A novel method to model read counts in genomic data to reduce false positive identification of heterozygotes

Steven H. Wu,[1] Rachel S. Schwartz,[1] David J. Winter,[1] Don Conrad,[3] Reed A. Cartwright[1, 2, *]

[1]The Biodesign Institute, [2]School of Life Sciences

Arizona State University

Tempe, AZ, USA

[3] Washington University in St. Louis

St. Louis, MO, USA

[*]Corresponding author: cartwright@asu.edu

Key words: next-generation sequencing, genotyping, genotype likelihood, mixture of Dirichlet multinomial.

Abstract

Accurate identification of genotypes is critical in identifying de novo mutations, linking mutations with disease, and determining mutation rates. To call genotypes correctly from short-read data requires modeling read counts for each base. True heterozygotes may be affected by mapping reference bias and library preparation, leading to a distribution of reads that does not fit a 1:1 binomial distribution, and potentially resulting failure to call the alternate allele. Homozygous sites can be affected by the alignment of paralogous genes and sequencing error, which could incorrectly suggest heterozygousity.

Previous work has modeled increased variance and skewed allele ratios to some degree. Here, we were able to model reads for all data as a mixture of Dirichlet multinomial distributions. This model has a better fit to the data than previously used models. In most cases we observed two distributions: one corresponds to a large proportion of heterozygous sites with a low reference bias and close-to-binomial distribution, and the other to a small proportion of sites with a high bias and overdispersion. The sites with high reference bias have not been previously identified as SNPs in extensive human genome research; thus, we believe these sites are not heterozygous in our data for the individuals studied here, and are falsely identified as heterozygous sites. We propose that this approach to modeling the distribution of NGS data provides a better fit to the data, which should lead to improved genotyping. Furthermore, the mixture of distributions may be used to suggest true and false positive *de novo* mutations. This approach provides an expected distribution of reads that can be incorporated into a model to estimate de novo mutations using reads across a pedigree.

Background

Identifying genotypes from next-generation sequencing (NGS) data is an important component of modern genomic analysis. Accurate genotyping is key to identifying sequence polymorphisms, detecting *de novo* mutations, linking genetic variants with disease, and determining mutation rates (Awadalla *et al*, 2010; Sayed *et al*, 2009). However, accurately identifying *de novo* mutations is a particular challenge, as true mutations are rare compared to errors in sequencing and downstream analyses.

Estimating genotypes from NGS data can be computationally and statistically complicated. A typical NGS experiment generates millions of short read fragments, 100 to 650 bp in length, that are aligned to a reference genome if available. For this reason, variant calling software typically uses a binomial distribution to model base-counts (although see Ramu *et al*, 2013). However, there are at least three experimental processes that affect the ratio of the alleles. (1) During library preparation, variation in amplification rates can cause some chromosomes to be replicated more than others (Heinrich *et al*, 2012). This variation is especially a concern if there is little starting material. (2) NGS technologies introduce sequencing errors into sequencing reads. Error-rates are on the order of 0.1–1% per base-call. While this may seem small, 0.1% error is equivalent to sequencing the wrong human genome, and 1% is equivalent to sequencing a chimpanzee instead of a human (Fox *et al*, 2014; Wall *et al*, 2014). (3) Bioinformatic methods that assemble reads with respect to a reference can misplace reads and penalize non-reference alleles (Degner *et al*, 2009). Together these processes shift the mean and increase the variance of sequencing-read distributions. Thus, it is possible for both homozygotes and heterozygotes to have an intermediate ratio of two alleles, making identification of true heterozygotes particularly difficult (Malhis and Jones, 2010).

These processes do not affect all parts of the genome equally. The genomic context of a site, including the presence of nearby indels, structural variants, or low-complexity regions, influences the probability that reads generated from a given site will be subject to these processes (Malhis and Jones, 2010). The mismatch between observed and expected read distributions created by the processes described above contributes to observed false positive single nucleotide polymorphism (SNP) discovery rates of 3 to 12% (Harismendy *et al*, 2009). Because putative SNPs are typically validated using another sequencing technology, high false positive rates increase the effort required for validation.

Previous approaches for accurate genotyping

Modeling systematic bias and variation in data has provided some improvements in statistical discrimination of true and false positive heterozygotes. The increased variance and skewed allele ratios produced from mismapped reads can be partially controlled for by including mapping quality data in a genotype-calling procedure. In the simplest approaches, reads with low quality scores are removed from an analysis. In Bayesian approaches to genotype calling, read quality data is included when calculating genotype probabilities (Li *et al*, 2009*b*). The increased variance caused by library preparation, sequencing, and errors in mapping reads to a reference genome can be accommodated by modeling read-counts as coming from a beta-binomial distribution (Ramu *et al*, 2013). The beta-binomial distribution acts as an over-dispersed binomial, allowing the excess variance to be handled in a standard statistical framework. All genotype calling procedures can combined with machine learning algorithms that attempt to differentiate between true variants and those caused by sequencing artifacts (DePristo *et al*, 2011). However, maximizing the true positive rate (i.e. maintaining high sensitivity) while minimizing the false negative rate (i.e. maintaining high specificity), remains a significant challenge (Greiner *et al*, 2000).

Our approach

In this study we introduce a new model for the distribution of reads produced from NGS, in which reads are assumed to come from a mixture of Dirichlet multinomial distributions. The Dirichlet multinomial distribution (DM) is the general case of the beta-binomial, allowing for overdispersion and modeling of more than two outcomes. By fitting a mixtures of DMs (MDM) we improve the beta-binomial models discussed above in two ways. First, we account for the context-dependent nature of genotyping errors by estimating multiple different DM models for a given dataset, each with different parameter values and levels of overdispersion. Second, we can explicitly model the presence of bases that are neither reference nor the likely alternative allele at a given site. This model allows us to directly estimate the probability of sequencing errors in a given DM model.

We first demonstrate the value of our approach by fitting MDMs to sequencing data derived from a haploid human cell line. The MDM produces a superior fit to this data compared to other methods, showing

that even relatively simple genetic datasets can be the result of heterogeneous processes, and thus benefit from a mixed-model approach. We then fit MDMs to diploid data generated by the 1000 Genomes Project (1000 Genomes Project Consortium *et al*, 2010, 2015). For this data, the MDM also improves the fits compared to other models. One component of the MDM model contains most of the true heterozygotes, while the other component contains primarily sites that have not been identified as heterozygous in any previous human research. Therefore we believe this model may be utilized to detect false positive heterozygous sites, leading to a significant reduction in the number of sites requiring validation.

Results

Haploid Dataset

We examined two genomic regions from the CHM1 (haploid human cell line) dataset: all of chromosome 21 and part of chromosome 10. For each region we further split sites into two subsets. The full dataset (FD), where reads were only filtered to exclude regions with unusually high coverage, and the reference dataset (RD), where only sites with at least 80% of reads matching the reference base were included.

Best fit models for haploid data

We fit seven models to each genomic region in each dataset: a multinomial, a DM, and MDM models with two to six components. The addition of model components increased the likelihood of the model for all cases (Table 1). Using Bayesian information criterion (BIC), the best fitting model for each dataset was the two component MDM. In all cases a single component contains a substantial majority of sites (approximately 75% of sites for the reference dataset from chromosome 21 and 95% of the sites for other datasets). We will refer the component to which the highest proportion of sites is assigned as the "major component" and all other components as "minor components".

The overdispersion parameter, φ, describes the degree which the expected variance of a given DM distribution is greater than that of a corresponding multinomial. φ can take values between 0 and 1, with

0 being identical to the multinomial and 1 being completely overdispersed. For the full dataset, the major component had relative little overdispersion ($\varphi = 0.00252$ and 0.00415 for chromosome 21 and chromosome 10 respectively). The minor component displayed strong overdispersion ($\varphi = 0.892$ and 0.948). For the reference dataset, there was also little overdispersion for the major component (φ = 0 for chromosome 21 and 0.00269 for chromosome 10). After removing sites with high proportions of reads in the error categories, the minor component was slightly overdispersed ($\varphi = 0.0153$ and 0.0475) (Table 1 and supplementary tables)

Visualizing model fitting for the haploid data

We examined the fit of the data to each model using quantile-quantile (QQ) plots, where the quantile of the observed read counts are plotted against the quantile of the estimated read counts. For the model with two components applied to the reference dataset, the reference and error counts fit the expected values (Figure 1 and supplementary figures).

Diploid dataset

We examined the same two genomic regions for NA12878, the daughter of the CEU trio. In order to investigate the impact of sequencing technology on parameter estimates from our our model we repeated our analysis for each of the four released datasets (1000 Genomes Project Consortium *et al*, 2010, 2015) . As potential heterozygotes present the greatest challenge to variant calling, we focused on these sites. Specifically, we identified potential heterozygous sites using the SAMtools heterozygote caller on NA12878 alone and using the trio caller. Sites were only included in the potential heterozygote (PH) dataset if they were called by both methods. We filtered the PH dataset to include only sites identified as SNPs by the 1000 Genomes Project. We considered these sites to be true heterozygous sites (TH dataset). The number of sites and the proportion of true heterozygous sites are summarized in Table 2.

Best fitting models

We fit eight models to each of the 16 CEU datasets (two genomic regions, four release years, PH/TH): a multinomial, multinomial with reference bias, a DM, and MDMs with two to six components. The addition of model components increased the likelihood of the model for all cases (Table 3).

For the TH dataset, the best-fitting model as selected by BIC had two or three components, and the best AIC had three or four components depending on the run year. The majority of the sites (88–99%) were assigned to one component in the model (Table 3 and supplementary tables). The major component of the model for each dataset (the component with the highest proportion of sites) had little overdispersion (φ = 0 to 0.00055). In addition, the major component also had a approximately equal proportion of reference and alternative alelles (49% to 51%), and a relatively small error term ($< 0.1\%$) for the 2011, 2012, and 2013 datasets. Thus, the majority of sites fall into a component that is approximately a binomial distribution. The 2010 dataset has a slightly larger error term (0.2% and 0.3%), and the reference and alternate terms are 53% and 46% respectively. For the datasets with a two component model, the minor component is similar to the major component but with greater overdispersion (φ = 0.06 – 0.1). For models with three components, the minor components had an elevated proportion of one of the reference, alternate, or error terms, and greater overdispersion. For instance, CEU2013 chromosome 21 has φ = 0.0656 and π_{error} = 0.278 for the third component.

For the PH dataset, the best fitting model had three to six components (Table 3). The major component of the model contains between 71% and 95% of sites for all years. As with the TH dataset, the major components all had little overdispersion (φ = 0 to 0.00135). Even when models with greater than four components were favored by BIC, the additional components contain a very small proportion of the data ($< 1\%$), and frequently produce estimates of sequencing error very close to zero. Thus, a model with more than three components is likely overfitting the data.

Visualizing model fit

We examined the fit of the data to each model using quantile-quantile (QQ) plots. When we examined the QQ plot for the MDM model with the lowest BIC, all three terms (reference allele, alternative allele, and

error term) fit closely to the expected values (Figure 2).

Assignment of sites to model components

We assigned each site from each of the eight PH datasets to a component in the MDM model based on the site likelihood. The minor or combined minor components were always enriched for false positive heterozygous sites; these false positive sites made up between 10% and 51% of the sites in these components. The major component contained only 3 – 10 % of the false positives (Table 3 and supplementary table).

We constructed a receiver operating characteristic (ROC) curve to examine the performance of the MDM model as a classifier of heterozygotes (Figure 3 and supplementary figures). The performance of this classifier on a given dataset can be summarized by the area under the ROC curve (AUC). AUC was between 0.634 and 0.81 (Table 4).

Based on the ROC curve, we selected a threshold value, the probability cutoff for assigning a site to the major component, with sensitivity close to 1 and specificity near 50%. Thus, we can use our model to filter out half of the false positive heterozygous sites without losing true heterozygotes.

Classification of sites as copy number variants

We tested the hypothesis that CNVs produce false positive heterozygous sites (Li, 2014). The proportion of the false positive heterozygous sites belonging to CNV regions is between 8% to 13% for chromosome 21, but $< 2\%$ for chromosome 10 (Table 5).

Discussion

We have developed a novel statistical approach to model the distribution of NGS reads. Using an MDM produces a better fit to haploid human cell line data than previous approaches (i.e. the multinomial and DM Ramu *et al*, 2013). This result demonstrates that NGS datasets from relatively simple biological

176 samples (i.e. no true heterozygotes and a high quality reference genome) can benefit from the approach
177 we describe here.

178 Similarly, our MDM model provide a better fit to more complex data, including potentially heterozygous
179 sites in data arising from the 1000 Genomes Project. Our goal in developing this model was to improve
180 the accuracy of genotype calling, and reduce the number of false-positive variant calls produced from
181 NGS data.

182 **Best fitting models**

183 Minor components of our MDMs tended to display bias toward reference or alternative alleles or higher
184 values for overdispersion of sequencing errors. These results suggest that most sites in an NGS experiment
185 match the idealized expectation of a binomial distributed of base-counts. On the other hand, a substantial
186 minority of sites appear to be generated by processes that differ from this expectation. Moreover, these
187 minor components are greatly enriched for apparently false positive heterozygous sites.

188 Our results are similar to that of Muralidharan *et al* (2012), who observed a high proportion of SNPs with
189 low error rates and a low proportion of SNPs with high error rates. This result was attributed to high
190 alignment error in repetitive regions. We now provide a way of using this approach to distinguish these
191 two types of sites.

192 **Assignment of sites to model components**

193 The MDM classifier shows promise in discriminating true and false positive heterozygotes, as illustrated
194 by our ROC curves. The two exceptions, CEU2010 chromosome 10 and CEU2013 chromosome 10, may be
195 due to an extremely low proportion of false positives in the dataset: with only 5% false positive sites, it is
196 a challenging task for the classification algorithm to identify these sites (Table 6). Thus, the modeling
197 approach described here can be used to remove sites that have been called as heterozygous but are likely
198 to be false positive calls, by selecting a probability cutoff for assigning sites to the major component and
199 filtering out sites belonging to the minor components. Removing such sites reduces the cost and time

200 required for validation.

201 **Copy number variants**

202 It is possible that there is a weak correlation between false positive heterozygous sites and copy number
203 variations in chromosome 21. We expected this correlation to be stronger across all regions. This also
204 suggested that there are several other multiple factors caused copy number variations. Species with
205 greater numbers of duplicated regions than humans may have greater numbers of sites incorrectly
206 identified as heterozygous, potentially affecting the identification of *de novo* mutations.

207 **Conclusion**

208 Our modeling approach is designed to accommodate the correlated and context-specific nature of errors
209 introduced in generating NGS datasets. The datasets we analyzed were produced using a variety of
210 different library preparations and sequencing technologies. These differences are partly reflected in
211 the different parameter values we estimate from our model. In particular, the 2010 dataset appears to
212 have a quite different profile than those from other years: the major component of this model has higher
213 overdispersion and stronger reference-bias than that of any other dataset.

214 Previous work has suggested that there is reference bias in mapping, and overdispersion due to bio-
215 logical factors (Meyer and Liu, 2014). We observed very little reference bias and overdispersion for true
216 heterozygotes. However, false positive heterozygote calls fall into a distribution with reference bias and
217 overdispersion. By using an MDM model we believe we are able to separate true heterozygotes from false
218 positive calls, which can significantly reduce the time and expense of subsequent validation work. In the
219 future this modeling approach will be incorporated into a pedigree-based approach for accurate genotype
220 calling (Cartwright *et al*, 2012)

Methods

Data

We extracted datasets from two types of data. The first dataset is the haploid human sequence from a hydatidiform mole cell line (CHM1hTERT SRR1283824 from SRP017546). We refer to this dataset as the CHM1 dataset in this paper.

Second, we obtained sequences from the 1000 Genomes Project for three individuals, a woman (NA12878) and both of her parents (NA12891 and NA12892). Sequencing was repeated for these individuals in different years using different technologies (2010, 2011, 2012, 2013). The 2010 dataset was generated during pilot 2 studies; the 2013 PCR free dataset was part of the phase 3 release; the 2011 and 2012 datasets were two non-official release datasets, aligned with a decoy genome that captures reads that failed to align to the standard reference genome (1000 Genomes Project Consortium *et al*, 2010, 2015).

We refer to this dataset as the CEU dataset. If the release year is appended, for example CEU2013, then we refer to the specific release in 2013. For each of these five datasets (CHM1 and each of the four releases of CEU), we analyzed two genomic regions, the whole chromosome 21 and a subregion of chromosome 10, from positions 85534747 to 135534747, which is approximately the same size as chromosome 21 (48 million base pairs).

For CHM1 we probabilistically called genotypes by first obtaining allele counts for each base at each site using the mpileup function in SAMTools v1.2 (Li *et al*, 2009*a*; Li, 2011) and the human reference genome (Genome Reference Consortium human genome build 37). We then used BCFTools v1.2 (Li *et al*, 2009*a*; Li, 2011) to identify potential heterozygous sites. For each of these sites we calculated the frequency of the reference allele and the frequency of all non-reference alleles (error). We filtered this dataset based on the read depth for each site: we removed sites with read counts of less than 10 or greater than 150. Sites with high numbers of reads are likely in copy number variable genes that have aligned to a single region of the genome. Apparent heterozygotes are more likely to be due to paralogs rather than variation within a gene. The low read filter limits the data to calls with enough coverage to provide a reasonably accurate call and proportion of reads for each base. We refer this dataset as the full dataset (FD). Additionally, we

removed sites for which less than 20% of the reads contained the reference allele. We refer to this dataset as the reference dataset (RD).

For the CEU data, we obtained allele counts as above for all three individuals. We then called genotypes as above on NA12878 (the daughter of the trio) and by using the BCFtools trio caller with the data from all three individuals. We limited the dataset we used for subsequent analyses to sites that were found by both methods. Sites that were found only in triocaller, but not in the individual caller were likely identified by the pedigree with limited data for the daughter; thus, these low coverage sites were not included in subsequent analysis. We removed sites with read counts of less than 10 or greater than 150, as for CHM1. We call this the Potential Heterozygote (PH) dataset. For each of these sites we calculated the frequency of the reference allele, the frequency of the alternate allele, and the frequency of any other alleles (error). We compared the frequencies of each allele category (reference, alternate, error) for each possible genotype combination. Because we found no differences in frequencies for different genotypes, all subsequent analyses were only performed on the general reference-alternate-error dataset.

We created an additional dataset by removing sites from the PH dataset not found to be heterozygous by the 1000 Genomes Project (1000 Genomes Project Consortium *et al*, 2010, 2015). We then discarded sites for which the alternate allele differed from the one previously identified by the 1000 Genomes Project. We call this the True Heterozygote dataset (TH). These datasets allow us to build a model that distinguishes sites found in the PH dataset but not in the TH dataset, which are likely false positive heterozygote calls.

Because the CHM1 dataset was larger than the CEU dataset, we randomly subsampled the CMH1 dataset to have an approximately equal number of sites (40,000 sites) as the CEU dataset.

Model fitting and parameter estimation

We fit seven models to each CHM1 dataset, and eight to each CEU dataset. The models included a multinomial, a multinomial with reference bias (CEU only), Dirichlet multinomial (DM) and mixtures of DM (MDM) distributions with various number of components, ranging from two to six. We estimated the parameters and calculated the genotype likelihood for each model.

The genotype likelihood measures the likelihood of a sample's genotype, G, given a set of base-calls, R, and is proportional to the probability of observing R if the genotype was G, i.e. $L(G|R) \propto P(R|G)$. We derived genotype likelihoods using MDM distributions. The Dirichlet multinomial distribution is a compound distribution generated when a Dirichlet distribution is used a prior for the probabilities of success of a multinomial distribution: $\boldsymbol{p} \sim Dirichlet(\boldsymbol{\alpha})$ and $\boldsymbol{x} \sim Multinomial(N, \boldsymbol{p})$ where $\boldsymbol{\alpha}$ is a vector of concentration parameters, $\boldsymbol{p}$ is a vector of proportions, $\boldsymbol{x}$ is a vector of counts, and N is the sample size. After integrating out $\boldsymbol{p}$, the resulting probability mass function can be trivially expressed as a product of ratios of gamma functions:

$$P(\boldsymbol{x}; \boldsymbol{\alpha}, N) = \binom{N}{\boldsymbol{x}} \frac{\Gamma(\sum \alpha_i)}{\Gamma(\sum \alpha_i + N)} \prod_i \frac{\Gamma(\alpha_i + x_i)}{\Gamma(\alpha_i)} \tag{1}$$

where $\sum x_i = N$ and $\alpha_i > 0$. Furthermore,

$$E(x_i) = N\pi_i \text{ and } Var(x_i) = N\pi_i(1-\pi_i)\frac{A+N}{A+1} \tag{2}$$

where $A = \sum \alpha_i$ and $\pi_i = \frac{\alpha_i}{A}$

It is helpful to reparameterize the distribution by letting $\alpha_i = \frac{1-\varphi}{\varphi}\pi_i$, where $\varphi = \frac{1}{(A+1)}$ represents the pairwise correlation between samples. As a result, $Var(x_i) = N\pi_i(1-\pi_i)(1+(N-1)\varphi)$ and $\varphi \in [0, 1]$ is a parameter controlling the amount of excess variation in the Dirichlet multinomial. When φ = 0, the DM reduces to a multinomial. Thus the Dirichlet multinomial can be interpreted as an over-dispersed multinomial distribution: as φ approaches 1, the distribution is completely overdispersed, the dataset is more heterogeneous than expected.

For a single-component Dirichlet multinomial, we computed the maximum likelihood estimate model starting with a method-of-moments estimation and optimizing using the Newton-Raphson method. For all other MDM, the maximum likelihood estimated was computed using an EM algorithm. This procedure was repeated 1000 times to search for the global maximum likelihood estimation. For each repetition, we started the search with the method of moments estimates of the parameters, then calculated the likelihood of the data for each component.

For the Dirichlet multinomial distribution, we estimated φ as a measure of the overdispersion of the data. In addition, we estimated ρ, the proportion of sites belongs to each Dirichlet multinomial component. For the CHM1 dataset we estimated the proportion of the reference allele and error term for each model or model component. For each CEU dataset we estimated the proportion of the reference allele, alternate allele, and the error term.

To determine the optimal number of components in the MDM model both the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were calculated for each dataset; the model with the lowest AIC or BIC is considered as the best model. We discussed about the model with the best BIC in the result section. The AIC and BIC for each model are calculated by the following formula;

$$AIC = -2\log(L) + 2\log(n) \tag{3}$$

$$BIC = -2\log(L) + k\log(n) \tag{4}$$

Where L is the maximum likelihood estimation from the model, k is the number of free parameters, n is the number of individual in the dataset.

Visualizing model fit

We visualized the fit of the data to each model and compared the fit between models using quantile-quantile (QQ) plots. The QQ plot plots the quantile of the observed read counts against the quantile of the estimated read counts. Parameters estimated from the EM were used to simulate the expected read counts for the plot. Two QQ plots, one for the reference allele and one for the error term, were used to illustrate the fit of models for the CHM1 dataset. Three separate QQ plots, one each for reference, alternate, and error terms, were used for the CEU datasets.

Assignment of sites to model components

To suggest the use of the MDM model as a classification method, we calculated the likelihood of every site under each component of the model in each of the CEU datasets. We assigned each site to a particular component in the MDM model by comparing the likelihood between all components. The likelihood for each component was reevaluated using the parameters estimated from the EM. The site was assigned to the component with the highest likelihood.

For the CEU PH dataset, we extracted all sites that assigned to the minor components. The number of true and false heterozygous sites and the proportion of false heterozygous sites were calculated using the 1000 Genome Project.

To determine the performance of the classification algorithm, we implemented an alternative way to assign each site. We recalculated the density of the probabilities of assignment to the major component of the model, and we interpreted that as the probability of being true heterozygous site. We used these probabilities to construct the receiver operating characteristic (ROC) curve, where sensitivity is plotted against specificity, to examine the performance of our model as a classifier across a range of classification thresholds. The area under the ROC curve (AUC) summarizes the performance of this classification method across a range of cutoff points. An AUC of 1 represents a perfect classifier, while an AUC of 0.5 suggests the prediction is close to random.

Classification of sites as copy number variants / paralogous

One possible cause of identifying false positive heterozygous site is that the site belongs to the region which is known for copy number variation (CNV). We extracted all the known CNV sites for NA12878 in the CEU dataset from the 1000 Genomes Project. We extracted all known false positive heterozygous calls, which are sites in the PH dataset but not in the TH dataset, and mapped them to the known CNV sites. We calculated the proportion of sites belong to the known copy number variation sites for each dataset and each genomic regions.

Figures

Figure 1: QQ plots from CHM1 RD dataset show that mixture of Dirichlet multinomial models with two components fits better than multinomial and Dirichlet multinomial model.

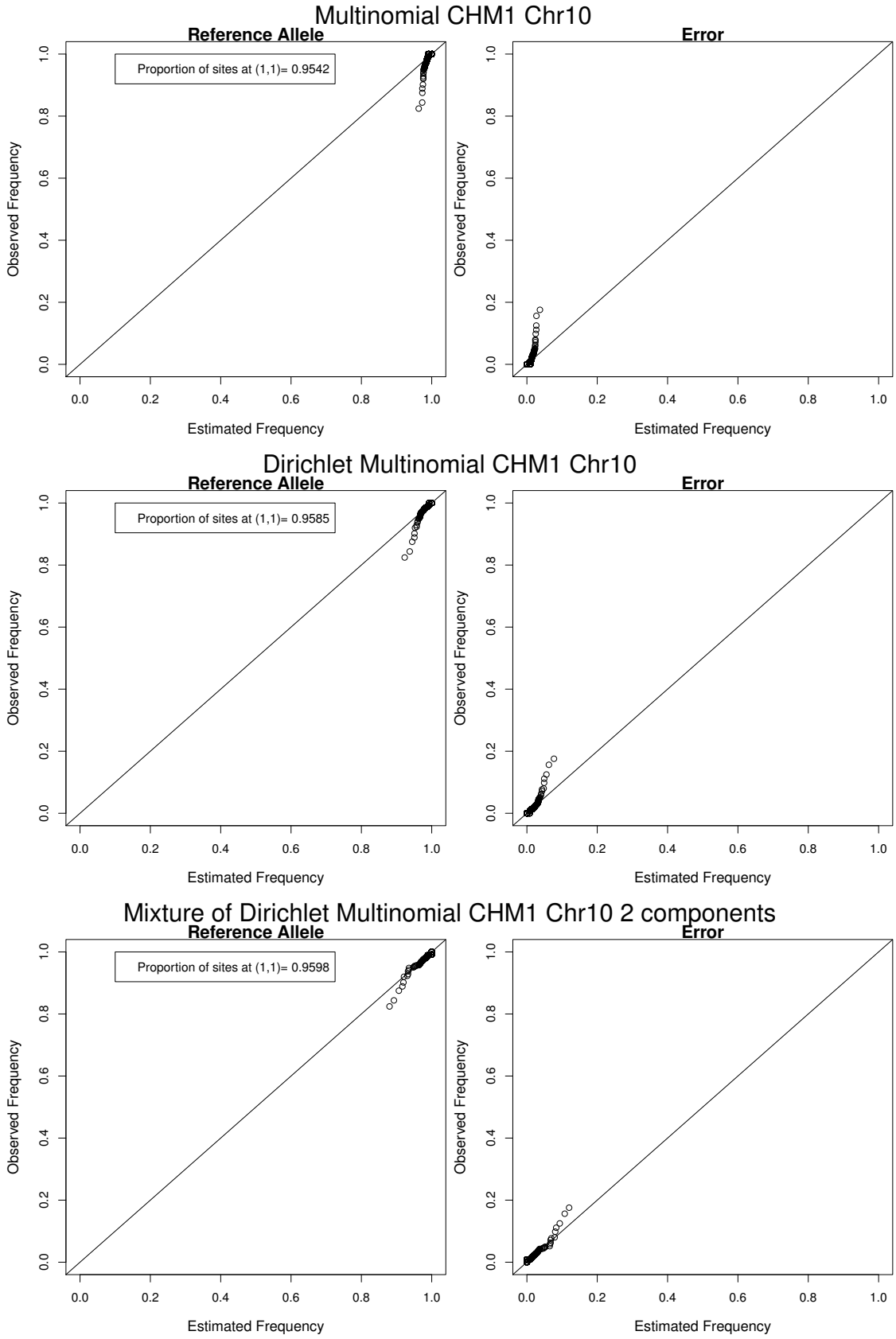


Figure 2: QQ plots for the CEU2013 TH dataset show that mixture of Dirichlet multinomial models with three components fits better than multinomial model.

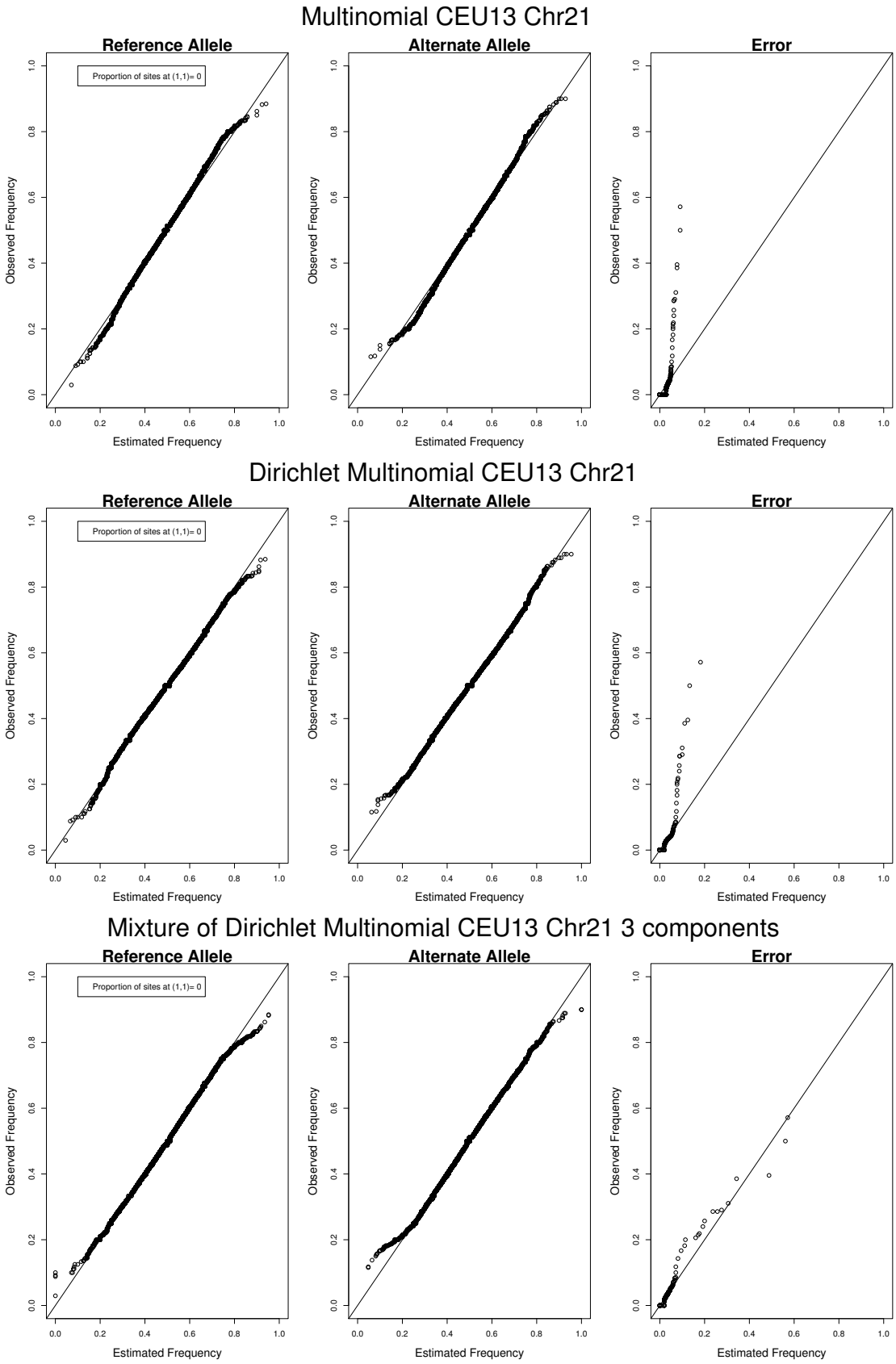
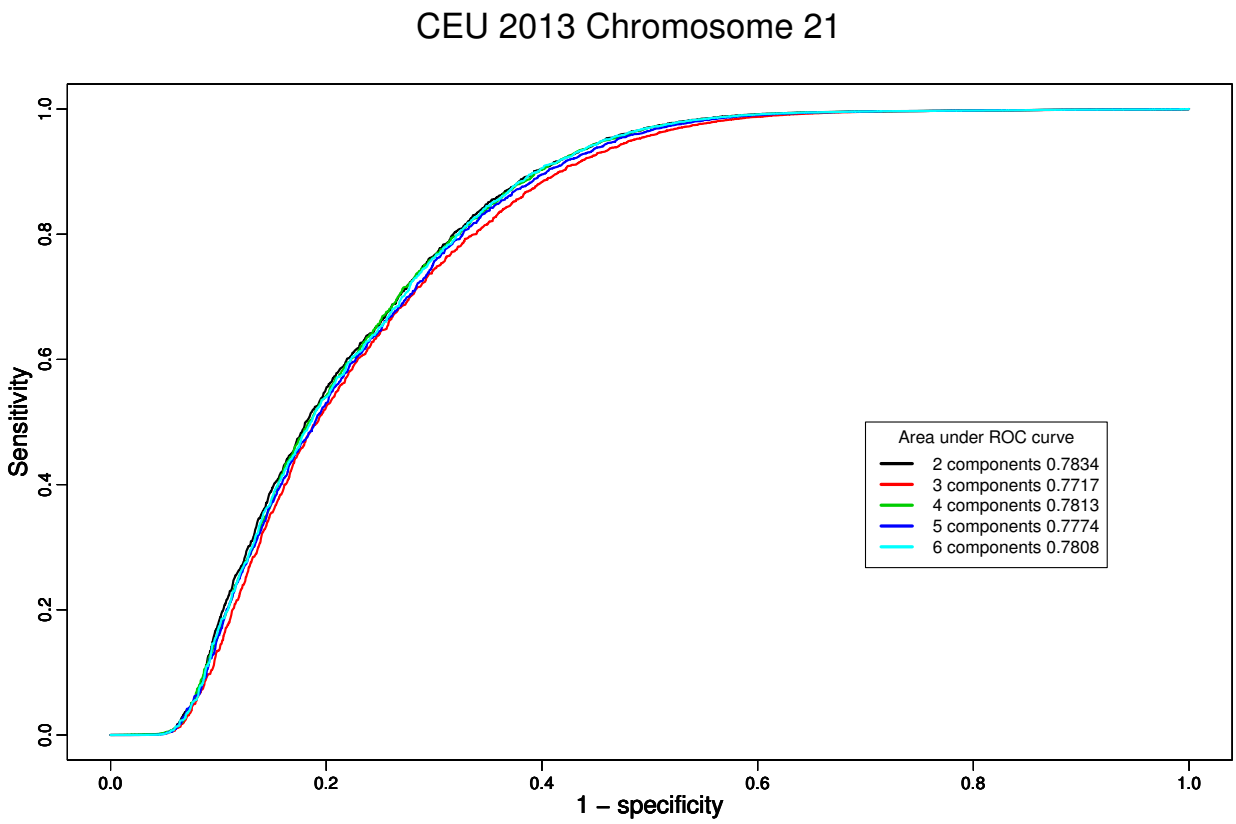


Figure 3: Receiver operating characteristic (ROC) curve with area under the curve for different components in the model.



Tables

Table 1: The number of components in the best MDM model according to AIC and BIC values for each CHM1 dataset and parameters estimated for the best BIC model. π_{ref} is the proportion of the reference term. π_{err} is the proportion of the error term. φ is the overdispersion parameter. When φ = 0, the DM reduces to a multinomial. As φ approaches 1, the distribution is completely overdispersed. p is the proportion of site in each component. ML-p is the proportion of sites assigned to each component using the likelihood.

Dataset	AIC	BIC	π_{ref}	π_{err}	φ	p	ML-p
Chr21 FD	4	2	1	0.000249	0.00252	0.972	0.995
			0.982	0.0176	0.892	0.0278	0.00485
Chr21 RD	2	2	1	0.00022	0	0.751	0.808
			1	0.000361	0.0153	0.249	0.192
Chr10 FD	3	2	1	0.000282	0.00415	0.984	0.999
			0.975	0.0254	0.948	0.0164	0.000978
Chr10 RD	2	2	1	0.00026	0.00269	0.942	0.997
			0.999	0.000833	0.0475	0.0585	0.00324

Table 2: Number of heterozygous sites identified by different methods and the number of true heterozygous sites from 1000 Genomes Project.

Dataset	Individual caller only	Trio caller only	Both callers	True heterozygotes (TH)	Proportion of TH
CEU13 Chr21	143	1604	40956	30120	0.735
CEU13 Chr10	1652	2645	38542	36590	0.949
CEU12 Chr21	151	1171	38180	29983	0.785
CEU12 Chr10	106	880	40144	36818	0.917
CEU11 Chr21	145	1190	38107	29991	0.787
CEU11 Chr10	114	867	40156	36825	0.917
CEU10 Chr21	6447	4971	31773	28197	0.887
CEU10 Chr10	194	1173	37108	35062	0.945

Table 3: The number of components in the best MDM model according to AIC and BIC values for each CEU dataset and parameters estimated for the best BIC model. π_{ref} is the proportion of the reference term. π_{alt} is the proportion of the alternative term. π_{err} is the proportion of the error term. φ is the overdispersion parameter. When φ = 0, the DM reduces to a multinomial. As φ approaches 1, the distribution is completely overdispersed. p is the proportion of site in each component. ML-p is the proportion of sites assigned to each component using the likelihood.

Dataset	AIC	BIC	π_{ref}	π_{alt}	π_{err}	φ	p	ML-p
CEU13 TH Chr21	4	3	0.504	0.496	0.000353	0.000246	0.939	0.783
			0.508	0.491	0.000526	0.0689	0.0604	0.213
			0.239	0.483	0.278	0.0656	0.000587	0.00412
CEU13 TH Chr10	3	2	0.502	0.497	0.000336	0.000428	0.992	0.848
			0.504	0.49	0.00621	0.124	0.00754	0.152
CEU12 TH Chr21	4	2	0.509	0.491	0.000315	0.000129	0.961	0.84
			0.541	0.457	0.00204	0.0773	0.039	0.16
CEU12 TH Chr10	4	2	0.508	0.491	0.000319	0.000553	0.986	0.851
			0.54	0.457	0.00305	0.076	0.0138	0.149
CEU11 TH Chr21	4	2	0.509	0.491	0.000315	0.000131	0.961	0.842
			0.541	0.457	0.00198	0.0782	0.0391	0.158
CEU11 TH Chr10	3	2	0.508	0.491	0.000319	0.000556	0.986	0.848
			0.542	0.455	0.00301	0.0737	0.0139	0.152
CEU10 TH Chr21	3	2	0.533	0.465	0.00225	0.00152	0.922	0.718
			0.67	0.327	0.0027	0	0.0783	0.282
CEU10 TH Chr10	4	2	0.534	0.463	0.00305	0	0.684	0.606
			0.55	0.449	0.000645	0.0129	0.316	0.394
CEU13 PH Chr21	6	6	0.501	0.499	0.000336	8.44e-05	0.709	0.46
			0.638	0.361	0.000751	0.00848	0.138	0.224
			0.372	0.627	0.00066	0.00219	0.0681	0.194
			0.781	0.219	0.000349	0	0.0662	0.0825
			0.215	0.785	0.000288	0.00219	0.0129	0.0245
			0.419	0.421	0.16	0.107	0.00549	0.016
CEU13 PH Chr10	3	3	0.502	0.497	0.000342	0	0.95	0.784
			0.523	0.476	0.00147	0.0639	0.0492	0.205
			0.299	0.485	0.216	0.0875	0.00121	0.011
CEU12 PH Chr21	6	5	0.507	0.492	0.000313	0	0.83	0.65
			0.643	0.356	0.00117	0.0157	0.0997	0.169
			0.381	0.618	0.00127	0.037	0.0433	0.125
			0.778	0.221	0.000457	0	0.024	0.0458
			0.439	0.375	0.186	0.102	0.00302	0.0101
CEU12 PH Chr10	6	4	0.508	0.491	0.000329	0.000834	0.957	0.8
			0.686	0.313	0.00144	0.0152	0.0371	0.13
			0.286	0.711	0.00207	0.0166	0.00317	0.0628
			0.424	0.388	0.188	0.0748	0.00232	0.00735
CEU11 PH Chr21	6	5	0.507	0.492	0.000311	0	0.831	0.648
			0.641	0.358	0.00113	0.0142	0.0979	0.17
			0.379	0.62	0.00125	0.0356	0.0428	0.125
			0.776	0.223	0.000423	0	0.0251	0.0463
			0.446	0.375	0.179	0.102	0.0031	0.0109
CEU11 PH Chr10	6	4	0.508	0.491	0.000331	0.000854	0.958	0.802
			0.687	0.312	0.00141	0.0141	0.0363	0.127
			0.288	0.71	0.00207	0.0165	0.00309	0.063
			0.427	0.387	0.186	0.0756	0.00233	0.00755
CEU10 PH Chr21	6	4	0.532	0.465	0.00221	0.00135	0.832	0.584
			0.696	0.301	0.00324	0.0089	0.145	0.247
			0.343	0.653	0.00423	0.0249	0.0193	0.122
			0.54	0.351	0.109	0.069	0.0038	0.0472
CEU10 PH Chr10	5	3	0.532	0.466	0.00227	0.000483	0.88	0.599
			0.66	0.338	0.00208	0.00326	0.0877	0.292
			0.508	0.482	0.00987	0.0637	0.032	0.109

Table 4: Summary of the area under receiver operating characteristic curve for each CEU PH dataset.

Dataset	Area under ROC curve
CEU13 Chr21	0.772
CEU13 Chr10	0.634
CEU12 Chr21	0.813
CEU12 Chr10	0.792
CEU11 Chr21	0.812
CEU11 Chr10	0.791
CEU10 Chr21	0.761
CEU10 Chr10	0.691

Table 5: CNVs are a small fractions of FP. Summary of the number copy number variants (CNV) sites within the false positive heterozygous sites in each CEU dataset.

Dataset	Not CNV	CNV	CNV proportion
CEU13 Chr21	8038	1074	0.1179
CEU13 Chr10	1710	31	0.01781
CEU12 Chr21	4901	440	0.08238
CEU12 Chr10	3023	32	0.01047
CEU11 Chr21	4845	439	0.08308
CEU11 Chr10	3020	32	0.01048
CEU10 Chr21	2600	399	0.133
CEU10 Chr10	1840	8	0.004329

Table 6: The major component has lower percentage of false positive and CNV. Proportion of true heterozygotes (TH) and false-positive heterozygote calls (FH) in each component for each CEU dataset. First category is the major component, all minor components are combined in the second category. .

Dataset	FH	TH	FH proportion	Not CNV	CNV	CNV proportion
CEU13 Chr21	3130	24209	0.1145	26924	415	0.01518
	5982	5642	0.5146	10851	773	0.0665
CEU13 Chr10	1167	30939	0.03635	32012	94	0.002928
	574	5322	0.09735	5864	32	0.005427
CEU12 Chr21	1778	25306	0.06565	26891	193	0.007126
	3563	4528	0.4404	7760	331	0.04091
CEU12 Chr10	1289	31361	0.03948	32560	90	0.002757
	1766	5306	0.2497	7035	37	0.005232
CEU11 Chr21	1765	25284	0.06525	26858	191	0.007061
	3519	4557	0.4357	7743	333	0.04123
CEU11 Chr10	1293	31335	0.03963	32539	89	0.002728
	1759	5339	0.2478	7060	38	0.005354
CEU10 Chr21	967	21110	0.0438	21968	109	0.004937
	2032	6812	0.2298	8518	326	0.03686
CEU10 Chr10	815	25614	0.03084	26380	49	0.001854
	1033	9190	0.101	10197	26	0.002543

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This was supported by NIH Grants R01-GM101352 to RA Zufall, RBR Azevedo, and RA Cartwright and R01-HG007178 to DF Conrad and RA Cartwright.

References

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Awadalla, P., J. Gauthier, R. A. Myers, F. Casals, F. F. Hamdan, A. R. Griffing, M. Côté, E. Henrion, D. Spiegelman, J. Tarabeux, A. Piton, Y. Yang, A. Boyko, C. Bustamante, L. Xiong, J. L. Rapoport, A. M. Addington, J. L. E. DeLisi, M.-O. Krebs, R. Joober, B. Millet, E. Fombonne, L. Mottron, M. Zilversmit, J. Keebler, H. Daoud, C. Marineau, M.-H. Roy-Gagnon, M.-P. Dubé, A. Eyre-Walker, P. Drapeau, E. A. Stone, R. G. Lafrenière, and G. A. Rouleau. 2010. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* 87:316–324.

Cartwright, R. A., J. Hussin, J. E. M. Keebler, E. A. Stone, and P. Awadalla. 2012. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. *Stat Appl Genet Mol Biol* 11.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. 2009. Effect of

read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25:3207–3212.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, and et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.

Fox, E. J., K. S. Reid-Bayliss, M. J. Emond, and L. A. Loeb. 2014. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* 1.

Greiner, M., D. Pfeiffer, and R. D. Smith. 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med* 45:23–41.

Harismendy, O., P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 10:R32.

Heinrich, V., J. Stange, T. Dickhaus, P. Imkeller, U. Krüger, S. Bauer, S. Mundlos, P. N. Robinson, J. Hecht, and P. M. Krawitz. 2012. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res* 40:2426–2431.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30:2843–2851.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. S. . 2009*a*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li, R., Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. 2009*b*. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19:1124–1132.

Malhis, N., and S. J. M. Jones. 2010. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics* 26:1029–1035.

386 Meyer, C. A., and X. S. Liu. 2014. Identifying and mitigating bias in next-generation sequencing methods
387 for chromatin biology. *Nat Rev Genet* 15:709–721.

388 Muralidharan, O., G. Natsoulis, J. Bell, D. Newburger, H. Xu, I. Kela, H. Ji, and N. Zhang. 2012. A cross-sample
389 statistical model for SNP detection in short-read sequencing data. *Nucleic Acids Res* 40:e5.

390 Ramu, A., M. J. Noordam, R. S. Schwartz, A. Wuster, M. E. Hurles, R. A. Cartwright, and D. F. Conrad. 2013.
391 DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* 10:985–987.

392 Sayed, S., D. R. Langdon, S. Odili, P. Chen, C. Buettger, A. B. Schiffman, M. Suchi, R. Taub, J. Grimsby, F. M.
393 Matschinsky, and C. A. Stanley. 2009. Extremes of clinical and enzymatic phenotypes in children with
394 hyperinsulinism caused by glucokinase activating mutations. *Diabetes* 58:1419–1427.

395 Wall, J. D., L. F. Tang, B. Zerbe, M. N. Kvale, P.-Y. Kwok, C. Schaefer, and N. Risch. 2014. Estimating genotype
396 error rates from high-coverage next-generation sequence data. *Genome Res* 24:1734–1739.