BMC Bioinformatics

SOFTWARE Open Access

Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm

José Ignacio Lucas-Lledó[1,2*], David Vicente-Salvador[1], Cristina Aguado[1] and Mario Cáceres[1,3]

Abstract

Background: Population genetics and association studies usually rely on a set of known variable sites that are then genotyped in subsequent samples, because it is easier to genotype than to discover the variation. This is also true for structural variation detected from sequence data. However, the genotypes at known variable sites can only be inferred with uncertainty from low coverage data. Thus, statistical approaches that infer genotype likelihoods, test hypotheses, and estimate population parameters without requiring accurate genotypes are becoming popular. Unfortunately, the current implementations of these methods are intended to analyse only single nucleotide and short indel variation, and they usually assume that the two alleles in a heterozygous individual are sampled with equal probability. This is generally false for structural variants detected with paired ends or split reads. Therefore, the population genetics of structural variants cannot be studied, unless a painstaking and potentially biased genotyping is performed first.

Results: We present *svgem*, an expectation-maximization implementation to estimate allele and genotype frequencies, calculate genotype posterior probabilities, and test for Hardy-Weinberg equilibrium and for population differences, from the numbers of times the alleles are observed in each individual. Although applicable to single nucleotide variation, it aims at bi-allelic structural variation of any type, observed by either split reads or paired ends, with arbitrarily high allele sampling bias. We test *svgem* with simulated and real data from the 1000 Genomes Project.

Conclusions: *svgem* makes it possible to use low-coverage sequencing data to study the population distribution of structural variants without having to know their genotypes. Furthermore, this advance allows the combined analysis of structural and nucleotide variation within the same genotype-free statistical framework, thus preventing biases introduced by genotype imputation.

Keywords: Structural variation, Population genetics, Maximum likelihood, Reference bias, Genotyping

Background

Ongoing efforts to discover genetic variation in humans and other species are yielding long lists of known variants [1-3]. The discovery of genetic variation is always the first step of population genetics or association studies. Once structural or nucleotide variation is revealed, individuals not present in the original sample are usually genotyped, for subsequent studies. Genotyping individuals is much easier than discovering new variants, because fewer loci have to be tested, and the prior probability of there being an alternative allele is higher than in sites not known to be variable. However, the presence of sequencing and mapping errors, and undersampling at heterozygous sites demand high coverage in order to infer individual genotypes accurately. Accurate genotypes are the basic information upon which most classic methods of population genetics depend. Our reliance on classic methods and the convenience of low-coverage data are such

*Correspondence: lucas.lledo@igb-berlin.de
[1]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
[2]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), 12587 Berlin, Germany
Full list of author information is available at the end of the article

that the population-level structure of the variation is being used to improve the genotype calls on low-coverage data [4,5], which are then expected to help us understand the population-level structure of the variation. This circularity prevents, for example, the use of genotypes imputed on the bases of known patterns of linkage disequilibrium to infer recombination rates. The problem stems from considering the genotypes as the ultimate result of a research program. For most studies, other than personalized medicine, the genotypes are just the means to gain insight into population-level processes: association between genotypes and phenotypes, history of migration and admixture, patterns of recombination, natural selection, etc.

When studying structural variants, there is too much to loose from relying on imputed genotypes. Polymorphic structural variants contribute to phenotypic diversity and disease susceptibility in humans [6], and other species [7-11]. The population genetics of structural variants is both a classic field, and, thanks to the new sequencing techniques, an emergent research venue [12,13]. In the last years, many programs have been developed to identify structural variants from sequence data, using diverse signatures from split reads [14,15], read depth [16,17], paired reads [18,19], or a combination thereof [20,21]. Among the hotest topics that await to harvest full benefits from the high-throughput sequencing technologies, is the interplay between structural and nucleotide variation. The nucleotide variation linked to structural variants can inform of their evolutionary history and their effects on fitness, and it can also reveal to what extent structural variants affect recombination patterns. However, these questions cannot be addressed with SNP genotypes imputed on the bases of assumed patterns of linkage disequilibrium.

An alternative to genotype imputation is to study the population-level structure of the variation with new methods that take genotype uncertainty into account in the analysis. This is what has been proposed before for single-nucleotide variation [22-24]. The idea is to calculate the genotype likelihoods at every site, instead of calling the genotypes. Genotype likelihoods can be used to obtain unbiased estimates of allele frequencies or site-frequency spectra, and likelihood-ratio tests can be used to address population-level hypotheses, such as Hardy-Weinberg equilibrium, population differentiation, linkage disequilibrium, genotype-phenotype association, etc [24,25]. However, this promissing approach is not applicable yet to structural variants, mostly because existing implementations assume even sampling probabilities of the two alleles from a heterozygous sample. While this assumption is a reasonable simplification for the analysis of SNP data, it is not for the analysis of structural variants, where one of the alleles may be observed much more frequently than the other in a heterozygous sample [26]. One source of bias is the conservative alignment of reads to the reference genome. By default, mapping tools are optimized for the discovery phase, when the prior of there being a structural variant is very low. Therefore, aligners favor concordant mappings, that is, those that do not report the variation [27-29]. Another source of allele sampling bias is the different detectability of the alleles. For example, using single reads, the presence of a polymorphic insertion may be revealed by any read mapping to either of the two breakpoints, or even to its inner sequence, if unique and known; while the absence of the inserted fragment can only be positively attested by reads mapping on a single breakpoint. Also, repetitive sequences are frequently present around a structural variant, and they can impair the detection of one allele more than the other. For example, when the sequence of a transposable element is broken by an inversion in one of the breakpoints. In summary, allele sampling bias is the rule, rather than the exception, when genotyping structural variation with sequence data; and the bias can be orders of magnitude higher than in the case of SNPs (see below). In addition, the few existing tools able to report genotype likelihoods [5,24] require a bam file [30] for their calculation, which is not designed for structural variation at all.

Here, we present *svgem*, a simple and flexible command-line application to infer genotype likelihoods and allele frequencies from counts of reference and alternative alleles, appropriate for structural variation data, with an arbitrarily high reference (or alternative) bias. This program is not concerned with the discovery phase of the structural variants, but with the post-discovery analyses. The only assumptions made about the type of variation that can be analysed are that it is bi-allelic, and that the two alleles can be distinguished from sequencing data. Thus, simple insertions, deletions, inversions, and translocations of any length can be analysed with *svgem*, while multiple, overlapping rearrangements, with more than two possible alleles are not. The sampling bias must be known in advance, and passed as a parameter. We offer some guidelines on how to estimate it, and explain how to test for Hardy-Weinberg equilibrium. Also, the ploidy of the samples is taken into account to infer genotype likelihoods and allele frequencies in sex chromosomes properly.

Implementation

Overview and requirements

The program *svgem* implements an expectation-maximization algorithm in C++. The source code is freely available in Additional file 1, and in [31]. It runs from the command line, analyses one structural variant at a time, and it takes as input a text file with three (optionally, four) columns: a sample identifier and the numbers of times the reference and the alternative alleles have been observed.

The optional fourth column represents the ploidy, for the case when the sample is composed of a mixture of males and females, and the variant being analysed is on the sex chromosome.

In order to make *svgem* compatible with virtually any way of counting the observations of reference and alternative alleles, only two qualities are distinguished: the average quality of the reference counts, and the average quality of the alternative counts. These are analogous to the base qualities of a sequenced read, and they can be passed as parameters to the program, in terms of the frequencies of erroneous counts, or estimated from the data. Alternatively, if the individual quality (probability of error) of each observation is known, the expected number of true counts or 'effective' number of times the alleles are observed, instead of the raw counts, can be used. This approach proved to be very useful in SNPTools [5].

The allele sampling bias is represented by one parameter, λ, defined as the odds of sampling the reference allele from a diploid, heterozygous genotype. Even though in some cases it could be estimated from the data, *svgem* requires λ to be passed as a parameter, in order to save degrees of freedom. Otherwise, it would be impossible to get accurate estimates of extreme allele frequencies in the presence of errors. Plus, because *svgem* is not designed to discover variants, but to analyse already known variants, it is fairly easy to estimate λ. There are two different situations. First, if the exact sequences of both structural alleles are completely known, the alleles can be distinguished by the single reads mapping specifically to one of them. All possible reads from the informative regions of both alleles can be extracted, as if they came from a heterozygous sample, and mapped back to the reference genome and to the alternative allele. Then, λ would be estimated as the ratio between the number of reference reads that map uniquely to the reference allele, and the number of alternative reads that map uniquely to the alternative allele.

Second, if the exact sequences of the structural alleles are not known (e.g., imprecise breakpoints), they must be distinguished by the pattern of paired-end mappings: concordant for the reference, and discordant for the alternative (methods based on the depth of coverage, usually requiring high coverage, are not considered here). Still, the most likely version of the alternative allele could be composed, and used to extract paired end reads from it. In this case, the complete extraction of all possible paired-end reads is not feasible, but extensive simulations can be done with available programs, such as *wgsim* [24], or *ART* [32]. If the same coverage is simulated in both alleles, λ can be estimated as the ratio of the number of reference paired ends that map concordantly to the number of alternative paired ends that map discordantly. To overcome the imprecision of the breakpoints, several alternative alleles could be simulated and averaged. Alternatively, if a subset of individuals are known to be heterozygous by other means (e.g., PCR evidence, or higher sequencing coverage), the ratio between their pooled numbers of reference and alternative reads can also be used as an estimate of λ. Inaccurate estimates of λ have a mild impact on genotype likelihoods, and they are always preferred to the default value of $\lambda = 1$, as long as they are closer to the true value of the allele sampling bias (see below).

Implementation of the expectation-maximization algorithm

Following the notation in [24] (see Table 1), we refer to a genotype by its number of reference alleles, $g \in \{0, 1 \ldots m\}$, where m is the ploidy, usually 2. We assume that variants are biallelic, so that $m - g$ is the number of alternative alleles in the genotype. Table 2 shows the likelihoods of the three diploid genotypes. The main difference with respect to Li's equation 2 [24] is the heterozygous genotype, the likelihood of which depends here on the allele sampling bias. If $\lambda = 1$, and allowing for all the observations of the same allele to have the same quality, the difference vanishes (see Additional file 2). The likelihoods of the hemizygous genotypes Alt/0 and Ref/0 are the same as those of the respective homozygous genotypes.

Treating the genotypes as missing values, we implement an expectation-maximization (EM) method to estimate either the alternative allele frequency, ψ, under the assumption of Hardy-Weinberg equilibrium, or the genotype frequencies ψ_g (with $g \in \{0, 1, 2\}$ for diploids) or ϕ_g (with $g \in \{0, 1\}$, for hemizygous individuals), and eventually the proportions of errors among reference (ϵ_r) and alternative (ϵ_a) counts. Note that ψ_g is the frequency of genotype g among diploids, and ϕ_g is the frequency of genotype g among hemizygous individuals. The EM algorithm is an iterative estimation of the parameters that gets closer to the maximum likelihood estimates in every iteration. Additional file 2 gives a summary of how the standard formulation of the EM algorithm

Table 1 Notation

k	Total number of allele observations, or counts, in one individual.
l	Number of times the reference allele is observed in one individual ($l \leq k$).
m	Ploidy.
g	Number of reference alleles in the genotype ($g \leq m$).
λ	Allele sampling bias in heterozygous individuals.
ϵ_r	Frequency of erroneous counts among reference counts.
ϵ_a	Frequency of erroneous counts among alternative counts.

Table 2 Likelihoods of the three diploid genotypes ($m = 2$)

Genotype (g)	Likelihood
0	$\epsilon_r^l (1-\epsilon_a)^{k-l}$
1	$\left(\frac{1}{1+\lambda}\right)^k (\epsilon_r + \lambda - \lambda\epsilon_r)^l (1-\epsilon_a+\lambda\epsilon_a)^{k-l}$
2	$(1-\epsilon_r)^l \epsilon_a^{k-l}$

is used to derive the next values of these parameters, namely:

$$\psi^{(t+1)} = \frac{2D_0^{(t)} + D_1^{(t)} + H_0^{(t)}}{2(D_2^{(t)} + D_1^{(t)} + D_0^{(t)}) + H_0^{(t)} + H_1^{(t)}}$$

$$\psi_g^{(t+1)} = \frac{D_g^{(t)}}{D}$$

$$\phi_g^{(t+1)} = \frac{H_g^{(t)}}{H}$$

$$\epsilon_a^{(t+1)} = \frac{A_2^{(t)}}{A_0^{(t)} + A_2^{(t)}}, \qquad \text{if} \quad \lambda = 1$$

$$\epsilon_r^{(t+1)} = \frac{R_0^{(t)}}{R_0^{(t)} + R_2^{(t)}}, \qquad \text{if} \quad \lambda = 1$$

In the equations above, $D_g^{(t)}$ is the t^{th} estimate of the total number of diploid individuals with genotype g, and $H_g^{(t)}$ is the t^{th} estimate of the total number of hemizygous individuals with genotype g. That is, they are the summations of the posterior probabilities of genotype g over the respective kind of individuals. D and H are the total number of diploid and hemizygous individuals, respectively, where $D+H = N$. $A_g^{(t)}$ is the t^{th} estimate of the total number of alternative counts coming from hemizygous and homozygous individuals for either the alternative ($g = 0$) or the reference ($g = 2$) allele. Finally, $R_g^{(t)}$ is the t^{th} estimate of the total number of reference counts that come from hemizygous and homozygous individuals for either the alternative ($g = 0$) or the reference ($g = 2$) allele.

When there is sampling bias in heterozygous individuals, $\lambda \neq 1$, and the next values of the proportions of errors among reference (ϵ_r) and alternative (ϵ_a) counts are the results of two quadratic equations (Additional file 2). In practice, it is assumed that the erroneous counts are a minority, and the program halts when $\epsilon_r \geq 0.5$ or $\epsilon_a \geq 0.5$. This can prevent the correct estimation of extreme allele frequencies in the presence of erroneous counts, as should be expected.

Output and applications

The output includes: maximum likelihood estimates of the parameters mentioned above, the likelihood of such estimates, the genotype likelihoods of all individuals, and the posterior probabilities of the genotypes of all individuals. The main purpose of *svgem* is to obtain unbiased estimates of allele and genotype frequencies, which are fundamental parameters in population genetics. From these estimates, several other population parameters can be estimated. The maximum likelihood estimate of the frequency of the heterozygous genotype ($\hat{\psi}_1$), estimated without assuming Hardy-Weinberg equilibrium, is a direct estimate of heterozygosity. An estimate of the inbreeding coefficient follows from comparing $\hat{\psi}_1$ with the expected frequency of heterozygous individuals under Hardy-Weinberg equilibrium: $\hat{F} = 1 - \hat{\psi}_1/(2\hat{\psi}(1-\hat{\psi}))$ (where $\hat{\psi}$ is the maximum likelihood estimate of the alternative allele frequency). The fixation index, F_{ST}, which measures genetic differentiation among populations, is also readily estimated from allele frequencies [33-35]. A test for Hardy-Weinberg equilibrium (HWE) can be performed by running *svgem* with and without the equilibrium assumption, and comparing the log-likelihoods of the estimated frequencies. Twice the difference between the log-likelihoods must be compared with a χ^2 distribution with 1 degree of freedom, if all individuals are diploid, or with 2 degrees of freedom if the frequencies of hemizygous genotypes are also being estimated.

Some analyses that used to require accurate knowledge of individual genotypes can be performed now using only genotype likelihoods. For example, it is possible to estimate the linkage disequilibrium between pairs of variants using genotype likelihoods, instead of individual genotypes [24]. At the end of the next section, we show how to estimate the linkage disequilibrium between a structural variant and the SNPs around it, without the biases typically associated with genotype imputation.

It is also possible to run genetic association tests from genotype likelihoods, without knowing the exact genotype of the individuals [24]. Associations between phenotypes and genetic variants are a significant difference in allele frequency between two samples (cases and controls), and they are routinely searched along the human genome to infer the causal variants of diseases. To compare the allele frequency of a variant between two samples, *svgem* must be run three times: once in each sample separately, and once in the joint dataset. Lets call ℓ_a and ℓ_b the log-likelihoods of the two independent estimates for samples a and b. The total log-likelihood of the hypothesis of two different frequencies is just the sum of the log-likelihoods of the two samples: $\ell_1 = \ell_a + \ell_b$. The log-likelihood of the hypothesis of one common allele frequency, ℓ_0, is obtained from the run on the joint data set. Because the two hypotheses are nested, the application of a likelihood ratio test is justified. Thus, if the null hypothesis of a common allele frequency is true, the statistic $2(\ell_1 - \ell_0)$ is expected to follow a χ^2 distribution with as many degrees of freedom as additional parameters the most complex model has, which is 1 in this case (see [36], page 137).

For other analyses, that may require knowledge of individual genotypes, we recommend using the genotype with the highest posterior probability, which is more accurate than the most likely genotype, because posterior probabilities take into account the information of the genotype frequency in the population (see below). *svgem* uses the maximum likelihood estimates of allele ($\hat{\psi}$) or genotype ($\hat{\psi}_g$) frequencies, and the genotype likelihoods ($\mathcal{L}(g)$) to calculate the genotype posterior probabilities, $P(g \mid \text{data})$:

$$P(g \mid \text{data}) = \begin{cases} \frac{\mathcal{L}(g)P(g|\hat{\psi})}{\sum_{g=0}^{m} P(g|\hat{\psi})\mathcal{L}(g)} & \text{if HWE is assumed} \\ \frac{\mathcal{L}(g)\hat{\psi}_g}{\sum_{g=0}^{m} \hat{\psi}_g \mathcal{L}(g)} & \text{if it is not} \end{cases}$$

Performance

Although *svgem* is designed to analyse one variant at a time, more than one variant can easily be analysed in multiple parallel runs. For the purpose of detailed population genetic analyses of specific variants of interest, *svgem* performs well, since its typical run time is always a tiny fraction of the time usually needed to obtain the allele counts. In absolute terms, it can analyse > 1000 individuals in, at most, a few seconds, in a standard PC. However, the run time is variable, just as in any expectation-maximization algorithm, and convergence may take longer if the information content of the input data set is limited.

Results and discussion

Estimates of λ from real structural variants

In order to assess the expected range of values of the λ parameter, we downloaded the BreakDB database [37], last accessed on January 26th 2014, and built a library with the known sequences of both alleles of 568 structural variants (54 inversions, 161 insertions, and 353 deletions). Then, we simulated the exhaustive sequencing of both alleles with single-end, 100-bp reads. Next, we mapped the sequenced reads first to the library built with the sequences of the reference and the alternative alleles, and afterwards to the whole reference genome (HG18), in order to discard any non-specific read. For this purpose, we used the pipeline BreakSeq [38], with minor modifications. We removed from BreakSeq a filter that required the reads to map on the breakpoints, in order to be able to use the inserted or deleted sequences as evidence of the presence or the absence of an insertion or a deletion. Finally, we counted how many reads mapped specifically on the reference or on the alternative allele, and estimated λ as the ratio of the two counts.

The observed, finite values of λ ranged between 0.002 and 10000 (11 inversions, and 158 deletions had infinite λ, meaning that one of the alleles was not detectable by sequencing with single-end, 100 bp reads, due to the presence of repeats around the breakpoints). Figure 1 shows

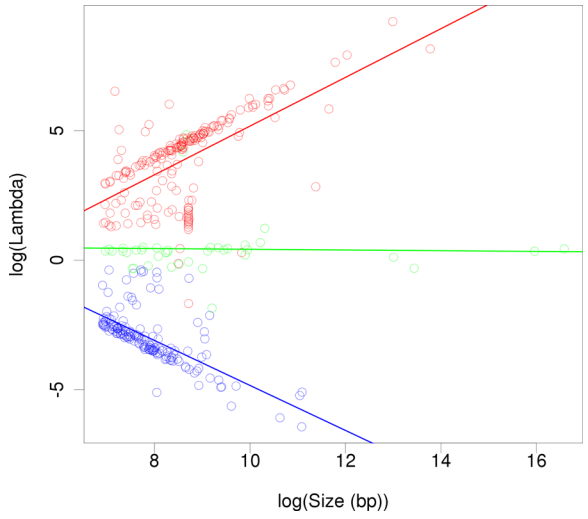

Figure 1 Estimates of the allele sampling bias, λ, of real variants. The allele sampling bias of 195 deletions (red), 43 inversions (green), and 161 insertions (blue) is plotted against their lengths, in logarithmic scales. The bias, λ, was estimated by simulating the sequencing of both alleles with 100 bp, single-end reads. Lines are linear regressions.

that the length of the insertions and deletions greatly contributes to the value of λ, as expected. Because the inserted or deleted sequence is known, and used as evidence of the presence of the longest allele (the reference in a deletion, the alternative in an insertion), longer insertions or deletions produce more unbalanced allele observations, favouring the reference allele in a deletion (high λ) or the alternative one in an insertion (low λ). The local sequence around and within the variant, and the method used to detect the alleles (read length, whether single or paired-end) must also influence the exact value of λ. However, the linear regressions between the logarithm of λ and the logarithm of the size of the insertion or deletion (in base pairs) allow for a rough, first approximation to λ, at least when detecting insertions or deletions with single-end sequenced reads: $\log(\lambda) = -4.22 + 0.94\log(\text{size})$ for deletions with respect to the reference allele (adjusted $R^2 = 0.37$), and $\log(\lambda) = 3.84 - 0.87\log(\text{size})$ for insertions (adjusted $R^2 = 0.44$). In the case of inversions, a $\lambda = 1$ is a fair assumption, in the absence of additional information. This approximations are not expected to hold when detecting structural variants flanked by segmental duplications or other repeats.

Analysis of simulated data

We run some simulations to test *svgem*. In the artificial datasets, genotypes followed Hardy-Weinberg equilibrium, with allele frequencies between 0.01 and 0.99, and the allele counts were sampled with a simulated error rate of 0.005. Coverage was Poisson-distributed, in order

to introduce variation in coverage among individuals, although the exact distribution is irrelevant for the calculation of genotype likelihoods. In all, 100 simulations were run with the same parameter values.

First, we checked that *svgem* is able to get unbiased estimates of the allele frequency with low coverage data, in the absence of any allele sampling bias ($\lambda = 1$). Figure 2 shows how estimates based on a sample of 100 individuals are accurate and as precise as they can be with mean coverages ranging from 0.2 to 4. The only biased estimates correspond to alternative allele frequencies lower than 0.01 (or higher than 0.99, not shown), targeted with sequencing coverages lower than 0.5. Not surprisingly, as the number of parameters to estimate increases, the precision and the accuracy drop. The frequencies of the three genotypes can still be estimated without bias in most cases, if the mean coverage is higher than 2 (Figures S2–S4 in Additional file 2). When comparing the true genotypes of all individuals simulated with the most likely and with the most probable genotypes (Figure 3), two results become aparent: 1) the benefit of using posterior probabilities, instead of just likelihoods (which do not require the EM algorithm), is higher when the coverage and the minor allele frequency are lower; and 2) applications that require accurate genotypes should use coverages higher than 4, unless the minor allele frequency is always very low. The very high levels of genotype errors observed when the minor allele frequency is high and the coverage is low are an instrinsic problem of the limited amount of data available to infer the genotype. Even if the allele frequency and the allele sampling bias were known with accuracy, up to 50% of the genotypes predicted by maximum posterior probability are expected to be wrong when the coverage is 1 and the minor allele frequency is 0.5 (Figure S5 in Additional file 2). The fact that allele and genotype frequencies can be estimated accurately under rampant uncertainty of individual genotypes strongly encourages the use and further development of genotype-free methods, that take full advantage of low-coverage sequencing data.

Next, we checked *svgem* performance with different sample sizes from 10 to 1000 and a fixed SV frequency of 0.5, which is the one with higher sampling variance. Figure S1 in Additional file 2 shows that smaller samples, with a mean coverage of 4, also yield unbiased estimates

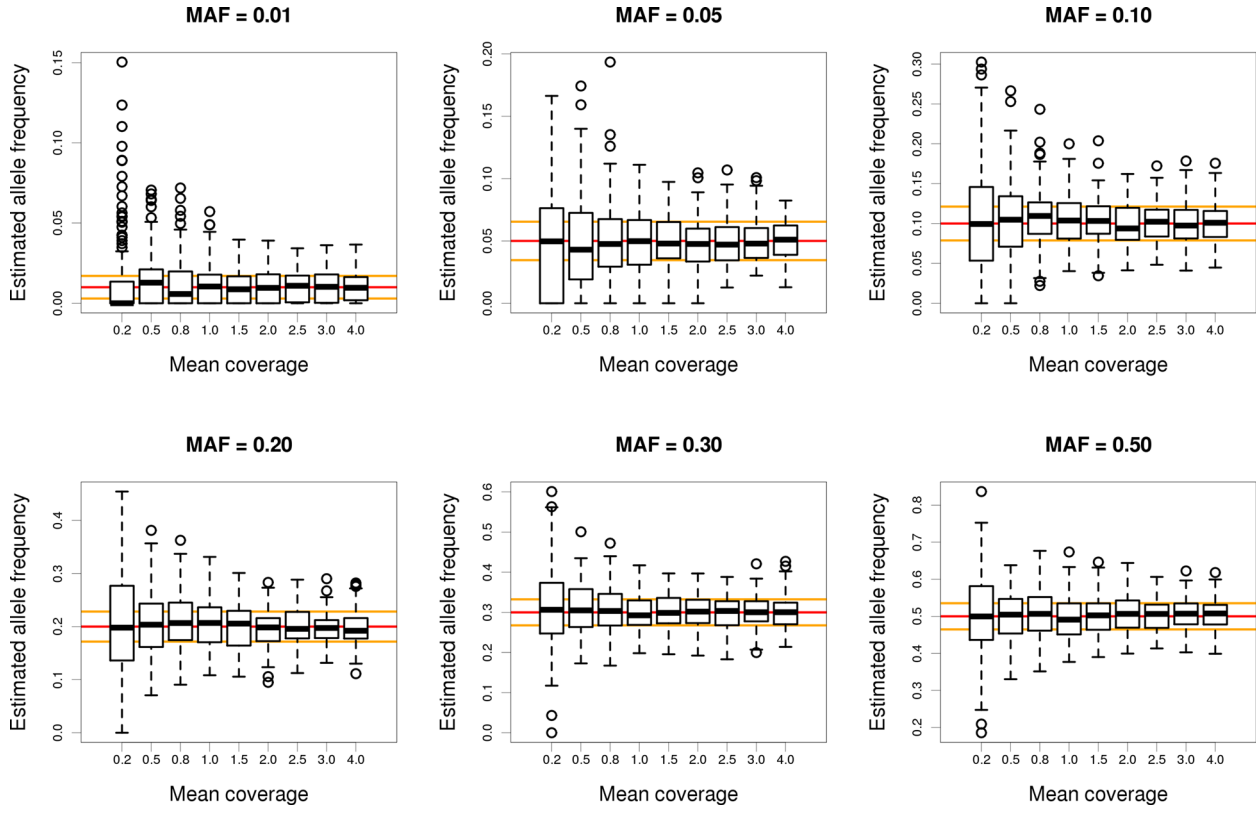

Figure 2 Accuracy and precision of allele frequency estimates with low-coverage data. Estimates obtained by *svgem* from random samples of 100 diploid individuals, from populations with alternative allele frequencies between 0.01 and 0.5 (upper label of each plot; MAF, minor allele frequency), for mean coverage depths between 0.2 and 4. Each combination of allele frequency and depth of coverage was simulated 100 times. The red line indicates the true allele frequency, and the orange lines indicate the expected standard error of the true frequency, from a sample of 200 chromosomes.

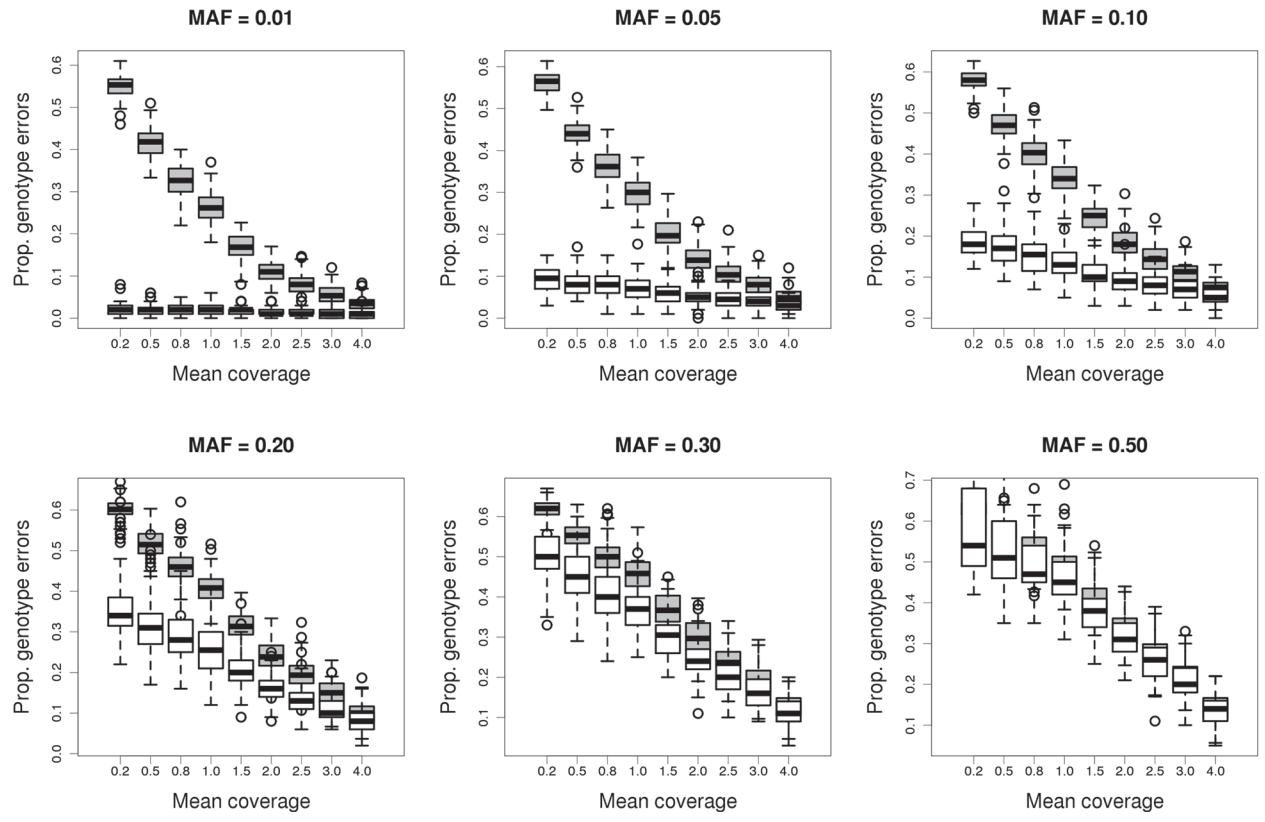

Figure 3 Accuracy of genotype calls with low-coverage data. Proportion of most likely (grey) or most probable (white) genotypes that do not match the true genotype among 100 individuals, in a series of simulations with different values of allele frequency and coverage. The minor allele frequencies (MAF) are shown above each plot. Each box corresponds to 100 simulations.

with a precision comparable to the expected under accurate knowledge of genotypes.

Finally, we prove that arbitrarily high reference bias does not deviate the estimates from the true values, if *svgem* is informed of the bias. Figure 4 represents the accuracy and the biases of the estimates of the allele frequency for several combinations of the true (λ) and the estimated ($\hat{\lambda}$) values of the allele sampling bias. The estimates are always unbiased if $\hat{\lambda} = \lambda$, as expected. Interestingly, the estimates are also unbiased when λ is very low (≤ 0.01) or very high (≥ 100), and $\hat{\lambda}$ is even lower, or even higher, respectively. This implies that extreme values of λ can be effectively approximated by a wide range of values. The reason of this nice property is that at low coverage, the outcome of an extreme λ is always the same: none of a few observations gets to sample the disfavoured allele from a heterozyous sample. It is also worth noticing that rough approximations to λ, in the range between $\lambda/2$ and 2λ produce only minor biases in allele estimates. Moreover, any estimated $\hat{\lambda}$ closer to the real value of λ than the default $\lambda = 1$ will improve the allele frequency estimates.

Figure 5 shows that the difficulty to predict individual genotypes varies in parallel with the difficulty in estimating the allele frequency in the presence of an uncertain allele sampling bias. While the accuracy of allele frequency estimates is mostly independent of the allele frequency (Figure 2), the accuracy in genotype prediction highly depends on the frequency of the genotypes, and therefore on the allele frequency. The heatmaps in Figure 5 represent the observed proportion of true genotypes that did not match the most probable genotype among 500 simulated diploid individuals, with a mean coverage of 4, as a function of the alternative allele frequency and the ratio between the estimated and the true allele sampling bias. The highest genotyping accuracy always happens when the true allele sampling bias is known ($\hat{\lambda} = \lambda$), and the most dramatic increase in genotyping errors happens when the estimated bias deviates in the direction opposite to the true bias.

It is also important to mention that the frequency of erroneous reference or alternative allele counts, ϵ_r and ϵ_a, need to be either known or co-estimated from the data to get accurate estimates of allele or genotype frequencies. An erroneous count is a false observation of an allele, which should not contribute to the estimate of allele frequency. They are assumed to be less frequent than

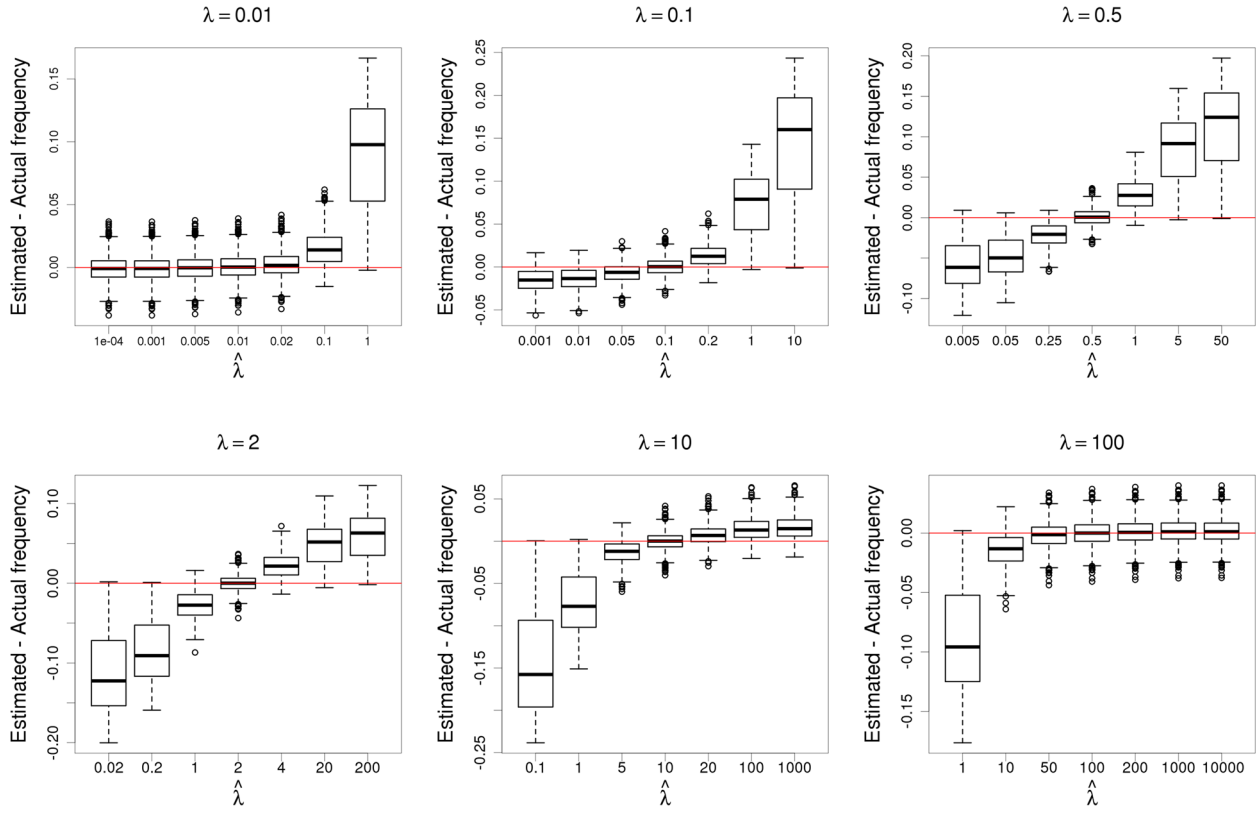

Figure 4 Effect of the allele sampling bias on allele frequency estimates. Accuracy and precision of the allele frequency estimates under biased sampling of alleles from heterozygous individuals, with accurate or inaccurate estimates of the allele sampling bias. The titles above the plots show the real values of λ. A red line is traced where the estimated and current allele frequencies are the same. Each combination of real and estimated λ values is simulated 1000 times, using random allele frequencies. All allele frequency estimates are run here with 1000 individuals sequenced at 4× coverage, in order to reduce the dispersion of the estimates and make the bias caused by the misspecification of λ more visible.

true counts. In practice, the accurate estimation of both the allele (or genotype) frequency and the frequency of erroneous counts is only feasible if there is enough information in the data. As a rule of thumb, when coverage is below 4, or the number of individuals is below 100, *a priori* estimates of ϵ_r and ϵ_a are highly recommended. They can be obtained from simulations, or estimated from a subset of individuals with high coverage, or empirically determined in a subset of homozygous individuals.

Analysis of real data

To test the performance of the algorithm on real data, we used as a model a previously unknown human inversion with simple breakpoints. Analysis of this inversion was part of a larger study to characterize and validate polymorphic inversions in human populations. In particular, the inversion selected (HsInv0201) is a 376 bp inversion in the Chr5q33.1 region, supported by paired-end mapping data of fosmids [39] or small DNA fragments [40,41]. By comparison of the HG18 Human Genome reference assembly [42] with the alternative human assemblies of Celera [43] and HuRef [44], it was found that the inverted allele includes two small deletions flanking the inversion and it was possible to locate the breakpoints (BP) to HG18 position chr5:147533233-147534432 (BP1) and chr5:147534809-147534971 (BP2), which correspond to the sequences deleted in the inverted chromosomes [45]. From there, we extracted 100 nucleotides-long *in silico* probes, in which the sequence change between the two orientations is located exactly in the middle (see Table S3 in Additional file 3). Then, we mapped the reads from 550 individuals from the 1000 Genomes Project [3] on these probes, using the program *BreakSeq* [38], and counted how many matched specifically the reference or the alternative breakpoints. In order to quantify the allele sampling bias, λ, we extracted all possibly informative reads from the 100 nucleotides probes (see Table S3 in Additional file 3), with the same length range as the real reads used (36–100 nucleotides), and used *BreakSeq* again to count how many of them mapped uniquely to either the reference or the alternative breakpoints. A negligible bias ($\lambda = 1$) was found. Erroneous counts were experimentally

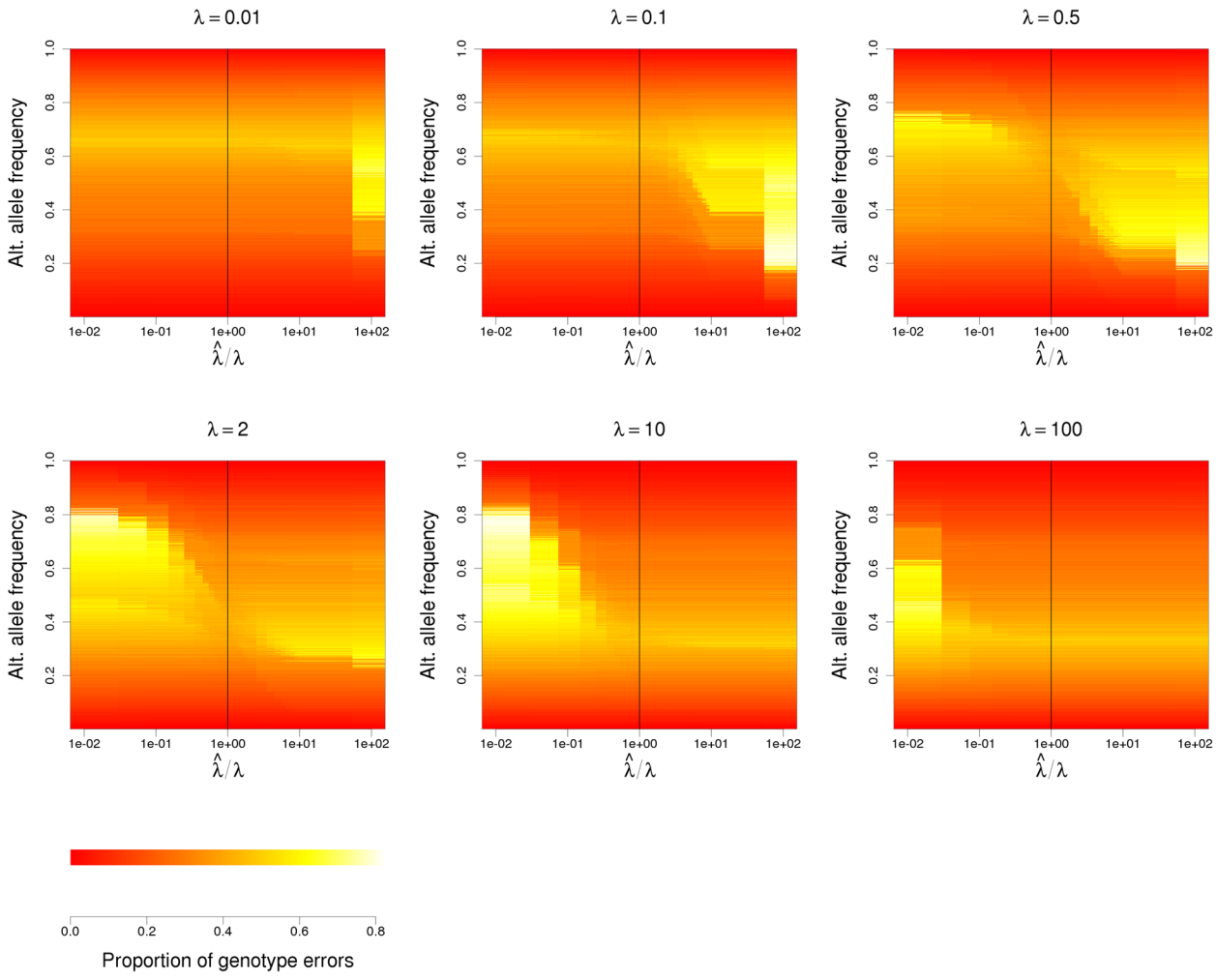

Figure 5 Effect of the allele sampling bias on the accuracy of the most probable genotypes. The heatmap represents the observed accuracy of the most probable genotypes in simulated datasets, as a function of the allele sampling bias (λ, above each plot), the ratio between the estimated and the true allele sampling bias ($\hat{\lambda}/\lambda$), and the frequency of the alternative allele. Each combination of real and estimated λ values is simulated 1000 times, using 500 individuals sequenced at 4× coverage and a random allele frequency.

determined to be also negligible (see below), and their frequencies were set to 1.0×10^{-5}, although several orders of magnitude of variation in this parameter did not alter the results significantly. From the allele counts and from these parameter values, *svgem* estimates a global alternative allele frequency of 0.55, and population-specific frequencies ranging from 0.45 to more than 0.65 (Figure 6). Among related individuals, only the oldest parents of a family were retained for the estimation of population parameters. Using a likelihood ratio test, we prove that Asian and native American populations have a significantly lower frequency of the alternative conformation than African and European populations (p-value $= 5.3 \times 10^{-5}$).

To genotype experimentally the inversion, we used different pairs of primers specific for the reference orientation (A2-B2) or the inverted orientation (A4-C3 and B2-D1; Table S3 in Additional file 3) and carried out duplicate PCRs of each individual, both in simplex and multiplex format. In total, the 270 individuals of the Phase II of the HapMap Project were analyzed, including 90 Yoruba (YRI), 90 from European origin (CEU), 45 Chinese (CHB) and 45 Japanese (JPT), and the PCR results can be accessed in the InvFEST database [45]. Table S4 shows the observed genotypes and the genotype posterior probabilities calculated with *svgem* for the 122 individuals that were both genotyped by PCR and analysed with *BreakSeq* and *svgem*. The alternative allele frequencies determined experimentally or estimated with *svgem* in this subsample were, respectively, 0.545 (standard error 0.045), and 0.541. The most probable genotype determined by *svgem* matched the true genotype in 111 (91%) individuals, with

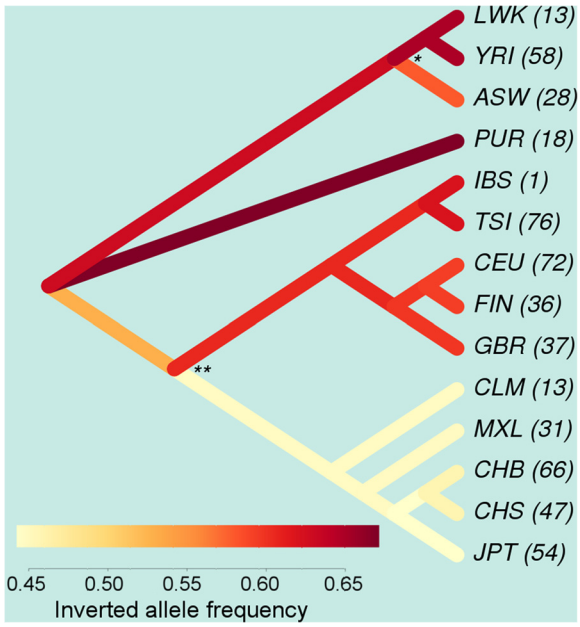

Figure 6 Estimated global frequency of inversion HsInv0201 in human populations. Allele frequencies of a polymorphic inversion on chromosome 5, calculated with *svgem*, are shown as colors on top of the topology of a Neighbor-joining tree of some human populations. The tree was built using the SNP-based, genome-wide, Hudson's F_{ST}s among the populations as distances [3]. In each bifurcation, the overall frequency was assiged to both sister clades if there were no significant differences. Significant changes are labeled with asterisks (*, *p*-value <0.05; **, *p*-value <0.001). The numbers of individuals analysed are shown in parentheses, and population labels are: LWK, Luhya (Kenia); YRI, Yoruba (Nigeria); ASW, African-American (Southwest, US); PUR, Puerto Rican; IBS, Iberian (Spain); TSI, Tuscan (Italy); CEU, Utah residents with Northern and Western European ancestry; FIN, Finnish; GBR, British (England, and Scotland); CLM, Colombian; MXL, Mexican; CHB, Han Chinese (Beijing, China); CHS, Southern Han Chinese; JPT, Japanese.

only 1 error (out of 86) when the coverage is higher than 2. From the allele counts of homozygous individuals, it can be seen that the opposite allele is never observed, confirming that the rate of erroneous counts is negligible.

Once having accurate genotype likelihoods of this short inversion, it is possible to calculate its linkage disequilibrium with nearby SNPs, without having to know the true genotypes, neither of the inversion, nor of the SNPs. This allows the study of the association between structural variants and SNPs without having to rely on imputed genotypes, and without having to exclude SNPs with arbitrary coverage thresholds. The method to calculate the pairwise linkage disequilibrium statistic r^2 from genotype likelihoods is implemented in *bcftools* [24], and requires an input VCF file with genotype likelihoods. We downloaded from the 1000 Genomes Project database a VCF file spanning 7 kb around the inversion, and manually combined it with the inversion itself, represented by two punctual variants at the positions of its breakpoints. Figure 7 shows how the two breakpoints are correctly determined to be in perfect linkage disequilibrium between them, and how the patterns of linkage among the inversion and the SNPs around it differ between European and Asian populations. Note that linkage disequilibrium estimates, let alone their comparison between populations, would be biased if imputed genotypes had been used, because imputation already assumes some linkage disequilibrium, not always measured in the population of interest.

Conclusions

The development of methods to discover structural variants in individual genomes is giving way to population-level analyses. The most recently developed discovery

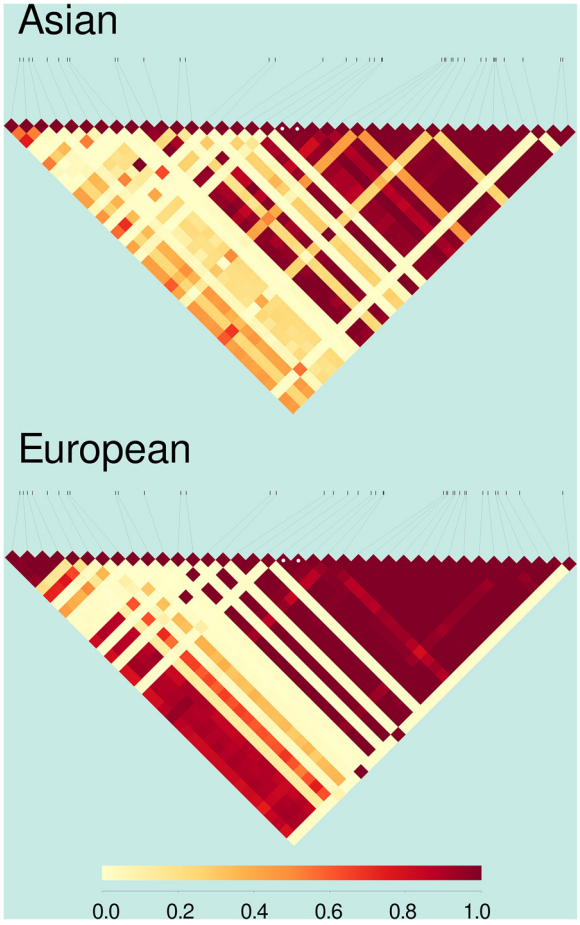

Figure 7 Linkage disequilibrium patterns around inversion HsInv0201 in Asian and European populations. The heatmap represents values of r^2, a measure of linkage disequilibrium between pairs of SNPs or between SNPs and the breakpoints of inversion HsInv0201. The two breakpoints are marked with white spots. The SNP positions span chromosome 5 (hg19) bases 147549663 to 147556698 in Asians, and 147549507 to 147556942 in Europeans.

tools, such as GASVPro [21] or CloudBreak [46], call the genotypes of the individuals analysed, instead of just reporting the variants discovered (a notable exception being ForestSV [20]). However, these individual-based methods require high coverage, and they are oblivious to the information present at the population level. While these methods are still useful in some applications, there is a current demand for efficient ways to analyse low-coverage population genomics data. Most studies still insist in genotyping the individuals, despite of the loss of data caused by arbitrary quality thresholds, and despite the circularity and biases associated with genotype imputation [1,47,48]. The alternative of using likelihood or Bayesian approaches, which take genotype uncertainty into account, is an optimal strategy to explore genetic diversity, since it does not require high coverage per individual, and allows the sequencing of more individuals at the same cost [49]. Not surprisingly, a new method to genotype indels from sequence data in polyploid genomes uses the same approach of likelihood calculation, frequency estimation through an expectation-maximization algorithm, and reporting of posterior probabilities [50]. However, this method does not consider any allele sampling bias, because it targets indels shorter than the reads used to distinguish the alleles. By including allele sampling bias in the genotype likelihood calculation, our program extends the applicability of these methods to the analysis of large structural variation. Furthermore, for the first time nucleotide and structural variation can be analysed in the same statistical framework, without having to rely on the accuracy of the genotypes.

Two of the key features of *svgem* are its simplicity and its few assumptions about the data, which make the program useful for a wide variety of data types. Any bi-allelic structural variant detected by sequenced paired-ends or split reads, including inversions, mobile element insertions, duplications, and deletions, can be analysed by *svgem*. Using simulations, we have shown that estimates of allele or genotype frequencies are accurate, even in the face of rampant allele sampling bias, that usually accompanies the detection of structural alleles. Finally, using data from the 1000 Genomes Project and PCR experiments, we prove its applicability to real data.

Availability and requirements

Project name: svgem.
Project home page: http://grupsderecerca.uab.cat/cacereslab/content/resources.
Operating system(s): Platform independent.
Programming language: C++.
Other requirements: None.
License: GNU General Public License.
Any restrictions to use by non-academics: None.

Additional file

Additional file 1: This is a plain text file containing the source code of *svgem*, in C++.

Additional file 2: Additional text including *svgem*'s manual and some details on how the expectation-maximization algorithm is implemented.

Additional file 3: Tables S3 and S4. Table S3. *In silico* probes and PCR primers. *In silico* probes including the two breakpoints, in both the reference and the inverted conformations, and primers used for PCR validations. **Table S4.** Experimental validation. Allele counts, genotype posterior probabilities obtained with *svgem*, and true genotypes determined by PCR, for inversion Hslnv0201 in 122 individuals from the 1000 Genomes Project.

Abbreviations

EM: Expectation-maximization; kb: kilobase pairs; PC: Personal computer; HWE: Hardy-Weinberg equilibrium; PCR: Polymerase chain reaction; SNP: Single-nucleotide polymorphism; MAF: Minor allele frequency.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JILL wrote the program, run it on the simulated and real data, and coordinated the writing of the manuscript. DV retrieved the allele counts from the 1000 Genomes Project database. CA designed and oversaw the PCR experiments. MC designed the study and supervised all steps. All authors contributed to the writing. All authors read and approved the final manuscript.

Acknowledgements

We thank David Izquierdo for help with the experimental genotyping of the inversion, and Xavier Estivill and Marta Morell for help with the CEU population lymphoblastoid cell cultures. This work was supported by the European Research Council [Starting Grant 243212 (INVFEST) to MC] under the European Union Seventh Research Framework Programme (FP7); and by the Commission for Universities and Research of the Ministry of Innovation, Universities and Enterprise of the Autonomous Government of Catalonia and the Cofund programme of the Marie Curie Actions of the FP7 ['Beatriu de Pinós' postdoctoral fellowship to JILL].

Author details

[1]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain. [2]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), 12587 Berlin, Germany. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain.

Received: 30 October 2013 Accepted: 14 May 2014
Published: 29 May 2014

References

1. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin CY, Luo R, et al.: **Mapping copy number variation by population-scale genome sequencing.** *Nature* 2011, **470**(7332):59–65.
2. Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HYK, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT, 1000 Genomes Project: **A comprehensive map of mobile element insertion polymorphisms in humans.** *PLoS Genet* 2011, **7**(8):e1002236.
3. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**(7422):56–65.
4. Handsaker RE, Korn JM, Nemesh J, McCarroll SA: **Discovery and genotyping of genome structural polymorphism by sequencing on a population scale.** *Nat Genet* 2011, **43**(3):269–276.

5. Wang Y, Lu J, Yu J, Gibbs RA, Yu F: **An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data.** *Genome Res* 2013, **23**(5):833–842.
6. Girirajan S, Campbell CD, Eichler EE: **Human copy number variation and complex genetic disease.** *Annu Rev Genet* 2011, **45:**203–226.
7. Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SAAC, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, Hubner N, Cuppen E: **Distribution and functional impact of DNA copy number variation in the rat.** *Nat Genet* 2008, **40**(5):538–545.
8. Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C, Marquès-Bonet T, Eichler EE, Navarro A: **Copy number variation analysis in the great apes reveals species-specific patterns of structural variation.** *Genome Res* 2011, **21**(10):1626–1639.
9. Berglund J, Nevalainen EM, Molin AM, Perloski M, The LUPA Consortium, André C, Zody MC, Sharpe T, Hitte C, Lindblad-Toh K, Lohi H, Webster MT: **Novel origins of copy number variation in the dog genome.** *Genome Biol* 2012, **13**(8):R73.
10. Muñoz Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N: **Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome.** *Genome Biol* 2013, **14**(6):R58.
11. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Great Ape GenomeProject, Marques-Bonet T, Eichler EE: **Evolution and diversity of copy number variation in the great ape lineage.** *Genome Res* 2013, **23**(9):1373–1382.
12. Corbett-Detig RB, Hartl DL: **Population genomics of inversion polymorphisms in *Drosophila melanogaster*.** *PLoS Genet* 2012, **8**(12):e1003056.
13. Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO: **Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing.** *Genome Res* 2013, **23**(3):568–579.
14. Wang J, Mullighan CG, Easton J, Roberts S, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Heatley SL, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J: **CREST maps somatic structural variation in cancer genomes with base-pair resolution.** *Nat Methods* 2011, **8**(8):652–654.
15. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE: **Detection of structural variants and indels within exome data.** *Nat Methods* 2012, **9**(2):176–178.
16. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res* 2011, **21**(6):974–984.
17. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, Park PJ: **Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion.** *Proc Natl Acad Sci U S A* 2011, **108**(46):E1128–E1136.
18. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6**(9):677–681.
19. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM: **Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome.** *Genome Res* 2010, **20**(5):623–635.
20. Michaelson JJ, Sebat J: **forestSV: structural variant discovery through statistical learning.** *Nat Meth* 2012, **9**(8):819–821.
21. Sindi SS, Önal S, Peng LC, Wu HT, Raphael BJ: **An integrative probabilistic model for identification of structural variation in sequencing data.** *Genome Biol* 2012, **13**(3):R22.
22. Keightley PD, Halligan DL: **Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans.** *Genetics* 2011, **188**(4):931–940.
23. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, Jorgensen T, Hansen T, Pedersen O, Wang J, Nielsen R: **Estimation of allele frequency and association mapping using next-generation sequencing data.** *BMC Bioinformatics* 2011, **12:**231.
24. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27**(21):2987–2993.
25. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J: **SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data.** *PLoS ONE* 2012, **7**(7):e37558.
26. Lucas Lledó JI, Cáceres M: **On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing.** *PLoS ONE* 2013, **8**(4):e61292.
27. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25**(15):1966–1967.
28. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26**(5):589–595.
29. **Novocraft technologies.** [http://www.novocraft.com]
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project DataProcessingSubgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
31. **Mario Cáceres Lab.** [http://grupsderecerca.uab.cat/cacereslab/content/resources]
32. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator.** *Bioinformatics* 2012, **28**(4):593–594.
33. Nei M: **Analysis of gene diversity in subdivided populations.** *PNAS* 1973, **70**(12):3321–3323.
34. Weir BS, Cockerham CC: **Estimating F-Statistics for the analysis of population structure.** *Evolution* 1984, **38**(6):1358–1370.
35. Bhatia G, Patterson N, Sankararaman S, Price AL: **Estimating and interpreting FST: the impact of rare variants.** *Genome Res* 2013, **23**(9):1514–1521.
36. Yang Z: *Computational Molecular Evolution*. New York: Oxford University Press; 2006.
37. **BreakDB.** [http://sv.gersteinlab.org/breakdb/]
38. Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB: **Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.** *Nat Biotech* 2010, **28:**47–55.
39. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, et al.: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**(7191):56–64.
40. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, Kim SJ: **The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group.** *Genome Res* 2009, **19**(9):1622–1629.
41. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, et al.: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19**(9):1527–1541.
42. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860–921.
43. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C,

Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304–1351.
44. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, et al.: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**(10):e254.
45. Martínez-Fundichely A, Casillas S, Egea R Ràmia, M, Barbadilla A, Pantano L, Puig M, Cáceres M: **InvFEST, a database integrating information of polymorphic inversions in the human genome.** *Nucleic Acids Res* 2013, **42**(D1):D1027–D1032.
46. Whelan CW, Tyner J, L'Abbate A, Storlazzi CT, Carbone L, Sönmez K: **Cloudbreak: accurate and scalable genomic structural variation detection in the cloud with MapReduce.** *arXiv:1307.2331 [q-bio]* 2013.
47. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, Catacchio CR, Ko A, Malig M, Baker C, Project GAG, Marques-Bonet T, Ventura M, Batzer MA, Eichler EE: **Rates and patterns of great ape retrotransposition.** *PNAS* 2013, **110**(33):13457–13462.
48. Li X, Chen S, Xie W, Vogel I, Choy KW, Chen F, Christensen R, Zhang C, Ge H, Jiang H, Yu C, Huang F, Wang W, Jiang H, Zhang X: **PSCC: Sensitive and reliable population-scale copy number variation detection method based on low coverage sequencing.** *PLoS ONE* 2014, **9:**e85096.
49. Buerkle CA, Gompert Z: **Population genomics based on low coverage sequencing: how low should we go?.** *Mol Ecol* 2013, **22**(11):3028–3035.
50. Shao H, Bellos E, Yin H, Liu X, Zou J, Li Y, Wang J, Coin LJM: **A population model for genotyping indels from next-generation sequence data.** *Nucl Acids Res* 2013, **41**(3):e46.

doi:10.1186/1471-2105-15-163
Cite this article as: Lucas-Lledó *et al.*: Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm. *BMC Bioinformatics* 2014 **15**:163.