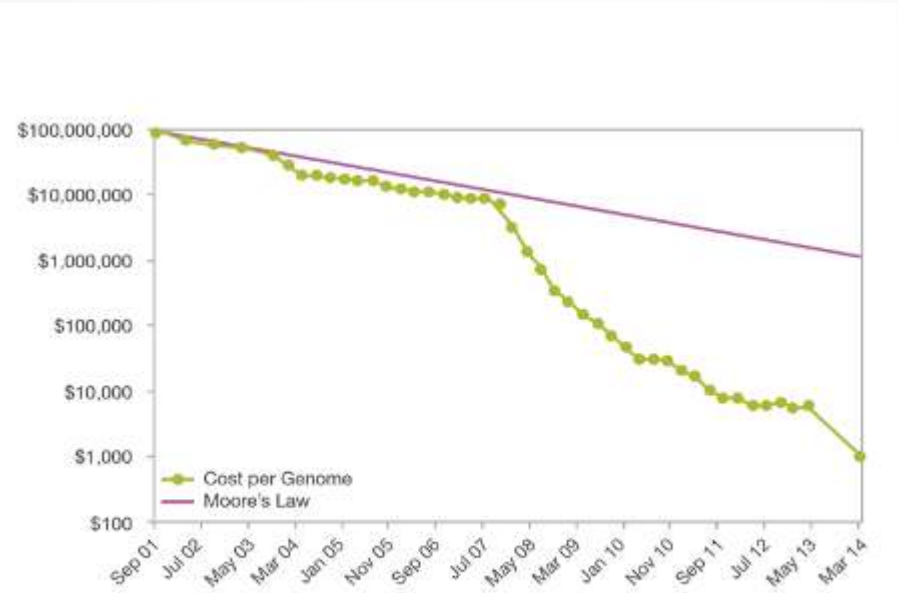
Next Generation Sequencing – NGS Genome Assembly

Pongor Lőrinc Sándor
II. sz. Gyermekgyógyászati Klinika
MTA-TTK Lendület Onkológiai Biomarker Kutatócsoport

Price of Genome Sequencing



Sanger Sequencing
>100,000,000 $ / Human Genome

~14 years

Illumina HiSeq X Ten
1,000 $ / Human Genome

NGS sequencers

Illumina

Ion Torrent

Pacific Biosciences

Oxford Nanopore

Roche 454 FLX
No more support

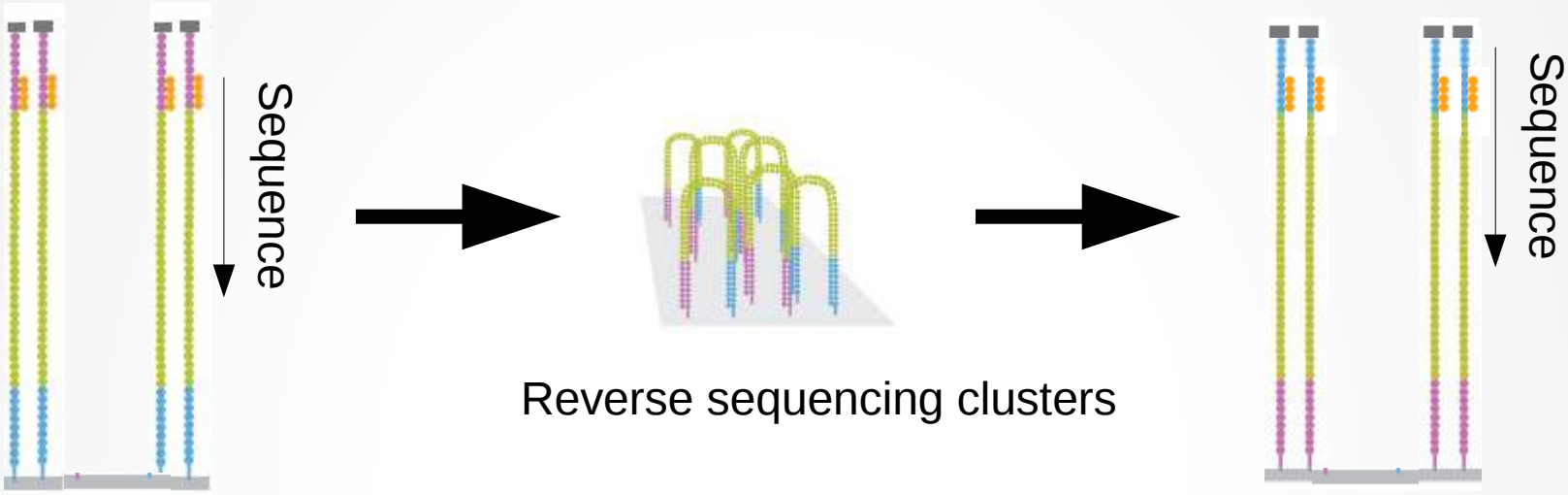
NGS sequencer comparison

	Sanger	***454***	***Ion Torrent***	***Illumina***	***PacBio***
Read length	**1000 bp**	max. 700 bp	200 bp	max. 300 bp	max. 1500 bp
Accuracy	**100,00%**	98,00%	98,00%	99,90%	87-99%
Run time	1(h)	7 (h)	2 (h)	1-10 (days)	2 (h)
Sequenced reads	few	1 million	max. 5 million	**3 billion**	45 thousand
Price (sequencer)	95k USD	500k USD	80k USD	690k USD	695k USD
Sequencing price	4 USD	7000 USD	350 USD	6000 USD	100 USD
Mb price	2400 USD	10 USD	1 USD	**0,07-0,5 USD**	2 USD
Generated data	1,9-84 Kb	10-100 Mb	1 Gb	**600 Gb**	

NGS applications

			Sequencing method		
	Category	**Example**	**Single-end**	**Paired-end**	**Mate-pair**
Genome sequencing	*de novo*	Create draft genome	(x)	x	x
	Resequencing	Mutation and / or structural variation	(x)	x	
	Metagenomic sequencing	Microbiome identification	x	x	
Targeted sequencing	Exome sequencing	Sequence DNA of all coding genes	x	x	
	Targeted sequencing	Sequence DNA of selected genes	x	x	
	RNA-seq / microRNA-seq	Identify expression levels	x	x	
	Barcoding	Sequence multiple patients in one run	x	x	
	ChIP-Seq / Faire-seq	Sequence open chromatin regions	x	x	

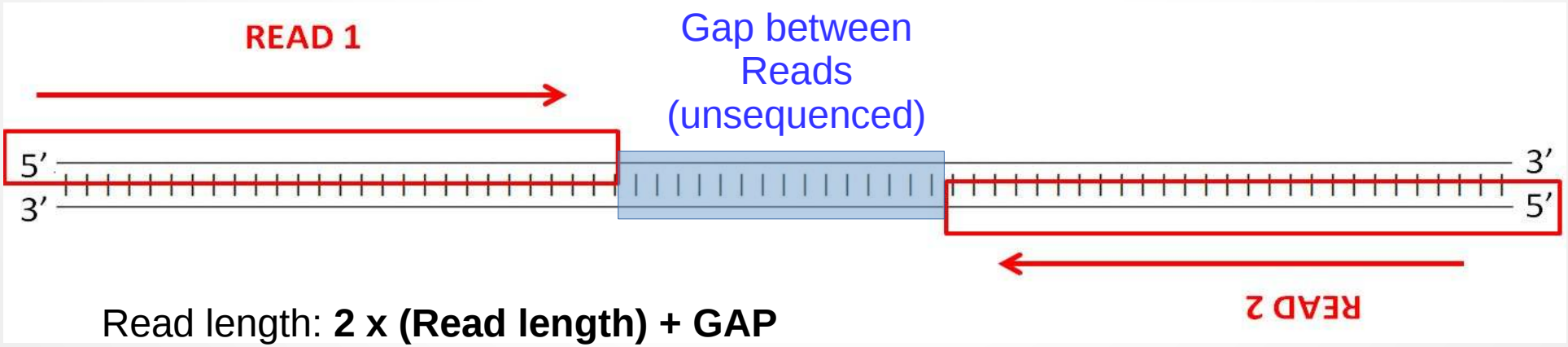
Sequencing – Paired-end



Reverse sequencing clusters

Forward orientation

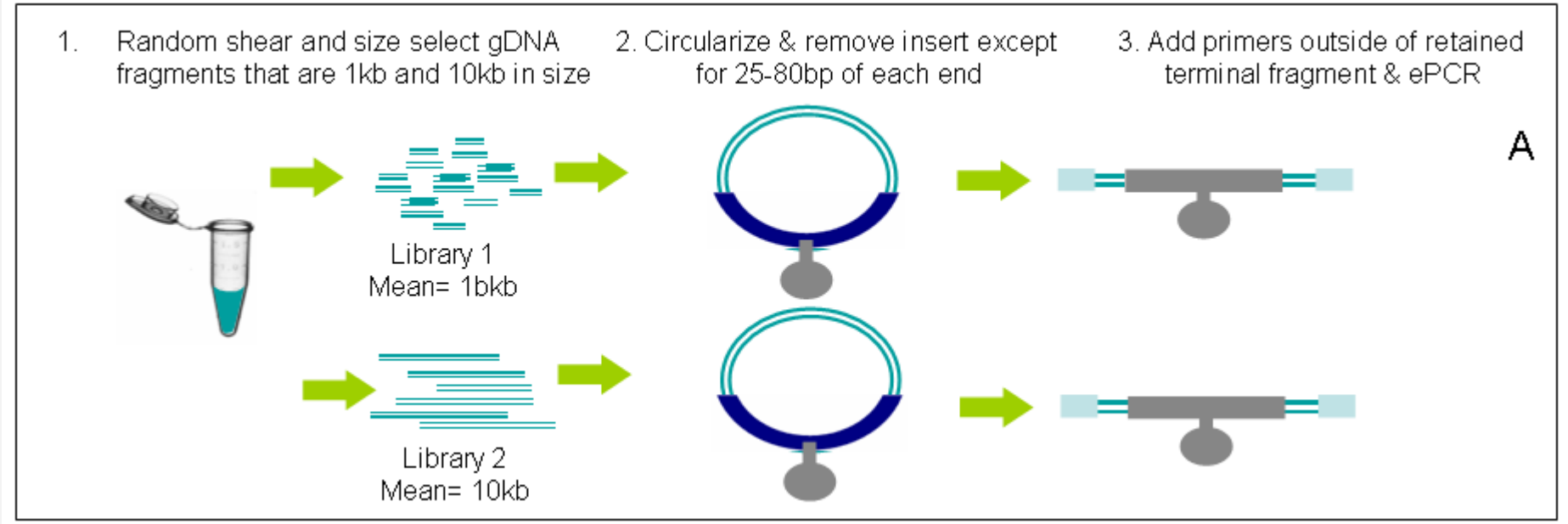
Reverse orientation



Read length: **2 x (Read length) + GAP**

Gap size: usually < 700 bp

Mate-pair sequencing

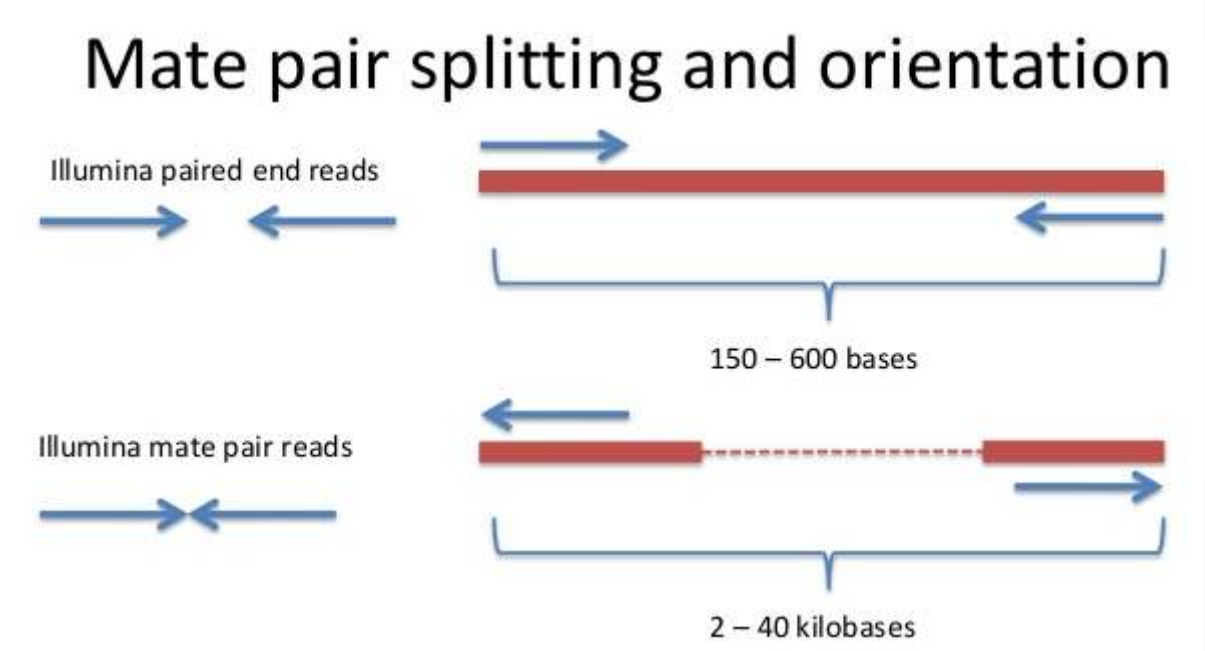


Orientation: First read is reverse, second read is forward (opposite to paired-end)

Read length: **2 x (Read length) + GAP**

Gap size: usually ~1,000 - 10,000

Paired-end vs. mate pair



Denovo Genome Assembly

Assembly – sequencing requirements

Minimum (only small genomes)
- Single end / Paired end library

Acceptable (small and large genomes)
- Paired end library
- Mate pair library

Optimal (mainly large genomes)
- Paired end library (1 or more)
- Mate pair library (1 or more)
- Long read library (PacBio sequencer)

IMPORTANT: Since we (bioinformatcs people) do the assembly, we should **advise** the type of library needed. The people from the wetlabs don't alway know the minimum requirements

Eg: Allpaths-LG: short insert sized paired-end library where the pairs overlap is **mandatory**

Genome Assembly

„What computer resources do I have? What genome will I assemble?”

Mammalian genome assembly: usually not CPU limited, but RAM limited

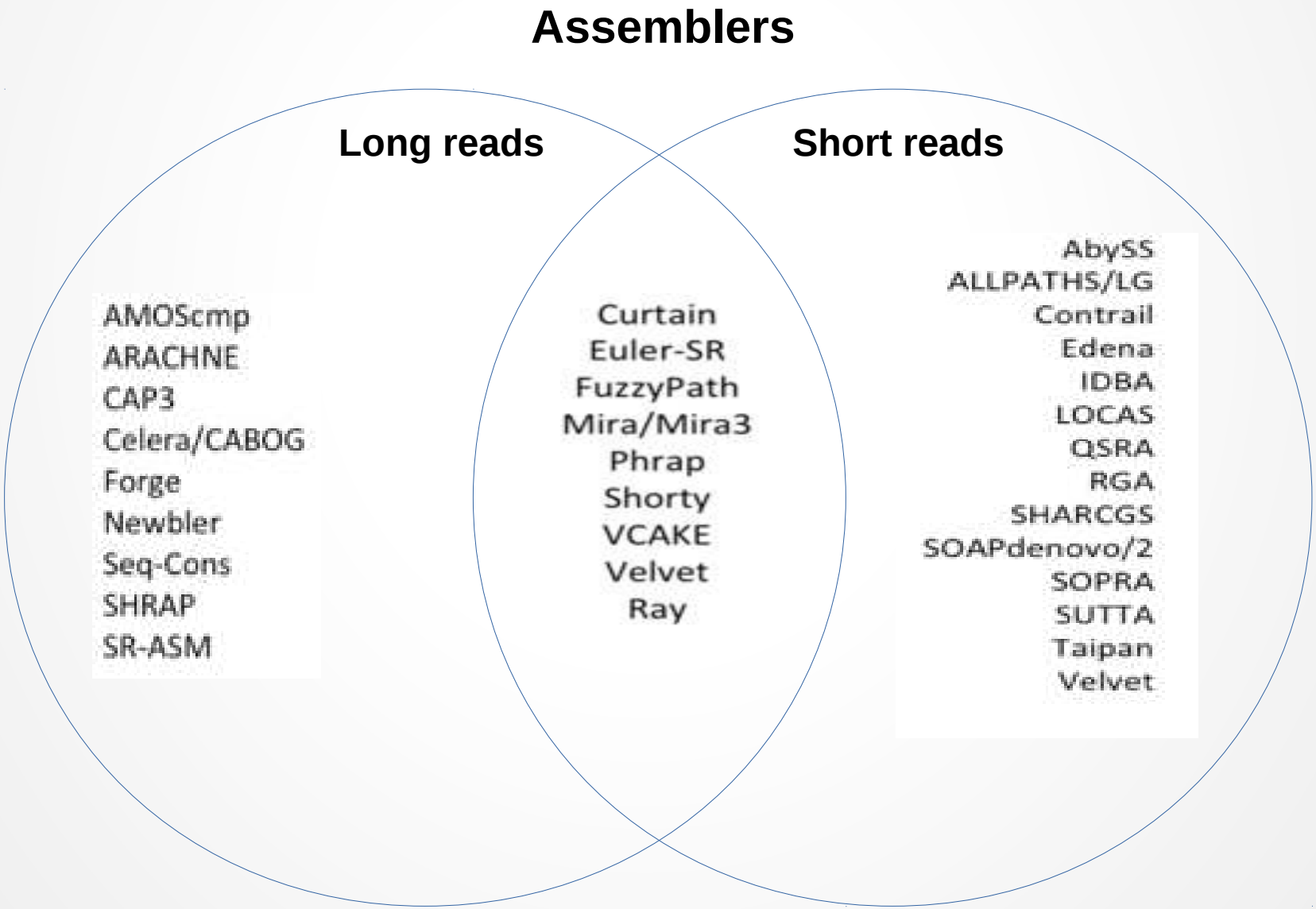
vs.

	Memory requirement	Assembly rank
minia	~2 Gb	3
SOAPdenovo	124 Gb	2
AllPaths-LG	~1 Tb	1

Microorganism genome assembly

	Memory requirement
Velvet	>8Gb
SPAdes	>8 Gb

Genome Assembly - „How long are my reads?”



What assembler should I use?

- http://gage.cbcb.umd.edu/
  - GAGE is an evaluation of the very latest large-scale genome assembly algorithms.
  - They compared assemblers, have the „recipe” (or commands) and datasets used for assembly.

GAGE

Recipe for Allpaths-LG

Staphylococcus aureus:

```
PrepareAllPathsInputs.pl DATA_DIR=$PWD PLOIDY=1
RunAllPaths3G PRE=. REFERENCE_NAME=. DATA_SUBDIR=. RUN=allpaths SUBDIR=run
```

Rhodobacter sphaeroides:

```
PrepareAllPathsInputs.pl DATA_DIR=$PWD PLOIDY=1
RunAllPaths3G PRE=. REFERENCE_NAME=. DATA_SUBDIR=. RUN=allpaths SUBDIR=run
```

Human Chromosome 14:

```
PrepareAllPathsInputs.pl DATA_DIR=$PWD PLOIDY=2
RunAllPaths3G PRE=. REFERENCE_NAME=. DATA_SUBDIR=. RUN=allpaths SUBDIR=run
```

Genome Assembly Gold-Standard Evaluations

Main page | Genome Assemblers | Data sets | Recipes | Results | Twitter

Assembly results of the human chromosome 14

3. Assemblies of Human chromosome 14 (ungapped size 88,289,540).

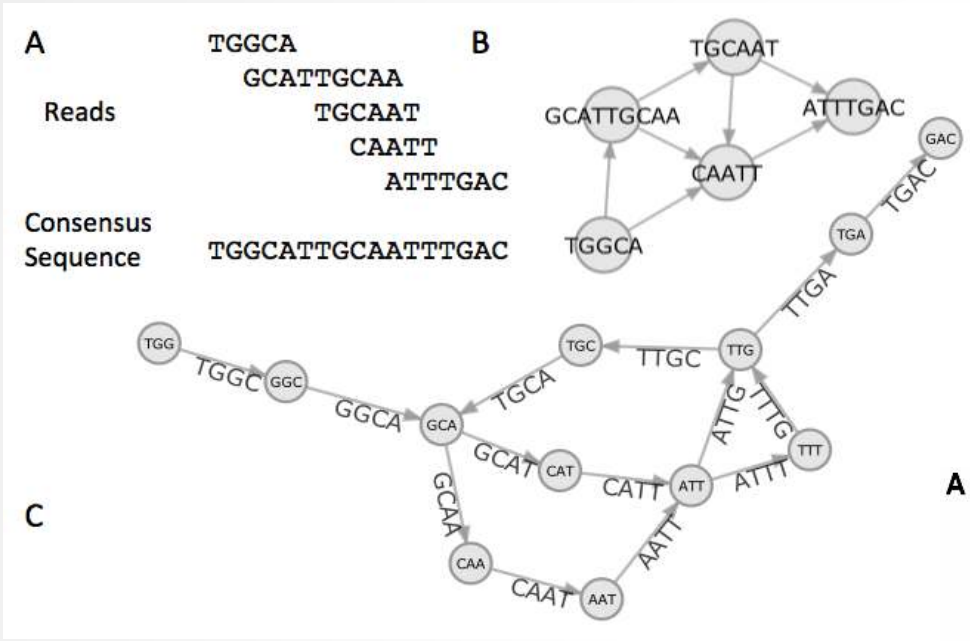
Assembler	Contigs				Scaffolds			
	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	51,924	2.0	**704**	2.0	51,301	2.1	**9**	2
Allpaths-LG	4,529	36.5	2,760	21.0	**225**	**81,647**	45	**4,702**
Bambus2	13,592	5.9	11,943	4.3	1,792	324	143	161
CABOG	**3,361**	**45.3**	3,181	**23.7**	479	393	597	26
MSR-CA	30,103	4.9	5,550	4.3	1,425	893	1068	94
SGA	56,939	2.7	981	2.7	30,975	83	19	79
SOAPdenovo	22,689	14.7	6,424	7.4	13,502	455	268	214
Velvet	45,564	2.3	4,910	2.1	3,565	1,190	9156	27

Assembly algorithm – Greedy algorithm

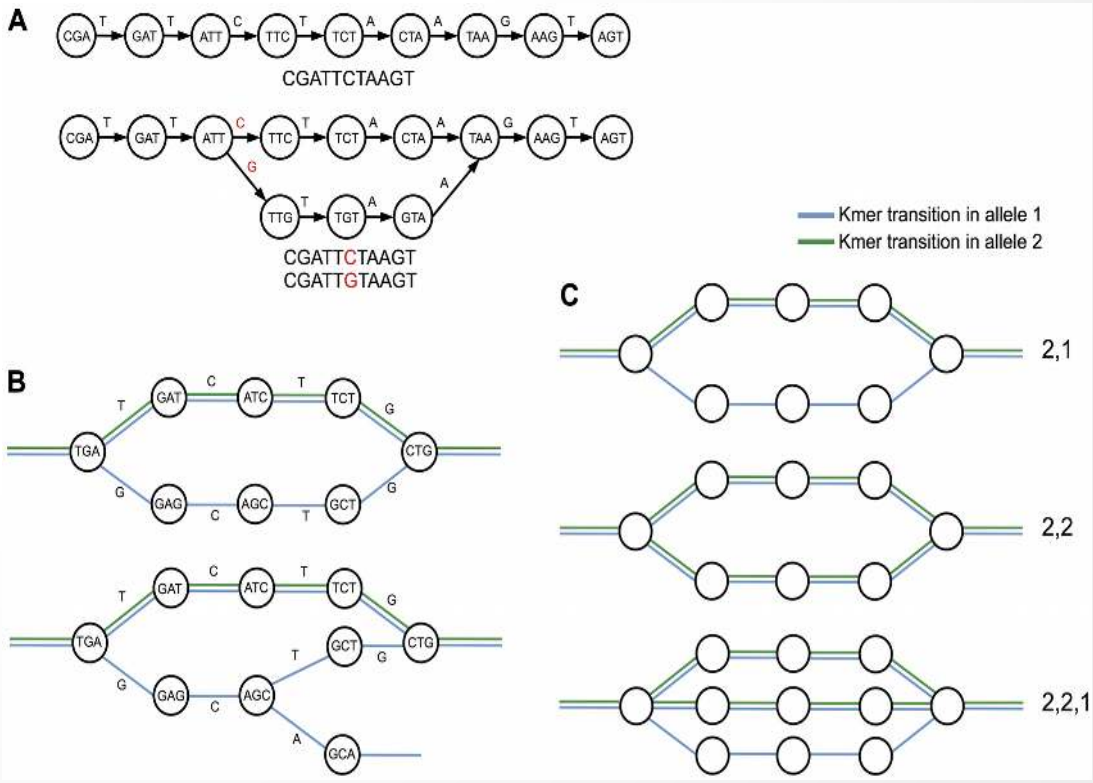
Given a set of sequence fragments the object is to find the shortest common supersequence.

1. Calculate pairwise alignments of all fragments.
2. Choose two fragments with the largest overlap.
3. Merge chosen fragments.
4. Repeat step 2 and 3 until only one fragment is left.
5. The result is a suboptimal solution to the problem.

Assembly algorithm – de Brujin Graph



Heterozygous positions



Genome assembly steps

- Check read **quality**
  - Question: Was there any problem with my sample preparation or sequencing?
  - Tool: FastX toolkit
- **Trimm** bad reads or read ends
  - eg. Trimmomatic
- Re-check read **quality**
  - Question: Do my reads have better quality? Did the trimming fix my problem (if there where any)?
- Genome assembly, usually with multiple parameters (eg. Different k-mers, change a bit the insert sizes) and / or multiple assemblers

Results: Draft genome!

Quality Control

- How good are my reads?

FastQC

Good sequencing reads

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html#M0

Bad sequencing reads

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html#M0

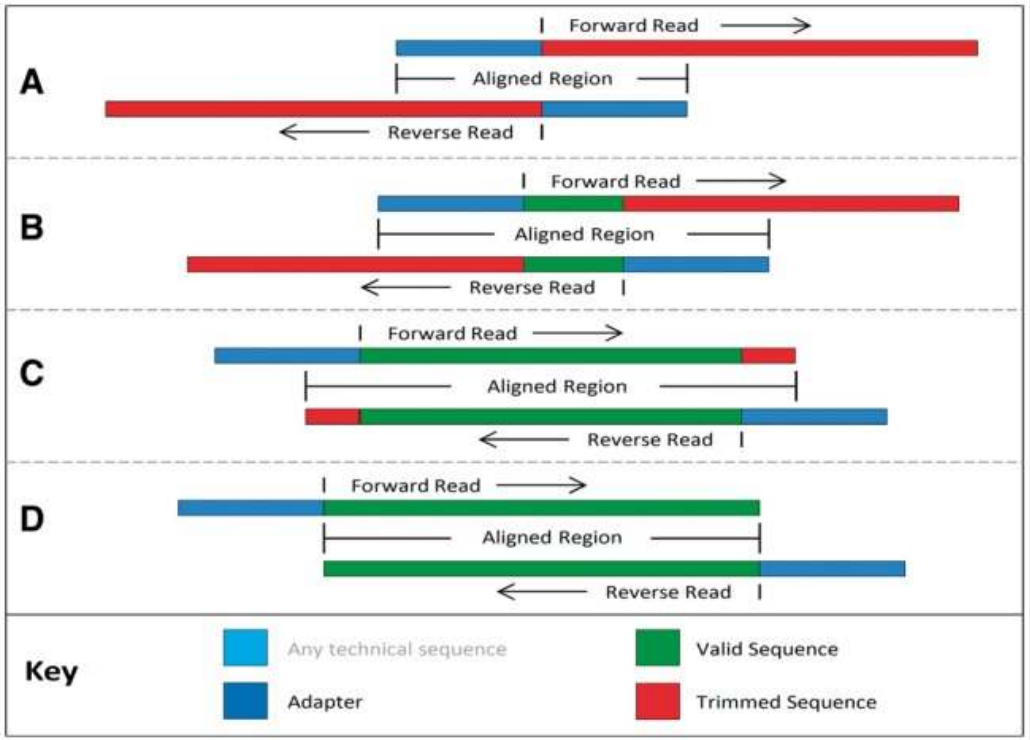
Trimming reads

- Remove bad quality reads
- Trimm bad read ends (beginning, end)

Trimmomatic

http://www.usadellab.org/cms/?page=trimmomatic

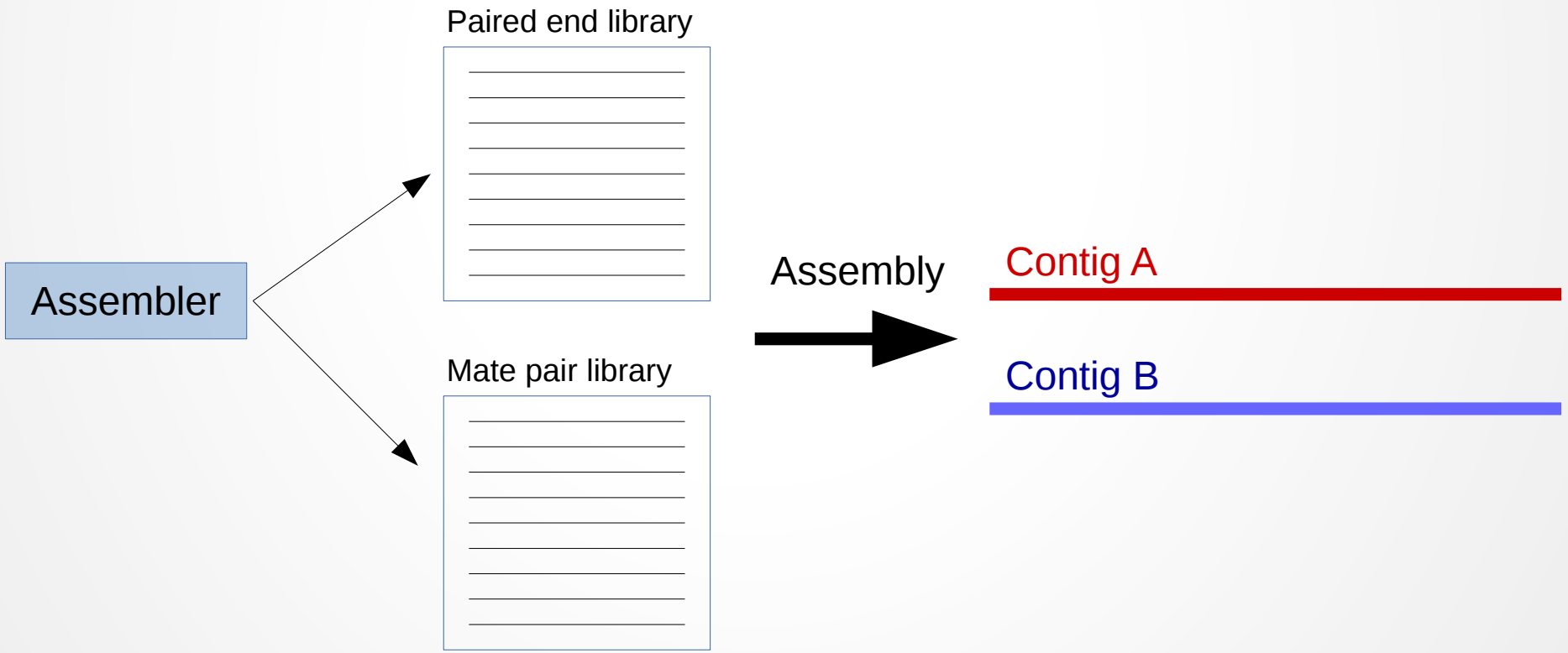
A few options:
1) Remove adapters
2) Remove leading low quality or N bases (below quality 3)
3) Remove trailing low quality or N bases (below quality 3)
4) Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15
5) Drop reads below the 36 bases long (MINLEN:36)



Assembly
Contigs vs. Scaffolds

A **contig** (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA.

Contigs are the result of the „primary” assembly.



Assembly
Contigs vs. Scaffolds

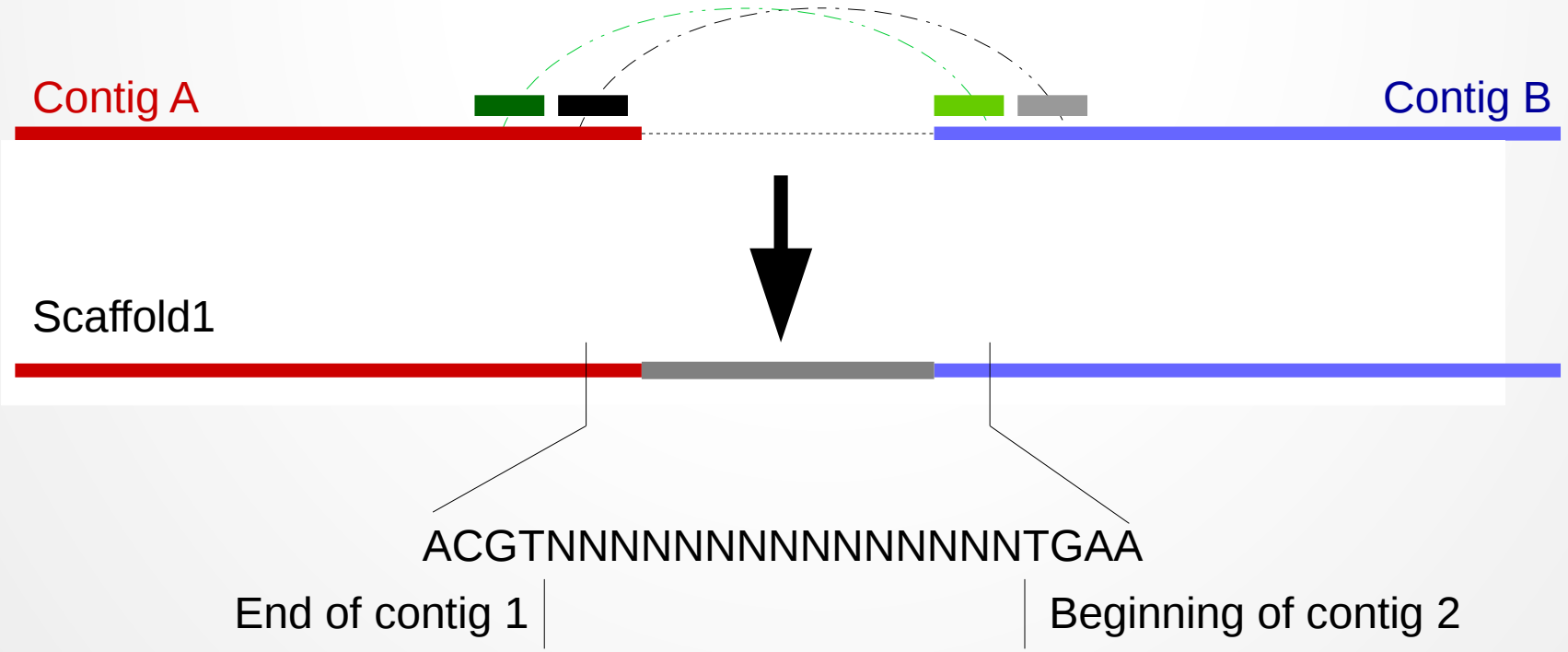
Scaffolds consist of overlapping contigs separated by gaps of known length.

Mate pair library

	pair1	pair2
read1		
read2		

Mate pair insert size: 2000 bases

Mate pair reads align to the edge of the contigs



Assembly
Contigs vs. Scaffolds (Example)

```
TTCCTGTAATGCTGTTAATGTCCGTGTTTCCTTCTTTAATCACAGTGATGCTCTGTTTTTTTCTGGACATTAACACTTCTATGATGTATTACTTTCAGTG
GCTCCTGCATGTGTAATTTAAGCACTAAACACACTGTTTTATTGTGTCACAGTGTAATTGGATAAACGATGCTATTTTTGCAGGCGACAGAGTACATATT
CAAATACATTATCCAATCTCGAAGGCTGTTTGCACTCGCCACCGGAGGCCAGAATGAAGAAGAGTTTCGCTGCTGTGTGCATGAACTGTTCATGTCCATC
CGCTTCTTCTTGTCGCAGGAGAGCAAAGGCTCTGGCCCTATCATGCACACACAGGTGAGGCCAGTGGATTATAGCTGCTTGATAACATCAGTTATCAATA
TTTATTAAAGATGATAAAAAAAGTTTCTTTTAAAAAGTGCTTTTAGACTGGACACGAGTAGGAAATTTATGAAGCCTACAAAGAAGTGTCAGTTATTATT
TTTTTAAGTTTTTGCTGTAACTTATAGCGTTGAAAAAAATCACATTTGTTCATACATTTTGCTTAGCAGCAGCAAGTAAAGAGTAAAAAAAAAATCTTGG
CTTATTGTTGCACTTTGGCAGGACTGCACATTTTTTTTGAATAATATCCTTAATTAATTATATAATCAAGGAGCGATAATTGATTCTTTATGGTCTGTTT
AAACAGAAATCTTAGCACCGCAGCTATTTTTTTTTTTTTTTTTTAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTCACTCCACCTCTCAGTCCTCAAACAAGACTAGTTCACATTCATGAGATCTGTATTGTGCTCATGTG
AAGAGATTCCCACAGAACAAACCGTTCTAATAGAAACTTTGAAGGACTGAGAAAGCAGAATCACCACAAAACTTGCGAAAACTAGAAGTAAAACTGGCCA
GAAAAATTGCAGTGACAGGAATTGCTGGAAAGTTGAGAGAAACTTGAGTATCCTTATGAAACAAGACATGACTCAGACTCCCAAACAGAAACTAAAAAAT
ATTTTGGGGTACCTTTTTCTTCTATAATTTATGTTTAATGTATAATTTTCCTGATTTTCATTTTTCCCTTCTTTAGGGCTGACTTATATATATATATATA
TACACACACACACACACACACTGTATATACTGTATGTGCATTCTGACCTGTAATGTGTACTTTATTTCATCTTAAATACAAATACAGCACATAAAAAAAC
AATAAAAGAGCACCCGTAGGTTTTAATCTAGATCGTGCCACATTTTTTAATTTATTTAAACAAAGCTGTCAAACCTGAATATTTTCATCTTTATACAGCA
CAGGACAAAAGCAATAAAGAGAATTCAGGTCAATGCTGTCCTAATCTGTACTAATCCATGGCCTTACTCTCCATCAGATTCCAGCATTTTGTTGTCTATT
TCTATTCACTTCCCTTCACTGTACACGGTCTGTAGTTGCAGTAATTCATGTTACACTGTAGTGCACTGTAGATCTGATGCTTTCTGTGGTTGGATGTCAG
```

End of contig 1

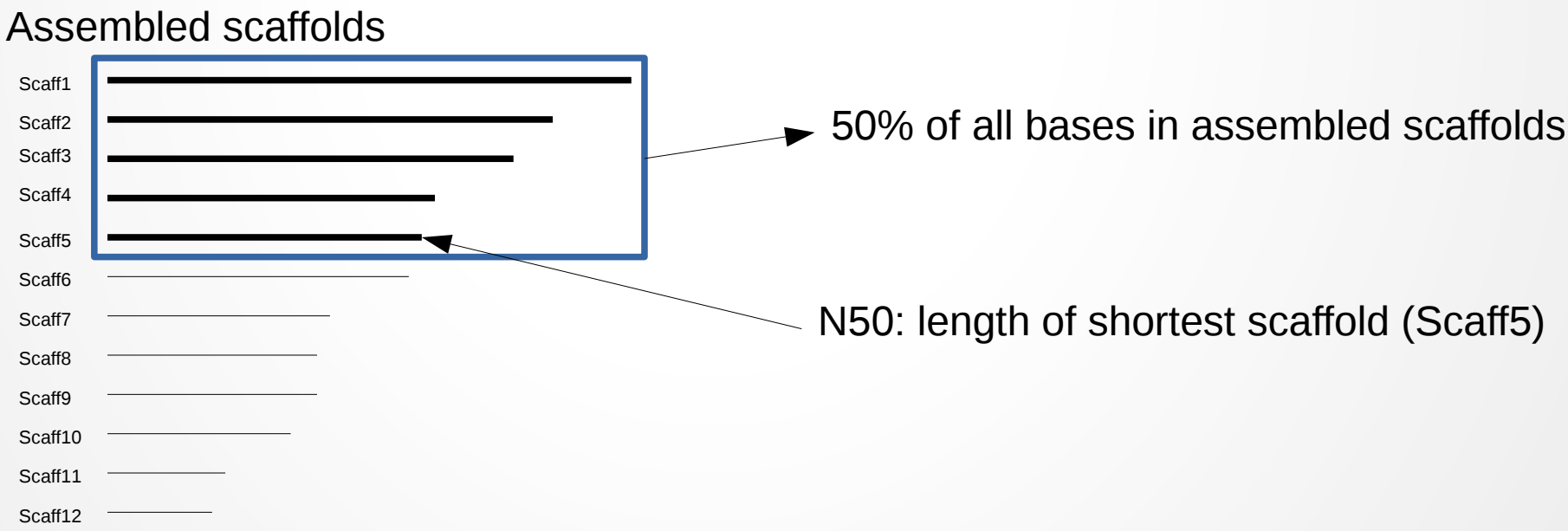
Beginning of contig 2

Assembly quality evaluation

- How can we tell, if we have a good assembly?
  - Assembled genome length?
  - No. of contigs and or scaffolds?
  - Predicted genes in genome?
  - **Length of contigs/Scaffolds??**

Assembly evaluation
N50 value

- N50: Given a set of sequences of varying lengths, the **N50** length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N.



Assembly quality evaluation - Quast

- Quast: Quality Assessment Tool for Genome Assemblies

Assembly	SOAPdenovo	Allpaths-lg	minia
# contigs (>= 0 bp)	1825968	12766	2652684
# contigs (>= 1000 bp)	115065	12477	101164
Total length (>= 0 bp)	1422762994	939385245	979673168
Total length (>= 1000 bp)	1104282542	939106079	630067803
# contigs	199019	**12766**	201002
Largest contig	953317	**5453839**	163814
Total length	**1162594395**	939385245	698425129
GC (%)	38.37	38.23	37.99
N50	49277	**600219**	10176
N75	6657	**191375**	3203
L50	5215	388	17287
L75	21312	1087	46128
# N's per 100 kbp	28709.12	26954.32	**19473.97**
# predicted genes (unique)	155876	96820	117608
# predicted genes (>= 0 bp)	157661	96943	118931
# predicted genes (>= 300 bp)	71251	**73645**	43897
# predicted genes (>= 1500 bp)	20441	**27712**	14325
# predicted genes (>= 3000 bp)	9045	**13131**	6258

Genome Assembly: Important Questions

- What genome do I have to assemble?
- What sequencing libraries do I have?
- What assembler should I use?
- Do I have the correct libraries for my assembler?
- Did the reads / libraries pass quality control? (FastX-toolkit)
- Is my assembly good? (Quast)
- Should I use another assembler?
- How can I validate my assembly?
  - eg. Do I have a Sanger sequenced marker database?
  - eg. Do I have RNA-seq / Exome-Seq data?

Our job today

- Patient with helicobacter pylori infection
  - Symptoms: abdominal pain, nausea and vomiting, fatigue…
- Biopsy → sample isolation and preparation

NGS sequencing

(Illumina paired-end and Pac Bio long reads)

Bioinformatics analysis

Many reads did not map to H. pylori genome → ***De novo*** genome assembly

Two different assemblers: SPAdes and Velvet