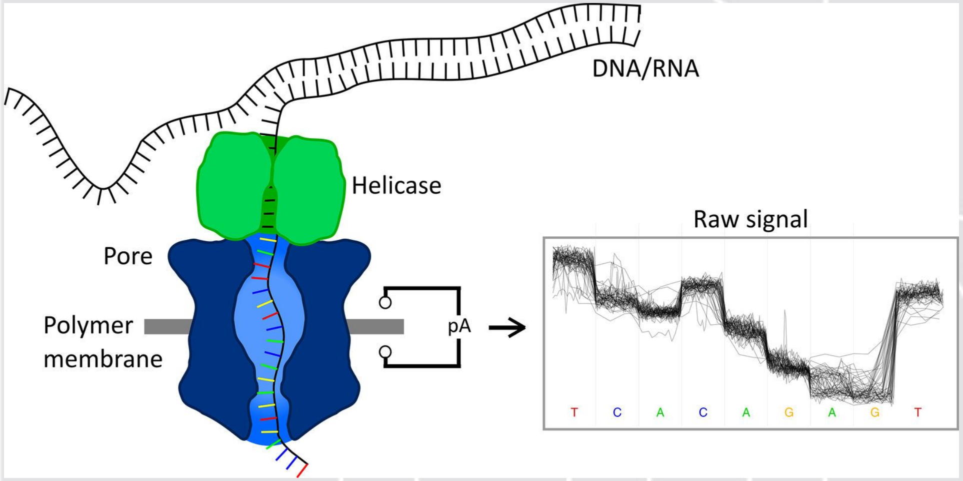
Single-End vs. Paired-End



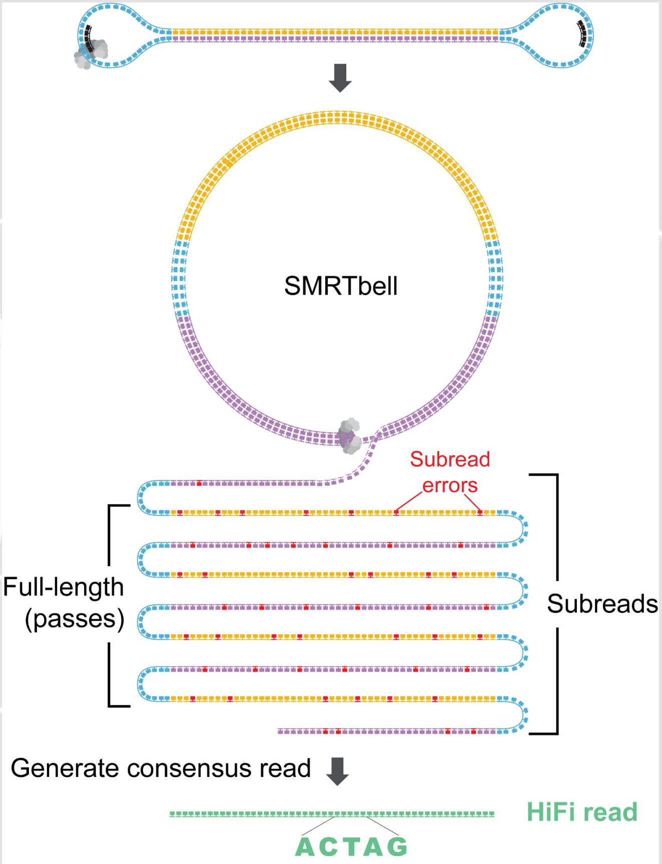
Read types

Single-End vs. Paired-End vs. Single Molecule

Oxford Nanopore



Pacific Biosciences

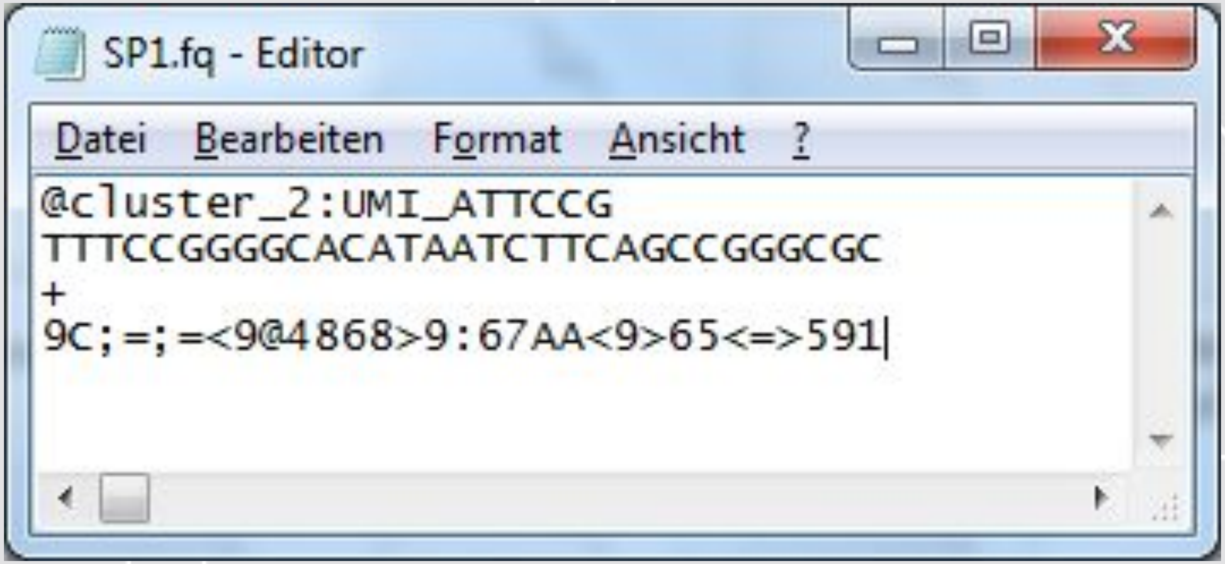


Next Generation Sequencing

Read technologies

Raw data

- (FASTA)
- FASTQ

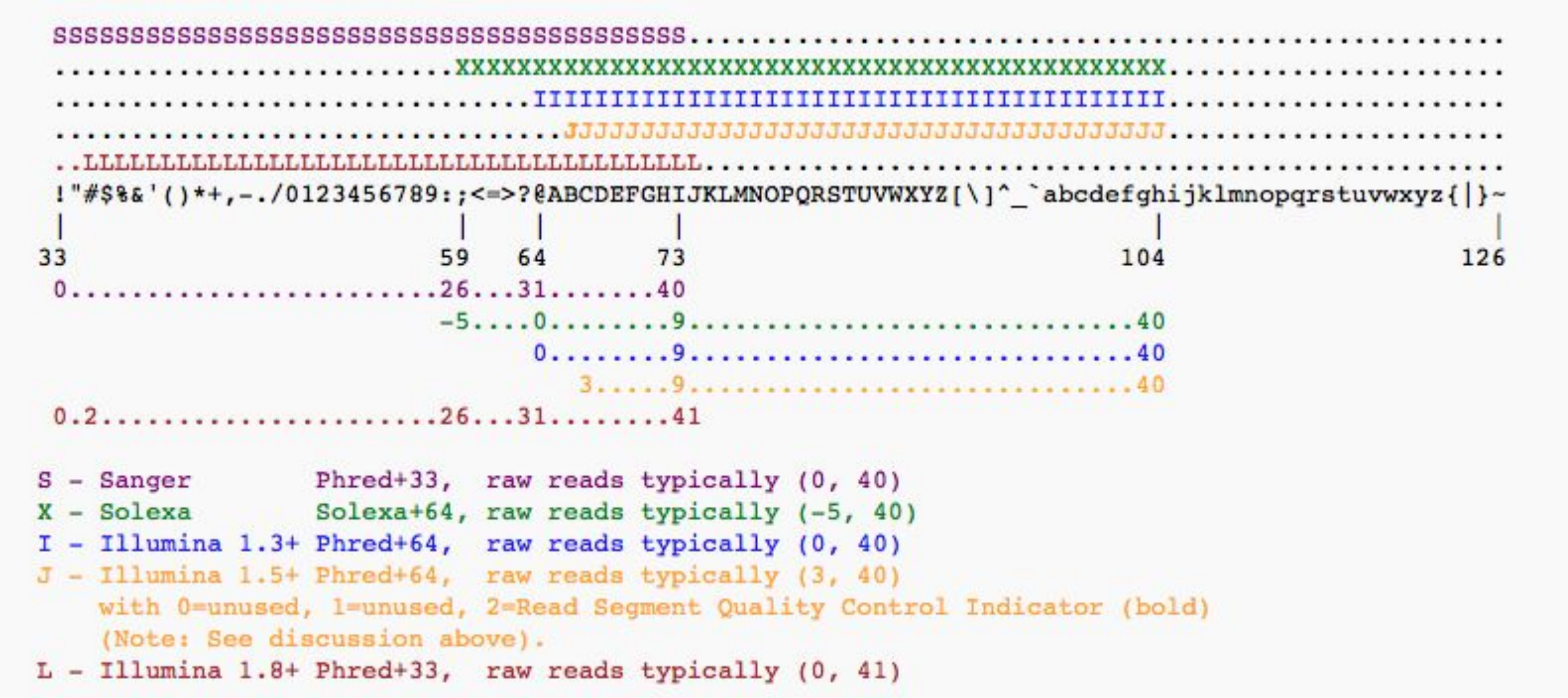


```
@cluster_2:UMI_ATTCCG
TTTCCGGGGCACATAATCTTCAGCCGGGCGC
+
9C;=;=<9@4868>9:67AA<9>65<=>591
```

Next Generation Sequencing

Read Quality

Phred Quality Scores



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Next Generation Sequencing

Read Quality

Phred Quality Scores

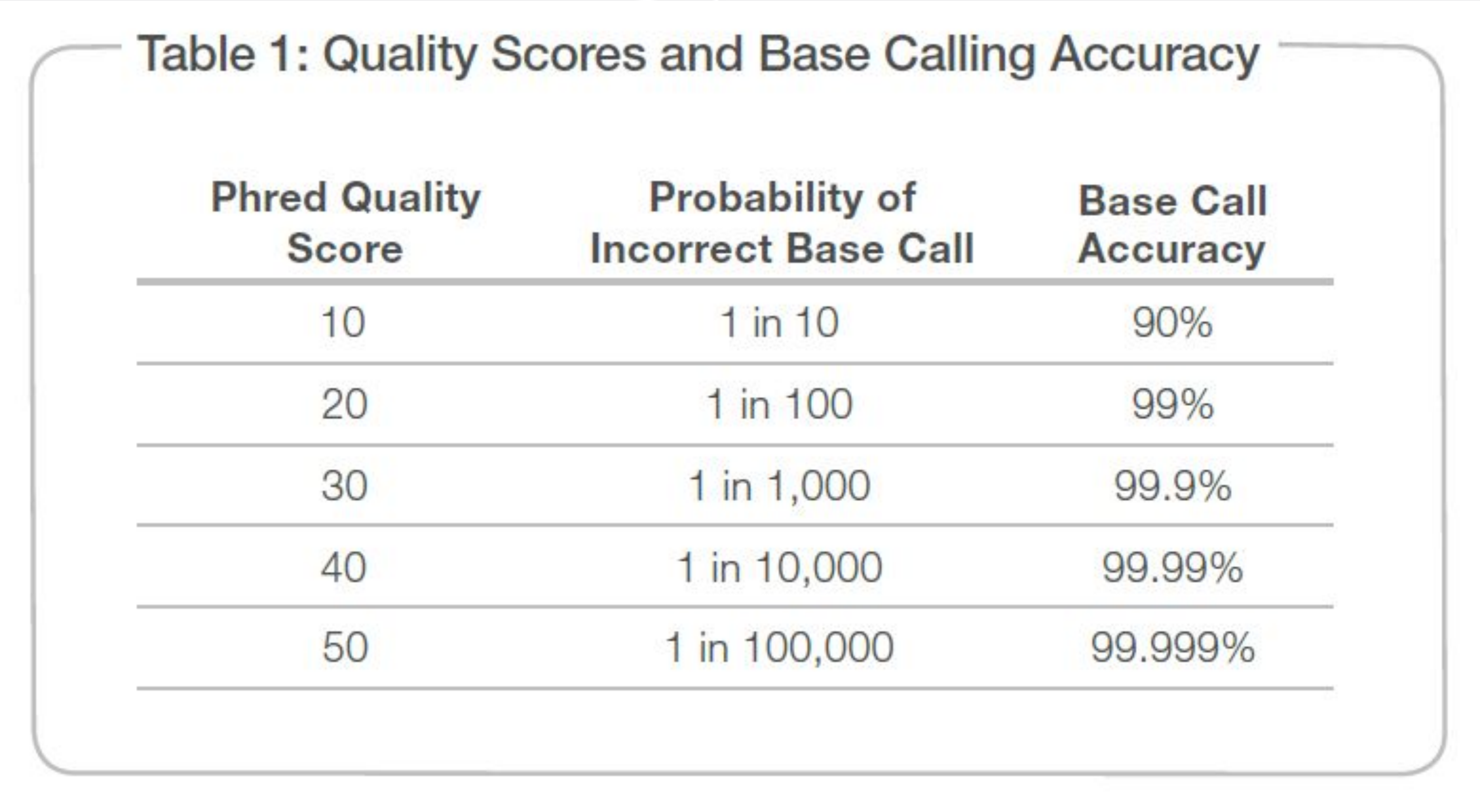


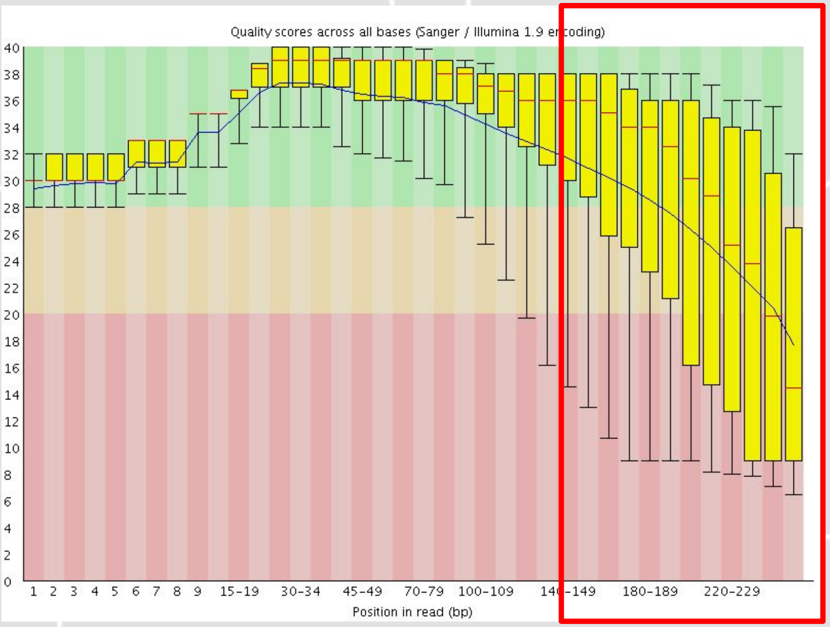
Table 1: Quality Scores and Base Calling Accuracy

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%

Data Quality

Short reads

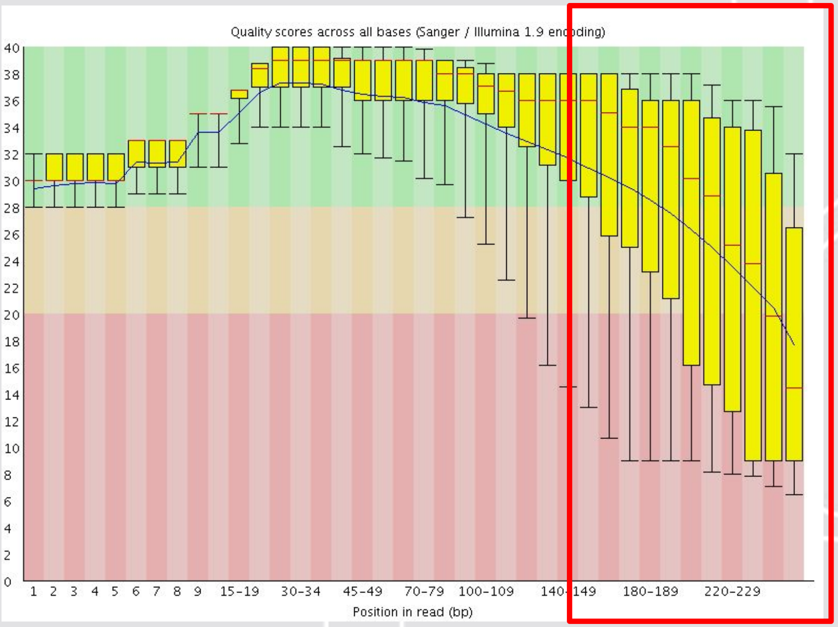
150 bp □ 250 bp

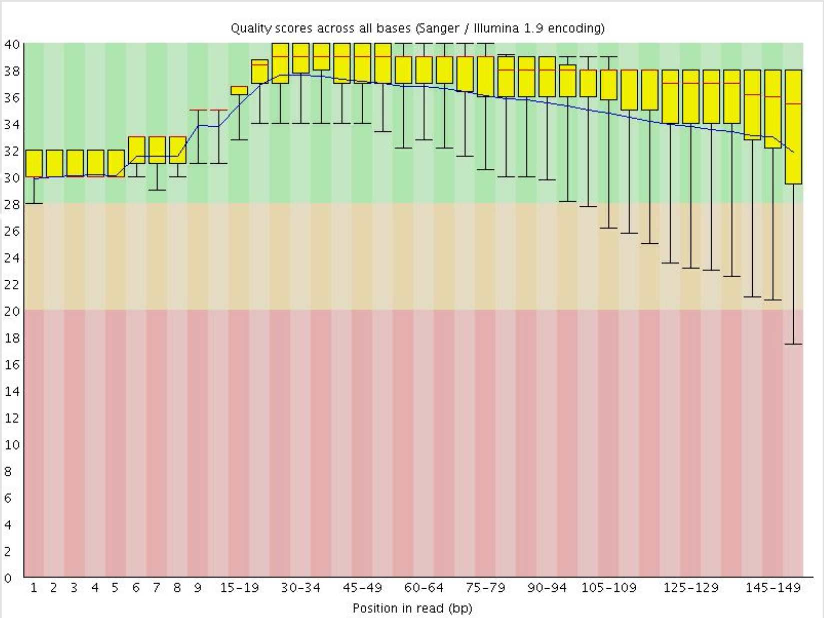


Data Quality

Short reads

Quality trimming (removal of low qual bases; adapters)

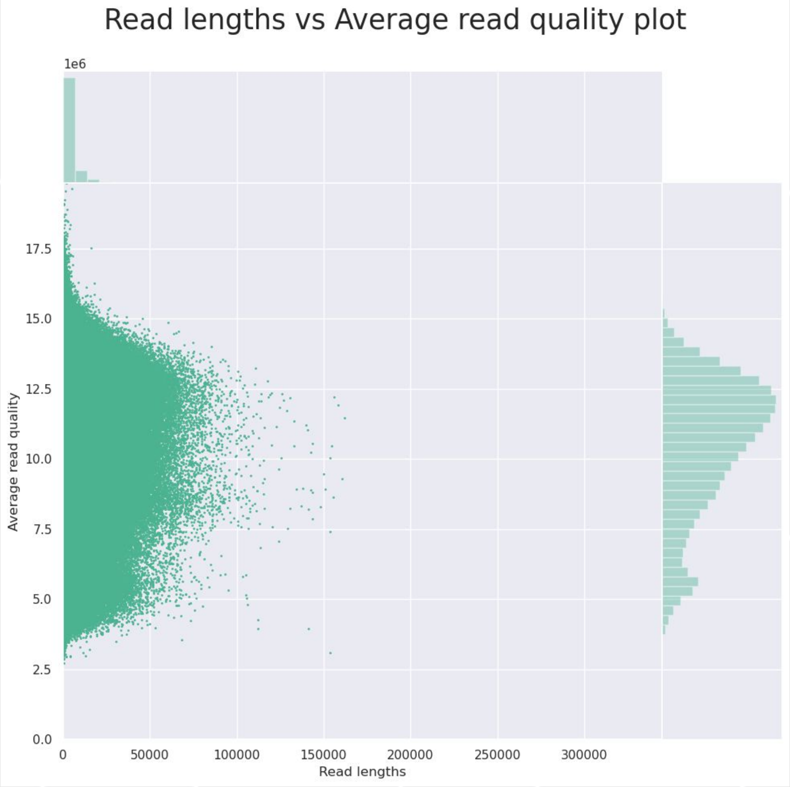




Data Quality

Long reads

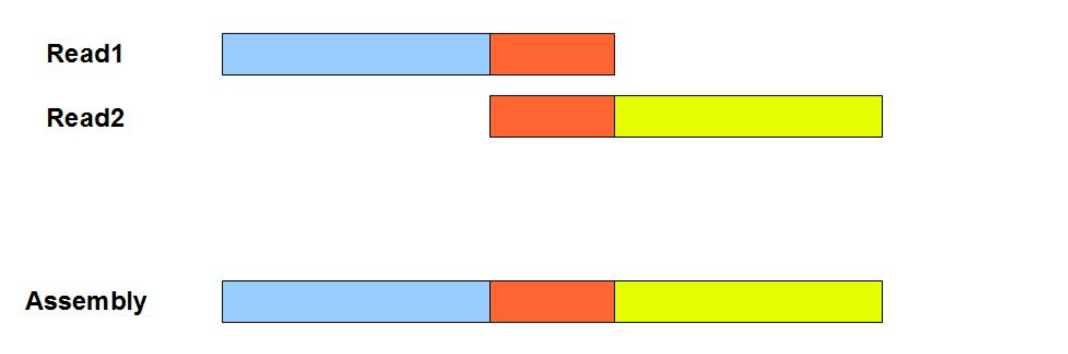
< 200,000bp



Assembly algorithms

(1) Overlap-consensus-layout (OCL) method

- e.g. *Flye* assembler (long reads)



and the gathering together

the gathering together of

gathering together of the

together of the waters

of the waters he called

he called Seas. And God

and the gathering together of the waters he**called Seas.** And God

And God called the firmament Heaven

And God said, Let the earth

And God said, Let there be

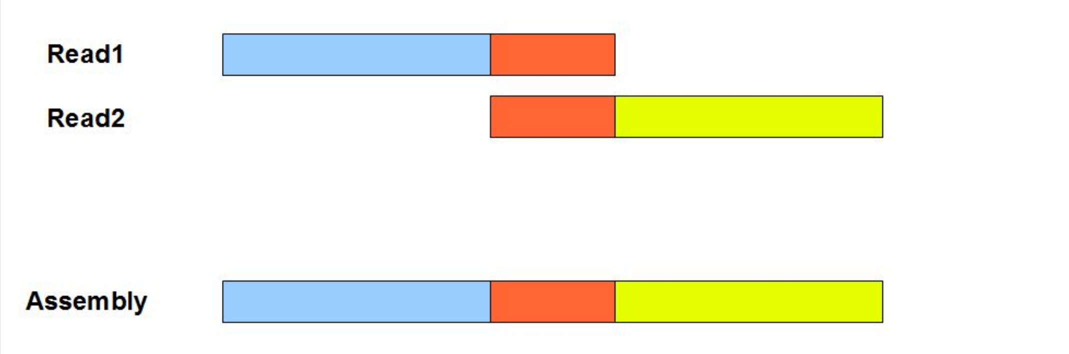
And God called the dry

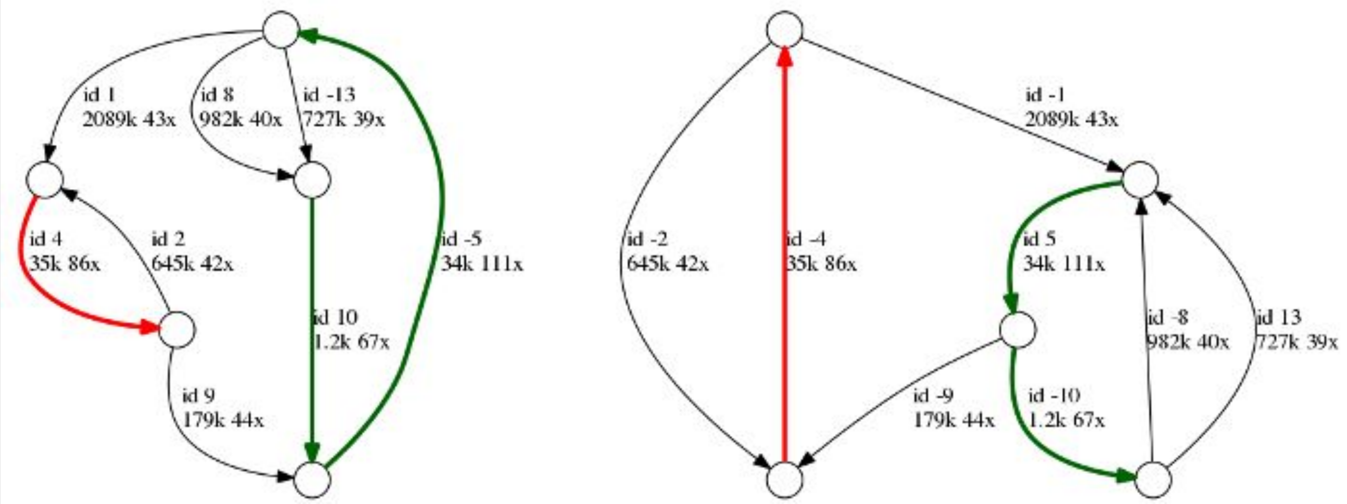
And God saw that it was good.

Assembly algorithms

(1) Overlap-consensus-layout (OCL) method

- e.g. *Flye* assembler (long reads)





https://github.com/fenderglass/Flye/blob/flye/docs/graph_example.png

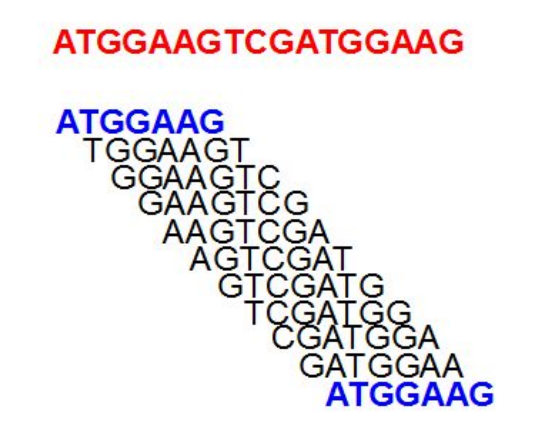
Assembly algorithms

(2) de-Bruijn-graph based method

- e.g. *SPAdes* assembler (short reads)

K-mer

The term k-mer refers to all the possible substrings of length k that are contained in a string



ATGGAAGTCGATGGAAG

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGA
     AGTCGAT
      GTCGATG
       TCGATGG
        CGATGGA
         GATGGAA
          ATGGAAG
```

# of overlapping k-mers = ***s - k + 1***

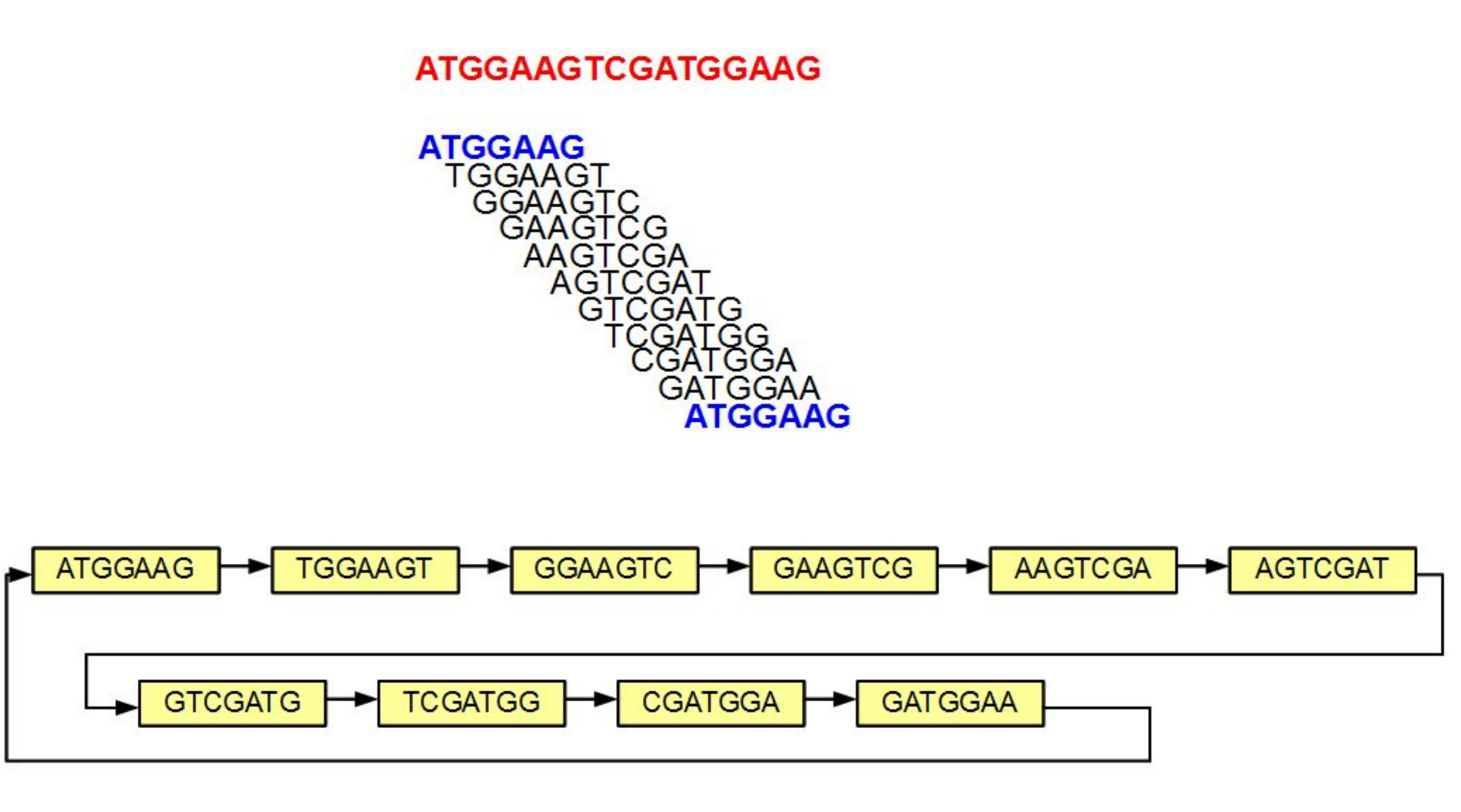
s = sequence length
k = k-mer length

NHMW Workshop IV

Assembly algorithms

(2) de-Bruijn-graph based method

- e.g. *SPAdes* assembler (short reads)



ATGGAAGTCGATGGAAG

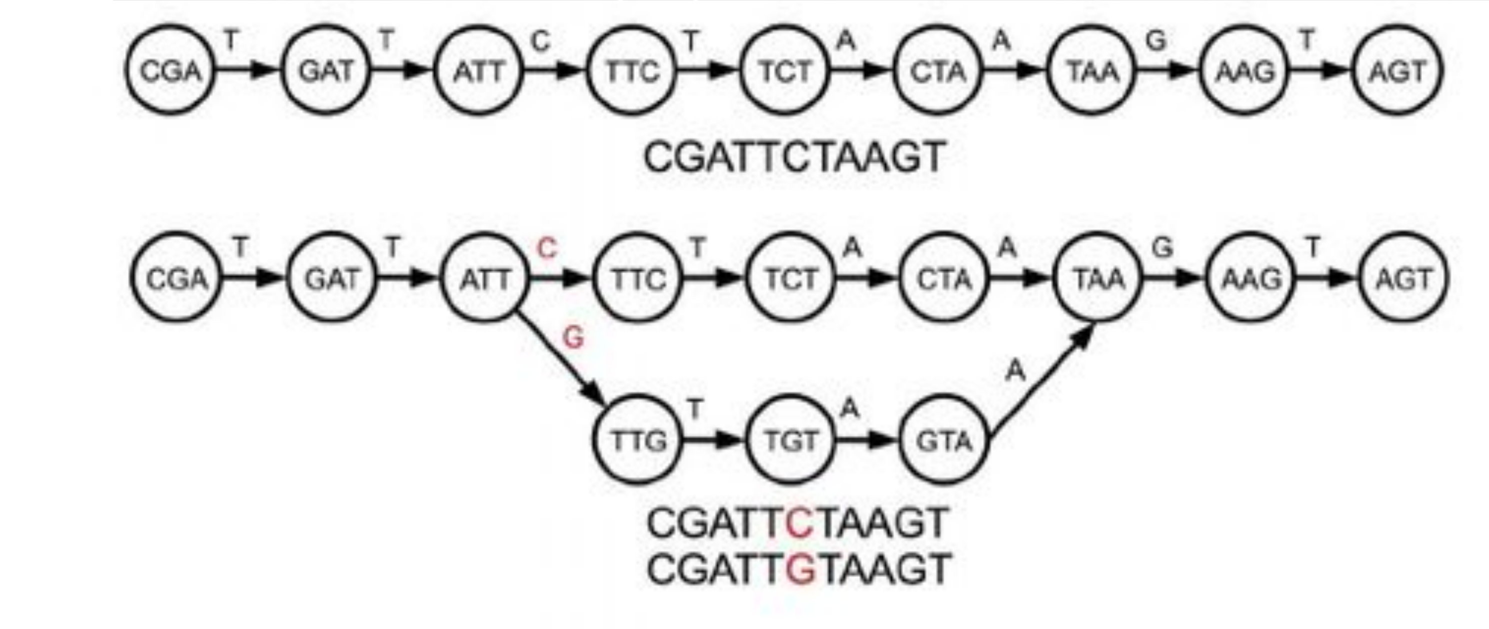
ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGA
AGTCGAT
GTCGATG
TCGATGG
CGATGGA
GATGGAA
ATGGAAG

ATGGAAG → TGGAAGT → GGAAGTC → GAAGTCG → AAGTCGA → AGTCGAT → GTCGATG → TCGATGG → CGATGGA → GATGGAA → ATGGAAG

Assembly algorithms

(2) de-Bruijn-graph based method

- e.g. *SPAdes* assembler (short reads)



Assembly algorithms

(2) de-Bruijn-graph based method

- e.g. *SPAdes* assembler (short reads)

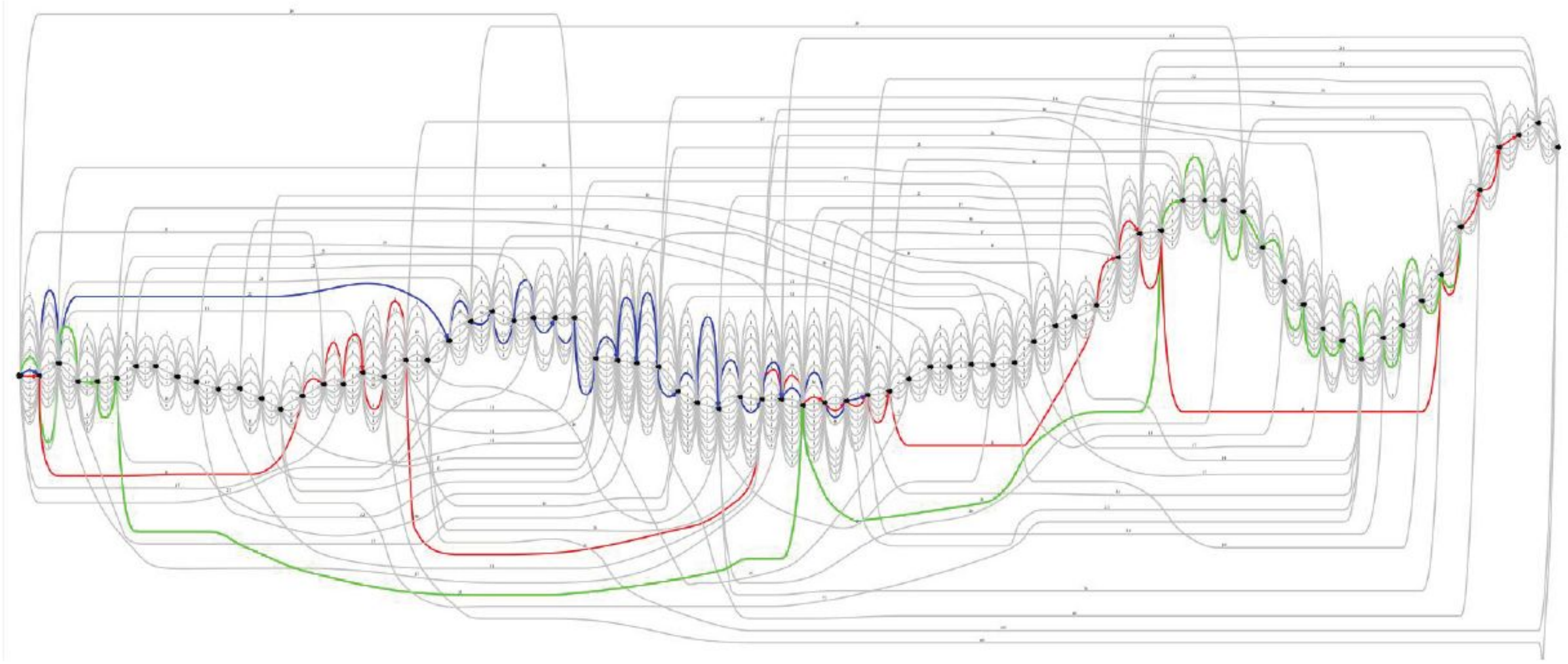
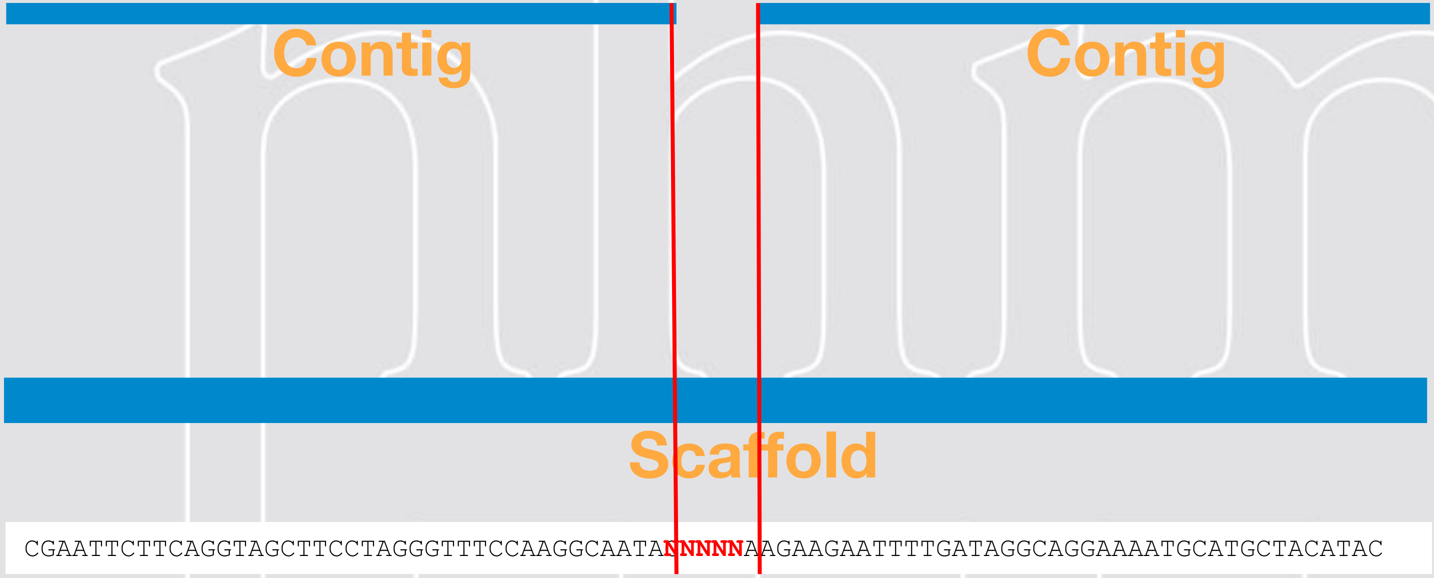


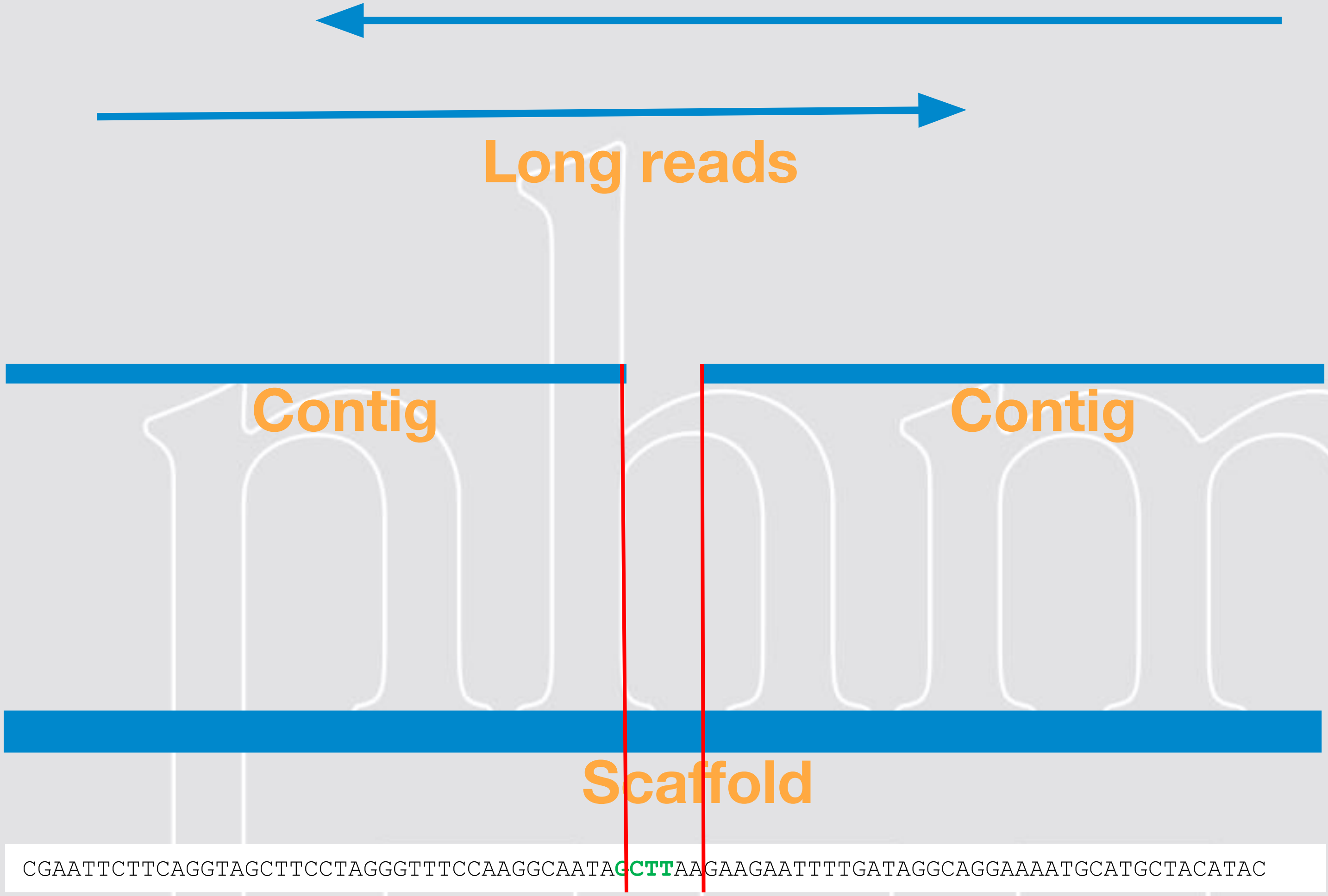
Fig. S1. A small subgraph of the A-Bruijn graph constructed from 76 (15,8)-mers appearing in segments of 55 reads covering a short 100-nucleotide region (starting at position 2,100,000 in *E. coli* genome). Three out of 55 read-paths are highlighted in blue, red, and green.

Assembled sequences



CGAATTCTTCAGGTAGCTTCCTAGGGTTTCCAAGGCAATANNNNNAAGAAGAATTTTGATAGGCAGGAAAATGCATGCTACATAC

Assembled sequences


Long reads
Contig
Contig
Scaffold
CGAATTCTTCAGGTAGCTTCCTAGGGTTTCCAAGGCAATA**GCTT**AAGAAGAATTTTGATAGGCAGGAAAATGCATGCTACATAC

K-mers revisited

A relationship between k-mers and genome-size

For sequences > 1mb:

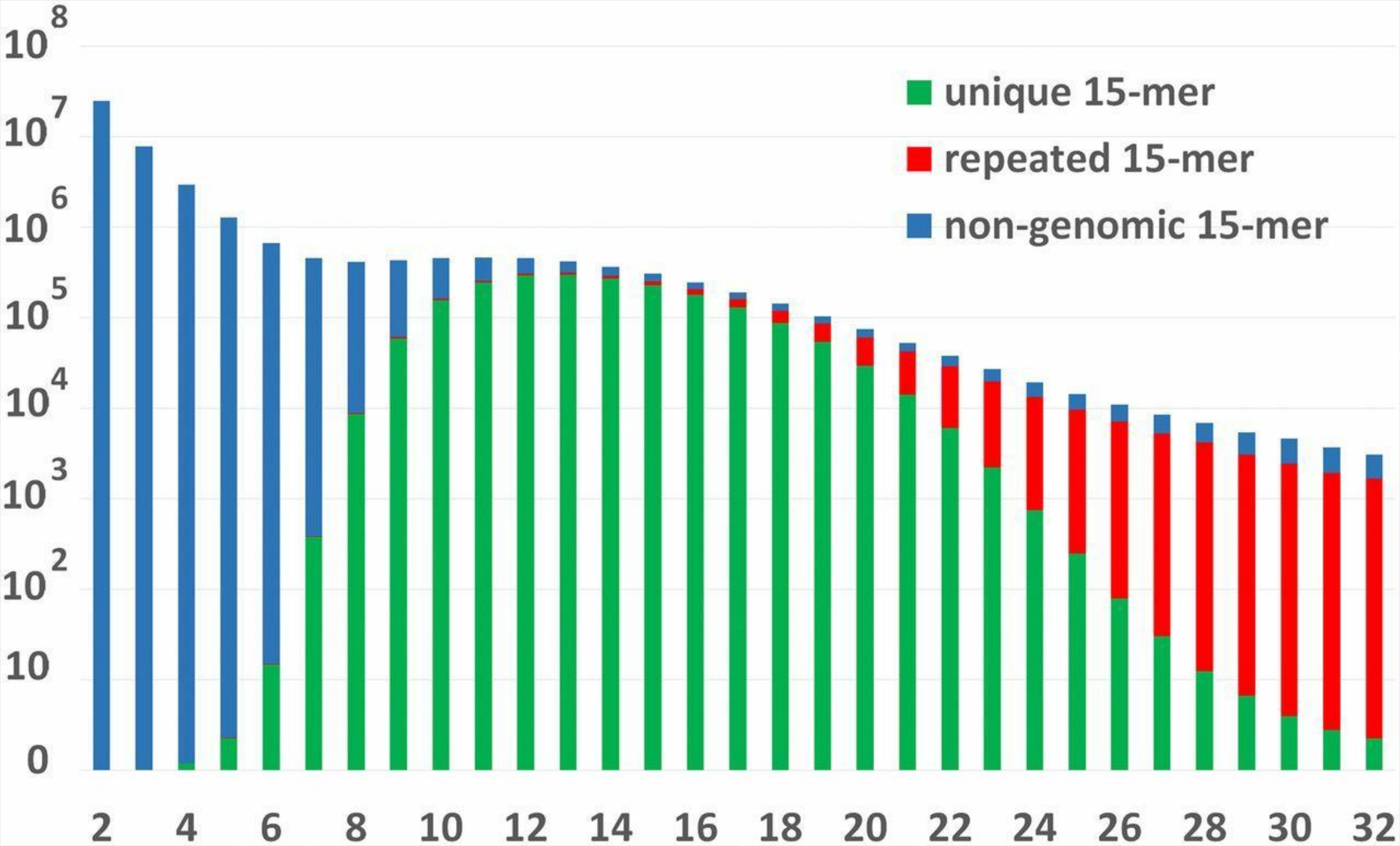
# unique k-mers ~ sequence length

Genomesize ~ #unique k-mers/coverage

BUT, Read Error, Repeats & Heterozygosity!!

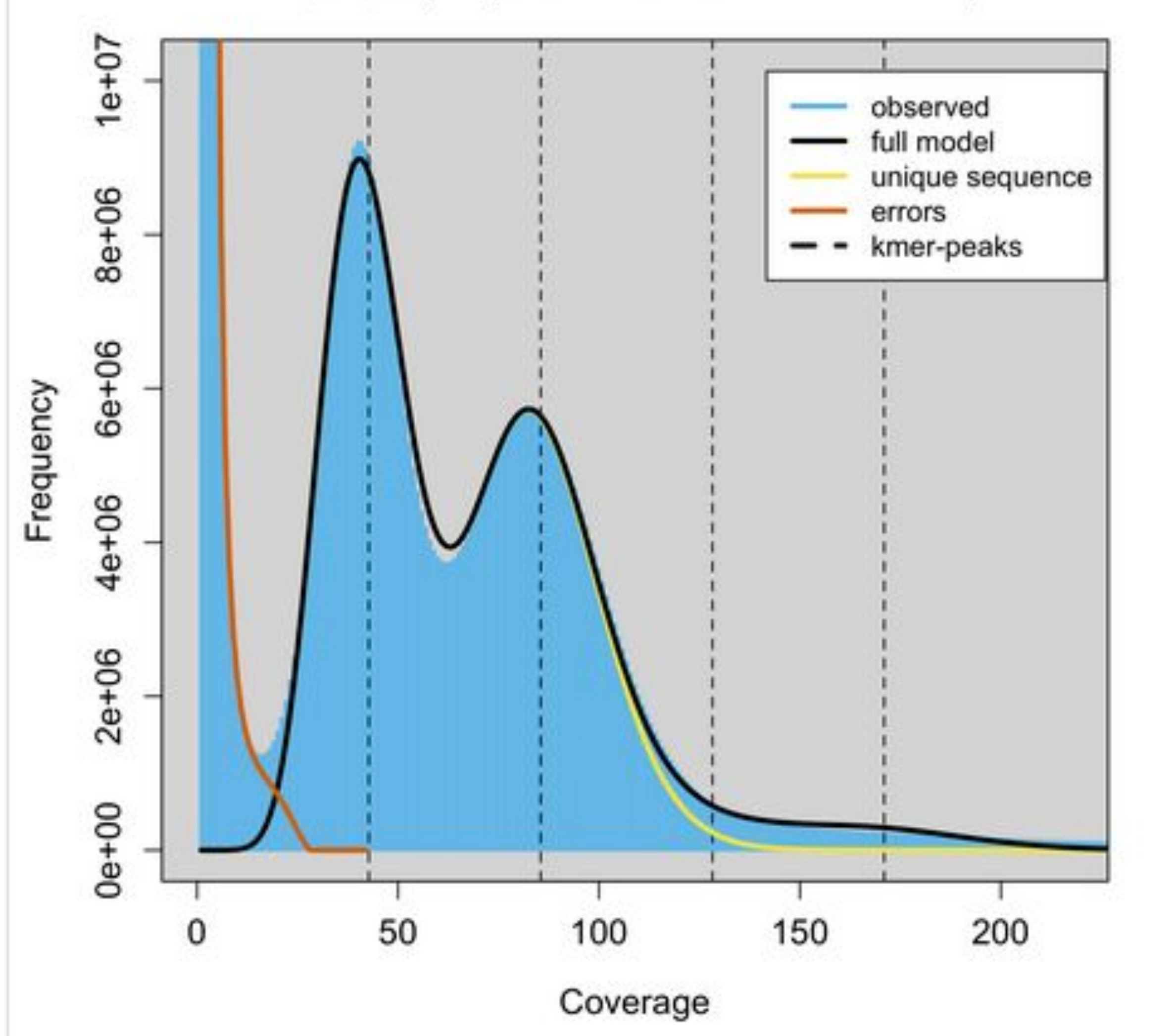
K-mers revisited

K-mer, Read Error, Repeats & Heterozygosity



K-mers revisited

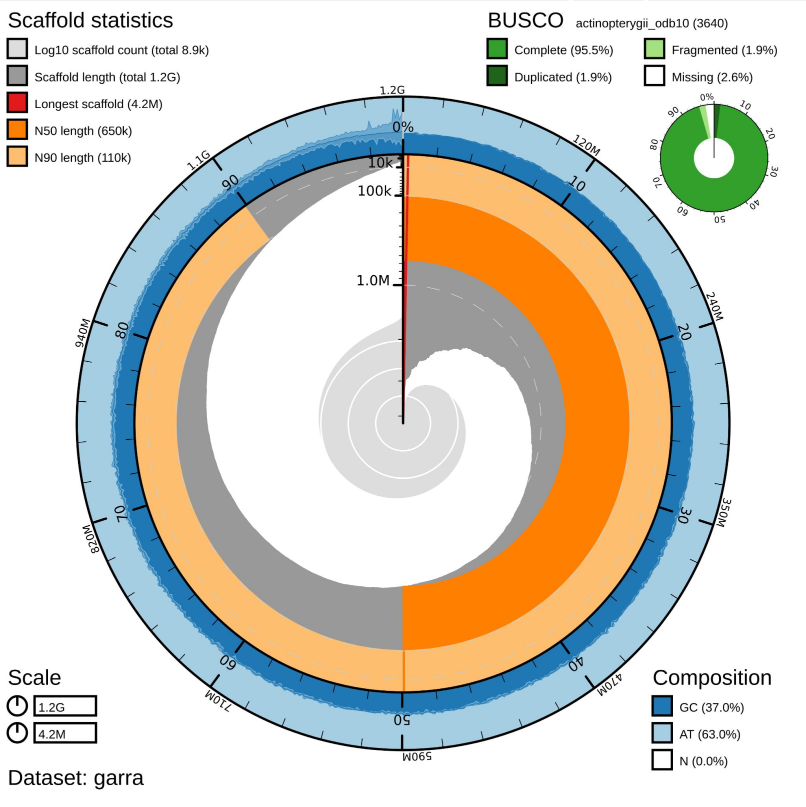
Model-based estimate of genome-size with *Genomescope*



Assembly QC metrics

Standard Metrics (*QUAST*):

- Longest Contig; longest Scaffold; N50; #contigs/scaffolds; (threshold)

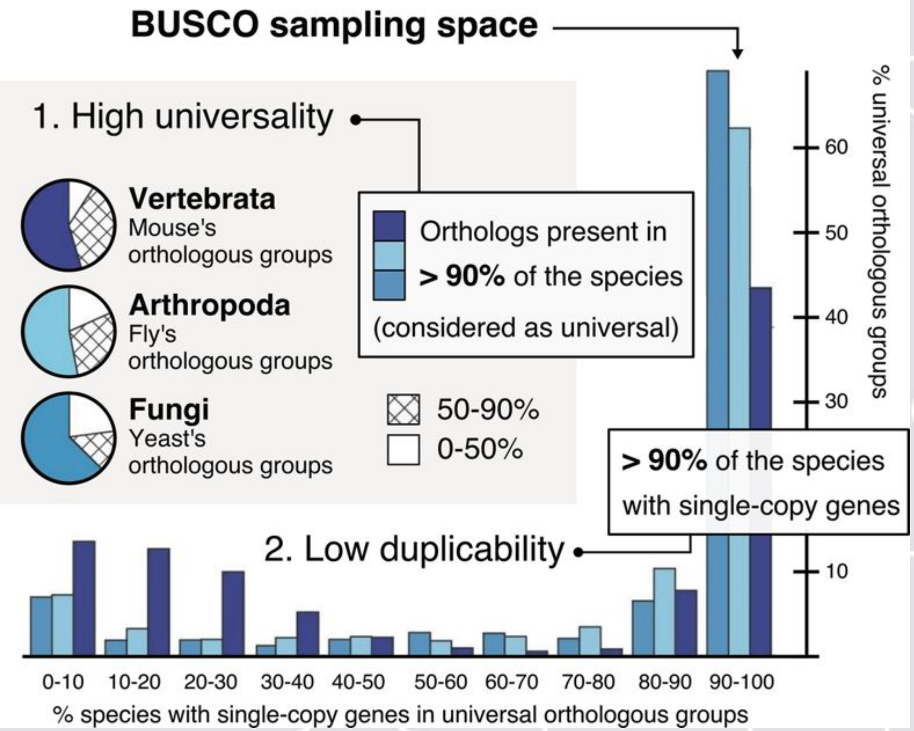


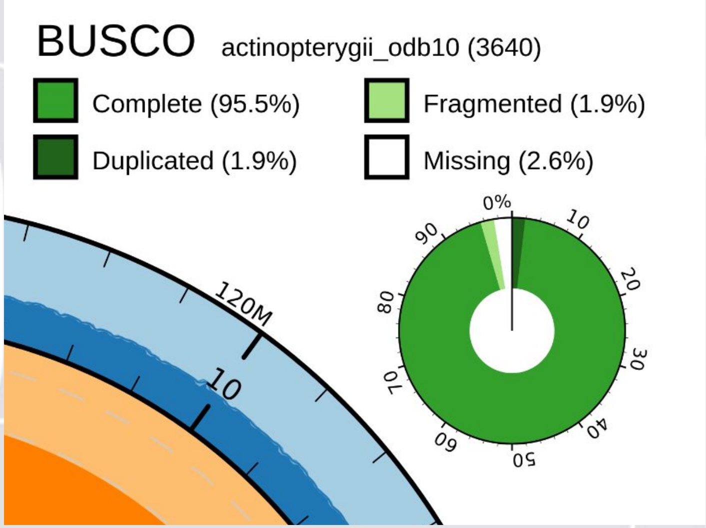
- light/dark blue: AT/GC content [%]
- light orange: N90 [y-axis: length]
- dark orange: N50 [y-axis: length]
- dark grey: [y-axis: scaffold lengths]
- light grey: [count of scaffolds]

Assembly QC metrics

Benchmarking Universal Single-Copy Orthologue (BUSCO)

- Estimate assembly quality based on presence of conserved single copy genes.

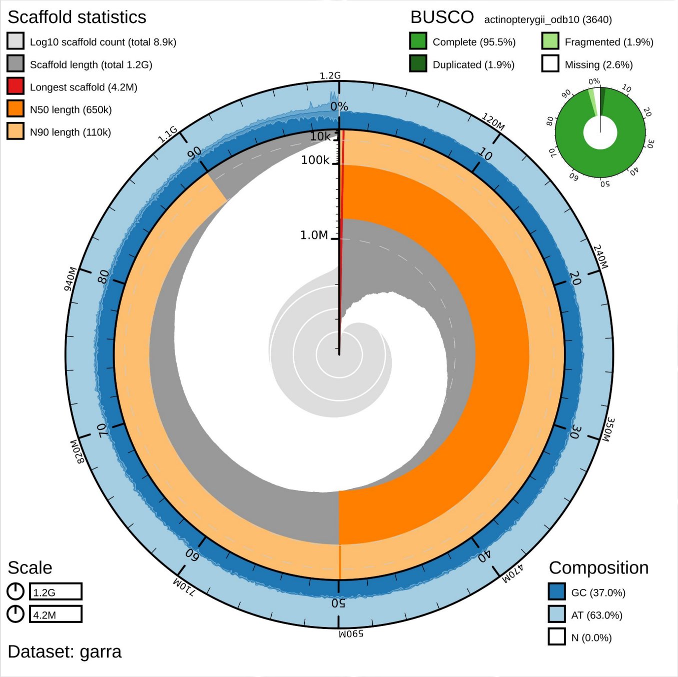


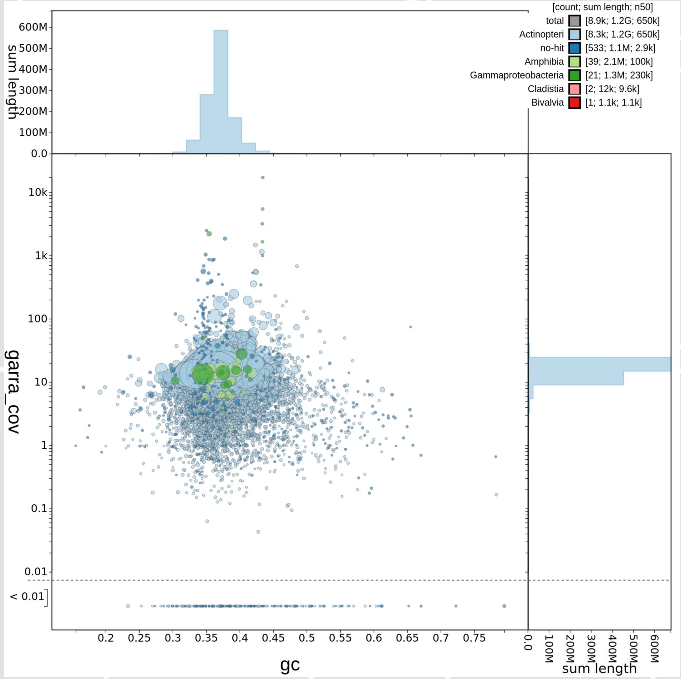


Assembly QC metrics

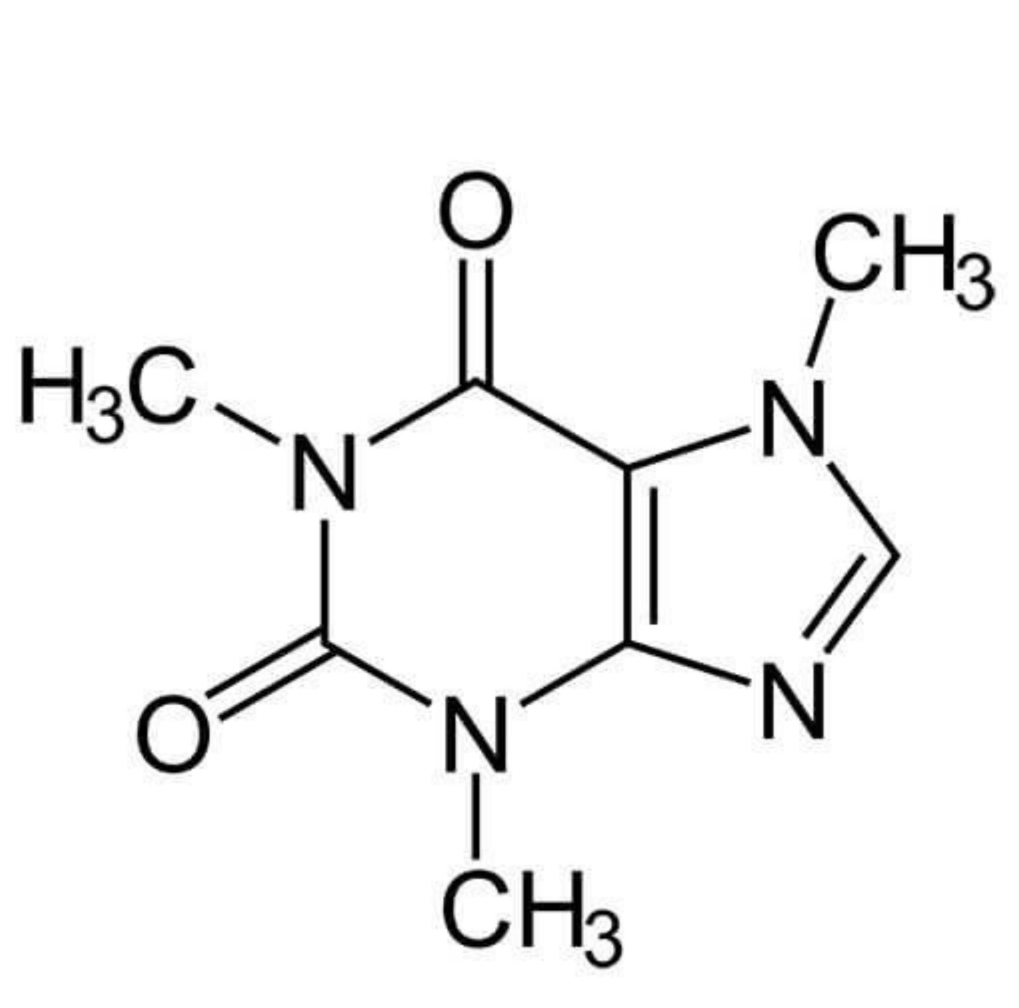
BLOBtools

- Combines various assembly-QC analyses
  - BLAST -> similarity of contigs to taxa
  - MAPPING -> coverage variation
  - BUSCO



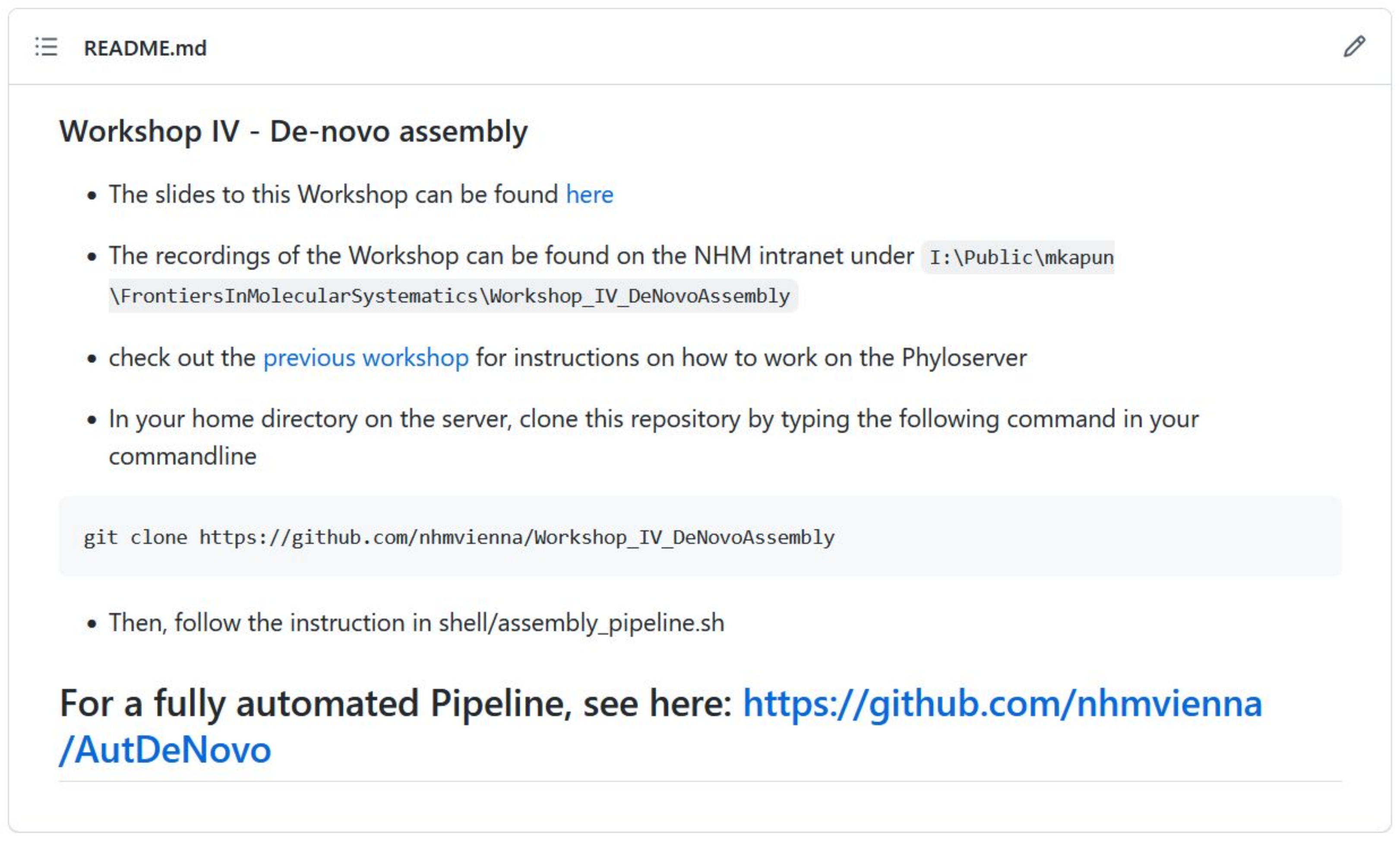


10’ Break!



Hands on…

https://github.com/nhmvienna/Workshop_IV_DeNovoAssembly



Workshop IV - De-novo assembly

- The slides to this Workshop can be found here
- The recordings of the Workshop can be found on the NHM intranet under `I:\Public\mkapun\FrontiersInMolecularSystematics\Workshop_IV_DeNovoAssembly`
- check out the previous workshop for instructions on how to work on the Phyloserver
- In your home directory on the server, clone this repository by typing the following command in your commandline

```
git clone https://github.com/nhmvienna/Workshop_IV_DeNovoAssembly
```

- Then, follow the instruction in shell/assembly_pipeline.sh

For a fully automated Pipeline, see here: https://github.com/nhmvienna/AutDeNovo

Thank you!!!

NHMW Workshop IV