De novo assembly of bacterial genomes from the seed microbiome

Journée annuelle Plateforme Get-PlaGe

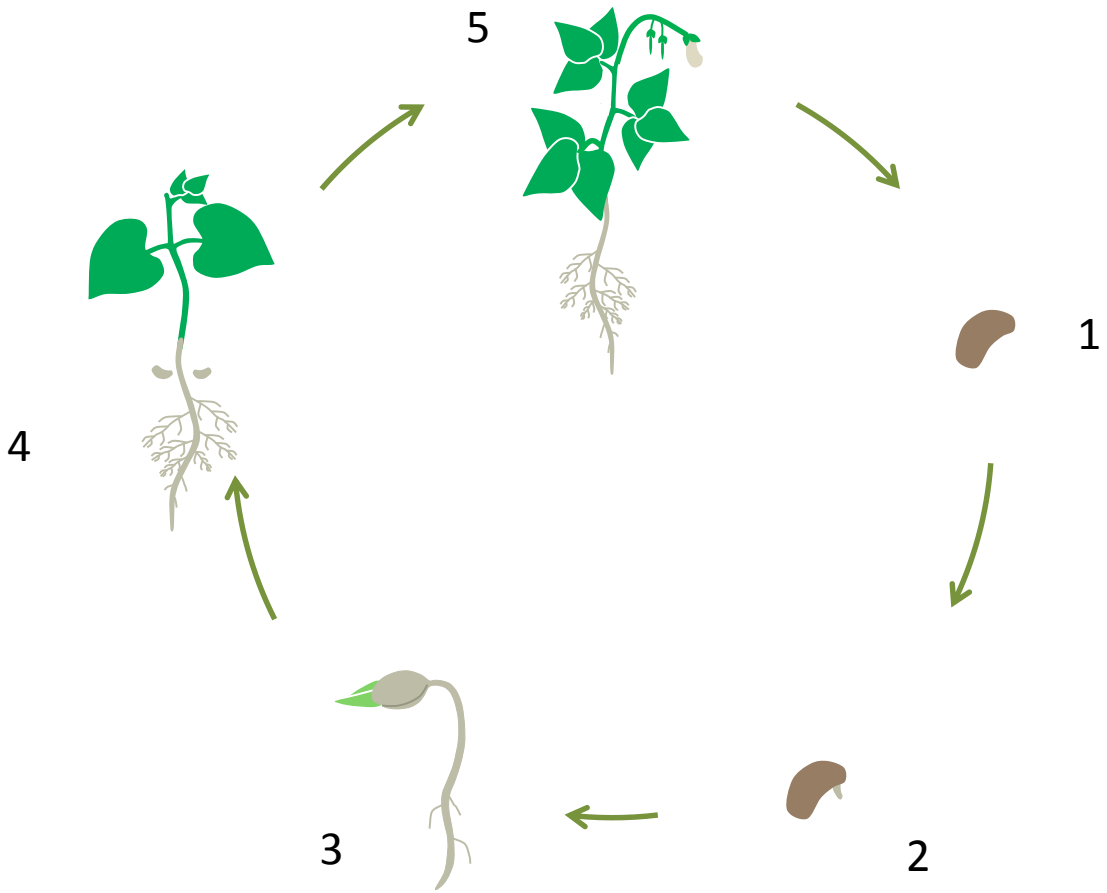
Matthieu Barret

27 05/ 2016

The plant microbiota

- Biomass accumulation (Sugiyama *et al*. 2013)
- Metabolite production (Badri *et al*. 2013)
- Flowering time (Panke-Buisse *et al*. 2015)
- Disease resistance (Mendes *et al*. 2011)

Assembly of the plant microbiota

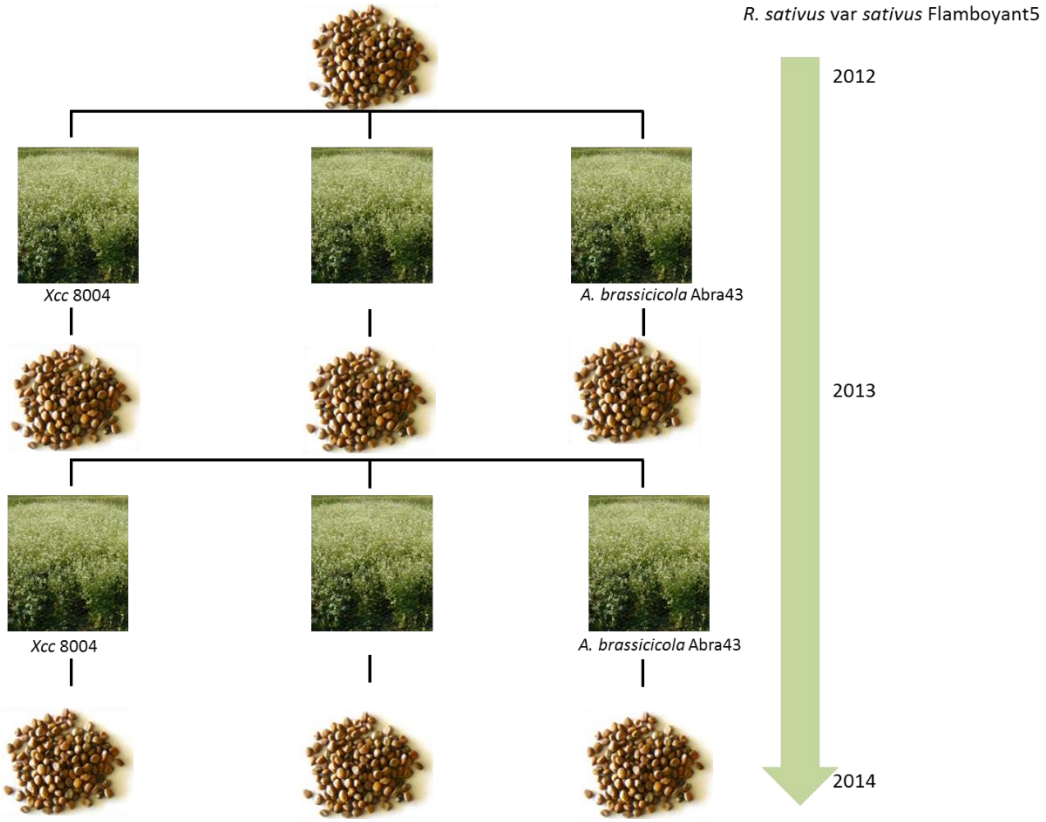


The seed microbiota

- Seed-associated microbial assemblages highly variable (10 OTUs – 300 OTUs)
- Production region shaped the structure of fungal-associated microbial assemblages
- No discernable effect of the plant genotype
- Neutral-based processes involved in assembly of seed-associated bacterial assemblages
- Lottery hypothesis : selection of functional equivalent species ?

Barret *et al*., 2015 AEM; Klaedtke *et a*l., 2015 Env. Microbiol; Barret *et al*. 2016 MPP

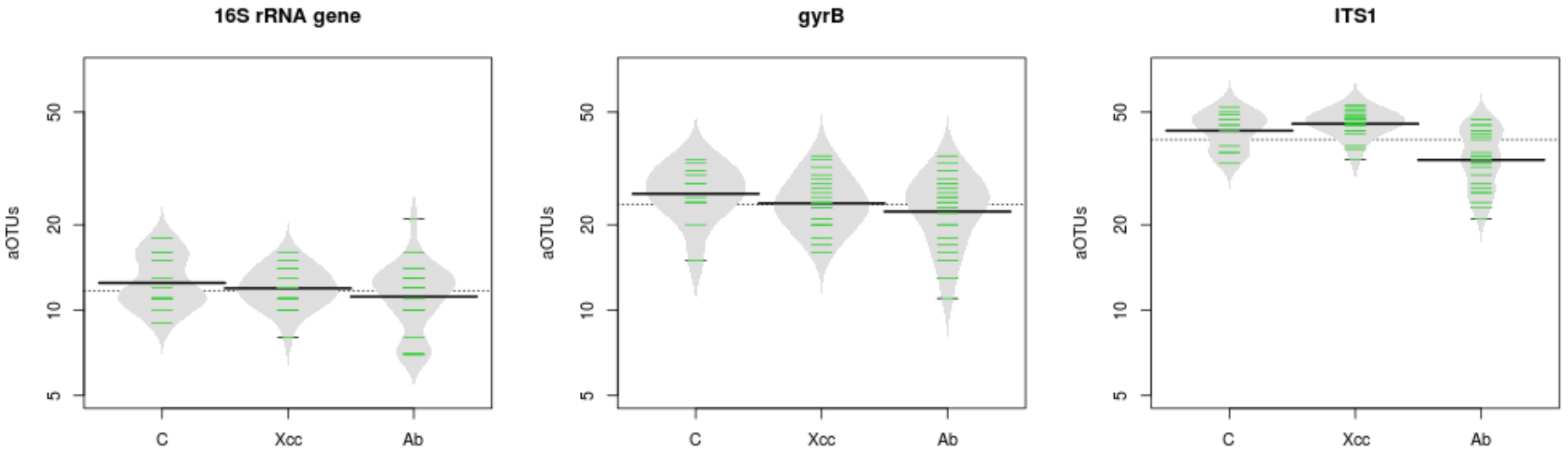
Seed microbiome : exp. design



Rezki *et al*. 2016 Peer J

Seed microbiome : richness

- Community profiling approaches (16S rRNA gene, *gyrB*, ITS1)



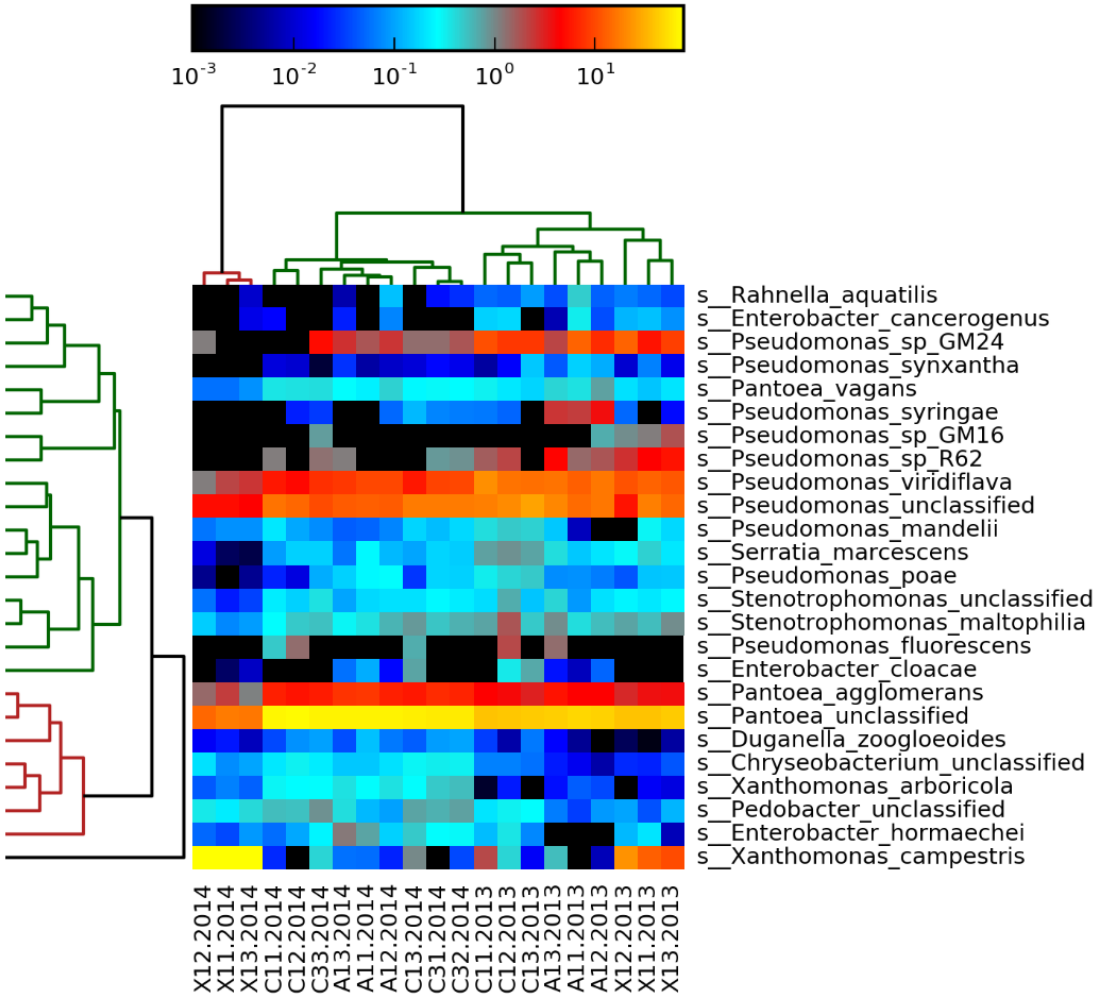
Rezki *et al*. 2016 Peer J

Seed microbiome : samples

- 21 samples
- HiSeq3000 (2 x 150; PF GeT-PlaGe)
- 20 million paired-end reads

Rezki *et al*. 2016 Peer J

Seed microbiome : taxonomic affiliation



Barret *et al*. unpublished

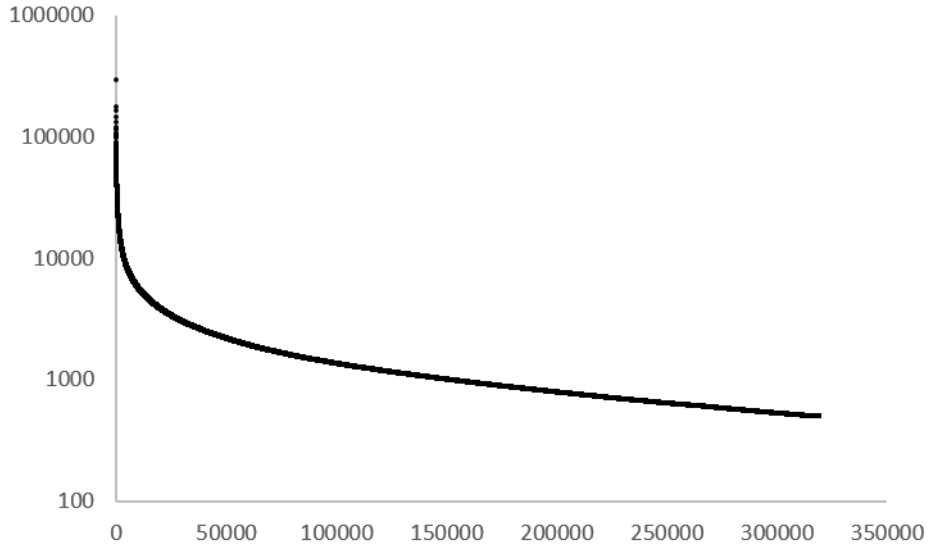
Estimation of sequencing depth needed

- Genome size x Coverage x Abundance = Seq depth

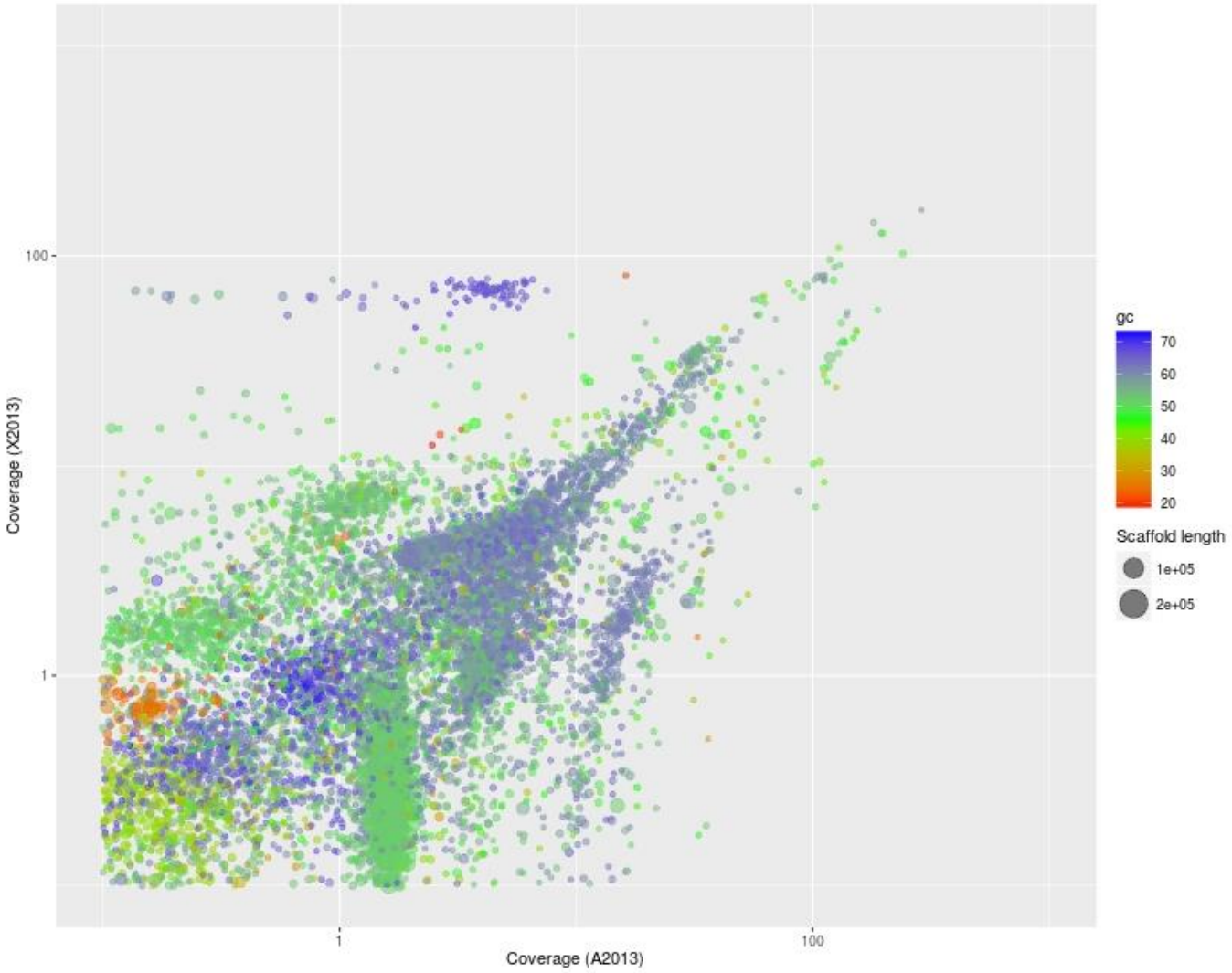
- Xcc = 5 x 50 x 1 % = 25 Gb

Seed microbiome : assembly

- IDBA-UD (Peng *et al*., 2016)
- 333,580,576 paired reads (150 b)
- 49 Gb
- contigs: 319,197 (>500 bases)
- median length : 917 bases
- max length: 295,168 bases (*Pseudomonas sp.*)



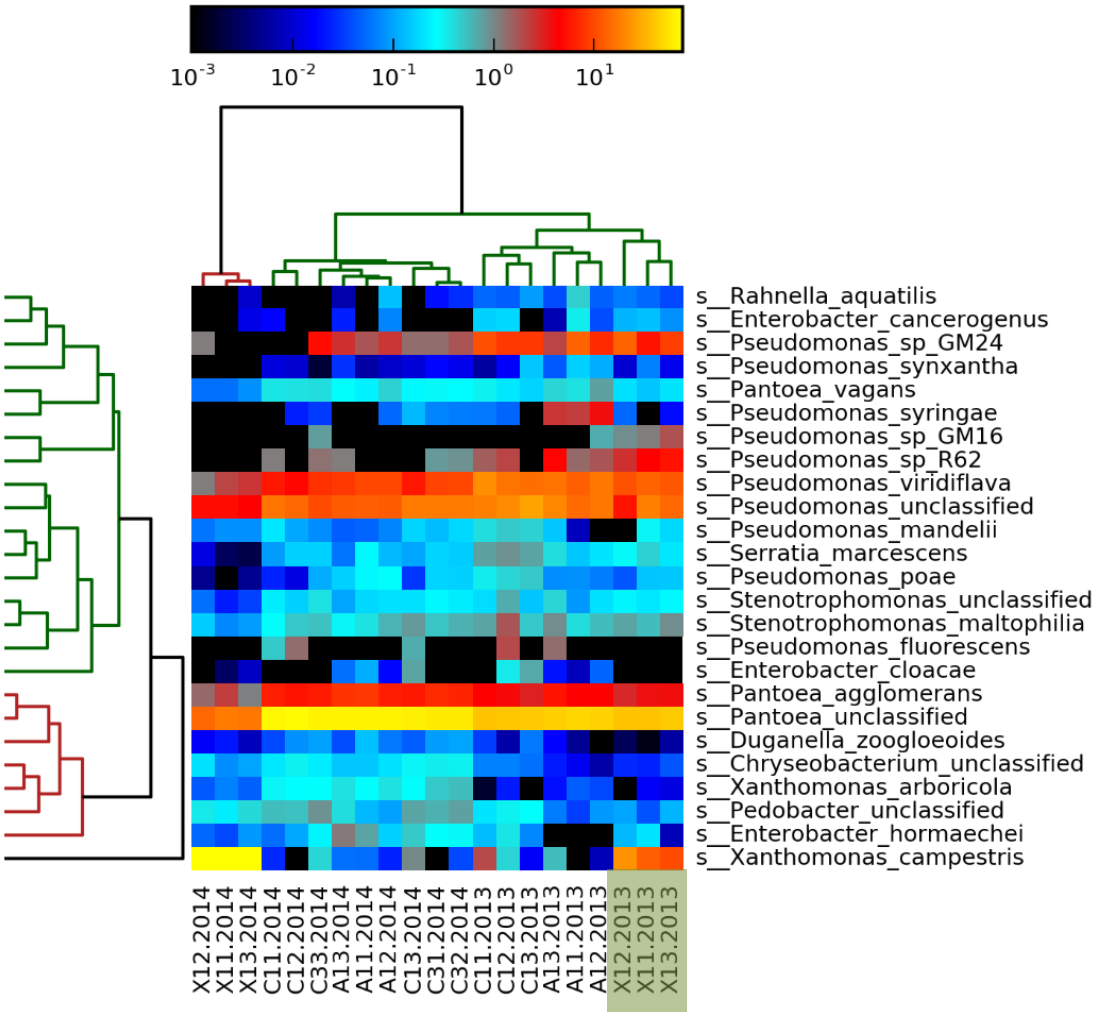
Contigs HiSeq (differential coverage binning)



Contigs HiSeq (differential coverage binning)

- 79 contigs affiliated to Xcc 8004 (> 5000 bases)
- 1,526,643 bases (ref genome = 5,148,708)

Increased assembly length with PacBio ?



Increased assembly length with PacBio ?

Sample	Nb reads	Mean read length	Gbases
X2013-PacBio	626,296	11,806	7,4
X2013-HiSeq	52,413,970	150 (paired)	15,7

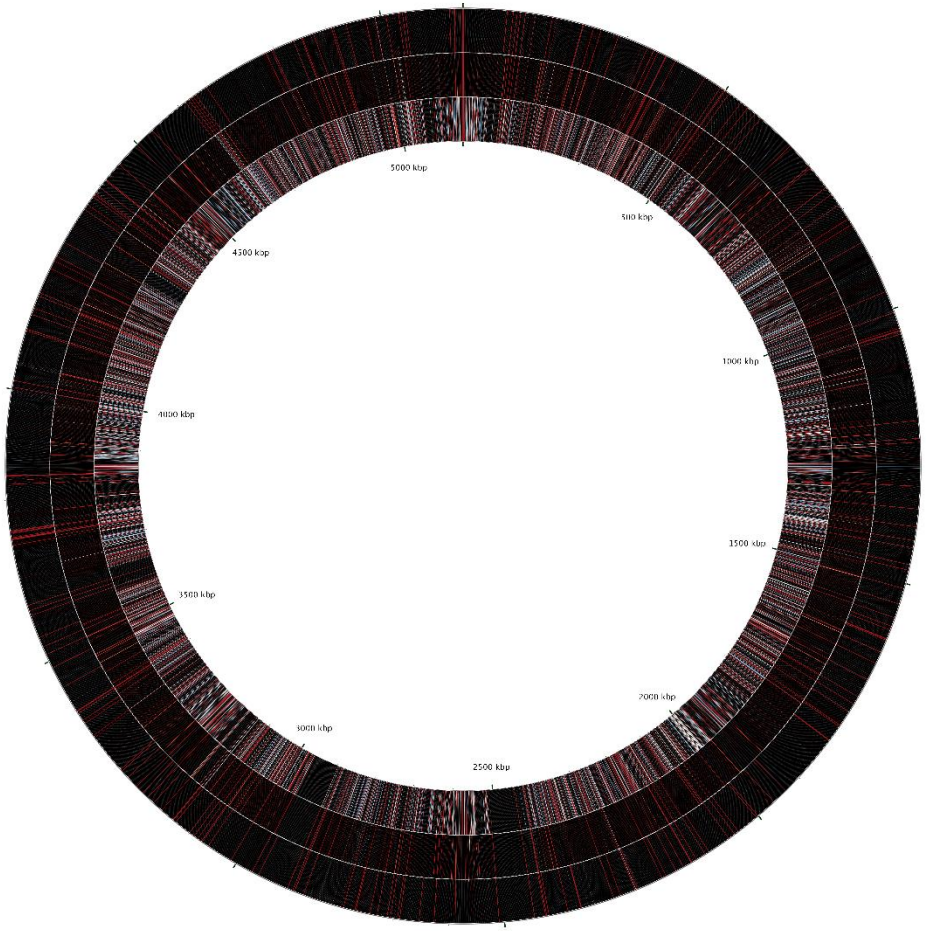
- Same biological sample
- 2 DNA extraction protocols

Assembly comparison

Sample	Method	Contigs	Median	Max
X2013-PacBio	HGAP3	4,491	7,148	3,967,155
X2013-HiSeq	IDBA-UD	4,610	907	82,788
X2013-HiSeq	BBRIC (small genome assembly	47,937	912	76,630

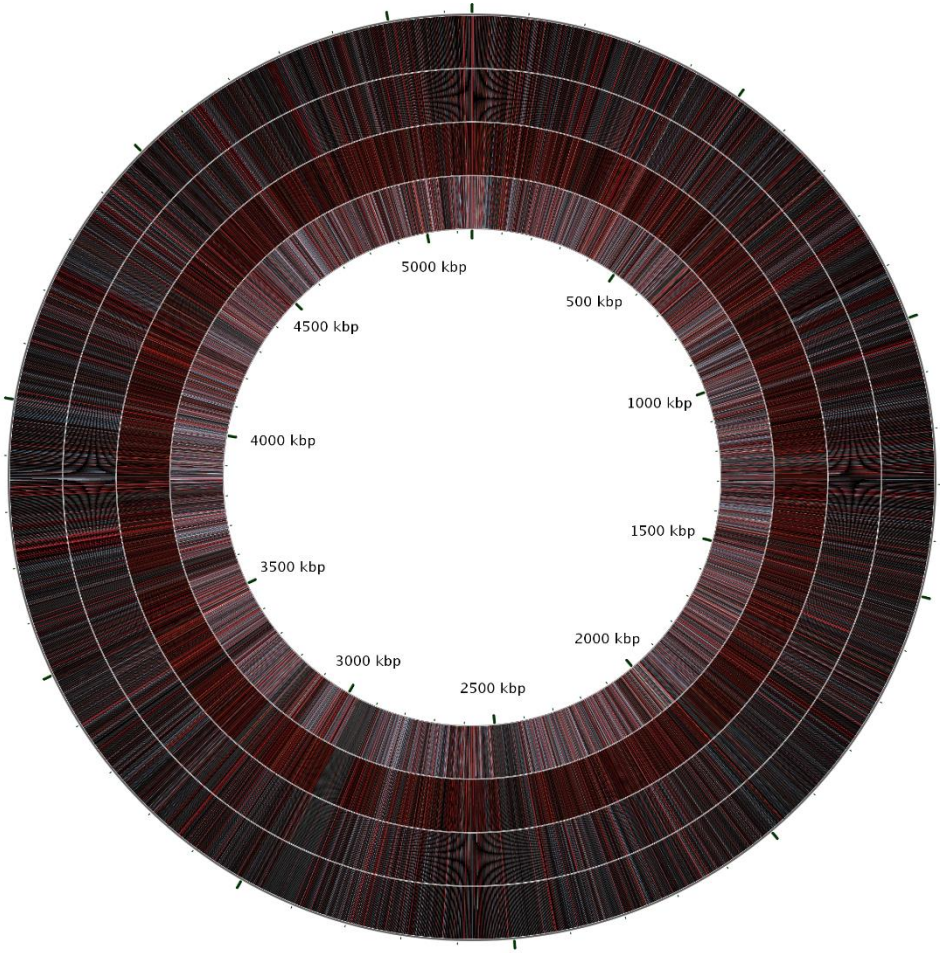
De novo assembly of Xcc 8004

- 3 contigs affiliated to Xcc 8004
  - 3,967,155
  - 652,031
  - 537,920
- Reference : 5,148,708
- delta = 8398 bases
- 4248 predicted CDSs (4271 ref genome)

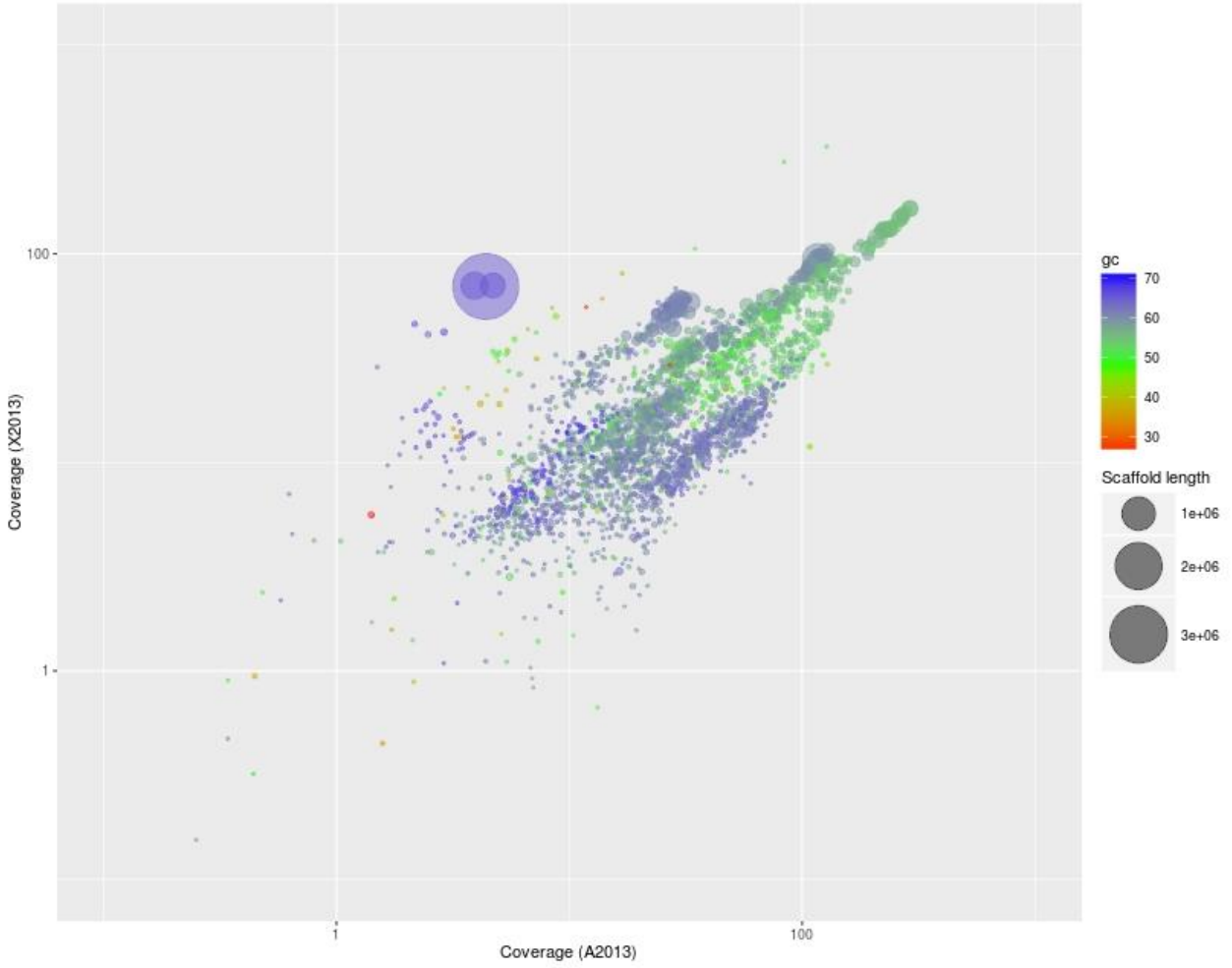


Genome assembly improvment

- Mapped HiSeq reads on PacBio assembly (Pilon, Walker et al., 2014)
- Coverage = 122
- SNPs corrected = 17
- small insertions = 102 (110 bases)
- small deletions = 9 (11 bases)

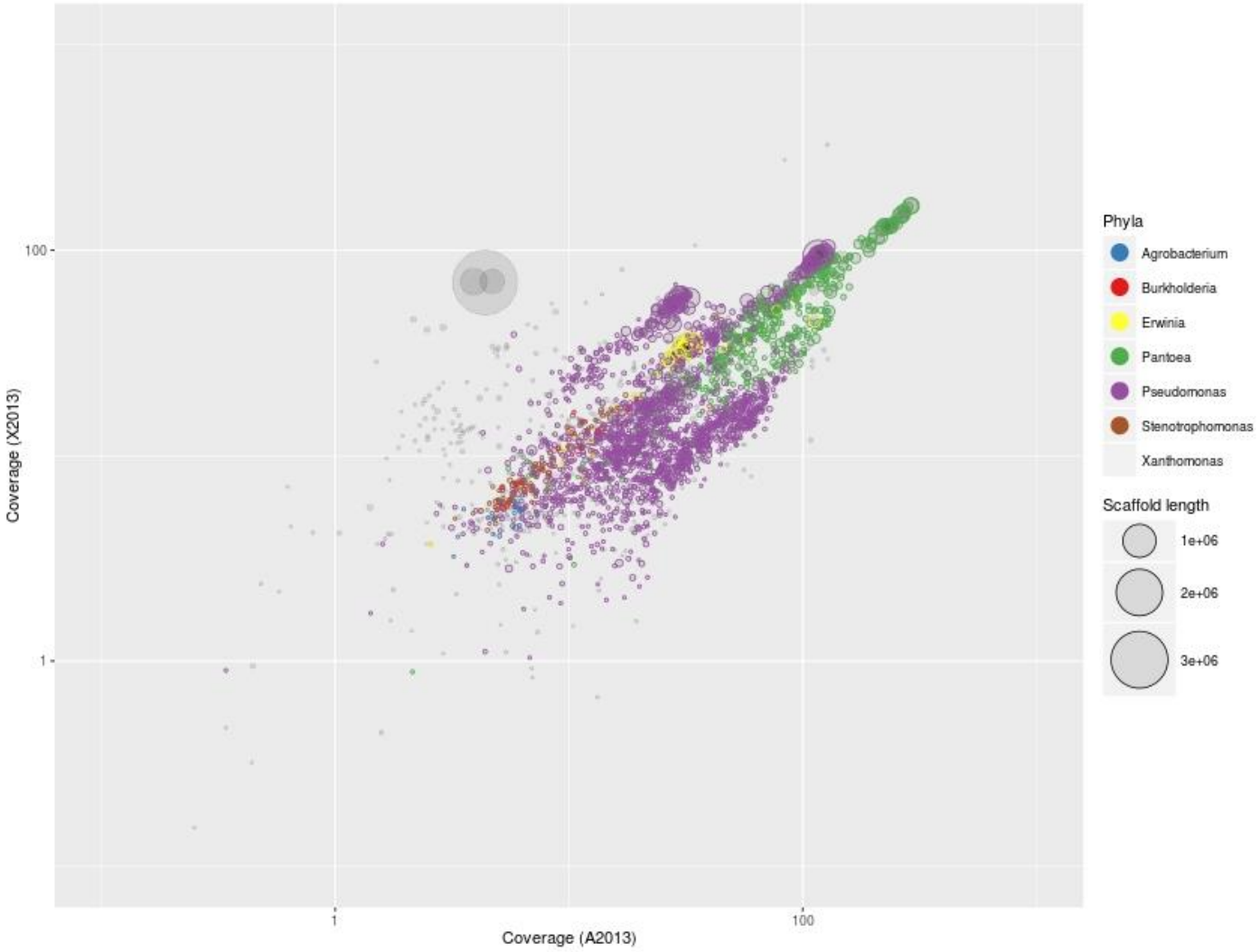


Contigs PacBio

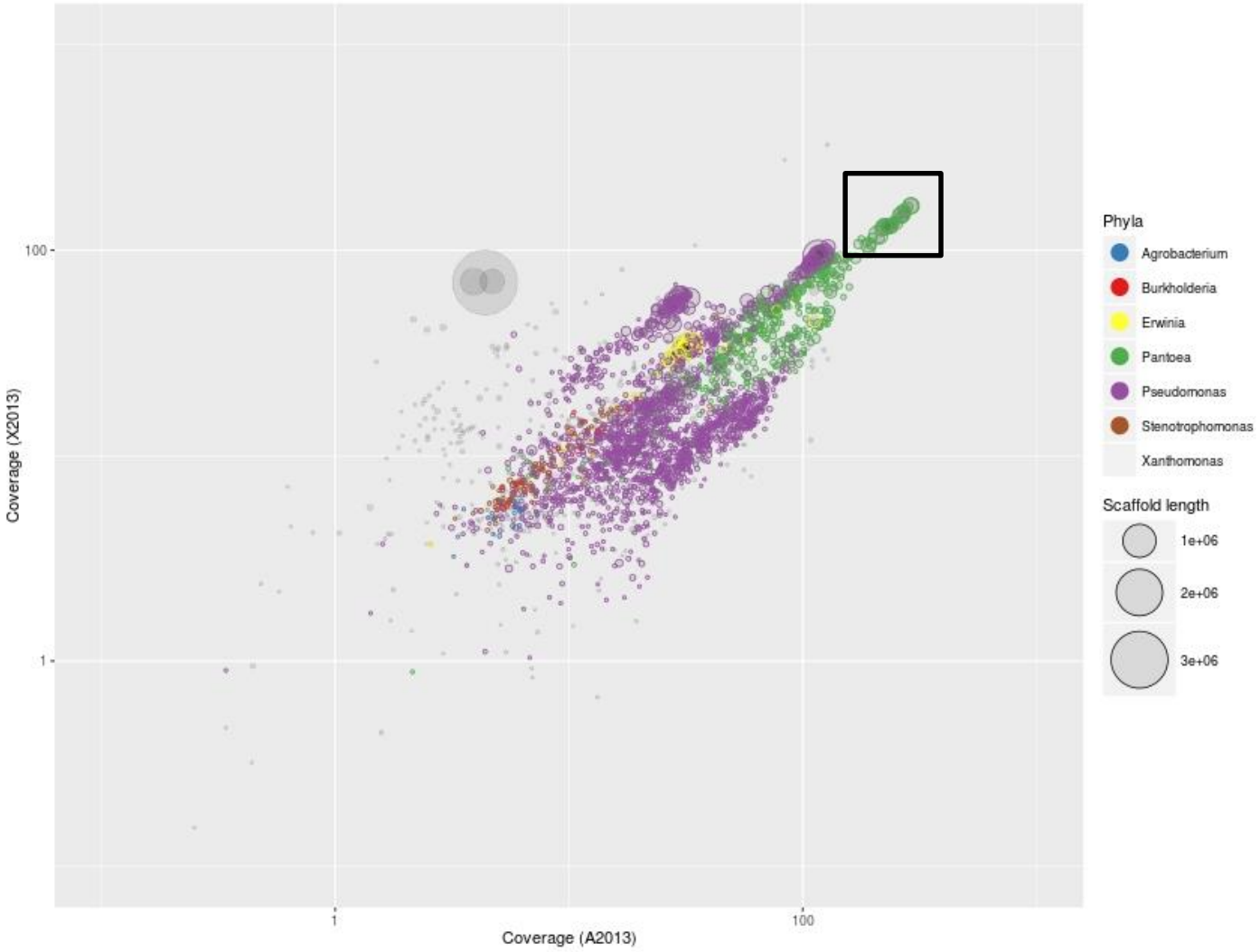


- 73% of HiSeq reads mapped on Pacbio contigs

Contigs PacBio



Extracted contigs PacBio



Extracted contigs PacBio

contigs	Median length	Mean GC	Mean A2013 cov	Mean X2013 cov
77	88723	56,8	165	110

- Partial reconstruction of 4 large contigs (*Pantoea agglomerans*, 40% genome sequence, 89% id)
  - 552,125 bases
  - 367,816 bases
  - 325,599 bases
  - 165,297 bases
- Microdiversity interferes with assembly of high abundance strains

Conclusions

- Combination of long-read + short reads promising approaches for de novo assembly of abundant bacterial strains (>1%)
- The Xcc genome assembled (3 contigs)
- Microdiversity impaired assembly of high abundant bacterial strain
- Improved by choosing samples with marked differences in RA of targeted organism
- Could not help to assemble genomes with marked strain diversity of the targeted species

Acknowledgements

MA. Jacques
M. Briand
S. Rezki

O. Bouchez
A. Roulet
C. Genthon

M. Bahut