Two examples of hard to assemble genomes even with long reads

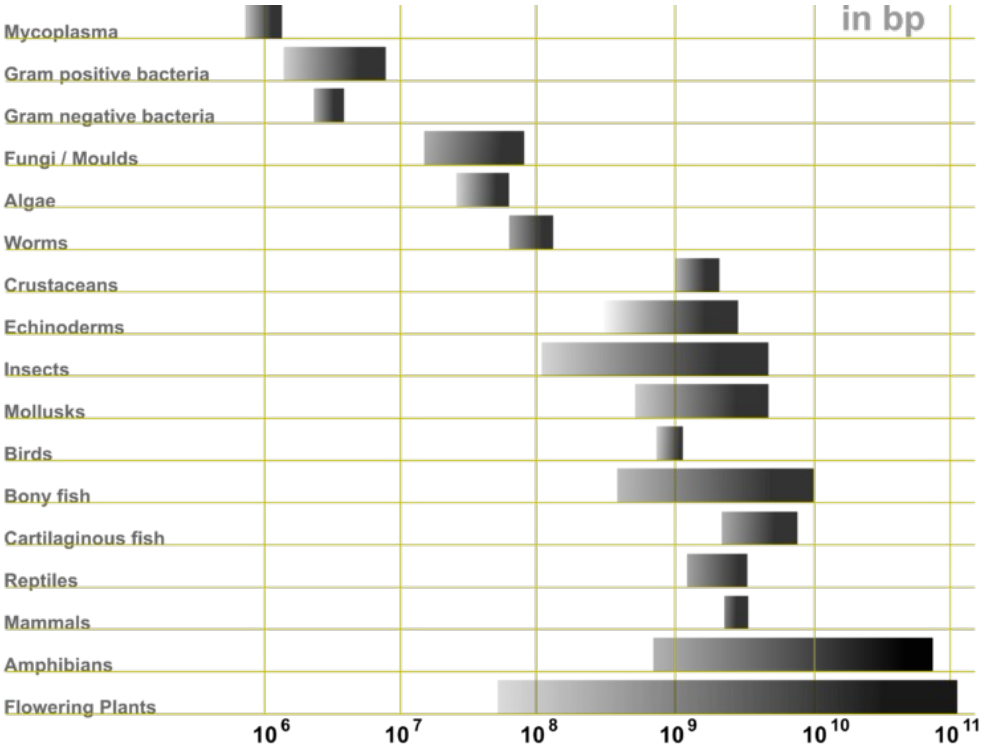
Christophe Klopp

http://bioinfo.genotoul.fr/
http://www.sigenae.org/

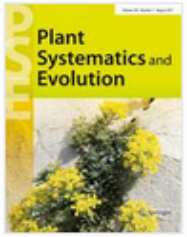
http://get.genotoul.fr
get@genotoul.fr
@get_genotoul

What makes genomes difficult to assemble ?

- Genome size
- Number of chromosomes
- Repeat content
- Repeat size (structure)
- Heterozygocity
- Ploïdy
- DNA conformation
- Contamination



https://en.wikipedia.org/wiki/Genome_size



Plant Systematics and Evolution
August 2017, Volume 303, Issue 7, pp 981–986 | Cite as

The largest fungal genome discovered in *Jafnea semitosta*

Authors | Authors and affiliations

Zuzana Egertová, Michal Sochor

Short Communication
First Online: 03 May 201

3.6 Gb

Abstract

Jafnea semitosta is an ascomycete (Pyronemataceae, Pezizales) originating from North America and spreading uncommonly in Europe. Its genome size was measured via flow cytometry of fruiting bodies from five localities in the Czech and Slovak Republic. The nuclear 1*C* DNA content was estimated at 3.706 ± 0.011 pg (~3.625 ± 0.011 Gbp) which represents the highest value ever reported for fungi and 100× higher than the average. Generally, the genome inflation in fungi appears to be driven mainly by proliferation of repetitive sequences, but polyploidy should also be considered in further studies on this greatly unexplored topic.

First example : Ganoderma boninense

- Long lasting project with CIRAD biologists in Montpellier
- Palm tree and coconut pest
- Incurable Basal Stem Rot disease
- No public genome available

http://mushroomobserver.org/observer/show_observation/158919

Ganoderma lucidum

- medicinal mushroom in traditional Chinese medicine
- Immune building properties
- **monokaryotic strain** G.260125-1 obtained by protoplasting
- 43.3-Mb genome
- 16,113 predicted genes

https://en.wikipedia.org/wiki/Lingzhi_mushroom

First tests

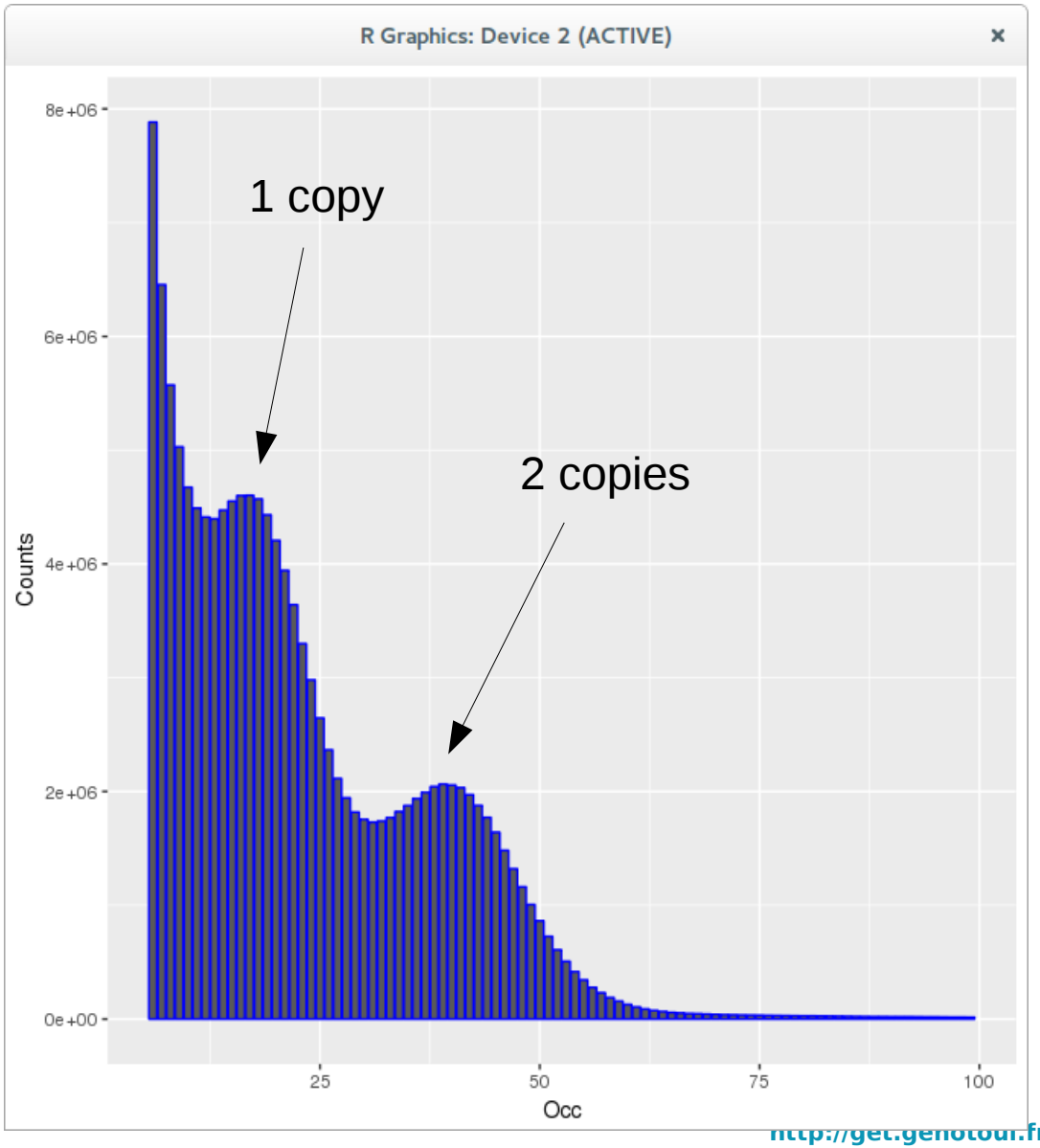
- 454
- Illumina MiSeq & HiSeq (paired-ends and mate pairs)
- PacBio P4C2

	454	+ HiSeq MP	+ PacBio
software	**Newbler**	**SSPACE**	**SSPACE-LR**
version	**2012**	**2.0**	**1.1**
Number of contigs	16,451	3,363	3,284
Nb of contigs not in scaffolds	2,278	2,203	1,948
Total size of contigs	60,455,132	75,263,670	75,167,085
N50 contig length	5,749	61,481	62,696
L50 contig count	2,783	337	330

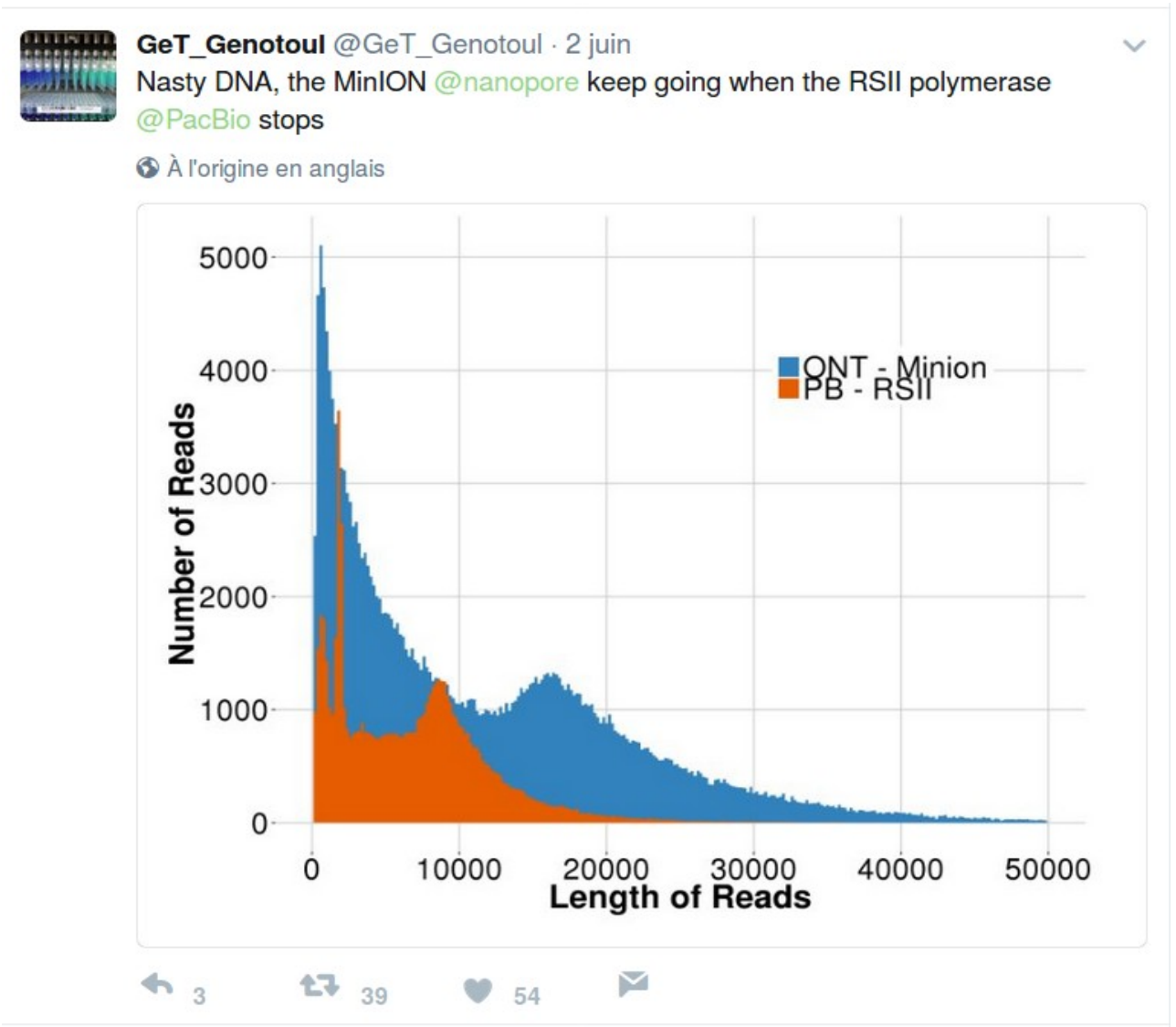
PacBio P6C4 assembly

	MiSeq	PacBio P6C4
software	**SSPACE-LR**	**CANU**
version	**1.1**	**1.5**
Number of contigs	3,284	915
Number of contigs not in scaffolds	2,203	915
Total size of contigs	75,263,670	84,137,414
N50 contig length	61,481	166,220
L50 contig count	337	136

Raw reads kmer content



Same DNA : ONT vers PacBio

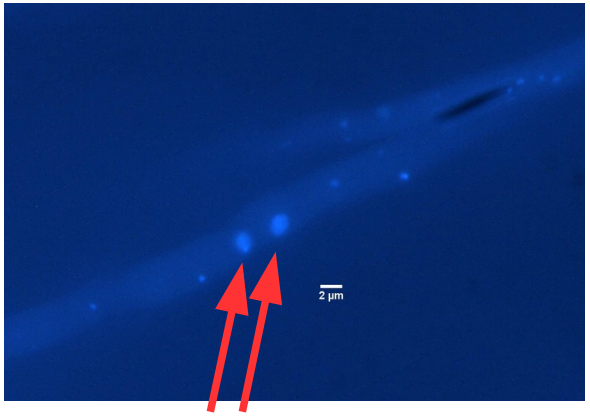


Nanopore assembly

https://github.com/marbl/canu/issues/569#issuecomment-319169640

	ONT MinION
software	**canu**
version	**1.5**
Number of contigs	400
Total size of contigs	67 Mb
N50 contig length	323 Kb
L50 contig count	50

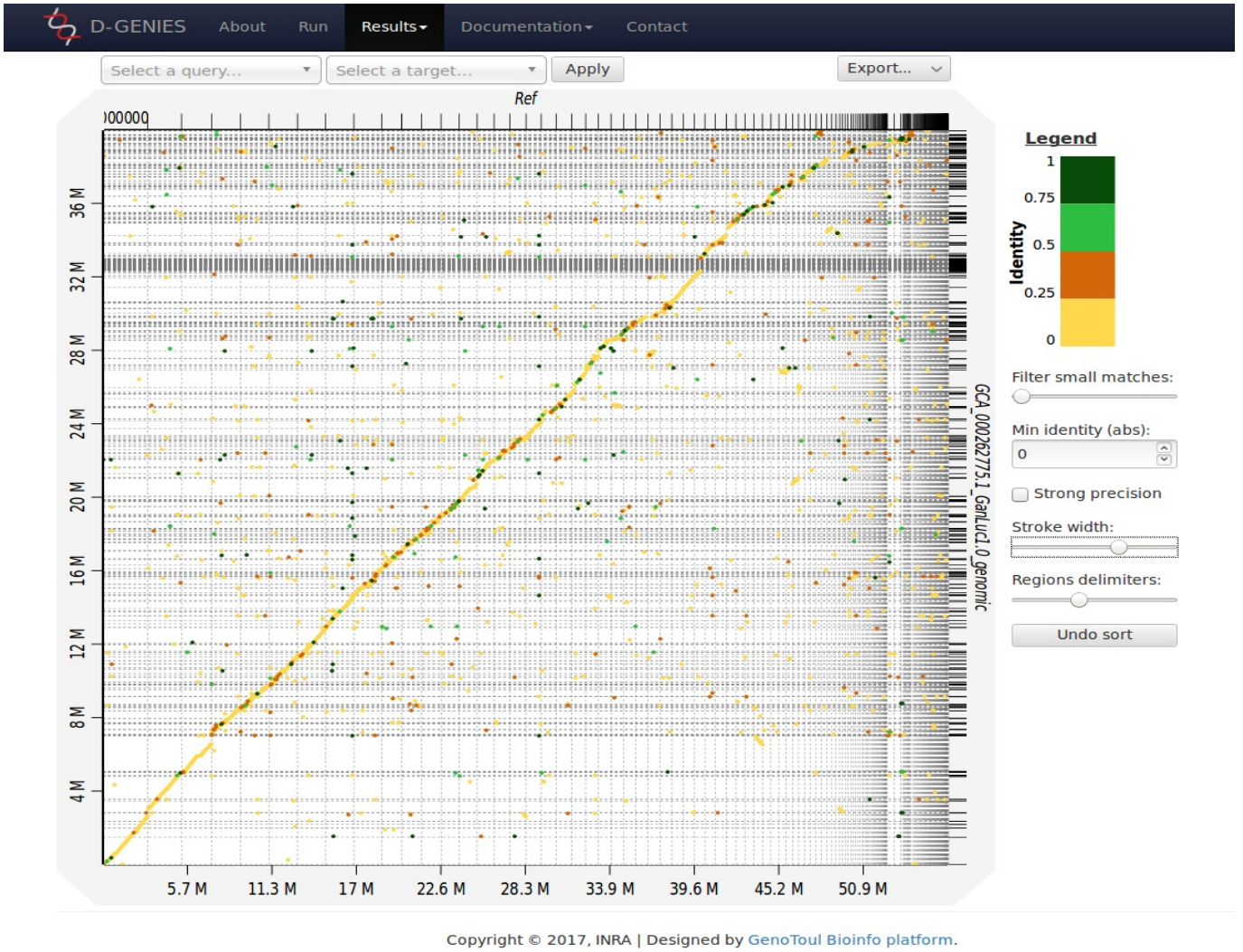
Haplomerging



dikaryotic stage

	PacBio + Nanopore	PacBio + Nanopore	PacBio + Nanopore
software	**CANU**	**haplomerger**	**haplomerger**
version	**1.5**	**2.0**	**2.0**
Number of contigs	436	218	218
Number of contigs not in scaffolds	436	218	218
Total size of contigs	76,821,474	56,529,789	49,372,845
N50 contig length	323,663	**1,068,631**	**889,398**
L50 contig count	55	**17**	**18**

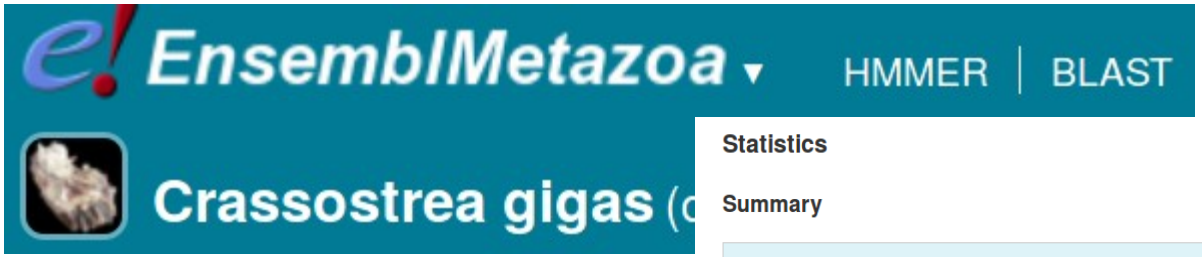
Assembly validations



BUSCO Gl: C:93.7%[S:93.0%,D:0.7%],F:5.6%,M:0.7%,n:1335
Gb: C:94.8%[S:13.4%,D:81.4%],F:4.1%,M:1.1%,n:1335

Second example : Crassostrea gigas

- New project with IFREMER biologists
- Estimated genome size 650 Mb



Statistics

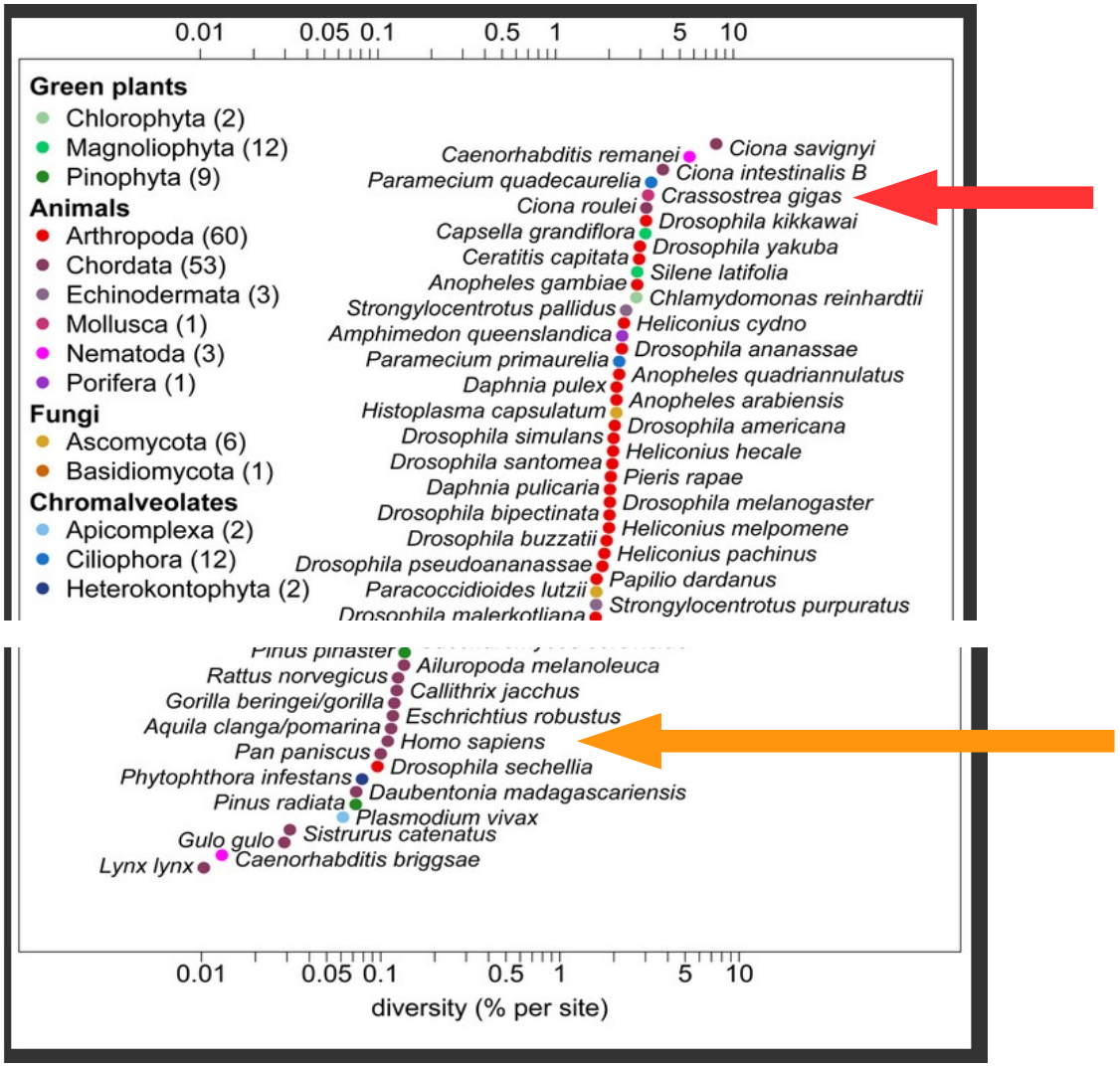
Summary

Assembly	oyster_v9, INSDC Assembly GCA_000297895.1, Sep 2012
Database version	90.1
Base Pairs	491,850,583
Golden Path Length	557,717,710
Genebuild by	ENA
Genebuild method	Imported from ENA
Data source	GigaDB

Gene counts

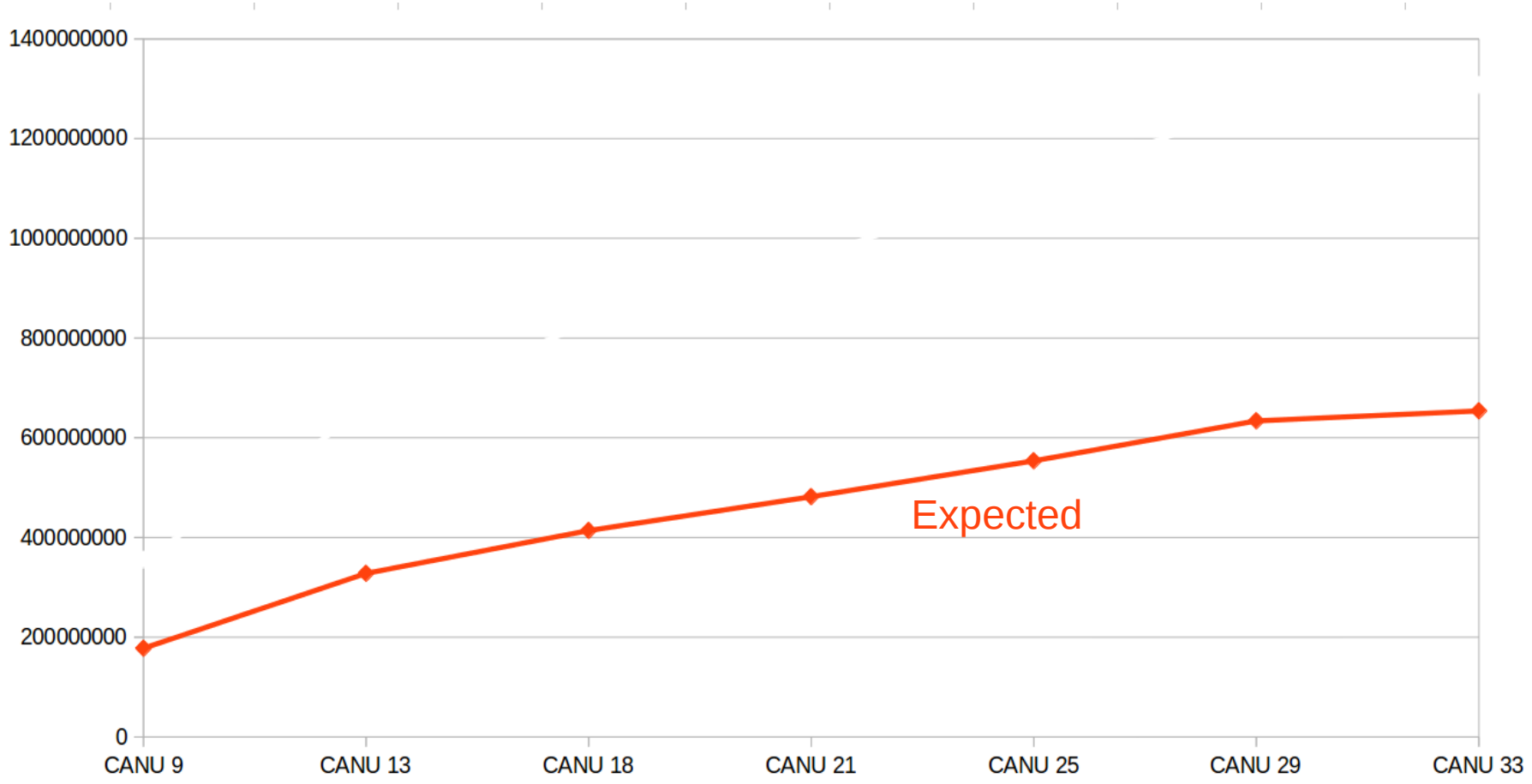
Coding genes	26,101
Non coding genes	497
Small non coding genes	494
Long non coding genes	3
Gene transcripts	26,598

Heterozygosity

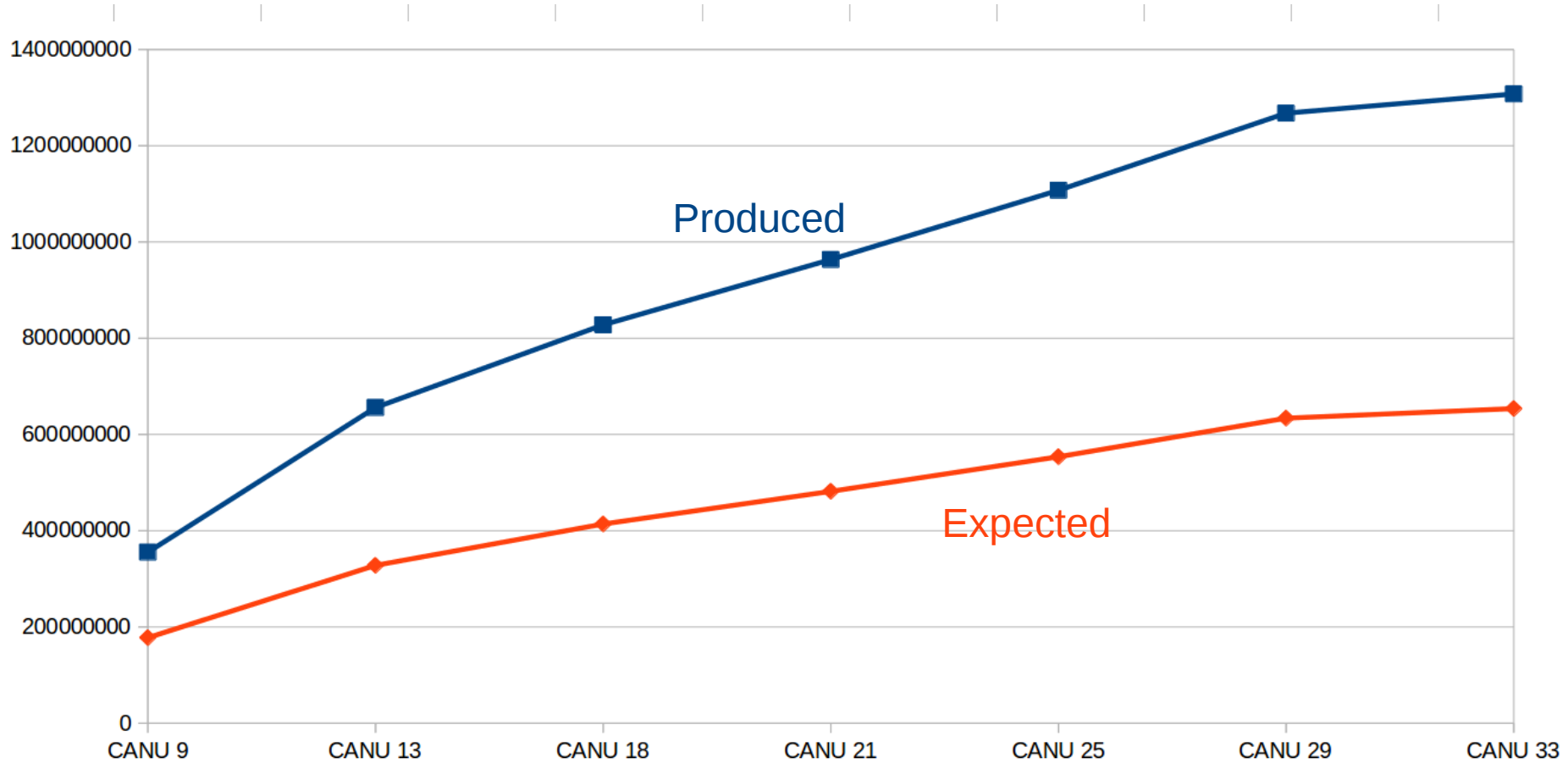


http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001388

assemblies sizes evolution



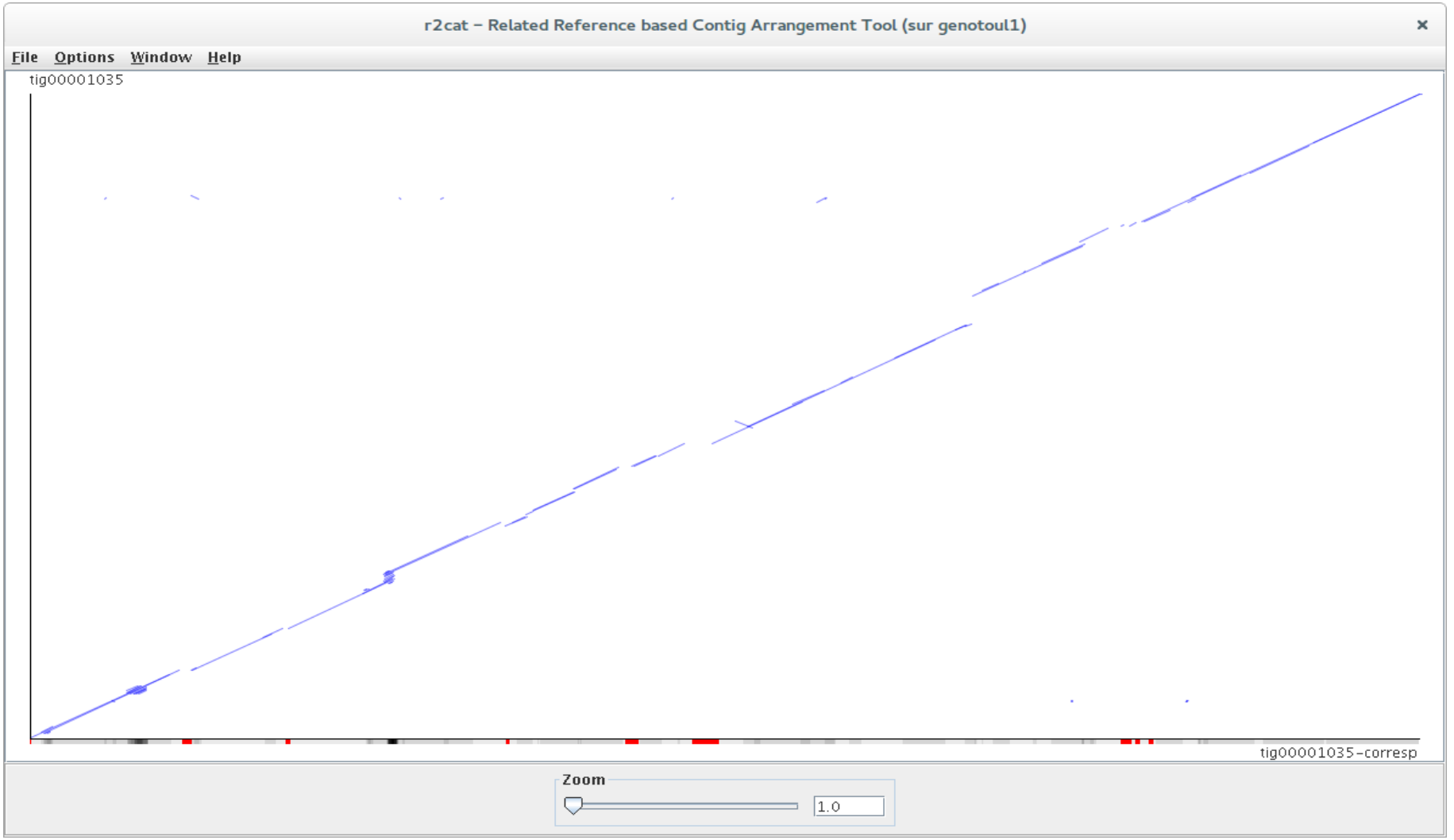
PacBio canu assemblies sizes



Comparing to reference

metrics	reference	CANU PacBio
Number of contigs	29,416	**19,188**
Total size of contigs	491,868,483	1,307,520,554
Longest contig	233,101	**2,388,545**
Shortest contig	200	1,158
Mean contig size	16,721	**68,143**
Median contig size	9,743	**45,702**
N50 contig length	32,564	**86,828**
L50 contig count	4,448	**3,779**

Canu contig realignment



tig00001035 unique match 57kb

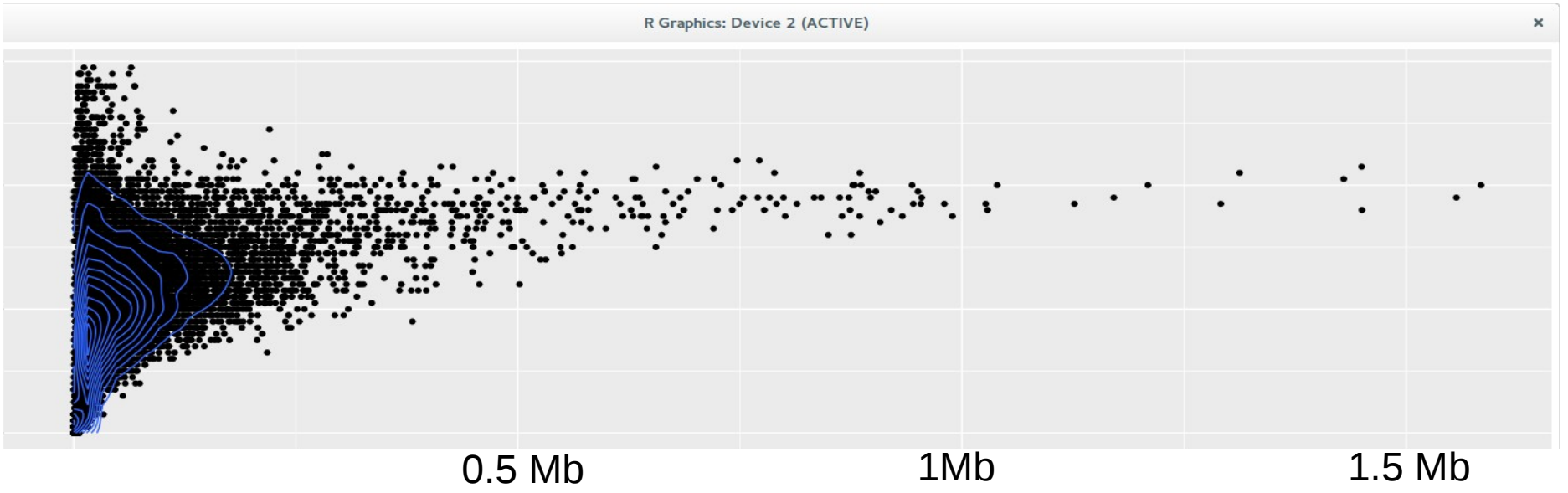
haplomerging

metrics	CANU PacBio	Ref	Alt
Number of contigs	19,188	18,714	18,714
Total size of contigs	1,307,520,554	1,286,245,159	1,283,968,615
Longest contig	2,388,545	2,388,545	2,388,545
Shortest contig	1,158	735	735
Mean contig size	68,143	68,732	68,610
Median contig size	45,702	45,881	45,874
N50 contig length	86,828	88,269	88,036
L50 contig count	3,779	3,641	3,651

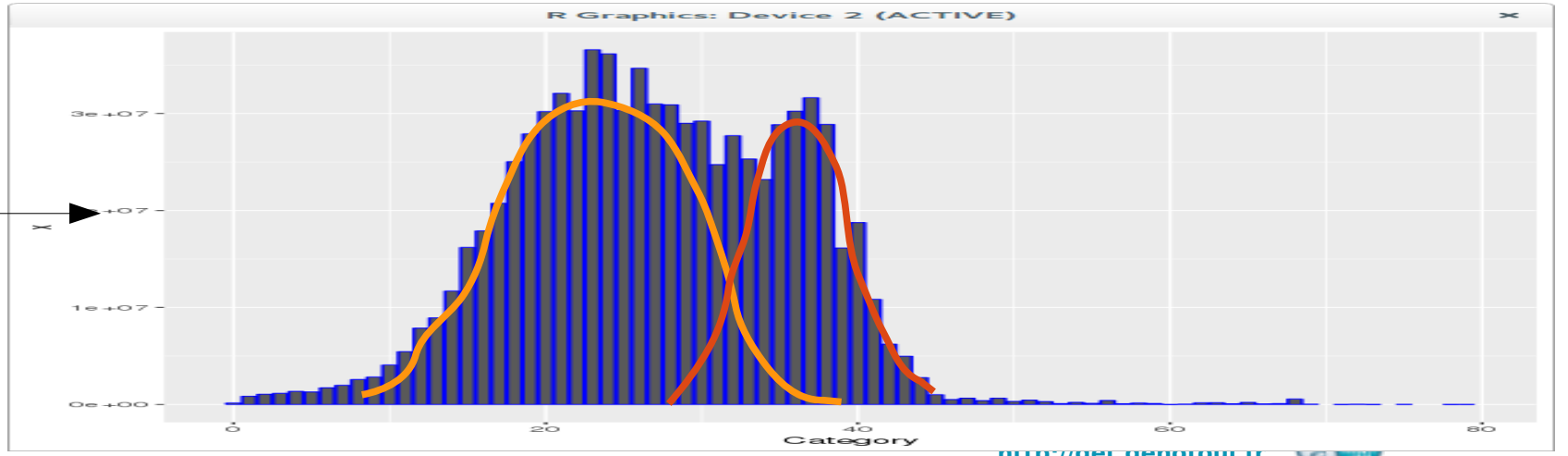
Other assemblies

metrics	reference	Falcon	smartdenovo	miniasm
Number of contigs	29,416	12,023	5,896	**4,651**
Total size of contigs	491,868,483	**806,703,370**	434,836,206	131,352,830
Longest contig	233,101	**1,584,246**	1,243,994	160,717
Shortest contig	200	88	9,298	756
Mean contig size	16,721	67,097	**73,751**	28,242
Median contig size	9,743	35,898	**44,482**	25,117
N50 contig length	32,564	**126,839**	110,701	30,840
L50 contig count	4,448	1,581	**1,045**	1,482

Falcon read realignments



2Mb



Conclusions

- In all cases long reads will help.
- Try to get information about the genome, related genomes, the DNA,...
- Heterozygocity can have a massive impact on assembly metrics.
- It is not because your genome seems simple to assemble that it will work out of the box.

Acknowledgments

- **CIRAD : Ganoderma boninense**
  - Letizia Camus
  - Frédéric Breton
- **IFREMER :** projet Gigastore (**Crassostrea gigas**)
  - Pierre-Alexandre Gagnaire (Univ Montpellier)
  - Jean-Baptiste Lamy
- **Get-Plage :**
  - Alain Roulet
  - Céline Lopez-Roques
  - Catherine Zanchetta

  & all the team