Population genetics of immune-related multilocus CNV in Native Americans

Journal:	*Journal of the Royal Society Interface*
Manuscript ID	rsif-2017-0057
Article Type:	Research
Date Submitted by the Author:	25-Jan-2017
Complete List of Authors:	Zuccherato, Luciana; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Schneider, Silvana; Universidade Federal de Minas Gerais, Departamento de Estatística, Instituto de Ciências Exatas Tarazona-Santos, Eduardo; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Hardwick, Robert; University of Leicester, Department of Genetics Berg, Douglas; Washington University in Saint Louis School of Medicine, Department of Molecular Microbiology; University of California San Diego , Department of Medicine Bogle, Helen; University of Leicester, Department of Genetics Gouveia, Mateus; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Machado, Lee; University of Leicester, Department of Genetics; University of Northampton, School of Health Machado, Moara; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Rodrigues-Soares, Fernanda; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Soares-Souza, Giordano; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Togni, Diego; Universidade Federal de Minas Gerais, Departamento de Estatística, Instituto de Ciências Exatas Zamudio, Roxana; Universidade Federal de Minas Gerais, Departamento de Biologia Geral Gilman, Robert; Johns Hopkins University, Johns Hopkins School of Public Health; Asociación Benéfica PRISMA; Universidade Peruana Cayetano Heredia Duarte, Denise; Universidade Federal de Minas Gerais, Departamento de Estatística, Instituto de Ciências Exatas Hollox, Edward; University of Leicester, Department of Genetics Rodrigues, Maira; Universidade Federal de Minas Gerais, Departamento de Biologia Geral
Categories:	Life Sciences - Mathematics interface
Subject:	Bioinformatics < CROSS-DISCIPLINARY SCIENCES, Computational biology < CROSS-DISCIPLINARY SCIENCES

Keywords:	immunity, population structure, Amerindians, profiled likelihood, genomic structural variation

Population genetics of immune-related multilocus CNV in Native Americans

Luciana W Zuccherato[1†], Silvana Schneider[2†], Eduardo Tarazona-Santos[1], Robert J Hardwick[3], Douglas E Berg[4,5], Helen Bogle[3], Mateus H Gouveia[1], Lee R Machado[6,3], Moara Machado[1], Fernanda Rodrigues-Soares[1], Giordano B Soares-Souza[1], Diego L Togni[2], Roxana Zamudio[1], Robert H Gilman[7,8,9], Denise Duarte[2], Edward J Hollox[3], Maíra R Rodrigues[1*]

[†] These authors equally contributed to the present paper.

*To whom correspondence should be addressed.

Abstract

While multiallelic copy number variation (mCNV) loci are a major component of genomic variation, quantifying the individual copy number of a locus and defining genotypes is challenging. Few methods exist to study how mCNV genetic diversity is apportioned within and between populations (i.e., to define the population genetic structure of mCNV). These inferences are critical in populations with a small effective size, such as Amerindians, that may not fit the Hardy-Weinberg model due to inbreeding, assortative mating, population subdivision, natural selection or a combination of these evolutionary factors. We propose a likelihood-based method that simultaneously infers mCNV allele frequencies and the population structure parameter f, which quantifies the departure of homozygosity from the Hardy-Weinberg expectation. This method is implemented in the freely available software CNVice, which also infers individual genotypes using information from both the population and from trios, if available. We studied the population genetics of five immune-related mCNV loci associated with complex diseases (beta-defensins, *CCL3L1/CCL4L1*, *FCGR3A*, *FCGR3B*, and *FCGR2C*) in 12 traditional Native American populations and found that the population structure parameters inferred for these mCNVs are comparable to but lower than those for SNPs studied in the same populations.

Keywords: immunity, population structure, Amerindians, profiled likelihood, genomic structural variation

1 Introduction

Multiallelic copy number variation (mCNV) is an underappreciated and complex component of genetic variation that has been challenging to detect for two reasons. First, they are not effectively tagged by flanking SNPs. Second, direct measurements of hybridization intensity from single nucleotide polymorphism (SNP) or comparative genome hybridization arrays are often noisy. Techniques based on PCR such as the paralogue ratio test [1] and, more recently, sequence read depth analysis from short-read second-generation sequencing have begun to allow analysis of mCNV across different human populations [2-5]. However, the number of human populations studied remains low and has focused mainly on Europeans.

Complex mCNV involves genes that are of biological and medical interest. For example, the immune system proteins beta-defensins (DEFB), macrophage inflammatory protein 1 α (MIP-1α) and Fcγ receptors (FCGRs) are encoded by mCNV loci that modulate susceptibility to infectious and autoimmune diseases. Beta-defensins are small cationic peptides with a role in innate immunity that interact with pathogens by depolarizing and rendering their cellular membrane permeable. Increases in copy number (CN) of beta-defensins are associated with psoriasis [6], as well as an increase in HIV load and impaired immune reconstitution following the initiation of Highly Active Antiretroviral Therapy [7].

MIP-1α, also known as the Chemokine ligand 3-like 1, (encoded by *CCL3L1*) binds the CC chemokine receptors CCR1, CCR3 and CCR5, and their CNV has been inconsistently associated with clinical parameters of HIV infection [8,9].

A cluster of Fc gamma receptors with low affinities for IgG shows mCNV (*FCGR3A, FCGR3B,* and *FCGR2C*), and there is some evidence of association with disease. Low *FCGR3B* copy number is associated with glomerulonephritis, systemic lupus erythematosus and rheumatoid arthritis [10,11], and the non-synonymous *FCGR2C* mutation rs10917661 results in an activated FcRIIc protein with cytotoxic effects [12].

The study of mCNV is challenging due to difficulties in quantifying the number of copies of a locus in an individual [13] and, consequently, quantifying how genetic diversity is apportioned between individuals and populations. Typing methods do not reveal true genotypes for CNV loci, as is the case of SNPs; rather, quantitative information about total copy numbers of a locus in both chromosomes (diploid copy number) is produced. Therefore, in the absence of information from segregation in pedigrees or physical information from molecular methods such as fibre-FISH, the true copy number genotype that identifies the allele carried by each chromosome can only be probabilistically inferred from the distribution of observed diploid copy numbers in a population. This limits the application of population genetics models developed for diploid genotypes, and indeed, specific methods have been proposed to (i) infer combined SNPs/CNV haplotypes [14]; (ii) quantify population differentiation by the F_{ST}-like V_{ST} statistic in which quantitative intensity ratios are directly obtained from the genotyping signals [15]; and (iii) infer CN allele frequencies assuming Hardy-Weinberg equilibrium [16]. However, some natural populations do not fit the Hardy-Weinberg model, due to inbreeding, assortative mating, population subdivision, natural selection or a combination of these evolutionary forces.

This article has two goals. The first is to propose a method (focused on mCNVs) to study the genetic structure of populations (i.e., the departure from the Hardy-Weinberg model due to the apportioning of genetic diversity within and between populations). To achieve that aim, we generalize the Gaunt et al. [16] algorithm, allowing for deviations from Hardy-Weinberg equilibrium. We implemented our more general approach in the R software CNVice, an acronym for *Inbreeding Coefficients Estimation for CNV data,* freely available at https://github.com/mairarodrigues/cnvice. Although CNVice measures the departure from Hardy-Weinberg equilibrium in general, and not only that due to recent identity by descent (i.e., inbreeding), the name of the software is reminiscent of the classical inbreeding studies of Wright [17].

We implement a method in CNVice that simultaneously infers for a population (i) the CNV allele frequencies and (ii) a multiallelic f , the most classical and widely used population genetics parameter [17,18]. f quantifies the departure of homozygosity from Hardy-Weinberg expectations, summarizing the genetic structure of natural populations. Formally, the probability of a homozygous genotype (h_k, h_k) for the allele k is $fp_k + (1-f)p_k^2$ and of a heterozygous genotype (h_k, h_q) for the alleles k and q is $(1-f)2p_kp_q$ [19], where p_k and p_q are the frequencies of k and q alleles in the population. The Hardy-Weinberg equilibrium corresponds to the specific case of $f = 0$. Moreover, our method infers individual genotype probabilities for CNV loci, based on the observed individual diploid copy number, the population allele frequency distribution, the inferred f parameter, as well as information on inheritance patterns from trios, if available.

The second goal of our article is to determine the population genetics and genetic structure of three immune-related mCNV loci in Native American populations, which have been rather neglected in human genome diversity studies. We studied twelve traditional villages from three ethnic South American native groups (Figure 1A) whose genetic structure has likely been affected by evolutionary factors, such as strong genetic drift and inbreeding, that are more relevant in small populations and that render methods assuming Hardy-Weinberg equilibrium suboptimal.

Using CNVice, we inferred copy number genotypes at the beta-defensin (*DEFB*) locus, the *CCL3L1/CCL4L1* locus and the low affinity Fc gamma receptor cluster locus (*FCGR*). mCNV at the *FCGR* locus can be further subdivided into copy number variation of the individual genes, namely *FCGR2C*, *FCGR3A* and *FCGR3B*. All mCNV loci studied show a high level of copy number diversity.

2 Results and Discussion

2.1 Assessing population structure with CNVice

We tested CNVice and compared it with CoNVEM [16], which also estimates mCNV allele frequencies using the expectation maximization algorithm but assumes Hardy-Weinberg equilibrium. For this, we used simulated data representing different levels of diversity at a mCNV locus (measured by expected heterozygosity H under Hardy-Weinberg equilibrium, and by the number of alleles, Table S1). The ranges of parameters were consistent with those previously observed at the mCNV loci analysed in this study. Both CoNVEM and our new CNVice software estimate allele frequencies for mCNV loci. Figure S1 compares allele frequencies estimated with CoNVEM and CNVice, under different levels of population diversity. In general, CoNVEM and CNVice produce similar allele frequency estimates. However, in some instances, CNVice estimates are more accurate than CoNVEM as the departure from Hardy-Weinberg equilibrium increases (Figure S1A, Table S2).

Tables 1 and S2 show CNVice inferences of allele frequencies and f_{CNV} on simulated data. In all cases, CNVice 95% CIs of allele frequencies contain the true allele frequencies. Moreover, CNVice estimations of allele frequencies become less accurate when the allele frequency spectrum is dominated by a very common allele (>.60). In this case, the frequency of the most common allele is underestimated and the frequencies of rare alleles are overestimated.

CNVice is novel in its estimation of the population structure parameter f from mCNV data. The estimator f_{CNV} captures the population structure for most mCNV loci when there is departure from Hardy-Weinberg equilibrium (Table 1 and Table S2), but, as in the case of allele frequency estimations, f is underestimated when there is a very common allele. When the mCNV locus diversity is very low, due to the presence of a predominant allele (frequency > 0.80 in Table S2), CNVice estimators have a large bias, a problem shared by most estimators of f statistics in scenarios of low genetic diversity [28].

2.2 The genetic structure of Native American populations for immune-related mCNV loci

The very low European or African admixture of the studied populations indicates that they are reasonable representatives of autochthonous Native American traditional populations (Figure 1B, average individual non-Native American ancestry: 2%). Previous studies of the genetic structure and evolution of mCNV have focused on the global level [15, 24, 25, 3]. This is the first population genetics study that assesses the level of population structure of mCNV loci (in this case, immune-related) in a set of autochthonous and traditional villages from different linguistic groups, residing in different environments (the Andean highlands and the Amazon Yunga forest) (Figure 1A). In this case, the 12 studied populations are scattered in an area of nearly 25,000 km^2 (similar to the size of Sardinia) that was peopled at least 10,000 years ago [29]. Therefore, our study is informative about how inbreeding, genetic drift, gene flow and probable selective pressures (associated with different environments) shape the genetic structure of traditional populations for mCNV loci.

Figures 1C-D and Supplementary Tables S3-S7 show the observed diploid copy number distributions, the inferred allele frequencies and f_{CNV} for the studied mCNV loci. Estimated f_{CNV} values are within the range of f estimates observed for unlinked SNPs in the same populations (Figure 1D). CNVice allows assessment of the uncertainty of the inference by examining the estimated f_{CNV} likelihood profile (Figure 1C). On the other hand, it has the limitation that it only allows positive values of f_{CNV}.

The five immune-related mCNV loci studied here have a level of population structure (mean f_{CNV} = 0.018) that is lower than the mean f = 0.136 observed for 695 unlinked SNPs. The two Matsiguenga villages, which are separated by nearly 80 km in the same Amazon Yunga valley, show the highest level of population structure (Figure 1C, mean f_{CNV}= 0.04).

DEFB mCNV [24] and *FCGR* mCNV [25] have low levels of genetic structure worldwide when compared with genome-wide estimates for CNV loci [15]. When we used CNVice to estimate allele frequencies and f_{CNV} based on published *DEFB* [24] and *FCGR3B* data [25], we confirmed the low level of worldwide genetic structure previously observed (Tables S8 and S9, Figure S3). In contrast, the genetic structure of *CCL3L1/CCL4L1* is higher worldwide [3]. The estimated f_{CNV} values for Native Americans studied here are partially consistent with the patterns of population structure observed for *FCGR* mCNV and *CCL3L1/CCL4L1* mCNV worldwide. Indeed, despite local variation in the f_{CNV} statistics between the 12 Native American populations, the inferred f_{CNV} values across the populations are 0 for beta-defensins, *FCGR3A* and *FCGR2C* and higher (0.016) for *CCL3L1/CCL4L1* (Tables S3-S7). *FCGR3B* had the highest observed f_{CNV} (0.076).

The following features of the genetic structure of the immune-related mCNVs in Peruvian Native Americans are noteworthy:

(i) For beta-defensins (Table S3), the Shimaa population exhibited an elevated frequency of a diploid copy number of 7 (11.4% vs. 1.7% globally, [24]). Considering that the 2-copy allele is most common worldwide, the Shimaa result likely derives from being the only studied population in which the 5-copy allele is common (frequency of 9% vs. <0.4% elsewhere), which may reflect the action of genetic drift.

(ii) For *CCL3L1/CCL4L1* (Table S4), our Native American populations share modal diploid copy numbers (3-4 copies) with West African Yoruba [9] and differ from Europeans (modal diploid copy number: 2 [30]). This is because Native Americans show allele 2-copies as modal. Native Americans are more diverse than Europeans [30] but less diverse than Africans [9].

(iii) Both the *CCL3L1/CCL4L1* genes occur in a single CNV block in most populations (although they are separate in a small number of Ethiopian samples) [9]. The high correlation (Pearson R^2>0.91, p<0.001) between the normalized copy number values across the pairs of the three amplified loci (*CCL3C*, *CCL4A*, and *LTR61A*, Figure 2) in Native American populations suggests that, in these populations, the region encompassing the *CCL3L1* and *CCL4L1* genes are on the same repeat unit.

(iv) For *FCGR3B*, the Amazonian Yunga populations have the highest known frequency of gene deletion, likely due to genetic drift with frequencies of homozygosity for this deletion of 2.4% and 11.6% in Ashaninkas and Matsiguengas, respectively, vs. <0.31% in the worldwide Human Genome Diversity Panel populations [25]. This trend is also seen, although less strikingly, for *FCGR2C* gene deletions (Tables S6-S7).

(v) Q57X is a relevant substitution in *FCGR2C* (rs10917661, Table S10). While the frequency of this sequence variant in South and Central Native American samples from the HGDP-panel is similar to that in many other global populations (0.24) [25], our samples from Western South Amerindians have the lowest frequency (0.07) of this allele (Table S10); only that in East Asians is lower (0.05).

2.3 Inferring CNV genotypes using trios

Defining individual copy number genotypes for mCNV may be challenging, but it is important, for example, to identify carriers of specific alleles to be re-sequenced to study their associated nucleotide diversity and genomic organization. This information may be particularly relevant in neglected populations with few studied individuals such as Native Americans, to compare the nucleotide diversity and genomic organization of CNV alleles with other well-studied populations. As an example of how CNVice uses both population diploid copy number frequencies and trio diploid copy numbers to infer individual mCNV genotypes, we consider data from an Ashaninka trio for the beta-defensins cluster (a mother with 2 copies, a father with 4 copies and an offspring with 4 copies) (Figure 3). In this population, the beta-defensins diploid copy number varies from 0 to 9 with the following respective absolute frequencies: (0, 0, 7, 45, 51, 23, 10, 5, 1, 1) for 143 individuals (i.e., 0 individuals with 0 copy, 0 individuals with 1 copy, 8 individuals with 2 copies and so forth) (Table S11, line 6, column Diplotype frequencies (CN)). The possible genotypes for the offspring ASH064, as for all individuals with 4 copies in the population, are (0, 4), (1, 3) or (2, 2). First, CNVice estimates f_{CNV} = 0 and the vector of allele frequencies for the population, P = (0, 0.25, 0.56, 0.12, 0.04, 0.02, 0, 0.002, 0, 0) for alleles from 0 to 7 copies, using the observed copy number distribution above (Figure 3A). Based on this information and following Equation (1) in section 3.2 of the Methods, the probabilities of each of the three genotypes are 0.839 for (2, 2), 0.161 for (1, 3) and 0 for (0, 4). Note that these genotype probabilities apply to all individuals with diploid copy number equal to 4 (Figure 3B).

Equation (1) (section 3.2) estimates a prior probability of a genotype, given population genetics information. If trio information is available, CNVice uses the parents' diploid copy number information applying the Bayes theorem, to estimate a posterior probability of a genotype, considering the prior probability estimated by equation (1) and also diploid copy number information from the parents. As shown in Figure 3C, for individual ASH06, with information of the parents' diploid copy number, CNVice infers that this individual carries the (1, 3) genotype and not the (2, 2) genotype for beta-defensins. This calculation takes into account the population allele frequencies, the individual's genotype frequencies and the parents' genotype frequencies. As Figure 3 shows, because there is no allele 0 in the population, the genotypes (0, 4) for individual ASH06, (0, 2) for the individual's mother and (0, 4) for the individual's father are eliminated from the calculation. As the only possible allele the mother can pass on to her offspring is 1, the only possible genotype for ASH06 is (1, 3). This reasoning is formalized in Equation (2) of section 3.2.

We tested this functionality of CNVice in 53 trios with unique diploid copy number combinations for beta-defensins, *CCL3L1/CCL4L1, FCGR3A*, *FCGR3B* and *FCGR2C*, and found that the inferences of offspring genotypes improve in 49% of cases, with the posterior probability of the most likely genotype increasing on average from 0.74 to 0.83 (Table S11 and Figure S2). This means that, in general, the information about the diploid copy number of the parents reduces the uncertainty about the individual genotype. However, a decrease in the probability of the most likely genotype, or an increase in the uncertainty of the individual genotype (as seen in 7.6% of cases in Figure S2), is still an optimized result and indicates that there was as overestimation in the prior probability. Because trio information limits the possible genotypes of an individual, by imposing additional constraints on its calculation, trio information corrects such over- or underestimates of prior probabilities.

3 Methods

In this section, we outline the Materials and Methods; further details are given in the Supplementary Material section.

3.1 Populations, Samples and Genotyping

We analysed between 324 and 375 individuals (depending on the locus) hierarchically sampled from 12 populations (hereafter called villages), belonging to three Peruvian ethnic groups (Figure 1A, Table S12): (a) 143 Ashaninka from five villages along the Junin River (Central Peru); (b) 113 Matsiguenga living in the villages of Monte Carmelo (n=24) and Shimaa (n=89), in Southern Peru; and (c) 120 Aymara highlanders from the Andean region, living in five villages near the Titicaca Lake (Southern Peru). The Ashaninkas and Matsiguengas are settled in the Amazon Yunga tropical forest environment, and their languages belong to the Arawak linguistic family, while the highlanders' language, Aymara, belongs to the Andean linguistic family. For population genetics analyses, we avoided genotyping parents, offspring or siblings for CNVs, except for the Ashaninkas. For the latter, 69 families composed of two parents and their offspring were further genotyped for 277 individuals, which include the 143 unrelated individuals considered for the main analyses (Table S12). The Institutional Review Boards of the participant institutions approved this study.

We used 103 ancestry informative marker SNPs [20] to estimate African, European and Native American ancestry in these villages, using the software *Admixture* v.1.2 [21].

To compare the genetic structure based on mCNVs with that based on SNP genotypes, we used 695 SNPs that were unlinked between them as well as with respect to the mCNV loci. These SNPs were genotyped in 124 individuals from the same studied populations. The 695 SNP genotypes are the intersection of two datasets obtained using different technologies: 1442 gene-centric and cancer-associated SNPs from an Illumina Golden Gate Oligonucleotide Pool Assay (genotyped by Dr. Stephen J Chanock's group at the National Cancer Institute), and 2.3M SNPs genotyped with the Illumina's HumanOmni2.5-8v1 array. We used the hierfstat R package [22] to estimate the *f* statistics for each SNP.

We measured diploid copy number in the *DEFB* region, *CCL3L1/CCL4L1* and for the *FCGR3A, FCGR3B* and *FCGR2C* genes, using the paralogue ratio test (PRT) approach, described previously for each locus [1,23,31]. The latter technique is a development of quantitative PCR that uses a single primer pair to simultaneously amplify both a test locus and a reference locus, allowing an accurate CN determination [1].

After amplification, the ratio of the amounts of amplification products (i.e., the corresponding areas under their capillary electrophoresis peaks) between the reference and test locus is calculated. Reference samples with known CN (Table S13) are then used to convert these ratios in an expected CN estimation, which is a continuous number. Next, a probabilistic model is used to obtain a maximum likelihood estimation of the discrete CN, combining information from different primers [23]. The relative proportions of each allele for SNPs on the promoter of *DEFB103* (rs2737902), *FCGR3A*/*FCGR3B* genes (rs1042207 C>T, Arg234X), *FCGR3B* HNA1a/1b allelotype (rs76714703 C>T, Asn468Ser) and *FCGR2C* null variation (rs10917661 C>T, Glu57X) were determined using multiplex restriction enzyme digest variant ratio (REDVR) [24,25].

The combined use of multiplex REDVR with PRT allows the determination of copy number, the relative proportion of each allele for SNPs and paralog genes, which adds an additional layer of complexity. For instance, in contrast to the CCL3L1 and beta-defensin CNs, which gene products are identical between copies, or almost so (and are very likely to have the same function), we can distinguish between deletions in FCGR3A or in FCGR3B. Considering their importance as functionally distinct genes, they are separately included in the analysis. This strategy also implies that FCGR regions is analysed at a higher resolution respect to the other studied mCNVs loci.

3.2. Algorithm

CNVice performs four analyses. First, it estimates allele frequencies for a given population assuming Hardy-Weinberg equilibrium, using the expectation maximization algorithm as implemented in CoNVEM [16], conditioned on the observed diploid copy number frequencies.

Second, CNVice implements a more general approach to study population structure by jointly estimating f_{CNV} (i.e., the population structure parameter *f*) and allele frequencies. For this, a profiled likelihood function and expectation maximization (EM) algorithm is used [26,27], where the set of allele frequencies is the main parameter and f_{CNV} is the perturbation parameter. The perturbation parameter is replaced by a maximum likelihood estimate in the original likelihood while maintaining fixed values for allele frequencies. The analytical derivation and CNVice algorithm are detailed in the Supplementary Material and in Figure 4, respectively.

Third, CNVice calculates individual genotype probabilities given a diploid copy number and estimates population genotype frequencies, based on both the observed diploid copy number distribution and the estimated f_{CNV} and allele frequencies, using the expression:

$$P_{\hat{f}_{ind}}(h_k, h_q \mid j) = \frac{2P_{\hat{f}}(h_k, h_q)}{\sum_{k=0}^{m}\sum_{q=0}^{m} P_{\hat{f}}(h_k, h_q)}, j = k + q, \qquad (1)$$

where $P_{\hat{f}_{ind}}(h_k, h_q \mid j)$ is the expected genotypic proportion given the estimated f_{CNV}.

Finally, CNVice uses trio information, if available, to improve the inference of mCNV genotype by considering diploid copy numbers of the parents and using Bayes Theorem. The offspring genotype probability given the diploid copy number and genotypic probability of the parents is:

$$P_{\hat{f}_{ind}}(h_{ks}, h_{qs} \mid j_s) = \frac{P_{\hat{f}}(h_{ks}, h_{qs} \mid h_{km}, h_{qm}, h_{kf}, h_{qf}) P_{\hat{f}}(h_{km}, h_{qm} \mid j_m) P_{\hat{f}}(h_{kf}, h_{qf} \mid j_f)}{\sum_{km=0}^{m}\sum_{qm=0}^{m}\sum_{kf=0}^{m}\sum_{qf=0}^{m} P_{\hat{f}}(h_{ks}, h_{qs} \mid h_{km}, h_{qm}, h_{kf}, h_{qf}) P_{\hat{f}}(h_{km}, h_{qm} \mid j_m) P_{\hat{f}}(h_{kf}, h_{qf} \mid j_f)}, \qquad (2)$$

where h_{ks} and h_{qs} denote the offspring's alleles; h_{km} and h_{qm} denote the mother's alleles; h_{kf} and h_{qf} denote the father's alleles; j_m denotes the mother's diploid copy number; and j_f denotes the father's diploid copy number. It is required that:

(1) k_s+q_s=j_s;
(2) ($h_{ks} = h_{km}$) or ($h_{ks} = h_{qm}$) or ($h_{ks} = h_{kf}$) or ($h_{ks} = h_{qf}$), and
(3) ($h_{qs} = h_{km}$) or ($h_{qs} = h_{qm}$) or ($h_{qs} = h_{kf}$) or ($h_{qs} = h_{qf}$).

For the Native American data, we estimated allele frequencies and f_{CNV} using CNVice. We also used CNVice to assess trios from the Ashaninka population and to infer individual genotypes using information about their parents' diploid copy number.

4 Conclusion

We present a novel likelihood approach, implemented in the CNVice software, which allows for the study of the genetic structure of natural populations using the classical *f* statistic framework. Using Monte Carlo simulations, we show that our method improves (1) the estimation of allele frequencies when departure from Hardy-Weinberg equilibrium increases and (2) the estimation of individual genotypes in comparison with existing methods, in particular when trio information is available. Using our approach, we observed a low level of genetic structure for three immune-related mCNV loci in a set of traditional Native American populations, settled for at least 10,000 years in the different Andean and Amazon Yunga environments. Our method, used here to infer the genetic structure of traditional Native American populations, is also applicable to the most diverse diploid species of animals and plants, where departure from the Hardy-Weinberg model due to drift, assortative mating or natural selection may be stronger and more frequent than in human populations.

5 Data and Software Accessibility

Details of the statistical methods presented here are in the Supplementary Material. CNVice is implemented in R, and the source code can be found at https://github.com/mairarodrigues/cnvice. The article's supporting data are also available as Supplementary Material.

The software run time depends mostly on the number of alleles and the sample size that form the distribution of observed diplotype frequencies given as input. For example, an input containing the distribution of 8 alleles in a sample size of 120 individuals takes 7.42 minutes to run in a computer with 12 GB of RAM, while the same allele distribution in a sample size of 30 individuals takes almost 7 times less (that is, 1 minute). Similarly, an input with 4 alleles and sample size of 120 takes only 13 seconds. CNVice also works for larger copy number ranges (it has been tested for distributions with up to diploid CN=20).

Acknowledgements

We thank Lavínia Schuler-Faccini, Alessandra Pontillo, Tiago Magalhães, and Cláudia Teixeira Guimarães for discussion and critiques.

Funding

This work was supported by the European Molecular Biology Organization (EMBO) travel grant and the Brazilian National Research Council (CNPq).

Conflict of Interest: none declared.

Authors' contributions

LWZ performed the copy number genotyping and sequence analysis; SS and DD developed the statistical methods and SS carried out experiments with simulated data and Monte Carlo simulations; LWZ, EJH, RJH, LRM, HB, GBSS, MHG, FRS and MM carried out data analysis; DEB, RZ and RHG participated in sample collection; DLT implemented the software; MRR participated in algorithm development and software implementation; ETS conceived the work; EJH, ETS, DD and MRR supervised the work; LWZ, SS, DD, DEB, EJH, ETS and MRR participated in the draft of the manuscript. All authors gave final approval for publication.

References

[1] Walker, S., Janyakhantikul, S. and Armour, J.A. Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. Genomics 2009;93(1):98-103.

[2] Handsaker, R., E. et al. Large multiallelic copy number variations in humans. Nat Genet 2015; 47:296-303.

[3] Sudmant, P., et al. Diversity of Human Copy Number Variation and Multicopy Genes. Science 2010;330(6004):641-646.

[4] Zarrei, M. et al. A copy number variation map of the human genome. Nat Rev Genet 2015 16:172-83.

[5] Forni, D. et al. Determining multiallelic complex copy number and sequence variation from high coverage exome sequencing data. BMC Genomics; 16(1):891.

[6] Hollox, E.J., et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 2008;40(1):23-25.

[7] Hardwick, R.J., et al. β-defensin genomic copy number is associated with HIV viral load and immune reconstitution in sub-Saharan Africans. J Infect Dis 2012; 206:1012-9.

[8] Gonzalez, E., et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307(5714):1434-1440.

[9] Aklillu, E., et al. CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans. BMC infectious diseases 2013;13:536.

[10] Willcocks, L., et al. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. Journal of Experimental Medicine 2008;205(7):1573-1582.

[11] Rahbari, R., Zuccherato, L.W., Tischler, G., Chihota, B., Ozturk, H., Saleem, S., Tarazona-Santos, E., Machado, L.R. & Hollox, E.J. 2016 Understanding the Genomic Structure of Copy Number Variation of the Low-Affinity Fcgamma Receptor Region Allows Confirmation of the Association of FCGR3B Deletion With Rheumatoid Arthritis. Hum Mutat. (doi:10.1002/humu.23159).

[12] Breunis, W., et al. Copy Number Variation at the FCGR Locus Includes FCGR3A, FCGR2C and FCGR3B but not FCGR2A and FCGR2B. Human Mutation 2009;30(5):E640-E650.

[13] Cantsilieris, S. et al. Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication. BMC Genomics 2014; 15:329.

[14] Su, S.Y., et al. Inferring combined CNV/SNP haplotypes from genotype data. Bioinformatics 2010; 26(11):1437-45.

[15] Redon, R., Global variation in copy number in the human genome. Nature 2006;444(7118):444-54.

[16] Gaunt, T., et al. An Expectation-Maximization Program for Determining Allelic Spectrum from CNV Data (CoNVEM): Insights into Population Allelic Architecture and Its Mutational History. Human Mutation 2010;31(4):414-420.

[17] Wright, S. Coefficients of inbreeding and relationships. American Naturalist 1922;56:330-338.

[18] Wright, S. The genetical structure of populations. Ann Eugen 1951;15(4):323-354.

[19] Hedrick, P.W. The Genetics of Populations. Jones and Barlett Publishers. US. 2005.

[20] Pereira, L., et al. Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population. PLoS One 2012;7(8):e41200.

[21] Alexander, D.H., Novembre, J. and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19(9):1655-1664.

[22] Goudet, J. hierfstat, a package for r to compute and test hierarchical F-statistics. Molecular Ecology Notes 2005;5:184–186.

[23] Hollox, E.J., Detering, J.C. and Dehnugara, T. An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus. Hum Mutat 2009;30(3):477-484.

[24] Hardwick, R., et al. A Worldwide Analysis of Beta-Defensin Copy Number Variation Suggests Recent Selection of a High-Expressing DEFB103 Gene Copy in East Asia. Human Mutation 2011;32(7):743-750.

[25] Machado, L.R., et al. Evolutionary History of Copy-Number-Variable Locus for the Low-Affinity Fcγ Receptor: Mutation Rate, Autoimmune Disease, and the Legacy of Helminth Infection. Am J Hum Genet 2012.

[26] Dempster, A.P., Laird, M.N. and Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 1977. 39(1): 1-38.

[27] Severini, T.A. Likelihood Methods in Statistics. Oxford University Press; 2000.

[28] Long, J.C. The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics 1986;112(3):629-47.

[29] Scliar, M.O., et al. Bayesian inferences suggest that Amazon Yunga Natives diverged from Andeans less than 5000 ybp: implications for South American prehistory. BMC Evol. Biol. 2014; 14(1):1-8.

[30] Field, S.F., et al. Experimental aspects of copy number variant assays at CCL3L1. Nat Med 2009;15(10):1115-1117.

[31] Aldhous, M.C. et al. Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. Hum Mol Genet. 2010;19(24):4930-8.

Table 1. Estimation of the population structure parameter *f* by CNVice. Monte Carlo simulations (1000 replications) were performed sampling from populations with different levels of diversity, assuming a sample size of 100 individuals and different values of the *f* parameter: 0.05, 0.10, and 0.20. Estimated f_{CNV} values correspond to the mean of the 1000 replications. Allele frequencies and diversity parameters for the six simulated populations are in Table S1. Allele frequencies estimated by CNVice together with f_{CNV} are listed in Table S3.

Pop	**Alleles**	**Genetic**	**Observed *f* = 0.05**		***f* = 0.10**		***f* = 0.20**	
		diversity	**Estimated** f_{CNV}	**[CI 95%]**	**Estimated** f_{CNV}	**[CI 95%]**	**Estimated** f_{CNV}	**[CI 95%]**
1	3	0.66	0.07150	[0.00000;0.23972]	0.10221	[0.00000;0.29497]	0.17170	[0.00000;0.40610]
2	4	0.38	0.03901	[0.00000;0.21126]	0.06026	[0.00000;0.27732]	0.10469	[0.00000;0.39309]
3	3	0.56	0.04858	[0.00000;0.16891]	0.08628	[0.00000;0.23701]	0.17951	[0.00000;0.37278]
4	9	0.88	0.05316	[0.00000;0.20805]	0.08369	[0.00000;0.27313]	0.17085	[0.00000;0.41368]
5	9	0.19	0.17217	[0.00000;0.69289]	0.18816	[0.00000;0.72492]	0.23368	[0.00000;0.81062]
6	7	0.61	0.03502	[0.00000;0.19410]	0.05859	[0.00000;0.26130]	0.12928	[0.00000;0.42511]

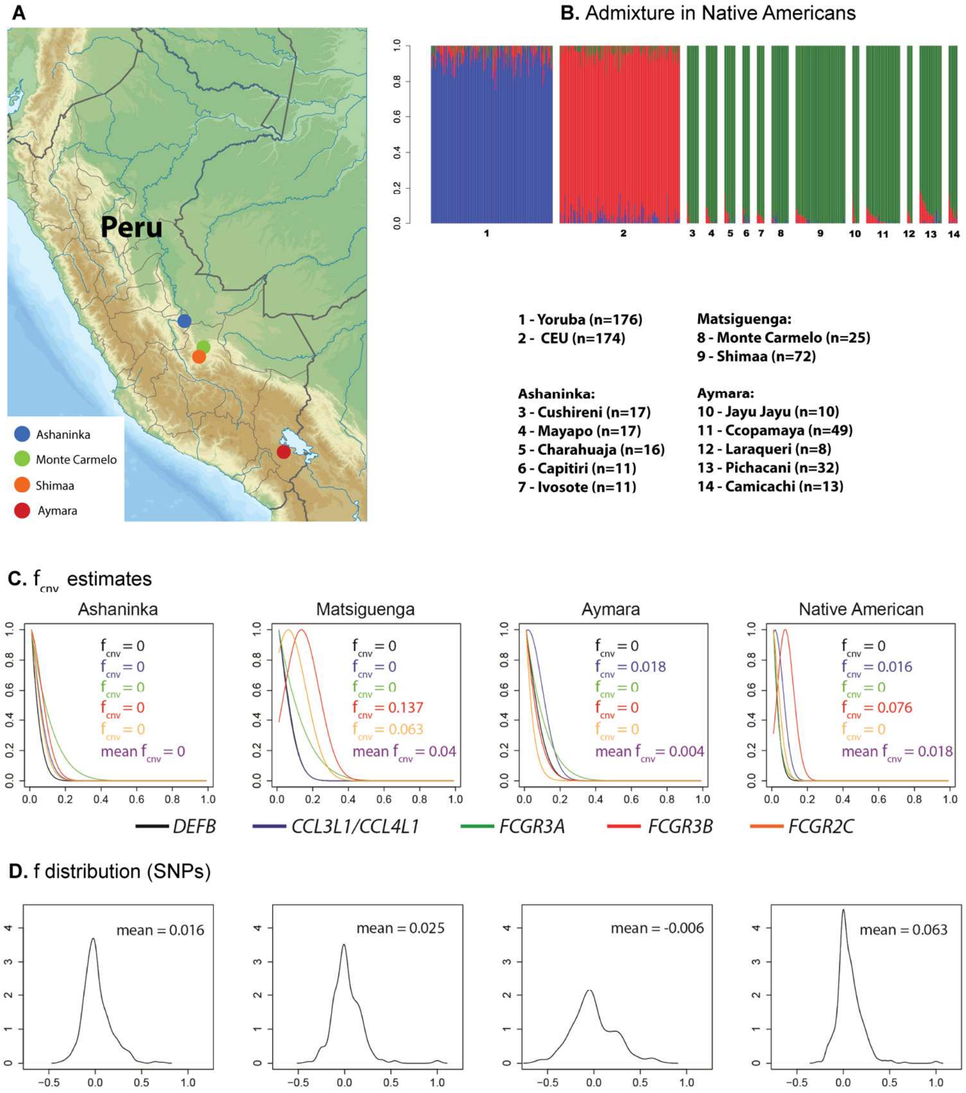


Figure 1. A. Geographic location of the three studied ethnic groups (Ashaninka, Matsiguenga and Aymara); B. Barplot of individual continental ancestry for the populations included in this study and for parental populations from public databases: (1) Yoruba (HapMap YRI, n=176); (2) European ancestry individuals (HapMap CEU, n=174); and the Peruvian populations from the present study. Asha-ninka villages: (3) Cushireni, n=17; (4) Mayapo, n=17; (5) Charahuaja, n=16; (6) Capitiri, n=11; (7) Ivosote, n=11. Matsiguenga villa-ges: (8) Monte Carmelo, n=25; (9) Shimaa, n=72; Aymara villages: (10) Jayu Jayu, n=10; (11) Ccopamaya, n=49; (12) Laraqueri, n=8; (13) Pichacani, n=32; (14) Camicachi, n=13. Ancestry colours: Blue: African; Red: European; Green: Native American. This analysis used 103 ancestry informative markers [19] and was performed with Admixture v.1.2 software; C. Profiled likelihood and maximum likelihood estimation of the population structure parameter fCNV for each multilocus CNV locus. The vertical axis of each locus is stand-ardized according to its maximum likelihood. Native American squares correspond to the entire set of studied individuals from the three ethnic groups. D. Empirical distribution of the f parameters estimated for 695 unlinked SNPs for each ethnic group.

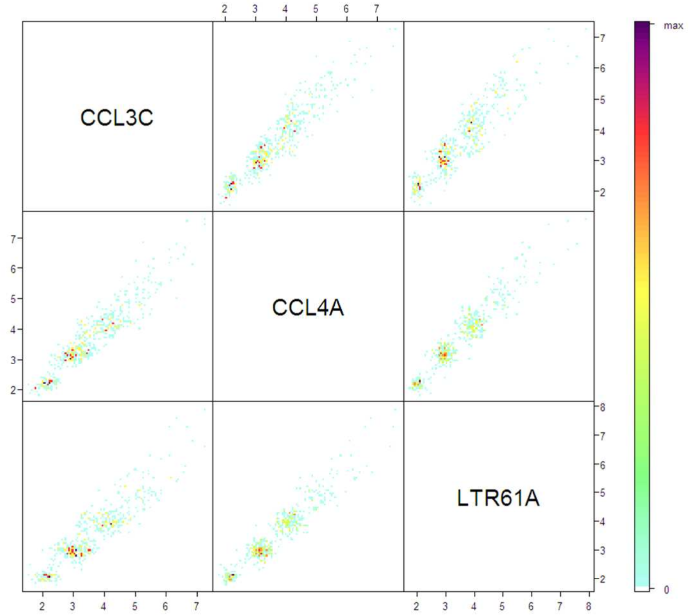
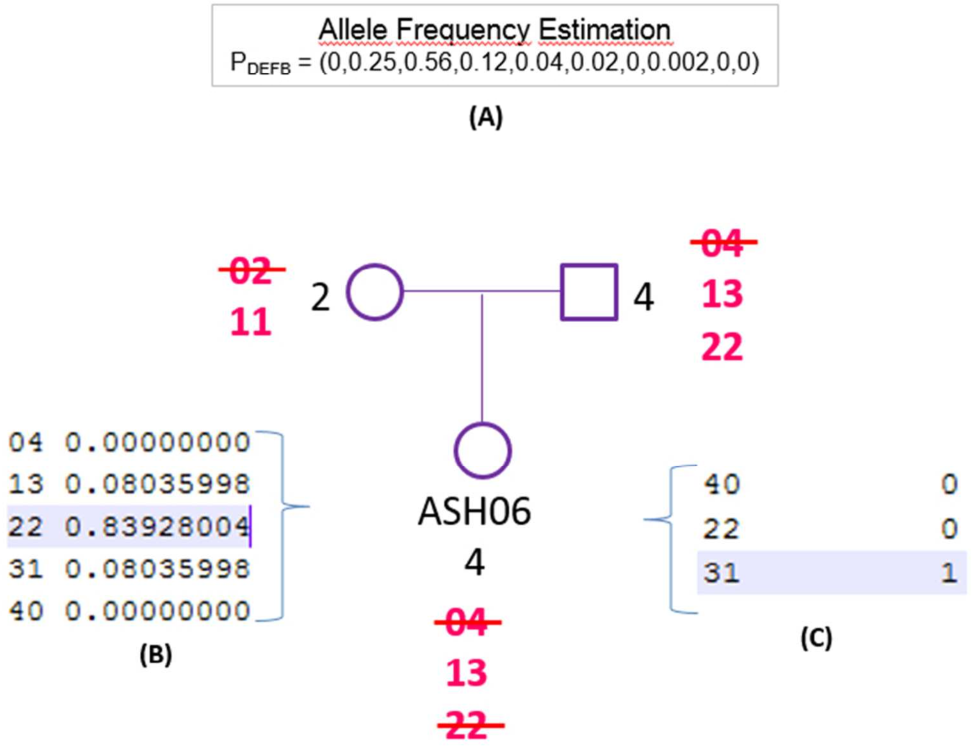


Figure 2. Correlation of copy numbers for the three markers used to infer copy number of the CCL3L1/CCL4L1 region. Dispersion matrix of the normalized copy number values for CCL3C, CCL4A and LTR61A markers for 522 Native American individuals genotyped for the regions CCL3L1/CCL4L1. Colours correspond to point density. The high correlation (Pearson R2>0.91, p<0.001) between the normalized CN values across the pairs of the three amplified loci suggests that in these populations, the region encompassing the CCL3L1 and CCL4L1 genes belongs to the same repetitive unit.

254x190mm (300 x 300 DPI)


Allele Frequency Estimation
P_{DEFB} = (0,0.25,0.56,0.12,0.04,0.02,0,0.002,0,0)
(A)
~~02~~
11
2
4
~~04~~
13
22
ASH06
4
~~04~~
13
~~22~~
(B)
(C)

```
04 0.00000000
13 0.08035998
22 0.83928004
31 0.08035998
40 0.00000000
```

```
40          0
22          0
31          1
```

Figure 3. Improving individual CNV genotype estimation with trio information. (A) vector of allele frequencies for the locus DEFB in the Ashaninka population; (B) individual genotype probabilities for ASH06 without using trio information but conditioning on observed diploid copy number, inferred allele frequencies and fCNV; (C) individual genotype posterior probabilities for ASH06 (i.e., considering parental CNV information).

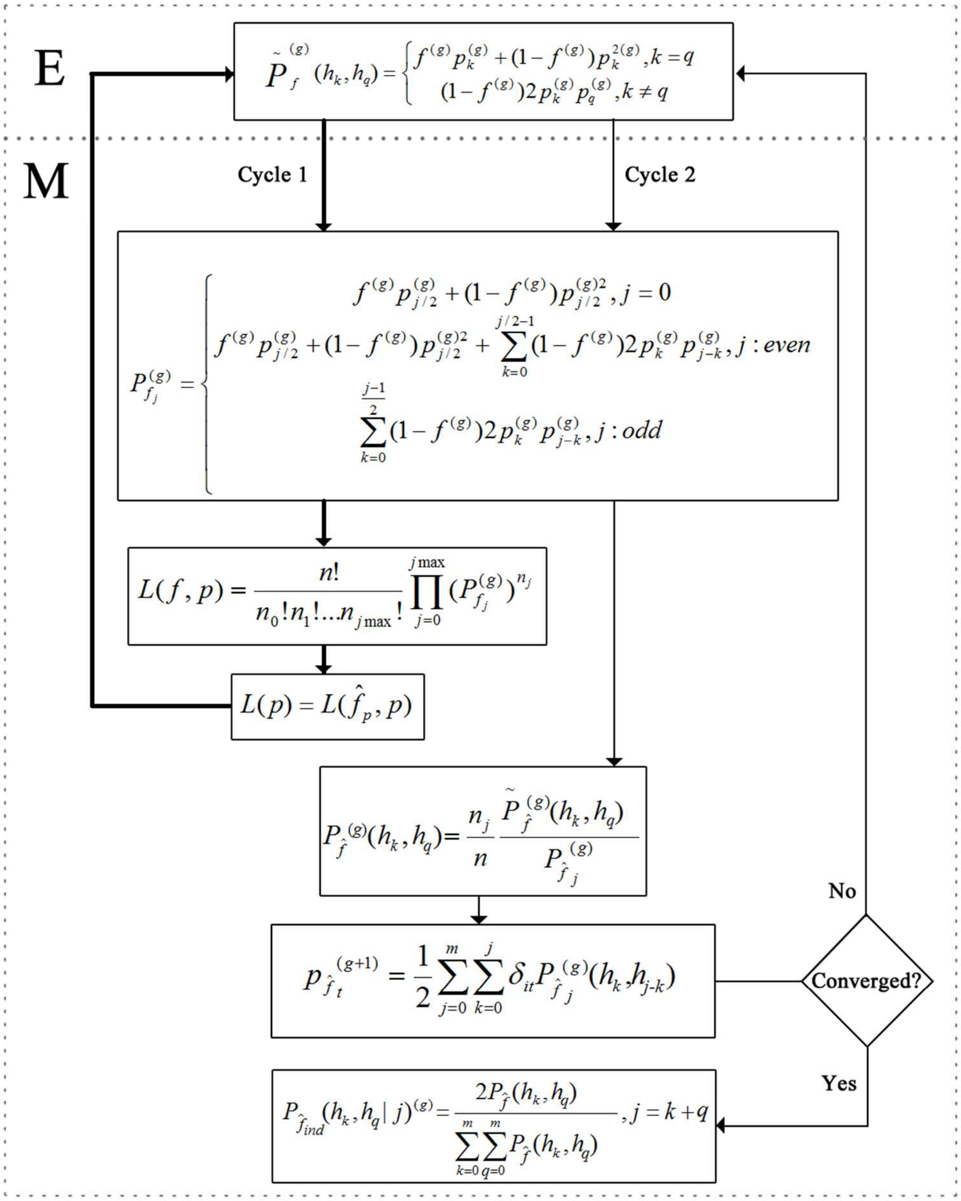

E
$\tilde{P}_f^{(g)}(h_k,h_q)=\begin{cases} f^{(g)}p_k^{(g)}+(1-f^{(g)})p_k^{2(g)},k=q \\ (1-f^{(g)})2p_k^{(g)}p_q^{(g)},k\neq q \end{cases}$
M
Cycle 1
Cycle 2
$P_{f_j}^{(g)}=\begin{cases} f^{(g)}p_{j/2}^{(g)}+(1-f^{(g)})p_{j/2}^{(g)2},j=0 \\ f^{(g)}p_{j/2}^{(g)}+(1-f^{(g)})p_{j/2}^{(g)2}+\sum_{k=0}^{j/2-1}(1-f^{(g)})2p_k^{(g)}p_{j-k}^{(g)},j:even \\ \sum_{k=0}^{\frac{j-1}{2}}(1-f^{(g)})2p_k^{(g)}p_{j-k}^{(g)},j:odd \end{cases}$
$L(f,p)=\frac{n!}{n_0!n_1!...n_{j\max}!}\prod_{j=0}^{j\max}(P_{f_j}^{(g)})^{n_j}$
$L(p)=L(\hat{f}_p,p)$
$P_{\hat{f}}^{(g)}(h_k,h_q)=\frac{n_j}{n}\frac{\tilde{P}_{\hat{f}}^{(g)}(h_k,h_q)}{P_{\hat{f}_j}^{(g)}}$
No
$p_{\hat{f}_t}^{(g+1)}=\frac{1}{2}\sum_{j=0}^{m}\sum_{k=0}^{j}\delta_{it}P_{\hat{f}_j}^{(g)}(h_k,h_{j-k})$
Converged?
Yes
$P_{\hat{f}_{ind}}(h_k,h_q|j)^{(g)}=\frac{2P_{\hat{f}}(h_k,h_q)}{\sum_{k=0}^{m}\sum_{q=0}^{m}P_{\hat{f}}(h_k,h_q)},j=k+q$

Figure 4. CNVice algorithm. hk and hq are alleles, where k and q represent the number of copies of a gene in each chromosome. The observed diploid copy number, or the total copy number, is denoted by j, where j = k + q, k and q being natural numbers. The popula-tion structure parameter is f (or fCNV), and the probability of an allele to be k is pk. For example, an individual with genotype (h3,h1) for a locus L has 3-copies of the allele in one chromosome, 1 copy on the other, and diploid copy number 4; the probability of the allele to be h3 in the population is given by p3 and of the allele h1 is given by p1. n is the total number of individuals in the sample, and g is the number of iterations. δit is a binary variable that indicates whether the t allele is present in genotype i=(hk,hq). Steps: E: Expectation, M: Maximization.