Detecting Structural Variants and Associated Gene Presence–Absence Variation Phenomena in the Genomes of Marine Organisms

Marco Sollitto, Nathan J. Kenny, Samuele Greco, Carmen Federica Tucci, Andrew D. Calcino, and Marco Gerdol

Abstract

As complete genomes become easier to attain, even from previously difficult-to-sequence species, and as genomic resequencing becomes more routine, it is becoming obvious that genomic structural variation is more widespread than originally thought and plays an important role in maintaining genetic variation in populations. Structural variants (SVs) and associated gene presence–absence variation (PAV) can be important players in local adaptation, allowing the maintenance of genetic variation and taking part in other evolutionarily relevant phenomena. While recent studies have highlighted the importance of structural variation in Mollusca, the prevalence of this phenomenon in the broader context of marine organisms remains to be fully investigated.

Here, we describe a straightforward and broadly applicable method for the identification of SVs in fully assembled diploid genomes, leveraging the same reads used for assembly. We also explain a gene PAV analysis protocol, which could be broadly applied to any species with a fully sequenced reference genome available. Although the strength of these approaches have been tested and proven in marine invertebrates, which tend to have high levels of heterozygosity, possibly due to their lifestyle traits, they are also applicable to other species across the tree of life, providing a ready means to begin investigations into this potentially widespread phenomena.

Key words Hemizygosity, Genome, Presence–Absence Variation, Structural Variation, Pangenome

1 Introduction

It is becoming apparent from whole genome resequencing that intraspecific genome variation, once mostly thought to be a prerogative of prokaryotes, is more widespread than previously expected in eukaryotes. These phenomena, collectively referred to as structural variations (SV), encompass any large scale difference in genome architecture and, in addition to translocations, duplications, inversions, and copy number variations, they also include

insertions and deletions [1]. Insertions and deletions are particularly consequential because they include regions of a species' genome which are subject to presence–absence variation (PAV) either between homologous chromosomes within an individual or between individuals. The outcome of such variation is that the genetic complement of two individuals belonging to a single species or population may differ from one another. This deviates from the commonly held idea that variation between individuals is primarily the result of allelic variation of a common genomic repertoire. In genome assemblies of diploid organisms, PAV can be determined by the observation of large genomic regions present in a hemizygous state, that is, only present in one of the two homologous chromosomes. These loci can be readily identified by their reduced read coverage which exists at approximately half that of flanking homozygous regions. On a broader population level, any genomic region affected by PAV could be present in zero, one, or two copies (found in a nullizygous, hemizygous, or homozygous state) across different individuals.

Unexpectedly, not only do genomic regions subject to PAV include intergenic or intronic sequence, but they may also include protein-coding genes which retain full functionality, with a potential impact on phenotypic diversity [2] and adaptation [3, 4]. In this respect, the gene repertoire of any given species can be divided into two different categories: (1) *core* genes, which are thought to carry out functions indispensable for survival and are shared by all individuals, and (2) *dispensable* genes, which may provide accessory functions and are only found in a subgroup of individuals. Altogether, the full complement of *core* and *dispensable* genes found across all the individuals belonging to the same species define the pangenome [5]. While this concept has been long applied to microbial and viral genomes [6, 7], broad scale genomic PAV has also been observed in a number of plants, microalgae and fungi [3, 8–11], where the accessory functions provided by *dispensable* genes have been linked with an improved ability to colonize new ecological niches, to withstand biotic and abiotic stress, or to escape host immune response in the case of pathogenic organisms. Even though pangenomic studies have been only very recently extended to metazoans, gene PAV has already been noted in a number of lineages, including in humans [12–14]. Following early observations collected for single genes [15], a subsequent genome-wide approach allowed for the first description among marine animals of the presence of a pangenome in the Mediterranean mussel *Mytilus galloprovincialis* [16].

The scale of PAV varies widely between species. In plants [10], and some animal species, variable regions, which can be readily identified thanks to their hemizygous state (i.e., for diploid species, regions with half the expected sequencing coverage), can encompass large proportions of the genome. For example, hemizygous

regions account for more than 35% of the reference genome assembly of *M. galloprovincialis*, which results in the presence of more than 20,000 *dispensable* genes [16]. However, in other animals, such as humans, PAV accounts for a much smaller fraction of the genome. Typically, less than 0.1% of the human genome varies between individuals and a small fraction of genes (i.e., 240 out of the ~20,000 genes annotated in the human genome) are potentially subject to homozygous deletions [12]. The rate of *dispensable* to *core* genes is generally used to describe the "openness" of a pangenome. Even though this definition is somewhat arbitrary, genomes such as those of the mussel, plants and bacteria, where a high proportion of *dispensable* genes have been detected, are considered "open" pangenomes. In contrast, genomes with a relatively lower *dispensable* to *core* gene proportion are defined as "closed" pangenomes. While the presence of "open" pangenomes has been previously associated with an improved potential for adaptation in bacteria [4], the possibility that accessory genes may also be associated with neutral or slightly deleterious effects on fitness [17] reveals that a consensus about the functional role of microbial pangenomes is still far from being reached by the scientific community. Although *dispensable* genes are characterized by an enrichment of functions linked with immune defense and survival in *M. galloprovincialis* [16], it is unclear whether this seemingly adaptive role for PAV is common to other marine species that share similar life traits. This remains an open question of great ecological and evolutionary relevance.

Due to the relatively recent inclusion of metazoans among the targets of pangenomic studies, no consolidated method for the analysis of SVs and PAV have been developed and broadly validated to date in these organisms, which (unlike bacteria) have complex diploid genomes. The standard method for detecting genomic regions subject to PAV between individuals is the use of whole genome resequencing (WGR). In this approach, a sample of individuals from a population or species are resequenced and compared to a reference genome assembly of the target species. This allows for the identification of genomic loci that are subject to presence–absence variation between each sequenced individual and the reference assembly. Through this iterative comparative approach, the true genomic complement of the population or species can be approached as more individuals are sampled.

While WGR remains the only method to determine the complete genomic complement of a population or species, it is expensive and time consuming, and is not an efficient option to gain a first insight into the level of PAV that may exist in a population. Moreover, it does not necessarily provide information on PAV between homologous chromosomes within an individual. We have established a pipeline which can be applied to any well-assembled genome, and which leverages the long reads used in the initial

assembly to identify genomic loci that are subject to PAV between the homologous chromosomes of diploid individuals rather than between individuals. As previously mentioned, PAV between homologous chromosomes results in stretches of hemizygous DNA that can be identified using preexisting reads. This means that no additional sequencing is required to gain a preliminary understanding of the level of hemizygosity in the sequenced individual [18]. This method accounts for regions of tandem duplication and false positives due to genome assembly artifacts, producing robust yet conservative estimates of hemizygosity through a mapping and coverage estimation-based approach.

The stark differences in PAV raise many interesting questions for evolutionary biology. In an era where genomes are generally generated from single individuals, it is useful to consider whether they are representative of populations as a whole, and the methods described here provide a ready initial assessment scheme. In concert with the method for the detection of hemizygous regions in diploid genome assemblies detailed in this chapter, we also describe here a protocol for the use of Illumina WGR data for the detection of gene PAV among individuals. While the costs of third generation sequencing technologies, such as those offered by Oxford Nanopore and Pacific Biosciences are rapidly dropping, price considerations still prevent the broad applicability of such methods to large scale resequencing projects. For many research groups the use of short reads remains the preferred choice for high-throughput nonmodel species genomics. The reliability of the short-read based method presented here has been proven on mollusks, and in particular in *M. galloprovincialis* [16]. However, this approach will be just as applicable to other species, as long as the quality of both the raw resequencing data and of the annotated reference genome are sufficient.

The ready nature of the approach we describe could be helpful in ascertaining the extent to which phenomena such as genomic and genic PAV are coupled to particular environments or life history traits. This approach could be particularly useful in understanding heterozygosity in marine broadcast spawners. Marine species have been shown to possess higher levels of heterozygosity than other species [19]. This could be due to large effective population sizes, rapid dispersal, high levels of genetic outcrossing and panmixture, coupled with local variations in environment, although factors such as local sperm viability do impact "real world" genetic diversity levels [20]. However, genetic variation is ultimately the result of mutational events, and when many mutations are retained rather than purged from genomes within a population over evolutionary timescales, this will result in high levels of heterozygosity. Mutation rate will be further impacted by the rate of transpositional activity or low fidelity error correction mechanisms, which we have previously shown to be likely involved in PAV in mollusks [18].

This approach provides new means for raising and addressing questions regarding the origin, maintenance and prevalence of PAV in nonmodel species, which may be more suitable for addressing these issues than more typical laboratory organisms. It also provides a means of linking these investigations with gene PAV, the adaptive nature of this phenomenon, and processes of gene family expansion and contraction. As genomic sequences become readily available across the tree of life, and as long read sequencing reduces in price, it will only become easier to gain a more representative understanding of the level of PAV in all species, and to reflect on the contribution of these phenomena to the evolution of marine life as we know it.

2 Materials

2.1 Target Selection

The procedures described in Subheading 3.1 involve the analysis of a near complete or complete (chromosomal) haploid genome assemblies produced from diploid species using both PacBio long reads (*see* **Note 1**) and Illumina short reads. Care should be taken to ensure that the sequenced individual was not the product of selective breeding as this would likely reduce the level of structural variation that would be expected from comparable individuals from outbreeding or wild populations. Target species/individuals should be selected based on the availability of the following.

1. A high-quality genome assembly (*see* **Note 2** for additional recommendations).
2. High quality PacBio long read libraries (good results were attained with >40× coverage, and we have not tested this to determine a minimum required level) derived from the same individual from which the genome assembly was constructed.
3. High quality Illumina short read libraries (good results were attained with >20× coverage) derived from the same individual from which the genome assembly was constructed. Please *see* **Note 3** for additional recommendations.

On the other hand, the analyses described in Subheading 3.2 can be applied to any fully sequenced and annotated genome, regardless of the use of long reads during the de novo assembly process. The choice of the target species should be based on the following criteria.

1. Availability of a full genome assembly with associated gene annotations in a .gff file format. Since the accuracy of gene model annotations will be crucial to allow a proper interpretation of PAV data, we suggest to evaluate the completeness of the set of the annotated genes in the target species with BUSCO [21], selecting the appropriate taxonomic dataset available from the most recent release of OrthoDB [22].

Availability of high quality Illumina WGR short read libraries (good results were attained with >20× coverage) derived from one or more individuals different from the one used for the generation of the genome assembly. Please *see* **Note 2** for additional recommendations.

2.2 Software

All computational procedures should be performed on a Linux-based system and the following software should be installed prior to commencement.

bbduk [23]: https://sourceforge.net/projects/bbmap

bedmap [24]: https://github.com/bedops

bedtools [25]: https://github.com/arq5x/bedtools2

BUSCO [21]: https://busco.ezlab.org

Bwa [26]: http://bio-bwa.sourceforge.net

bwa mem [27]: https://github.com/lh3/bwa

fastp [28]: https://github.com/OpenGene/fastp

FastQC [29]: https://github.com/s-andrews/FastQC

jellyfish [30]: https://github.com/gmarcais/Jellyfish

mosdepth [31]: https://github.com/torfinnnome/mosdepth

Numpy [32]: https://numpy.org

Pandas [33]: https://pandas.pydata.org

pbmm2 [34]: https://github.com/PacificBiosciences/pbmm2

pbsv [35]: https://github.com/PacificBiosciences/pbsv

Python: https://www.python.org

samtools [36]: https://github.com/samtools/samtools

Scipy [37]: https://www.scipy.org

Tandem Repeats Finder [38]: https://tandem.bu.edu/trf/trf.html

3 Methods

3.1 Allelic Structural Variation Detection Within Assembled Genomes

Identify tandem repeats in genome assembly and convert output to zero-based six field bed file.

1. `trf genome.fa 2 7 7 80 10 50,500 -d -h`
2. `TRFdat_to_bed.py --dat genome.fa.2.7.7.80.10.50.500_mod.dat --bed genome.fa.2.7.7.80.10.50.500.bed`
3. `awk '{print $1"\t"$2-1"\t"$3"\t"$4}' genome.fa.2.7.7.80.10.50.500.bed | sed 's/Sequence://' | awk '{print $0"\t"$3-$2"\t1"}' >genome.fa.2.7.7.80.10.50.500_0based_6field.bed`

Align PacBio reads to genome assembly with the minimap2 [39] wrapper pbmm2.

1. `pbmm2 align -j 16 genome.fa pacbio_reads.fofn genome.aligned.bam --sort --median-filter --sample sample1`

Identify structural variants in genome assembly (*see* **Note 4**).

1. `pbsv discover --tandem-repeats genome.fa.2.7.7.80.10.50.500_0based_6field.bed genome.aligned.bam genome.svsig.gz`

2. `pbsv call -j 16 genome.fa genome.svsig.gz genome.var.vcf`

Extract deletions (DELs) and insertions (INSs) for which two alleles can be detected to avoid likely false positives that occur due to genome assembly errors.

1. `grep DEL genome.var.vcf | grep PASS | grep -v '1/1' | awk '{print $1"\t"$2-1"\t"$2+length($4)-1"\t"$3}' | awk '{print $0"\t"$3-$2"\t1"}' >genome.var.DEL.6field.bed`

2. `grep INS genome.var.vcf | grep PASS | grep -v -P '1/1' | awk '{print $1"\t"$2-1"\t"$2+length($5)-1"\t"$3}' | awk '{print $0"\t"$3-$2"\t1"}' >genome.var.INS.6field.bed`

Remove adapters and low quality regions from Illumina libraries. This step could be carried out using several different tools as an alternative to bbduk, reported in the example below (*see* **Note 5** for additional recommendations).

1. `bbduk.sh in1=reads_1.fastq in2=reads_2.fastq out1=reads_clean_trimmed_1.fq out2=reads_clean_trimmed_2.fq ref=adapters.fa ktrim=r k=25 mink=11 hdist=1 qtrim=r trimq=30 tpe tbo threads=8`

Map processed reads to genome assembly, merge resulting bam files if more than one library is independently mapped (**step 2**) and convert to fasta format.

1. `bwa mem -t 16 genome.fasta reads_clean_trimmed_1.fastq reads_clean_trimmed_2.fastq | samtools sort -@16 -o reads_bwa_aligned.bam -`

2. `samtools merge mergedBamFile.bam *.bam`

3. `samtools view -@ 8 -F 4 -h mergedBamFile.bam >all_mapping.sam`

4. `reformat.sh in=all_mapping.sam out=all_mapping.fa`

Produce kmer histogram of mapped reads.

1. `jellyfish count -t 8 -C -m 21 -s 16G all_mapping.fa -o all_reads.jf`

2. `jellyfish histo -o all_reads.histo all_reads.jf`

Extract reads that map to deletions and that are completely contained within a deletion.

1. `samtools view -@ 8 -F 4 -h -b -L genome.var.DEL.6field.bed mergedBamFile.bam >del_mapping.bam`

2. `bedmap --echo --fraction-map 1 <(bam2bed <del_mapping.bam) genome.var.DEL.6field.bed >del_reads.bed`

3. `cut -f1-6 del_reads.bed >del_reads.6field.bed`

4. `fastaFromBed -fi genome.fa -bed del_reads.6field.bed -fo del_reads.fa`

Produce kmer histogram of deletion mapping reads to compare against kmer histogram of all mapped reads.

1. `jellyfish count -t 8 -C -m 21 -s 16G del_reads.fa -o del_reads.jf`

2. `jellyfish histo -o del_reads.histo del_reads.jf`

Produce bam file of reads mapped to deletions.

1. `cut -f4 del_reads.bed | sort -u >del_reads.names`

2. `samtools view mergedBamFile.bam | fgrep -w -f del_reads.names >del_reads.sam`

3. `samtools view -H mergedBamFile.bam >del_reads_header.sam`

4. `cat del_reads.sam >>del_reads_header.sam`

5. `samtools view -@8 -S -b del_reads_header.sam >del_reads_header.bam`

6. `samtools index -@8 del_reads_header.bam`

Determine median coverage of deletion mapped reads.

1. `mosdepth -t 8 -m -b del_reads.6field.bed genome del_reads_filtered.bam`

Produce histogram of deletion coverages (rounded to whole number) so that each deletion has its coverage counted only once.

1. `zcat genome.regions.bed.gz | awk 'BEGIN{OFS=FS="\t"} {$6=sprintf("%.0f",$5) }1' | cut -f6 | sort -n | uniq -c | sed -e 's/^ *//' -e 's/\ /\t/' | awk '{print $2"\t"$1}' | head -200 >deletion_coverage.txt`

Determine read coverage at every nucleotide of the genome and split the output file by chromosome.

1. `genomeCoverageBed -d -ibam mergedBamFile.bam >genome.cov`

2. `awk '{print>$1".cov"}' genome.cov`

Calculate median coverage using a 1000 bp sliding window and a step of 1 nt then count how many windows have each coverage value (rounded to whole number). Script (median_sliding_window.gawk) can be found in **Note 6**.

1. `for i in *cov ; do cat $i | sh median_sliding_window.gawk | sed 's/\ /\t/g' >`echo $i | sed 's/cov/median/'` ; done`

2. `cat *median | awk 'BEGIN{OFS=FS="\t"}{$4=sprintf("%.0f", $3) }1' | cut -f4 | sort -n | uniq -c | sed -e 's/^ *//' -e 's/\ /\t/' | awk '{print $2"\t"$1}' >coverage_count.txt`

3.2 Gene Presence–Absence Variation Detection and Analysis

Some useful scripts to run the pipeline of analysis described below can be found online at the following link: https://github.com/Carmen-Tuc/PAV_pipeline.

Recover all the short read Illumina libraries derived from WGR experiments carried out on the target species of interest (*see* **Note 1** for further recommendations about the selection of the data to be analyzed). Although the procedure described below could be applied, with some modifications, to other types of sequencing reads, the reliability of this analysis pipeline has only been tested so far with Illumina short reads (*see* **Note 7** for further details).

Perform quality control on reads and trim them accordingly. This task can be performed with a number of different tools, including FastQC [40], that we are using in the example below. *See* **Note 5** for further recommendations.

```
# perform quality control
fastqc read1.fq read2.fq

# perform trimming
fastp -i read1 file.fq -I read2 file.fq -o read1.trimmed.fq -O
read2.trimmed.fq --detect_adapter_for_pe -V -f 5 -t 3 -F 5 -T
3 -h report.html -w 16 -x -n f -5 -3 -p -l 75 -M 24
```

Map the reads on the reference genome, making sure to allow the nonunique mapping of reads (*see* **Note 8**). Each sequencing library should be mapped separately from the others. We recommend the use of bwa [27], paired with samtools [41], for this task.

```
# index the genome
bwa index genome.fna
```

```
# map trimmed reads on the reference, piping the result into
# samtools view for bam conversion
bwa mem -M -t 64 genome.fna read1.trimmed.fq read2.trimmed.fq
| samtools view -bS - > mapping.bam

# sort the bam file
samtools sort -@ 64 -O bam -o mapping.sorted.bam mapping.bam
```

Produce a coverage file for each set of Illumina reads mapped. This file should link each position of the genome with read mapping coverage. Make sure to include sites that display coverage = 0.

```
# produce the depth file
samtools depth -aa mapping.sorted.bam > genome.depth
```

Extract the coordinates for all genes and corresponding exons from the genome annotation.

```
# extract meaningful rows and columns from the gff annotation
awk '{ if ($3 == "exon") print ($0) }' genome.gff | cut -f
1,4,5,9 > filtered_annotation.tsv

# clean the last column, reducing it to the exon id only.
# this command may vary depending on the format of the
annotation

cut -d "|" -f 1,7 filtered_annotation.tsv | sed 's/ID=exon-
gnl|//g' | cut -d ";" -f 1 > exons_coordinates

# input row example:
MTYJ01000001.1 Genbank exon 13344 13541 . + .
ID=exon-gnl|WGS:MTYJ|mrna.BV898_00003.1-1;Parent=rna-gnl|WGS:
MTYJ|mrna.BV898_00003.1;gbkey=mRNA;locus_tag=BV898_00003;or-
ig_protein_id=gnl|WGS:MTYJ|BV898_00003.1;orig_transcrip-
t_id=gnl|WGS:MTYJ|mrna.BV898_00003.1;partial=true;product=hy-
pothetical protein;start_range=.,13344

# output row example:
MTYJ01000001.1 13344 13541 BV898_00003.1-1
```

Run a BUSCO analysis [21] on the extracted transcript sequences obtained from the genome, making sure to select the most appropriate dataset available in OrthoDB, and create a list of genes flagged as "complete".

```
busco -m transcriptome -l $lineage -c 64 -i transcript.fa -o
busco_exons -f
```

```
awk '$2 == "Complete"' full_table.tsv | cut -f 3 | cut -d ":"
-f 1 > complete_busco.list
```

Load the coverage file in a data analysis environment such as R or python. In the following example, python will be used with the numpy [32] and pandas [33] modules. Estimate the average coverage for the exon regions of each gene (if alternatively spliced exons are present, make sure to include the longest exons available. *See* **Note 9** for a simple python code to achieve this computation and **Note 10** for an explanation about the need to exclude intronic regions from these calculations.

Plot the distribution of the exon coverages of the complete set of BUSCO genes. As this set only includes single-copy genes, the graph produced should display a gaussian shaped curve, with its peak approximately corresponding to the expected genome sequencing depth (*see* **Note 11** for additional details). Estimate the median coverage of the BUSCO genes and divide it by eight to obtain a coverage threshold for calling absent genes (*see* **Note 12** for additional details).

```
def readlist(filename):
    with open(filename, "r") as input_list:
        return([x.rstrip() for x in input_list.readlines()]

complete_busco_genes = read_list("complete_busco.list")

exons_busco_coverage = total_exons[total_exons.ID_gene.isin
(complete_busco_genes)].groupby("ID_gene").apply(flatten_ex-
ons)
sns.distplot(exons_busco_coverage)
threshold = exons_busco_coverage.median()/8
```

Plot the exon coverage for all genes annotated in the reference genome and observe the peaks in the generated graph. The presence of genes with zero coverage and of a hemizygous peak in addition to the homozygous peak indicates the detection of gene PAV (*see* **Note 13** for additional details).

```
# Update xlim accordingly to the sequencing depth
xlim = 100
sns.distplot(coverage_total_exons[coverage_total_exons<xlim])
```

For troubleshooting purposes, we recommend to run the same analysis using the Illumina short reads library originally used for the de novo assembly of the reference genome itself. In this case, no absent gene should be identified (*see* **Note 14** for additional details).

Please note that this protocol has some limitations in PAV detection for genes subjected to copy number variations and, in particular, for genes associated with transposable element activity (*see* **Note 15**).

The users might want to optionally collect unmapped reads resulting from the bwa mem mapping step. As these may belong to regions which are not found in the reference genome, the de novo assembly of unmapped paired-end reads can be used to generate a collection of alternate contigs, building the pangenome of the species of interest. *See* **Note 16** for some suggestions.

3.3 PAV Gene Functional Enrichment Analysis

Perform enrichment analysis by hypergeometric test [42] on the union set of the *dispensable* genes identified in all the analyzed resequenced genomes, using the complete set of genes annotated in the reference genome as a background (this will constitute the "universe" dataset, see below). To achieve this, at least one type of annotation will be needed. While multiple alternative functional annotation resources could be used, we recommend Gene Ontology (GO) terms, which can be further subdivided in three main categories: Biological Process, Molecular Functions, and Cell Component [43, 44]. For functional enrichment analysis in nonmodel species, we suggest using PFAM [45] in addition to GO terms (*see* **Note 17**).

While these annotations are usually already available from the genome annotation file, they can be also obtained with InterProScan [46, 47], which should be run on the proteome resulting from the translation of the coding sequences associated with gene model annotations.

A simple python script to perform functional enrichment tests is available online (https://gitlab.com/54mu/enrichment_test). This analysis requires a "universe" dataset (the complete set of genes annotated in the reference genome) and a subset of genes to test for enrichment (which in this case comprises all the *dispensable* genes identified with the previous steps). The universe dataset needs to be a table matching gene ids and feature ids, one per line, as reported in the example below. It is also important for each match to be unique.

extract from a universe file for GO enrichment.

```
...
LOC105326593    GO:0065003
LOC105326593    GO:0070062
LOC105326593    GO:1903561
LOC105344258    GO:0000212
LOC105344258    GO:0001516
LOC105344258    GO:0001525
...
```

```
Tyr     GO:0042470
Tyr     GO:0046872
Tyr     GO:0050149
Tyr     GO:0055114
Pepck   GO:0000287
Pepck   GO:0003729
Pepck   GO:0004550
Pepck   GO:0004611
Pepck   GO:0004613
...
```

The subset file must contain the list of unique ids of the *dispensable* genes subject to PAV. The last parameter required to run the analysis is the number of unique features in the universe (in this case the total number of genes annotated in the reference genome).

Enriched annotations can be filtered by the user based on arbitrary p-value thresholds, which may be also combined with a threshold of observations for any given annotation in the *dispensable* gene subset, to filter out annotations linked with a very low number of genes.

4 Notes

1. This procedure may potentially also work with long reads generated with Oxford Nanopore technologies (ONT), that is, obtained with MinION, GridION, or PromethION platforms. Since we have not tested the performance of the protocol with this type of data, in case of availability of ONT long reads, we suggest testing alternative SV detection tools in addition to psvb. Among these, NanoSV [48], SVIM [49] and cuteSV [50] have been previously indicated as suitable for SV detection using ONT data.
2. We here define a "high quality" genome as a genome whose assembly approaches a chromosome-scale quality. Ideally, such a genome would display a low amount of genomic sequence located in unplaced scaffolds and a very high completeness, which can be generally estimated either with gene model-centric tools, such as BUSCO [21], or with k-mer based methods, such as Merqury [51].

   Most importantly, a high quality reference genome analyzed with the hemizygous regions and PAV detection protocols must be entirely devoid of exogenous contaminations, that is, contigs and scaffolds deriving from other associated organisms, such as symbionts, parasites, and pathogens, as these may lead to artifacts (i.e., contaminant genes might be wrongly

identified as *dispensable* genes part of the pangenome of the species of interest). The overwhelming majority of the reference genomes deposited in public repositories, such as Ensembl, NCBI Genomes and others, are expected to meet these quality requirements. Nevertheless, we recommend extra caution, especially when no other genomes from closely related species are available as a reference, since the discrimination between novel orphan genes, horizontally transferred genes and exogenous contaminants is often not a trivial task [52].

3. Illumina genomic DNA libraries can be obtained with a number of different commercial kits and sequenced using different strategies. These can affect in a significant way both the accuracy of read mapping and the evenness of read coverage across the reference genome assembly. The protocols described in this chapter were extensively tested on regular paired-end libraries sequenced on HiSeq and NovaSeq series platforms with a 2×100, 2×125 or 2×150 strategies, which usually allow the attainment of a relatively uniform distribution of reads on the full length of chromosomes, with the exception of local peaks with very high coverage in highly repetitive regions. The use of particularly short reads (e.g., <50 base pairs) generated with single-end sequencing might lead to a decrease in mapping accuracy, determining an increase in the fraction of reads mapping to multiple genomic sites. This might have a negative effect on the quantification of the sequencing coverage of some loci, most notably those that include paralogous genes.

   At the same time, we discourage the use of paired-end libraries which derive from chromosome conformation capture approaches (such as Hi-C or Dovetail Genomics Omni-C™) and include long-range connectivity information, as these may lead to an uneven distribution of mapped reads, not suitable for downstream PAV analyses. On the other hand, we have successfully tested this protocol on 10× Genomics Illumina libraries [18].

4. Due to the limitations of pbsv, deletions larger than 100 kb will be missed. We recommend checking the length distribution of the insertions identified in the target genome to evaluate whether deletions larger than this threshold are likely to be present. In the case of genomes including a high amount of large hemizygous regions, the output of this protocol might result in an underestimate of total hemizygous genomic DNA content.

5. The trimming parameters may be modified based on the quality of the raw sequencing data available. In particular, note that Illumina short reads generated with different library preparation kits might include different adapter sequences and

barcodes. Make sure to check the technical documentation provided by the manufacturers to identify the most appropriate list of adapter sequences to be used. We recommend using stringent trimming parameters to discard all possible sources of bias. Discarding all the reads whose length, following the trimming procedure, is lower than 50 nucleotides, might be also beneficial in some cases, as these reads may result in ambiguous mappings on multiple sites.

6.

```
median_sliding_window.gawk script
#!/usr/bin/sh
gawk -v wsize=1000 '
   BEGIN {
    if (wsize % 2 == 0) { m1=wsize/2; m2=m1+1; } else {
    m1 = m2 = (wsize+1)/2; }
   }
   function roundedmedian() {
     asort(window, a);
   return (m1==m2) ? a[m1] : int(0.5 + ((a[m1] + a[m2])
   / 2));
   }
   function push(value) {
     window[NR % wsize] = value;
   }
   NR < wsize { window[NR]=$3; next; }
   { push($3);
     $3 = roundedmedian();
     print $0;
   }'
```

7. Short reads generated with other sequencing platforms (e.g., with BGISEQ platforms) may work as well, as long as the sequencing error rate is in line with those expected from Illumina approaches. We do not recommend using reads generated with sequencing methods that are known to suffer from relatively high error rates, in particular in correspondence with homopolymeric sequence stretches, such as 454 Life Sciences pyrosequencing and Thermo Fisher Scientific Ion Torrent, as they may introduce significant biases in read mapping profiles. This protocol might be implemented in the future to also allow the mapping of resequencing data obtained with third generation methodologies (i.e., nanopore sequencing by Oxford Nanopore and SMRT sequencing by Pacific Biosciences), even though significant modifications might be required to take into account their different length and error rates of the reads generated with these approaches.

8. This can be achieved by selecting the -M option in bwa mem. We have previously shown that this is necessary in order to avoid erroneously obtaining long stretches of genomic sequences with coverage = 0 in the presence of repeats. Not selecting this option would therefore result in an inflation of PAV calls.

9.
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

def flatten_exons(gene):

    # by using a coverage map this function returns the average
    # coverage of the exonic regions of a gene (passed as a
    # pandas groupby object)

    global cov_map
    contig = gene.contig.values[0]
    start = min(gene.start.values)-1
    end = max(gene.end.values)
    size = end - start
    mask = np.full([size], False)
    pairs= np.array([gene.start.values, gene.end.values])
    startendarray = pairs.flatten('F')
    for i in range(0, startendarray.size, 2):
        this_exon = np.arange(startendarray[i]-start,starten-
darray[i+1]-start)
        mask[this_exon] = True
    sub_cov_map = cov_map[contig][start:start+size]
    try:
        return sub_cov_map[mask].mean()
    except Exception as e:
        return np.nan

coverage_file = pd.read_csv("/path/genome.depth", sep = "\t",
names = ["contig", "position", "coverage"])
cov_map = coverage_file.groupby("contig")["coverage"].apply
(np.array).to_dict()
total_exons = pd.read_csv("/path/exons_coordinates", sep =
"\t", names = ["contig", "start", "end", "ID_gene"])

coverage_total_exons = total_exons.groupby("ID_gene").apply
(flatten_exons)
```

10. While the fragments included in Illumina paired-end sequencing data derive from the random fragmentation of genomic

DNA, reads are not necessarily expected to be evenly distributed across the genome assembly. In particular, due to the mapping strategy used in this protocol (*see* **Note 8**), a significant number of reads may be aligned to multiple genomic locations, resulting in local mapping spikes. We expect such spikes to be found in genomic regions encompassing repeats, and therefore to be mostly associated with intergenic regions. However, some species are known to include a considerable amount of heterochromatic introns, which may also include short repeats that could lead to the nonspecific mapping of short reads [53]. Since anomalous mapping may lead to an artefactual inflation of read coverage estimates at the gene level, we recommend to perform such calculations based on exonic regions only.

11. While the theoretical sequencing depth of any resequenced genome can be easily calculated by dividing the total amount of sequence data generated (i.e., the total number of reads multiplied by their average read) by the size of the genome, we recommend using the strategy explained in this protocol to obtain a more reliable estimate of the mapping coverage that would be expected for any single-copy *core* gene in the genome of a diploid organism. As a matter of fact, several factors may cause some discrepancies between the theoretical and actual genome sequencing coverage: namely, the presence of a high amount of reads mapping on mitochondrial (and plastidial) genomes, the presence of exogenous contamination and the occurrence of low quality, unmappable reads. By only taking into account the coverage observed in exonic regions of validated single-copy genes, the method we propose disregards the aforementioned confounding factors and provides a much more reliable estimate.

    The distribution of the coverage of BUSCO genes is expected to follow a Gaussian curve, centered on the actual genome-wide sequencing depth, which identifies the "homozygous peak," as shown in the example provided in Fig. 1.

    Note that this could be also achieved by the use of an approach based on genome-wide k-mer distribution, as explained in Subheading 3.1. However, we noted that the identification of a clear homozygous peak may become difficult in genomes characterized by high heterozygosity rates and resequenced with low coverage (i.e., $<25\times$).

12. This is an arbitrary threshold, which has been previously shown to work well for PAV detection in *M. galloprovincialis*, as it displayed high correlation between in silico gene presence–absence calls and PCR confirmation [16]. In any given resequenced genome of a diploid organism, gene coverage

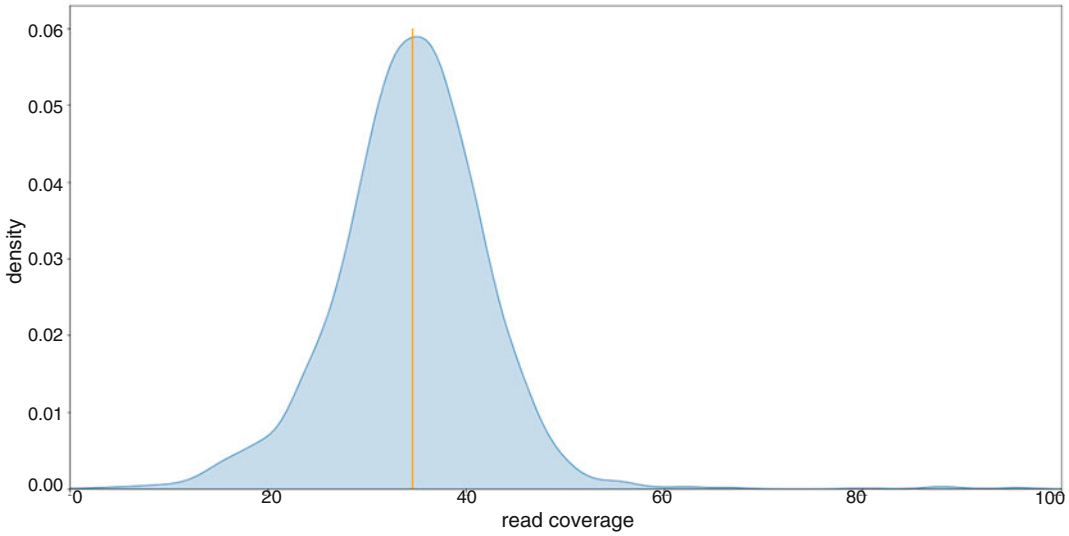

Fig. 1 Expected distribution of the coverage of BUSCO genes in a diploid organism resequenced with a paired-end Illumina short reads library and 36× coverage. The vertical bar indicates the median coverage observed for all BUSCOs. Note the Gaussian distribution around the "homozygous peak"

(calculated on exons only) would be expected to follow a distribution similar to the one shown in the example below (Fig. 2).

In brief, a main "homozygous peak" of coverage, corresponding to the one previously identified from the analysis of BUSCOs, should be observed. This peak indicates genes present with two alleles in the diploid genome. A second "hemizygous peak" should be observed at a coverage equal to exactly half of the "2n" peak. This peak indicates genes present with a single allele in the diploid genome. The relative height of the two peaks can be used to estimate the rate of genes encoded by hemizygous genomic regions in any resequenced individual. Note that species where PAV is very rare are not expected to display a visible hemizygous peak. Finally, a third peak should be observed at zero coverage, marking *dispensable* genes that are present in the reference genome, but absent in the resequenced individual.

We have empirically noted that whenever a relatively low sequencing coverage is used for resequencing (e.g., <50×), the two Gaussian curves cannot be well separated, but rather result in a valley where the upper- and lower-end tails of the two curves are partially overlapping, which does not permit to discriminate with certainty the homozygous or hemizygous state of a given gene (*see* Fig. 2). With sequencing coverage <30×, the hemizygous peak will simply appear as a "shoulder" on the side of the homozygous peak. The same consideration

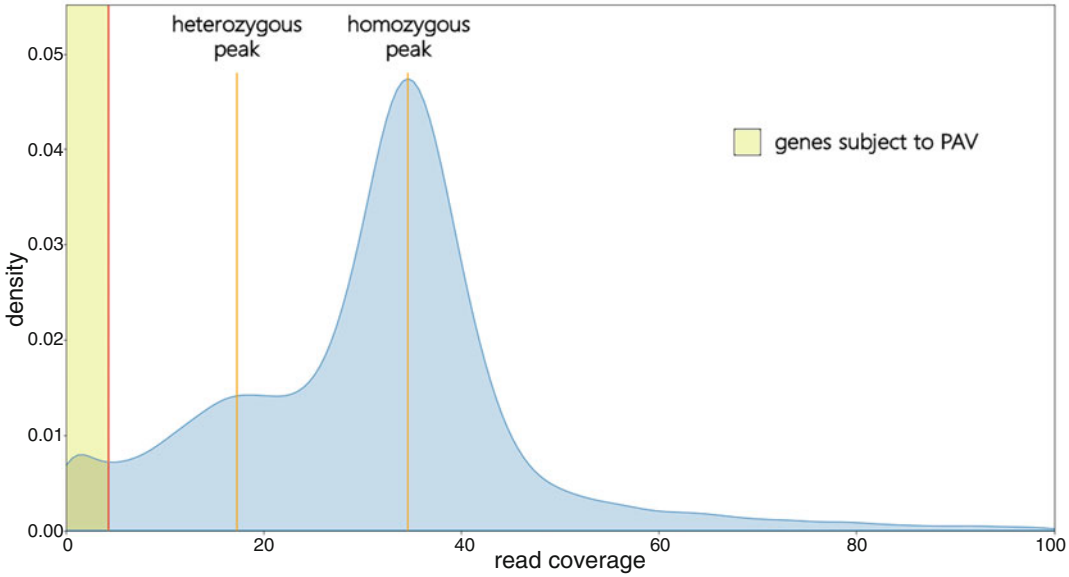

Fig. 2 Expected distribution of the sequencing coverage of all genes in a resequenced genome of a diploid species with widespread PAV. This example shows a genome resequenced with 36× coverage using a paired-end Illumina short read library. The homozygous and hemizygous peaks are placed at 36× and 18×, respectively. The threshold for gene PAV calling is set at 4.5× (i.e., all genes showing a coverage lower than 4.5× are called as absent)

applies to the ability to discriminate between the lower-end tail of the hemizygous peak and the "gene absence peak". This issue is most likely linked with the cross-mapping of a few reads derived from paralogous *core* genes on local regions with high pairwise sequence homology. While higher sequencing depths increase the confidence level of presence–absence calls, we here provide a conservative way of estimating absent genes by using as a threshold 1/8th of the coverage of the homozygous peak. This means that only genes showing an exon sequencing coverage lower than 25% of the expected coverage of a *dispensable* gene found in a hemizygous genomic region will be called as absent.

13. This protocol has been extensively tested on the genome of diploid organisms only. In the case of target genomes with different ploidy levels, multiple peaks might be observed (e.g., four peaks, denoting genes present with one, two, three or four alleles, should be present in a tetraploid species). In such cases, we recommend setting the threshold for PAV detection at 1/4th of the "single allele" peak coverage. Please note that small peaks can be occasionally observed in diploid genomes at coverages which are multiples of the homozygous peak. These may indicate the presence of nearly identical, recently duplicated, paralogous genes. We do not expect this

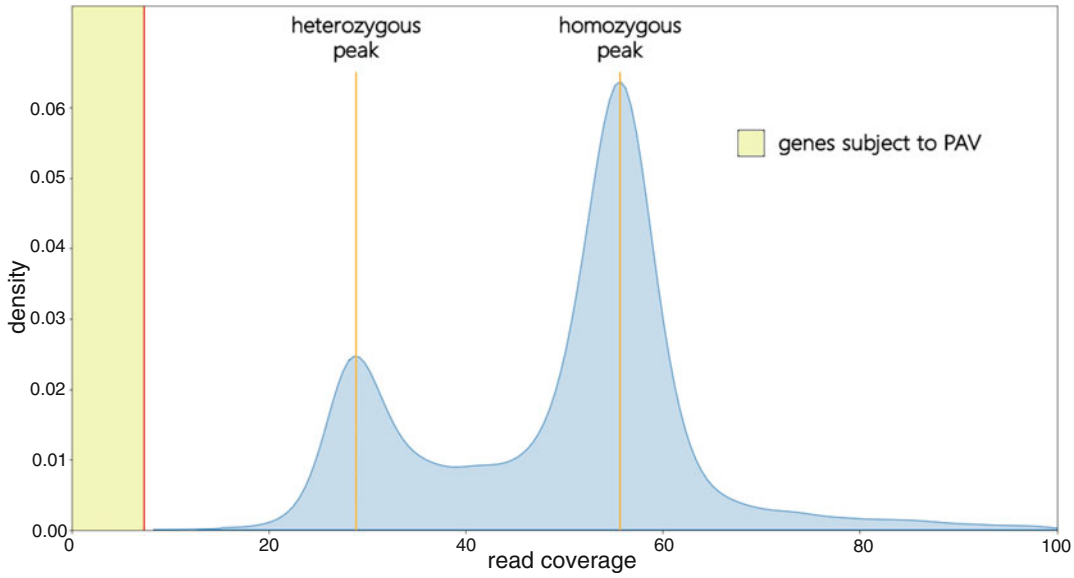

Fig. 3 Expected distribution of the sequencing coverage of all genes in the reference genome of a diploid species with widespread PAV, based on the mapping of a paired-end Illumina short read library generated from the same individual used for the de novo genome assembly. This example shows a genome resequenced with 56× coverage. The homozygous and hemizygous peaks are placed at 56× and 28×, respectively. The threshold for gene PAV calling is set at 7× (i.e., all genes showing a coverage lower than 4.5× are called as absent). Note that no absent genes can be detected in the reference genome

factor to represent a relevant issue in most cases, unless the target species has been subjected to recent whole genome duplication events.

14. We recommend running this test to check the correct setting of all parameters. The mapping of Illumina short read libraries used to perform the de novo assembly of the reference genome against the reference genome itself should, by definition, not produce any gene with zero coverage. On the other hand, depending on the level of hemizygosity of the sequenced individual, the hemizygous peak might be visible. *See* an example of the expected sequencing coverage in Fig. 3.

15. The allowance of read multimapping (*see* **Note 8**), applied in this protocol to avoid the artefactual identification of PAV within highly repetitive genomic regions, may lead to the impossibility of detecting PAV for genes present with multiple identical or nearly identical paralogous gene copies, which may be subject to copy number variation. Although we have previously reported that the activity of transposable elements is likely associated with hemizygous genomic regions in several molluscan reference genomes [18], no significant enrichment of gene families linked with reverse transcriptase, integrase, transposase and other TE-related enzymatic activities has

been identified associated with *dispensable* genes in *M. galloprovincialis* with the protocol explained in Subheading 3.2 [16]. In our interpretation, this discrepancy is linked with the cross-mapping of reads that originated from identical gene copies placed in distinct genomic locations. The protocol described in Subheading 3.1 might be more indicative of PAV detection in these particular cases, thanks to the possibility of exploiting flanking unambiguous sequence to accurately map the reads to the correct genomic locations.

16. Unmapped reads, collected from all the individual resequenced genomes analyzed, can be de novo assembled using several different algorithms, using a recursive reassembly approach [16]. For this task, we recommend paying attention to the assembly of contaminant contigs, which may derive from a series of different sources, in particular in marine filter-feeding organisms. Considering that most pangenomic contigs assembled using unmapped reads are expected to be present in hemizygous genomic regions in the resequenced individuals, the estimated coverage of the hemizygous peak (*see* **Note 12**) can be used as a guidance to set appropriate coverage thresholds to identify contigs which belong to the target species. Both contigs showing excessively high or particularly low coverage compared with expectations should be flagged as suspect and discarded. In addition, nucleotide composition, and GC content in particular, can be used as a complementary information to further detect possible contaminants. In this respect, BlobTools [54] can be very useful. While it is unlikely that contaminant contigs will have both the same coverage and the same GC content expected for hemizygous regions of the target species, we recommend extra caution to avoid including suspect sequences in the pangenome assembly. Therefore, further controls, such as the use of Kraken 2 [55] or BLASTn-based filtering against the complete genome assembly of known contaminants could be applied.

    We also recommend including in the collection of pangenomic contigs only those exceeding a length of 1 kb, as those shorter than this threshold may correspond to local intergenic or intronic regions characterized by high heterozygosity.

    The obtained pangenomic contigs may be then subject to gene annotation, making sure to apply the same annotation pipeline used for the reference genome, providing a list of *dispensable* genes absent from the reference assembly, but present in one or more resequenced genomes of the same species.

17. Considering that GO terms tend to be strongly biased toward model species [56] and that orphan, taxonomically restricted genes without detectable primary sequence homology are highly abundant in the genomes of marine invertebrates [57],

we suggest to use Hidden Markov Model (HMM)-based conserved protein domain annotations to improve the inference of functional enrichment in nonmodel species. Several alternative resources, such as PFAM [45], InterPro [46, 47], and others, may be used for this purpose. Due to the BLAST-independent nature of HMM searches, these allow to improve the annotation rate of genes lacking significant primary sequence homology with entries deposited in public repositories.

References

1. Feuk L, Marshall CR, Wintle RF et al (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15:R57–R66
2. Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18:31–36
3. Read BA, Emiliania huxleyi Annotation Consortium, Kegel J et al (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499(7457): 209–213. https://doi.org/10.1038/nature12221
4. McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40
5. Medini D, Donati C, Tettelin H et al (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594
6. Vernikos G, Medini D, Riley DR et al (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154
7. Aherfi S, Andreani J, Baptiste E et al (2018) A Large Open Pangenome and a Small Core Genome for Giant Pandoraviruses. Front Microbiol 9:1486. https://doi.org/10.3389/fmicb.2018.01486
8. Song J-M, Guan Z, Hu J et al (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 6:34–45
9. Alonge M, Wang X, Benoit M et al (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182:145–161.e23
10. Golicz AA, Bayer PE, Bhalla PL et al (2020) Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet 36:132–145
11. McCarthy CGP, Fitzpatrick DA (2019) Pan-genome analyses of model fungal species. Microb Genom 5:e000243
12. Sherman RM, Forman J, Antonescu V et al (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51:30–35
13. Tian X, Li R, Fu W et al (2020) Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci China Life Sci 63:750–763
14. Li R, Li Y, Zheng H et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63
15. Rosa RD, Alonso P, Santini A et al (2015) High polymorphism in big defensin gene expression reveals presence–absence gene variability (PAV) in the oyster Crassostrea gigas. Dev Comp Immunol 49(2):231–238. https://doi.org/10.1016/j.dci.2014.12.002
16. Gerdol M, Moreira R, Cruz F et al (2020) Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol 21:275
17. Vos M, Eyre-Walker A (2017) Are pangenomes adaptive or not? Nat Microbiol 2:1576–1576
18. Calcino AD, Kenny NJ, Gerdol M (2021) Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philos Trans R Soc Lond Ser B Biol Sci 376: 20200153
19. Martinez AS, Willoughby JR, Christie MR (2018) Genetic diversity in fishes is influenced by habitat type and life-history variation. Ecol Evol 8:12022–12031
20. Olsen KC, Ryan WH, Winn AA et al (2020) Inbreeding shapes the evolution of marine invertebrates. Evolution 74:871–882
21. Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and

annotation completeness. Methods Mol Biol 1962:227–245
22. Zdobnov EM, Tegenfeldt F, Kuznetsov D et al (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res 45:D744–D749
23. Bushnell B. et al. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. No. LBNL-7065E. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA.
24. Neph S, Kuehn MS, Reynolds AP et al (2012) BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919–1920
25. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
26. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760
27. Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. http://github.com/lh3/bwa
28. fastp, Github. https://github.com/OpenGene/fastp
29. Andrews S FastQC, Github. https://github.com/s-andrews/FastQC
30. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770
31. Pedersen BS, Quinlan AR (2018) Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34:867–868
32. Harris CR, Millman KJ, van der Walt SJ et al (2020) Array programming with NumPy. Nature 585:357–362
33. McKinney W (2010) Data Structures for Statistical Computing in Python. Proceedings of The 9th Python in Science Conference, pp. 51-56. https://doi.org/10.25080/majora-92bf1922-00a
34. Pacific Biosciences (2017) pbmm2, Github. https://github.com/PacificBiosciences/pbmm2
35. Pacific Biosciences (2017) pbsv, Github. https://github.com/PacificBiosciences/pbsv
36. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
37. Virtanen P, Gommers R, Oliphant TE et al (2020) Author correction: SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17:352
38. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580
39. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100
40. Wingett SW, Andrews S (2018) FastQ screen: a tool for multi-genome mapping and quality control. F1000Res 7:1338
41. Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008
42. Falcon S, Gentleman R (2008) Hypergeometric testing used for gene set enrichment. Analysis:207–220. https://doi.org/10.1007/978-0-387-77240-0_14
43. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
44. Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOld mine. Nucleic Acids Res 49:D325–D334
45. Mistry J, Chuguransky S, Williams L et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419
46. Jones P, Binns D, Chang H-Y et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240
47. Blum M, Chang H-Y, Chuguransky S et al (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49:D344–D354
48. Stancu MC, van Roosmalen MJ, Renkens I et al (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun 8:1–13
49. Heller D, Vingron M (2019) SVIM: structural variant identification using mapped long reads. Bioinformatics 35:2907–2915
50. Jiang T, Liu Y, Jiang Y et al (2020) Long-read-based human genomic structural variation detection with cuteSV. Genome Biol 21:189
51. Rhie A, Walenz BP, Koren S et al (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21:245
52. Bemm F, Weiß CL, Schultz J et al (2016) Genome of a tardigrade: horizontal gene transfer or bacterial contamination? Proc Natl Acad Sci U S A 113(22):E3054–E3056
53. Espinas NA, Tu LN, Furci L et al (2020) Transcriptional regulation of genes bearing intronic heterochromatin in the rice genome. PLoS Genet 16:e1008637

54. Laetsch DR, Blaxter ML (2017) BlobTools: interrogation of genome assemblies. F1000Res 6:1287
55. Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with kraken 2. Genome Biol 20:257
56. Gaudet P, Dessimoz C (2017) Gene ontology: pitfalls, biases, and remedies. Methods Mol Biol 1446:189–205
57. Khalturin K, Hemmrich G, Fraune S et al (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet 25:404–413