Microbial *de novo* genome assembly:
Comparison of CLC Genomics & VELVET for assembly of contaminated but deeply covered Illumina 1.5 single reads.

CNRS UPMC
Station Biologique Roscoff

Gregory FARRANT [1,2], Erwan CORRE[2], Mark HOEBEKE[2], Wilfrid CARRE[2], Christophe CARON[2], Frédéric PARTENSKY[1] and Laurence GARCZAREK[1]

[1] UMR7144 CNRS, Station Biologique, Place Georges Teissier, 29680, Roscoff, France, {gregory.farrant, laurence.garczarek, frederic.partensky}@sb-roscoff.fr
[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France, {erwan.corre, mark.hoebeke, wilfrid.carre, christophe.caron,}@sb-roscoff.fr

EQUIPE PROCARYOTES PHOTOSYNTHETIQUES MARINS

1 - CONTEXT

Project **METASYN** (Collaboration with the Genoscope):

- ➔ Sequencing and *de novo* assembly of 25 genomes of marine picocyanobacteria
- ➔ Illumina Sequencing (version 1.5)
- ➔ Pilot project, 2 strains: MINOS11 and BOUM118
  - ❖ Two libraries:
    - Single Read 100 bp
    - Mate-Pair 2 x 50 bp with ~10 Kb insert
  - ❖ Coverage: ~ 2 500 x
  - ❖ Estimated culture contamination: 16 and 22%

2 - FIRST STEPS IN *de novo* ASSEMBLY

First assembly realized by the Genoscope using VELVET:

	MINOS11		BOUM118	
	Contigs	Scaffolds	Contigs	Scaffolds
Number	**295**	26	**710**	14
Average length	7 653	95 866	2 912	280 085
Total length	2 257 586 bp	2 492 520 bp	2 067 493 bp	3 918 003 bp
%N	-	10 %	-	46 %

➔ Contigs number consistent with literature but scaffolds oversized

Second assembly using CLC Workbench Assembly Cell on SINGLE-READS only:

	MINOS11		BOUM118	
Trimming parameter	0.005	0.001	0.005	0.001
Number	152	**81**	**35**	54
Average length	14 785	28 041	66 081	42 862
Total length	2 249 368 bp	2 271 322 bp	2 312 832 bp	2 314 542 bp

- ➢ **CLC gives better results but proprietary (i.e. black box) and commercial software**
- ➢ **Can we get similar results using a widely used assembler like VELVET ?**

3 - AVENUES TO EXPLORE

5 strategies tested to improve assembly using VELVET:

PRETREATMENTS BEFORE VELVET

1. Cleaning based on read quality (*AdaptiveTrim.pl*)
2. Reducing the coverage (by random sampling)
3. Getting rid of sequencing errors (Quake)
4. Getting rid of contamination (Geneious mapping)

TUNABLE VELVET PARAMETER

1. Choice of k-mer size

4 - EXPERIMENTS AND RESULTS

Test material: MINOS11 single reads

EXPERIMENT 1: EFFECTS OF READS CLEANING BASED ON QUALITY AND K-MER SIZE

- Cleaning with *adaptativeTrim.pl* using default parameters : quality drop threshold = 10; minimal size of trimmed reads = 20 bp.
- k-mer sizes tested : 41, 51 and 61 bases.

K-mer	41	51	61
Contigs	5 319	677	468
MaxLength	31 680	42 888	62 194
AverageLength	637	3 429	4 850
TotalLength	3 387 169	2 321 218	2 269 658

➔ Low assembly coverage, ~ 87% (when contigs are mapped to reference)
➔ Larger K-mers seem to lead to larger contigs, yet too many small contigs

EXPERIMENT 2: EFFECT OF COVERAGE REDUCTION AND MORE STRINGENT READS CLEANING

- Coverage reduced to ~200x by random sampling
- More stringent cleaning with *adaptativeTrim.pl* (quality threshold = 28; read length cutoff at 75 bp)

➔ No significant improvement

EXPERIMENT 3: EFFECT OF SEQUENCING ERRORS CLEANING

- Cleaning with *QUAKE[2] (k-mer coverage cutoff: 80x; see box 5)*

➔ Much better assembly coverage (99.5%), yet too numerous contigs

EXPERIMENT 4: EFFECT OF READ DECONTAMINATION

- Read decontamination by mapping on genomes of known contaminants using Geneious[3]
- Decontamination of unmapped reads using Quake

	Raw	Quake > Decon	Decon > Quake	DeconOnly
Contigs mapping on MINOS11 (from CLC)	1 102	1091	1244	676
Taux de couverture	98,7%	98,7%	98,4%	96,5%

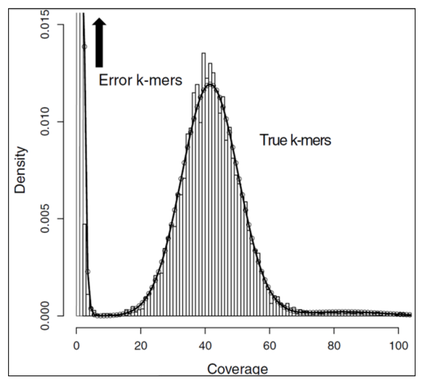
➔ Contigs number similar to the Genoscope assembly

5 – QUAKE PROGRAM

Quake[2] : reduces read variability by eliminating erroneous k-mers (error and low-coverage contaminants k-mers)

Principle of the JELLYFISH algorithm:

- Scattering of reads in 14-mers
- Histogram of the distribution of all 14-mers (see graph)
- True and low coverage contaminants depict a Gauss curve around the sequencing coverage
- Error k-mers are rare and numerous



➔Selection of a coverage threshold allows to eliminate errors k-mers
➔ Contaminant peaks may be removed by discarding k-mers corresponding to their coverage range.

6 - CONCLUSION

A fairly complex combination of strategies was needed to obtain a reasonable number of contigs using the VELVET assembler. Yet, quality of the assemblies were still much lower than the one obtained using the proprietary assembler CLC with very little optimization. Moreover, contigs generated using CLC were found to be non-chimeric when compared to a reference genome.

We speculate that the reason for this difference is that VELVET is too cautious when confronted to an assembly branching choice when CLC tends to take a ruthless decision.

FURTHER READING:

1. Zerbino DR, Birney E. *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*. Genome Res. 2008 May;18(5):821-9. Epub 2008 Mar 18. PMID: 18349386
2. Kelley DR, Schatz MC, Salzberg SL. *Quake: quality-aware detection and correction of sequencing errors*. *Genome Biol*. 2010;11(11):R116. Epub 2010 Nov 29.
3. Drummond,A.J., Ashton,B., Burton,S., Cheung,M., Cooper,A., Heled,J., Moir,R., Stones-Havas,S., Sturrock,S., Thierer,T. et al. (2010) *Geneious v5.1,* http://www.geneious.com/

Microbien - PELICAN - Génomique ANR
AGENCE NATIONALE DE LA RECHERCHE ANR
UPMC SORBONNE UNIVERSITÉS
cnrs dépasser les frontières