Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea
Majda Arif, Jérémy Gauthier, Kevin Sugier, Daniele Iudicone, Olivier Jaillon, Patrick Wincker, Pierre Peterlongo, Mohammed-Amin Madoui

To cite this version:

HAL Id: hal-01987200
https://inria.hal.science/hal-01987200
Submitted on 29 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea

Majda Arif¹, Jérémy Gauthier², Kevin Sugier¹, Daniele Iudicone³, Olivier Jaillon¹, Patrick Wincker¹, Pierre Peterlongo², Mohammed-Amin Madoui¹

¹ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
² Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes
³ Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy
Abstract
Comparison of the molecular diversity in all plankton populations present in geographically distant water columns may allow for a holistic view of the connectivity, isolation and adaptation of organisms in the marine environment. In this context, a large-scale detection and analysis of genomic variants directly in metagenomic data appeared as a powerful strategy for the identification of genetic structures and genes under natural selection in plankton.

Here, we used DiscoSnp++, a reference-free variant caller, to produce genetic variants from large-scale metagenomic data and assessed its accuracy on the copepod Oithona nana in terms of variant calling, allele frequency estimation and population genomic statistics by comparing it to the state-of-the-art method. DiscoSnp++ produces variants leading to similar conclusions regarding the genetic structure and identification of loci under natural selection. DiscoSnp++ was then applied to 120 metagenomic samples from four size fractions, including prokaryotes, protists and zooplankton sampled from 39 Tara Oceans sampling stations located in the Atlantic Ocean and the Mediterranean Sea to produce a new set of marine genomic markers containing more than 19 million of variants.

This new genomic resource can be used by the community to relocate these markers on their plankton genomes or transcriptomes of interest. This resource will be updated with new marine expeditions and the increase of metagenomic data (availability: http://bioinformatique.rennes.inria.fr/taravariants/).
**Abbreviations:** BWA/samtools/bcftools (BSB), Mediterranean Sea (MS), Atlantic Ocean (AO), B-allele frequency (BAF), Marine Genomic Variants (MGVs), Variant Calling Format (VCF)
Introduction
The identification of population connectivity, isolation and adaptation is of great interest to understand the current and future ecological responses of plankton communities to environmental variations such as the rise of water temperature and acidity (Freer et al. 2018; Pelejero et al. 2010), especially in a climate change context (Beaugrand et al. 2003; Beaugrand et al. 2002). To understand the impact of these changes on living organisms, the study of plankton populations at the molecular level is a valuable option since it allows us not only to characterize genetic structures but also to determine which genes and biological functions are under natural selection (Avise 2004; Peijnenburg & Goetze 2013). Previous studies performed on plankton were based mostly on a few molecular markers, such as ribosomal DNA or mitochondrial genes (Blanco-Bercial et al. 2014; Cepeda et al. 2012). An alternative capture-based approach based on RAD-seq has also been proposed (Blanco-Bercial & Bucklin 2016). These approaches permitted the construction of population genetic structures using only a subset of the whole genomic variability. Furthermore, as the loci under selection represent only a very small fraction of a genome, the lack of resolution of these methods does not allow a comprehensive view of the natural selection occurring on plankton. To be able to capture the entire genomic variability of these organisms, whole genome sequencing of individuals could be the ideal strategy. However, due to the small size of certain major zooplankters and their large genome size (Wyngaard & Rasch 2000; Wyngaard et al. 2005), the current DNA extraction methods applied on a single individual do not permit us to retrieve a sufficient amount of genomic DNA that captures the whole genome complexity and that is needed to build genomic DNA libraries (without random genomic amplification) usable for high-throughput sequencing.
Recently, the use of metagenomic data has been proposed to identify natural selection in prokaryotes (Costea et al. 2017; Delmont et al. 2017; Schloissnig et al. 2013). A similar approach has also been applied to the widespread marine copepod *Oithona* (Madoui et al. 2017) to establish a population genomic analysis at the whole-genome level. The methods used in these studies were all based on metagenomic reads mapping to reference genomes, followed by several filtering steps based on the nucleic identity cut-off and depth of sequencing coverage prior to the variant calling step. This allowed the detection of polymorphic loci and the estimation of allele frequencies in each sample that were followed by a wide range of analyses to characterize the nucleic variations and to identify selection using population genetic metrics such as $F_{ST}$ (Wright 1951), LK (Lewontin & Krakauer 1973), and FLK (Bonhomme et al. 2010). In these previous studies, the arbitrary nucleic identity cut-off was used to decrease the amount of false positive variants that can be generated by the alignment of metagenomic reads provided by a closely related species that can be present in the sample. Although the use of such a filter is justified, reads harbouring more variation (< 97% identity) but belonging to the studied organism are *de facto* discarded. Moreover, the time and computational resources needed for metagenomic read alignments increase with the number of reference genomes included in the analysis. Finally, methods based on read alignments suffer from bias due to the incompleteness and imperfectness of reference genome sequences unless reference genomes are exhaustively and correctly assembled, which is rarely the case.

To bypass these problems, the use of an alignment-free variant calling method could be a solution. Therefore, in the present study, we used *DiscoSnp++* (Peterlongo et al. 2017; Uricaru et al. 2015), a reference-free variant caller, and compared its performance to the one obtained with bwa/samtools/bcftools (BSB) (Li et al. 2009) first using simulated data and then the *Tara Oceans*
metagenomic data (Karsenti et al. 2011; Pesant et al. 2015) on the O. nana reference genome as a case study to determine *DiscoSnp++* accuracy for variant calling, allele frequency estimation and downstream population genomic analysis. Then, we applied *DiscoSnp++* to *Tara* Oceans metagenomic data from the Atlantic Ocean (AO) and the Mediterranean Sea (MS) to provide a new genomic resource that contains more than 19 million marine genomic variants (MGVs) that can be used as is or directly mapped on plankton genomes and transcriptomes of interest for population genomic analysis using the provided *DiscoSnp++* module.

**Material and Methods**

**Metagenomic data and genome reference**
To compare *DiscoSnp++* to BSB, we used metagenomics reads from the MS collected by the *Tara* Oceans expedition (Alberti et al. 2017) that correspond to the 20-180 µm fraction size from the surface (≤20 m) water layers of Mediterranean stations TARA_8, 10, 11, 12, 24 and 26 (Supplementary Notes S1) (Pesant et al. 2015). Only the metagenomic data from the two stations TARA_8, and 11 were used to compare the performances of *DiscoSnp++* versus BSB for variant calling and B-allele frequency accuracy. Data from the five stations TARA_10, 11, 12, 24 and 26 were used to compare the two approaches in order to perform population genomic analysis. The *O. nana* genome was downloaded from NCBI (accession number: GCA 900157175.1). To build the marine genomic variants sets (MGVs), we used *Tara* Oceans metagenomic reads generated from samples corresponding to four size fractions (0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm) collected from stations located in the AO and MS (Supplementary Notes S1).

**The BSB pipeline**
The *bwa mem* (Li & Durbin 2009) command was used to align the metagenomics reads on the *O. nana* genome with a 17 bp seed, and alignments were stored in one sorted BAM file per station.
To avoid spurious read alignments, Dust was applied with default parameters to discard reads with low complexity. The reads with an identity under 97% with the O. nana genome were discarded. For the variant (in this study, we will systematically use the term ‘variant’ to refer to a single nucleotide polymorphism) calling step, we used the samtools mpileup and bcftools call -m commands (Li et al. 2009) with default parameters. Loci with a maximum of two alleles were kept. Only positions with a vertical coverage between the median coverage ± two standard deviations were kept with a minimum of 4x coverage (Supplementary Notes S2).

**DiscoSnp++ method overview**

*DiscoSnp++* was originally designed for genomic data analysis, however, the core of the programme also applies to cases of metagenomic data. The tool is based on the analysis of the *de Bruijn* Graph (*dBG*). In the genome assembly context (Pevzner et al. 2004), a *dBG* is a graph in which nodes are words of length *k* (*k*-mers), and each edge connects two *k*-mers that share a *k*-1 overlap. For assembling purposes, the *dBG* is constructed from *k*-mers of a read set, and contigs are obtained by finding paths in this graph. In practice, *k*-mers are counted and those having an unexpected low abundance are removed as they are considered to contain sequencing errors. The *dBG* is constructed with the remaining *k*-mers. Basically, in a *dBG*, a bubble denotes a path in the graph which diverges into two distinct paths before they reunite. Any couple of distinct sequences that exists in the data, starting and finishing with the same *k*-nucleotides, generates a bubble in the *dBG*. In particular, small indels and SNPs generate such a topological pattern. The *DiscoSnp++* algorithm detects bubbles whose couple of paths is of equal length (generated by substitutions in the data) and bubbles whose couple of paths have a difference of length ≤D (generated by insertion or deletion of size at most D). The detection of bubbles in the *dBG* can be performed through different methods corresponding to different stringencies: parameter –b 0 or 1, with –b 0 providing high precision, and lower recall and conversely (see Uricaru et al. 2015 for
more details). In a second step, raw reads are mapped back on the sequence of these paths. This step provides a way to remove non-coherent sequences (Myers 2005) and to supply read coverage per variant and per input read set, whatever the number of input read set(s). This allows the simultaneous analysis of large metagenomic data sets. When a reference genome is available, sequence variants can be mapped to it. Thus, mapped predicted variants have a genomic position, provided in a VCF file.

*DiscoSnp++* was run using the default parameters but avoiding indels (-D 0) for the methods comparison, and additionally using –k 51 to build the MGVs. Using a large k value (here k=51) decreases the method sensitivity, increases the precision and, by simplifying the *de Bruijn* graph, allows faster computing and lower memory use on very large datasets. Depending on the situation, *DiscoSnp++* was run using –b 0 (default) or –b 1, and this parameter is specified and motivated in the text. For the BSB pipeline, biallelic loci were kept and only positions with a vertical coverage between the median coverage ± two standard deviations were kept with a minimum of 4x coverage (Supplementary Notes S2).

**Comparison of the variant calling methods on simulated data**

We simulated a first population of 20 *O. nana* genomes having 99% identity with the *O. nana* reference genome and a second population of 20 genomes of “Oithona2” based on a new reference having 95% identity to the *O. nana* genome and with a 99% identity within the population. SNPs were simulated to reproduce their natural distribution along the genome (Supplementary Notes S3). We generated 100x of Illumina reads on each population and created 20 read datasets by mixing the two populations in different proportions and each dataset contained a total of 30X of simulated Illumina reads. We applied the two approaches to these simulated datasets using the *O. nana* genome as a reference for reads mapping and variant calling
for BSB and variant relocating for DiscoSnp++. The variants found by the two methods (DiscoSnp++ was run in relaxed mode only) were compared to the simulated ones (methodology presented in Figure 1). Considering O. nana as the organism of interest, the signal-to-noise ratio was calculated, i.e., the ratio of the number of O. nana variants over the number of Oithona2 variants.

\[ \text{Signal-to-Noise Ratio} = \frac{\text{Number of O. nana variants}}{\text{Number of Oithona2 variants}} \]

**Figure 1: Workflow for BSB and DiscoSnp++ methods comparison.**

**Comparison of the variant calling methods on real data**

The two methods were compared on their performance to identify intra-species variants present in the O. nana genome. The TARA_8 sample was known to contain an abundant species closely related to O. nana with a median identity percentage of 95% and very few O. nana (<10% of total Oithona based on the 28S relative abundance) (Madoui et al. 2017). The variants predicted from this sample by any method and remapped on the O. nana genome can be considered as enriched in inter-species variants. The TARA_11 sample was known to contain a large majority of O. nana (>60% of total Oithona) with other Oithona species that are not closely related to
O. nana (here O. similis and O. atlantica). The variants predicted from this sample by any method and remapped on the O. nana genome can be considered enriched in intra-species variants. The stations TARA_8 and 11 were used to compare the two approaches (methodology presented Figure 1), in terms of variant calling, allele frequency accuracy and population genomics statistics (DiscoSnp++ was run in relaxed and stringent mode). To evaluate the possible biases on the coverage of biallelic loci that could be introduced by the variant calling methods, the read depth of the biallelic loci was fitted to a negative binominal distribution and the expected skewness of the distribution was calculated and compared to the observed one. The significance of the method’s impact on the coverage skewness was tested by Wilcoxon signed-rank tests.

**Population genomics analysis**

The B-allele frequencies (BAFs), also named alternative allele frequencies compared to a haploid reference genome, were calculated from the VCF files generated by the two methods (DiscoSnp++ in relaxed and stringent mode). Only loci with at least a BAF ≥ 0.05 in one population were selected. To identify populations having the same genomic variant pattern, a PCA was performed based on the BAF of the five populations. To measure the genetic differentiation between the populations, we used the $F_{ST}$ (Wright’s fixation index): $F_{ST} = \frac{V(p)}{E(p)(1-E(p))}$, with $p$ being a set of BAFs observed in $n$ populations at the same biallelic locus, $E$ the mean and $V$ the variance. For each locus, the $F_{ST}$ was calculated between each population (pairwise $F_{ST}$). The median pairwise $F_{ST}$ was then used to estimate the genetic differentiation between each population.

To evaluate the use of DiscoSnp++ to identify population differentiation, we calculated the median pairwise $F_{ST}$ using four sets of BAFs; set1: BAFs inferred from BSB variants found in
common with *DiscoSnp++*; set2: BAFs inferred from *DiscoSnp++* variants found in common with BSB; set3: BAFs inferred from all BSB variants; set4: BAFs inferred from all *DiscoSnp++* variants. For the four sets, we used the *DiscoSnp++* variants called using the –b 0.

**Detection of loci under selection**

To detect loci under selection, we calculated the Lewontin-Krakauer (LK) statistic, which is an improvement of the $F_{ST}$ that can be used for testing the neutrality of polymorphic genes,

$$LK = \frac{(n-1)F_{ST}}{E(F_{ST})}.$$  

To be able to detect loci under selection, the LK distribution must follow a chi-square distribution $\chi^2(n - 1)$ with $n$ being the number of different populations. The fitting between the theoretical $\chi^2$ distribution and the observed LK distribution obtained from BSB and *DiscoSnp++* (with –b 0 and –b 1 options) was observed to validate the neutral model, i.e., the majority of the biallelic loci are not under selection (Supplementary Notes S4). The FLK statistics (Bonhomme et al. 2010) were also calculated; this metric is an extension of the LK test that uses a kinship matrix of the populations based on the BAF to correct genetic distance biases due to population structure. The FLK statistics were also tested for the neutral model. The first hundred loci having the highest LK or FLK values higher than expected (with a $p$-value $\leq 0.05$) were considered to be under selection. The annotation of the variants and their possible effect on protein structure was performed with *SnpEff* (Cingolani et al. 2012).

**Results**

**Variant calling**

The BSB and *DiscoSnp++* pipelines were compared for variant detection (methodology presented in Figure 1) using simulated data representing an admixture of *O. nana* and a closely related species in different proportions. Here, we considered *O. nana* as the organism of interest,
its variants were considered as true positives and the variants of Oithona2 as false positives. BSB found more true positives than DiscoSnp++ in all admixtures especially for low O. nana content (between 5 and 50% of O. nana) (Figure 2.a, Supplementary Notes S5). BSB also identified more false positives than DiscoSnp++ especially for admixtures with O. nana lower than 90%. Based on these simulations, DiscoSnp++ was less sensitive in any admixture but more specific than BSB when dealing with an admixture of two closely related species. However, the signal-to-noise ratio was higher for BSB, especially for admixtures with more than 90% of O. nana (Figure 2.b, Supplementary Notes S5).

**Figure 2:** Comparison of variant calling between DiscoSnp++ and BSB on simulated data. a. Variants recall for increasing proportion of O. nana in the admixture. b. Methods efficiency for increasing proportion of O. nana in the admixture.

The BSB and DiscoSnp++ pipelines were also compared using Tara Oceans metagenomic data from the stations TARA_8 and 11 and the O. nana genome. Compared to DiscoSnp++ in stringent mode, BSB found approximately 14 times more intra-species variants (TARA_11) and eight times more variants enriched in inter-species variants (TARA_8) (Figure 3). Compared to DiscoSnp++ in relaxed mode (-b 1), BSB found 3.5 times more intra-species variants, and 1.3
Figure 3: Comparison of variant calling between DiscoSnp++ and BSB on Tara Oceans metagenomic data. Variants found by each method are written under the method name. Variants found in common with DiscoSnp++ and BSB in populations from TARA_8 and 11. DiscoSnp++ parameters –b 0 (stringent mode) and –b 1 (relaxed mode) were tested. The percentages correspond to the fraction of variants found by the two methods (in blue for BSB and red for DiscoSnp++).

- 5 times more variants enriched in inter-species variants. On real metagenomic data, the results provided the same trend given by the comparison on the simulated data, showing that DiscoSnp++ is less sensitive but more specific for intra-species variant detection in a population admixture, even in relaxed mode. The effect of the variant calling methods on the skewness of the depth of coverage distribution was not significant but still present (p-value = 0.06, Wilcoxon signed-rank test) (Supplementary Notes S2.b) and the skewness obtained using the –b 0 option of
"DiscoSnp++" was closer to the expected one (Supplementary Notes S2.c).

**Figure 4: B-Allele frequency correlation between DiscoSnp++ and BSB.** The x-axis of the scatter-plots corresponds to BAFs obtained with BSB and the y-axis corresponds to BAFs obtained with DiscoSnp++. a. y-axis is BAFs computed from TARA_8 with DiscoSnp++ –b 0. b. y-axis is BAFs computed from TARA_8 with DiscoSnp++ option –b 1. c. y-axis is BAFs computed from TARA_11 with DiscoSnp++ –b 0. d. y-axis is BAFs computed from TARA_11 with DiscoSnp++ –b 1.

**Allele frequency accuracy**
The BAFs obtained for variants found by the two calling methods were compared (Figure 4) and we observed a strong correlation between the two methods in *O. nana* populations from TARA_8 and 11 ($R^2 \geq 0.95$). However, we found that 7.5% of the variants had a higher BAF difference than expected between the two methods (i.e., with a BAF difference higher/lower than the median difference plus/minus two standard deviations, see Supplementary Notes S6.a). For variants having a higher BAF with DiscoSnp++ (6.3% of the total variants found in common), we explained the difference by the identity cut-off of 97% used in the BSB pipeline (Supplementary Notes S7). The variants presenting a strong BAF deviation between the two
methods were annotated based on their genomic location (i.e., intronic, exonic, UTR and intergenic) and compared to (i) the genomic location of the variants presenting no significant BAF differences, and (ii) a random distribution of the variants on the genome. Significant differences ($p$-value < 0.001, chi-square test) were found, with an increase of biallelic loci having higher BAFs with DiscoSnp++ located in the non-coding regions of the genome (Supplementary Notes S6.b and c). This result suggests that DiscoSnp++ can recruit more reads than BSB in non-coding regions of the genome. These regions are indeed expected to contain more polymorphisms than coding regions within populations. Therefore, filters that are applied in BSB tend to discard reads that should be aligned at a reduced similarity threshold. Consequently, DiscoSnp++ seems to provide a better estimation of the allele frequency in more variable regions of the genome compared to BSB applied with a 97% identity cut-off.

Figure 5: *O. nana* genetic structure in the Mediterranean Sea obtained with DiscoSnp++ and BSB. a. PCA on five *O. nana* populations on the MS based on BAFs obtained with BSB. b. PCA on five *O. nana* populations on the MS based on BAFs obtained with DiscoSnp++. c. Differences of the median pairwise $F_{ST}$ between BSB and DiscoSnp++.

**Population genomic analysis**

Five *O. nana* populations from sampling stations (TARA_10, 11, 12, 24 and 26) were clustered by PCA based on their BAFs (Figure 5.A and 5.B). For the two methods, the clustering showed a
Table 1: Median pairwise $F_{ST}$ between *O. nana* populations obtained from the four BAFs sets. Set1: BAFs inferred from BSB for all variants found in common with *DiscoSnp++*; set2: BAFs inferred from *DiscoSnp++* for variants found in common with BSB; set3: all BAFs inferred from BSB; set4: all BAFs inferred from *DiscoSnp++*.

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARA_10 vs TARA_11</td>
<td>0.074</td>
<td>0.074</td>
<td>0.065</td>
<td>0.075</td>
<td>0.0046</td>
</tr>
<tr>
<td>TARA_10 vs TARA_12</td>
<td>0.077</td>
<td>0.084</td>
<td>0.065</td>
<td>0.086</td>
<td>0.0096</td>
</tr>
<tr>
<td>TARA_10 vs TARA_24</td>
<td>0.096</td>
<td>0.096</td>
<td>0.077</td>
<td>0.099</td>
<td>0.01</td>
</tr>
<tr>
<td>TARA_10 vs TARA_26</td>
<td>0.109</td>
<td>0.125</td>
<td>0.096</td>
<td>0.133</td>
<td>0.016</td>
</tr>
<tr>
<td>TARA_11 vs TARA_12</td>
<td>0.077</td>
<td>0.099</td>
<td>0.071</td>
<td>0.1</td>
<td>0.0149</td>
</tr>
<tr>
<td>TARA_11 vs TARA_24</td>
<td>0.100</td>
<td>0.124</td>
<td>0.089</td>
<td>0.128</td>
<td>0.0189</td>
</tr>
<tr>
<td>TARA_11 vs TARA_26</td>
<td>0.114</td>
<td>0.142</td>
<td>0.099</td>
<td>0.143</td>
<td>0.0216</td>
</tr>
<tr>
<td>TARA_12 vs TARA_24</td>
<td>0.096</td>
<td>0.1</td>
<td>0.077</td>
<td>0.105</td>
<td>0.0121</td>
</tr>
<tr>
<td>TARA_12 vs TARA_26</td>
<td>0.096</td>
<td>0.1</td>
<td>0.077</td>
<td>0.111</td>
<td>0.0141</td>
</tr>
<tr>
<td>TARA_24 vs TARA_26</td>
<td>0.105</td>
<td>0.116</td>
<td>0.095</td>
<td>0.125</td>
<td>0.013</td>
</tr>
</tbody>
</table>

similar grouping of the populations by geographic location, separating the ones from the Western MS (WMS) from the ones of the Eastern MS (EMS). We estimated the genetic differentiation between the *O. nana* populations by calculating the pairwise $F_{ST}$ using four different sets of variants and related BAFs (Supplementary Notes S6) and compared the median $F_{ST}$ values to evaluate any biases that could be introduced by *DiscoSnp++* ran in stringent mode (Table 1). Using only variants detected by the two methods (i.e., using BAF from set 1 and set 2), we found no $F_{ST}$ difference over 0.024 and the average difference between the median pairwise $F_{ST}$ was 0.012 ± 0.01 (Figure 5.C). A higher difference was observed for high $F_{ST}$ values. We found a negligible difference between pairwise $F_{ST}$ computed by *DiscoSnp++* in relaxed mode versus stringent mode (mean=0.003, sd=0.008). For the selected variants, the two methods allowed the identification of the same genetic pattern between the five *O. nana* populations. The genetic distance observed using all *DiscoSnp++* or all BSB variants also produced a similar genetic
pattern with an absence of genetic structure within the WMS and a weak differentiation between the two MS basins and within the EMS (Table 1). Compared to previously published results (Madoui et al. 2017), there was a lower genetic distance between the population of TARA_26 and the four other populations. This difference can be due to the more stringent filtering on reads coverage used in the current study (see Materials and Methods) to consider valid variants compared to the previous study where biallelic loci with a read coverage up to 80x were kept. The current coverage filters may have discarded reads provided by repeated regions or a closely related species possibly present in the TARA_26 sample.

![Diagram showing loci under natural selection found in common between DiscoSnp++ and BSB.](image)

**Figure 6:** **Loci under natural selection found in common between DiscoSnp++ and BSB.** a. Loci in common using the 100 LK highest values ($p$-value<0.001, chi-square test). b. Loci in common using the 100 highest FLK values ($p$-value<0.001, chi-square test).

**Detection of loci under natural selection**
To identify loci under natural selection, the LK and FLK statistics were computed from the BAFs of sets 1 and 2. For each variant set and statistics, the hundred loci with the highest LK and FLK were compared to estimate the congruence between the two variant calling methods (Figure 6).
We found more loci in common with LK than FLK and by using the –b 0 option of DiscoSnp++ suggesting a more accurate detection of loci under selection as being more stringent in the variant calling of DiscoSnp++.

The functional annotation of the 79 variants under natural selection (detected by DiscoSnp++ -b 0 and using the LK outliers) that were found in common with BSB (Supplementary Notes S7) showed 16 non-synonymous variants and 14 synonymous variants. Compared to the previous study (Madoui et al. 2017), we found four new Lin12 Notch Repeat (LNR) domain-coding genes. These domain-coding genes are of particular interest in O. nana where they were found to be over-abundant compared to other metazoans and one of them detected under positive selection was male-specific based on expression data (Madoui et al. 2017). Among the four new LNR domain-coding genes found to be under selection, one (GSONAT00015400001) codes a metallopeptidase domain protein, another (GSONAT00015380001) codes an LNR protein associated with a Kelch domain and two others (GSONAT00013822001, GSONAT00015410001) code only LNR domain proteins without association to other known domains. These new results reinforce the highly evolutionary potential of LNR domain-containing proteins and their importance in the O. nana biology.
Figure 7: Geographic and size fraction distribution of MGVs.

Plankton genomic variant resources from the *Tara* Oceans metagenomic data

We produced the new set of MGVs by running *DiscoSnp++* in relaxed mode (to optimize the number of MGVs) on more than 40 billion metagenomic 100 bp reads from 39 *Tara* stations located in the AO and the MS (Figure 7 and Supplementary Notes S8). These MGVs correspond to genomic variants (SNVs and indels) found from natural populations of prokaryotic, protist and animal plankton that were sampled during the three-year expedition of *Tara*. For the four different size fractions, we generated more than nineteen million MGVs (Table 2). The amount of input data was relatively similar among all size fractions (~11-12·10^9 of 100 bp reads) but the
computation time globally increased with the size fraction and all had the same very low memory usage (~100 Gb). The amount of MGVs found in the different fraction sizes was at the same scale (5.2-6.2\times10^6 variants) except for fraction 5-20 µm that presented half the MGVs and had the lowest computation time. This may be because of less genomic complexity in this fraction size, as shown previously (Carradec et al. 2018). The MGVs can be downloaded and directly used by the scientific community in order to perform new analyses of genomic diversity on any organism of interest as demonstrated on *O. nana*.

Table 2: Marine genomic variants produced by *DiscoSnp++* on *Tara Oceans* metagenomic data from the Atlantic Ocean and the Mediterranean Sea.

<table>
<thead>
<tr>
<th>Fraction size (µm)</th>
<th>Number of stations</th>
<th>Number of reads used</th>
<th>Number of variants</th>
<th>Computation time (hours)</th>
<th>Max memory used (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8-5</td>
<td>25</td>
<td>11.3 x 10^9</td>
<td>5.5 x 10^6</td>
<td>64</td>
<td>107</td>
</tr>
<tr>
<td>5-20</td>
<td>27</td>
<td>11.8 x 10^9</td>
<td>2.3 x 10^6</td>
<td>60</td>
<td>107</td>
</tr>
<tr>
<td>20-180</td>
<td>31</td>
<td>11.2 x 10^9</td>
<td>5.2 x 10^6</td>
<td>105</td>
<td>110</td>
</tr>
<tr>
<td>180-2000</td>
<td>31</td>
<td>11.2 x 10^9</td>
<td>6.2 x 10^6</td>
<td>124</td>
<td>120</td>
</tr>
</tbody>
</table>

**Discussion**

Like any reference based variant detection method, *DiscoSnp++* limitations are mainly due to genomic approximate repeats. Reads from approximate repeats and, in the metagenomic framework, reads from similar inter-species genomic regions contain the same signal as those from regions containing intra-species variants. As shown by the results from this study, those imperfect predictions are not an insurmountable limitation for population genomics analysis where alignment-based and reference-free–based approaches provide similar conclusions in terms of population differentiation and overlapping results in terms of natural selection. Moreover,
DiscoSnp++ is an order of magnitude faster and uses fewer resources (Peterlongo et al. 2017; Uricaru et al. 2015). In the case of the admixture of two closely related species, neither the alignment-based nor the reference free-based approach allows the removal of inter-species variants which reduce the number of populations that can be integrated into the population genomic analysis focused on a single species.

The MGVs detected de novo with DiscoSnp++ from Tara Oceans data can be downloaded from [http://bioinformatique.rennes.inria.fr/taravariants/](http://bioinformatique.rennes.inria.fr/taravariants/) and used directly on any genome or transcriptome provided by the users to create VCF files without computation of the variant calling. This can be done by running only the final DiscoSnp++ step ‘run_VCF_creator.sh’ that can be done on a laptop computer. This allows the community to avoid (i) the systematic downloading of the whole Tara Oceans metagenomic data set that needs investment in large infrastructure for data storage and backup and (ii) the alignment of the reads to their genomes and transcriptomes of interest that needs investment in computational power. As demonstrated in this study, the MGVs allow an accurate analysis of the molecular diversity of the plankton present in the AO and MS that were captured during the Tara Oceans expedition. In addition to the lack of reference sequences for plankton, depending on the genome size and abundance of the studied plankton in the Tara Oceans samples, the use of the MGVs collection may have some limits. Analyses focusing on small-size genomes (<100 Mb) and abundant protists such as green algae are more likely to provide interesting results compared to those focusing on copepods with large-size genomes (>1 Gb).

The increasing number of large collections of marine plankton samples and their related metagenomic dataset forces a rethinking of the way population genomics can be performed. This can push the community towards the use of a universal genomic resource of variants that can be
updated with the accumulation of newly released metagenomic data. From this perspective, the use of DiscoSnp++ offers a great advantage by providing a uniform method to generate community shared markers that store all the information needed to perform robust downstream population genetics analyses of plankton.

**Acknowledgements**

We wish to thank the individuals and sponsors who participated in the *Tara* Oceans Expedition 2009–2013: Centre National de la Recherche Scientifique, European Molecular Biology Laboratory, Genoscope/Commissariat à l’Energie Atomique, the French Government “Investissements d’Avenir” programmes OCEANOMICS (ANR-11- BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08) and HYDROGEN (ANR-14-CE23-0001). This is *Tara* Oceans contribution number 83.

**Data availability**

The metagenomic data from *Tara* Oceans are available at ENA (Supplementary data S1). The *Oithona nana* genome sequence and annotation are available at ENA with the study Accession no. PRJEB18938. The MGVs files and their corresponding tutorial are available at [http://bioinformatique.rennes.inria.fr/taravariants/](http://bioinformatique.rennes.inria.fr/taravariants/).

**Authors’ contribution**

MA, KS, PP, JG and MAM performed the analyses. OJ, DL, PP and MAM designed the study. MAM wrote the manuscript, and all authors accepted its final version.
References


Delmont TO, Kiefl E, Kilinc O, et al. (2017) The global biogeography of amino acid variants within a single SAR11 population is governed by natural selection.


Peterlongo P, Riou C, Drezen E, Lemaitre C (2017) DiscoSnp++: de novo detection of small variants from raw unassembled read set(s) bioRxiv 209965;.


