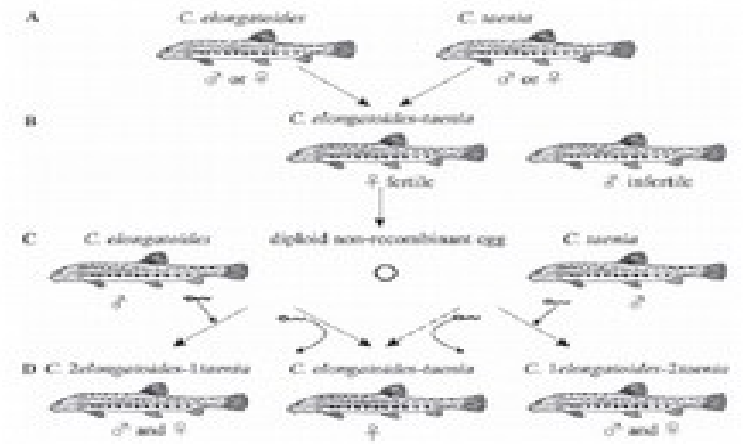
Assembling diploid genome of *Cobitis taenia* using Illumina short reads

November 7 2018

Martin Mokrejš

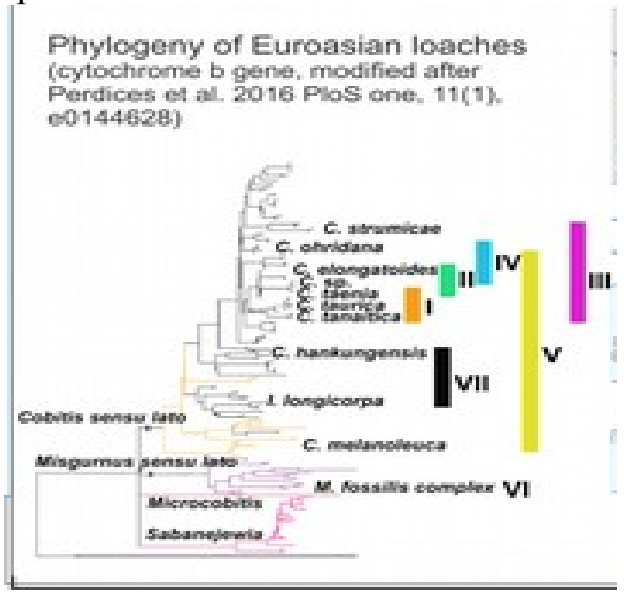
IT4Innovations
national supercomputing
center

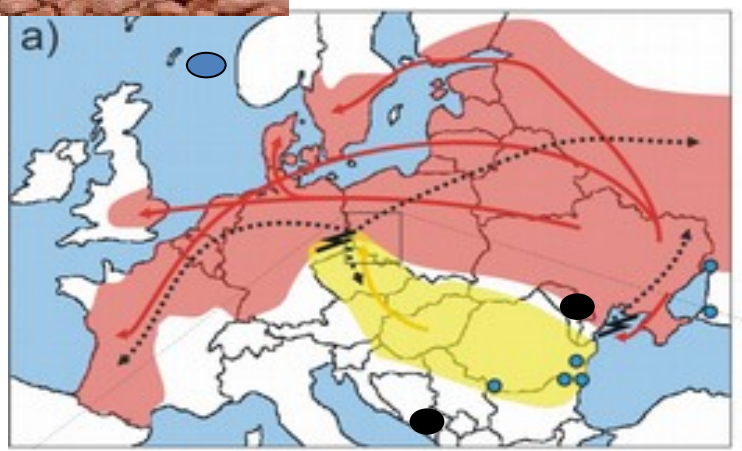


European loaches – suitable model case
Hybridisation among *Cobitis* species is known to produce clonal lineages;
Gynogenetic reproduction of hybrids

Hybrids exist in following forms:
ET, EET, ETT, EN, EEN, ENN,ETN, EP, EEP

Polyphyletic and dynamic origin of clones





***Cobitis elongatoides* E**

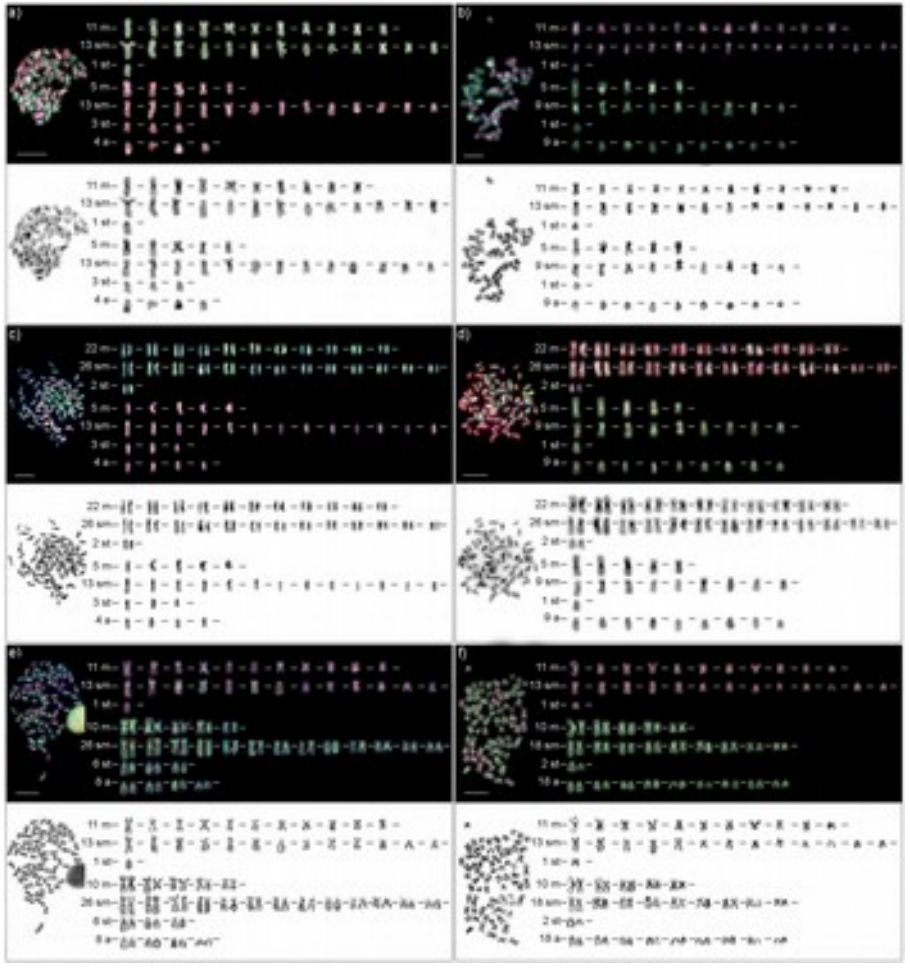
***Cobitis taenia* T**

***Cobitis tanaitica* N**

***Cobitis pontica* P**

Janko et al. 2006 Mol. Ecol.; Choleva et al. 2012 Evolution

Cobitis fishes have 49 or 50 chromosomes



How to sequence and assemble a genome

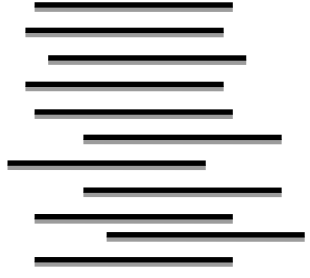
Apply some assembly algorithm

Genome

Fragmented, size-selected

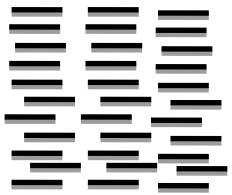
Some sequencing technology

“Shotgun” reads

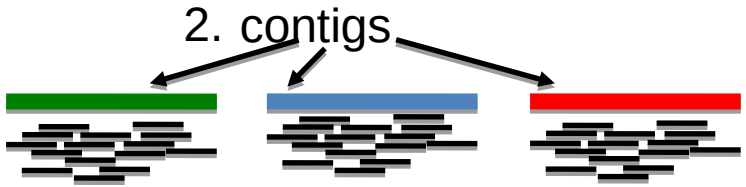


De novo assembly

1. reads



2. contigs



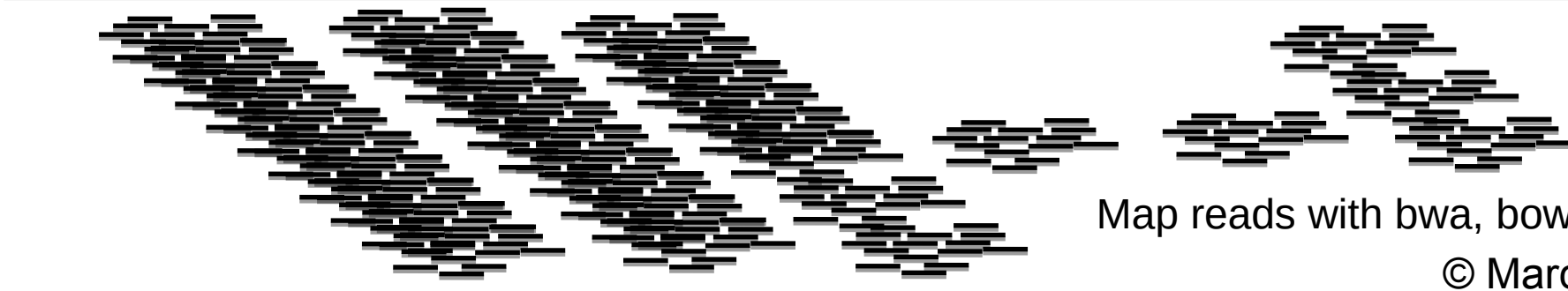
3. scaffolds

PE reads with known distances

Must assemble from scratch

Reference-based assembly

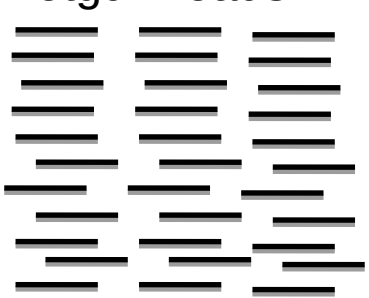
Reference genome

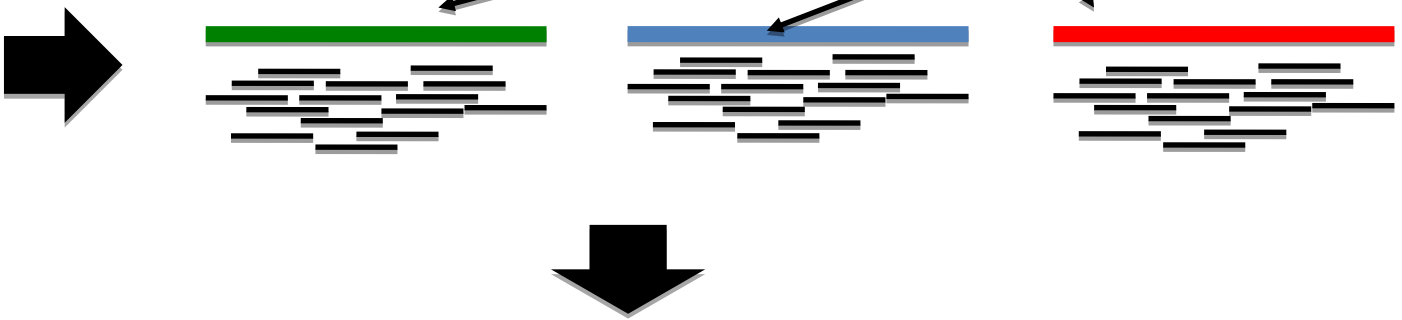


Map reads with bwa, bowtie etc…

De novo Assembly Basics

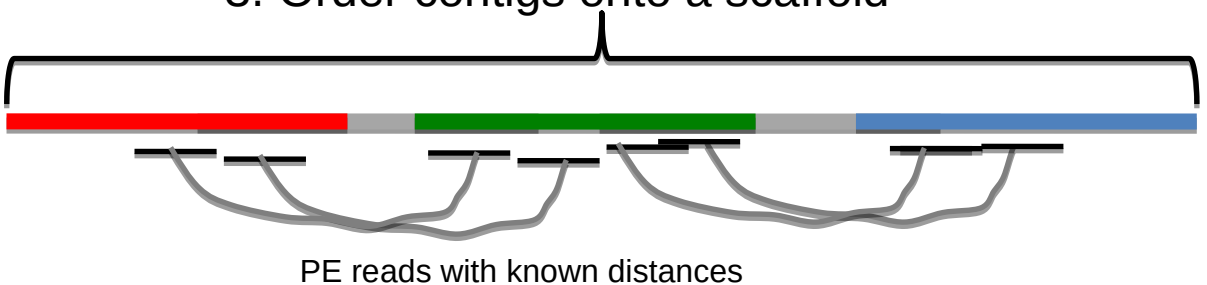
1. Shotgun reads





3. Order contigs onto a scaffold

Use paired-end info to determine order of (and distance between) contigs



4. (optional) gaps between contigs are filled in by mapping reads back to the scaffolds

Illumina Paired-end and Mate-pairs

Paired-end (PE) “short insert library” sequencing

orientation

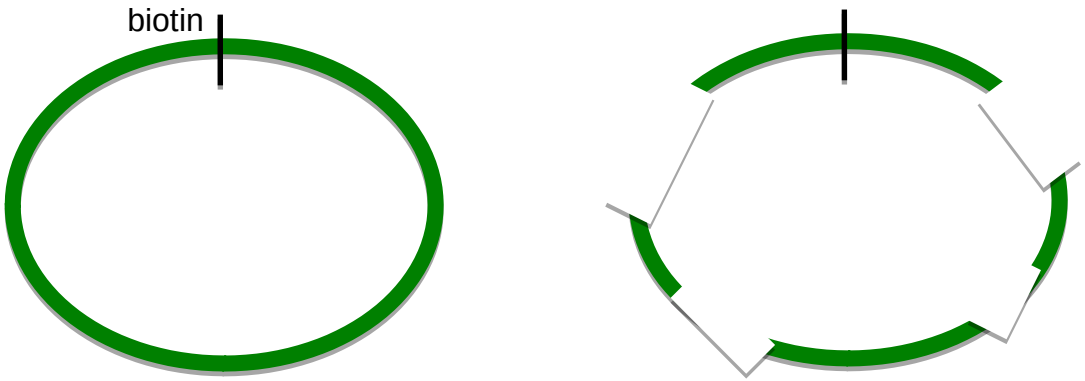
- Genome is fragmented to desired lengths
- Reads one end of the molecule, flips and then reads the other end
- Generates read pairs with a known distance between them

500bp

FR

Mate-pair (MP)”jumping library” sequencing

- Circularizes longer molecules (2kb-25kb)
- Biotinylated, fragmented, enriched, and sequenced



RF

Repeats Resolved

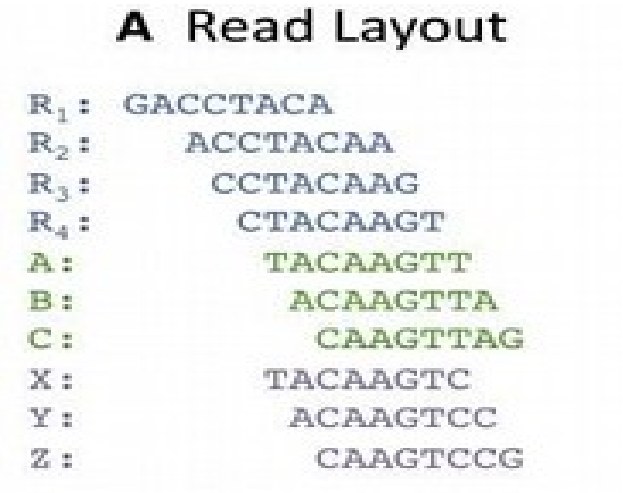
- Repeats can be resolved using paired-end information
- If one end of a read is unique, then you can map both reads.

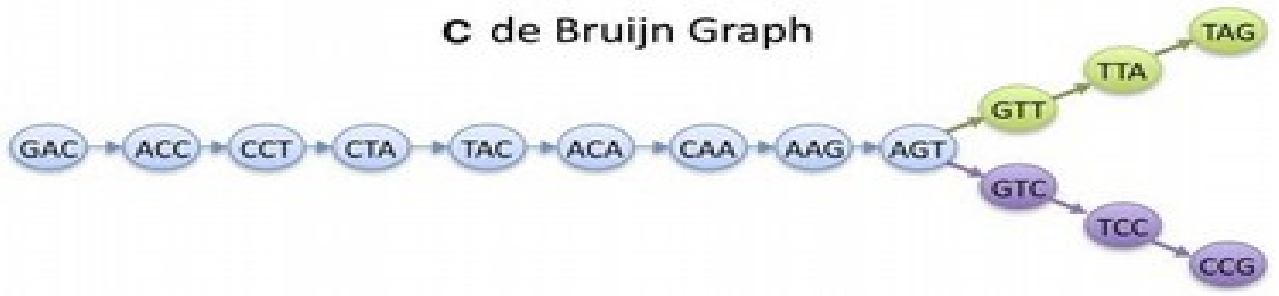
- However, for longer repeats (*i.e.* LINEs) this will not work.
- Hence Illumina-based genomes tend to be fragmented

- Chromosomes
- fragments
- K-mers
- unitigs
- contigs
- scaffolds

de Bruijn Graph Construction

- Reads are decomposed in to *k*-mers
- *K*-mers become nodes in a graph.
- Edges are drawn between *k*-mers which overlap by *k*-1 bases.
- Non-branching paths in the graph form unambiguous stretches of sequence.





Expected *Cobitis* genome size?

1.3Gbp (1.3giga characters [ATGC])

3.2Gbp (human genome)

Animal Genome Sizes

Genome Size in Gigabases

http://genomesize.com/

Sequence datasets of the Cobitis (fish) genome

Illumina 2x250nt reads paired-end, inserts ~600bp

Illumina 2x250nt reads mate-pair:

5kbp

8kbp

5kbp

About 5 000 Euro were spent in sequencing with sample preparation

Quality control steps

analyze frequencies of [ATGC], cross-compare dataset properties

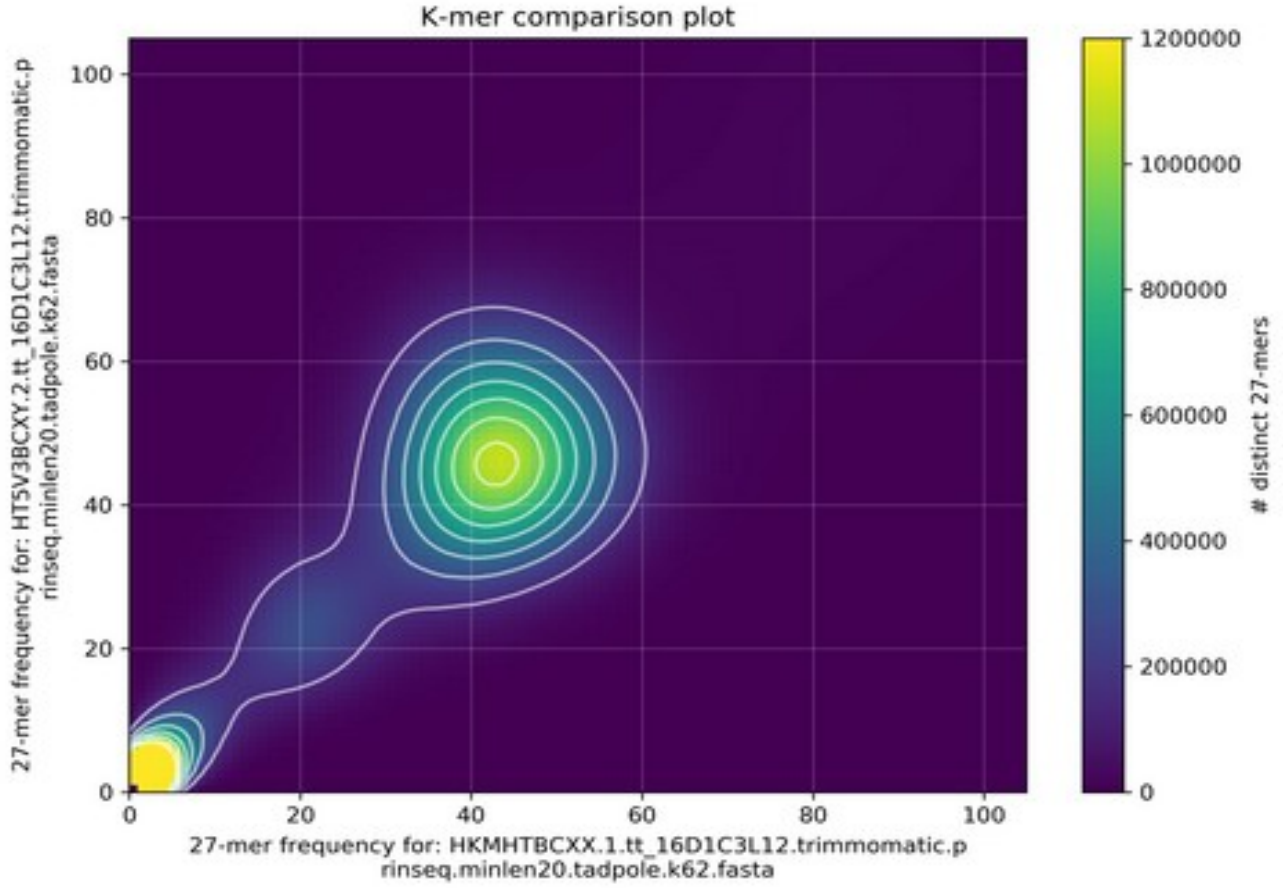
raw vs. trimmed FASTQ files can be inspected

- FASTQC/MultiQC
- trimmomatic
- KAT

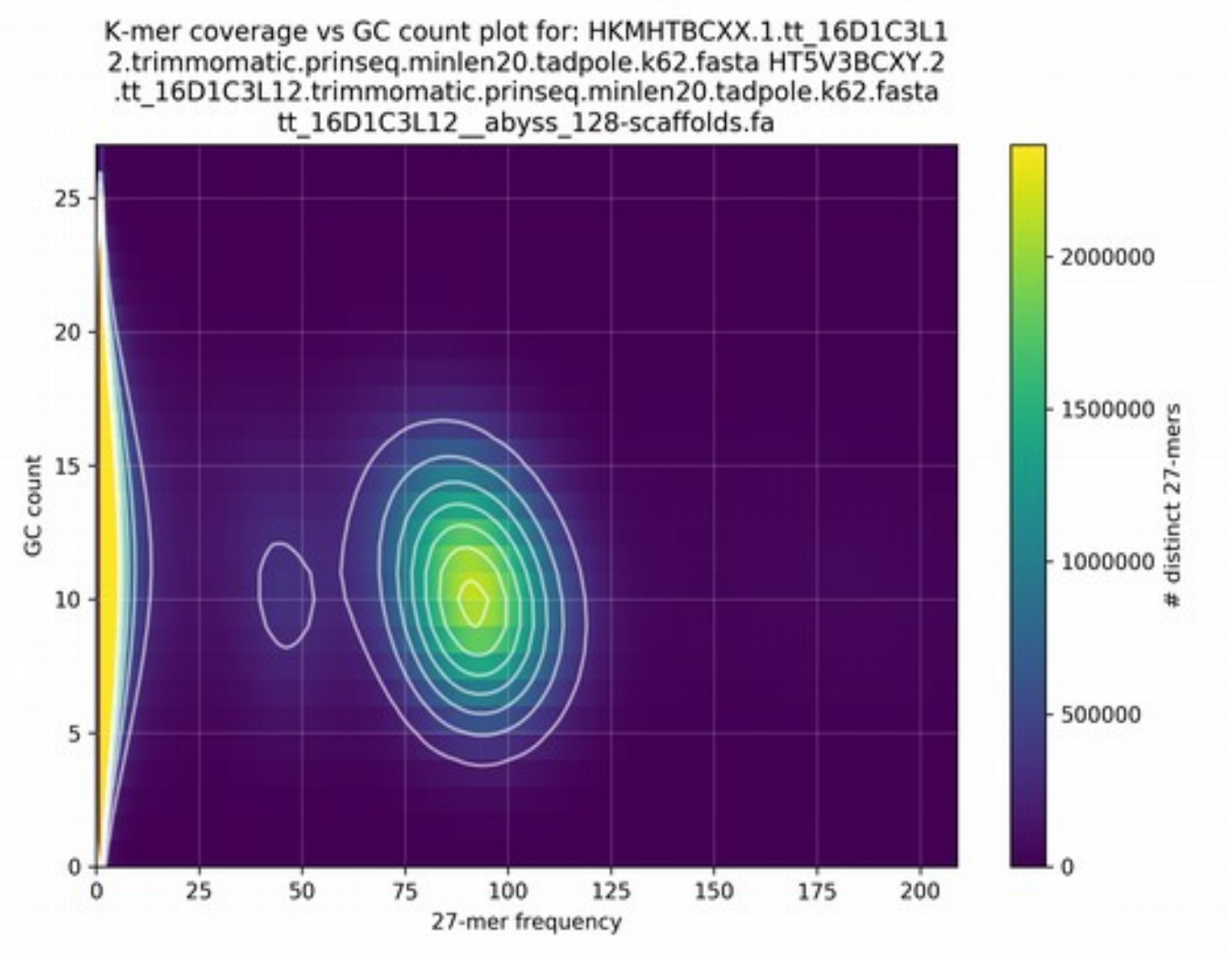
assembled unitigs/contigs/scaffolds can be inspected

- ntCard/jellyfish
- GenomeScope
- KAT

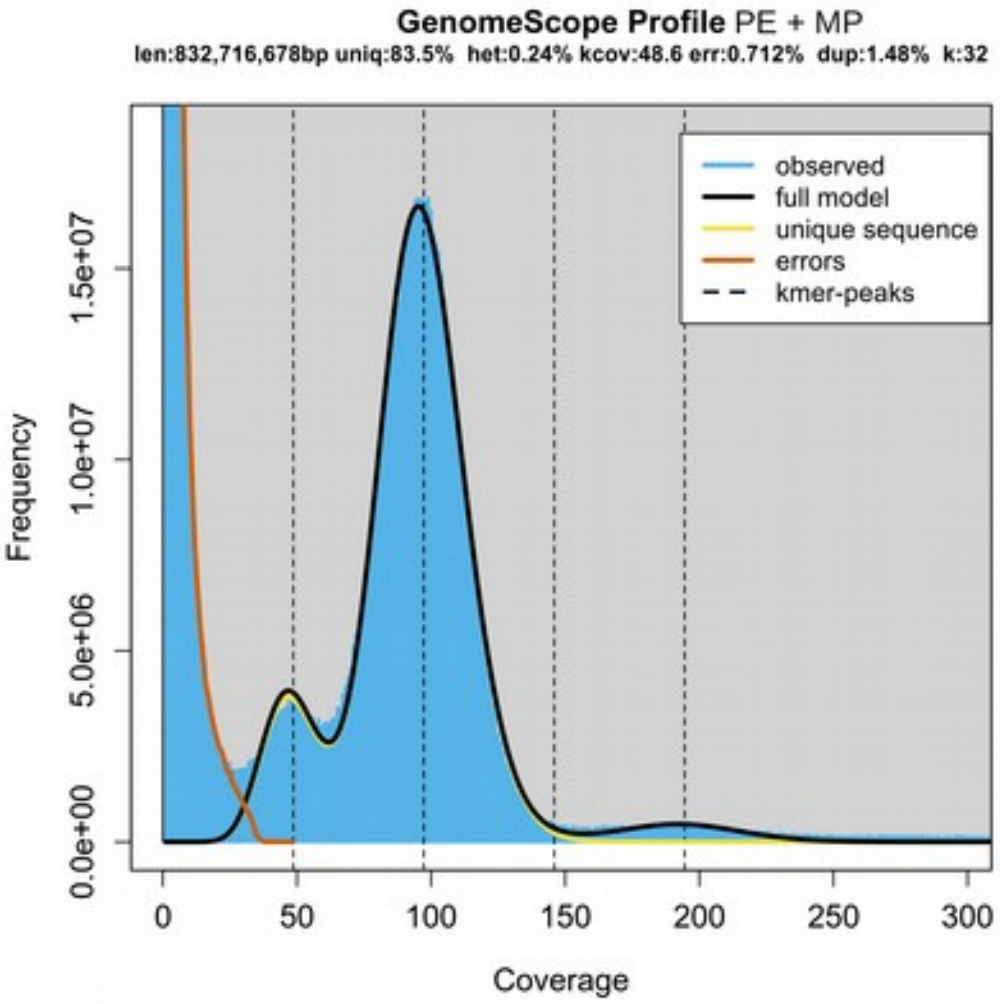
Basic quality-control checks, PE2016 vs. PE2017 libraries, kmer freq. vs. kmer freq.

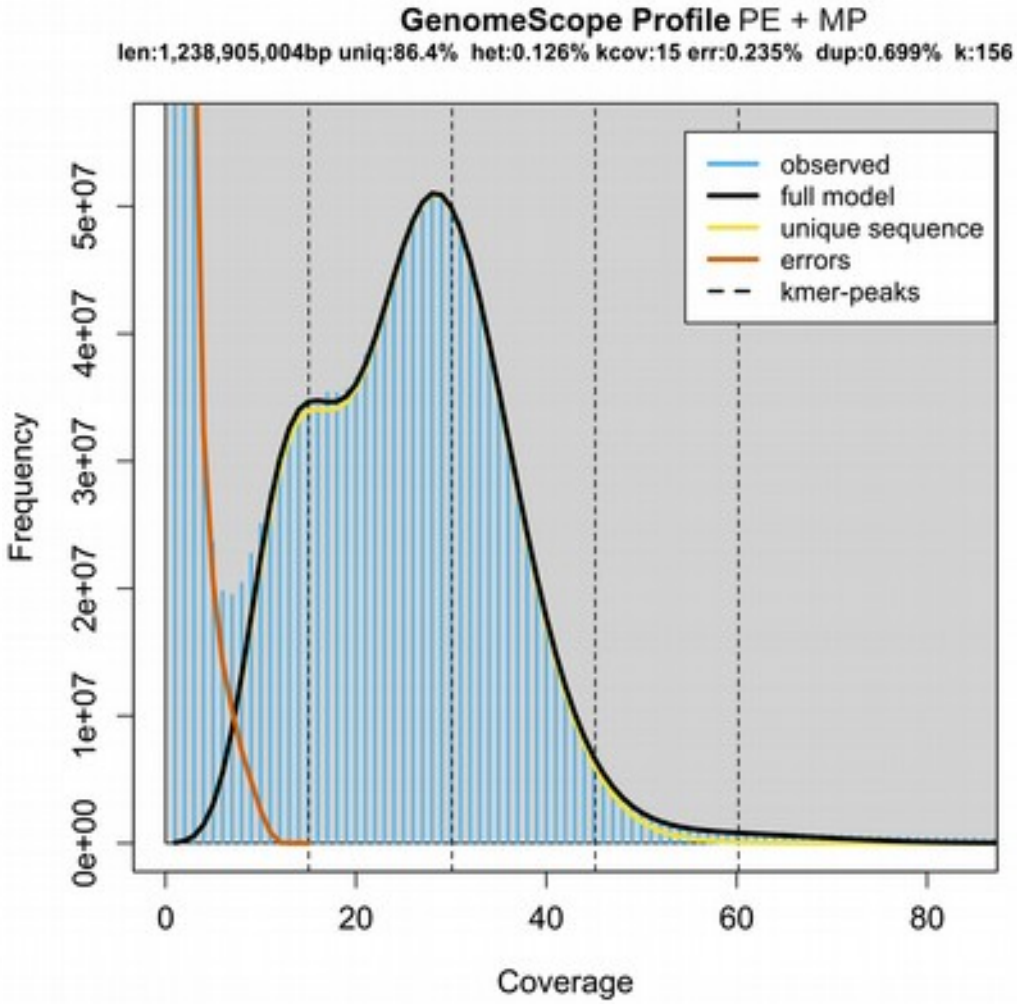


Basic quality-control checks, PE2016 vs. PE2017 libraries, kmer freq. vs. GC-content



K-mer frequency analysis of cleaned reads reveals diploid and haploid portions of the genome (k-32 and k=156)





Computational resources spent so far

- 32 TB working dataset (FASTQ, BAM, VCF) (25TB compressed)
- 731 GB working 454 cDNA datasets
- 723 856 CPU hours burned under OPEN-9-41 project
- 91 306 CPU hours burned under OPEN-13-42 project (out of 448 000 core hours available until 2019-02-20

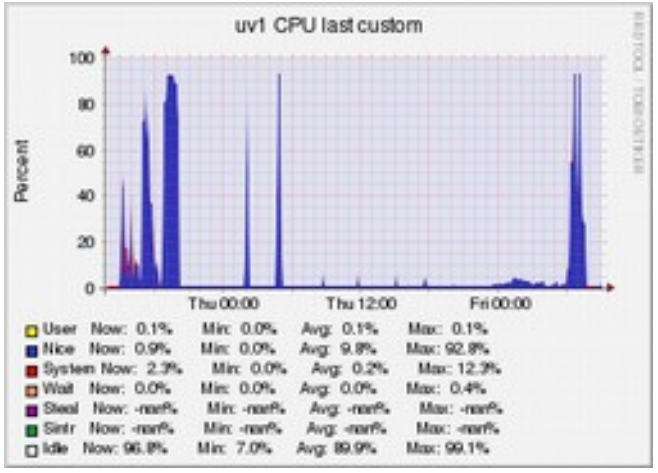
IT4Innovations
national
supercomputing
center

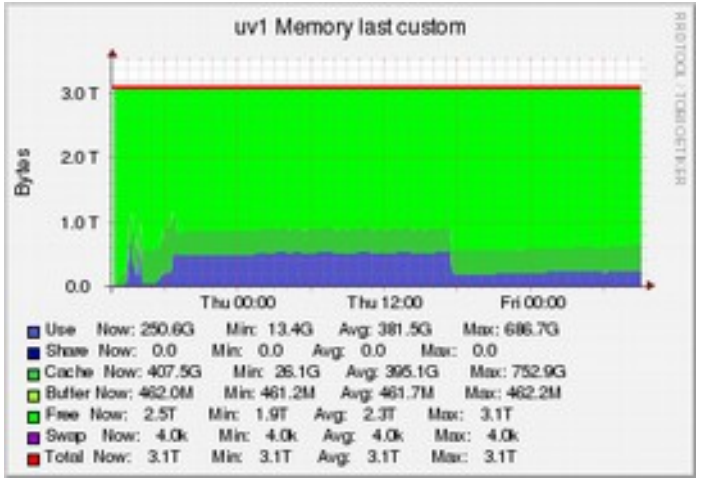
Genome assembly programs tested

- **SOAPdenovo**
- **SPAdes-3.11.1**
- **abyss-2.0.2 and 2.1.0, 2.1.2**

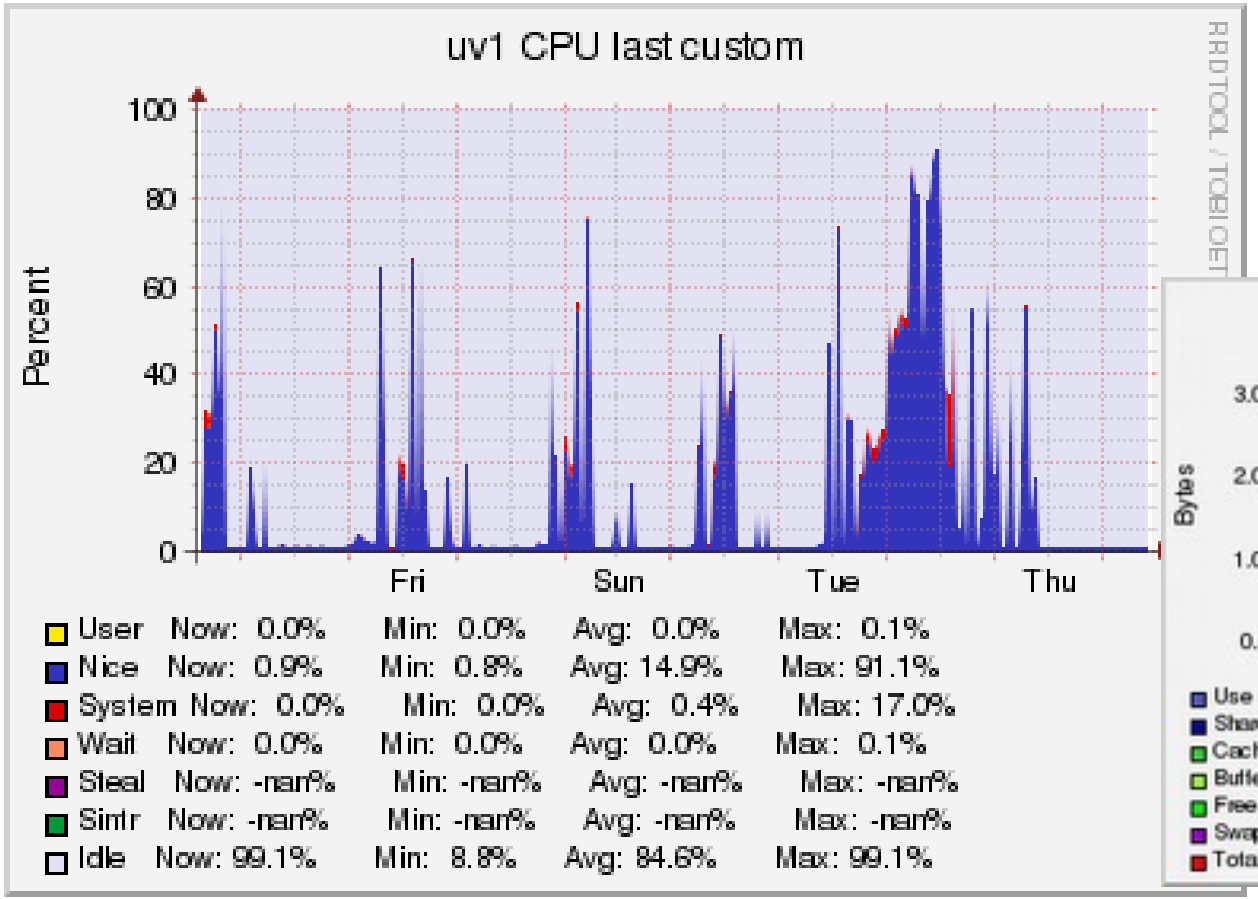
SPAdes 3.11.1 assembly attempts

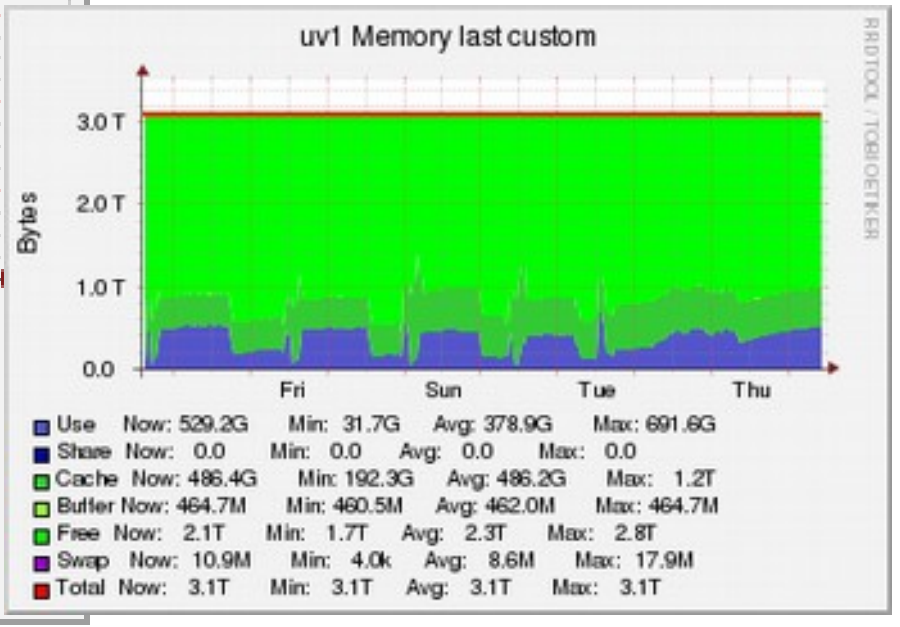
- no good results nor performance, placing input data + $tmpdir into /*dev/*shm is a must
- had to use --read-buffer-size and --tmp-dir options
- badly scaling, k-mer splitting/counting is single-threaded
- no expected remaining computations time is printed
- scaffolding crashed
- error-correction using builtin hammer tools uses k=21 (suboptimal)
- can perform several incremental assemblies but after the last k-mer size moves to gap closing (without outputting intermediate files)





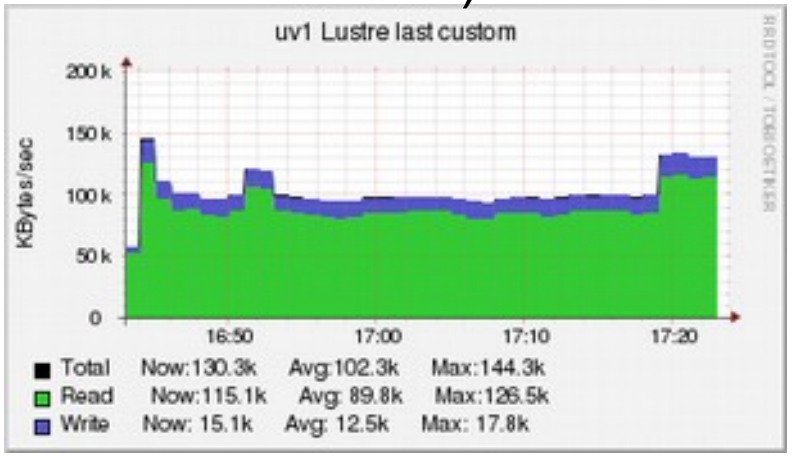
SPAdes: Series of several incremental assemblies with increasing k-mer sizes

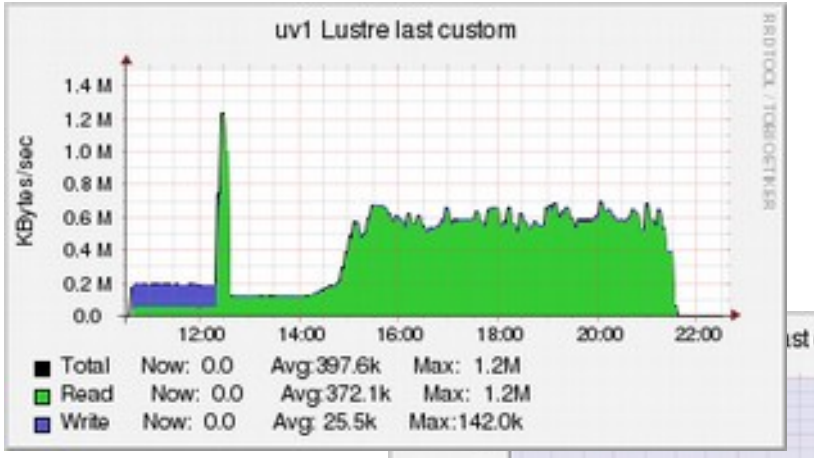


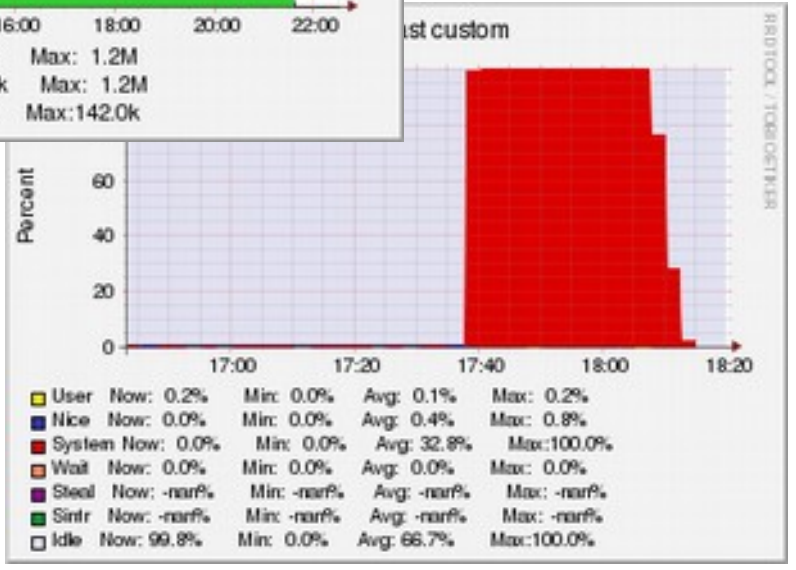


SPAdes: reading/writing of output files is slow (1 MBps vs. 0.6 – 1.2 GBps)

SPAdes supposedly writes out data in too small chunks (should be 1MB or even 10MB in size)







too many of too small chunks kill the LustreFS filesystem servers and cause data loss

Results

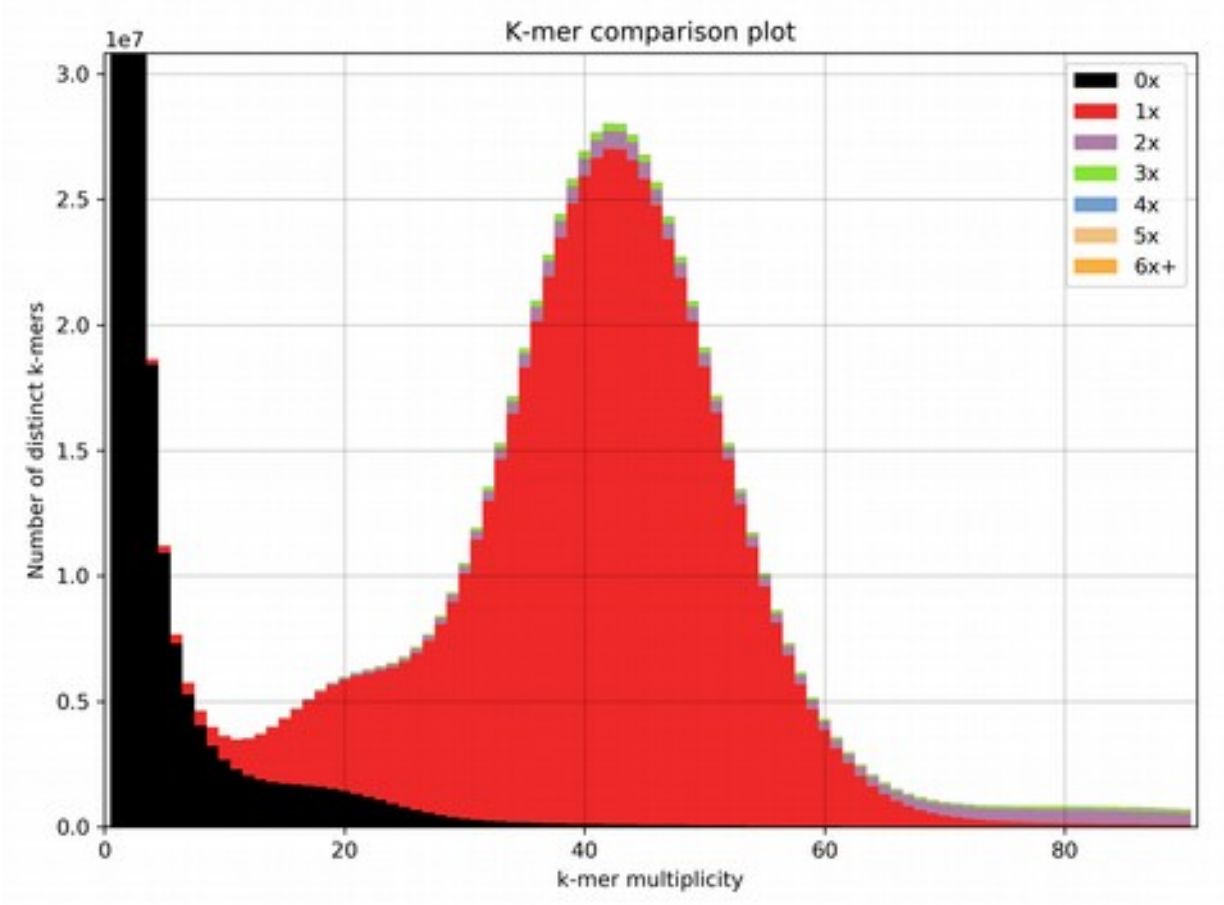
Cobitis genome assembly

- abyss-2.0.2
  - ECC correction using tadpole.sh, k=63
  - tested k-mer sizes 64, 96, 128, 144, 156, 160, 192
  -
  - 517042 contigs >=500nt
  - 223325 (scaffolds+remaining_contigs) >=500nt

$ abyss-fac -G 1267403131 tt_16D1C3L12__abyss_160-?.fa

n	n:500	L50	LG50	NG50	min	N75	N50	N25	E-size	max	sum	name
5397779	618531	106862	157307	1919	500	1173	2615	5048	3810	42190	1.041e9	tt_16D1C3L12__abyss_160-1.fa
2419093	517042	85266	125235	2482	500	1537	3344	6266	4688	57811	1.037e9	tt_16D1C3L12__abyss_160-3.fa
1814706	349105	47307	63838	5107	500	3111	6369	11680	8712	82251	1.079e9	tt_16D1C3L12__abyss_160-6.fa
1672273	**223325**	4552	6810	34920	500	13093	52459	132681	91701	721116	1.074e9	tt_16D1C3L12__abyss_160-8.fa

Basic quality-control checks, PE2016 vs. abyss-k160 assembly
Is the genome assembly at k=160 inflated?

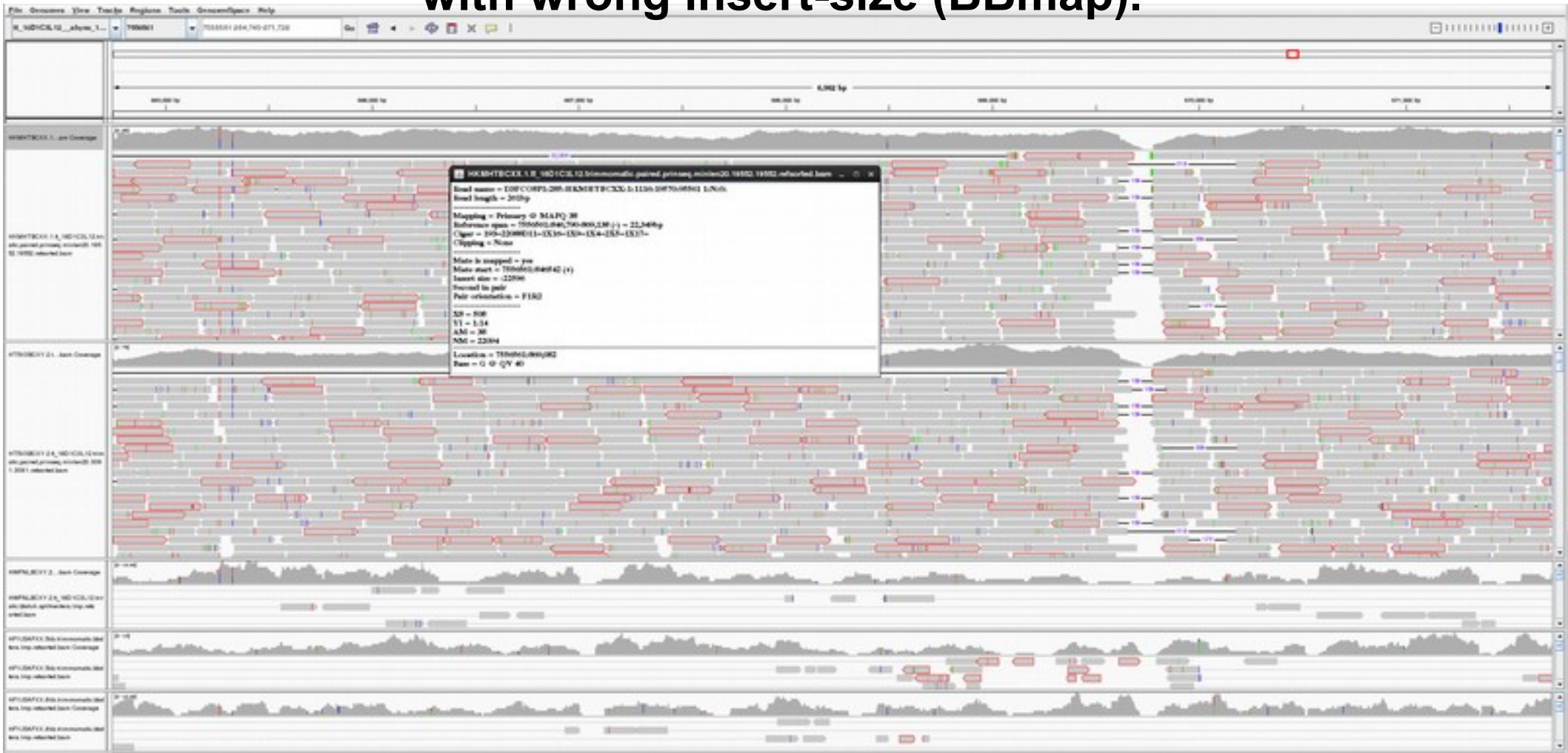


Conclusion: The assembly is not inflated by redundant contigs, which is good.

So why do we have the genome still in 223 325 pieces instead of just 50?

There are gaps due repetitiveness of the genome and conflicting long-distance evidence from mate-pair datasets.

Gaps in the abyss_k128_ecc_N10 assembly and lots of read pairs with wrong insert-size (BBmap).




6,962 bp
667,000 bp
668,000 bp
669,000 bp
670,000 bp
22,093
HKMHTBCXX.1.tt_16D1C3L12.trimmomatic.paired.prinseq.minlen20.19552.19552.refsorted.bam
Read name = D3FCO8P1:285:HKMHTBCXX:1:1116:19670:95561 1:N:0:
Read length = 261bp
Mapping = Primary @ MAPQ 38
Reference span = 7556561:846,790-869,138 (-) = 22,349bp
Cigar = 193=22088D11=1X16=1X9=1X4=2X5=1X17=
Clipping = None
Mate is mapped = yes
Mate start = 7556561:846542 (+)
Insert size = -22596
Second in pair
Pair orientation = F1R2
X8 = 508
YI = 1.14
AM = 38
NM = 22094
Location = 7556561:869,082
Base = G @ QV 40

Future plans

Oxford Nanopore datasets

2017:

1D-reads, RapidSequencing kit

3kbp avg. reads

2018:

Cobitis taenia, spleen

1D-reads, RapidSequencing kit

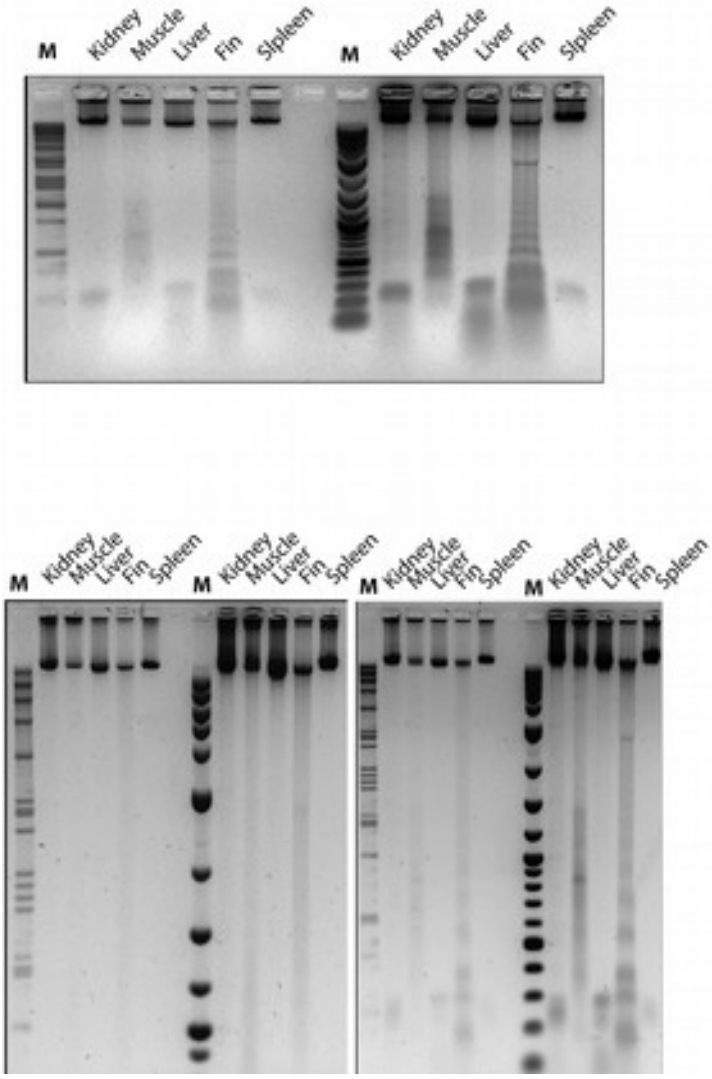
7 runs, 0.7Gbp per run

~ 60 000 reads >=10 kbp

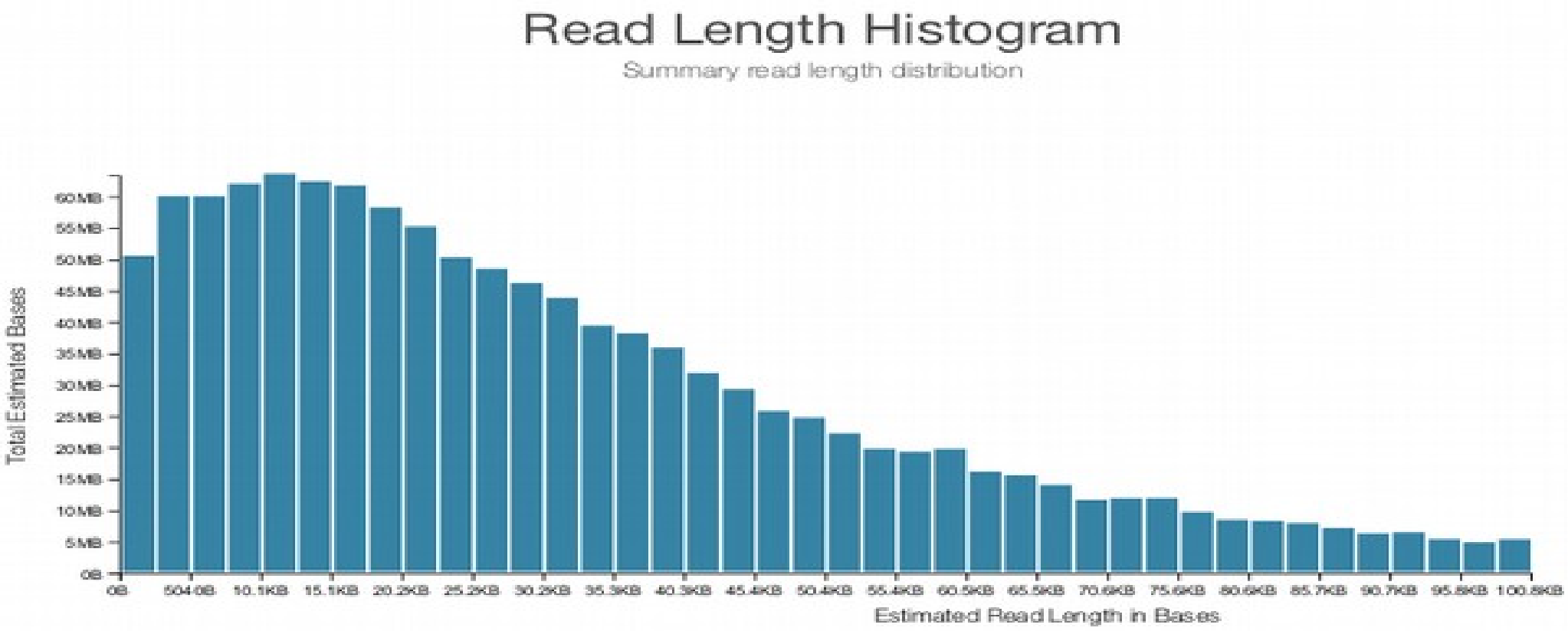
proovread

albacore

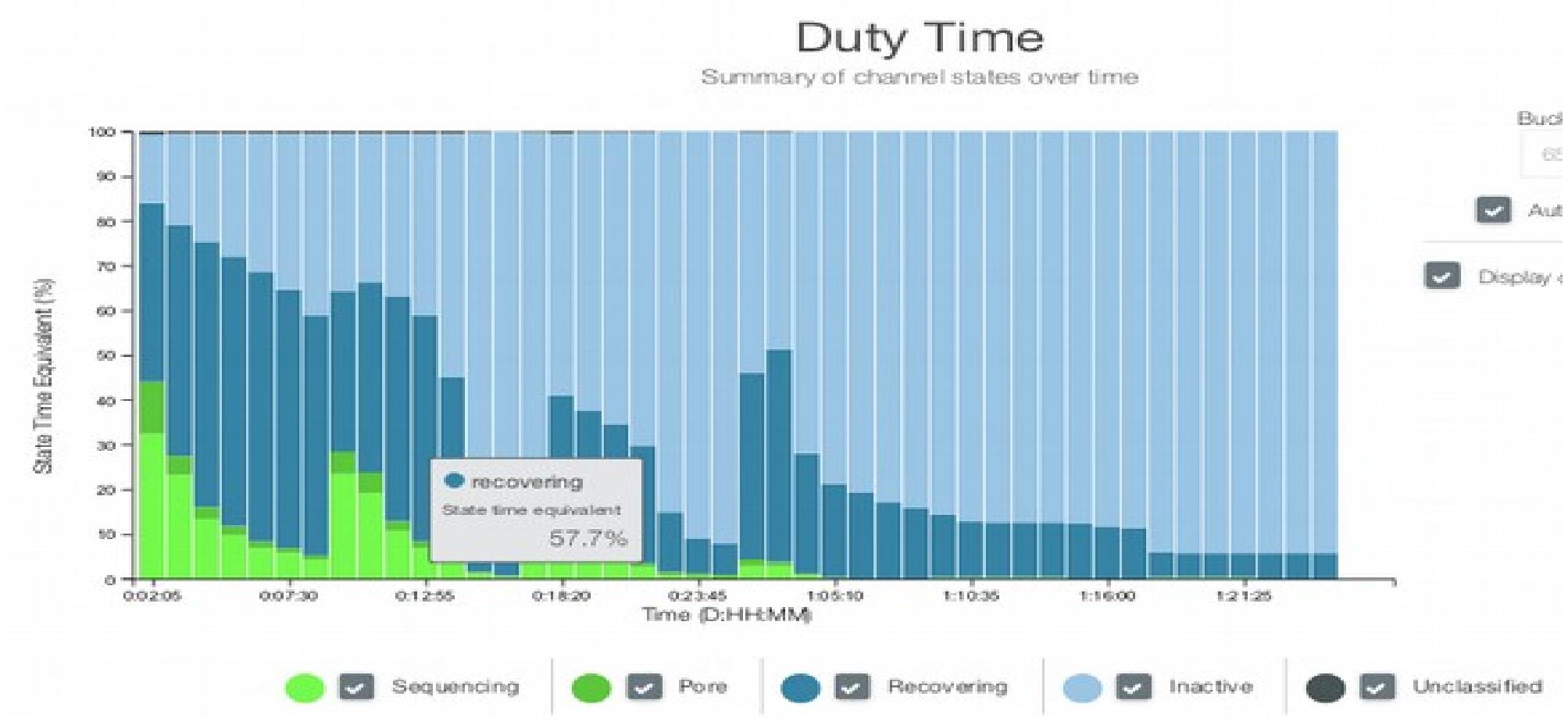
future: chiron



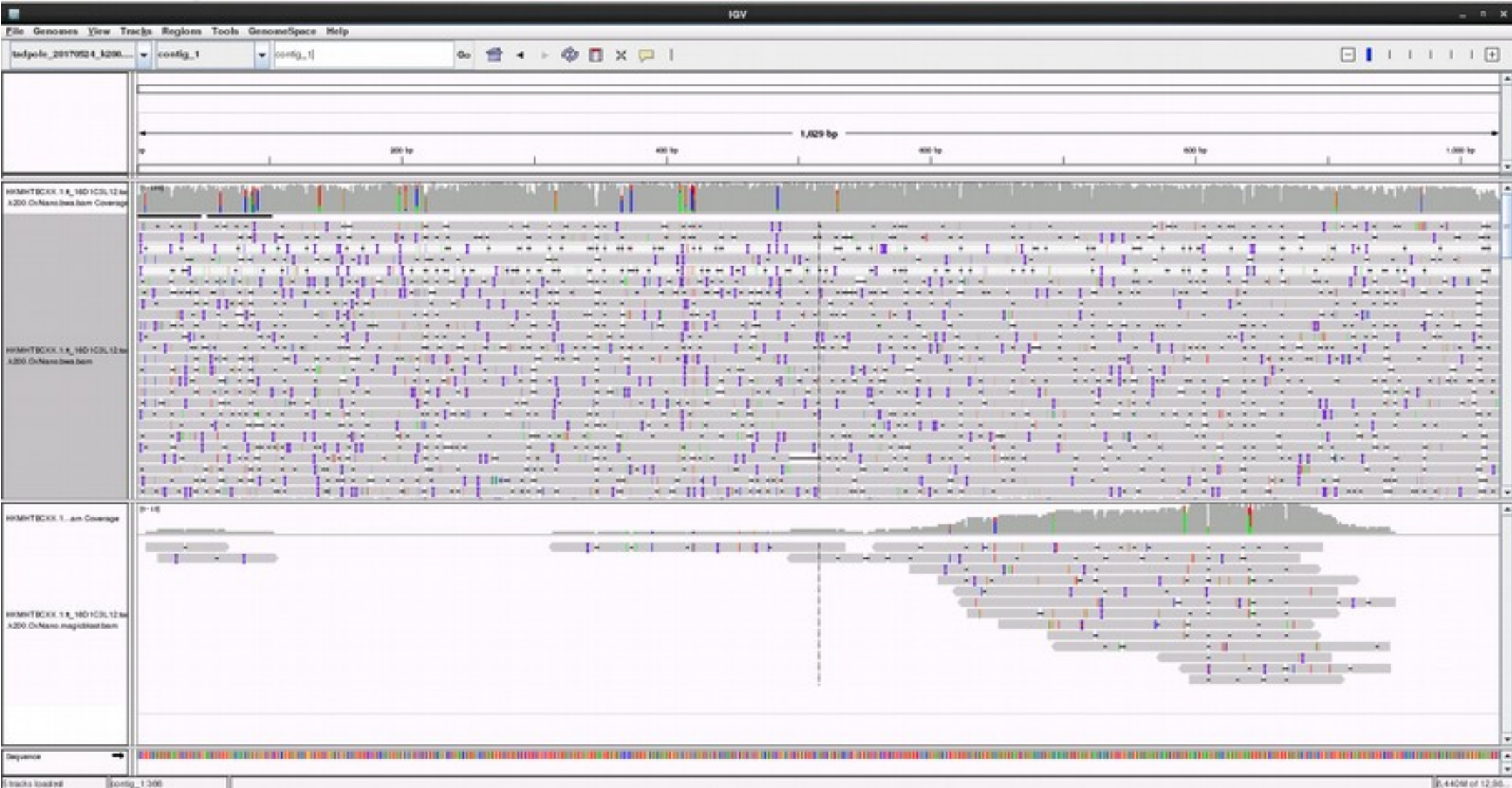
2018: Oxford Nanopore sequencing for 1 day 21 hrs (SEKA1)



2018: Oxford Nanopore sequencing for 1 day 21 hrs (SEKA1)



Raw quality of OxfordNanopore 1D reads after alignment to an Illumina-based contig



Vladimir Nikolić (IT4I, SoHPC student)
Karel Janko (ÚŽFG AV ČR)
Jan Kočí (OSU)
Jan Roslein (OSU)
Oldřich Bartoš (ÚŽFG AV ČR)
Vladimír Beneš (EMBL)
Dinko Pavlinič (EMBL)

Jan Pačes (ÚMG AV ČR)

Petr Pajer (ÚMG AV ČR)

Thank vou!

IT4Innovations national supercomputing center

Several very good assemblies were prepared by abyss-2.0.2 using error-corrected reads using different k-mer values

n	n:500	L50	LG50	NG50	min	N50	E-size	max	sum
3282085	479704	87706	162732	1801	500	2936	4112	57779	921.7e6
2524721	320183	47968	79411	3892	500	5684	7801	84004	972.2e6
2383961	**195687**	4463	8913	23272	500	48678	83684	**910150**	**968.8e6**

n	n:500	L50	LG50	NG50	min	N50	E-size	max	sum
2809379	496455	86074	141270	2144	500	3151	4426	56465	980.4e6
2130364	332146	47269	70164	4540	500	6090	8324	109348	1.027e9
1988503	**207153**	4441	7594	29632	500	51296	89693	**730307**	**1.023e9**

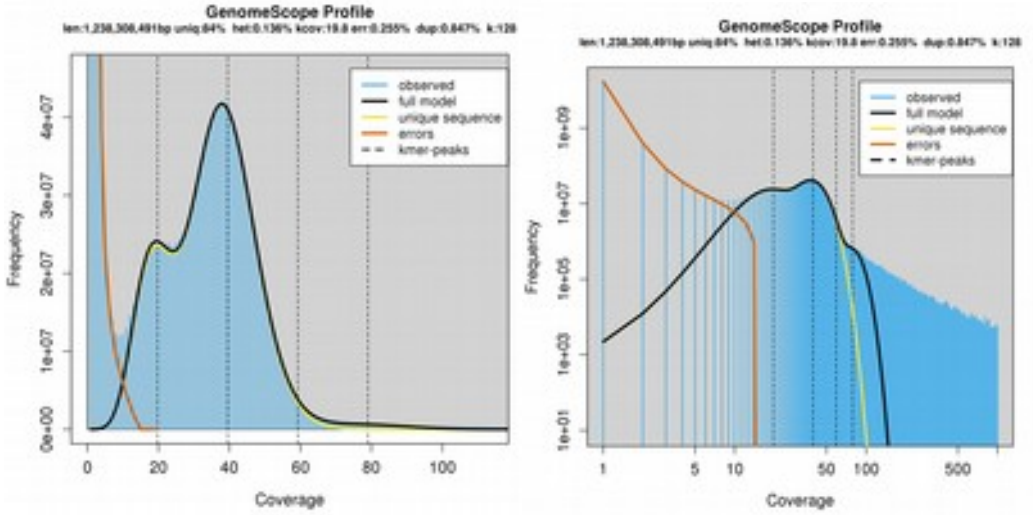
n	n:500	L50	LG50	NG50	min	N50	E-size	max	sum
2411515	516769	85211	125222	2482	500	3345	4688	57811	1.037e9
1807627	348775	47265	63826	5107	500	6370	8717	82251	1.078e9
1665496	**223249**	4585	6863	34706	500	51976	91040	**724617**	**1.074e9**

n	n:500	L50	LG50	NG50	min	N50	E-size	max	sum
1731479	596980	93440	114146	2829	500	3294	4686	60880	1.141e9
1361846	455052	61113	71142	4723	500	5287	7278	98989	1.167e9
1213455	**319826**	6263	7774	30914	500	38588	74019	**751657**	**1.163e9**

N50 is the length of the contig, and L50 is the number of the contigs whose size is N50 or larger. Yes it's weird, but that's the way it is.
See https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics
and http://quast.bioinf.spbau.ru/manual.html#sec3.1.1

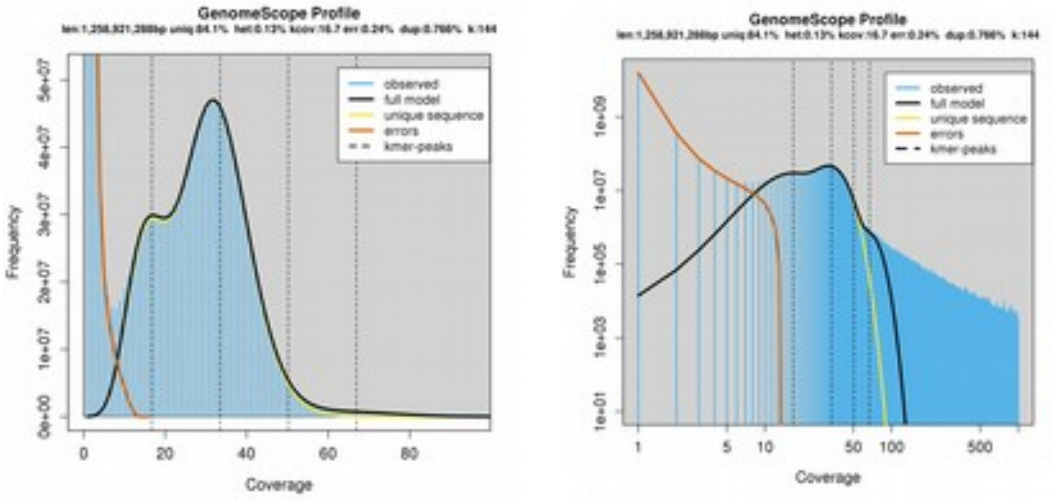
tt_16D1C3L12__PE-only__ntCard_k128.histo http://qb.cshl.edu/genomescope/analysis.php?code=HBnsyGEI1MNQDu1sB3fW

```
GenomeScope version 1.0
k = 128
property                  min                 max
Heterozygosity            0.135367%           0.136532%
Genome Haploid Length     1,236,859,142 bp    1,238,308,491 bp
Genome Repeat Length      197,837,241 bp      198,069,067 bp
Genome Unique Length      1,039,021,900 bp    1,040,239,424 bp
Model Fit                 96.0452%            98.0828%
Read Error Rate           0.255411%           0.255411%
```



tt_16D1C3L12__PE-only__ntCard_k144.histo http://qb.cshl.edu/genomescope/analysis.php?code=2ZkJ2eyTW83ZdmDl98mk

```
GenomeScope version 1.0
k = 144
property                  min                 max
Heterozygosity            0.129235%           0.13013%
Genome Haploid Length     1,257,611,366 bp    1,258,921,288 bp
Genome Repeat Length      200,553,631 bp      200,762,527 bp
Genome Unique Length      1,057,057,735 bp    1,058,158,761 bp
Model Fit                 96.5393%            98.5629%
Read Error Rate           0.239662%           0.239662%
```



tt_16D1C3L12__PE-only__ntCard_k156.histo http://qb.cshl.edu/genomescope/analysis.php?code=2ZkJ2eyTW83ZdmDl98mk

GenomeScope version 1.0
k = 156

property	min	max
Heterozygosity	0.127693%	0.128482%
Genome Haploid Length	1,267,403,131 bp	1,268,723,749 bp
Genome Repeat Length	201,426,908 bp	201,636,793 bp
Genome Unique Length	1,065,976,223 bp	1,067,086,957 bp
Model Fit	96.8822%	98.8092%
Read Error Rate	0.230913%	0.230913%