Lecture 6

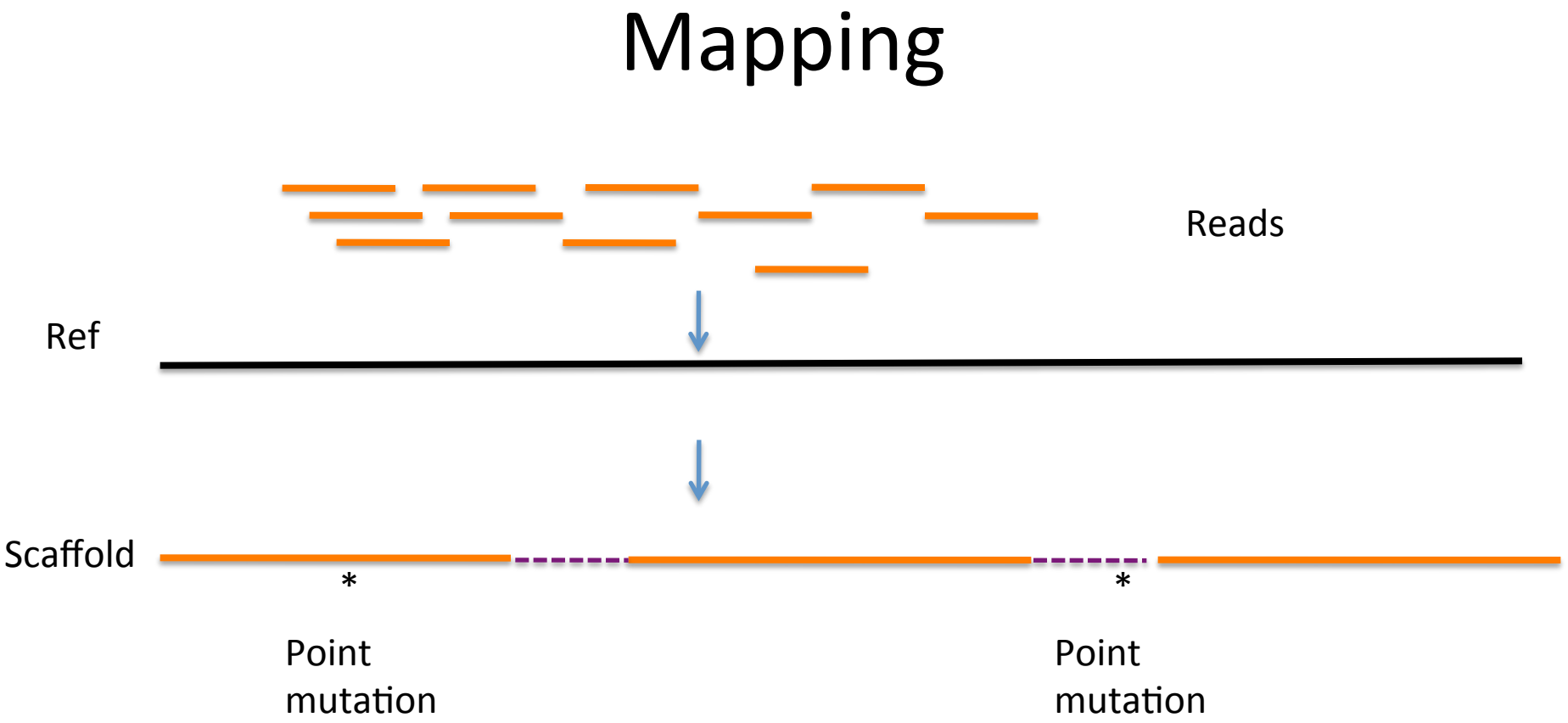
Genome data analysis 2: de novo assembly

Dealing with millions of small reads

Museum of the
Inquisition – Lima

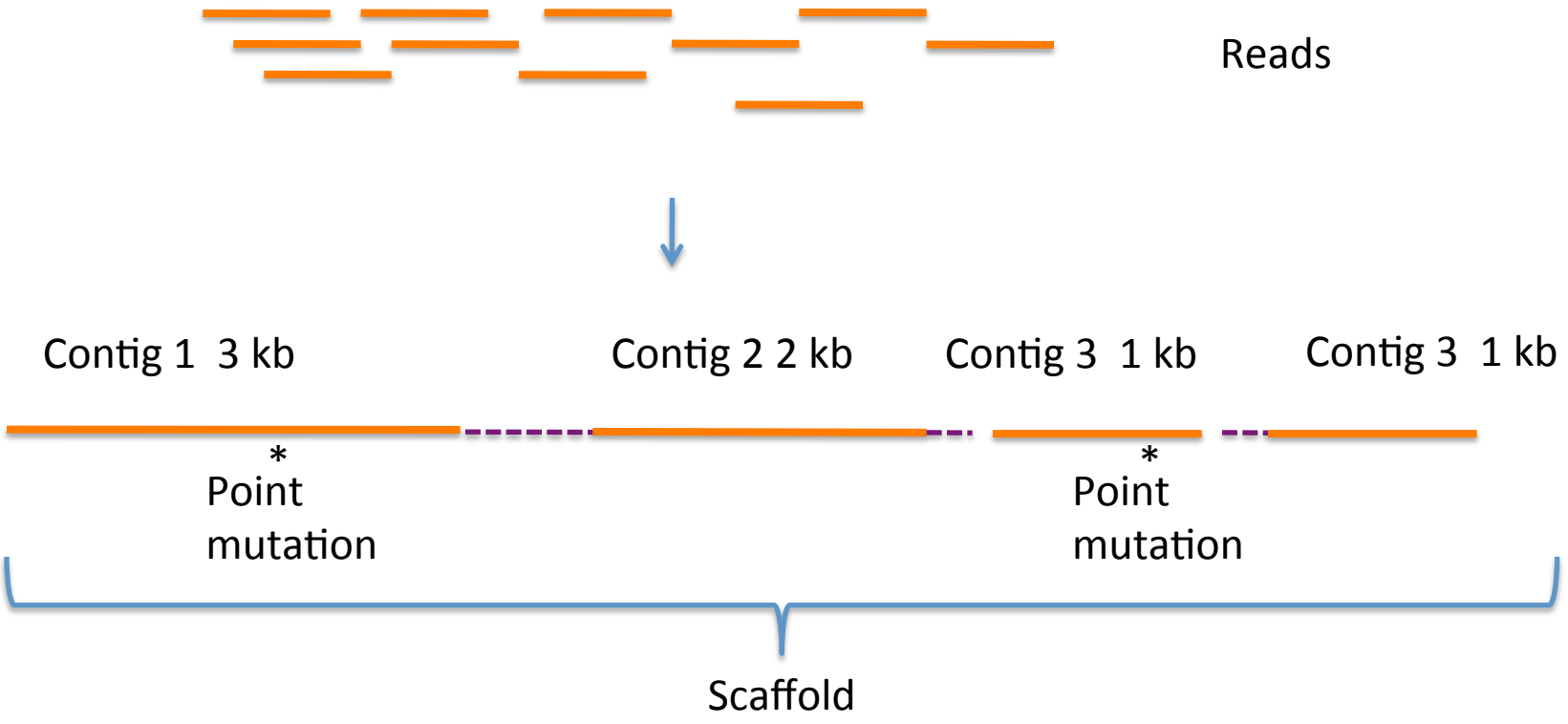
- Dealing with millions of small reads

Dealing with millions of small reads



Dealing with millions of small reads

de novo assembly



N50 is the contig length such that using equal or longer contigs produces half the bases of the assembly. This can be thought of as the point of half of the mass of the distribution.

Coverage = L x N/ G
L = Length reads
N = number of reads
G = Length of the genome

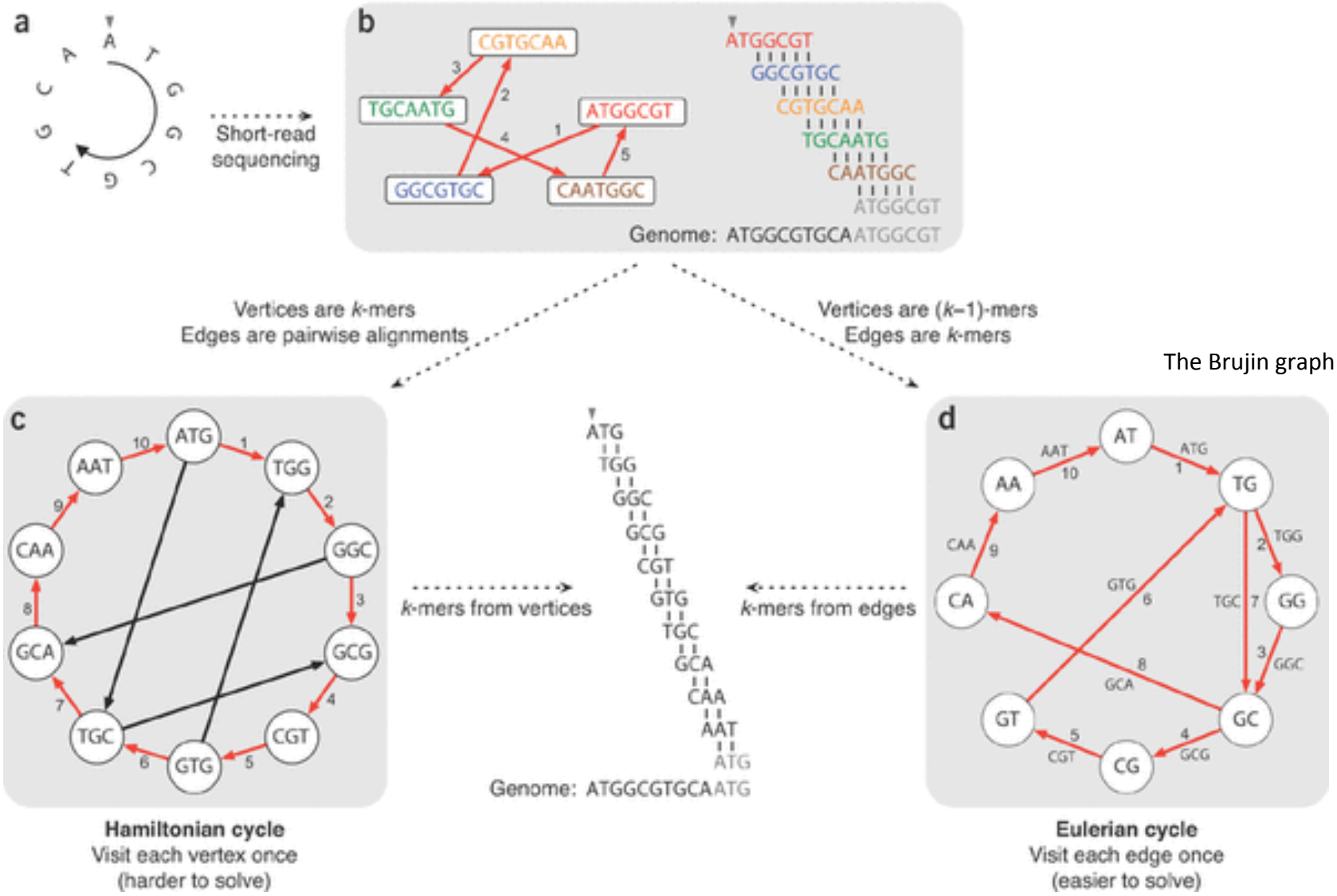
Examples

- We have a collection of contigs with sizes 7, 4, 3, 2, 2, 1, and 1 kb .
- Calculate the N50 length….

Mapping and *de novo* assembly

	Annotation re-sequencing	Point mutation studies	Gene content studies	Computational cost
Mapping	good	good	good	Cheap-moderate
De Novo	-	good	better	Moderate-high

Two strategies for genome assembly

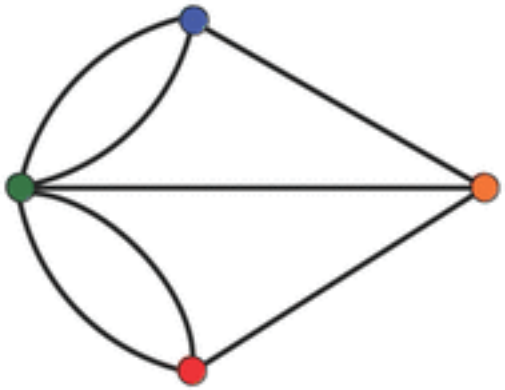


Phillip E C Compeau, Pavel A Pevzner & Glenn Tesler
Nature Biotechnology 29, 987–991 (2011

Bridges of Königsberg problem

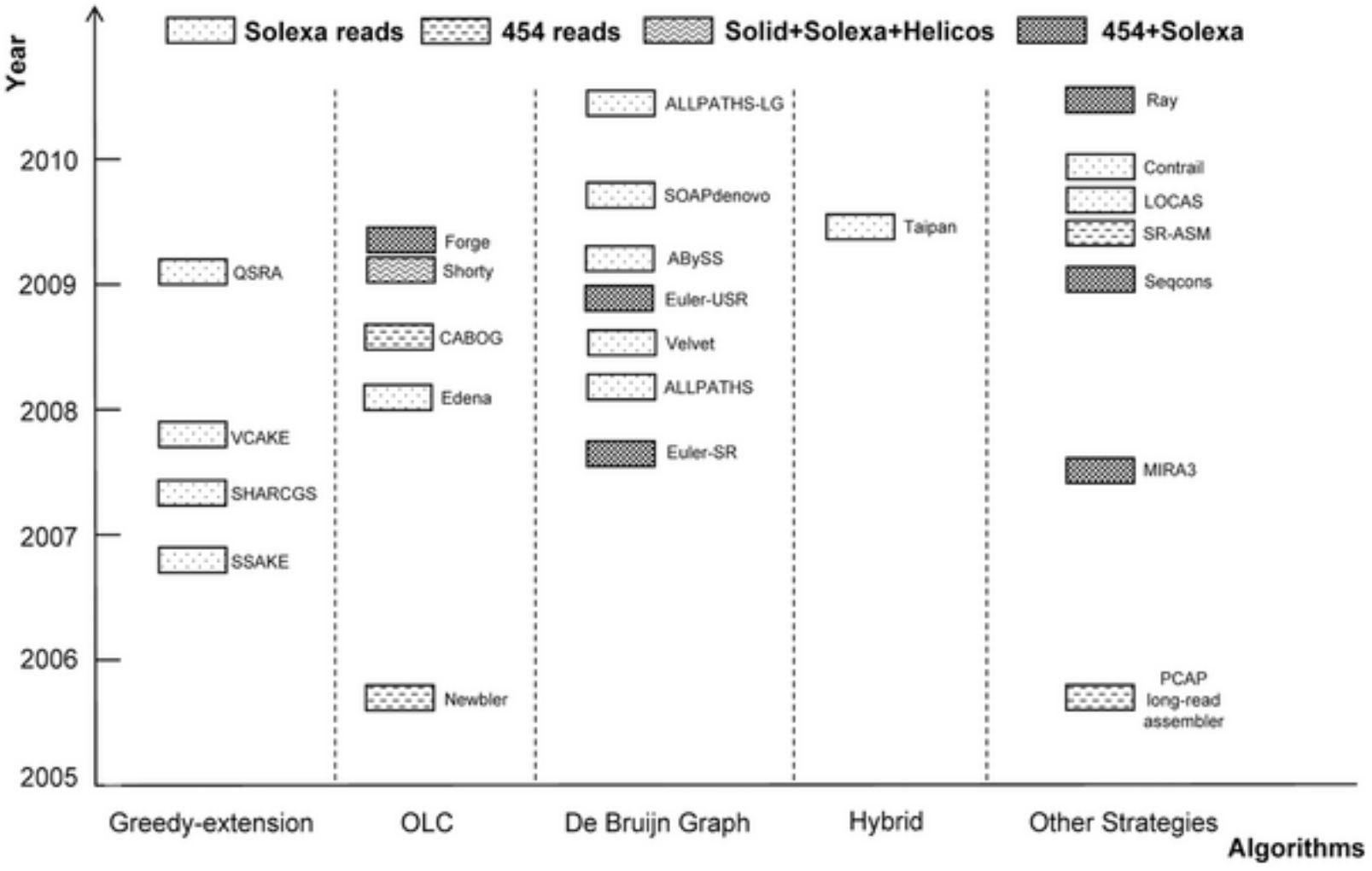
a

b



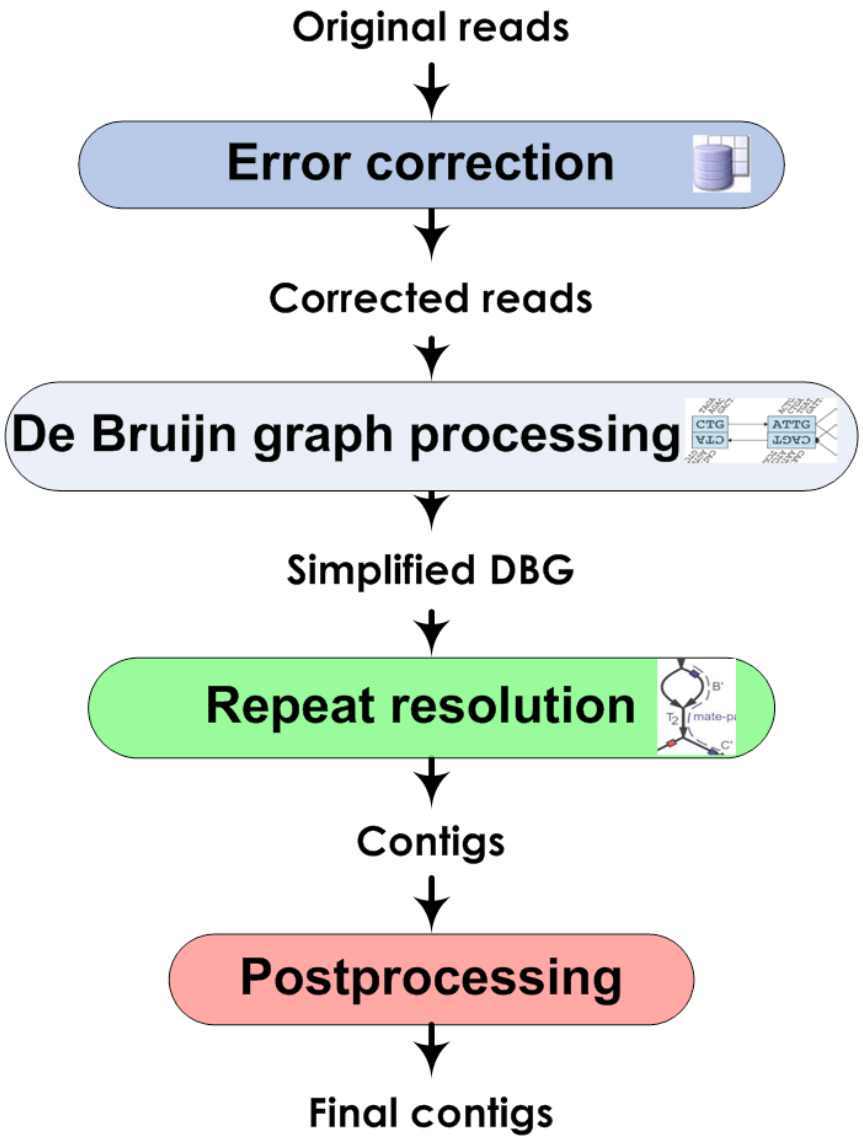
Phillip E C Compeau, Pavel A Pevzner & Glenn Tesler
Nature Biotechnology 29, 987–991 (2011

Figure 1. Overview of de novo short reads assemblers.



Zhang W, Chen J, Yang Y, Tang Y, et al. (2011) A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. PLoS ONE 6(3): e17915. doi:10.1371/journal.pone.0017915
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0017915

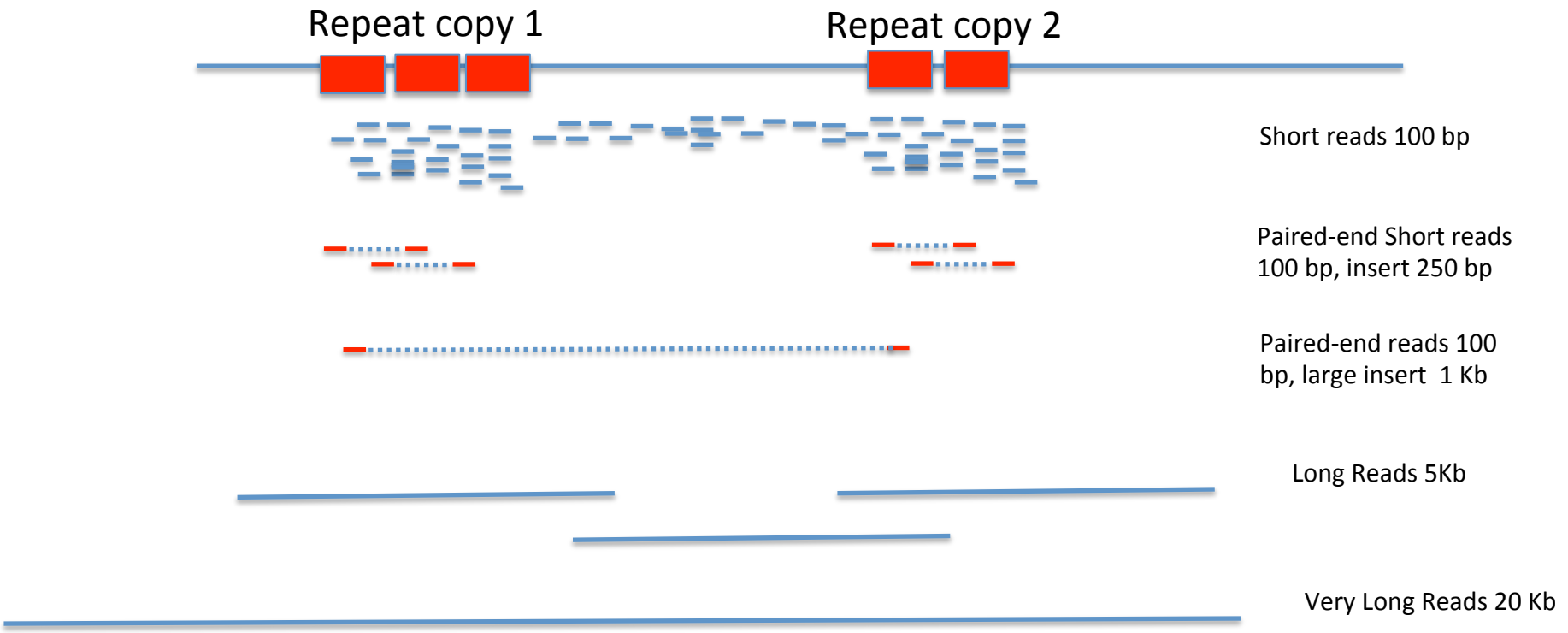
SPADES


Original reads
Error correction
Corrected reads
De Bruijn graph processing
CTG
CTA
ATTG
CAGT
Simplified DBG
Repeat resolution
Contigs
Postprocessing
Final contigs

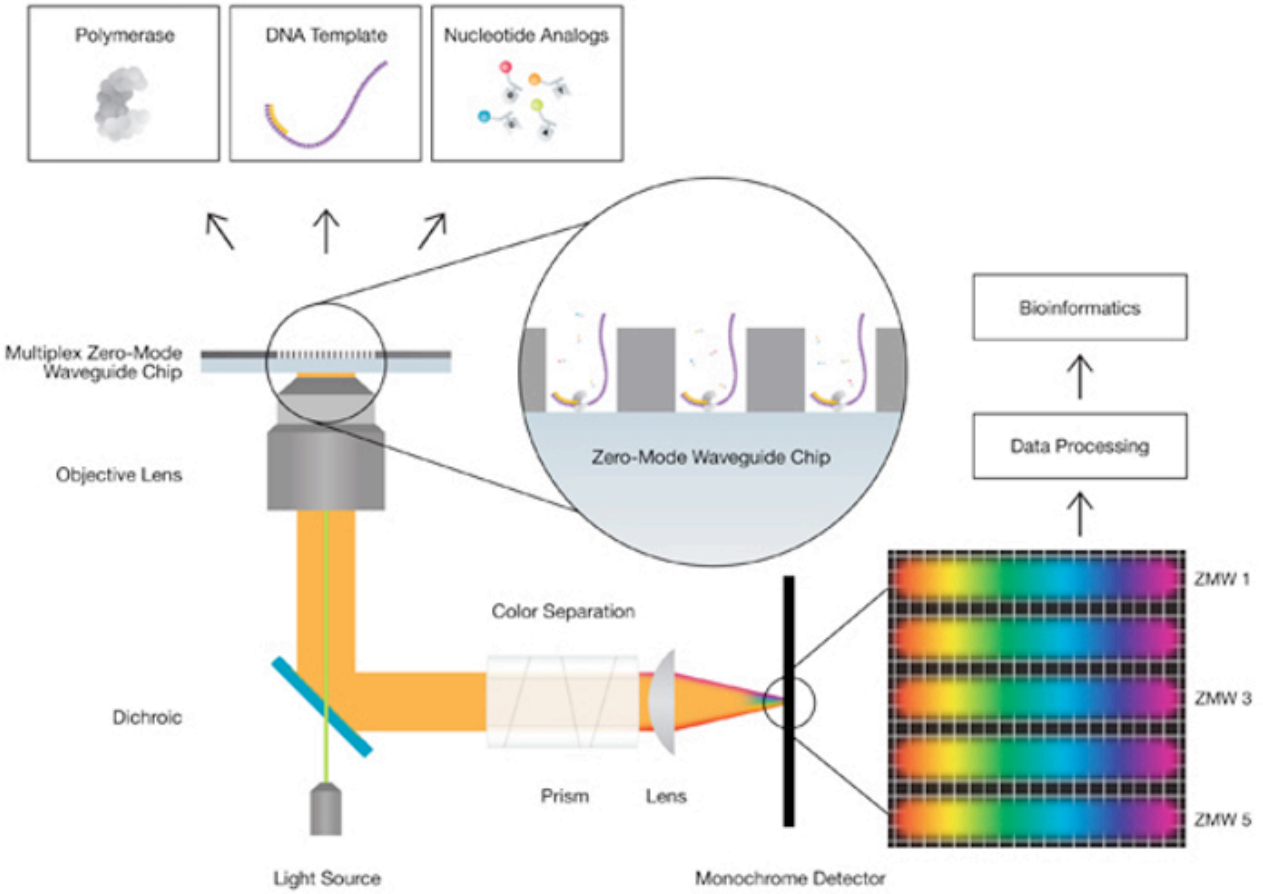
Basic command

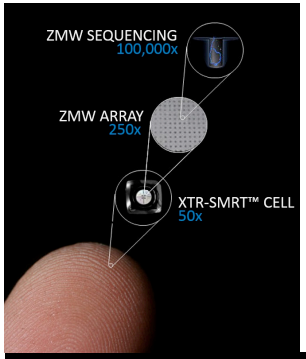
- spades.py -1 read1.fq -2 read2.fq –o assembly

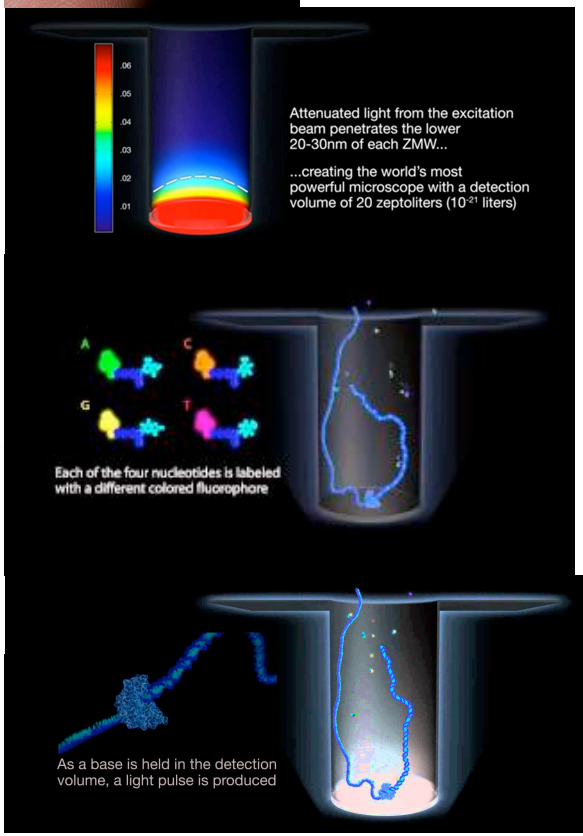
Long reads for assembly
“The return of OLC”



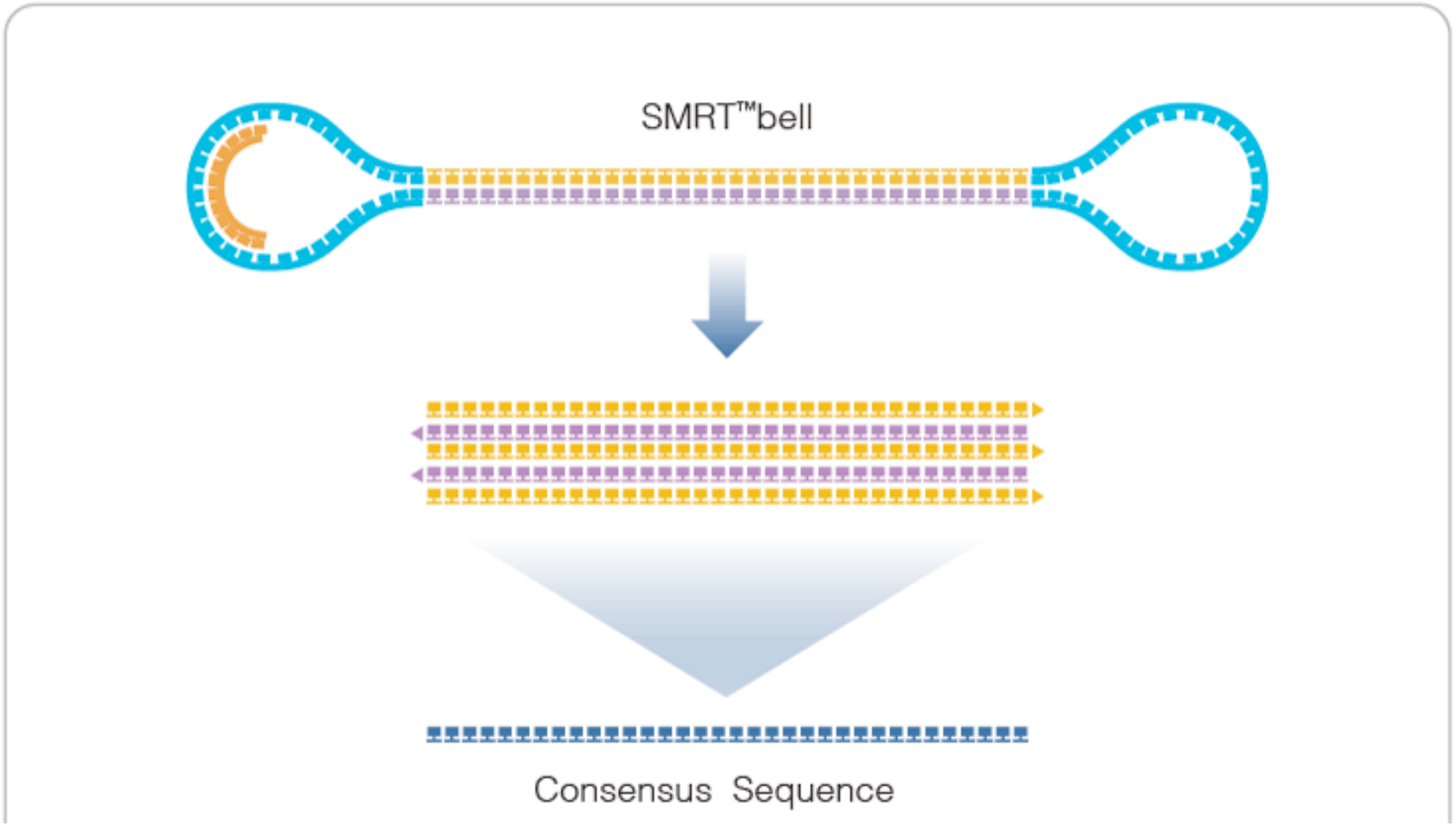
SMRT sequencing –PacBio





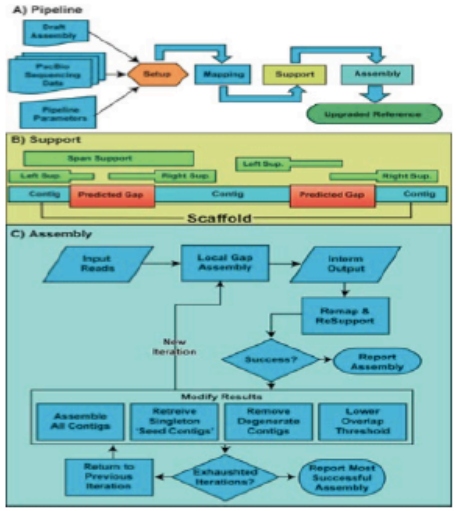


Pacbio results: Circular Consensus Reads (CCS) vs Continuous Long Reads (CLR)



PacBio Assembly Algorithms

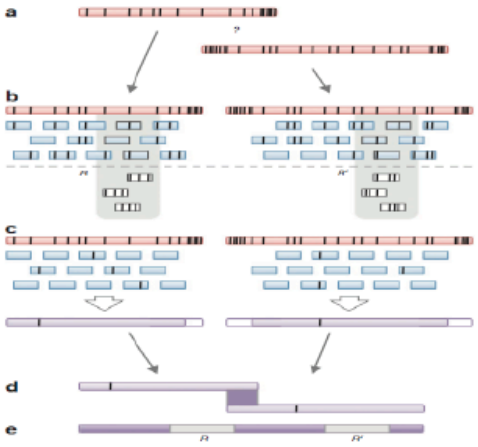
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

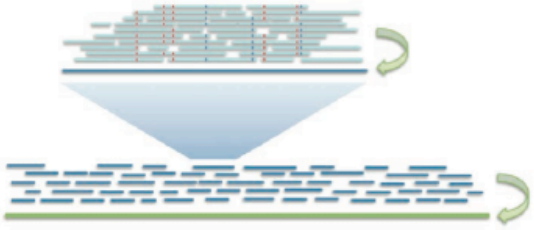
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

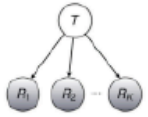
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(*M. ruber* ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

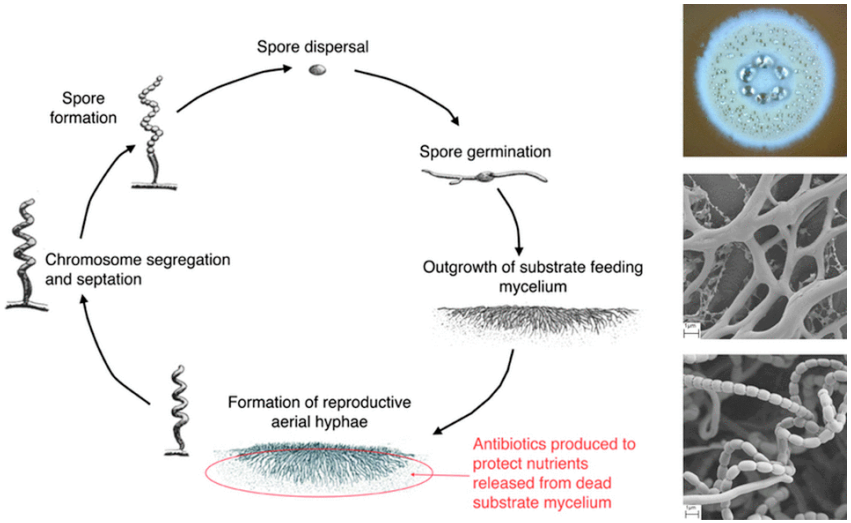
PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

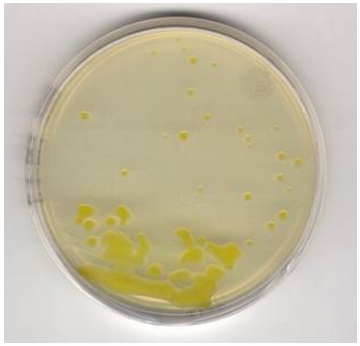
< 5x PacBio Coverage > 50x

SMRT-PacBio

- Two Examples

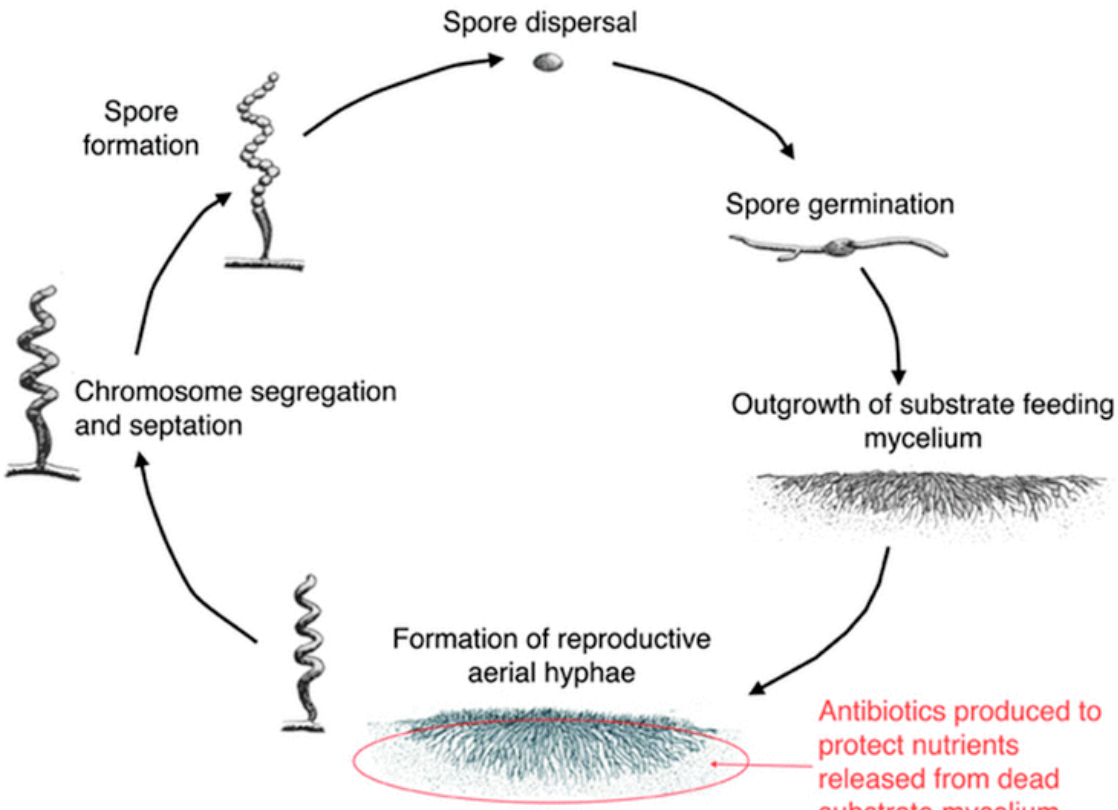


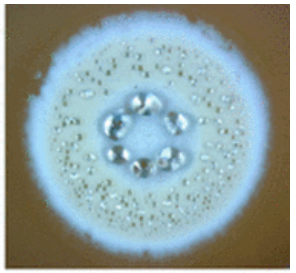
Streptomyces

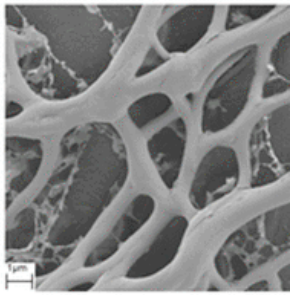


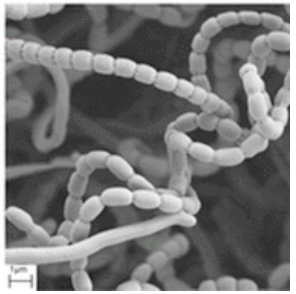
Xanthomonas

Streptomyces species

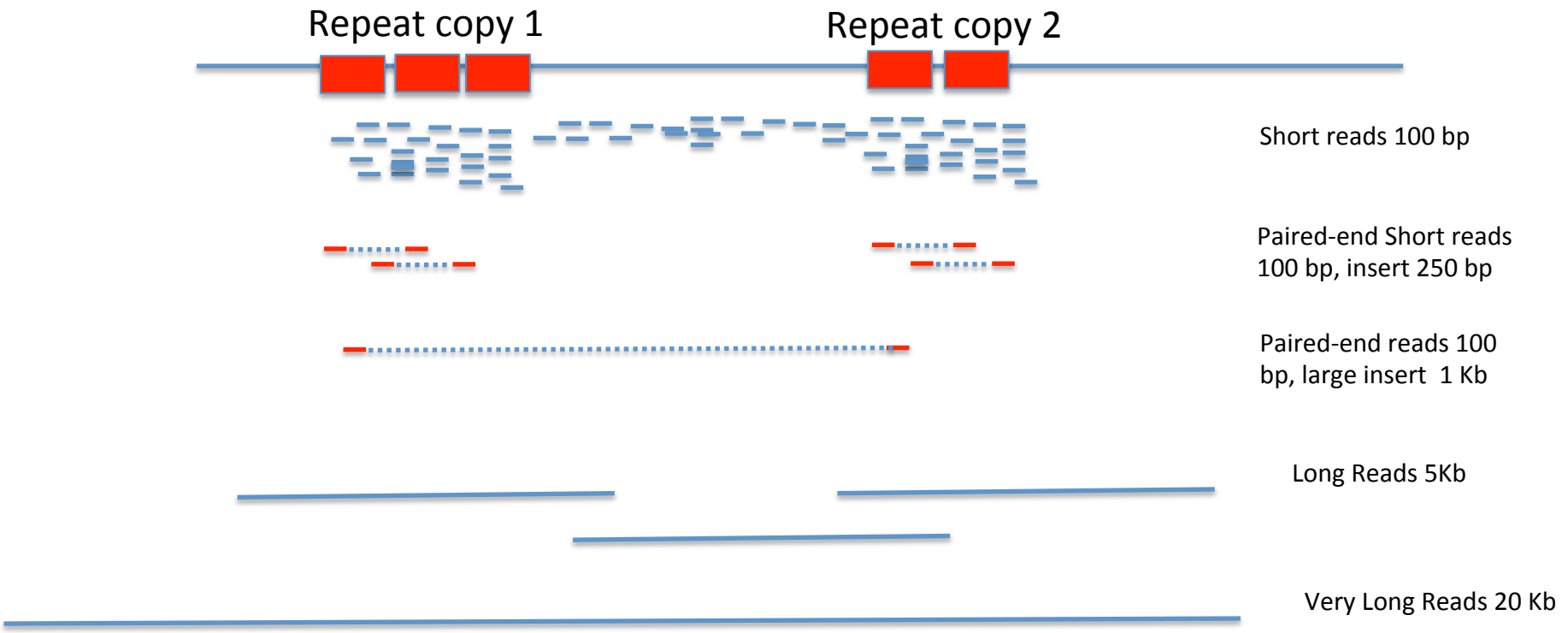








Long reads for assembly



Complete genomes of Streptomyces in Genebank 2015

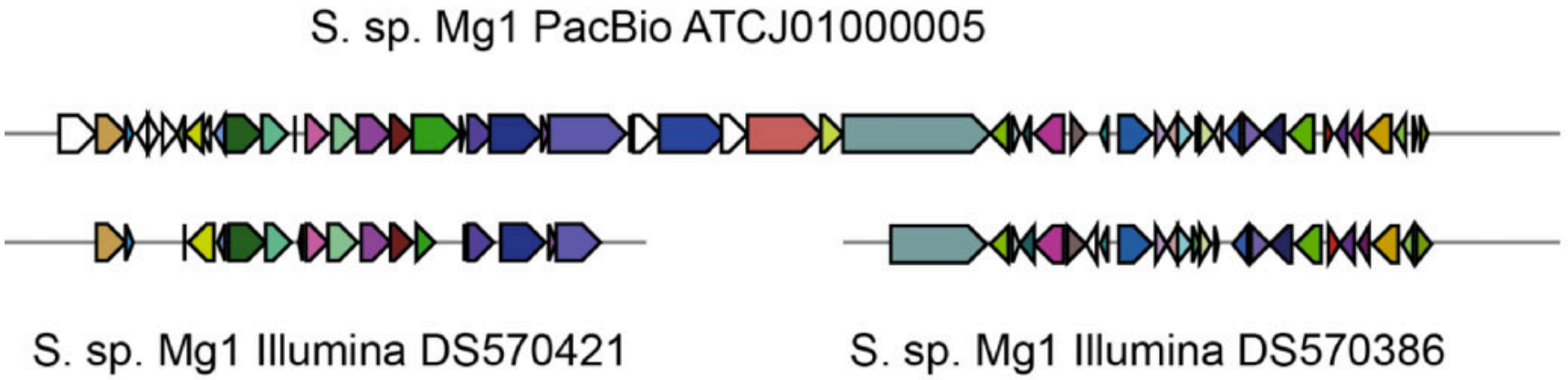
1. Streptomyces albus J1074 (Zaburannyi et al., 2014) CP004370
2. Streptomyces avermitilis (Omura et al., 2001; Ikeda et al., 2003) AP005645, BA000030
3. Streptomyces bingchenggensis BCW 1 (Wang et al., 2010) CP002047
4. Streptomyces cattleya NRRL 8057 (no publication) CP003219, CP003229
5. Streptomyces cattleya NRRL 8057 (Barbe et al., 2011) FQ859184, FQ859185
6. Streptomyces coelicolor A3(2) (Bentley et al., 2002) AL589148, AL645771, AL645882
7. Streptomyces collinus Tu 365 (Rückert et al., 2013) CP006259, CP006260, CP006261
8. Streptomyces davawensis JCM 4913 (Jankowitsch et al., 2012) HE971709, HE971710
9. Streptomyces flavogriseus ATCC 33331 (no publication) CP002475, CP002476, CP002477
10. Streptomyces fulvissimus DSM 40593 (Myronovskyi et al., 2013) CP005080
11. Streptomyces griseus NBRC 13350 (Ohnishi et al., 2008) AP009493
12. Streptomyces hygroscopicus jinggangensis 5008 (Wu et al., 2012) CP003275, CP003276, CP003277
13. Streptomyces hygroscopicus jinggangensis TL01 (no publication) CP003720, CP003721, CP003722
14. Streptomyces sp. PAMC26508 (no publication) CP003990, CP003991
15. Streptomyces rapamycinicus NRRL 5491 (Baranasic et al., 2013) CP006567
16. Streptomyces scabiei 87 22 (Bignell et al., 2010) FN554889
17. Streptomyces sp. SirexAA E (no publication) CP002993
18. Streptomyces venezuelae ATCC 10712 (Pullan et al., 2011) FR845719
19. Streptomyces violaceusniger Tu 4113 (no publication) CP002994, CP002995, CP002996

Assembly of *Streptomyces* Mg1

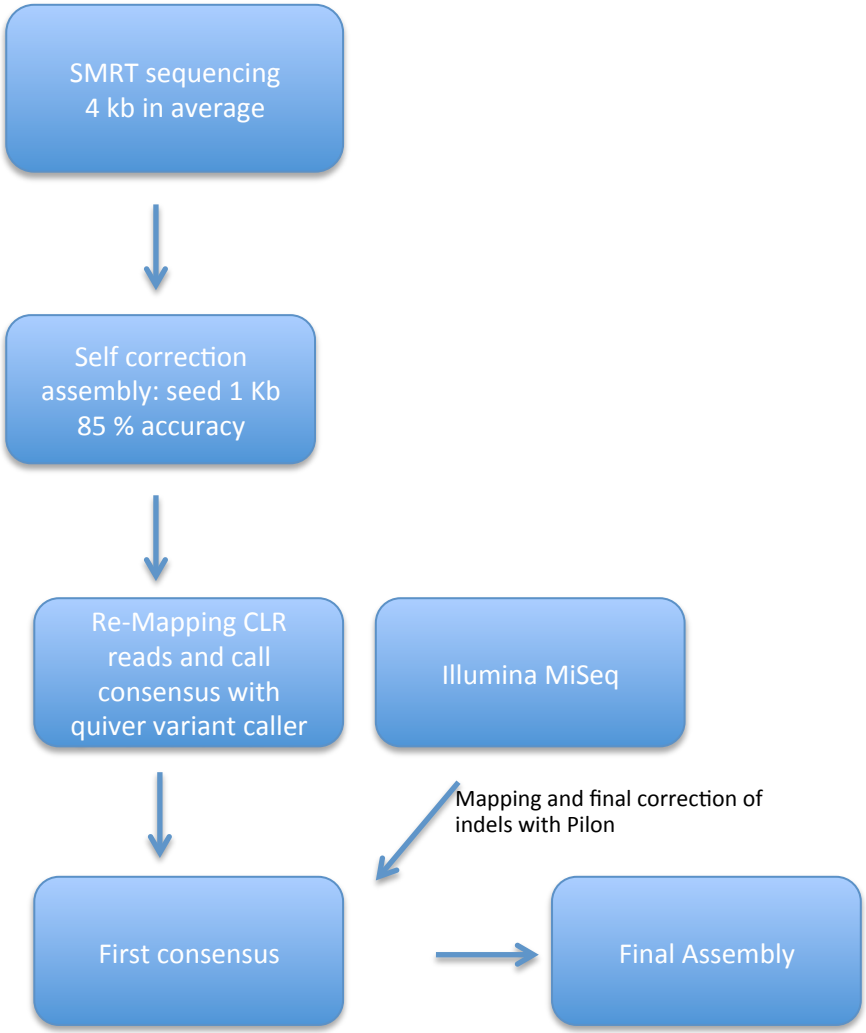
- Illumina/454 reads
- 7,2 Mb total genome in 466 contigs

- Eight SMART Cells
- 8 – 10 Kb insert library
- 8.7 Mb total genome in seven contigs

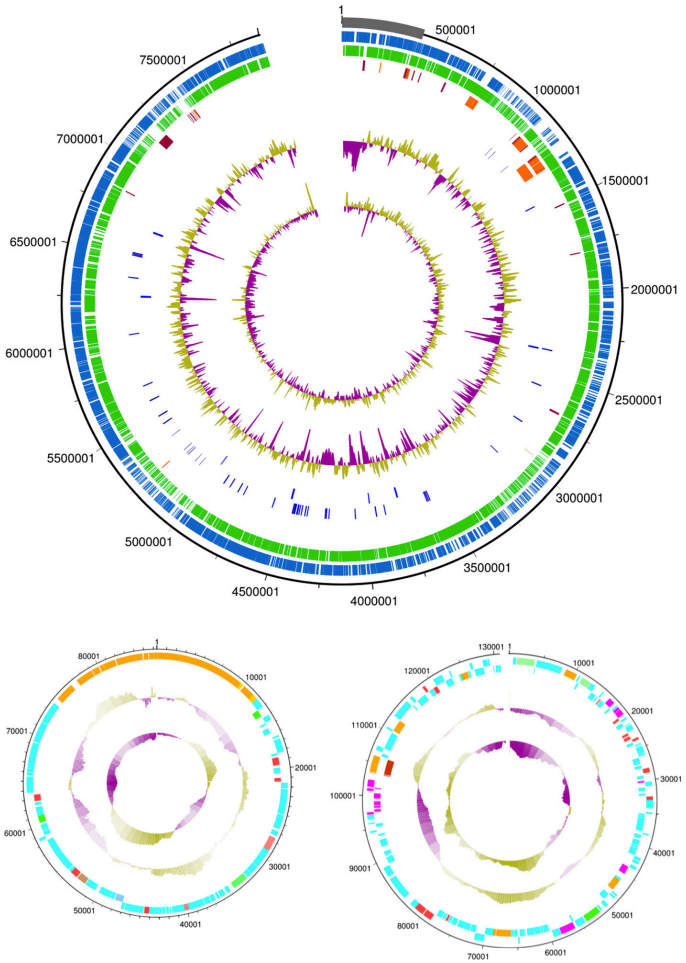
Recovering whole novel secondary-metabolism gene cluster with PaCBio



Assembly pipeline



Streptomyces leeuwenhoekii

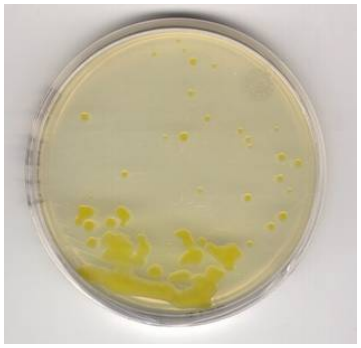


3 SMRT Cells
One linear chromosome
2 plasmids
8 Mb Total genome

35 gene clusters for biosynthesis of specialized metabolites,

Xanthomonas oryzae pv oryzae

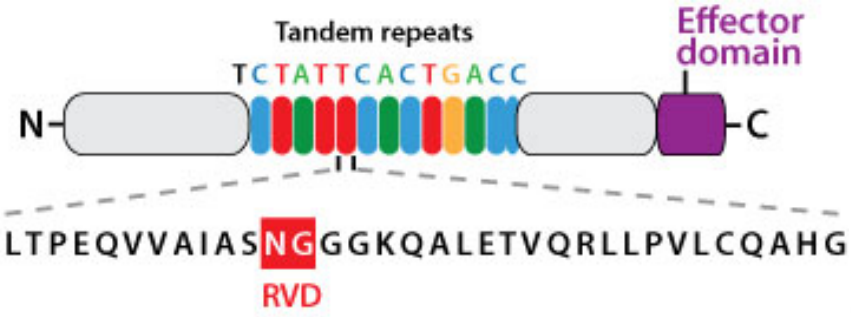
Xanthomonas oryzae pv. oryzae (Xoo)



Bacterial leaf blight on rice

TAL effectors

Di-Amino acid		Nucleotide bound
NI	=	A
NG	=	T
HD	=	C
NN	=	G/A



Functional convergence in TAL effectors and plant targets

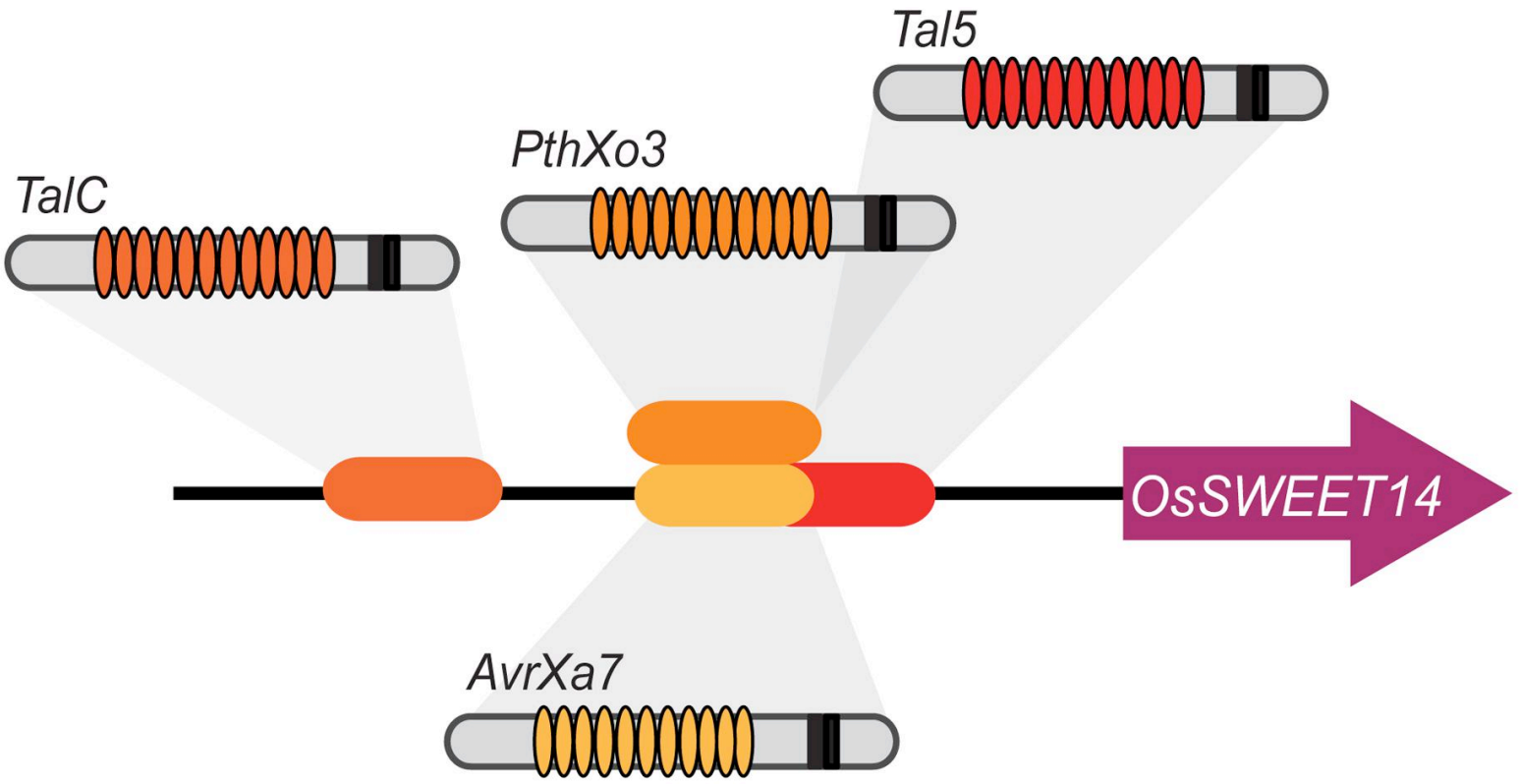
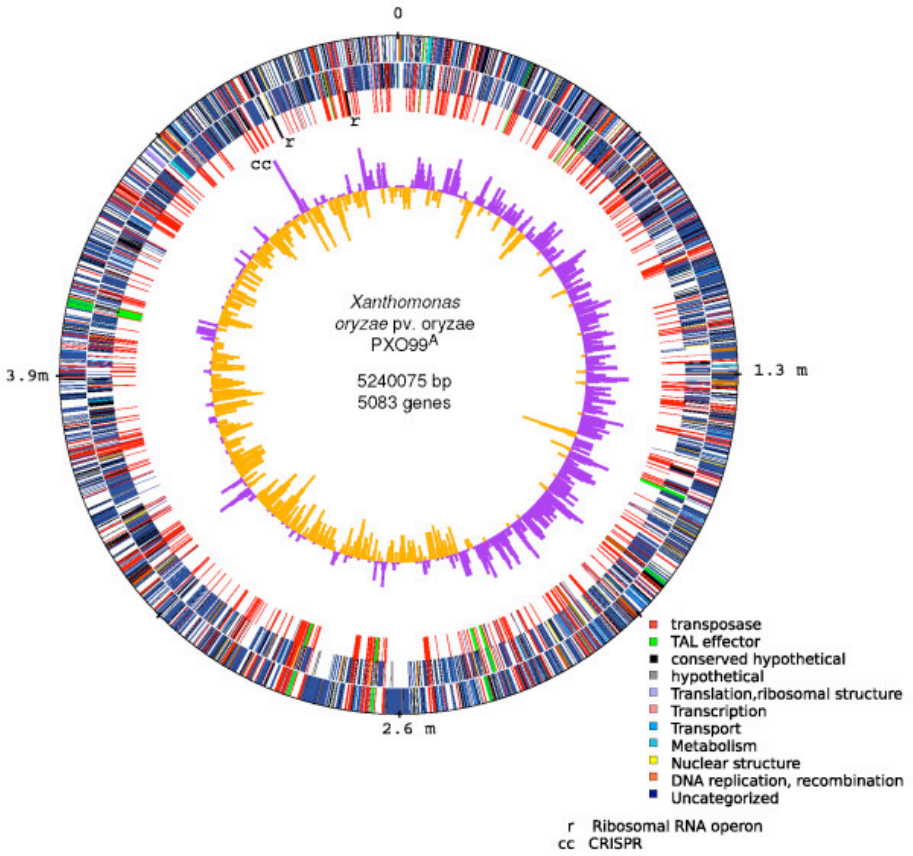


Figure adapated from Hutin et al, 2015

Strains and TalE information from：Yang and White, 2004, Chu et al., 2006 ,Antony et al., 2010; Yu et al., 2011; Streubel et al., 2013; Zhou et al., 2015)

Assembly of Xoo genomes

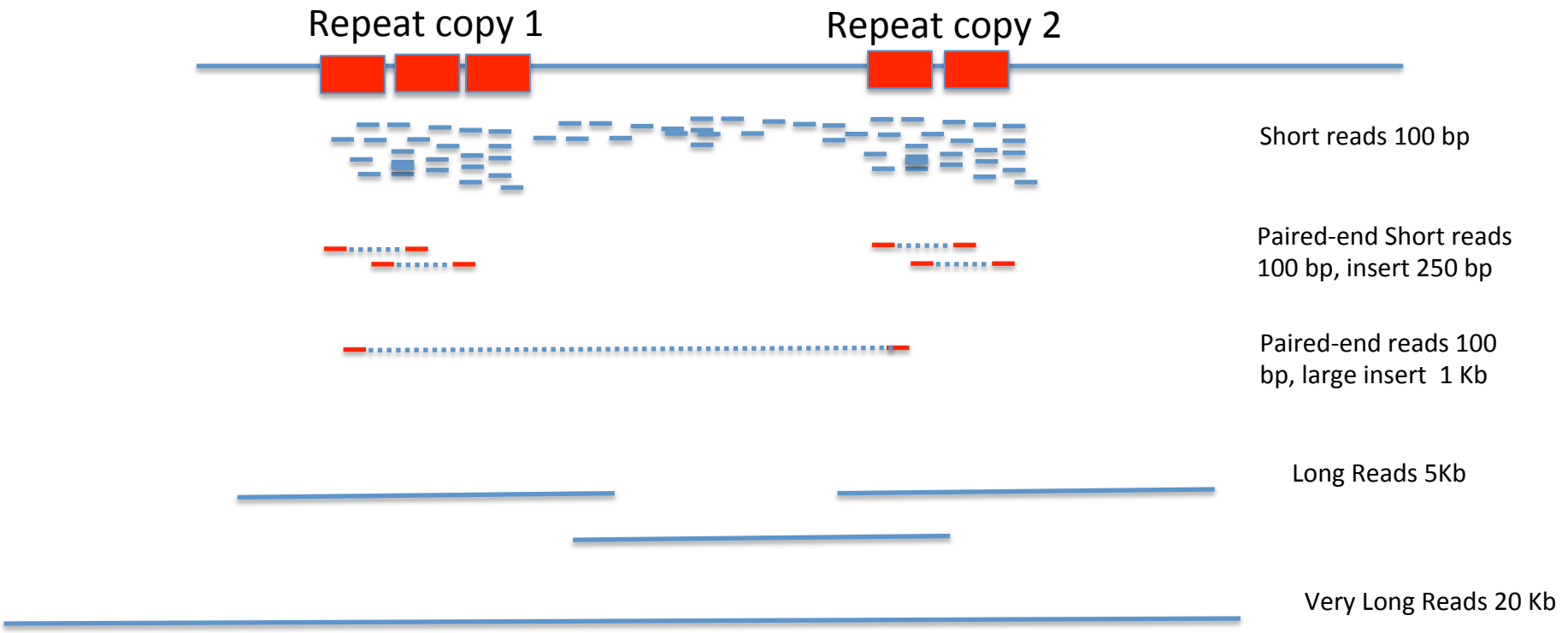
Xoo KACC10331	2004
Xoo MAFF311018	2006
Xoo PXO99A	2008
Xoc BLS256	2011



Genomes of Xoo have are complex structures :

-TAL effectors
-Insertion sequences/transposases

Long reads for assembly

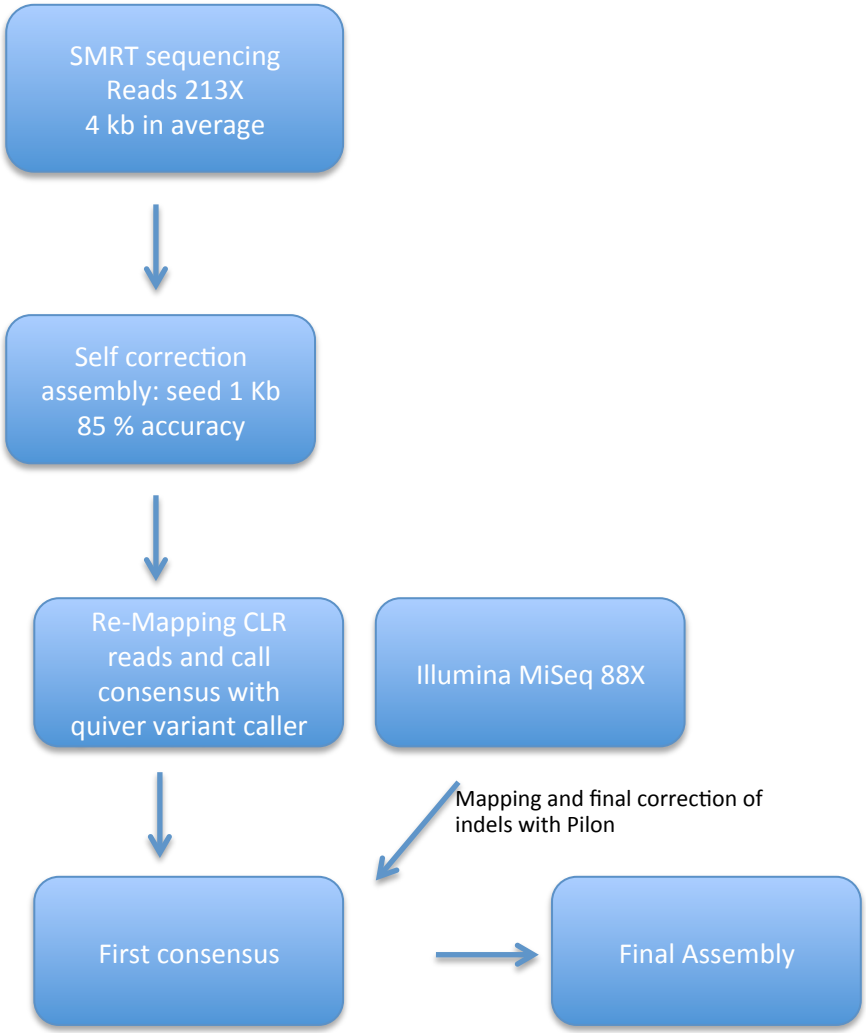


Assembly of Xoo AXO1947

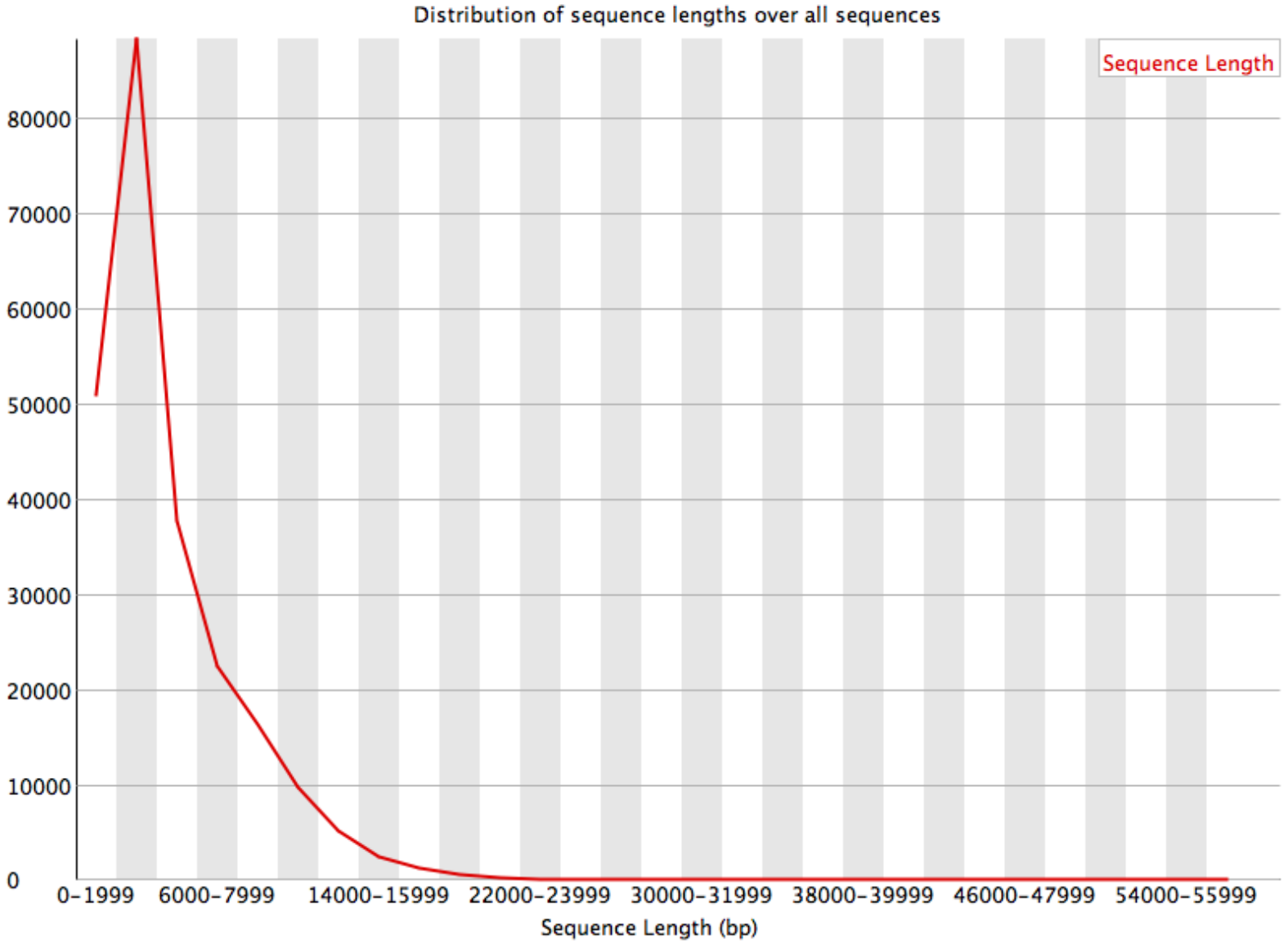
Previous attempts failed to assemble the genome

	Number of contigs	N50	Total genome
MiSeq illumina	1630	35,251	4,838,609 bp
454 GX FLX	384	35,322	4,323,533 bp

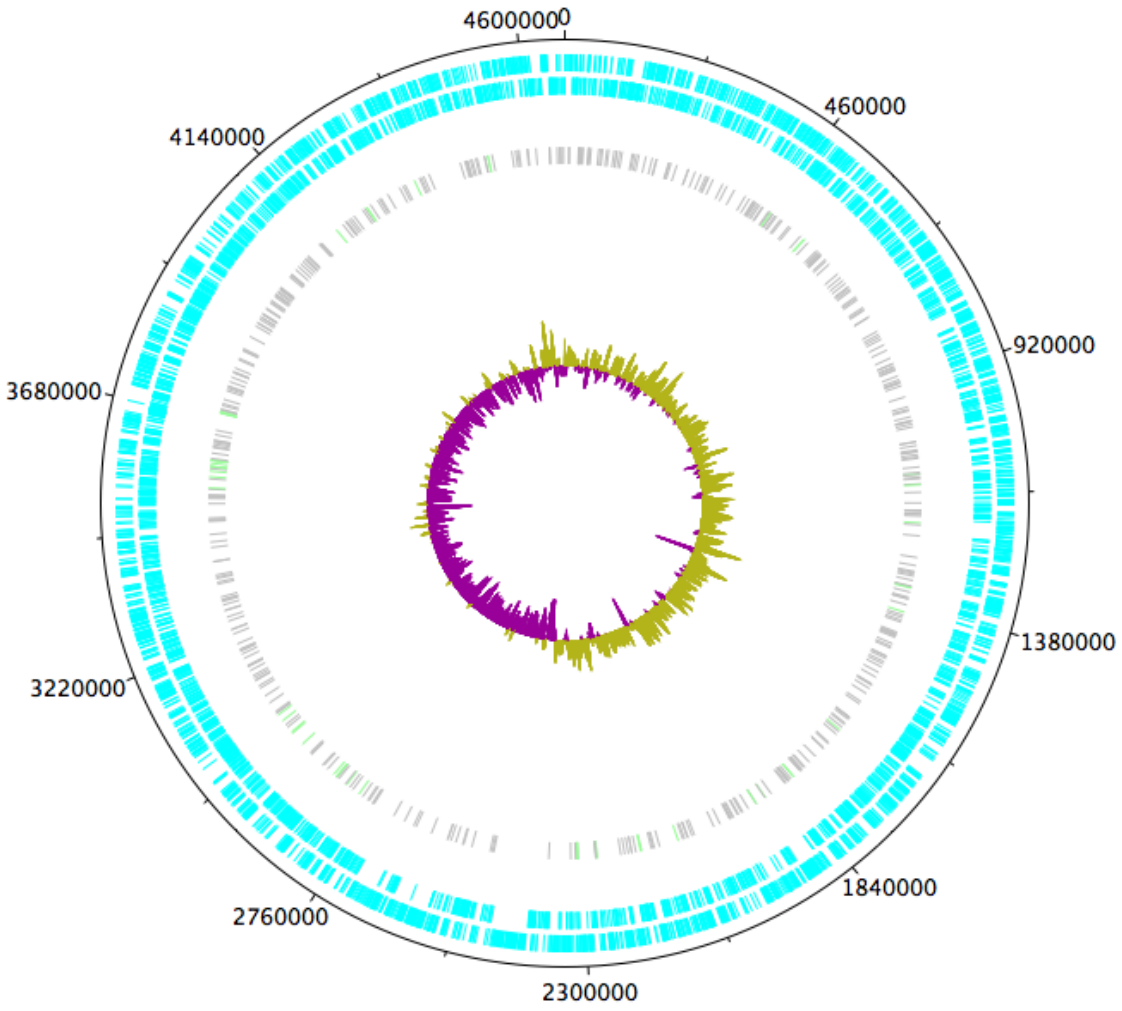
Assembly of Xoo AXO1947



Distribution of reads in sequencing of AXO1947



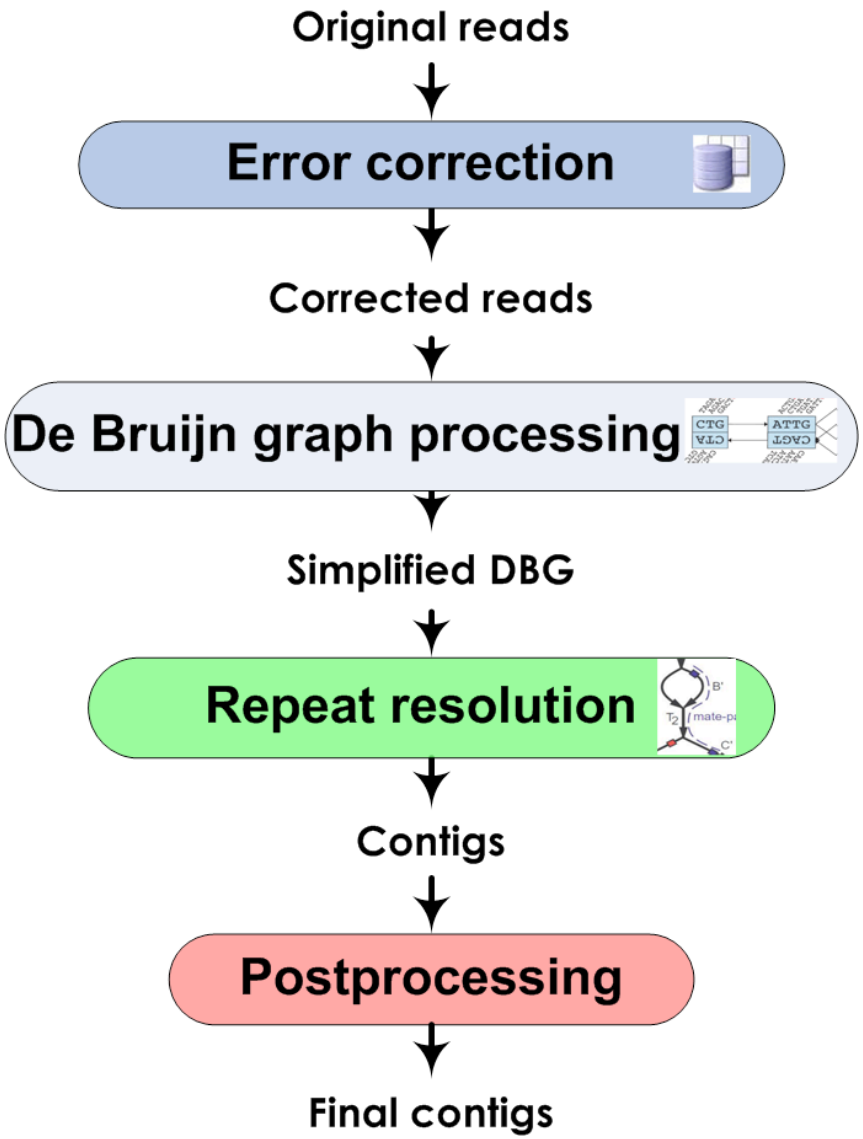
AXO1947 genome



-4,674,975 bp
-63.89% GC
-3,706 genes coding proteins
-54 tRNA
-9 TAL effectors

Huguet-Tapia et.al 2016

SPADES


Original reads
Error correction
Corrected reads
De Bruijn graph processing
CTG
CTA
ATTG
CAGT
Simplified DBG
Repeat resolution
Contigs
Postprocessing
Final contigs

Basic command

- spades.py -1 read1.fq -2 read2.fq –o assembly

Exercise 1

- Use reads of cassava_virus to conduct the assembly of the molecule.

- Check the results.
- How many contigs ?
- N50?
- Open the contigs.fa in artemis

Exercise 2

- Conduct de novo assembly of *Candidatus liberibacter*
- Two paired end files:
- LB1.fastq
- LB2.fastq
- Do not run it in your PC. I will do a demo in the workstation.