Published in final edited form as:

Genet Epidemiol. 2012 September ; 36(6): 549–560. doi:10.1002/gepi.21648.

Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next Generation Sequencing of Pooled Samples

Xiaowei Chen¹, Jennifer B Listman², Frank J Slack³, Joel Gelernter^{2,4,5}, and Hongyu Zhao^{1,5,6}

¹Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

²Department of Psychiatry, Yale School of Medicine, New Haven, CT

³Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT

⁴VA CT Healthcare Center, West Haven, CT

⁵Department of Genetics, Yale School of Medicine, New Haven, CT

⁶Division of Biostatistics, Yale School of Public Health, New Haven, CT

Abstract

Next generation sequencing is widely used to study complex diseases because of its ability to identify both common and rare variants without prior single nucleotide polymorphism (SNP) information. Pooled sequencing of implicated target regions can lower costs and allow more samples to be analyzed, thus improving statistical power for disease-associated variant detection. Several methods for disease association tests of pooled data and for optimal pooling designs have been developed under certain assumptions of the pooling process, e.g. equal/unequal contributions to the pool, sequencing depth variation, and error rate. However, these simplified assumptions may not portray the many factors affecting pooled sequencing data quality, such as PCR amplification during target capture and sequencing, reference allele preferential bias, and others. As a result, the properties of the observed data may differ substantially from those expected under the simplified assumptions. Here, we use real datasets from targeted sequencing of pooled samples, together with microarray SNP genotypes of the same subjects, to identify and quantify factors (biases and errors) affecting the observed sequencing data. Through simulations, we find that these factors have a significant impact on the accuracy of allele frequency estimation and the power of association tests. Furthermore, we develop a workflow protocol to incorporate these factors in data analysis to reduce the potential biases and errors in pooled sequencing data and to gain better estimation of allele frequencies. The workflow, Psafe, is available at <http://bioinformatics.med.yale.edu/group/>.

Keywords

pooled sequencing; allele frequency estimation; next-generation sequencing; disease association tests

INTRODUCTION

With the ability to identify both common and rare variants, next generation (next-gen) sequencing has become a widely used approach to identify disease-associated variants [Lee, et al. 2010; Shah, et al. 2009; Stephens, et al. 2009]. However, large-scale studies based on individual whole genome sequencing are still expensive, and sometimes not necessary when screening large numbers of samples for rare variants, particularly when only a few regions are of interest to the investigators. When targeted resequencing is used, the cumulative throughput (or sequencing depth) of a single run is often much higher than the sequencing depth needed for each individual, so it may be more cost effective to sequence multiple individuals in a pool after target region capture. Such pooled sequencing of targeted regions provides a cost-effective way both to identify novel variants, and to identify variants associated with human traits. Several common and rare variants for diseases such as inflammatory bowel disease, breast cancer and retinitis pigmentosa, have been successfully identified through pooled sequencing strategies [Benaglio, et al. 2011; Out, et al. 2009; Rivas, et al. 2011]. There are two general pooling paradigms: pooling DNA from different individuals without tagging or other methods to distinguish whose DNA contributed a particular observed polymorphism; and pooling tagged samples (multiplexing samples with barcodes prior to pooling [Smith, et al. 2010]), which allows the identifications of the individuals in a pool. Because tagging is both laborious and more expensive, pooling without barcoding is still a popular choice when the target regions are small (too small to use sequencing lane space efficiently) or when the project cost is a main consideration. In this paper, pooling refers to pooling without barcoding, unless otherwise specified.

Because of the practical utility of pooling, it is not surprising that several research groups have considered both analytical methods [Kim, et al. 2010; Wang, et al. 2010] and design issues for pooling [Lee, et al. 2011]. These methods are usually based on a set of simplified assumptions for pooled sequencing, such as a distributional assumption on individual contributions to the pool and a constant sequencing error rate (e.g. 1%) across the sequenced regions (i.e., error rate identical for all sites and all pooled individuals). Most of the published work also assumes that the numbers of reads from the constituent pooled individuals follow the multinomial distribution with parameters equal to the DNA contributions of the individuals in the pool and that the number of reads carrying a certain allele at a heterozygous site for an individual follows a binomial distribution with $p = 0.5$. In reality, the pooling process is far more complex than that described by these assumptions, which cannot capture the many factors contributing to the observed variation in pooled sequencing. Empirical data suggest that amplification during target capture and sequencing, formation of pools, varying base-calling errors, and alignment of reads after sequencing all contribute to biases and variation in the observed pooling data. There is usually a big gap between the variation observed in real data and that predicted based on the simplified assumptions. Therefore, analytical methods and optimal pooling schemes developed under these simplified assumptions may not perform well in practice and can lead to biased results and inefficiently designed studies.

The main purposes of the present study are to (1) identify and characterize the main factors contributing to variations in the observed pooling data; (2) quantify the magnitudes of their effects, and (3) design a workflow to minimize their effects on data analysis and interpretation. These will allow us to gain a better understanding of the observed variation resulting from pooling processes and offer a practical pipeline for pooled next-gen sequencing data analysis. We considered a pooled sequence data set generated from barcoded samples that were sequenced together in a lane. We analyzed these data either based on individuals using the barcode information or as a combined set ignoring barcoding. We also had SNP chip genotype data for some of the same subjects.

The paper is organized as follows. We first introduce the simplified assumptions of pooled sequencing data in the existing literature and then discuss the main factors contributing to variation in pooled samples from empirical data. These factors include varying sequencing error rates, reference allele preferential bias and amplification biases at both individual and pooled levels. After the effects of these factors are inferred from the empirical datasets, we investigate their impact on the accuracy of allele frequency estimation and the power of disease association studies. Finally, we describe a workflow to analyze the pooled sequencing data for better allele frequency estimates, leading to an increase in the power of association tests.

METHODS

SIMPLIFIED ASSUMPTIONS

As discussed above, a number of assumptions have generally been made on the pooling process to facilitate the design and simulation studies of disease association methods [Kim, et al. 2010; Wang, et al. 2010] and optimal pooling designs [Lee, et al. 2011]. These assumptions include: individual contributions to the pool (that is, percent of each sample's sequencing in the final pool); region capture efficiency; read depth for individuals at each position; and read depth of different nucleotides at heterozygous sites of one individual. Although these assumptions are introduced to describe various factors contributing to the observed variations of real data in allele frequencies, they are nonetheless simplified and there may be discrepancies between these simplified assumptions and reality. In the following, we use n to denote the number of individuals in a pool and R to denote the average read depth of the pool. To facilitate discussion, we define *site*, *base* and *nucleotide* as follows: “site” refers to one position in the genome; “base” refers to one base in a read; and “nucleotide” refers to the nucleotide type A, T, C or G.

Individual contributions—It is desirable to have equal contributions from the individuals in a pool so as to minimize the bias of allele frequency estimation, and efforts are made to achieve equal contributions. However, variations not only exist but can be substantial across individuals based on discrepancies that commonly occur throughout the sample preparation process [Lee, et al. 2011]. To model variations across individuals, the Dirichlet distribution is often used

$$p_1, p_2, \dots, p_n \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_n) \quad (1)$$

where p_i is the i^{th} individual's contribution to the pool, α_i is the parameter in the Dirichlet distribution, and it is often assumed that $\alpha_1 = \alpha_2 = \dots = \alpha_n = \alpha$ due to the exchangeability of the samples in a pool where a larger α implies more homogeneity among the individual contributions.

Total read depth at one site—Read depth at a site is often modeled by a Poisson distribution with mean equal to the average coverage across all sites. However, sequence composition and complexity introduce more variation in read depth across the genome than that can be modeled by a Poisson distribution. One approach is to first model the relative read depth of a region by the Gamma distribution with mean equal to 1 and shape parameter equal to $1/b$. Conditional on the depth for a given region, denoted by T , the read depth at a site in the region is then modeled by a Poisson distribution the region specific average read depth, i.e.

$$T_k \sim \text{Gamma}\left(b, \frac{1}{b}\right) * R \quad \text{and} \quad S_j \sim \text{Poisson}(T_k) \quad (2)$$

where T_k is the average coverage of the k^{th} region, and S_j is the read depth of the j^{th} site.

Read depth of individuals at one site—The read depth of individual i is related to the proportion of individual DNA sample in the pool, denoted by p_i in Eq. (1), and the total sequencing depth at the site, denoted by S_j in Eq. (2), when differential amplification of alleles and other factors are not considered. The read depth of the i^{th} individual at this site, denoted by C_i , can be modeled by a multinomial distribution.

$$C_1, C_2, \dots, C_n \sim \text{Multinomial}(S_j, p_1, p_2, \dots, p_n) \quad (3)$$

Read depth of alleles of individuals at heterozygous sites—For diploid organisms such as humans, the read depth of a given allele at a heterozygous site is usually modeled by a binomial distribution with $p = 1/2$. [Li, et al. 2009]. Let X_m be the read depth of the allele m of individual i .

$$X_m \sim \text{Binomial}(C_i, \frac{1}{2}) \quad (4)$$

EXPERIMENTAL DATASETS

In the following, using two real data sets, we show that the above simplified assumptions do not capture all the variations in the observed pooled sequence data. We also use these two datasets to estimate the empirical biases and variation in the sequencing data and evaluate the performance of our proposed workflow for allele frequency estimation.

Pooled sequencing datasets—A total of 23 individuals were analyzed through pooled sequencing. Pooled dataset 1 included 12 individuals, where 7 individuals were genotyped using Illumina HumanOmni1-Quad v1.0 chip genotyping arrays. Pooled dataset 2 included 11 individuals, where 2 individuals were genotyped using the genotyping arrays. Multiplexing with barcodes was used to obtain sequencing data for each individual in the same pool. These barcodes give us the ability to analyze variations both at the pool and at the individual levels. See supplementary methods for detailed data production.

The process of experimental datasets—Raw sequencing reads were mapped to the human reference genome (hg18) by Burrows-Wheeler aligner (BWA) [Li and Durbin 2009] with three mismatches and no gaps allowed. Reads outside the targeted regions and clipped mapped reads were discarded. After mapping, GATK [DePristo, et al. 2011] was used to recalibrate the base quality scores of mapped reads, and SAMtools v0.1.16 [Li, et al. 2009] was then used to call the genotype of each site with the following parameters: mpileup -C50 -m2 -F0.0005 -d 10000 -P ILLUMINA. Additional filters were applied, such as requiring SNP calling quality to be > 10 and coverage > 8 , to gain more accurate variant calling results. For paired-end sequencing (sequencing both forward and reverse ends of each DNA fragment in turn), we kept only read pairs for which both ends mapped. For the two pooled datasets, barcoding assigns a unique identifier (6bp tags added to the 3' end adapter) to each individual, so the sequence data of each individual can be gathered by isolating sequences with unique tags. This enabled us to process the pooled datasets either by individuals or by

the whole pool. The former method together with the microarray genotyping data for some of the same subjects is used to detect errors and biases at the level of individual sequencing; the latter allowed us to test aspects of pooling independent of tagging information. In the following sections, individual sequencing refers to tag-isolated individual samples from pools; pooled sequencing refers to the original pools without separation. The genotypes called by SAMtools from individual sequencing are combined to estimate the ‘true’ allele frequency of variants in one pool, and then the “true” allele frequency is compared with the frequency estimated from the pooled analysis, which allows us to evaluate the workflow for allele frequency estimation of pooled sequencing.

VARIABILITY ESTIMATION

In this section, we consider the main factors contributing to the observed variations in pooled sequencing data. These factors include sequencing error rate, reference allele preferential biases, and unequal amplification (within-sample and cross-sample amplification biases).

Sequencing error rate—The average sequencing error rate ranges from 1% to 3% for the current Illumina sequencing platform [Nielsen, et al. 2011; Richter, et al. 2008]. The error rate may vary across sites according to sequence complexity, base pair composition and sequencing read depth. To estimate error rate distribution, we used homozygous SNP genotyping data as the “gold standard” and evaluated the sequencing calls at these sites. For each homozygous site, we classified those bases for which the nucleotide type called by sequencing was different from that by genotyping as errors. More formally, the estimated error rate at a site, denoted by er_j is defined as

$$er_j = \frac{e_j}{S_j} \quad (5)$$

where e_j is the number of incorrect bases by sequencing at site j and S_j is the read depth at this site. Because SNP genotyping may also be incorrect with the estimated genotyping error rate 0.1% (Illumina Technical Note, http://www.illumina.com/Documents/products/technotes/technote_genotyping_rare_variants.pdf), the estimated sequencing error rate at those sites where there is a genotyping error can be extremely high, e.g. >50%. Therefore, to reduce the impact of inaccurate SNP genotype calls of microarray genotyping, we only consider sites with an estimated sequencing error rate <10%.

Reference allele preferential bias—Preference in reference allele designation might be caused by target capture bias or read mapping bias. Target capture methods are mainly hybridization-based or PCR-based. For hybridization-based methods, sequencing including the non-reference allele is less likely to be captured efficiently because of capture mismatch; for PCR-based methods, polymorphism that occurs in a primer sequence would affect capture efficiency as well. These two situations may lead to the preferential capture of the reference allele. However, the bias may not be significant for PCR-based methods since the primer sequences are usually much shorter than the whole target region and primers are typically designed to exclude regions including known SNPs. Regarding read mapping bias, the first step of sequence data analysis is either mapping the raw reads to a reference genome by an aligner, or de novo construction of fragments with the raw reads. Mapping to a reference genome is faster and more accurate than de novo mapping, if the reference genome is of high quality and the individual reads do not have many SNPs or sequencing errors that result in mismatches for alignment. However, the reference genome only represents one of the two alleles at a polymorphic site. Therefore, for heterozygous sites in

any individual, it is a critical point that the reads harboring the reference allele are more likely to be mapped than the reads containing the other allele. This is because the reads with the alternative allele might be discarded by the aligners due to additional mismatches compared to the reference genome. Thus, aligning reads to the reference genome may lead to preferential mapping of reference allele. Both preferential capture and preferential mapping lead to reference allele preference bias, so we use a single parameter to estimate the total preference rather than consider them separately.

To estimate the bias in favor of the reference allele (which, on a chance basis that is proportional to allele frequency, may not necessarily be the major allele at a polymorphic site), we considered the individual sequencing data at heterozygous sites from SNP genotyping. The observed distribution of the proportion of reference alleles at each site was compared to the binomial distribution with mean equal to 0.5. The bias is represented by the new sampling parameter of the reference allele, which is obtained by the mean of the proportions of reads that carry the reference allele mapped across all heterozygous sites:

$$B_r = \frac{\sum_{j=1}^m rf_j / C_j}{m} \quad (6)$$

where B_r is the bias towards the reference allele, rf_j is the number of reads mapped to the reference allele at site j , C_j is the total coverage at this site, and m is the number of the heterozygous sites of sequencing data. Based on the properties of capture method used and sequencing methods, B_r represents the preference of capture and/or mapping for sequencing of hybridization-based capture targets. We note that B_r is mostly due to mapping bias for PCR-based capture targets and for whole-genome sequencing.

Because the mapping bias is aligner-dependent, we also examined whether we could quantify the bias based on sequencing data alone without the corresponding SNP genotyping array, in the case that a new aligner is used or there is no SNP genotyping array presented. To quantify the bias from sequencing data, we extracted the robust heterozygous positions called from Samtools. We define robust SNP positions as the positions with SNP calling quality score larger than 40 and the reference allele frequency within 0.05–0.95. For these positions, we calculated the reference preferential bias as the method we used for the estimation from the SNP sets of SNP genotyping array. Then, we examined whether the biases estimated from these two independent methods were similar. If they agree, this would suggest that we may be able to estimate the reference allele preferential bias well for any aligner based on the sequencing data we have.

Within-sample amplification bias—Amplification of alleles is a stochastic process in target capture and sequencing, and this generates additional variance to sequencing as well. To test the difference in the amplification of alleles at one site in an individual, we examine the level of amplification of each of the two alleles at heterozygous sites in the individual sequencing data. This is accomplished by obtaining the set of reads before amplification through removing potential duplicates of alleles and adjusting read count by Phred base quality scores (details are in workflow section). The amplification level is then calculated by the number of raw reads of the allele divided by the number of read counts before amplification.

Cross-sample amplification bias—Besides the directly observable instances where there are two different alleles at a site where we have individual sequencing data to estimate relative amplification rates, individuals in different sequencing pools may also have different amplification rates. The difference of amplification rates across individuals can be estimated from the pooled datasets. The set of reads before amplification for each individual

is obtained by the same method described for that of within-sample amplification bias. The amplification level is then calculated by the comparison of the counts before and after amplification.

SIMULATIONS

To understand the effects of biases and errors (estimated from the empirical data) on the accuracy of allele frequency estimation and the power of disease association tests of pooled samples, two simulation scenarios were considered. In the first scenario, the data set was simulated under the simplified assumptions used in the previous literature; whereas the second scenario incorporated the empirical errors and biases described above, which we call the ‘full model’ in this paper. The reference allele preferential bias was incorporated into the full model by replacing the binomial parameter 0.5 with the biased parameter B_r defined above. Within-sample amplification bias, cross-sample amplification bias and sequencing error rate were sampled randomly from those of the real data instead of assuming constant bias or error parameters. Since the sequencing errors depend on the read depth (Supp Figure 1.), the variability cannot be sufficiently captured if a constant error rate is used or the error rate is sampled from the same distribution regardless of the sequencing depth. Therefore, in our study, we tried to mimic the variability of the errors observed in real sequencing data in our simulations. More specifically, we divided the read depths into bins and sampled the sequencing error at a specific position according to the empirical error distributions from the bin covering the read depth at this position. In the simulation, we consider pooled sequencing with a total of 12 individuals in a pool.

To evaluate the accuracy of the allele frequency estimation, we simulated 1000 pools for common variants having population allele frequencies from 0.1 to 0.9 with a step size of 0.05 and rare variants with frequencies 0.01 and 0.05, respectively. We also considered allele frequencies of 0.95 and 0.99 for the case where the reference genome includes low frequency variants. Individual genotypes were generated assuming Hardy Weinberg Equilibrium with the specified allele frequencies. The true allele frequencies in the pools were calculated from the underlying individual genotypes. The estimated allele frequency of a pool was obtained by the ratio of the read depth of the allele to the total depth. For each population allele frequency, root mean squared error (RMSE) was calculated by comparing the true and estimated allele frequencies. The estimator $\hat{\theta}$ for θ_0 , which is the difference between the true and the estimated allele frequency in the pools is

$$\hat{\theta} = E_a - T_a \quad \text{and} \quad RMSE = \sqrt{Var(\hat{\theta}) + \{E(\hat{\theta} - \theta_0)\}^2} \quad (7)$$

where E_a is the estimated allele frequency from raw counts, T_a is the true allele frequency from the true genotypes, and θ_0 is equal to 0. To assess the contribution of each factor to the difference of simplified model and full model, we added one factor at a time and compared RMSE of simplified model and the model with one factor added while controlling for others.

To assess the power of disease association tests, we simulated 600 individuals (50 pools) for the cases and controls, respectively. For disease model, we assumed that $P(\text{disease}|AA) = 0.01$, $P(\text{disease}|Aa) = P(\text{disease}|AA) * r$, and $P(\text{disease}|aa) = P(\text{disease}|AA) * r^2$, where r is the relative risk. We varied allele frequencies from 0.05 to 0.5 with a step size of 0.05, and the value of r from 1.1, 1.2, 1.3, 1.5, to 1.7 to evaluate power under different relative risks. We estimated the power of the association tests with 1000 replicates at the statistical significance level of $\alpha = 0.05$. We also assessed the type I error rate through 1000 replicates. Although the χ^2 test is commonly used in association analysis, directly applying it to the sequencing count data is not appropriate. One important assumption for the χ^2 test is that

the observations are independent. However, the allele counts from pooled sequencing data do not arise from independent observations, since the read counts in the observed data can be multiple amplification copies of a single allele in an individual. This can significantly inflate the type I error. [Kim, et al. 2010; Wang, et al. 2010]. Therefore, in our analysis, we used the t-test with the estimated allele frequency in each pool as a sample to study associations between SNPs and disease.

WORKFLOW FOR REAL DATA PROCESSING

Based on our observations of the factors contributing to the observed variations in allele frequency estimation, we developed a workflow (Figure 1) to reduce potential biases and errors from real data and improve the estimation of allele frequency, with the object of obtaining more accurate allele calling and therefore more powerful disease association tests. The first two steps of the workflow use existing programs, BWA and GATK, to process the raw sequencing data. In this section, we mainly describe the methods that we developed to reduce amplification bias, sequencing errors and reference allele preferential bias. The reference allele preferential bias is estimated from individual sequencing data described in the “datasets” section, where it is mainly due to the specific aligner used and should be similar for all of the datasets mapped with the same aligner; other factors (amplification biases and sequencing errors) are directly adjusted or removed from the datasets of interest without any prior information.

Amplification biases removal—Amplification biases considered in the workflow are amplification of alleles at one site in one individual or across individuals. Considering that the sequence properties (complexity and composition) of a certain site are almost the same for different nucleotide types and for different individuals, amplification biases considered here usually result from unequal PCR duplications. To reduce potential amplification biases, accurate detection of PCR duplicates is required. In the existing programs [DePristo, et al. 2011; Li, et al. 2009], PCR duplicates are usually defined by reads starting from the same position or reads that are identical. However, these two definitions do not fit our needs here well. For reads starting from the same position, we may lose variant information. Considering reads that are completely identical does not take sequencing errors into account. In our work, we introduce the idea of ‘unique base’ illustrated in Supp Figure 2 to reduce the amplification bias site by site. The properties of a base include the nucleotide type, the position in the read, the direction of the read, and the second end of read pairs if applicable. Bases with the same properties are grouped together and only one of them is kept as the unique base for subsequent analyses. In this way, we may reduce the potential amplification bias in the dataset site by site.

MLE of allele counts incorporating sequencing errors—Phred quality scores Q , which estimate the accuracy of base calling, are logarithmically related to the probabilities of error, denoted by e , as shown in the following equation:

$$Q = -10 \log_{10} e \quad (8)$$

We are not aware of any method presently available to incorporate base quality scores to adjust allele counts. Here, we propose to take advantage of the quality score for each base to estimate more accurately the frequency of each allele. For a given site j , let P_{mj} denote the probability that a sequence carries nucleotide type m at this site within a single individual or within a particular pool. For a given base i at site j , the probability that the observed nucleotide (denoted by O_{ij}) is m would be

$$P(O_{ij}=m)=P(O_{ij}=m|T_{ij}=m)P_{mj}+P(O_{ij}=m|T_{ij} \neq m)(1-P_{mj})=(1-e_{ij})P_{mj}+e_{ij}(1-P_{mj}) \quad (9)$$

where T_{ij} is the true allele at the nucleotide and e_{ij} is the error rate of the i th nucleotide at site j . Considering all the bases at site j , the likelihood of the observed nucleotides for site j can be represented by

$$P(P_{Aj}, P_{Cj}, P_{Tj}, P_{Gj} | (O_{1j}, e_{1j}), (O_{2j}, e_{2j}), \dots, (O_{nj}, e_{nj})) = \prod_{m=A} [(1-e_{ij})P_{mj}+e_{ij}(1-P_{mj})] \prod_{m=T} [(1-e_{ij})P_{mj}+e_{ij}(1-P_{mj})] \prod_{m=C} [(1-e_{ij})P_{mj}+e_{ij}(1-P_{mj})] \prod_{m=G} [(1-e_{ij})P_{mj}+e_{ij}(1-P_{mj})] \quad (10)$$

We use the maximum likelihood estimation (MLE) method to estimate P_{Aj} , P_{Cj} , P_{Tj} and P_{Gj} in Eq. (10) with the Nelder-Mead numerical optimization. P_{mj} multiplying the read depth at site j gives us the adjusted allele count of nucleotide m after considering the sequencing errors.

Reference allele preferential bias—The bias of the reference allele capture and alignment is estimated as B_r as described in Eq. (6). We define the non-reference allele factor, denoted by f , for the adjustment of allele counts of non-reference alleles as

$$\frac{r}{r+(1-r)f} = B_r \quad (11)$$

where r is the probability that a base is mapped to the reference allele without bias and $1-r$ is then the probability that it is mapped to the non-reference allele. For a heterozygous site, r is equal to 0.5, so $f = \frac{1}{B_r} - 1$ ($B_r > 0.5$ and $f < 1$). We adjust the count of non-reference allele as N/f , where N is the raw count for non-reference allele.

Assessment of the proposed workflow—We evaluate the performance of the proposed workflow by comparing the accuracy of allele frequency estimation before and after adjustments. The SNP calling results for barcoded individuals are treated as ‘true’ genotypes which can then be considered in the context of the pools. Estimated allele frequency is calculated by raw allele count and adjusted allele count, respectively, to evaluate the improvement in accuracy of allele frequency estimation accomplished through the proposed workflow. RMSE is calculated for each group of true allele frequency of heterozygous sites. For example, if a pool has 12 individuals, 22 RMSE’s are calculated for allele frequency from 1/24, 2/24, ..., 23/24.

Implementation of the workflow—The workflow is implemented with Perl and R, available at <http://bioinformatics.med.yale.edu/group/>. The software can be run through command lines. Amplification bias removal, sequencing error adjustment and alignment bias adjustment can be run selectively according to users’ needs. A few parameters can be defined by users as well. See Software Manual for details.

RESULTS

VARIABILITY ESTIMATION

We used the barcoded individual sequencing data together with corresponding SNP genotype calls from SNP arrays to characterize the errors and biases at the individual sequence level. We treated the SNP array data as the ‘gold standard’ as discussed in the Methods section. A total of 288 sites were genotyped via the SNP array in each individual,

for a total of 2592 sites from 9 individuals. More than 95% of the genotypes called by sequencing and SNP arrays match for each sample, suggesting that the data are of good quality (Supp Table. 1.). We analyzed all the 2592 sites together for error and bias estimation. After removing the sites with low genotyping quality, 550 heterozygous and 1935 homozygous sites remained.

To infer the sequencing error rates across sites, we considered the 1935 homozygous positions in the target regions. The average sequencing error rate was 1.2%, with most <5%, but some reached a very high level (>10% likely due to low read depth at the sites). Figure 2A shows the distribution of error rates with error rate <5% and this figure suggests that sequencing error rates are not constant and can be very different across sites. We therefore chose to sample error rates from this empirical error distribution in our simulation, rather than assuming a constant error rate.

Heterozygous sites called by SNP genotyping were used to estimate the reference allele preferential bias and within-sample amplification bias. For the former bias, the distribution of the proportion of reads coming from reference allele has its center shifted to the right, compared to a symmetrical 0.5-centered distribution (Supp Figure 3), which shows that the reference allele is indeed more preferred. More specifically, the mean of the preferential representation toward the reference allele is 0.514, which gives an estimated bias B_r in Eq. (6) equal to 0.514, and the non-reference allele factor $f = 0.946$ in Eq. (11). This bias parameter likely applies to sequencing data mapped with the BWA program, since the bias is aligner-dependent. As we mentioned in the Methods section, we also examined the reference allele preferential bias quantified from sequencing data alone without genotyping array information. The bias from sequencing data alone is 0.514 as well (Figure. 2B), which means the bias can be estimated from the sequencing data of interest when a new aligner is used.

Within-sample amplification bias introduces additional variability of sequencing. Amplification is randomly performed on two alleles and not directly controlled, so the two alleles at the heterozygous sites may not be amplified equally (Figure 2C.). To investigate whether the reference allele alignment bias is due to amplification bias, we studied whether amplification had any bias towards the reference allele in our datasets by checking the distribution of the amplification proportion of the reference allele. We found that the reference and non-reference alleles have about the same amplification likelihood since the distribution is symmetrical around 0.5, in contrast to the alignment bias.

The cross-sample amplification bias is estimated from pooled dataset 1. We calculated the amplification rates of individuals in a pool site by site and summarized the mean and standard deviation of the amplification rates across the individuals in Figure 2D, which shows substantial variation as well, e.g. the amplification of individual 5 is about 1.5 times of that of individual 1.

SIMULATION STUDY

To investigate the magnitude of the effects of biases and errors on pooled dataset analysis, two simulation scenarios were generated, one based on the simplified assumptions used in the existing literature, and one based on the empirical data for biases and errors (full model). We compared the accuracy of allele frequency estimation and the power of disease association studies under these two scenarios. For the power analysis, we added a comparison with tests using individuals' actual genotypes, which represent the highest achievable power.

The accuracy of allele frequency estimation is summarized by RMSE, shown in Figure 3A. As expected, allele frequency can be better estimated if the simplified assumptions are true, than the full model that better represents realistic settings. On average, the RMSE under the full model is about 1.5 times of that of the simplified assumptions. Variation for allele frequency estimation is higher when the population allele frequency is higher, since more minor alleles presented in a pool introduce larger variance. In general, the biases and errors present in empirical data lead to a substantial increase in variation of the estimated allele frequencies, compared to the expectations under the simplified assumptions. The results of each factor's contribution (Supp Figure 4) show that the differences between the two models are mainly caused by sequencing errors in the case of rare variants (0.005, 0.01, 0.995 and 0.99) and by amplification biases in the case of common variants. The reference allele preferential bias contributes to the difference of two models for common variants where the reference allele has a high frequency.

More variation in allele frequency estimation would result in reduced power in disease association analysis. In Figure 3B–F, we compared the power under the two scenarios with low, moderate and high relative risks, including 1.1, 1.2, 1.3, 1.5 and 1.7. Analysis based on the individuals' true genotypes has the highest power, as expected. The power decrease under the full model is in general twice that under the simplified assumptions. The reduction is greatest for common alleles and weak effect sizes. For example, when the relative risk is low or moderate (Figure 3C and D), the biases and errors under the full model decrease the power by as much as 10% for common variants. On the other hand, when relative risk is larger (Figure 3E and F), the power is not affected as much. There is relatively less reduction in power for a higher relative risk (1.5 and 1.7) because the difference between cases and controls is larger and there is relatively less information loss with additional variations in allele frequency estimates under the full model. However, most variants associated with common diseases have relatively weak effects, so the biases and errors in pooled sequencing can have significant effects on the estimated power compared to what is expected under simplified assumptions. We also note that the decrease of power under very low relative risk (1.1, Figure 3B) is not very large because the risk is so low that the association signal is hardly identified by single marker methods (Even the power for directly-assigned genotypes is already very low.).

As we mentioned in the methods section, a χ^2 test is not appropriate for sequencing data if it is directly applied to the raw counts. It can be seen in Supp Figures 5 and 6 that the type I error is indeed much higher than the nominal significance level when the χ^2 test is used.

From both comparisons – accuracy of allele frequency estimation and power of association test -- we found that the effects of biases and errors observed in the empirical data are not negligible; they need to be incorporated in association analysis of pooled samples and for optimal design of pooled sequencing, with specific steps taken to reduce biases and errors in the analysis and design comparisons.

REAL DATA PROCESSING WITH PROPOSED WORKFLOW

In this section, we show that our proposed workflow can achieve improved allele frequency estimation from pooled sequencing data.

To assess whether we could reduce the bias in our analysis, we plotted the histogram for the p -values from the goodness of fit statistics used to test whether the two alleles from the heterozygous sites within an individual have equal chance to be represented in the sample (Supp Figure 7A) and whether the individual contributions to the pool across the region are similar to the contribution of individual DNA samples. (Supp Figure 7B). The former is based on a binomial distribution with $p = 0.5$ and the latter is based on a multinomial

distribution with parameters according to the relative DNA contribution from each individual. These figures are plotted with pooled dataset 1, and results from pooled dataset 2 are similar (data not shown). These figures demonstrate that the raw data from the pooled sequencing show a significant departure from equal representations of the two alleles at heterozygous sites, and also show a significant departure from constant contributions to the pool across the region. These large deviations lead to a decrease in the accuracy of allele frequency estimation and power of association tests. However, after we adjusted the sequencing read counts with the proposed workflow, there was an improved fit to the predicted equal allelic representation at heterozygous sites and constant contributions to the pool. Our results suggest that the proposed method can substantially improve the analysis of pooled samples. These results also suggest that we are able to capture the main factors introducing biases and errors in pooled sequencing data analysis, and that the simulated dataset under the full model may well represent the variation of realistic sequencing of pooled samples.

In addition to the above goodness-of-fit tests, we also checked the validity of our method on sequencing error rate adjustment (Figure 4). With the MLE of allele counts based on the recalibrated base quality scores, the average sequencing error rate decreased from 1.2% (Figure 2A) to 0.8%. With adjusted data, about 60% of the polymorphic sites have no errors at all; while this is the case for fewer than 25% of sites for the raw data. Thus, the proposed workflow does lead to very effective control of these kinds of errors. The simulation results above show that the variability of allele frequency estimation of rare variants in pooled data is mainly due to sequencing errors. Therefore, our proposed workflow can address the critical problem of rare variants analysis in pooled sequencing data.

To evaluate the improvement in allele frequency estimation with the proposed workflow, we tested our workflow on two independent realistic pooled sequencing datasets with 12 individuals and 11 individuals in each pool, respectively (Figure 5). The total number of non-overlapping high quality variants combined from individuals in a pool is 1607 for dataset 1 and 1547 for dataset 2. We considered both pools with barcoding and pools without barcoding. When pools with barcoding were considered (Figure 5A dataset 1 and Figure 5B dataset 2), we adjusted the allele counts for each barcoded individual first and then combined the allele counts of individuals in the pool to gain the adjusted allele frequency estimation. Raw allele frequency estimation was derived by combining raw counts in the individual sequencing pileup. Figures 5A and 5B show that the adjusted data can estimate the allele frequencies substantially better, since almost all RMSEs of adjusted data were smaller than that of raw data. More than 75% of them were improved by 0.01 or more in RMSE. Variance and bias – two components of RMSE – are examined individually in Supp Figure 8. The biases for raw data and adjusted data were similar and the improvements in RMSE were mainly gained from reduced variance. When pools without barcoding were examined, we directly applied our workflow on pooled sequencing reads without individual identification. Pools without barcoding present the more challenging and immediate problem, since estimating allele frequency through raw counts is currently the only way to analyze pools without barcodes currently. The RMSE results are shown in Figure 5C for dataset 1 and Figure 5D for dataset 2. The majority of RMSEs of adjusted data were smaller than that of raw data, and about 30% of dataset1 and 50% of dataset2 improved RMSE by at least 0.01; those that did not decrease had very similar RMSE values to the raw data. Variance (Supp Figure 9A and C) and bias (Supp Figure 9B and D) – two components of RMSE – were examined individually as well. As expected, our method reduced the variance of allele frequency estimation. Pools without barcodes had a smaller decrease of RMSE than pools with barcodes, because the adjusted data had higher bias than the raw data on the allele frequency estimation. The reason for the higher bias might be that some individuals in a pool present the same ‘unique base’ at a site by chance. When we remove

the potential duplicates from a pool, we cannot determine whether they are duplicates from one individual or alleles from different individuals, so we simply treat those bases as duplicates and only keep one of them. However, when we consider bias and variance together, the variance is the dominating factor. The workflow can reduce the variance on each allele frequency substantially, compensating for the increase in bias.

DISCUSSION

In this study, we investigated the performance of allele calling based on targeted sequencing of pooled samples in disease association tests, and provided a workflow to control the errors and biases in pooled sequencing data analysis. We were able to compare different approaches because we had both SNP genotyping from chip genotyping microarrays and individual-tagged sequence data from the same set of individuals. Ignoring tags in the pooled data analysis allowed us to investigate the effect of untagged pooling.

We first identified the main factors that can contribute to the variations in the allele frequency estimation in targeted sequencing of pooled samples. These factors have significant impact on the accuracy of allele frequency estimation and the power of disease association tests. The factors identified include varying sequencing error rates, reference allele preferential bias, within-sample amplification bias, and cross-sample amplification bias. Our simulations compared the results from two different scenarios – one under the simplified assumptions commonly used in the existing literature and one under the full model that incorporates the empirical biases and errors extracted from the empirical pooled datasets. Our results showed that the RMSE of allele frequency estimates under the full model is about 50% higher than that under the simplified assumptions. For disease association analysis, the simulation results suggested that these factors can reduce the statistical power, especially when the variants have relatively small effect on disease risk. Considering that current analysis methods for pooled sequencing do not take into account the main factors of variations considered in this paper, we believe that it is necessary to design a workflow to consider these factors in the analysis. In addition, for optimal pooling design, investigators might want to take into account these factors.

Here, we have described a workflow that can substantially adjust or nearly remove the biases and errors for pooled sequencing data to better estimate allele frequencies and thus increase the power of disease association analysis. The workflow consists of potential amplification bias removal, sequencing error rate adjustment by MLE with recalibrated base quality scores, and reference allele preferential bias adjustment. The results of the adjusted pooled data from the proposed workflow showed that the workflow can increase accuracy of allele frequency estimation for pools with or without barcodes. The strength of disease association tests is highly dependent on the accuracy of allele frequencies of cases and controls; improved allele frequency estimation will increase ability to extract association signals from pooled samples. Another improvement made by our workflow is that the MLE sequencing error adjustment reduces the impact of sequencing errors by incorporating base quality scores. High sequencing error rate is a problem for identification of rare variants from pooled sequence data. With the workflow, we provide an effective pipeline, particularly for rare variant analysis. In the workflow, we adjusted allele frequency by BWA alignment bias of reference allele; and the bias can be accurately estimated by individual sequencing dataset when a new aligner is used, as we showed in the section of ‘variability estimation’. However, it would be challenging if the sequencing data are pooled without barcoding because the variability of other factors would affect the estimation of reference allele bias. Since the bias is aligner-dependent, one approach is to use publicly available individual sequencing dataset, e.g. 1KG project, to estimate the bias first and then apply it to the analysis of pooled sequencing.

While the workflow may be used for pools with or without barcodes (or individual sequencing), the processing of pools without barcodes shows higher bias. This is possibly due to some individuals in a pool presenting the same ‘unique bases’ at a site by chance. Thus, one limitation of the workflow for pools without barcodes is that it is less suitable for pooled datasets with many individuals or pooled samples of single-end sequencing, since it is more difficult to distinguish the duplicated reads from real individual copies in these two situations.

The proposed workflow is not restricted to the analysis of pooled sequencing. For example, for RNA-seq data, it may not be desirable to remove all the potential duplicates, but sequencing errors and reference allele preferential bias need to be adjusted. The last two steps of the proposed workflow can be used separately for RNA-seq data analysis.

In summary, we have identified the main factors contributing to the variance in pooled sequencing data and demonstrated the large effect of these factors on the accuracy of allele frequency estimation and the power of disease association tests. A workflow has been designed to reduce the impacts of these factors. The workflow can also be applied in other sequencing contexts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Joanne Weidhaas and Trupti Paranjape for a discussion of pooling protocols. We thank John Ferguson and Gengxin Li for helpful discussions. We thank Zach Pincus for careful reading this manuscript. We also thank Yale University Biomedical High Performance Computing Center (funded by NIH RR19895) for data storage and computation runs. Genotyping services were provided by the Center for Inherited Disease Research (CIDR) [funded through a federal contract from the National Institutes of Health to Johns Hopkins University (contract number N01-HG-65403)]. XC was supported by a fellowship from the China Scholars Council. JG and JBL were supported by NIH grants DA12849, DA12690, AA017535, AA12870, and DA028909. FJS was supported by a grant from an anonymous foundation. HZ was supported by NIH R01 GM59507.

References

- Benaglio P, McGee TL, Capelli LP, Harper S, Berson EL, Rivolta C. Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa. *Hum Mutat.* 2011; 32(6):E2246–58. [PubMed: 21618346]
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011; 43(5):491–8. [PubMed: 21478889]
- Kim SY, Li Y, Guo Y, Li R, Holmkvist J, Hansen T, Pedersen O, Wang J, Nielsen R. Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet Epidemiol.* 2010; 34(5):479–91. [PubMed: 20552648]
- Lee JS, Choi M, Yan X, Lifton RP, Zhao H. On optimal pooling designs to identify rare variants through massive resequencing. *Genet Epidemiol.* 2011; 35(3):139–47. [PubMed: 21254222]
- Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature.* 2010; 465(7297):473–7. [PubMed: 20505728]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25(14):1754–60. [PubMed: 19451168]
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25(16):2078–9. [PubMed: 19505943]

- Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 2011; 12(6):443–51. [PubMed: 21587300]
- Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PE, Tops CM, et al. Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat.* 2009; 30(12):1703–12. [PubMed: 19842214]
- Pierucci-Lagha A, Gelernter J, Feinn R, Cubells JF, Pearson D, Pollastri A, Farrer L, Kranzler HR. Diagnostic reliability of the Semi-structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend.* 2005; 80(3):303–12. [PubMed: 15896927]
- Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One.* 2008; 3(10):e3373. [PubMed: 18841204]
- Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burt N, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet.* 2011; 43(11):1066–73. [PubMed: 21983784]
- Shah SP, Morin RD, Khattri J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature.* 2009; 461(7265):809–13. [PubMed: 19812674]
- Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* 2010; 38(13):e142. [PubMed: 20460461]
- Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature.* 2009; 462(7276):1005–10. [PubMed: 20033038]
- Wang T, Lin CY, Rohan TE, Ye K. Resequencing of pooled DNA for detecting disease associations with rare variants. *Genet Epidemiol.* 2010; 34(5):492–501. [PubMed: 20578089]

Figure 1.

Workflow for pooled sequencing data processing. Reads are aligned by BWA and then Phred quality of bases is recalibrated by GATK (optional). Potential amplification biases are removed and allele counts are adjusted with maximum likelihood estimation. Then reference allele preferential bias is adjusted. Finally, allele frequency of pooled samples can be estimated from adjusted read counts of sequencing data.

Figure 2.

Empirical errors and biases estimated from realistic data. A–C are on the level of individual sequencing; D is on the level of pooled sequencing. A: Empirical distribution of sequencing error rate (only $\leq 5\%$ are shown); B: Distribution of reference allele preferential representation. Dashed line: symmetry parameter = 0.5. The mean of reference allele preferential bias is 0.514, while the median is 0.511. C: Amplification rate of two alleles (log value) at heterozygous positions. D: Amplification rate of individuals in one pool.

Figure 3.

Comparison of simulated data with simplified assumptions and simulated data with full model of errors and biases. A: Root mean square error (RMSE) of allele frequency estimation of pooled samples on population allele frequency ranging from 0.01 to 0.99; B – F: Power of disease association test of pooled samples on population allele frequency ranging from 0.01 to 0.5 with different relative risks (B: RR 1.1; C: RR 1.2; D: RR 1.3; E: RR 1.5; F: RR 1.7). Red lines represent simulated data with simplified assumptions; blue lines represent simulated full model with errors and biases added; black lines represent test results with assigned genotypes.

Figure 4.
Empirical distribution of sequencing error rate ($\leq 5\%$) after adjustment.

Figure 5.

The RMSE of allele frequency estimation of two realistic pooled datasets with or without barcodes. X-axis represents the RMSE of the estimation from raw data; Y-axis represents the RMSE of the estimation from adjusted data. A: dataset1 with barcodes; B: dataset2 with barcodes; C: dataset1 without barcodes; D: dataset2 without barcodes. Each dot represents the result for one certain true allele frequency of variants in a pool. True allele frequencies of variants in dataset 1 with 12 individuals in one pool are $1/24, 2/24, \dots, 23/24$; true allele frequencies of variants in dataset 2 with 11 individuals in one pool are $1/22, 2/22, \dots, 21/22$.