

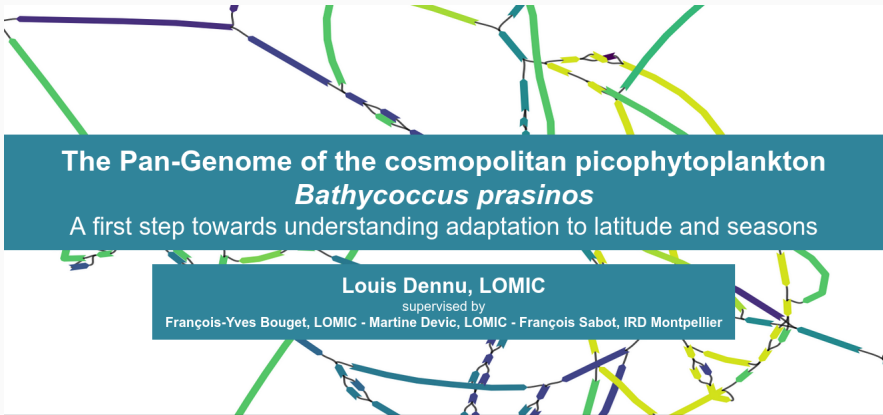
Mapping and SNP

Christine Tranchant-Dubreuil & Francois Sabot

4th of October, 2023

IRD

- Learn to manipulate NGS data
- Having a critical look on *Mapping*
- Learn to launch a *SNP Calling* and having a critical look



The Pan-Genome of the cosmopolitan picophytoplankton *Bathycoccus prasinus*

A first step towards understanding adaptation to latitude and seasons

Louis Denu, LOMIC

supervised by

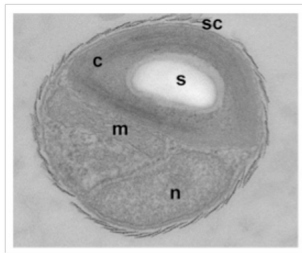
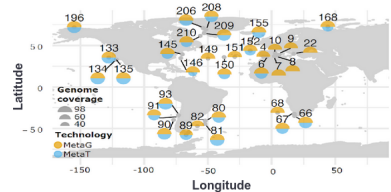
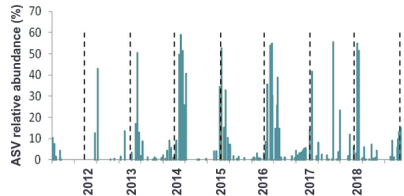
François-Yves Bouget, LOMIC - Martine Devic, LOMIC - François Sabot, IRD Montpellier

Observatoire Océanologique de Banyuls-sur-mer
Laboratoire d'Océanographie Microbienne



Bathycoccus prasinos A cosmopolitan model to study adaptation to latitude and seasons

Bathycoccus prasinos

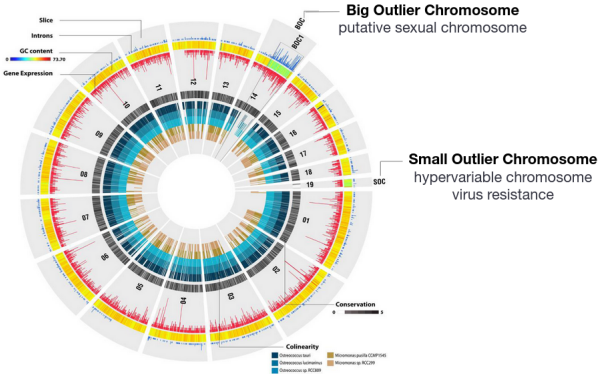


***Bathycoccus prasinus* single reference genome**

RCC1105
Banyuls-sur-Mer, France

Size
15 Mbp

19
Chromosomes



1. 18 samples

1. 18 samples
2. One reference genome to rule them all

1. 18 samples
2. One reference genome to rule them all
3. Illumina Sequencing for all 18, 150-PE

1. 18 samples
2. One reference genome to rule them all
3. Illumina Sequencing for all 18, 150-PE
4. ONT sequencing for 2 of 18

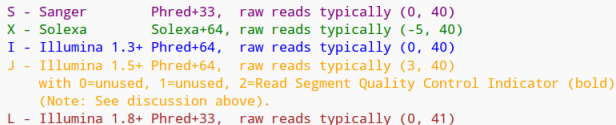
1. 18 samples
2. One reference genome to rule them all
3. Illumina Sequencing for all 18, 150-PE
4. ONT sequencing for 2 of 18
5. ~~Torturing students with these data~~

```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTGTAAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGAAATTTCTAGAATAAGGGTAACTT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
TCTCAGAAACTTGTTTGTGATGTGTGATTCAACTA
+HWI-EAS236_3_FC_20BTNAAXX:2:1:363:717
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
TTGATTAACTCTGACAAAATAAACAAAGTCTTAGG
+HWI-EAS236_3_FC_20BTNAAXX:2:1:208:209
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhGh
```

Sequencing info

Nucleotide sequence

Quality score in ASCII



“Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bwa-mem2, ...

“Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bwa-mem2, ...
2. *Cleaning mapping*: samtools, picard-tools,...

“Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bwa-mem2, ...
2. *Cleaning mapping*: samtools, picard-tools,...
3. *Realigning and Duplicates*: GATK, picard-tools,...

OPTIONAL!

“Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bwa-mem2, ...
 2. *Cleaning mapping*: samtools, picard-tools,...
 3. *Realigning and Duplicates*: GATK, picard-tools,...
- OPTIONAL!**
4. *SNP calling and Cleaning*: BCFtools, vcftools, GATK,...

“Classic” launch

1. *Mapping*: bwa aln/sampe, bwa mem, bwa-mem2, ...
 2. *Cleaning mapping*: samtools, picard-tools,...
 3. *Realigning and Duplicates*: GATK, picard-tools,...
- OPTIONAL!**
4. *SNP calling and Cleaning*: BCFtools, vcftools, GATK,...

Between 8 and 15 different commands...

We will

1. Index the reference genome

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*
3. Look at the SAM file

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*
3. Look at the SAM file
4. Sort it in BAM

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*
3. Look at the SAM file
4. Sort it in BAM
5. Obtain basic stats on mapping

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*
3. Look at the SAM file
4. Sort it in BAM
5. Obtain basic stats on mapping
6. Filter the non-mapped reads

We will

1. Index the reference genome
2. Map the reads of 1613 sample upon the *reference*
3. Look at the SAM file
4. Sort it in BAM
5. Obtain basic stats on mapping
6. Filter the non-mapped reads
7. Call the SNP

SAM format : <http://samtools.sourceforge.net/samtools.shtml>

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: <i>unsorted</i> , <i>queryname</i> or <i>coordinate</i> .
	GO	Group order (full sorting is not imposed in a group). Valid values are: <i>none</i> , <i>query</i> or <i>reference</i> .
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
PG - Program	PL	Platform/technology used to produce the read.
	ID*	Program name
	VN	Program version
CO - comment	CL	Command line
		One-line text comments

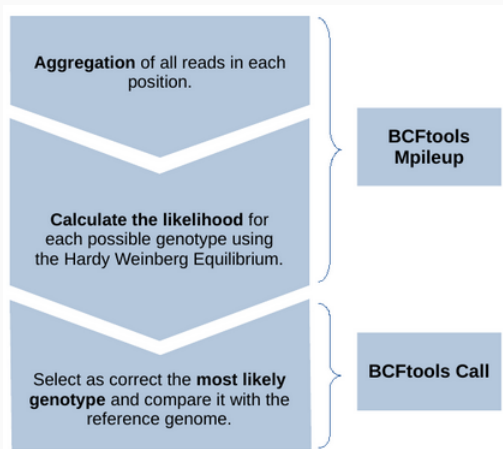
SAM format : <http://samtools.sourceforge.net/samtools.shtml>

Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)
7	NRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-based leftmost Mate POSition
9	ISIZE	inferred Insert SIZE
10	SEQ	query SEquence on the reference
11	QUAL	query QUALity (ASCII-33)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

SAM format: FLAG field

numeric	binary	description
1	00000001	template has multiple fragments in sequencing
2	00000010	each fragment properly mapped according to aligner
4	00000100	fragment is unmapped
8	00001000	mate is unmapped
16	00010000	sequence is reverse complemented
32	00100000	sequence of mate is reversed
64	01000000	is first fragment in template
128	10000000	is second fragment in template



The Variant Call Format (VCF) used in bioinformatics for storing gene sequence variations

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines (indicated by a red arrow pointing to `##fileformat=VCFv4.0`)

Optional header lines (meta-data about the annotations in the VCF body) (indicated by a red arrow pointing to the remaining header lines)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0) (indicated by a blue arrow pointing to the first column of the body)

Alternate alleles (GT>0 is an index to the ALT column) (indicated by a blue arrow pointing to the ALT column)

Deletion (indicated by a blue arrow pointing to the `<DEL>` variant)

SNP (indicated by a blue arrow pointing to the `A,AT` variant)

Large SV (indicated by a blue arrow pointing to the `T,CT` variant)

Insertion (indicated by a blue arrow pointing to the `G` variant)

Other event (indicated by a blue arrow pointing to the `SVTYPE=DEL` variant)

Phased data (G and C above are on the same chromosome) (indicated by a blue arrow pointing to the `1/1:12:3` genotype)

A DEMAIN!!

- Using GATK Variant Filtration, a flag per filter

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
DP<10 or DP>20000

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
 $DP < 10$ or $DP > 20000$
- MQ0 filter:
 $MQ0 < 4$ or $MQ0 < 0.1 DP$

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
 $DP < 10$ or $DP > 20000$
- MQ0 filter:
 $MQ0 < 4$ or $MQ0 < 0.1 DP$
- QUAL filter:
 $QUAL < 200$

- Using GATK Variant Filtration, a flag per filter
- Depth filter:
 $DP < 10$ or $DP > 20000$
- MQ0 filter:
 $MQ0 < 4$ or $MQ0 < 0.1 DP$
- QUAL filter:
 $QUAL < 200$
- SNPcluster filter:
more than 3 SNP per 10b

Problems with manual launches

- Long
- Fastidious
- Error prone
- Tracability and reproducibility not ensured

Problems with manual launches

- Long
- Fastidious
- Error prone
- Tracability and reproducibility not ensured

Solution $\implies$ Workflow Manager (or scripts...): SnakeMake, TOGGLE, NextFlow

Thanks for your attention

