A new framework for detecting copy number variants from single nucleotide polymorphism data: 'rCNV', a versatile R package for paralogs and CNVs detection

Piyal Karunarathne[1]*, Qiujie Zhou[1], Klaus Schliep[2], and Pascal Milesi[1,3†]

1. Plant Ecology and Evolution, Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden
2. Institute of Computational Biotechnology, Graz University of Technology, Graz, Styria, Austria
3. Science for Life Laboratory (SciLifeLab), Uppsala, Sweden

* piyal.karunarathne@ebc.uu.se; piyalkarumail@yahoo.com

† pascal.milesi@scilifelab.uu.se

Abstract

Recent studies show that copy number variants (CNVs), due to their ubiquitous presence in eukaryotes, contribute to phenotypic variation, environmental adaptation, and fuel species divergence at a previously unknown rate. However, the detection of CNVs in genomes, especially in non-model organisms is challenging due to the need for costly genomic resources and complex computational infrastructure. Therefore, to provide researchers with a low-cost and easily accessible resource, we developed a robust statistical framework and an R software package to detect CNVs from SNPs.

The methods implemented in the framework use read-depth variation from high throughput sequence data, and relies on the proportion of heterozygotes in populations and their allelic ratios. The framework consists of two main steps: 1) flagging SNP deviants using allelic ratio deviations from the expectation and excess of heterozygotes under Hardy-Weinberg Equilibrium; followed by 2) filtering putative CNVs from deviant SNPs using data-specific statistical approaches.

Our framework is well-catered for most GBS technologies (e.g., RADseq, Exome-capture, WGS). As such, it allows calling CNVs from genomes of varying complexity. The framework is implemented in the R package "rCNV" which effortlessly automates the analysis. We tested our models on four datasets obtained from different sequencing technologies (i.e., RADseq: Chinook salmon – *Oncorhynchus tshawytscha*, American lobster – *Homarus americanus*, Exome-capture: Norway Spruce – *Picea abies,* and WGS: Malaria mosquito - *Anopheles gambiae*). Our models detected CNVs with substantial accuracy and were significantly efficient with statistically low confident data, where average read-depth and number of samples are limited.

Keywords: Paralogs, CNVs, SNPs, GBS, R statistics

Background

One main outcome of a decade of massive sequencing efforts is that macro-mutations (e.g., inversions, deletions, insertions, duplications), are prevalent in eukaryotic genomes. Such rearrangement in genomes, known as structural variants (SVs), have been found to be crucial for evolution (Chain & Feulner, 2014; Emerson, Cardoso-Moreira, Borevitz, & Long, 2008; Fan & Meyer, 2014; Ohno, 1970). When such structural change occurs in an unbalance process (i.e., insertions, deletions), it leads to copy number variation (CNV) (Collins et al., 2020; Holland, Marlétaz, Maeso, Dunwell, & Paps, 2017). While single nucleotide polymorphism (SNP) is most commonly observed (Savolainen, Lascoux, & Merilä, 2013) and used in genetic studies, recent studies found that the number of base pairs affected by structural variants, and especially by CNVs are multiple times higher than that by SNPs (Catanach et al., 2019; Sudmant et al., 2013). It means that CNVs can introduce large biases in SNPs calling, and thereby biases in population genetic estimates (Nadukkalam Ravindran, Bentzen, Bradbury, & Beiko, 2018; Verdu et al., 2016). On the other hand, CNVs are increasingly acknowledged as a source of genetic polymorphism and for their direct role in organism's short-term evolution (Clop, Vidal, & Amills, 2012; Wellenreuther, Mérot, Berdan, & Bernatchez, 2019) and references therein). While most studies considered paralogs (i.e., multiple copies) as difficult to study and hence removed from analyses, others highlighted CNVs as a source of intraspecific variation during early divergence of species (e.g. (Lauer & Gresham, 2019; Lucek, Gompert, & Nosil, 2019; North, Caminade, Severac, Belkhir, & Smadja, 2020; Wellenreuther & Bernatchez, 2018), phenotypic complexity under fluctuating selection: *Mimulus guttatus* – (Nelson et al., 2019), and adaption (Kondrashov, 2012). However, the full spectrum of the role CNVs play in shaping evolutionary trajectories of species remains grossly unknown (Wellenreuther & Bernatchez, 2018; Wellenreuther et al., 2019, Mérot et al. 2019).

With the giant leap of high-throughput sequencing technologies (e.g., RADseq, GBS, Exome capture, etc.) millions of genomic sequences allowed researchers to study evolutionary process at multi-population scale (Ghosh et al., 2018; Savolainen et al., 2013). However, complex structural variations such as CNVs were overlooked, if not removed, in most population and quantitative genetics analyses (Limborg, Larson, Seeb, & Seeb, 2016; Garrett J. McKinney, Waples, Seeb, & Seeb, 2017). One main reason for the lack of such studies is the challenge of detecting CNVs in genomes. Several methodologies have been developed to detect CNVs but are mostly restricted to whole genome sequencing (WGS) or whole exome sequencing (WES) data and require well assembled and annotated reference genomes (e.g. in clinical studies, Gabrielaite et al., 2021). Such data are expensive to produce impeding large-scale evolutionary studies. This is particularly true for non-model organisms, where reference genomes are often highly fragmented (when available) and poorly annotated. Few recent studies have ventured this useful type of polymorphism and demonstrated their advantage when investigating evolutionary changes (e.g., Cayuela et al., 2021; DeBolt, 2010; Dorant et al., 2020; Garrett J. McKinney et al., 2017). For instance, Dorant et al. (2020) successfully implemented a method to call CNVs, originally developed by McKinney et al. (2017) to detect and remove paralogous regions in SNP data. They showed that local adaptation in American lobster (*Homarus americanus*) populations in southern Gulf of St. Lawrence is better explained by CNVs than by SNP variations. Similarly, Cayuela et al. (2021) used the same approach to show the temperature-dependent clinal shift in Spotted frog (*Rana luteiventris*) in Columbia. Nevertheless, the implementation of their method is limiting due to the specificity of sequencing technology and the lack of a solid analytical pipeline. Therefore, in the present work, we present a robust and an easy-to-use methodology for detecting CNVs from SNP data.

Here we present the R software package "rCNV", a robust framework to identify SNPs deviating from expected allelic-ratio while taking into account the errors linked to sample size and depth of coverage from any type of short-read data. Depending on the downstream analysis, the user can either i) remove

the "deviant" SNPs to produce a clean dataset free of positions lying in putative paralogous or CNV regions for population/quantitative genetic studies and/or ii) call putative CNVs to use them as genetic markers. We also provide further filtering methods combining depth-variation with tests for local enrichment assessments using "deviant" SNPs. "rCNV" is a stand-alone comprehensive package that also includes pre-processing functions that allow handling of raw unfiltered variant call format files (VCF). Downstream analyses using CNVs as genetic markers are also proposed (e.g. V_{ST}, an analogous of F_{ST} ; Redon et al., (2006)). The approach implemented in the R package has been optimized to handle large number of samples and SNPs as well as accounting for sequencing anomalies. The methodology and its implementation are explained in the **Approach** below. The software package is hosted on CRAN [link] and GitHub [link] with a comprehensive tutorial [link].

Approach

Our detection method is a two-step process involving, 1) flagging SNPs deviating significantly from expected allelic ratio (i.e., ratio between the respective read-depth of both alleles in heterozygote individuals) or showing an excess of heterozygotes under Hardy-Weinberg equilibrium (hereafter "deviant SNPs") and then 2) identifying SNPs lying in putative CNV regions (hereafter putative CNVs) from flagged deviant SNPs.

Deviant SNPs (Deviants):

1. *Deviation from expected allelic ratio*: In heterozygous individuals, in absence of sequencing biases or other error-prone factors, one would expect equal number of reads supporting each allele and hence a theoretical allelic ratio of 0.5 (see Box 1). This can be seen as the sampling of N_A reads carrying the alternative allele in a binomial distribution of parameter $\mathcal{B}(N, p)$ where, N is the total number of reads covering a given SNP in a heterozygote and, p, the probability of sampling a read carrying the alternative allele (p = 0.5 in absence of capture bias but can vary between 0 and 1). Therefore, at a given SNP, both the significance of deviation from expected allelic ratios and a higher-than-expected variance in allelic-ratios, can thus be tested; the stringency of the tests increases with both read-depths, N, and the number of heterozygotes, n (see Box 2). This method however, only allows the detection of CNVs where a given mutation is not present evenly among paralogous copies. For instance, one copy carries an adenine (A) while two or more copies carry a thymine (T; see Box. 1) and hence induces a deviation from the expected allele ratio.
2. *Excess of heterozygotes under Hardy-Weinberg equilibrium (HWE)*: Expected genotype frequencies are computed from observed allele frequencies under HWE (Box. 1). We then perform a Chi-square (χ^2) test to test for the significance of an excess of heterozygote between expected and observed genotype frequencies (Fig. i in Box 1). We test for excess of heterozygotes across all samples irrelevant of potential population structure. If population structure were to be present, it should lead to a deficit of heterozygotes (Wahlund effect, Wahlund, (1928)). This is rather cautious of false positives.

Putative CNVs

As mentioned above, factors other than changes in the copy number can explain allelic-ratio deviation (e.g., sequencing technology, mean coverage, quality of reference genome, methodology used for SNP calling). Hence, not all SNPs flagged as "deviants" may lie in a multi-copy region (putatively CNV region). Therefore, we provide the users with different alternative methods to further filter the dataset to identify putative CNVs;

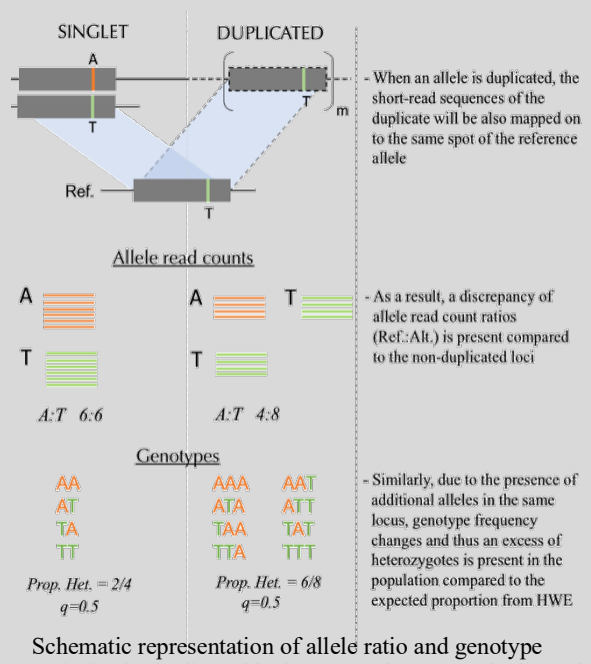
i) *Intersection set:* In this case, a "deviant" SNP is categorized as located in a putative paralog or CNV region if it is supported by at least two of the statistics tested above (i.e., excess of

heterozygotes, *Z*-scores and χ^2-scores: Boxs 1 and 2); see below for our recommendation on using intersection set.

ii) *Unsupervised clustering:* Using *K*-means clustering, the SNPs are grouped into two different clusters (K=2) based on the distribution of Z-scores, χ^2scores, excess of heterozygotes, and read-depth coefficient of variation ($cv_x = \sqrt{\sigma_x^2}/\mu_x$ where, σ_i^2 and μ_i are the variance and the mean coverage of SNP *x* across all individuals.

Note that the *intersection set* approach is based on the significance of the *p*-values for the various scores we computed. The *K-means clustering*, on the other hand, is independent of thresholds or cutoff values as it is conducted on the actual distribution of the statistics and not the *p*-values. For both approaches, users are free to use their own metrics (e.g., allelic probability values) in addition or in place of the ones already computed. The main purpose of the filtering of putative duplicates from the deviant SNPs is to allow the user to use the best statistic to detect putative CNVs based on their data (see recommendations table for more details).

Box 1. Allelic frequency variation with copy number variation (allele duplication)



Schematic representation of allele ratio and genotype variation in duplicated loci compared to a non-duplicated site in short-read sequencing.

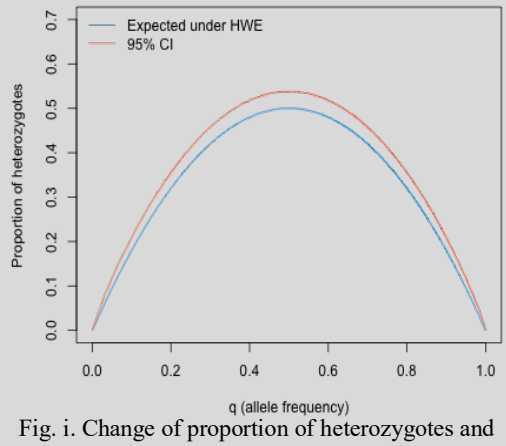


Fig. i. Change of proportion of heterozygotes and alternative homozygotes under Hardy-Weinberg equilibrium; 95% confidence interval is for 1000

Materials and Methods

DATA

In order to evaluate the performance of our approach and showcase the versatility of the *rCNV* package, we used previously published dataset from different sources. The datasets consist of SNPs called from short-reads data acquire through different sequencing technologies (e.g., RADSeq, dd-RADseq, Exome-capture and WGS) and of varying size in terms of number of samples and SNPs. They are as follows:

Chinook Salmon (Larson et al., 2014) – this dataset contains 19,299 SNPs for 266 individuals from five populations. The RADseq generated data was previously used in (G J McKinney et al., 2016; Garrett J. McKinney et al., 2017) for paralogs detection. This dataset is particular in the sense that it includes loci located on presumed ongoing residual tetrasomy, hence loci for which the number of copies is known. We used this annotation as a comparison to our CNVs detection.

American Lobster (Dorant et al., 2020) Dryad dataset (Waggitt, 2019) – this dataset includes 1,113 lobster samples collected from 21 locations in southern Gulf of St. Lawrence in Canada, and contain 44,374 SNPs (ddRADseq). After filtering and matching with SNPs used in Dorant et al. (2020) data,

the data set had 24,193 SNPs from 1,079 samples.
Norway spruce (a subset of Chen et al., 2019) – *Picea abies* dataset contains over 400,000 SNPs from 453 individuals belonging to 23 populations collected from European natural populations and Swedish breeding programs. The data was generated from exome-capture technology with over 40,000 probes.
Anopheles gambiae (The-Anopheles-gambiae-1000-Genomes-Consortium, 2017) - *Anopheles gambiae* data set is a subset of 80 samples from the phase 2 AR1 release (MalariaGen, 2017) consisting of whole genome sequencing data (illumina, paired-ends, short read) of 1,142 wild-caught mosquito specimens from 13 countries. Out of the 80 samples we selected, 40 carry an extensive tandem duplication of ~203kb (Assogba, Milesi et al. 2016). We thus used 108,295 SNPs from a 450Kb region centered on the duplication.

ANALYSIS

We followed all the steps of the workflow in Fig. 1 with all the data analyzed with *rCNV* (numbers 1-6 mentioned here in **bold** refer to the steps in the workflow Fig. 1). The raw (unfiltered) VCF files, obtained from the aforementioned databases, were imported to the R environment using the function *readVCF* in *rCNV* (**1**).

1. *Genotypes and variant filtering (not mandatory):*
   a. Missingness per sample and/or per SNP can be calculated and plotted with *get.miss* function to remove low coverage data; user specific thresholds can be passed on to the function.
   b. Individuals' pairwise relatedness can be calculated according to Yang et al., (2010) with the *relatedness* function. Individuals with a too high relatedness can, for instance, represent clones, technical replicates, siblings but also be sample DNA contamination during library preparation (> 0.9, Yang et al., 2010).
   c. Inbreeding coefficient (F_{IS}) can be computed using the function *h.zygosity* implementing the following method of moment:

$$F = \frac{(O-E)}{(N-E)} \qquad \text{.....eq. 3}$$

   where O is the observed percentage of heterozygous alleles, E is the expected from HWE and N is the total number of alleles. Extremely low F_{IS} values, often interpreted as a signal for outbreeding, can also indicate potential DNA contamination.

*NOTE: If the user has already done VCF filtering to their desired optimization, especially with above mentioned parameters, they can start the workflow at number **2**.*

2. *Allelic-ratio and metrics computations*
   a. Filtered VCFs are used to generate tables with depth of coverage values per SNP per allele (**3**). These tables are used as the input to compute various metrics our detection is based on. As the variant calls can contain mismatches between the allele depth (AD) and the genotype (GT) - for instance, the GT called is homozygote (0/0) but AD is non-null for both alleles (e.g. 40:1) - we correct the AD values according to the genotype with *ad.correct* function. This function can also be used to correct for the biases introduced by uneven total depth at a given locus in heterozygote individuals, which introduces additional deviation from expected allele ratio; this is done by adding one to the allele with the lowest depth value. Note that doing so, we are being precautious as we tend to even the allele ratio.
   b. Tables with normalized depth are then generated from the corrected allele depth table (**3**). The aim of normalization procedure is two-fold: i) checking for potential constitutive bias toward the reference or the alternative allele (i.e., a deviation from expected $p = 0.5$, see box 2) and ii) allowing the quantification of copy number in step **5.** We provide four normalization methods

Box 2: Test for deviation from expected allelic ratio and variance in expected allelic ratio distribution

Deviation from expected allelic ratio (Z-score):

$$Z_x = \sum_{i=1}^{n} Z_i \quad eq.1.0$$

$$Z_i = \frac{((N_i \times p) - N_{A_i})}{\sqrt{N_i \times p(1-p)}} \quad eq.1.1$$

N_i - total read-depth for heterozygote i at SNP x, N_A – alternative allele read-depth for heterozygote i at SNP x, p – probability of sampling allele A for SNP x (p = 0.5 if no bias), n – number of heterozygotes at SNP x.

With Z_i being *n* independent variables following normal distributions, $Z_i \sim N(\mu_i, \sigma_i^2)$, Z_x also follows a normal distribution $Z_x \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$. When $N_A \sim B(N, p)$, no deviation from expected allelic ratio is expected and $Z_i \sim N(0,1)$. The significance of the deviation from expected allelic-ratio can thus be tested by computing the quantile $q = Z_x$ of the probability density function given by $N(0, \sqrt{n})$.

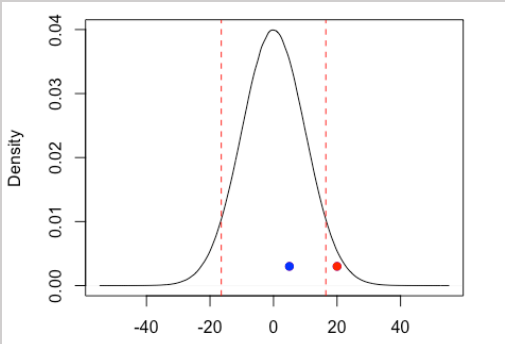


Fig.ii. Density distribution of expected *Z*-score. Blue point shows a non-deviant SNP falling inside the significant threshold corresponding to the quantile 0.025 and 0.975 of the probability density function of $N(0,100)$. Red point shows a deviant SNP as Z_x correspond to a quantile > 0.975.

Deviation from expected allelic ratio (Chi-square):

$$\chi_x^2 = \sum_{i=1}^{n} \chi_i^2 \quad eq.2.0$$

$$\chi_i^2 = \left(\frac{(N_i \times p - N_{A_i})^2}{N_i \times p} + \frac{(N_i \times (1-p) - (N_i - N_{A_i}))^2}{N_i \times (1-p)}\right) \quad \text{eq. 2.1}$$

N_i - total read-depth for heterozygote i at SNP x, N_A – alternative allele read-depth for heterozygote i at SNP x, p – probability of sampling allele A for SNP x (p = 0.5 if no bias), n – number of heterozygotes at SNP x.

With χ_i^2 being *n* independent variables following a chi-square distribution with one degree of freedom, χ_x^2 follows a chi-square distribution with *n* degree of freedom. When $N_A \sim B(N, p)$, no deviation from expected allelic ratio is expected and the significance of the deviation from expected allelic-ratio can thus be tested by computing the quantile $q = \chi_x^2$ of the probability density function given by $\chi_x^2 \sim X(n)$.

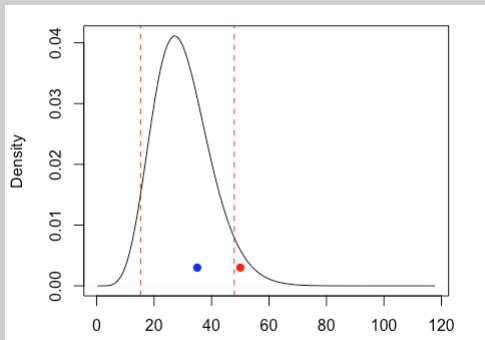


Fig.iii. Density distribution of expected X^2-score. Blue point shows a non-deviant SNP falling inside the significant threshold corresponding to the quantile 0.025 and 0.975 of the probability density function of $X(100)$. Red point shows a deviant SNP as χ_x^2 correspond to a quantile > 0.975.

in our package (Median ratio normalization – "MedR", Quantile normalization – "QN", Trimmed means of M-values – "TMM", and PCA based normalization – "pca") so that the user can use the optimal normalization method depending on the data and downstream analyses (*cpm.normal* function). Normalized depth values are important for determining the probe capture biases, correct allelic ratios, and downstream analyses as they minimize the biases introduced by sample-effect on the depth of coverage. See the recommendation section for more details on normalization methods. In addition to normalization of the depth values, *cpm.normal* function also flags SNPs and samples with extreme depth values and exports a list of outliers for further filtering.

c. Allele depth table (**3**) is used to calculate several statistics for each SNP independently, among which are, *Z*-score and χ^2-score both of which measure the strength of deviation from expected allelic ratio (box 2), the excess of heterozygotes, and the associated *p*-values (*allele.info* function) (**4**). Additional information generated are the proportion of heterozygotes per SNP, the median allelic ratio across all heterozygotes per SNP, and allele frequencies computed either from genotype or depth values.

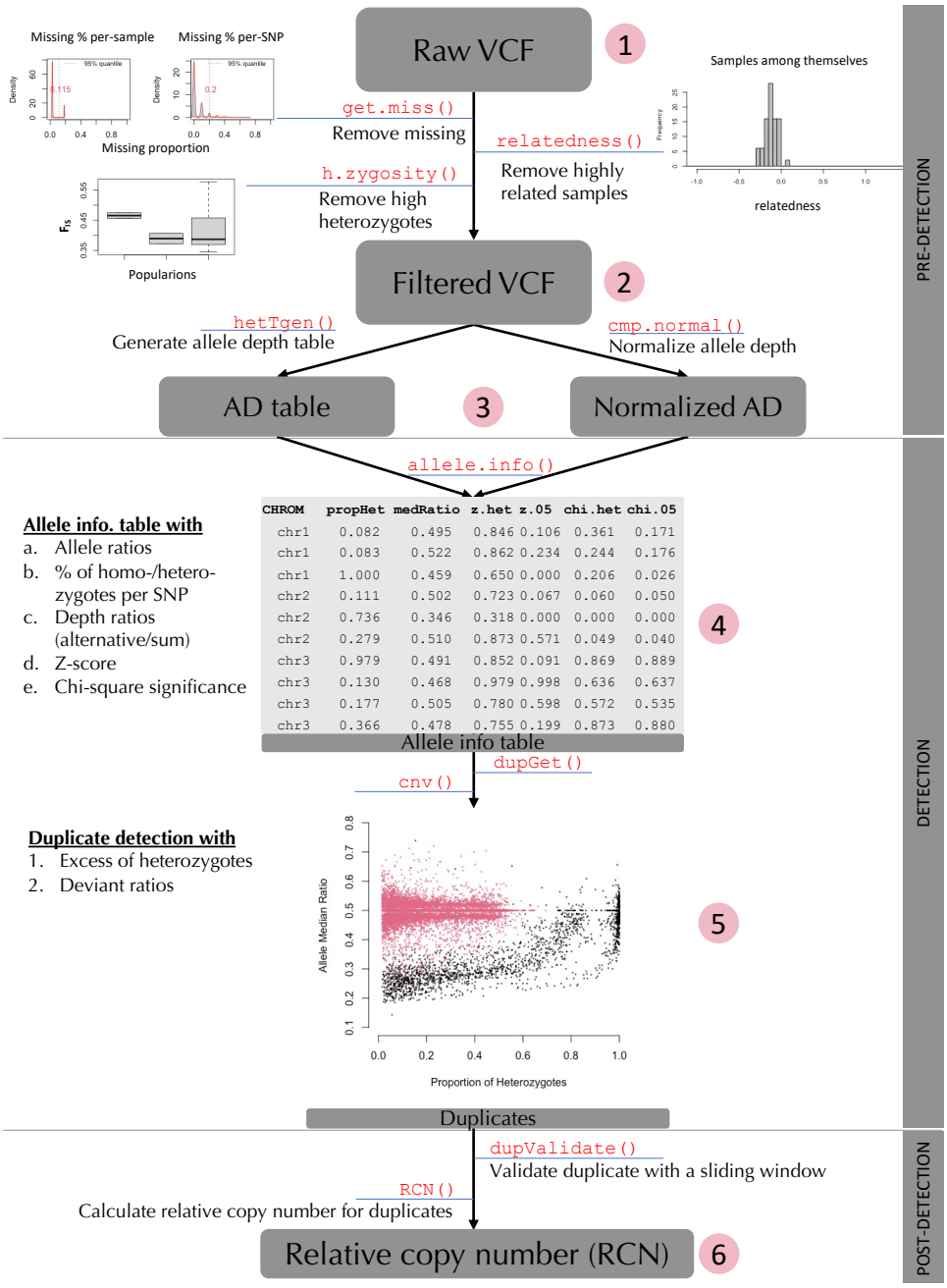


Figure 1. The compact workflow of rCNV, from importing raw data to detecting putative CNV loci and getting the relative copy number, included in the R package.

3. *Deviant SNPs and putative CNVs detection*
   a. In the last steps of detection stage (**5**), deviant SNPs are identified based on significance of the statistics mentioned above (i.e., Z-score, χ^2-score and excess of heterozygotes) with the in-built *dupGet* function. *dupGet* combines deviant ratios and excess of heterozygotes based on the expected allele ratio; by default, it is 0.5 but can vary between 0 and 1 and can also be provided by the user. Depending on the technology used for sequencing, some reference bias can exist leading to an over-representation of reads carrying the reference allele, which is often the case for probe-based target capture, for instance. Users are thus recommended to use the expected probe bias as expected allelic ratio. In addition, depending on the number of SNPs and the distribution of the *p*-values, user can choose different methods to control for false positive detection propensity associated with multiple testing, see recommendation section for more details.
   b. The candidate CNVs can then be identified using one of the two methods presented in the approach section "putative CNVs detection" using *cnv* function. See the recommendation section for more details on using different statistics.
   c. If the SNPs are ordered along a reference genome or a genetic map for instance, the user can adopt a density-based approach using a sliding window to further refine the list of putative CNVs with the *dupValidate* function. Briefly, if the deviation in expected allele ratio and other metrics is actually caused by the presence of multiple copies at a given locus (vs sequencing or technical bias), one would expect several close-by SNPs bearing a signal for CNVs.
   d. Finally, a table made of relative copy number is generated from the normalized allele depth table for the putative CNVs using the *RCN* function.

Downstream analyses: We have implemented the comparative statistic V_{ST} (variance fixation index – Redon et al. 2006), calculated from depth values of putative CNV loci to assess the genetic variation among populations. V_{ST} is analogous to fixation index (F_{ST}) in population genetic analyses, and calculated as $V_{ST} = V_S/V_T$ where V_T is the variance of normalized read depths among all individuals from two populations and V_S is the average of the variance within each population, weighed for population size (Dennis et al., 2017; Weir & Cockerham, 1984). We further provide the option to plot a network graph to visualize the among-sample association based on the V_{ST} distance matrix. The high-confident putative CNVs filtered through post-detection are then used to calculate the relative copy number (RCN) across samples per locus (see prospects) and can be used in further downstream analyses such as genome-wise association analysis (GWAS) and environmental association analysis (EAA).

Results

All the datasets analyzed here were filtered (initial variant filtering) for low-heterozygote SNPs (<4 heterozygotes per SNP) and outlier samples and SNPs where extreme depth values are recorded. All the removed SNPs had less than 10% heterozygotes in all the datasets. Such low levels of heterozygosity are uninformative and can negatively affect the CNVs detection (see Approach), and therefore were removed from our analyses. The SNPs with a median read depth of < 5 were also not considered for CNV classification and were removed from all the analyses (classified as low confidence in simulations: Supplementary Fig. 1).

Chinook salmon:

Out of 19,299 SNPs, only 18,309 were kept for the analysis after initial variant filtering. Out of these SNPs, 14.8% were identified as *deviant* SNPs, of which 11.3% were recognized as putative CNVs. Because of the recent tetrasomy, an allelic ratio of 1:3 is expected for SNPs located in the tetratomic regions. This is what we observed, with a clear definition in the allele ratio plot (Fig. 2) between the

SNPs located in multi-copy regions (red) and those that are in single copy regions (blue). Out of all the common SNPs (8,220) analyzed in both our analysis and McKinney et al. (2016), the latter identified 1,223 (14.87 %) SNPs as duplicates while *rCNV* detected 1,117 (13.58 %) as putative CNVs. Nevertheless, from 931 SNPs classified as putatively duplicated/CNVs by either analysis, about 83% are common to both (Table 1 & Supplementary Fig. 2A). We labeled putative CNVs as pseudo false positives (*pFP*) or pseudo false negatives (*pFN*) where the classification by either analysis (i.e., McKinney et al. and *rCNV*) does not match the other (Supplementary Figs 2B&C). The *pFP* in rCNV classification (i.e., classified as putative CNVs by rCNV but classified as non-duplicated by McKinney et al.) have allelic ratios expected from a putatively duplicated SNP (i.e., 3:1 or 1:3) while *pFN* (i.e., classified as duplicated in McKinney et al. analysis but classified as non-duplicated by rCNV) are in the low confidence zone (i.e., where either the number of heterozygotes are low e.g., 5 or less per SNP, or the median depth values are on the low side for confident detection (e.g., <6), as shown in simulations: Supplementary Fig. 1).

American lobster:

After filtering, *rCNV* only kept 24,047 SNPs for the duplicate detection. In the American lobster dataset, we flagged 7,125 (29.6%) SNPs as *deviants*, of which we detected 3,596 SNPs (14.95%) as putative CNVs, whereas Dorant et al. (2020) identified 39.92% (9,659 out of 24,193) to be located in duplicated regions (Table 1 and Supplementary Fig. 3). Out of all SNPs detected as putatively duplicated by either analysis (8,162 SNPs), 41% were detected in both analyses, but only 64 SNPs were detected as located in putatively CNV regions with *rCNV* that were not detected in Dorant et al. analysis, here again we demonstrate a higher stringency in our detection approach; whereas their study detected additional 4,593 alleles that we did not detect. We labeled these as pseudo false positives (*pFP*: detected as putative CNVs by rCNV but classified as non-duplicated by Dorant et al.) and pseudo false negatives (*pFN*: classified as duplicated by Dorant et al. and non-CNVs by rCNV): *pFP* – 64, pFN – 4,593. To visualize this overlap (and the lack thereof), we plotted *pFP* and *pFN* on allele ratio plots (Supplementary Fig. 4). Almost all the *pFN* detection (purple points in the plot) by *rCNV* are either in the low-heterozygosity zone (< 10%) or have an allelic ratio close to 1:1 (blue cloud), and they were not detected as deviants in our analysis with any of the three statistics tested, likely indicating high false positive detection rates in Dorant et al. (2020).

Norway spruce

After initial filtering, 245,758 SNPs were kept for the analysis. Out of which, 10.51 % (25,826 SNPs) were flagged as *deviants*. Among them, 4,201 SNPs (1.71%) were located in putatively CNV regions (Supplementary Fig. 5). Comparing mean read coverage between homozygotes for the reference or the alternative allele clearly indicated the presence of sequencing probe-bias towards the reference allele (Supplementary Fig. 6). Therefore, at a given SNP, we used the average allelic ratio computed across all samples as expected allelic ratio (instead of $p = 0.5$ in eq. 1.1 and 2.1 in Box. 2) for computing both Z-score and χ^2-score (*z.all* and *chi.all* in the functions *dupGet* and *cnv*). This resulted in a more accurate classification of SNPs than using $p = 0.5$ (*z.05*, *chi.05*) where additional 1,196 SNPs were classified as putative CNVs. Upon visualization of this additional SNPs, we observed that they had allelic ratios expected for SNPs located in non-CNV regions (1:1). This clearly demonstrates that probe-biases can cause false classification of SNPs and that rCNV framework is efficient at minimizing the false classification arising from probe-biases in exome-capture data.

Table 1. Copy number variation detection by rCNV for different datasets with their respective sequencing technology, size (i.e., samples & SNPs), deviant SNPs, putative CNVs, and overlap with previous studies.

Species	Technology	# Samples	# SNP	Deviants #	Deviants %	CNVs #	CNVs %	Overlap† (%)
Chinook Salmon	RAD-Seq	226	18,309	2,717	14.8	2,070	11.3	83
American Lobster	ddRAD-Seq	1,074	24,047	7,125	29.6	3,596	14.95	41
Norway Spruce	Exome capture	423	245,758	16,064	6.5	5,397	2.19	-
Anopheles gambiae	WGS	80	16,471	759	4.6	565	3.4	-

#: number; †: overlap compared to respective previous study

Anopheles gambiae

The target region (400 kb) of chromosome arm 2 (2R) with the known duplicated region consisted of ~145,000 SNPs after filtering out multi-allelic loci. Out of this, only 16,471 SNPs were kept for the analysis after filtering for low number of heterozygotes and low depth of coverage. We found that 565 SNPs bear a signal for being located in the putatively duplicated region (Supplementary Fig. 7) among the759 that were classified as deviant SNPs. While most (68%) of the SNPs detected as putative CNVs are actually in the region *AgamP4_2R:3436927-3639836*, which was previously described as a duplicated chromosome fragment (Assogba et al. 2016), we also found SNPs that are putative CNVs in the flanking the regions of *AgamP4_2R:3436927-3639836* (Fig. 3A); this observation is not surprising as the flaking regions are rich in transposable elements (Assogba milesi et a. 2016). However, out of all the SNPs in the putatively duplicated region (excluding the SNPs in the flanking regions), 84% were detected as putative CNVs in our analysis.

For all datasets, we observed higher density distribution of average read depth coverage in SNPs detected in putatively CNV regions (Fig. 4). Since we did not use average read depth coverage in any metrics, this is further in favor of our classification of CNVs as SNPs with higher copy numbers should carry higher read depth coverage than in non-CNV alleles. Nevertheless, different levels of overlap observed in the density distribution in Fig. 4 are arising from two major factors: 1) SNPs misclassified as putative CNVs due to low number of heterozygotes and/or SNPs misclassified as non-CNV due to absence of deviation in actual CNVs where all the individuals show homozygosity for that locus, and 2) CNVs are by definition polymorphic in the population and all samples do not carry the same number of copies of a given locus; this is evident in the *Anopheles* mosuqito dataset. Furthermore, the local enrichment of putative CNVs in specific genomic regions where the average depth coverage is higher (i.e., peaks) for the Norway spruce and mosquito dataset (Fig. 3B) further support a low false positive detection rate and corroborate the variation in the density distribution of the average dead depth coverage in putative CNVs.

Discussion

Efficient and more accurate detection of copy number variants from genomic data can immensely improve genetic analyses by i) generating a clean dataset free of SNPs that are in multiple copy regions and low confident SNP calls due to technical issues, and ii) detecting putative CNVs and their relative copy numbers that can be used in quantitative and population genomic studies. Generating such datasets with minimal errors is almost impossible without whole genome sequencing data and well annotated reference genomes. Yet, the vast majority of genetic data consists of SNPs datasets generated through genotyping by sequencing or target sequencing methods. With our framework, we provide a robust filtering method for SNPs showing unexpected pattern in allelic ratio (because of paralogous copy or sequencing bias) and a robust CNVs candidate detection pipeline, both from single nucleotide polymorphism data. Our framework is user-friendly, easily accessible and low demanding of programing skills as it is implemented within the R infrastructure. The approach implemented in the

rCNV R package is statistically sound, and the results tested in our case studies show substantial accuracy, especially with high-throughput non-whole genome sequencing data.

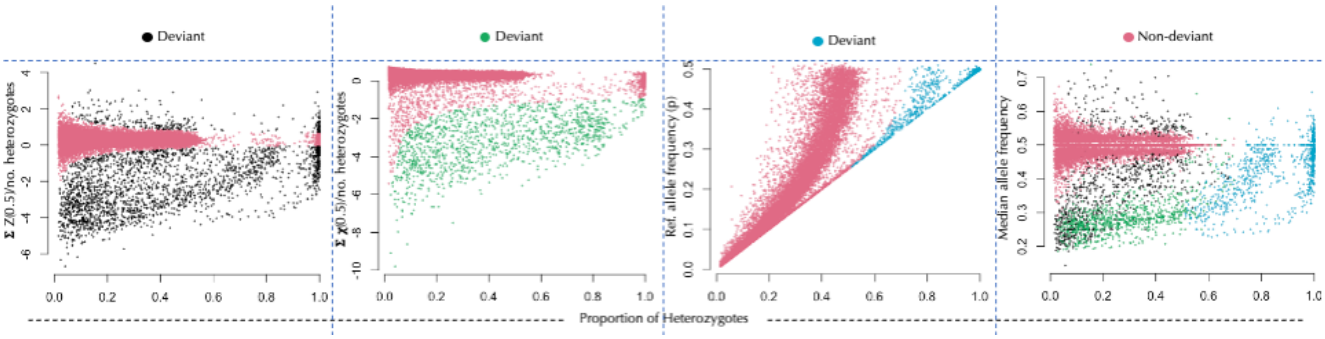


Figure 2. Different ranges of deviant SNP detection by (plots left to right) Z-score distribution of depth values, chi-square distribution of depth values, excess of heterozygotes, and the combination of all statistical methods on Chinook Salmon data (Larson et al. 2016).

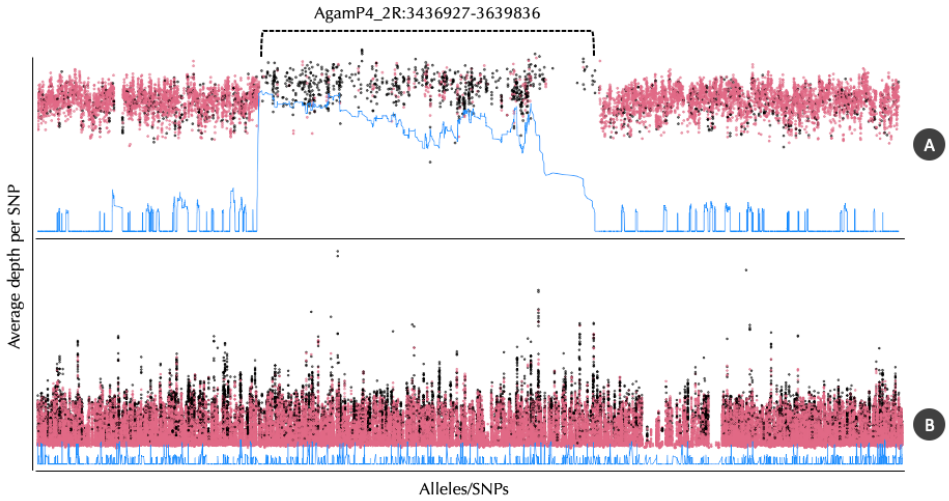


Figure 3. The average read coverage depth variation of alleles sequentially arrange on the genomes of A) *Anopheles gambiae* and B) Norway Spruce. Black points indicate the detected putative CNVs by rCNV; the blue line shows the moving window average of the depth values. AgamP4_2R:3436927-3639836 is the previously recognized region od duplication in the right chromosome arm of *Anopheles gambiae*.

Thresholds and false positive detection propensity

One of the biggest challenges in CNVs detection using SNPs data is the propensity for false detection due to, among other factors, sequencing and probe-biases (see example in chapter 4 of Leandro 2013), as well as large variation in read depth and number of heterozygous samples, that can introduce deviation from expected patterns. For a given SNP, what should be considered as a significant deviation from the expected allelic ratio would depend on both the number of heterozygotes and the depth of coverage across the heterozygotes. We thus recommend eliminating SNPs with low number of heterozygotes (< 4) and low total read depth (< 5) (the default parameters in *rCNV*) in CNVs detection. With the function *depthVsSample* in the *rCNV* package, it is possible to visualize how the confidence intervals around a given expected allele ratio vary with varying number of heterozygotes and depth of coverage values (Supplementary Fig. 1). In our R package, however, the users are given the choice to adjust these parameters to best fit their data.

Analogous methods to rCNV framework, i.e., *HD* plots: McKinney et al. (2017) and Dorant et al. (2020), have used a single threshold value defined through visualization of plots of median allele ratio in heterozygotes (medRatio), proportion of rare homozygotes (PropHomRare), and inbreeding

coefficient (F_{IS}). While this approach can serve as a preliminary assessment, it completely overlooks the variations in number of heterozygotes and depth of coverage from one SNP to another, which could lead to a high false positive detection rate. This can be clearly observed in the false positive detection of CNVs for a proportion of heterozygote below 0.2 in McKinney et al. (2016, 2017) and Dorant et al. (2020) (Supplementary Figs. 2B-C & 4 respectively).

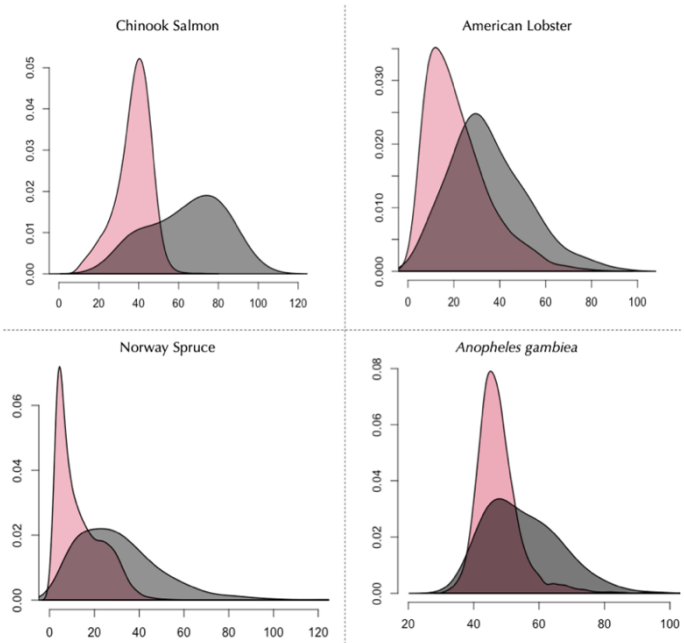


Figure 4. The density distribution of average read coverage depth of putative CNVs and non-CNVs in the studied SNP datasets (red: non-CNVs, black: putative CNVs).

In our framework we have addressed these shortcomings by computing a Z-score and a χ^2-score allowing to test the significance of the deviation from the expected allele ratio for each SNP independently. It allows taking into account the SNPs-specific parameters (i.e., number of heterozygotes and depth of coverage) thereby minimizing false positive detection arising from low sampling size, for instance. Still, the users have the flexibility of adjusting the thresholds for the *p.value* to better suit their data as well as the stringency for false positive detection rate. Note also that levels of stringency for the "*deviant*" SNPs detection may affect the detection of the SNPs putatively located in CNV regions using the *intersection-set* approach. Therefore, we also provide the user with a threshold-free approach using *K*-means clustering. We obtain highly overlapping putative CNVs using both methods, specifically for RAD-seq data, where sequencing bias is minimum compared to probe-based methods.

Deviants and putative CNVs regions

It is a fair evaluation that the methods used in our framework may not result in the same accuracy for all data sets. In fact, the accuracy is heavily influence by the quality of the data as the reliability of SNPs calling and read depth analysis are influenced by the quality of the raw sequences (Barbitoff et al., 2020; Borges, Rocha, Carvalho, & Lopes-Cendes, 2020; Gabrielaite et al., 2021). With real-world data, we show in our detection of *deviants* and putative CNVs, that our framework is efficient at encapsulating both undesired loci (deviants) to be removed in SNPs-based genetic analyses and important structural variants (CNVs) for downstream analyses. Further, depending on the quality of the data (e.g., probe-bias) and the purpose of downstream analysis (e.g., extracting CNVs or excluding possible paralogs), stricter filtering criteria in the deviant detection will eliminate false positives from the putative duplicates. We have high confidence that the appropriate use of filtering criteria (i.e., *test* and *filter* argument of the *cnv* function in the rCNV package) will filter out more than 90% of the false positive detection of duplicates from deviants. This means that the list of deviants is the maximum number of putative duplicates one can detect from a provided dataset.

Different sequencing technologies

While DNA sequencing technologies are rapidly improving with better quality and more accurate generation of sequences, most non-whole genome sequencing techniques carry certain degrees of error and anomalies (Jones & Good, 2016; Krumm et al., 2012; Meynert, Ansari, FitzPatrick, & Taylor, 2014). Studies show that exome-capture technologies are approximately 95% accurate at detecting the alternative allele at a mean depth of 40 reads compare to WGS (Lelieveld, Spielmann, Mundlos, Veltman, & Gilissen, 2015; Meynert et al., 2014) thus contributing to *probe-bias* towards the reference allele. Generally, we refer to as *probe-bias* when the (exome-capture)-PCR probes have a higher tendency to capture reads from one allele of the same locus than the other alleles. Among many factors, the GC content, DNA concentration, and annealing temperature are the most commonly known causes of such biases (Barbitoff et al., 2020; Borges et al., 2020; Neves, 2013). While these differences can be dramatic, most importantly, the captured reads for two regions and for two alleles in a heterozygote can be significantly different with probe-bias, leading to significant deviation from theoretical allele ratio of 0.5, and therefore, can add to false detection of CNVs. We show this tendency with the Norway spruce exome capture dataset. Although we have addressed this issue in our framework by assessing the observed allele frequency (p) across all SNPs, our *recommendation* is that, if a dataset is more prone to probe bias and there is prior knowledge of the bias ratio, the user input that information when detecting deviants and CNVs (e.g., *test=z.all/chi.all* option in *dupGet* and *cnv* functions). Nevertheless, one should keep in mind that in cases where poor sequencing and high probe-bias are prevalent, the detection of CNVs may be less reliable as seen in the simulations (see Supplementary Fig. 8) when the number of heterozygotes and the median allele depth are significantly low. rCNV has been developed to use variant calls from any sequencing technology to detect CNVs. However, unlike other CNV detection methods, we have incorporated model training to test for probe bias and adjust the duplicate detection accordingly in rCNV.

Prospects

Further annotation of CNVs

When the SNPs are ordered along a reference, we provide the user with the possibility to use a density-based sliding window approach to further refine the list of SNPs putatively located in CNV regions. Ideally, if a locus is truly located in a multi-copy region, close-by SNPs should also be classified as *deviants* while deviation caused by sequencing error should be more scattered along the genome. The function *dup.validate* in the rCNV package is dedicated for detecting regions enriched for deviant SNPs within a sliding window along a chromosome, scaffold or a sequence of any given length. User can define both, the length of the window and the threshold above which a locus is considered a *true-deviant*. We demonstrate this with Norway spruce and *Anopheles* mosquito datasets where deviant SNPs with a frequency threshold less than 0.2 within a moving window of 10,000 are flagged as low confident. This step of filtering is based on the assumption that at least 20% of the loci within a duplicated region must be detected as deviants/CNVs for the loci in the respective region to be considered a deviant/CNV of high confidence.

Relative Copy Number (RCN) and Environmental Association Analysis (CNV-EAA)

The depth values of putative CNVs can be used for population genetic analyses such as V_{ST} or multivariate and ordination analyses such as PCA to examine the population structure and local adaptation using CNV (e.g., Fig. 3&4, Supplementary Fig. 8). We show this application with the American lobster data; to compare our output with Dorant et al. results we performed two downstream analyses; the ordination of putative CNVs from the lobster data (Supplementary Fig. 8) showed a strong spatial structure among individuals that is not explained by the geographical distribution but rather by

the distribution of Sea Surface Temperature (SST) as was observed by (Dorant et al., 2020). Further, variance fixation index (V_{ST}) analysis using the normalized depth values of detected putative CNVs showed a correlation to F_{ST} calculated with the genotypes of the non-CNVs (putatively non-duplicated) in the analyzed data set ($r^2 = 0.82$), yet the population structure obtained from CNVs was much stronger than that obtained from non-CNVs. However, categorizing the samples or populations into relative copy number variants (RCN) will allow capturing a different set of genetic variation that is more meaningful for population structure and adaptation analyses. Further RCN can be incorporated into environmental association analysis (EAA) since the relative copy number can capture quantifiable genetic variation to adaptation (e.g., North et al., 2020; Prunier et al., 2019). However, correctly deciphering the RCN from the depth values is challenging, especially when the depth values vary continuously among samples. Clustering methods (e.g., Fraley & Raftery, 2002), breakpoint analysis (e.g., Chong et al., 2017; Faust & Hall, 2012), and Bayesian criterion (e.g., Fraley & Raftery, 2007) are potential candidates for categorizing such data. These methods however do not determine the relative copy number in accordance with the depth increment. For instance, clustering methods may assign an RCN value of one (RCN=1) for a depth value of 50 and two (RCN=2) for a value of 150. However, the correct relative copy number for such an increment should be three (RCN=3) as the second sample has three times the read depth value as the first one. Therefore, the development of a statistical method to correctly categorize the relative copy number from the depth values is underway for rCNV.

One major use of CNV from large-scale population genetic analyses is to detect spatial genetic variation and local adaptation. Environmental association analysis (EAA) with currently available methods (see Manel & Holderegger, 2013) relies on the allele frequency calculated from genotypes to assess the genetic variation corresponding to the environmental variables. However, allele ratios calculated with depth values in duplicated loci are not representative of the actual copy number variation among individuals as two different copy numbers can produce the same allele frequency. An alternative method used in previous studies (e.g., Cayuela et al., 2021; Dorant et al., 2020) is to use the composite values of a PCA performed on the normalized total depth values of the duplicates in a multivariate analysis with the environmental variables. Nevertheless, use of normalized depth values undermines the relative variation among populations and individuals. Therefore, the relative copy number (RCN) will provide a comparable analysis with local adaptation.

Benchmarking rCNV

In order to compare the efficiency of detecting CNVs from SNPs data using rCNV package, we explored the scientific literature and the major bioinformatics platforms to find comparable methods and software. Despite the availability of a range of methods and software for detecting CNVs from whole genome sequences (see Gabrielaite et al., 2021), we found only two methods comparable (see below) to our framework. While there is an array of methods developed for detecting CNVs from short-read sequencing such as read-pair split-read (SR), read-pair (RP), and assembly (AS), algorithms based on read coverage differences in the genome, also referred to as read-depth (RD) is the most common method of CNV detection in most software packages available (Gabrielaite et al., 2021; Kosugi et al., 2019). However, most if not all, such algorithms rely on complete and precise reference genomes (e.g., GATK gCNV, ExomeDepth) or require whole exome capture or whole genome sequences (e.g., CNVnator, LUMPY). Further, the raw data used in each software package is specific to their own method (e.g., STACK) or require a great deal of computation power and memory (e.g., CLC Genomics Benchmark – fastq; cnMOPS – bam files) (see Gabrielaite et al., 2021) for a detailed comparison). In contrast, most available methods are dedicated for exhaustive exploration of structural variants including CNVs from WGS data whereas our framework is ideal when such methods are not available, for instance for studies involving non-model organisms. Finally, most of the available software/methodologies require at least intermediate level of programing skills. Therefore, in essence,

our framework and rCNV package stand out from these methods in three main aspects: 1) use of VCF files with SNP calls, which are relatively small in size compared to other input data formats, 2) not requiring whole genome sequences or precise reference genomes, and 3) the detection is statistically robust and does not demand expert programming skills. Overall, the framework implemented in rCNV is reliable, cost effective, easy to use and significantly fast.
In addition to the robust and easy detection of CNVs, rCNV offers effective pre-processing of the data (e.g., filtering of SNPs based on missingness, heterozygosity, relatedness, outliers) as well as basic post detection downstream analyses such as VST calculation, RCN, and sliding window-based CNV validation. Therefore, we only compare the similarities of our framework others below.

Comparison with HDplot method (McKinney et al. 2017):
The method proposed by McKinney et al. (2017) aims at identifying SNPs in which the median allelic-ratio (ratio between the depth of coverage of two alleles at a given position) in heterozygote individuals deviates from 1:1 (expected ratio in heterozygote when both alleles have equal affinity to sequencing probes). This method was developed to filter out SNPs located in paralogous regions, and shows limitations when used to call CNVs. First, it is a threshold-based approach reliant on the expectation that allelic ratio can be drawn from a pre-determined number of copies. For CNVs however, the allelic ratio variation is more continuous as it depends on the number of copies in a given individual and their frequencies across all individuals; two values that can vary from one locus to another. Second, despite being effective as a priori, the use of median and a threshold is sensitive to sampling effect (variation of depth and number of heterozygotes among SNPs) that can lead to a high false-positive detection rate of CNVs, especially when removing putative paralogous. Third, a striking challenge in implementing the McKinney et al. method is the specificity of the computer pipeline to RADseq generated data (see also McKinney et al. methods and its development in Dorant et al. 2020). Applying their methodology on other data types (e.g., exome-capture, WGS) requires advanced programing skills and therefore the accessibility is limited to those who are familiar with the pipeline. Finally, sequencing technologies are prone to errors, especially, widely used target capture technologies such as exome-capture and SPET where capture-biases can occur toward the reference allele. In such cases, the expected ratio between the reference and alternative alleles is no longer 1:1 and should instead reflect the capture-bias. *rCNV* has addressed such shortcoming by statistically scrutinizing the probe-bias and implementing methods to correct for them while accounting for sampling effect inherent to depth variation and number of heterozygotes between SNPs.

Comparison with STACK workflow method in Dorant et al. (2020)
We compared the STACK workflow (https://github.com/enormandeau/stacks_workflow) used by Dorant et al. (2020) for a comparison of a similar detection method to *rCNV*. The method implemented in STACK is the same *HD*-plot approach developed by McKinney et al. 2017 and thus suffers the same drawbacks as mentioned above. However, to test the efficiency of the pipeline, we parallelly followed their workflow steps with our methodology (as in Fig. 1) on the American Lobster data set. From filtering raw VCF files to the detection of putative duplicates, Dorant et al. (2020) used three python scripts and two R scripts, which all together accounted for a total of 12 minutes 23 seconds for the complete analysis. However, this time is without the time spent on visual observation of the various allelic ratio-based statistic plots they used to determine the thresholds for sorting duplicates from "singletons". In contrast, *rCNV* package only took 45.82 seconds to complete the detection process with statistic-based setting of thresholds. Therefore, compared to the method used in the STACK workflow, *rCNV* is not only computationally faster, but also detects the duplicates based on a robust statistical approach with minimal intervention from the user. Further, several python scripts used in STACK

workflow are specific to VCF files generated from STACK software package for calling SNPs and usable with only RADseq data.

Other R packages

We found only one more R software package that implements a similar approach to detecting CNVs as rCNV. The R package "radiator" (Gosselin, 2020) is dedicated for exploration, manipulation, visualization, imputation, and exportation of RADseq and GBS data. The current version (v.1.2.2) is only hosted on GitHub (https://github.com/thierrygosselin/radiator/) and indicated to be "maturing." The function *detect.paralog* in "radiator" has been developed to detect duplicated loci from SNP data using the *HDplot* method described by McKinney et al. (2017). However, we were not able to test the "radiator" package or the *detect.paralog* as the examples provided in the function itself does not work and the author of the package has not provided any other alternative to test the function as it only accepts data formats internal to the "radiator" package. Our attempts to contact the author were not successful either. Overall, the "radiator" package is not usable for detecting CNVs in the current state.

Acknowledgments

We thank Swedish National Infrastructure for Computing (SNIC) for resource allocation for high-power computing under the project numbers SNIC 2021/6-313 and SNIC 2021/5-540. The Nilsson-Ehle grant by the Royal Physiographic Society of Lund awarded to Piyal Karunarathne (42191) was immensely helpful for acquiring the resources to test the features of the R package locally. We wish to thank Freja Lindstedt for her generous work in further testing of the CNVs detection on real-world data.

Data Accessibility

Data Accessibility

All the genetic data used in the paper are mentioned in the text with their references where applicable. All the scripts and simulation outputs can be accessed from the GitHub repository (link)

Author Contributions

The framework design was conceived by PM; the methods and statistical approached developed by PM, PK, and QZ; PK developed the software package and analyzed the data. KS improved and tested the functions and PM and QZ beta-tested the R software package. PK wrote the manuscript with input from all the authors.

Reference

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., … Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports*, *10*(1), 1–13. doi: 10.1038/s41598-020-59026-y

Borges, M. G., Rocha, C. S., Carvalho, B. S., & Lopes-Cendes, I. (2020). Methodological differences can affect sequencing depth with a possible impact on the accuracy of genetic diagnosis. *Genetics and Molecular Biology*, *43*(2), 1–6. doi: 10.1590/1678-4685-GMB-2019-0270

Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost Chrysophrys auratus. *Molecular Ecology*, *28*(6), 1210–1223. doi: https://doi.org/10.1111/mec.15051

Cayuela, H., Dorant, Y., Forester, B. R., Jeffries, D. L., Mccaffery, R. M., Eby, L. A., … Chris Funk, W. (2021). Genomic signatures of thermal adaptation are associated with clinal shifts of life history in a broadly distributed frog. *Journal of Animal Ecology*, (May), 1–17. doi: 10.1111/1365-2656.13545

Chain, F. J. J., & Feulner, P. G. D. (2014). Ecological and evolutionary implications of genomic structural variations. *Frontiers in Genetics*, *5*, 125–138. doi: 10.1038/nrg3373

Chen, J., Li, L., Milesi, P., Jansson, G., Berlin, M., Karlsson, B., … Lascoux, M. (2019). Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce.

Evolutionary Applications, *12*(8), 1539–1551. doi: 10.1111/eva.12801
Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., … Chen, J. (2017). novoBreak: local assembly for breakpoint detection in cancer genomes. *Nature Methods*, *14*(1), 65–67.
Clop, A., Vidal, O., & Amills, M. (2012). Copy number variation in the genomes of domestic animals. *Animal Genetics*, *43*(5), 503–517. doi: 10.1111/j.1365-2052.2012.02317.x
Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., … Consortium, G. A. D. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. doi: 10.1038/s41586-020-2287-8
DeBolt, S. (2010). Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. *Genome Biology and Evolution*, *2*, 441–453. doi: 10.1093/gbe/evq033
Dennis, M. Y., Harshman, L., Nelson, B. J., Penn, O., Cantsilieris, S., Huddleston, J., … Eichler, E. E. (2017). The evolution and population diversity of human-specific segmental duplications. *Nature Ecology and Evolution*, *1*(3), 1–10. doi: 10.1038/s41559-016-0069
Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., … Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Molecular Ecology*, *29*(24), 4765–4782. doi: 10.1111/mec.15565
Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O., & Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science*, *320*(5883), 1629–1631. doi: 10.1126/science.1158078
Fan, S., & Meyer, A. (2014). Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes . *Frontiers in Genetics* , Vol. 5. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2014.00163
Faust, G. G., & Hall, I. M. (2012). YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, *28*(19), 2417–2424. doi: 10.1093/bioinformatics/bts456
Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. doi: 10.1198/016214502760047131
Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*(2), 155–181.
Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., … Bagger, F. O. (2021). A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*, *13*(24), 1–21. doi: 10.3390/cancers13246283
Ghosh, M., Sharma, N., Singh, A. K., Gera, M., Pulicherla, K. K., & Jeong, D. K. (2018). Transformation of animal genomics by next-generation sequencing technologies: a decade of challenges and their impact on genetic architecture. *Critical Reviews in Biotechnology*, *38*(8), 1157–1175. doi: 10.1080/07388551.2018.1451819
Gosselin, T. (2020). *radiator: RADseq Data Exploration, Manipulation and Visualization using R*. Retrieved from https://thierrygosselin.github.io/radiator/. doi : 10.5281/zenodo.3687060
Holland, P. W. H., Marlétaz, F., Maeso, I., Dunwell, T. L., & Paps, J. (2017). New genes from old: Asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1713). doi: 10.1098/rstb.2015.0480
Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. doi: https://doi.org/10.1111/mec.13304
Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 5048–5057.
Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 1–18.
Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., … Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research* , *22*(8), 1525–1532. Retrieved from http://genome.cshlp.org/content/22/8/1525.abstract
Larson, W. A., Seeb, L. W., Everett, M. V, Waples, R. K., Templin, W. D., & Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). *Evolutionary Applications*, *7*(3), 355–369.
Lauer, S., & Gresham, D. (2019). An evolving view of copy number variants. *Current Genetics*, *65*(6), 1287–1295. doi: 10.1007/s00294-019-00980-0
Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., & Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, *36*(8), 815–822. doi: https://doi.org/10.1002/humu.22813
Limborg, M. T., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2016). Screening of duplicated loci reveals hidden divergence patterns in a complex salmonid genome. *Molecular Ecology*, *26*(17), 4509–4522. doi: 10.1111/mec.14201

Lucek, K., Gompert, Z., & Nosil, P. (2019). The role of structural genomic variants in population differentiation and ecotype formation in Timema cristinae walking sticks. *Molecular Ecology*, *28*(6), 1224–1237. doi: https://doi.org/10.1111/mec.15016

MalariaGen. (2017). MaleriaGen. Retrieved from https://www.malariagen.net/data/ag1000g-phase-2-ar1

Manel, S., & Holderegger, R. (2013). Ten years of landscape genetics. *Trends in Ecology and Evolution*, *28*(10), 614–621. doi: 10.1016/j.tree.2013.05.012

Maza, E., Frasse, P., Senin, P., Bouzayen, M., & Zouine, M. (2013). Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments. *Communicative & Integrative Biology*, *6*(6), e25849. doi: 10.4161/cib.25849

McKinney, G J, Seeb, L. W., Larson, W. A., Gomez-Uchida, D., Limborg, M. T., Brieuc, M. S. O., … Seeb, J. E. (2016). An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (Oncorhynchus tshawytscha). *Molecular Ecology Resources*, *16*(3), 769–783.

McKinney, Garrett J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. doi: 10.1111/1755-0998.12613

Meynert, A. M., Ansari, M., FitzPatrick, D. R., & Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, *15*(1), 247. doi: 10.1186/1471-2105-15-247

Nadukkalam Ravindran, P., Bentzen, P., Bradbury, I. R., & Beiko, R. G. (2018). PMERGE: Computational filtering of paralogous sequences from RAD-seq data. *Ecology and Evolution*, *8*(14), 7002–7013. doi: https://doi.org/10.1002/ece3.4219

Nelson, T. C., Monnahan, P. J., McIntosh, M. K., Anderson, K., MacArthur-Waltz, E., Finseth, F. R., … Fishman, L. (2019). Extreme copy number variation at a tRNA ligase gene affecting phenology and fitness in yellow monkeyflowers. *Molecular Ecology*, *28*(6), 1460–1475. doi: https://doi.org/10.1111/mec.14904

Neves, L. G. (2013). *Exome sequencing for high-throughput genomic analysis of trees*. University of Florida.

North, H. L., Caminade, P., Severac, D., Belkhir, K., & Smadja, C. M. (2020). The role of copy-number variation in the reinforcement of sexual isolation between the two European subspecies of the house mouse: CNV and reinforcement in the house mouse. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806). doi: 10.1098/rstb.2019.0540

Ohno, S. (1970). *Evolution by gene duplication*. New York, USA: Springer Science & Business Media.

Prunier, J., Giguère, I., Ryan, N., Guy, R., Soolanayakanahally, R., Isabel, N., … Porth, I. (2019). Gene copy number variations involved in balsam poplar (Populus balsamifera L.) adaptive variations. *Molecular Ecology*, *28*(6), 1476–1490. doi: https://doi.org/10.1111/mec.14836

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Chen, W. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), 1–9.

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, *14*(11), 807–820. doi: 10.1038/nrg3522

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., … Persengiev, S. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, *23*(9), 1373–1382.

The-Anopheles-gambiae-1000-Genomes-Consortium. (2017). *Ag1000G phase 2 AR1 data release. MalariaGEN*.

Verdu, C. F., Guichoux, E., Quevauvillers, S., De Thier, O., Laizet, Y., Delcamp, A., … Mariette, S. (2016). Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in Robinia pseudoacacia L. *Ecology and Evolution*, *6*(20), 7323–7333. doi: https://doi.org/10.1002/ece3.2466

Waggitt, J. (2019). Data from: Distribution maps of cetacean and seabird populations in the North-East Atlantic. Retrieved from https://doi.org/10.5061/dryad.mw6m905sz

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Ererbungslehre aus Betrachtet. *Hereditas*, *11*(1), 65–106. doi: https://doi.org/10.1111/j.1601-5223.1928.tb02483.x

Weir, B. S., & Cockerham, C. C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution*, *38*(6), 1358–1370. doi: https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Wellenreuther, M., & Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. doi: https://doi.org/10.1016/j.tree.2018.04.002

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural

genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203–1209. doi: 10.1111/mec.15066

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi: 10.1038/ng.608

Supplementary information

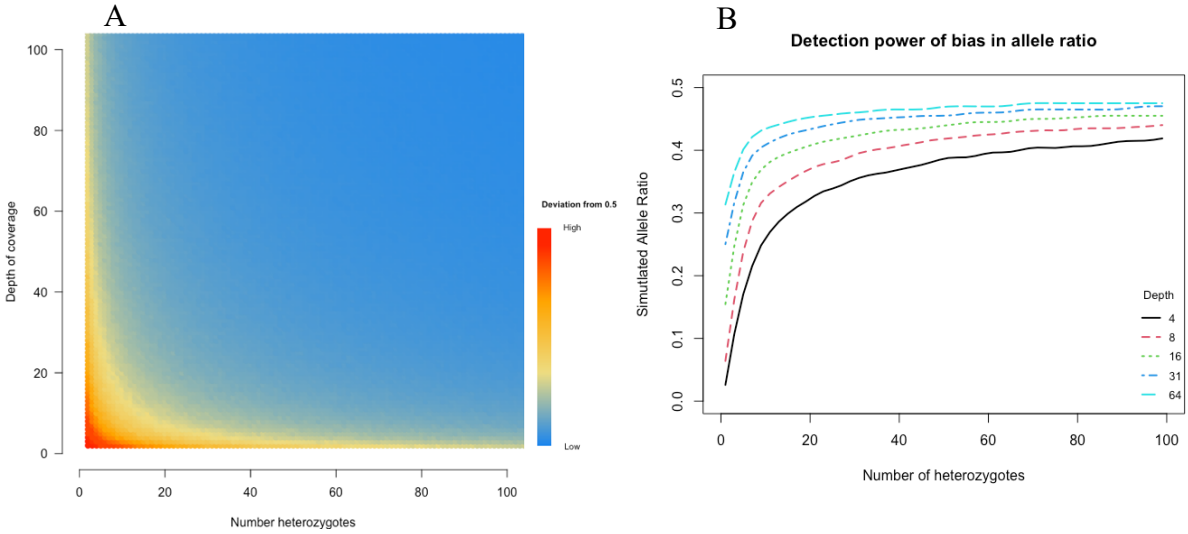


Figure 1. A. The median allele ratio variation and how the allele ratio deviates from the expected (e.g., 0.5 with no sequencing biases) in heterozygotes depending on the number of samples and different depth values, B. chage of expected allele ratio variation for different number of samples and depth values; colored lines show where the cut-off ratio lies at 95% confidence interval for both Z-score and Chi-square distribution. Both plots were generated using *depthVsSample* function in rCNV from a simulation of 10000 allele depth read values.

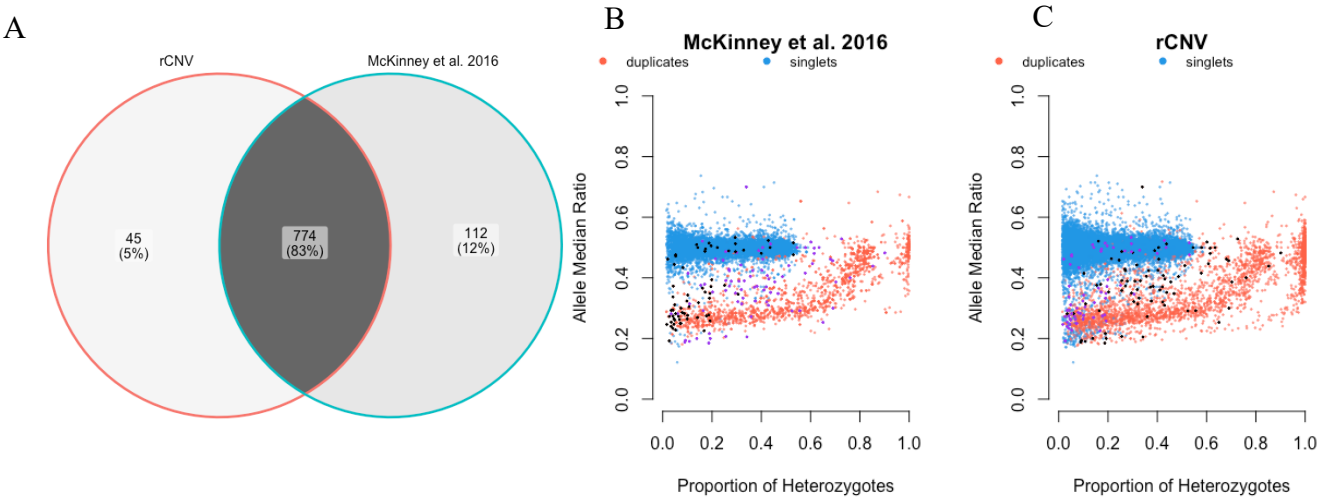


Figure 2. Comparison of CNVs detection in Chinook salmon by rCNV and McKinney (2016): A – Overlap of putatively duplicated alleles for the two methods, B&C – overlap of pseudo false positive and pseudo false negatives (black – *pFP*, purple – *pFN*)

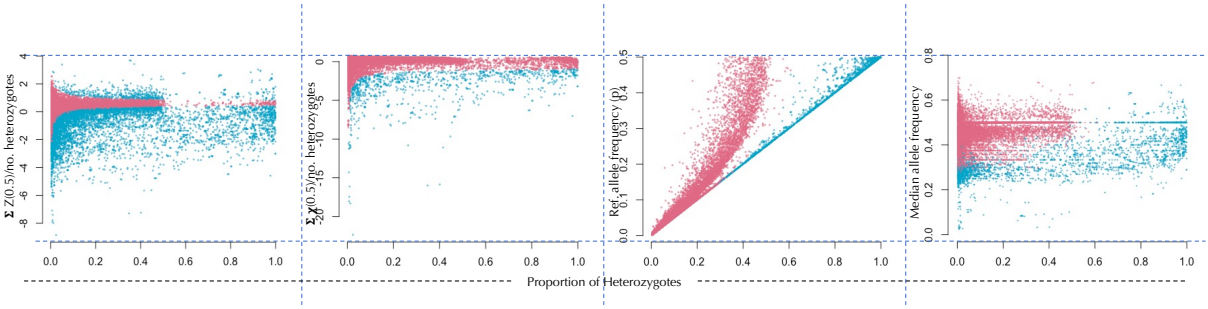


Figure 3. Different ranges of deviant SNP detection by (plots left to right) Z-score distribution of depth values, chi-square distribution of depth values, excess of heterozygotes, and the K-means clustering on American lobster data (Dorant et al. 2020)

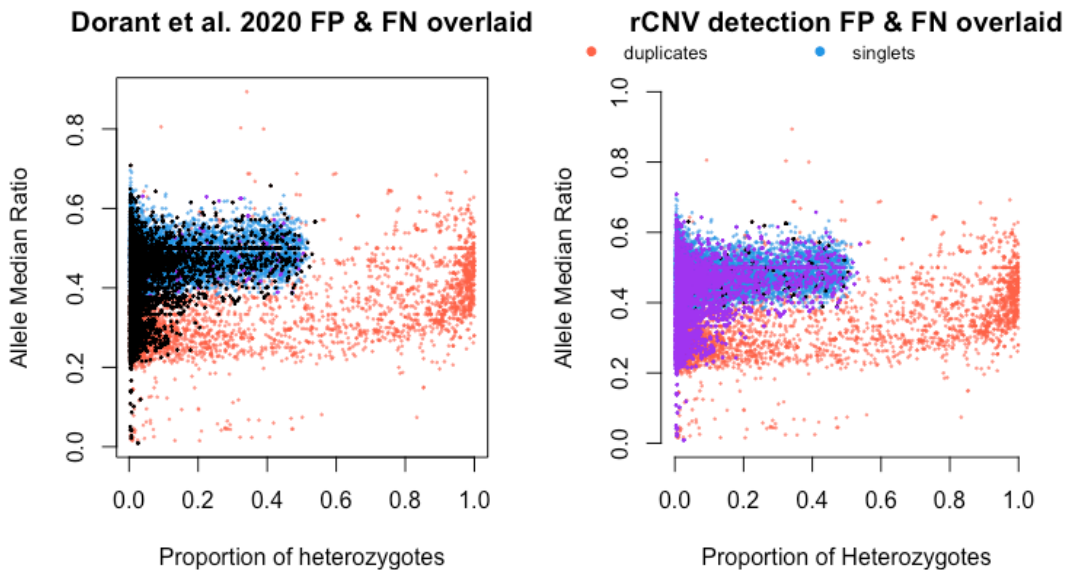


Figure 4. Overlap of false positive and false negative detections by rCNV and Dorant et al. 2020. Black – pseudo false positive, purple – pseudo false negative

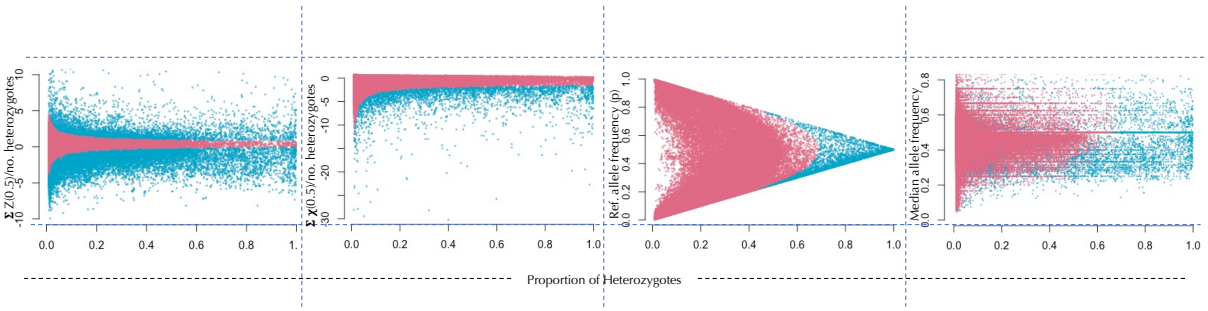


Figure 5. Different ranges of deviant SNP detection by (plots left to right) Z-score distribution of depth values, chi-square distribution of depth values, excess of heterozygotes, and the K-means clustering on Norway spruce data (subset from Chen et al. 2019).

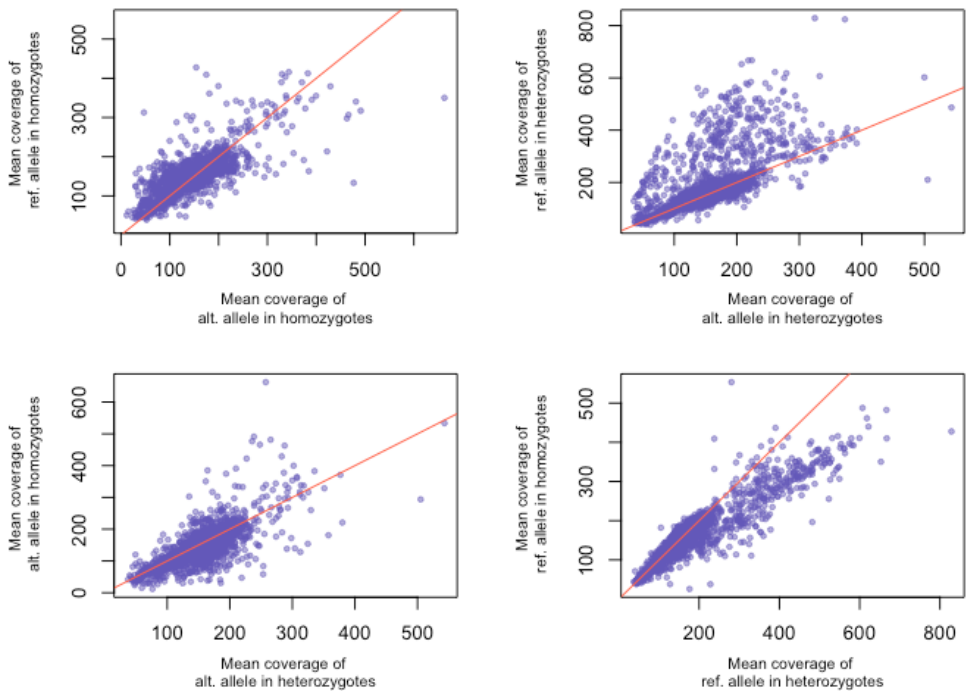


Figure 6. Correlation between coverage values of reference and alternative alleles (data from Norway spruce) showing probe bias towards the reference allele.

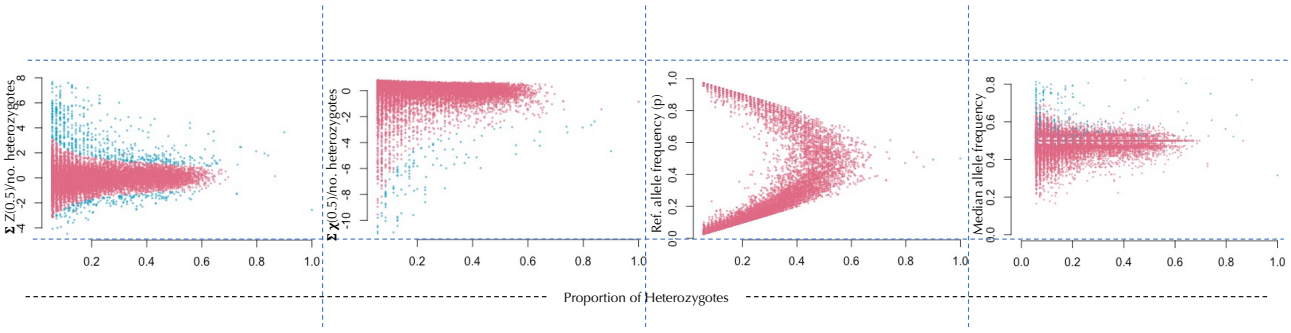


Figure 7. Different ranges of deviant SNP detection by (plots left to right) Z-score distribution of depth values, chi-square distribution of depth values, excess of heterozygotes, and the K-means clustering on *Anopheles gambiea* data

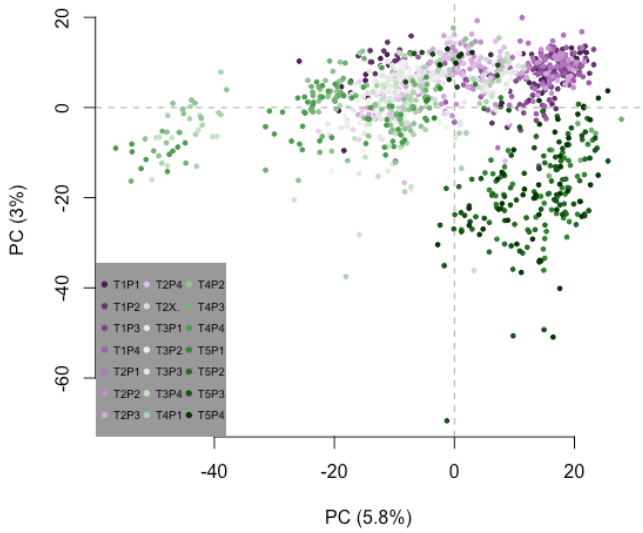


Figure 8. PCA of normalized allele depth values of putatively CNV sites of the American Lobster dataset (data - Dorant et al. 2020). The color scale shows different populations collected from different locations.

Supplementary Table 1. Recommendations for using different options in important functions included in rCNV

Step/function	**Description**	**Recommendation**
Normalization	Trimmed means of M-values (*TMM*) was proposed by (Robinson & Oshlack, 2010) calculates the scaling factors for individual samples, using which the effective library size can be calculated and thus normalized depth values. We adopted median ratio normalization (*MedR*)and quantile normalization (*QN*) methods from normalization methods for differential gene expression analysis of RNAseq data (Maza, Frasse, Senin, Bouzayen, & Zouine, 2013). The PCA normalization method assumes that the variation of samples from a principal component analysis is due to library size and removes the top PCs following a modified Kaiser's rule (eigen value < 0.7).	Although all normalization methods used in rCNV does a fairly good job in adjusting the batch effect/sequencing anomalies, the method of choice depends on the spread of the library size. Therefore, it is recommended to review the sample library size before normalization. If the user cannot decide on a method, median ratio (MedR) is a safe method.
dupGet/deviant detection	*dupGet* function efficiently combines deviant detection and allows users to use either Fisher's exact test or Chi-square test to determine the SNPs with excess of heterozygotes. With deviation from expected allelic ratios, users are given the options to use the Z-score values with or without probe biases (i.e., z.all or z.05 respectively) and the chi-square test on either heterozygotes (*chi.het*) or all (including homozygotes) (*chi.all*).	The choice of statistics for deviant detection should be based on probe bias. Putative probe biases can be assessed using the plots generated in the *allele.info* function.
Putative CNVs/ intersection set	If the underlying data was generated with sequencing methods that are less prone to probe-bias, such as RADseq, the excess of heterozygotes in combination with either *Z*-score or X^2-score for allele occurrence probability p = 0.5 will filter the optimal number of putative duplicates from the deviant SNPs. On the other hand, the sequencing technologies such as exome capture are more prone to sequencing bias and thus Z-score and chi-square tests with p determined on the distribution of allele frequency should be used in the filtering steps.	We recommend using both a less stringent approach such as p = 0.5 and more stringent z.all and chi.all for smaller datasets to compare the different.

Putative CNVs/ K-means clustering	The K-means clustering is independent of threshold or cutoff values for clustering the duplicated and non-duplicated SNPs and hence unsupervised. Since Z-score, Chi-square and delta are calculated on the read-depth values of alleles in heterozygotes, clustering can filter CNVs based on read-depth variation without being sensitive to extreme read-depth values. In contrast, Z-score and Chi-square, despite being highly effective in detecting deviants especially at lower depth values and lower heterozygosity, can produce false positives and false negatives due to strict threshold values. However, K-means clustering lacks power at higher heterozygosity in combination with lower read-depth values. Therefore, we combine excess of heterozygosity and K-means for filtering putative CNVs.	Although K-means clustering is always a safer method for filtering, we recommend the user pay attention to the proportion of heterozygosity and the distribution of depth values to avoid false negatives. Intersection set performs better for datasets with lower depth and higher proportion of heterozygotes (e.g., mean depth < 10 & propHet > 0.7)