Detecting and Correcting Contamination in Genetic Data

by

Matthew Flickinger

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2016

Doctoral Committee

Professor Michael L. Boehnke, Chair
Professor Gonçalo Abecasis
Professor David T. Burke
Assistant Professor Hyun Min Kang

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter 1
Introduction

Technological innovation in the past decade has dramatically increased the amount and variety of genomic data available to geneticists. While it took over a decade to sequence the first human genome, a new sample today can be sequenced in a few hours according the product manuals for “next-generation” sequencing (NGS) technologies. Additionally, array-based genotyping methods can assay millions variants in a matter of minutes. With faster data generation, we also see lower unit costs. The long awaited $1000 genome is finally within reach. These advances have enabled large-scale genomic studies that would have been impossible just a few years ago.

While high-throughput technologies have increased the number of samples that can be analyzed, they have also increased the opportunities for errors to occur. Even if per-experiment errors are relatively rare, the large number of experiments performed means it is likely errors will occur. Imperfect methods and protocols may result in systematic biases or errors in the generated data. If ignored, these errors may result in inaccurate genotypes which could lead to false associations between genotypes and a trait of interest or reduced power to detect such associations. For this reason, proper quality assessment is an important part of any data processing pipeline. If possible, test for errors should be easy to execute and interpret in an automated way as early in the experimental pipeline as possible.

One potential source of inaccurate genomic data is sample contamination. We define contamination as the accidental mixture of DNA or RNA from two or more individuals from the same species. Contamination between different species is also possible[1,2], but we will focus on the more challenging problem of contamination among human samples. Throughout the course of a study, physical samples may be handled and manipulated in the laboratory and it is possible for two DNA or RNA samples to become mixed. Anytime samples are pipetted (e.g. during collection, storage, or extraction) there is an opportunity for contamination. Improper shipment of samples in well plates from a repository to a processing center may also result in contamination. Additionally, some protocols require forms of PCR amplification and if multiple, barcoded samples are processed together, there is a risk of amplification errors that may result in a portion of the DNA being paired with the wrong barcode. One final example involves improper data merging. Some protocols sequence samples across many batches and then combine the data before processing. Even if all sequenced samples were uncontaminated, incorrect data merging may result in a data file that appears to be contaminated.

There is a need for methods to detect sample contamination for both array-based and sequencing-based data. For array-based genotypes, samples are often filtered or excluded based on the number of missing genotypes. It is possible that the cause of missing genotypes is contamination. However there are other possible causes as well and no attempt has been made to test or quantify contamination specifically for array-based genotypes. For sequencing data, there are methods that look at cross-species contamination by filtering out sequence during alignment[3], and there is a Bayesian method that additionally requires you to know the true genotypes (presumably from an array-based method)[4]. Because contamination increases

the diversity of alleles observed at a particular variant site, contaminated samples will generally have more heterozygous genotypes than expected for uncontaminated samples. Filtering individuals with a large number of heterozygous SNPs is a useful quality filter but it does not help quantify contamination and it also requires prior running of a genotype caller on the sample, an additional time-consuming step which could mean delays on the order of months before contamination is detected, allowing a contamination-prone process to continue.

In this thesis, I propose a comprehensive set of tools for dealing with contamination in modern genetic data. In chapter 2, I will look at detecting and quantifying contamination in both array-based genotyping data and NGS data. In chapter 3, I focus specifically on NGS data and propose a novel genotype calling algorithm that can produce accurate genotypes even when samples are contaminated. In chapter 4, I extend the methods for contamination detection to RNA sequencing (RNA-Seq) data. Finally in chapter 5, I reflect on the usefulness of these methods and describe possible future extensions.

Chapter 1 References

1) Tanner, M. A., Goebel, B. M., Dojka, M. A., Pace, N. R. (1998). Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. Appl Environ Microb. *64*, 3110-3113.
2) Longo, M. S., O'Neill, M. J., O'Neill, R. J. (2011). Abundant human DNA contamination identified in non-primate genome databases. PLoS One. *6*, e16410.
3) Schmieder, R.A.E., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE *6*, e17288
4) K. Cibulskis, A. McKenna, T. Fennell, E. Banks, M. DePristo, G. Getz. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601–2602

Chapter 2
Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data

Abstract[A]

DNA sample contamination is a serious problem in DNA sequencing studies, and may result in systematic genotype misclassification and false positive associations. While methods exist to detect and filter out cross-species contamination, few methods to detect within-species sample contamination are available. In this paper, we describe methods to identify within-species DNA sample contamination based on (1) a combination of sequencing reads and array-based genotype data; (2) sequence reads alone; and (3) array-based genotype data alone. Analysis of sequencing reads allows contamination detection after sequence data is generated but prior to variant calling; analysis of array-based genotype data allows contamination detection prior to generation of costly sequence data. Through a combination of analysis of *in-silico* and experimentally contaminated samples, we show that our methods can reliably detect and estimate levels of contamination as low as 1%. We evaluate the impact of DNA contamination on genotype accuracy, and propose effective strategies to screen for and prevent DNA contamination in sequencing studies.

[A] This work has been published: Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91, 839–848.

Introduction

Advances in array-based genotyping and next-generation sequencing have resulted in higher throughput, lower costs, and reduced error rates. These technologies enable increasingly comprehensive genetic studies for a wide range of human diseases and traits. While constantly improving, genotyping and sequencing technologies are not perfect, and careful attention must be paid to ensure high data quality. Sensitive and efficient methods to screen data for potential artifacts are critical.

One potential source of error is DNA sample contamination. Because samples are often processed in batches and genotyping and sequencing protocols require multiple steps of sample handling and manipulation in the lab, it is not surprising that DNA from more than one individual may end up in the same well or prepared library. In this paper, we focus on within-species contamination in which DNA from more than one individual is present, either from another individual in the same study or from an unknown individual. Note that cross-species contamination can often be detected and filtered out during the alignment of sequence reads[1]. Within species contamination is harder to detect, and can result in greatly reduced genotype quality for sequencing studies; the problem is most severe for low pass sequencing studies (where each allele is typically supported by only a few reads), but can affect even deep sequencing studies.

In a recent type 2 diabetes sequencing study, we identified a subset of individuals with unusually large numbers of heterozygous genotypes and high ratios of heterozygous genotypes to non-reference allele homozygous genotypes (HET/HOM ratio) (Figure 2-1AB). We

hypothesized that some DNA samples might be contaminated, resulting in poor genotype estimates and inflated heterozygosity and, therefore, set about to develop methods to identify such contamination and estimate its extent.

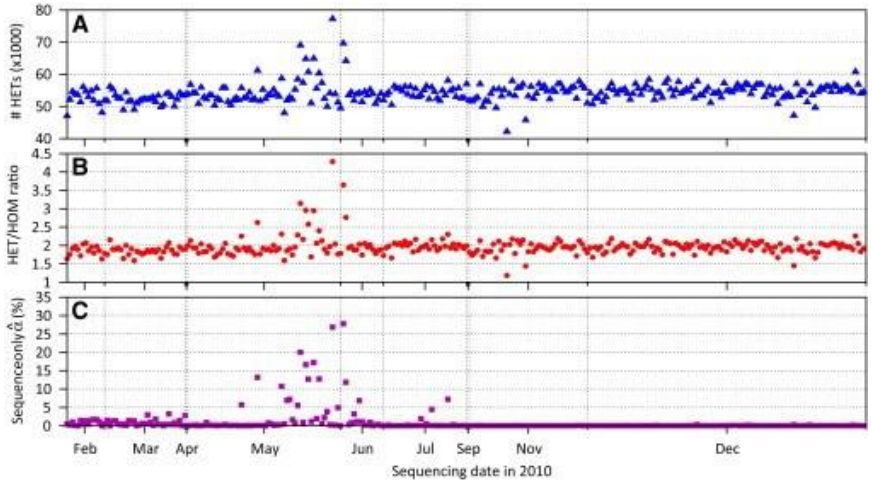

Figure 2-1 - ***SNP genotype calling and estimation of contamination from 299 European sequenced samples: chromosome 20 -*** *A. Numbers of heterozygous genotypes. B. Ratio of the numbers of non-reference homozygous genotypes to heterozygous genotypes (HET/HOM ratio). C. Estimated level of DNA sample contamination estimated from sequence data only.*

Here, we describe methods to detect DNA sample contamination based on sequencing and/or array-based genotype data. We demonstrate that when sequencing is carried out on DNA samples for which array-based genotypes are available, it is possible to estimate the level of sample contamination, and to identify the source of the contamination (see Web Resources)[2]. We further demonstrate that even with low-pass sequencing data alone, we can detect and estimate the degree of contamination. Finally, and perhaps most important, we demonstrate that it is possible to detect even modest levels of DNA sample contamination from array-based genotype data alone, allowing DNA samples to be pre-screened for possible

contamination prior to sequencing. Software based on our methods is already in use by major sequencing projects, including the 1000 Genomes Project, and is publicly available (see Web Resources).

Materials and Methods

In this section, we first describe a series of methods to evaluate DNA sample contamination and then outline a series of experiments carried out to evaluate our ability to identify contaminated samples. We present three likelihood-based methods that detect DNA sample contamination using (a) sequence data and array-based genotype data, (b) sequence data alone, and (c) array-based genotype data alone. We also present a regression-based method that uses array-based genotype data alone. For each of these methods, we assume that if DNA from a "contaminating sample" represents a fraction α of the observed data, then the same fraction α of sequence reads and genotype array intensity will be contributed by the contaminating sample. Initially, we also assume the presence of no more than one contaminating DNA sample (but see Discussion).

Detecting sample contamination using sequence data and array-based genotype data jointly

We first consider the simplest situation where a set of genotypes for each sequenced sample is known and we wish to investigate whether sequencing reads all originate from the targeted sample with no evidence for contaminating reads from a different sample. For each site , let g_i be the true genotype, b_{ij} $(1 \leq j \leq R_i)$ be the base call for the j^{th} overlapping base (among R_i total reads overlapping site i and passing mapping and base quality thresholds), and e_{ij} be a latent indicator variable that takes value 0 when b_{ij} is called correctly and 1 otherwise.

Assuming that sequencing errors are equally likely to result in any of the three alternate bases, the conditional probabilities of observing a specific overlapping base given the true genotype and error status $P(b_{ij}|g_i, e_{ij})$ can be calculated easily (Table 2-1). The conditional likelihood of a single overlapping base can then be written as the two-sample mixture model

$$P\left(b_{ij}\middle|g_i^1, g_i^2, e_{ij}; \alpha\right) = (1-\alpha)P\left(b_{ij}\middle|g_i^1, e_{ij}\right) + \alpha P\left(b_{ij}\middle|g_i^2, e_{ij}\right)$$

where g_i^1 and g_i^2 are the genotypes of the targeted and contaminating DNA samples at site i and α is the sample contamination level. Note that, in this section, we assume array based genotypes are error-free and therefore g_i^1 is known. In later sections, our methods that use either sequence or array-based data alone remove this restriction.

True Genotype g_i	Base Calling Error Event e_{ij}	Pr(b_{ij} = A)	Pr(b_{ij} = B)	Pr(b_{ij} = E)
g_i = AA	e_{ij} = 0	1	0	0
	e_{ij} = 1	0	1/3	2/3
g_i = AB	e_{ij} = 0	1/2	½	0
	e_{ij} = 1	1/6	1/6	2/3
g_i = BB	e_{ij} = 0	0	1	0
	e_{ij} = 1	1/3	0	2/3

Table 2-1 Conditional probability $P\left(b_{ij} \mid e_{ij},\ g_i\right)$ of read b_{ij} given true genotype g_i and read error e_{ij}. (AA: A allele homozygote, AB: heterozygote, BB: B allele homozygote, E: alleles other than A or B)

In the absence of knowledge of the identity of the contaminating individual, we formulate the likelihood

$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{\varepsilon_i} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{R_i} \sum_{e_{ij}} P\left(b_{ij}\middle|g_i^1, g_i^2, e_{ij}, \varepsilon_i; \alpha\right) P\left(e_{ij}\right) \right\} P(g_i^2)\, P(g_i^1|\varepsilon_i; G_i)\, P(\varepsilon_i)$$

(Eqn 1)

Here, M is the number of genotyped sites for the targeted individual, G_i is the array-based genotype for the targeted individual at site i, and ε_i is a binary indicator of genotyping error events. In Equation 1, we calculate genotype probabilities $P(g_i^2)$ from population allele frequency estimates assuming Hardy-Weinberg equilibrium, and error probabilities $P(e_{ij} = 1) = 10^{-Q_{ij}/10}$ and $P(e_{ij} = 0) = 1 - 10^{-Q_{ij}/10}$, where Q_{ij} is the phred-scale base quality score. For simplicity, we assume $P(g_i^1 = G_i | \varepsilon = 0; G_i) = 1$ and $P(g_i^1 = (G \neq G_i) | \varepsilon = 1; G_i) = 0.5$. We estimate the contamination fraction α by maximizing the likelihood in Equation 1, first using a grid search on the interval [0, 1], and then applying Brent's algorithm[3].

To identify the contaminating individual among the N study individuals with array-based genotype data, we consider the likelihood function

$$\mathcal{L}(\alpha, k) = \prod_{i=1}^{M} \sum_{\varepsilon_i^1} \sum_{\varepsilon_i^k} \sum_{g_i^1} \sum_{g_i^k} \left\{ \prod_{j=1}^{R_i} \sum_{e_{ij}} P(b_{ij} | g_i^1, g_i^k, e_{ij}, \varepsilon_i^1, \varepsilon_i^k; \alpha) P(e_{ij}) \right\} P(g_i^1 | \varepsilon_i^1; G_i) P(g_i^k | \varepsilon_i^k; G_i) \, P(\varepsilon_i^1) P(\varepsilon_i^k)$$

for individuals $2 \leq k \leq N$. Using maximum likelihood across α and k, we estimate the most likely contaminating individual k and contamination level α. By comparing the maximum likelihoods (over α) for the most likely and next most likely contaminating samples, including the generic individual represented by population allele frequencies (as in Equation 1), we obtain a measure of support for the inferred contaminating individual.

Detecting sample contamination using sequence data alone

Next, we consider the problem of identifying contamination when prior genotype data are not available. In the absence of prior genotype data, both g_i^1 and g_i^2 are unknown, and the likelihood for the contamination level α becomes

$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{R_i} \sum_{e_{ij}} \left((1-\alpha) P(b_{ij} | g_i^1, e_{ij}) + \alpha P(b_{ij} | g_i^2, e_{ij}) \right) P(e_{ij}) \right\} P(g_i^2) P(g_i^1) \qquad \text{(Eqn 2)}$$

Equation 2 can be maximized using an initial grid search followed by Brent's algorithm. In contrast to Equation 1 in which array-based genotype data are available, Equation 2 is symmetric with respect to the targeted and contaminating individuals. In this situation, with sequence data alone and without previously known genotypes, our method cannot detect sample swaps. Further, since $L(\alpha) = L(1-\alpha)$, here we restrict attention to $0 \leq \alpha \leq ½$.

Detecting sample contamination using array-based genotype data alone

We next turn to the problem of detecting DNA sample contamination using array-based genotype data alone, an analysis which can be carried out to identify contaminated samples prior to sequencing. We assume the availability of relative intensity information, as produced for example by the Illumina Infinium assay. The Infinium assay measures the relative intensities of fluorescently labeled probes associated with arbitrarily labeled alleles A and B. After normalizing intensities, the Illumina software reports (1) the genotype as AA, AB, BB, assigning a missing genotype to individuals with intensities outside the expected clusters; and (2) the estimated abundance of the B allele, called the B allele frequency (BAF). We expect BAF close to 0, ½, or 1 for genotypes AA, AB, and BB, respectively. We describe two types of contamination detection and estimation methods in this setting: two likelihood-based mixture-

model methods based on the intensity values, and a regression-based method using BAF as input.

Detecting sample contamination using array data alone: mixture models for intensity data

We implement our mixture model on the genotype intensity data in two ways. One implementation estimates model parameters by examining signal intensity distributions for each marker across all samples; a second implementation estimates signal intensity distributions by examining all markers for a single sample. Both implementations use genotype intensity values normalized by the GenomeStudio software as input, to reduce technical differences across samples and markers.

In the multi-sample implementation, for each marker i, we model the normalized A and B allele intensity data $\boldsymbol{x_i} = (x_{,A}\, x_B)$ for an uncontaminated DNA sample as a bivariate Gaussian distribution:

$$p_i(\mathbf{x_i}|g_i) \sim \mathcal{N}\left(\boldsymbol{\mu_i^{g_i}}, \boldsymbol{\Sigma_i^{g_i}}\right), \quad g_i = \{AA, AB, BB\},\ 1 \le i \le M$$

Here, g_i is again the true genotype at marker i, $\boldsymbol{\mu_i^{g_i}}$ is the intensity mean vector for marker i given g_i, and $\boldsymbol{\Sigma_i^{g_i}}$ is the covariance matrix of the A and B allele intensities. We estimate $\boldsymbol{\mu_i^{g_i}}$ and $\boldsymbol{\Sigma_i^{g_i}}$ using observed signal intensities and called genotypes at marker i across all genotyped individuals. To reduce the impact of genotype misclassification, we exclude samples with call rate < 99% and markers with minor allele frequency < 1%. Assuming the observed DNA sample is a mixture of two unrelated DNA samples, we can model the intensity values as a bivariate Gaussian mixture:

$$p_i(\mathbf{x_i}|g_i^1, g_i^2; \alpha) \sim \mathcal{N}\left(\alpha\boldsymbol{\mu}_{\boldsymbol{i}}^{g_i^1} + \alpha\boldsymbol{\mu}_{\boldsymbol{i}}^{g_i^2}, \alpha^2\boldsymbol{\Sigma}_{\boldsymbol{i}}^{g_i^1} + (\mathbf{1}-\alpha)^2\boldsymbol{\Sigma}_{\boldsymbol{i}}^{g_i^2}\right) \quad 1 \le i \le M$$

where g_i^1 and g_i^2 are the genotypes of the two samples at marker i. Given data on M independent markers, we formulate the likelihood of a sample using the intensity distribution estimated across multiple samples as

$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{g_i^1} \sum_{g_i^2} p_i(\mathbf{x_i}|g_i^1, g_i^2) P(g_i^1) P(g_i^2) \qquad \text{(Equation 3)}$$

Genotype probabilities $P\left(g_i^k\right)$ in Equation 3 can be calculated assuming Hardy-Weinberg equilibrium using allele frequencies estimated from the called genotypes or from external data. As before, we estimate α by maximum likelihood using a grid search on the interval [0, ½] followed by Brent's algorithm. With genotype array data alone, we cannot detect sample swaps.

The single-sample implementation is analogous to the multi-sample implementation. In the multi-sample implementation, the bivariate Gaussian parameters for p_i at each marker are estimated across all N samples, while in the single-sample implementation, parameters for p_k are estimated across all M markers called in the individual. The corresponding likelihood of single-sample implementation follows

$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{g_i^1} \sum_{g_i^2} p_k(\boldsymbol{x_i}|g_i^1, g_i^2) P(g_i^1) P(g_i^2)$$

where $p_k(\boldsymbol{x}|g_i^1, g_i^2)$ is mixture of bivariate Gaussians whose parameters are estimated across all markers for individual k.

The multi-sample implementation is appropriate when many samples have been genotyped and can be used to estimate the distribution of signal intensities for each marker. The single-sample implementation can be used when data are available on only one or a few samples.

Detecting sample contamination using array data alone: regression-based method

Our second genotype-array-based method detects contamination by identifying systematic shifts between the expected and observed BAF in sites called as homozygous. Consider an individual with genotype AA whose DNA sample is contaminated. As the population frequency of the B allele increases, the sample is increasingly likely to be contaminated with the B allele (Figure 2-2). In the case of no contamination, we expect BAF values close to 0, ½, and 1 for genotypes AA, AB, and BB, respectively. In the presence of contamination, we expect for AA and BB homozygotes that

$$E[BAF \mid g = AA; \alpha, p_B] = \alpha p_B$$

$$E[BAF \mid g = BB; \alpha, p_A] = 1 - \alpha p_A$$

where p_A and p_B are the population frequencies of A and B and α is again the contamination level. To estimate contamination, we fit the linear regression model

$$BAF = \gamma + \alpha p + \tau I(g = AA) + \varepsilon \quad \text{(Equation 4)}$$

where γ is the intercept,

$$p = \begin{cases} p_{B,} & \text{if } g = AA \\ -p_A, & \text{if } g = BB \end{cases}$$

τ is the difference in expected BAF between AA and BB genotypes, and ε is a normally distributed error term. This regression framework allows us to estimate the contamination level α and to test for contamination by evaluating the null hypothesis that the slope $\alpha = 0$ against the one-sided alternative $\alpha > 0$.

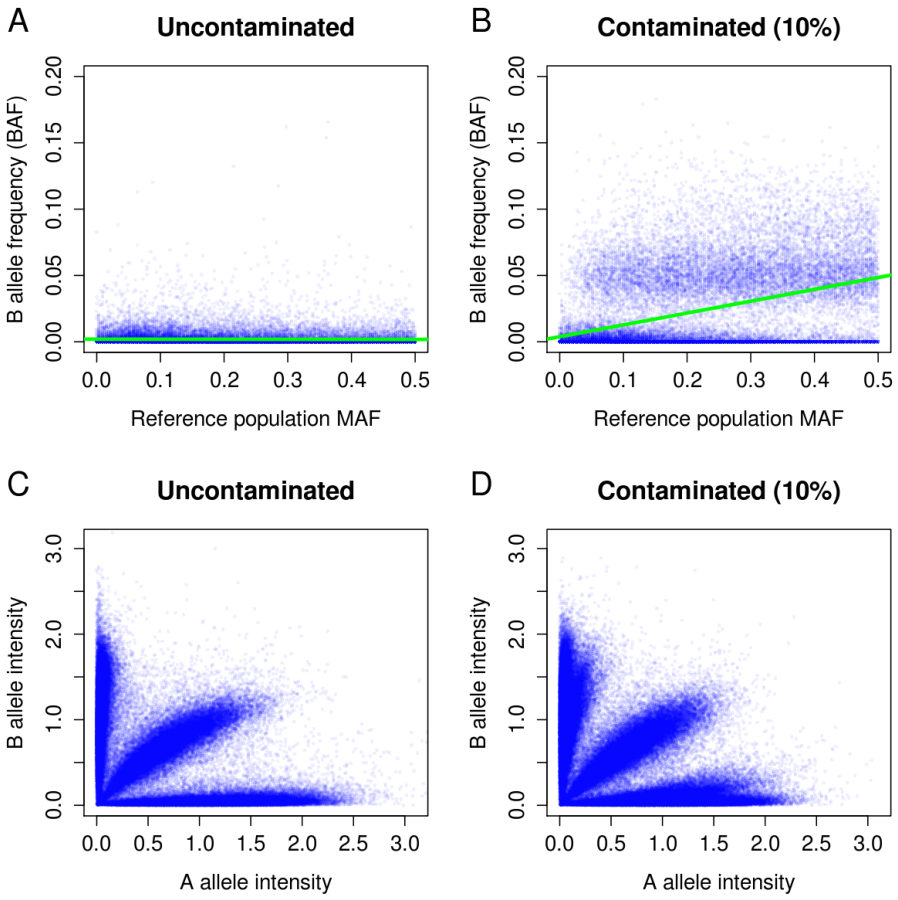

Figure 2-2 - ***B allele frequency (BAF) versus population minor allele frequency (MAF)*** *(A) uncontaminated (α=0) and (B) contaminated (α=10%) samples. Normalized intensity plots for (C) uncontaminated (α=0) and (D) contaminated (α=10%) samples.*

Instead of using the A or B allele frequency as covariate in the regression, we instead use the population minor allele frequency (MAF). This avoids the need to convert Illumina A/B allele calls to actual A/G/C/T alleles. Letting f be the MAF

$$\Pr(B \text{ is minor allele} | g = AA; f)$$

$$= \frac{\Pr(B \text{ is minor allele}, g = AA; f)}{\Pr(B \text{ is minor allele}, g = AA; f) + \Pr(A \text{ is minor allele}, g = AA; f)}$$

$$= \frac{(1-f)^2}{(1-f)^2 + f^2}$$

so that

$$E[BAF \mid g = AA; \alpha, f] = \alpha \frac{f(1-f)}{(1-f)^2 + f^2}$$

Although the relationship between MAF f and contamination level α is not linear, we found that using a regression model of the form

$$BAF = \gamma + \alpha f + \tau I(g = AA) + \varepsilon$$

produces nearly identical results to using the model in Equation 4 which requires knowledge of population allele labels and replaces f with p (data not shown). Thus, it is possible to detect contamination using only AB genotypes and without decoding the correspondence between labels A and B and the underlying A, C, G and T alleles. This ability to avoid decoding the A and B allele labels is important for early steps of data analysis and quality control which, in this way, can proceed without worrying about vagaries of specific genome builds and other informatics challenges that must be tackled before later rounds of analyses.

Assumptions

For ease of computation and notation, our models make several assumptions. The likelihood methods compute likelihoods over multiple markers and/or aligned base positions, as simple products of single marker and/or single base call likelihoods. As written, the resulting

likelihoods are strictly correct when sequencing errors are independent at each aligned base and markers are in linkage equilibrium; when these assumptions are violated, the likelihoods are approximate[4]. In practice, violation of these assumptions can be reduced by: (a) trimming overlapping ends of reads generated from the same template before analysis; (b) ensuring that variant sites considered in analysis are adequately spaced (so that it is unlikely that multiple base calls originating from a single DNA template are used in analysis); and (c) further trimming marker lists so they include only markers that are in linkage equilibrium. In the next section, we discuss empirical assessments of our method using real data demonstrating that our methods are highly accurate in real data settings.

Experimental data

We assessed our contamination estimation and testing methods using *in-silico* contaminated samples and intentionally contaminated real samples.

To evaluate our sequence-based methods, we constructed *in-silico* contaminated sequence data by randomly mixing aligned sequence reads from 21 CEU individuals sequenced at ~4x coverage on an Illumina platform as part of the 1000 Genomes Project. We retained reads from the targeted sample with probability $1 - \alpha$ and from the contaminating sample with probability α ranging from 0.1% to 50%. To avoid artifacts from intrinsic contamination of the original sequence data, we chose as targeted samples those with estimated contamination $\hat{\alpha} <$ 0.1%. Because samples had slightly different mean genome coverage and coverage varied across each genome, the nine levels of intended contamination α actually varied slightly across the samples. For all mixture-model-based methods, we estimated α using both joint and sequence-only methods. In both cases, we calculated likelihoods based on sites with MAF > 5%

(across 87 CEU samples) assayed on the Illumina HumanOmni2.5 array, using sequence reads above phred-scale mapping and base quality thresholds of 13. We based analyses on the entire genome (~1.2M SNPs), chromosome 20 alone (~30K SNPs), or thinned sets of 1,000 to 100,000 evenly spaced SNPs. We also estimated α using our sequence-only methods based on allele frequency estimates from 89 British (GBR), 93 Finnish (FIN), 381 European (CEU, GBR, FIN, TSI, IBS), or 246 African (YRI, LWK, ASW) samples to evaluate the impact of errors in estimated SNP allele frequencies.

To evaluate our genotype-array-only methods, we experimentally constructed contaminated DNA samples by combining pairs of HapMap CEU individuals and pairs of HapMap YRI individuals. We targeted six contamination levels, ranging from α = 0 to 10%. For each contamination level, we targeted three pairs of CEU individuals and three pairs of YRI individuals. We genotyped the 36 resulting samples with the Metabochip, an Illumina genotype array that assays ~200,000 SNPs of interest for studies of cardio-metabolic traits[5]. We used normalized array intensity values, BAF, and genotypes produced by the Illumina's GenomeStudio software run with default options.

Finally, to evaluate empirically our sequence-based methods, we examined potential contamination in 299 actual DNA samples sequenced genome-wide by a large sequencing center at ~4x average coverage in a study of type 2 diabetes. 150 samples were sequenced before a change in the sample handling process in August 2010; the remaining 149 samples were sequenced after the change. 227 of the 299 samples also were genotyped with the Illumina HumanOmni2.5 array. After quality control of the array data, call rates for each sample and each SNP were > 98%. We applied our sequence-based mixture methods to these

data across all SNPs with estimated MAF > 5%. For these samples, we called genotypes from the sequence data using glfMultiples[6] followed by refinement using BEAGLE[7]. From these sequence-based genotype data, we calculated the ratio of heterozygous genotypes to homozygous non-reference genotypes (HET/HOM ratio) and genotype discordances with the HumanOmni2.5 data. All procedures above were approved by the institutional review boards of the University of Michigan.

Results

Detecting sample contamination using sequence data

We estimated α for the 189 samples constructed with *in-silico* contamination ($0.1\% \leq \alpha \leq 50\%$) based on random pairings of 1000 Genomes Project CEU samples (see Materials and Methods). The estimated contamination level $\hat{\alpha}$ conformed well to the intended contamination level α, with Pearson correlation coefficient r = .9996 for the joint method and r = .9840 for the sequence-only method (Figure 3). Both methods tended to overestimate contamination, especially when $\alpha < 1\%$. Generally, absolute error $|\hat{\alpha} - \alpha|$ increased with α and relative error $|\hat{\alpha} - \alpha|/\alpha$ decreased with α. For example, the absolute error was 0.038% ± 0.024% for the joint method and 0.037% ± 0.021% for the sequence-only method when $\alpha \approx 0.1\%$, but increased 0.41% ± 0.30% and 0.56% ± 0.55% when $\alpha \approx 10\%$ (Figure 2-3). In contrast, the relative error of the estimated contamination was .380 ± .257 (mean ± SD) for the joint method and .390 ± .241 for the sequence-only method when $\alpha \approx 0.1\%$, but it was reduced to .044 ± .035 and .056 ± .055 when $\alpha \approx 10\%$. Finally, for the sequence-only method, because $\hat{\alpha}$ is bounded at 50%, we observed a downward bias for α near 50%.

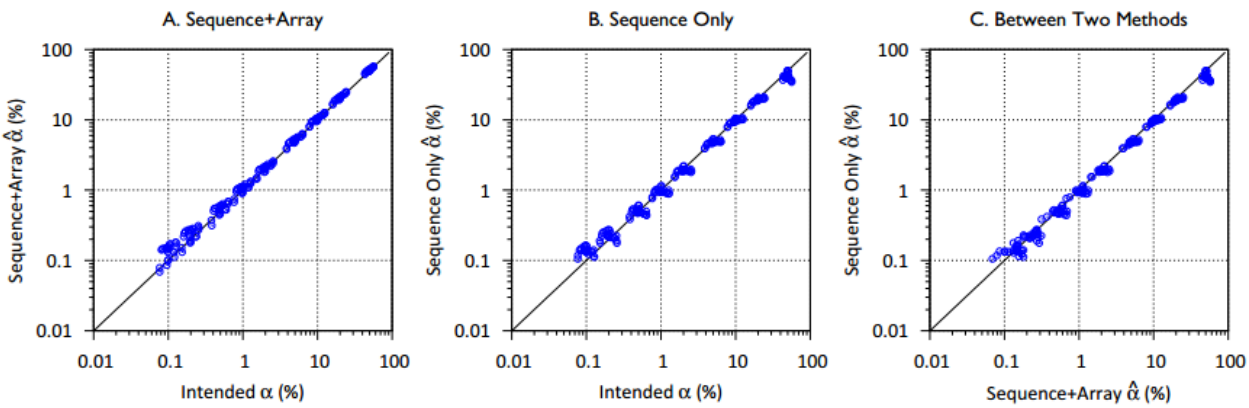

Figure 2-3 Estimated contamination levels for in-silico contaminated samples using (A) joint sequence and array-based method, (B) sequence-only method, and (C) between these two methods.

We evaluated the impact of estimated population allele frequencies on accuracy of contamination estimates (Appendix Figure 2-S1). Compared to the original sequence-only estimates of $\hat{a}$ that used CEU allele frequencies, using allele frequencies from the GBR samples resulted in reduced estimates of $\hat{a}$ (mean ratio ± SD for $\hat{a}_{GBR}/\hat{a}_{CEU}$ = .884 ± .083). Allele frequencies from the more distantly related FIN samples resulted in further reduced contamination estimates (mean ratio ± SD for for $\hat{a}_{FIN}/\hat{a}_{CEU}$ = .804 ± .135). Allele frequencies from the broader European (EUR) continental population (CEU, GBR, FIN, IBS, and TSI) performed better (mean ratio ± SD for $\hat{a}_{EUR}/\hat{a}_{CEU}$ = .926 ± .054), while allele frequencies from the very different African (AFR) samples (YRI, LWK, and ASW) resulted in severe reduction in contamination estimates (mean ratio ± SD for $\hat{a}_{AFR}/\hat{a}_{CEU}$= .160 ± .121).

Next, we evaluated the impact of the number of sites analyzed on contamination estimates using thinned sets of 1,000, 10,000, or 100,000 evenly spaced markers, and using only chromosome 20 sites. These smaller numbers of sites resulted in less accurate estimates

of contamination, particularly at lower levels of contamination (Appendix Figure 2-S2). For example, when α = 1%, the mean relative errors $|\hat{\alpha} - \alpha|/\alpha$ for the joint method were .414, .135, .103, and .099 for 1,000, 10,000, 100,000, and all 1.2M sites, and .112 when using the 30,471 chromosome 20 sites. Since computation times scale linearly with the number of sites analyzed, an (initial) analysis based on 10,000 sites or on all chromosome 20 sites requires 120- to 40-times less computing effort than an analysis of 1.2M sites.

We also compared our joint method to ContEst[2] (April 2012 version), which uses genotype and sequence data together to estimate contamination levels in a likelihood framework. We obtained very similar results for their method and ours when α > 1%; when α < 1%, ContEst tended to overestimate contamination levels to a larger degree than ours (Appendix Figure 2-S3).

Estimation and testing of sample contamination from genotype array data only

Next, we applied our genotype array-only methods to our deliberately constructed contaminated samples genotyped with the Metabochip. Applying the single-sample and multi-sample mixture model methods produced contamination level estimates that matched our constructs, except for two YRI samples with 3% intended contamination (Figure 2-4). Estimates from the regression-based method also showed very strong concordance except for these same two samples. We observe that the two mixture-model methods tend to over-estimate α, while the regression-based method tends to underestimate α.

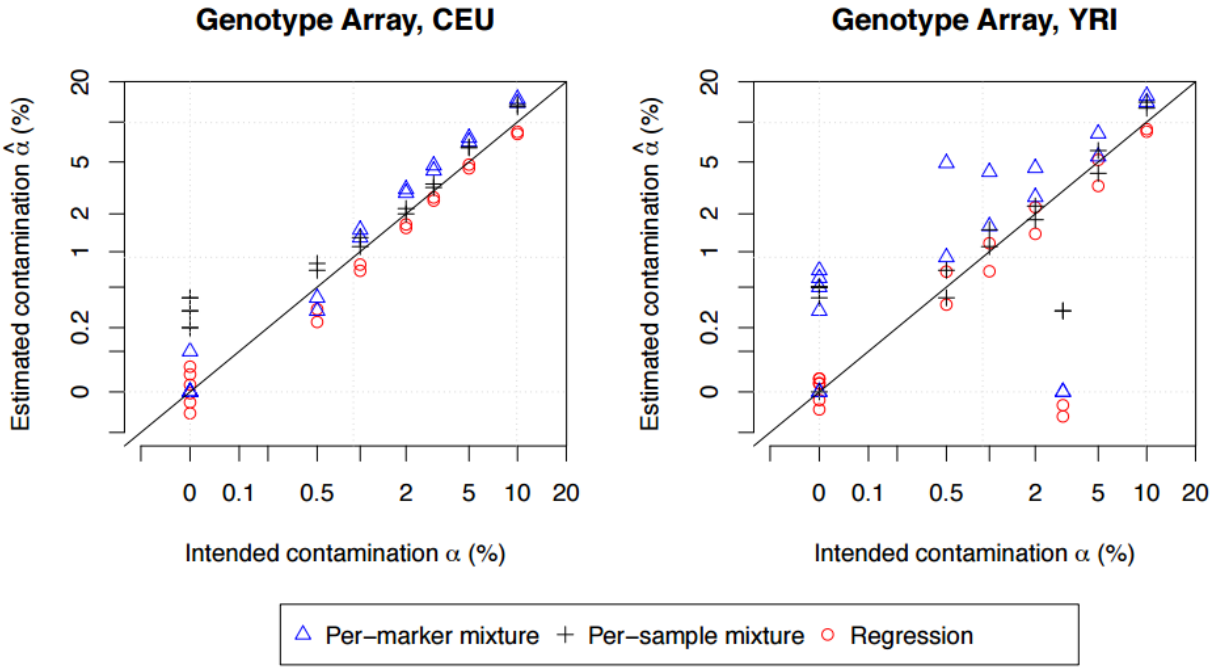

Figure 2-4 ***Estimated versus intended contamination levels from the experimentally contaminated array intensity data****, using (A) regression-based method, (B) multi-sample mixture model method, and (C) single-sample mixture model method.*

Using the mixture-model methods, 0 of the 6 uncontaminated CEU samples were identified as contaminated, while 3 of 6 uncontaminated YRI samples were identified as slightly ($0 < \hat{\alpha} < 1\%$) contaminated. We suspect this misclassification is due at least in part to not having had Metabochip cluster data for African samples and therefore having used our available Finnish samples for defining the clusters used in genotype calling. The mixture-model methods correctly identified 22 of 24 intentionally contaminated samples, the exceptions being the two YRI samples with 3% intended contamination.

Using the regression-based method, we tested the hypothesis of no contamination across 24 contaminated and 12 uncontaminated samples at significance level .05/36=.0013; the

results correctly identified the contamination state of 34 of the 36 experimental samples except for the two YRI samples with intended α = 3%. Given our consistent results across our three different methods, we suspect this pair of YRI samples was not successfully contaminated during the experimental process.

We evaluated a modified version of our regression-based method by including data on heterozygous sites in addition to homozygous sites or by binning SNPs by MAF; these modified approaches performed less well on both simulated and experimental data. The additional noise in the BAF at heterozygous sites made the estimation of contamination less accurate. Attempts to smooth out the uneven MAF distribution of SNPs on a genotype array by binning and averaging over BAF simply reduced power and failed to improve estimation. We also evaluated the regression method restricting analysis to various MAF bins and observed that the method performed best when SNPs across the entire MAF spectrum were included (data not shown).

Type 2 diabetes study

As described in the Introduction, in a recent sequencing study, early in the study we identified a subset of individuals with unusually large numbers of heterozygous genotypes and high HET/HOM ratios compared to other sequenced individuals Figure 2-1AB). We applied our sequence-based and sequence-only methods to these samples. Since HumanOmni2.5 genotype data were available on only 227 of these 299 individuals, we display results for the sequence-only method (Figure 2-1C); contamination level estimates for the sequence and array data jointly were very similar, particularly for individuals with higher contamination levels (Figure 2-5). Consistent with our impression based on genotype calls and HET/HOM ratio, our methods

identified a cluster of contaminated samples among the 150 samples sequenced before August 2010, with 45, 24, and 16 of these 150 samples estimated to have contamination levels of $\hat{\alpha} \geq 1\%$, $\geq 2\%$, and $\geq 5\%$, respectively (

Table 2-2).

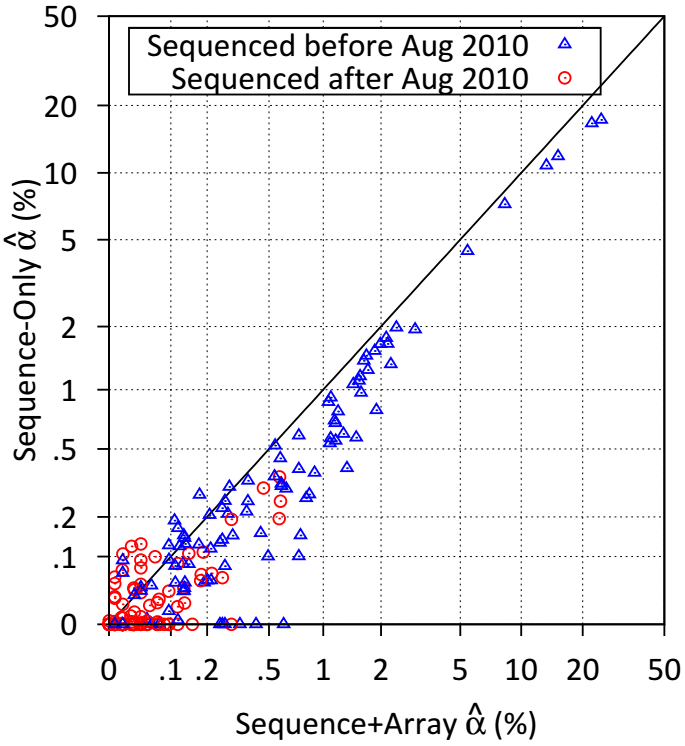

Figure 2-5 Comparison of estimated contamination levels using sequence data with and without array genotype data for type 2 diabetes sequencing study.

Comparison of results (Figure 2-1,

Table 2-2, Appendix Figure 2-S3) suggests that our contamination estimates were more sensitive than heterozygosity and HET/HOM ratio for detecting contaminated samples, particularly at lower levels of contamination. For example, the average HET/HOM ratios among the ten samples with $2\% \leq \hat{\alpha} < 5\%$ and the 254 samples with $\hat{\alpha} \leq 1\%$ were nearly

identical: 1.92 and 1.91. Investigation by the sequencing center suggested that contaminating samples were often in adjacent lanes to the targeted samples during library construction. Following modification of the library construction process in August 2010, none of the 149 samples sequenced later that year had estimated contamination level $\hat{\alpha} \geq 0.5\%$, (Figure 2-1C).

Array Genotypes?	Measure	$\hat{\alpha}$ (sequence only)			
		<1%	1-2%	2-5%	≥5%
Yes	Number of samples	208	13	1	5
(n=227)	− Before August 2010	81	13	1	5
	− After August 2010	127	0	0	0
	RR discordance[1]	.0021	.0030	.0071	.0492
	RA discordance[2]	.0154	.0157	.0172	.0300
	AA discordance[3]	.0085	.0143	.0377	.176
	HET/HOM ratio[4]	1.92	1.84	2.16	2.66
No	Number of samples	46	8	7	11
(n=72)	− Before August 2010	24	8	7	11
	− After August 2010	22	0	0	0
	HET/HOM ratio[4]	1.87	1.88	1.88	2.64

1. RR discordance: Genotype discordance when array-based genotype is homozygous reference
2. RA discordance: Genotype discordance when array-based genotype is heterozygous
3. AA discordance: Genotype discordance when array-based genotype in homozygous non-reference
4. HET/HOM ratio: Ratio of number of heterozygous genotypes to homozygous non-reference genotypes

Table 2-2 Summary of estimated contamination levels $\hat{\alpha}$, ratio of the numbers of heterozygous to non-reference allele homozygous genotypes, and genotype discordance with array data for 299 samples (227 with HumanOmni2.5 genotype array data) from type 2 diabetes study using sequence data only.

To assess the impact of DNA sample contamination on genotyping accuracy, we compared genotypes called from the diabetes sequence data to the HumanOmni2.5 genotypes.

As expected, discordance between the sequence-based genotypes and the highly accurate array genotypes increased with increasing estimated contamination. For homozygotes, average genotype discordance rates doubled in samples with $1\% \leq \hat{\alpha} \leq 5\%$ compared to those with $\hat{\alpha} \leq 1\%$, and increased by a factor of ~20 for $\hat{\alpha} \geq 5\%$ (

Table 2-2, Figure 2-6). The impact of contamination was less strong for heterozygous sites, but genotype discordance rates were still nearly doubled when $\hat{\alpha} \geq 5\%$ compared to those in samples with $\hat{\alpha} \leq 1\%$. The stronger effect of contamination on homozygous genotypes occurs because even modest numbers of contaminating sequence reads may result in calling a homozygote as a heterozygote.

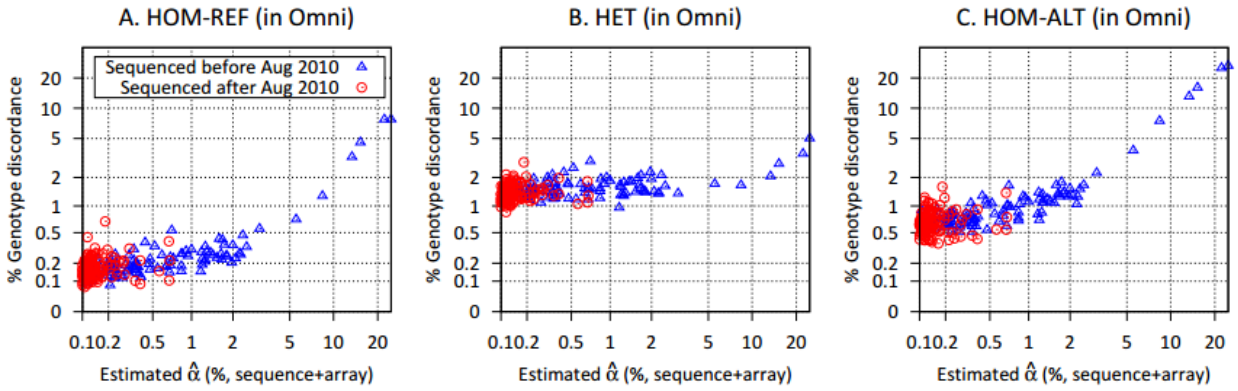

Figure 2-6 ***Genotype discordance between sequence-based and array-based genotypes*** *as a function of estimated contamination level $\hat{\alpha}$ in the type 2 diabetes sequencing study; contamination level estimates based on the combined sequence and genotype array data.*

Discussion

In this paper, we describe several methods to identify within-species DNA sample contamination based on the analysis of sequence read data and/or array-based genotype data. We first describe a mixture-model method that uses both sequence reads and array-based

genotypes, and then show that this method can be extended naturally to identify contaminated samples when only sequence reads are available. Both these sequence-based methods are highly sensitive, allowing detection of DNA sample contamination of 1% or less even with low-coverage (4x) sequence data. As expected, the combination of sequence reads and array-based genotypes results in greater sensitivity than sequence data alone, but the difference is modest (Figure 2-3). Both our sequence-based mixture-model methods are more sensitive than traditional checks that test for an excess of heterozygous genotypes or an unusually high ratio of heterozygous to non-reference homozygous genotypes (HET/HOM ratio) – both of which can only detect contamination rates of >5-10% (Appendix Figure 2-S3). A further advantage of our sequence-based methods is that they operate directly on the sequence reads (or BAM files), and so can be applied prior to variant calling. In sequencing studies, the availability of array-based genotypes for all samples allows identification of contaminating DNA samples and resolution of sample swaps.

As with other analyses of short read sequence data, the sequence-based mixture-model methods are computationally intensive. Given low-coverage (4x) whole-genome sequence data and focusing on sites with MAF > 5% from the Illumina 2.5M genotype array, our sequence-based analyses required ~1.6 hours compute time per DNA sample on a single 2.8GHz processor. Increasing sequence coverage results in an approximate linear increase in compute time. To reduce computational burden, or if sequence read data come in large batches, we often do initial DNA contamination checking using a subset of the genome. For example, analysis limited to chromosome 20 requires only ~2% the compute time, thus permitting rapid real-time early quality control and timely feedback to the sequence production group; for

contamination levels >1% and when the targeting and contaminating samples are unrelated, chromosome 20 analysis is also nearly as sensitive as analysis of the entire genome (Appendix Figure 2-S2).

While our analysis of sequence-based methods focused on low-coverage whole-genome sequences, we have found that our sequence-based methods robustly identify contamination in other types of sequencing data. For example, our methods have been successfully applied to targeted whole exome sequence data in the 1000 Genomes Project in addition to the low-coverage sequence data. We also found that our sequence-based methods robustly detect contamination in RNA-seq data with or without external genotypes. In these data sets, focusing on exonic or on-target sites provided more accurate estimates of contamination levels than using all sites (data not shown).

The models on which we base these methods (of course) do not capture all features of the sequencing experiment. One such feature is reference bias, in which more reference-sequence bases are observed than expected at a variant site, potentially resulting in an upward bias in estimated contamination levels. Poorly aligned bases, inaccurate base quality scores, and asymmetric calling errors between bases may have the same effect. Currently, both our sequence-based methods assume that the population from which the contaminating sample is drawn is known, and we observed reduced sensitivity with incorrect population allele frequencies. When the population of the contaminating DNA sample is unknown, our method could be extended to iterate over alternative population allele frequencies to identify the most likely source population for a contaminant and to more precisely estimate the level of

contamination. Our implementation uses a simple error model. Preliminary evaluations of more sophisticated genotype error models made little difference to our results.

In several sequencing studies, including the type 2 diabetes study described above, we have observed that our methods estimate a large fraction of samples to be contaminated at very low but non-zero levels, and likelihood ratio tests of $\alpha = 0$ against the alternative $\alpha > 0$ result in apparent "contamination detection" for most samples. In contrast, when we simulated uncontaminated DNA samples consistent with all our model assumptions, we found $\hat{\alpha} > 0$ for only 33% of samples as opposed to 50% expected by a 1:1 mixture between χ_0^2 and χ_1^2 distributions[8]. Furthermore, although both our likelihood-based methods naturally lead to confidence intervals for the level of estimated contamination, we generally find these intervals to be too narrow and do not recommend their use. These contrasting findings likely reflect the impact of not modeling some of the sequencing experiment features described above. Careful examination of the impact of uncertainty in population allele frequency, of variation in read depth by genotype, of the fraction of duplicate reads, and of runs of homozygosity, could help to identify important features that are missing from the model. We are working to include some of these features in our models, methods, and software.

Identifying contaminated samples using array data alone provides the opportunity to avoid sequencing contaminated samples. Both of our genotype-array-only methods – whether mixture model or regression based -- result in enhanced sensitivity compared to previous strategies that identify likely contaminated samples as those with low genotype call rates. Low genotype call rates can identify heavily contaminated DNA samples as well as those that fail for other technical reasons. However, in our experimentally contaminated samples genotyped

with the Metabochip, even at 5% contamination, all four samples had genotype call rates > 99.5%, and even at 10% contamination, call rates were still between 96.8% and 97.9%. Our mixture- and regression-based methods allowed accurate detection of contamination levels as low as 1%.

In contrast to the sequence-based methods, our genotype-array-only methods have modest computational requirements. For example, analysis of 36 samples genotyped at 200,000 SNPs required <100 seconds on a single 2.8GHz processor for either the mixture-model or regression-based methods. Further, these genotype-array-only methods were remarkably sensitive for contamination detection even with modest numbers of SNPs. For example, using our experimentally contaminated samples and defining contamination detection as $\hat{\alpha} \geq 1\%$, power to detect contamination using the regression method based on 1000 random subsets of 50, 100, 500, and 1000 homozygous SNPs was 37.3%, 59.6%, 99.0%, and 100%, respectively (Appendix Table 2-S1). A confidence interval for the estimated contamination level can also be obtained from a simple linear regression model, ignoring uncertainty in key parameters such as the site-specific allele frequencies. We found that, unlike the likelihood-based methods, the regression-based method provides reliable p-value and confidence interval with even a modest number of SNPs. Of course, neither genotype-array-based method eliminates the possibility of introducing contamination during subsequent library preparation or sample sequencing.

Our genotype-array-based mixture-model methods rely on good estimates of the means and variances of the genotype intensity clusters. Estimation can be carried out across multiple samples (for each marker) or using a single sample (and pooling estimates across markers). The single-sample method has the obvious advantage that it can be applied to one or a few

samples, permitting analysis to be carried out for small studies or on-the-fly as each sample is processed; a further advantage is that the method can analyze rare genotypes for which intensity distributions may be poorly estimated in methods that examine intensity distributions one site at a time, even across many individuals. The single-sample method also has disadvantages. The distribution of intensities across all SNPs for a given sample generally has larger variance than that for a given SNP across many samples[9]; for contamination detection, this larger variance leads to somewhat less sensitive contamination detection when small numbers of markers are available. Regularizing parameters that share information across sites could increase the performance of the intensity-based mixture models for array data. Compared to the mixture-model method, the regression method has the advantage of providing a better calibrated hypothesis test for contamination. In practice, running multiple methods on the array data will increase the confidence in analysis results.

All our contamination detection methods assume the targeted DNA sample is contaminated by DNA from one other unrelated individual. Given a fixed total contamination level α, contamination from two or more individuals increases the likelihood that multiple alleles will be observed at a marker and typically results in inflated estimates of α. For example, when we simulated contaminating reads originating from two, three, and four contaminating samples, we observed 1-9%, 4-11%, 8-14% relative increases in the estimated contamination levels compared to actual contamination (Appendix Table 2-S2). The joint sequence and array-based method, which relies mostly on genotype concordance rather than increased heterozygosity, showed only a small loss of precision with multiple contaminating samples. In contrast, if a DNA sample is contaminated with DNA from a relative of the targeted

individual, the genetic similarity between the targeted and contaminating sample will result in an underestimate of α. Simulation results suggest that given contamination at level α from an individual sharing a fraction f of genes with the targeted sample results in an estimated contamination level of $(1-f)\alpha$, for example, $\alpha/2$ for sibling or parent-offspring pairs (data not shown).

There are additional applications not yet covered by our method. We have implemented and evaluated our genotype-array-only methods for Illumina genotyping platform only. In principle, our methods can also support Affymetrix intensity data, as used in tools such as Birdseed[10] or PennCNV[11] which work with both Affymetrix and Illumina platforms. For the sequence-based mixture models, an interesting application would be detection of heterogeneous cell populations within tumors. Our experience suggests that even small contamination levels can be detected using only a small number of informative sites, so that this might well be practical.

We have described an efficient set of methods to detect DNA sample contamination that should be useful for investigators planning or carrying out large-scale sequencing studies. For studies based on DNA samples with prior GWAS or other large-scale genotype data, we recommend using the genotype array-only methods to detect contaminated samples prior to sequencing. These methods are useful even for small genotyping arrays with only 1000s of SNPs. Based on results for the genotype-array analysis, an investigator may decide to obtain new DNA samples when there is evidence of contamination, or to eliminate those individuals from the study. Whether or not the genotype-array-based contamination pre-screening is carried out, we recommend using the sequence-based methods to screen DNA samples for

contamination. Based on the results of this sequence-based contamination analysis, the investigator might choose to eliminate from downstream analyses substantially contaminated samples, or to resample and resequence those individuals; for example, the 1000 Genomes Project chose to eliminate all DNA samples with estimated contamination $\hat{\alpha} > 2\%$[12].

Application of these DNA contamination detection methods provides a sensitive method to identify contaminated samples and to maximize sequence data quality. In addition, it may prove helpful to develop analysis methods that explicitly incorporate detection and estimation of DNA sample contamination into variant calling and/or downstream analysis.

Acknowledgements

This work was supported by NIH grants DK088398 and HG000376 (to M.B.), by NIH grants MH084698, HG006513 and HG005214 (to G.R.A.), and by NIH contract numbers HHSN268200782096C and HHSN268201100011I to support the Center for Inherited Disease Research.

Web Resources

The URLs for data presented herein are as follows:

Our initial description on sample identity verification (April 29, 2010)
http://genome.sph.umich.edu/wiki/Verifying_Sample_Identities_-_Implementation

Contamination detection software package
http://genome.sph.umich.edu/wiki/ContaminationDetection

Chapter 2 References

1) Schmieder RAE, Robert (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLoS ONE 6:e17288
2) Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27:2601-2602
3) Brent RP (2002) Algorithms for minimization without derivatives. Dover Publications, New York
4) Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V (2004) Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Statistical Applications in Genetics and Molecular Biology 3:Article 26
5) Voight BF, Kang HM, Ding J, Palmer C, Sidore C, Chines P, Burtt N, Fuchsberger C, Li Y, Erdmann J, Frayling TM, Held IM, Jackson AU, Johnson T, Kilpeläinen TO, Lindgren C, Morris AP, Prokopenko I, Randall JC, Saxena R, Soranzo N, Speliotes EK, Teslovich TM, Wheeler E, Maguire J, Parkin M, Potter S, Rayner WN, Robertson N, Stirrups K, Winckler W, Sanna S, Mulas A, Nagaraja R, Cucca F, Barroso I, Deloukas P, Loos RJ, Kathiresan S, Munroe PB, Newton-Cheh C, Pfeufer A, Samani NJ, Schunkert H, Hirschhorn JN, Altshuler DA, McCarthy MI, Abecasis GR, Boehnke M (2012) The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genet 8:e1002793
6) Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21:940-951
7) Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85:847-861
8) Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Statist Assoc 82:605-610
9) Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC (2008) GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics 24:2209-2214
10) Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253-1260
11) Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665-1674
12) 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073

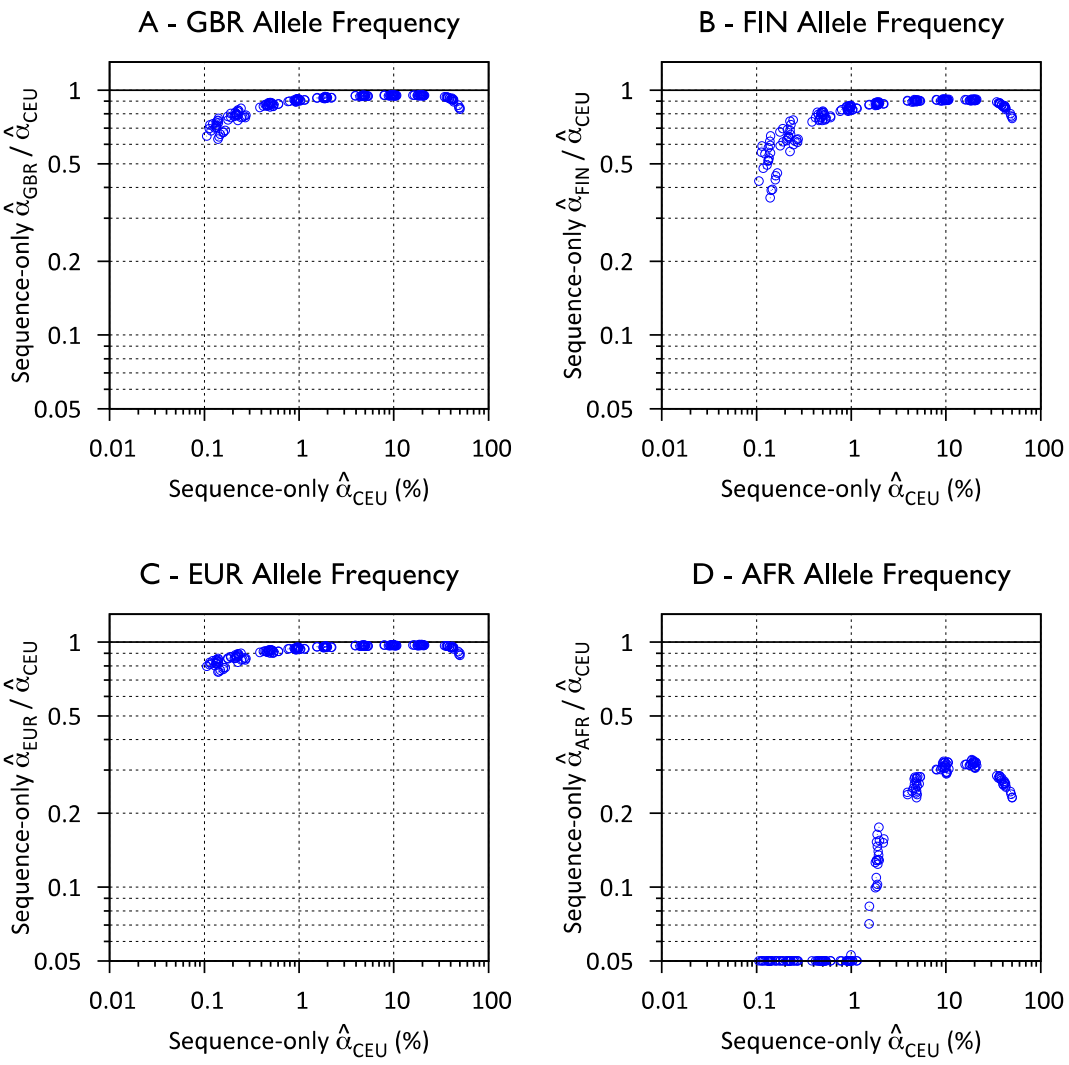

Figure 2-S1. Impact of Population Allele Frequency on Estimated Contamination Levels. Ratio between estimated contamination levels using different population allele frequencies with the sequence-only method.

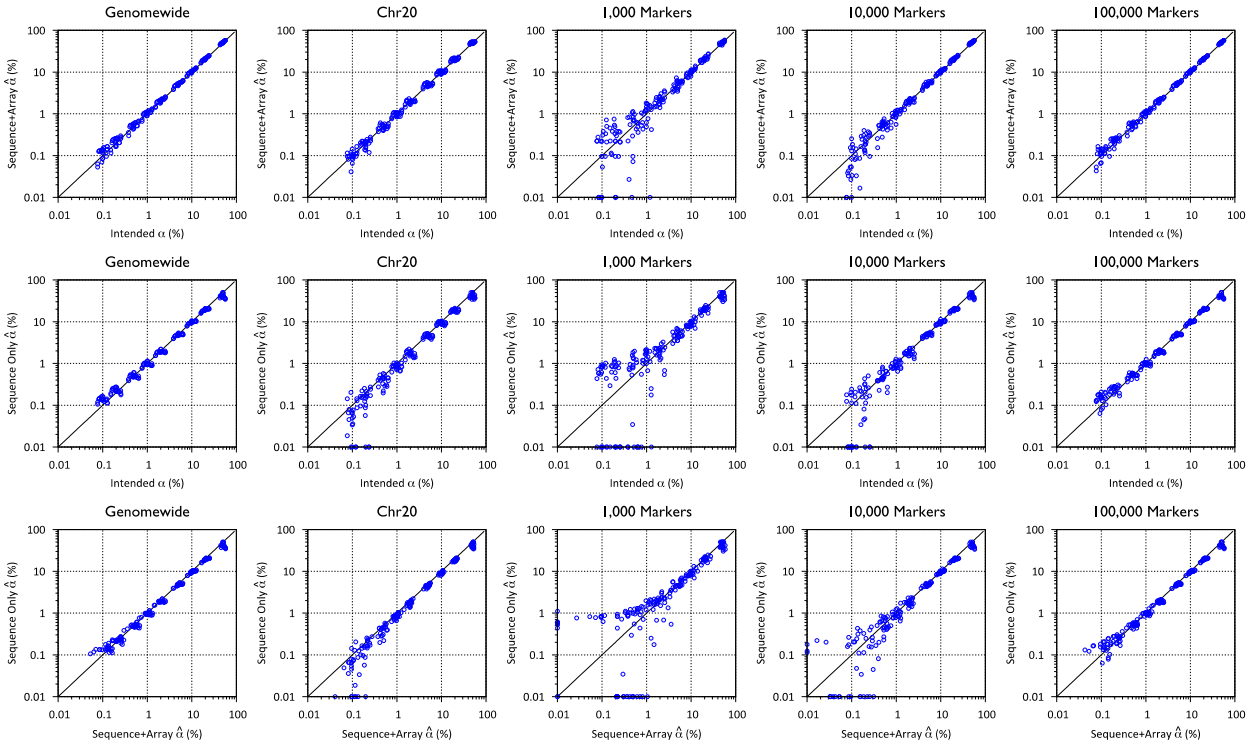

Figure 2-S2. Estimated Contamination Levels Across Different Number of Markers. Comparison between each pair of intended contamination level, estimated contamination levels $\hat{\alpha}$ using joint sequence and array-based method and $\hat{\alpha}$ using sequence-only method across different number of markers.

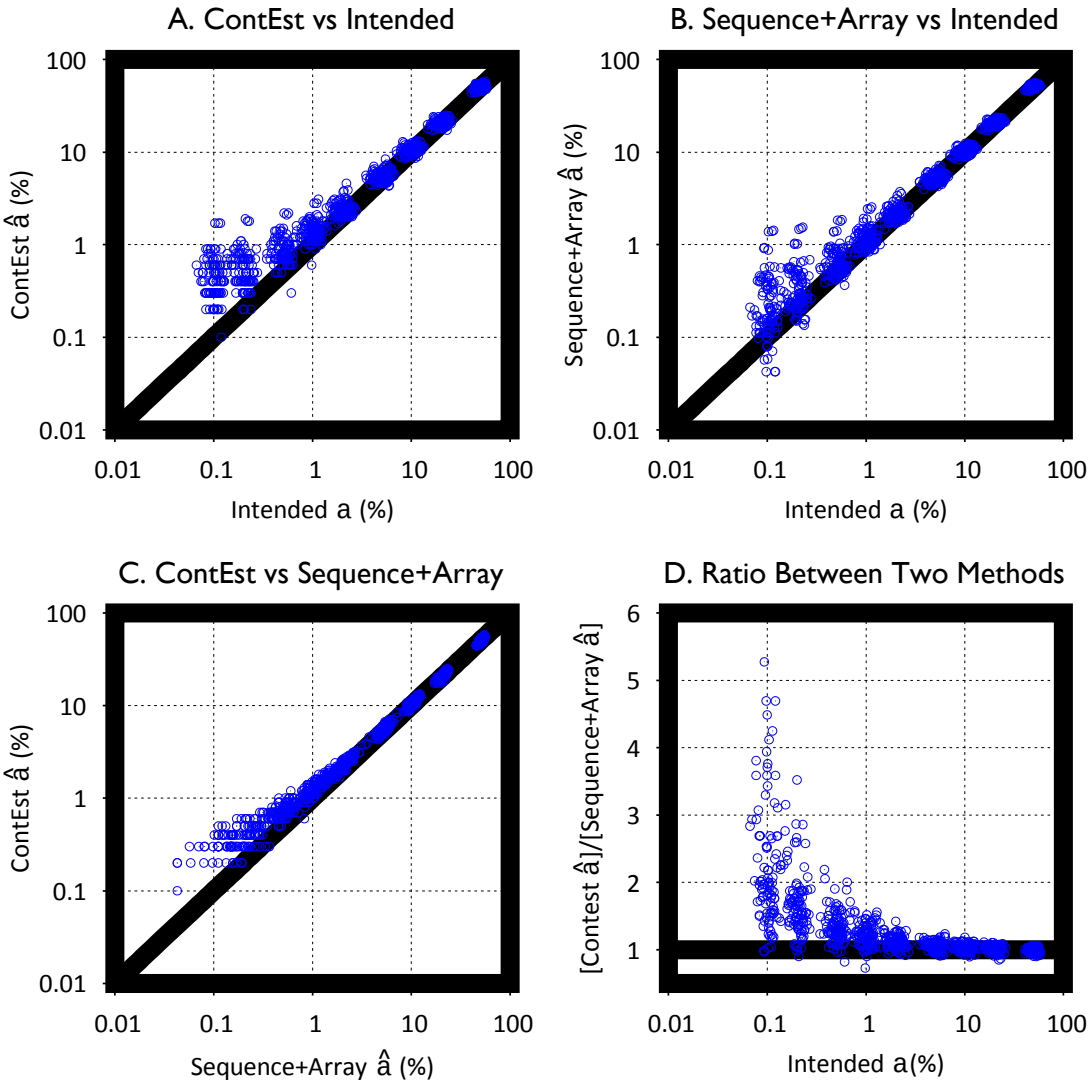

Figure 2-S3. Comparison of Our Methods with ContEst Software

Comparison of estimated contamination levels between joint sequence and array-based method and ContEst on the *in-silico* simulated data for chromosome 20. (A) intended contaminations versus ContEst estimates (B) Our joint sequence and array-based method versus ContEst estimates (C) ratio between the two estimates.

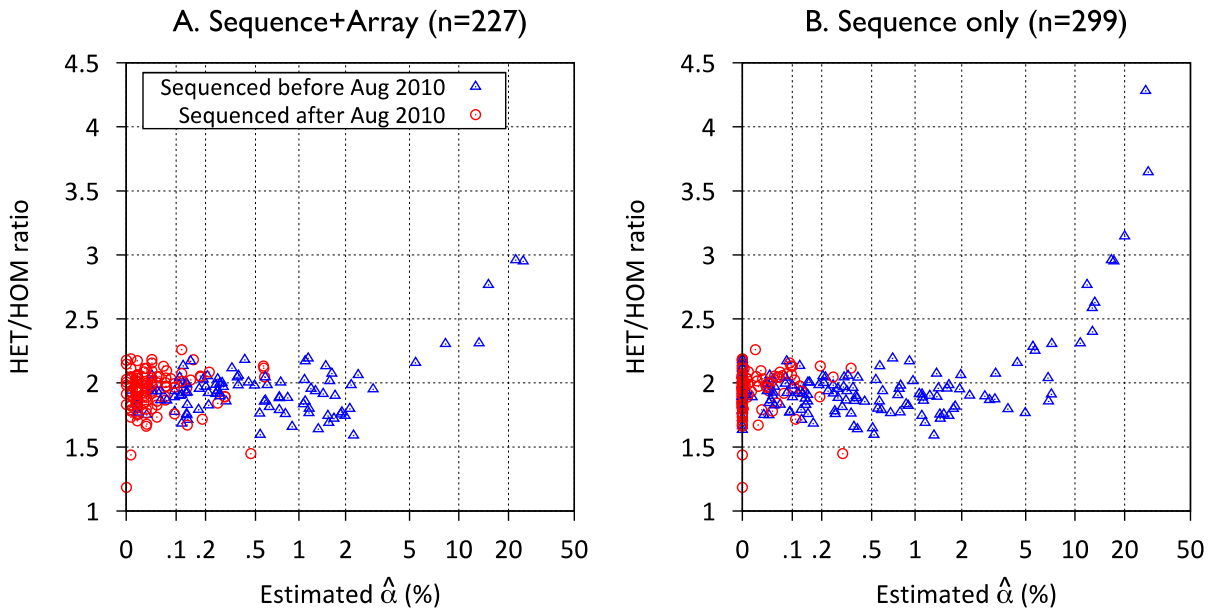

Figure 2-S4. Excess Heterozygosity in relation to Estimated Contamination
Comparison of HET/HOM ratio to estimated contamination level $\hat{\alpha}$ in the type 2 diabetes sequencing study based on analysis of (A) sequence and genotype array data (n=227) and (B) sequence data only (n=299).

Table 2-S1. **Power and Type 1 Error of Genotype-Array Only Regression Method**

# Homozygous SNPs	α=0	α=0.5%	α=1%	α=2%	α=3%	α=5%	α=10%
50	0.053	0.160	0.373	0.739	0.861	0.943	0.970
100	0.060	0.228	0.596	0.946	0.994	1.000	1.000
500	0.071	0.620	0.990	1.000	1.000	1.000	1.000
1000	0.076	0.853	1.000	1.000	1.000	1.000	1.000

For our experimentally contaminated sample, we selected different subset of homozygous SNPs and ran our regression method on those subsets. We then repeated this 1,000 times for each sample. The true level of contamination is shown at the top of the table. This values in the table show the proportion of tests which rejected the hypothesis of α=0 at the 0.05 level.

Table 2-S2. **Impact of multiple contaminating samples on estimated contamination**

Intended Contamination (Fixed Total)	Sequence-only			Sequence+Array		
	$\hat{\alpha}_2/\hat{\alpha}_1$	$\hat{\alpha}_3/\hat{\alpha}_1$	$\hat{\alpha}_4/\hat{\alpha}_1$	$\hat{\alpha}_2/\hat{\alpha}_1$	$\hat{\alpha}_3/\hat{\alpha}_1$	$\hat{\alpha}_4/\hat{\alpha}_1$
α=1%	1.01	1.04	1.14	1.03	1.03	1.07
α=2%	1.02	1.04	1.10	1.03	1.02	1.04
α=5%	1.03	1.05	1.08	1.01	1.01	1.01
α=10%	1.06	1.08	1.11	1.00	0.99	0.99
α=20%	1.09	1.11	1.13	0.97	0.95	0.95

The intended contamination was equally distributed across 2, 3, and 4 CEU samples. $\hat{\alpha}_k$ represents estimated contamination obtained from *k* contaminating samples, and the fold-enrichment of estimated contamination is average across 100 different runs. The results suggest that the sequence-only estimate of contamination tend to increase with multiple contaminating samples. In joint sequence and array-based method, multiple contaminating samples leads to slight overestimation of contamination when the contamination is small (α≤5%), and to underestimation when the contamination large (α≥10%).

Chapter 3
Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data

Abstract[B]

DNA sample contamination is a frequent problem in DNA sequencing studies, and may result in genotyping errors and reduced power for association testing. We recently described methods to identify within-species DNA sample contamination based on sequencing read data, showed that our methods can reliably detect and estimate contamination levels as low as 1%, and suggested strategies to identify and remove contaminated samples from sequencing studies. Here we propose methods to model contamination during genotype calling as an alternative to removal of contaminated samples from further analyses. We compare our contamination-adjusted calls to calls that ignore contamination and to calls based on uncontaminated data. We demonstrate that, for moderate contamination levels (5%-20%), contamination-adjusted calls eliminate 48-77% of the genotyping errors. For lower levels of contamination, our contamination correction methods produce genotypes nearly as accurate as those based on uncontaminated data. Our contamination correction methods are useful generally, but are particularly helpful for sample contamination levels from 2 to 20%.

[B] This work has been published: Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., Kang, H. M. (2015). Correcting for sample contamination in genotype calling of DNA sequence data. Am J Hum Genet *97*, 284-290.

Introduction

Advances in next-generation sequencing have resulted in higher sequencing throughput and lower sequencing costs, enabling a wide range of large-scale genomic studies. While the quality of sequence data is generally improving, methods and protocols are imperfect and errors inevitably occur. One such error is DNA sample contamination, in which DNA from two or more individuals is accidentally mixed.

DNA sample contamination is a common occurrence in large-scale sequencing studies and can arise at many steps of the experiment: during sample collection; any time a sample is placed into or taken out of storage; during shipping, particularly if plates are not properly sealed or kept frozen; and during the many steps of preparing DNA sequencing libraries. For example, if barcoded samples are amplified in pools, template switching may occur if amplification conditions result in templates that are only partially extended at the end of each round, resulting in DNA from one sample paired with the barcode of another. Even if samples are sequenced without contamination on a particular run, a sample may be included in multiple runs and merged afterwards. If samples are improperly labeled or there are errors in the processing pipeline, reads from multiple samples may be combined in error.

Screening for sample contamination is becoming a standard quality control step for DNA sequencing projects, and the patterns of contamination identified vary greatly. In the 1000 Genomes Project, DNA samples were screened for contamination[1] using our method[2]. Out of 1166 sequenced samples, 39 had an estimated contamination level >3% and were dropped from analysis. In a psychiatric genetics study, we detected 64 DNA samples each with estimated

contamination >25%. These samples were traced back to two 96-well plates in which contamination likely occurred during shipping. In a type 2 diabetes exome sequencing study, ~20% of a set of DNA samples had estimated contamination rates from 10-15%. Here, the apparent cause was a change in the library preparation protocol to allow processing of two samples at a time. Even in the most challenging contamination scenarios we have encountered, a subset of DNA samples show no evidence of contamination, so that most studies include a mixture of contaminated and uncontaminated DNA samples.

If left uncorrected, contamination results in systematic genotype misclassification with a bias in favor of heterozygotes. This bias arises since when a mixture of two DNA samples is sequenced, the presence of the contaminating sample DNA makes it more likely that reads supporting different alleles at the same site will be present. The impact of contamination typically increases with the contamination level and decreases with sequencing depth.

Here we propose likelihood-based methods that improve genotyping accuracy by explicitly modeling DNA sample contamination during genotype calling. We apply these methods to *in-silico* contaminated samples based on low-pass and high-depth sequence data from the 1000 Genomes Project and to actual contaminated samples from a type 2 diabetes exome sequencing project. We demonstrate that over a wide range of contamination levels and sequencing depths, modeling contamination can dramatically increase concordance between genotype calls and the true underlying genotypes, resulting in larger effective sample sizes for downstream genetic association studies than is possible by either ignoring contamination or dropping contaminated samples from the analysis.

Materials and Methods

Outline

First, we introduce notation and assumptions, and review our methods to detect DNA sample contamination[2]. Second, we describe our model for calling genotypes from sequence read data and propose a generalization of that model to account for DNA sample contamination. Third, we extend our model and method to provide even better results when the source of contamination is known and the corresponding sample is also sequenced. Finally, we describe a series of experiments and datasets used to evaluate the performance of our proposed methods.

Detecting and estimating DNA sample contamination

Consider the case where one DNA sample is contaminated by another[2]. Let $g_i^{(1)}$ and $g_i^{(2)}$ be the genotypes for the intended and contaminating samples at variant site i ($1 \le i \le M$). Let b_{ij} be the observed base at position i for read j $(1 \le j \le R_i)$ and e_{ij} be a latent variable indicating whether a base calling error occurred ($e_{ij} = 1$) or did not ($e_{ij} = 0$). Finally, let α be the proportion of reads from the contaminating sample and π be the proportion of samples that are contaminated. We assume that sites are independent, that reads at each site are independent, and that sequencing errors are equally likely to result in any of the three incorrect bases.

To model the probability of observing a particular base, we employ the mixture model

$$P\left(b_{ij}\middle|g_i^{(1)},\ g_i^{(2)};\alpha\right) = (1-\alpha)P\left(b_{ij}\middle|g_i^{(1)}\right) + \alpha P\left(b_{ij}|g_i^{(2)}\right) \qquad (1)$$

where

$$P(b_{ij}|g_i) = P(b_{ij}|g_i,\ e_{ij} = 1)\ P(e_{ij} = 1) + P(b_{ij}|g_i,\ e_{ij} = 0)\ P(\ e_{ij} = 0)$$

We present the read probabilities allowing for error $P(b_{ij}|g_i,\ e_{ij})$ in Table 3-1. We estimate the probability of a read error as $P(e_{ij} = 1) = 10^{-Q_{ij}/10}$ and $P(e_{ij} = 0) = 1 - P(e_{ij} = 1)$, where Q_{ij} is the phred-scaled base quality score for the sequence data[3]. To estimate the genotype probability (g_i), we use allele frequencies from the population from which the sample was drawn and assume Hardy-Weinberg equilibrium. Allele frequencies can be estimated from a closely related reference population (for example, HapMap or 1000 Genomes), from array-based genotypes from the same population, or even from the proportion of reads that carry each allele across all sequenced samples.

Taking expectations over the unknown genotypes and assuming all reads and loci are independent, we write the likelihood for contamination level α in a sample as

$$L(\alpha) = P(B|\alpha) = \prod_{i=1}^{M} \sum_{g_i^{(1)}} \sum_{g_i^{(2)}} \left\{ P\left(g_i^{(1)}\right) P\left(g_i^{(2)}\right) \prod_{j=1}^{R_i} \left[(1-\alpha) P\left(b_{ij} \middle| g_i^{(1)}\right) + \alpha P(b_{ij}|g_i^{(2)}) \right] \right\}$$

For each sample, we first maximize $L(\alpha)$ using a grid search in the interval $0.0 \leq \alpha \leq 0.5$ and then apply Brent's[4] algorithm to obtain the maximum likelihood estimate of α. By using information across a large number of variants M, we determine if the observed reads are better explained by a single sample or a combination of two samples with mixing proportion α. Even if not all markers are independent, there is little impact on the estimation of α.

True genotype	Base Read Error Indicator	$\boldsymbol{P(b_{ij} = \mathbf{A})}$	$\boldsymbol{P(b_{ij} = \mathbf{B})}$	$\boldsymbol{P(b_{ij} = \mathbf{E})}$
g=AA	e=0	1	0	0
	e=1	0	1/3	2/3
g=AB	e=0	1/2	1/2	0
	e=1	1/6	1/6	2/3
g=BB	e=0	0	1	0
	e=1	1/3	0	2/3

Table 3-1 Conditional probability $\boldsymbol{P(b_{ij} \mid e_{ij},\ g_i)}$ of read $\boldsymbol{b_{ij}}$ given true genotype $\boldsymbol{g_i}$ and read error $\boldsymbol{e_{ij}}$ Assumes a biallellic site with alleles A and B; E represents any base other than A or B. $e_{ij} = 0$ corresponds to a sequencing error; or 1 corresponds to a correct base call.

Genotype likelihoods for contaminated sequence data: source unknown

Having estimated the contamination level α for sample *k*, we explicitly model contamination during genotype calling using the estimated sample-specific contamination rate $\hat{\alpha}_k$. Treating the genotypes of the intended and contaminating genotypes as the unknowns, we calculate the likelihood for the combination of genotypes using the probability (1) as

$$L\left(g_i^{(1)}, g_i^{(2)} | B_i; \hat{\alpha}_k\right) = P\left(B_i | g_i^{(1)}, g_i^{(2)}; \hat{\alpha}_k\right) = \prod_{j=1}^{R_i} \left[(1-\hat{\alpha}_k) P\left(b_{ij} \middle| g_i^{(1)}\right) + \hat{\alpha}_k P\left(b_{ij} \middle| g_i^{(2)}\right)\right]$$

where $B_i = \left\{b_{ij} | \, j = 1 \ldots R_i\right\}$ is the set of bases overlapping position i in the sequence reads that cover the variant site. Usually, we do not know the genotype of the contaminating sample, and so we sum over this unknown variable to obtain the genotype likelihood

$$L\left(g_i^{(1)} | B_i; \hat{\alpha}_k\right) = P\left(B_i | g_i^{(1)}; \hat{\alpha}_k\right) = \sum_{g_i^{(2)}} \left[P\left(g_i^{(2)}\right) P\left(B_i | g_i^{(1)}, g_i^{(2)}; \hat{\alpha}_k\right)\right].$$

In contrast to the analysis in which we identified contaminated samples and estimated contamination level α for each sample k using a list of known variant sites and allele frequencies, during genotype calling we examine every site. This step requires allele

frequencies at each site, which we estimate using the EM algorithm[5] to maximize the above likelihood. Thus, we estimate the allele frequency as:

$$\widehat{f_i} = \arg\max_{f_i} \prod_{k=1}^{n} \left[\sum_{g_{ik}} P(g_{ik}|f_i) P(B_{ik}|g_{ik};\ \hat{\alpha}_k) \right]$$

where g_{ik} is the true genotype for individual k $(1 \leq k \leq n)$ at site i. Given the allele frequency estimate $\hat{f}_i$, we estimate the genotype probabilities assuming Hardy-Weinberg equilibrium.

Finally, to call a genotype for an individual at locus i, we select the value of $g_i^{(1)}$ with the highest likelihood. We calculate the corresponding genotype dosage (D_i ranging from 0 to 2) for bi-allelic sites by taking a weighted average of the number of alternative alleles for each of the possible genotypes $g_i^{(1)}$

$$D_i = \frac{P\left(g_i^{(1)} = \text{AR} \mid B_i; \hat{\alpha}, \widehat{f_i}\right) + 2 \cdot P\left(g_i^{(1)} = \text{AA} \mid B_i; \hat{\alpha}, \widehat{f_i}\right)}{P\left(g_i^{(1)} = \text{RR} \mid B_i; \hat{\alpha}, \widehat{f_i}\right) + P\left(g_i^{(1)} = \text{AR} \mid B_i; \hat{\alpha}, \widehat{f_i}\right) + P\left(g_i^{(1)} = \text{AA} \mid B_i; \hat{\alpha}, \widehat{f_i}\right)} \qquad (2)$$

where R and A are the reference and alternate alleles and

$$P\left(g_i^{(1)} \mid B_i; \hat{\alpha}, \widehat{f_i}\right) \propto P\left(B_i | g_i^{(1)}; \hat{\alpha}\right) P\left(g_i^{(1)}; \widehat{f_i}\right).$$

Genotype likelihoods for contaminated sequence data: source known

If the identity of the contaminating sample is known, as in the type 2 diabetes example described in the Introduction, we can use that information to improve genotype calls. In that case, we examine all available data from the paired DNA samples and call their genotypes simultaneously by considering all potential 3 x 3 = 9 genotype pairs ($g_i^{(1)}$, $g_i^{(2)}$). Let $B_i^{(1)} = \left\{ b_{ij}^{(1)} | \, j = 1 \ldots \, R_i^{(1)} \right\}$ and $B_i^{(2)} = \left\{ b_{ij}^{(2)} | \, j = 1 \ldots \, R_i^{(2)} \right\}$ be the observed bases for reads labeled

as originating from samples 1 and 2, respectively, and let $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ be the estimated contamination levels for those two samples. We then write the joint likelihood for the paired samples as

$$L\left(g_i^{(1)}, g_i^{(2)} | B_i^{(1)} B_i^{(2)}; \hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}\right)$$

$$= \prod_{j=1}^{R_i^{(1)}} \left[(1-\hat{\alpha}^{(1)}) P\left(b_{ij}^{(1)} \middle| g_i^{(1)}\right) + \hat{\alpha}^{(1)} P\left(b_{ij}^{(1)} \middle| g_i^{(2)}\right)\right]$$

$$\times \prod_{j=1}^{R_i^{(2)}} \left[\hat{\alpha}^{(2)} P\left(b_{ij}^{(2)} \middle| g_i^{(1)}\right) + (1-\hat{\alpha}^{(2)}) P\left(b_{ij}^{(2)} \middle| g_i^{(2)}\right)\right]$$

This likelihood can also be calculated for different possible contaminating samples and compared to find the most likely source contamination (assuming both samples were sequenced).

When inconvenient to work with the joint likelihood (such as when calculating per-individual dosages), we calculate per-sample genotype likelihoods by marginalizing over the partner genotype.

$$L\left(g_i^{(1)} | B_i^{(1)} B_i^{(2)}; \hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}\right) = \sum_{g_i^{(2)}} \left[P\left(g_i^{(2)}\right) P\left(B_i^{(1)} B_i^{(2)} | g_i^{(1)}, g_i^{(2)}; \hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}\right)\right]$$

We also calculate these individual likelihoods prior to genotype refinement for low-pass sequence data (see below).

LD refinement for low-pass sequence data

Genotype refinement using linkage disequilibrium (LD) on low-pass sequence data leverages information about surrounding markers to help infer haplotypes and improve genotype accuracy[6,7]. After adjustment for contamination, we use Beagle[6] on our genotype likelihoods for low-pass (4-6x) whole genome data to refine and improve genotype calls. Such an adjustment is less important for exome sequence data because of insufficient flanking markers to infer haplotypes accurately.

Experimental data

To construct *in-silico* contaminated samples to test our methods, we chose 198 European 1000 Genomes Phase 1 samples[1] with (a) low-pass (4-6x) genome sequence data, (b) high-depth (50-150x) whole exome sequence data, (c) Illumina HumanOmni2.5 and HumanExome chip data, and (d) estimated contamination levels $\hat{\alpha}$ < 0.5% for chip and sequence data. We chose two samples at a time (without replacement) and combined sequence reads to achieve synthetic contamination levels α from 2% to 30%. We paired samples with similar depths so as to approximately preserve total read counts and varied the proportion of contaminated samples π in each simulation from 0 to 100%.

We also analyzed 1,503 samples from a type 2 diabetes exome sequencing project (average sequencing depth ~100x), 1,009 of which (67%) were estimated to have contamination level $\hat{\alpha}$ > 5%. In this study, we learned after sequencing was completed that changes to sequencing library preparation protocols that were designed to improve efficiency and reduce cost resulted in contamination due to template switching during PCR amplification

of pairs of barcoded samples. In this case, we could reconstruct the identity of the contaminating sample by checking experimental records to identify samples that were amplified together.

Evaluation

For both examples, we compared sequence-based best-guess genotypes and genotype dosages to available array-based genotypes to estimate genotype concordance and squared Pearson's correlation r^2 between true genotypes and estimated genotype dosages. The genotypes for the *in-silico* contaminated low-pass samples were LD-refined, and then compared to all 41,847 Illumina HumanOmni2.5 genotype array chromosome 20 SNPs. Genotypes for *in-silico* contaminated high-depth samples were compared to all 33,884 SNPs from the Illumina HumanExome array that were polymorphic within the 198 1000 Genomes Project samples. Genotypes for the type 2 diabetes example were compared to all 3,881 SNPs from the Affymetrix 6.0 array that overlapped the targeted sequence regions and were variable within the sequenced samples.

Results

In-silico contaminated data: contaminating sample unknown

When we did not model contamination, increasing DNA contamination levels (α) resulted in decreasing concordance between sequence and array genotypes. For low-pass whole genome sequence data, as α increased from 2% to 30%, total genotype concordance decreased from 98.1% to 83.8%, compared to an average concordance of 98.9% for uncontaminated samples (Figure 3-1A; Appendix Table 3-S1). For high-depth exome sequence

data, total concordance decreased from 99.6% to 92.9% over the same contamination range compared to 99.8% for uncontaminated samples (Figure 3-1B; Appendix Table 3-S1). Similarly, r^2 values for genotype dosages decreased from >0.96 to <0.75 as α increased from 2% to 30% (Figure 3-1CD). Genotyping errors resulted in an increase in heterozygous calls roughly equal to α for the high-depth data and α/2 for the low-pass data (Appendix Figure 3-S1). The impact of contamination was greater for common variants than for rare ones (Appendix Table 3-S1), corresponding to the greater probability of a contamination resulting in a false heterozygote.

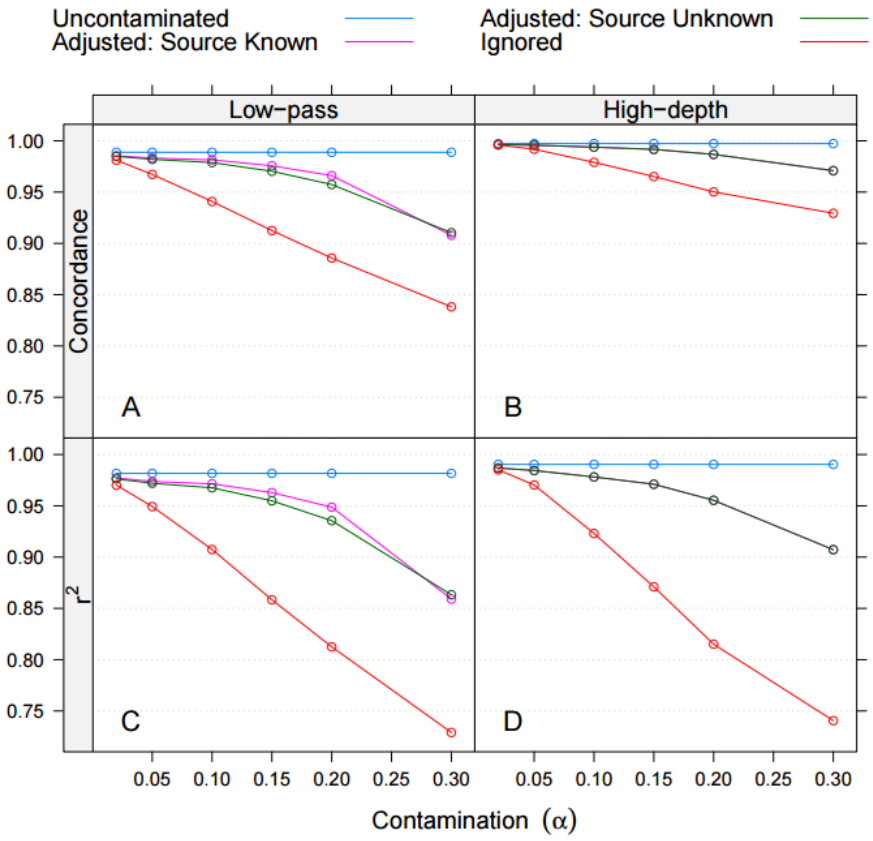

Figure 3-1 ***Effects of contamination adjustment on constructed contaminated DNA samples: genotype concordance and r^2.*** *Each point represents overall genotype concordance or dosage r^2 for contaminated samples when the proportion of contaminated samples π=50%*

Applying our method to these contaminated samples markedly increased genotype concordance and genotype dosage r^2. Estimated sample contamination levels $\hat{\alpha}_k$ closely matched intended α values (Appendix Table 3-S2). By accurately modeling contamination, we reduced the difference in genotype concordance rates between the contaminated and uncontaminated samples by up to 60-80% for the high-depth exomes and up to 50-80% for the LD-refined low-pass genomes (Figure 3-1AB) for contamination levels 5%-20%. We observed a similar pattern for r^2 (Figure 3-1CD). For the low-pass data, these improvements were seen only after LD-refinement (Appendix Figure 3-S2).

Joint calling uncontaminated samples with contaminated samples had little effect on the genontypes for the uncontaminated samples. For low-pass data, when the proportion of contaminated samples π=50% and contamination levels $\alpha \leq 30\%$, the largest observed reduction in genotype concordance for uncontaminated samples was 0.4%; average reductions were ~0.2%. Results changed only slightly as we varied the proportion of contaminated samples from π=5% to 90% (Figure 3-2). For high-depth data, the effect using our contamination-aware likelihoods when calling genotypes for uncontaminated samples was negligible for all π and α (Appendix Table 3-S3).

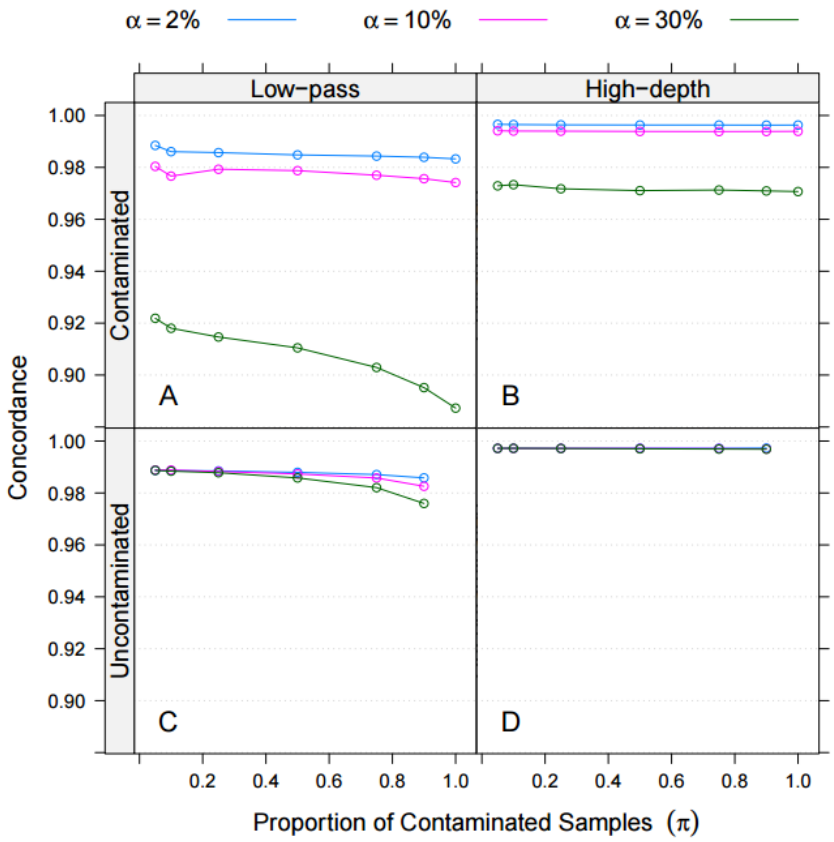

Figure 3-2 Effects of increasing proportion of contaminated samples π on genotype concordance for various levels of contamination α

In-silico contaminated data: contaminating sample known

When the source of the contaminating DNA sample was known and sequence data for both samples was available, modeling this information explicitly further improved concordance with array genotypes. For low-pass data, adding the pair information reduced the difference in concordance by an additional ~25% as α increased from 2% to 20% (Figure 3-1A). However, at α=30%, concordance was actually slightly lower. This reduction in concordance appears only after LD-adjustment on the data; it may be the result of a loss of information from marginalizing our pairwise genotype likelihoods as required for analysis with Beagle. Improvements to r^2

ranged from 0.1% to 1.3% for α=2% to 20%. For high-depth data, we did not see a meaningful difference in concordance or r^2 when using the known pair information (Figure 3-1B).

In-silico contaminated data: association information

Ultimately we wish to use the sequence-based genotypes to test for disease or trait association. In association analysis, we can choose one of three strategies: (1) ignore contamination, (2) exclude highly contaminated samples from analysis, or (3) adjust for contamination. To estimate the relative efficiencies of these three strategies, we note that effective sample size scales linearly with nr^2, the product of sample size and the squared correlation between the true genotype and the sequence-based genotype dosages[8]. Since even contaminated samples provide information about the true underlying genotype (r^2>0), including contaminated samples could provide association information even when contamination is ignored. The reduction in sample size due to contamination is at least 80% smaller when applying our correction compared to dropping contaminated samples (Table 3-2). In our evaluations, we maximized effective sample size when adjusting for contamination and using all samples, whether contaminated or not. For example, when all samples are contaminated at $\alpha = 10\%$, association information for the low-pass data is reduced by 10.6% if we ignore contamination and 4.0% if we correct for contamination (compared to 8.0% and 2.5% respectively for high-depth data). In this example, where all samples are contaminated, it would have been impractical to exclude contaminated samples from association analyses.

	Low-Pass						
	% of Samples Contaminated						
Method	**5%**	**10%**	**25%**	**50%**	**75%**	**90%**	**100%**
Adjusted	194	194	193	192	191	190	190
Ignored	193	193	190	186	182	179	177
Dropped	184	174	144	96	47	18	0
	High-Depth						
	% of Samples Contaminated						
Method	**5%**	**10%**	**25%**	**50%**	**75%**	**90%**	**100%**
Adjusted	195	195	195	194	194	193	193
Ignored	195	194	192	189	186	184	182
Dropped	186	176	146	98	48	18	0

Table 3-2 ***Effective sample size for association test*** *Shown here are the effective sample size estimates when α=10% and total sample size is 198 under three scenarios: all samples included and adjusted with our method ("adjusted"), all samples included but contamination ignored ("ignored"), and contaminated samples (α̂>0.01) removed from analysis ("dropped")*

In-silico contaminated data: impact of over- or underestimating contamination

To evaluate whether misspecified values of α could result in decreased genotype quality, we ran simulations in which we scaled the contamination estimate $\hat{\alpha}$ by 0.5, 0.75, 1.5, and 2 for samples in which the true α=5%, 10%, or 15%. Overestimating $\hat{\alpha}$ had little impact on total concordance and r^2 while underestimating contamination more negatively affected both statistics (Figure S3). For the low-depth data, overestimating $\hat{\alpha}$ by 1.5x actually resulted in better concordance then using the "true" $\hat{\alpha}$; this effect was only observed after LD-refinement The difference in concordance when reducing α by half was at least 40% greater than difference from of doubling α for the low-pass samples; there was very little difference for the high-depth samples. The negative impact of inflated $\hat{\alpha}$ estimates for samples that were not contaminated was very modest compared to the benefits of modeling contamination for the remaining samples.

Type 2 diabetes data

Convinced of the value of adjusting for contamination, we next applied our method to data from the type 2 diabetes exome sequencing project. In these data, $\hat{\pi}$=67% of samples were contaminated and we knew the likely contaminating sample. When we applied our correction methods, concordance with array genotypes dramatically improved: the average per-sample concordance increased from 94.5% to 99.4% (a 9-fold reduction in discordance), further increasing to 99.6% (a 14-fold reduction in discordance) when we both modeled contamination and used knowledge of its source. Similar patterns were observed for non-reference concordance and r^2 (Table 3-3).

	$\hat{\alpha}$	# Samples	Ignored	Adjusted	Paired
Total Concordance	0-1%	202	0.998	0.998	0.998
	1-5%	293	0.996	0.998	0.998
	5-10%	218	0.958	0.997	0.998
	10-15%	591	0.920	0.993	0.996
	15-20%	169	0.878	0.984	0.992
	>20%	30	0.841	0.950	0.971
	ALL	*1503*	*0.945*	*0.993*	*0.996*
Non-Ref Concordance	0-1%	202	0.996	0.997	0.997
	1-5%	293	0.992	0.995	0.995
	5-10%	218	0.908	0.993	0.994
	10-15%	591	0.833	0.985	0.991
	15-20%	169	0.760	0.964	0.983
	>20%	30	0.702	0.890	0.936
	ALL	*1503*	*0.882*	*0.985*	*0.991*
r^2	0-1%	202	0.997	0.998	0.998
	1-5%	293	0.994	0.996	0.996
	5-10%	218	0.929	0.995	0.996
	10-15%	591	0.863	0.990	0.994
	15-20%	169	0.791	0.977	0.989
	>20%	30	0.725	0.930	0.946
	ALL	*1503*	*0.905*	*0.990*	*0.994*

Table 3-3 GWAS concordance for type 2 diabetes exome sequencing data (Mean per-sample genotype accuracy with the GWAS data when we ignore contamination, adjust without regard for the source of contamination, and adjust using known contamination source)

Discussion

We have shown that genotyping accuracy for contaminated samples can be dramatically improved by modeling contamination using a mixture model. For example, in the type 2 diabetes exome sequencing example, our method reduced genotype discordance by 14-fold (4.2% to 0.3%) for α=5-10% contaminated samples. Consistent with our previous study, we observed that even low levels of contamination (e.g., α=2-5%) can result in increases in genotype discordance of >2-fold. Our correction method nearly eliminates the impact of low levels of DNA contamination (α=2-5%) and reduces by >80% genotype discordance incurred by

moderate level of DNA contamination (α=5-15%) in the type 2 diabetes exome sequencing examples. We expect our method to be particularly useful when a large fraction of sequenced samples are contaminated at small to moderate levels (α=2-15%). Below we discuss the robustness of our approach when model assumptions are not met, and explore other scenarios where these or similar modeling approaches might be useful.

We demonstrated (Appendix Figure 3-S3) that genotype calling methods that model contamination perform best when the contamination level α is well estimated and that underestimating α is more detrimental than overestimating it. Situations that may lead to deflated contamination estimates are (1) the use of misspecified allele frequency estimates (incorrect population as well as systematic overestimates or underestimates; data not shown), (2) contamination from related individuals[2], or (3) limited sequencing library complexity which results in decreased heterozygosity. If one or more of these situations is suspected, modestly inflating (2-5%) the estimated contamination level $\hat{\alpha}$ when correcting for contamination may improve overall genotype accuracy.

As long as contamination affects case and control samples similarly, we do not expect contamination adjustments to increase the rate of false positive findings in downstream association studies. For single-variant associations, results depend on accurate estimations of allele frequency differences in cases and controls. As long as contamination patterns do not differ drastically in the cases and controls and there are no issues of population stratification, we can accurately estimate allele frequencies after correction (Appendix Figure 3-S4). For rare-variant association, contaminated samples may appear to carry high numbers of rare heterozygous variants when analyzed with standard protocols. Our proposed correction will

decrease the number of false positive heterozygotes (Appendix Figure 3-S5), so false positive associations will be less likely.

While we have focused on sequencing genomic DNA, in principle our methods can be used for other sequencing studies as well. For example, we have used our methods to identify contamination in RNA-seq experiments. Using our existing method and restricting analyses to expressed exons in protein-coding genes, we detected that 11 of 249 RNA-seq samples were contaminated by >2%. Detection and estimation of contamination in these experiments may be made more robust by accounting for allele-specific expression (ASE), where gene transcription varies based on allele; we are exploring this possibility.

We described the methods in this paper specifically in the context of biallelic SNPs. Extension to multiallelic SNPs is straightforward, requiring only that we sum over a larger number of possible genotypes. Genotyping of other variant types, such as indels and structural variants, is also affected by contamination. We expect that the same principles, focused on modeling the observed data as a mixture of two samples, can be usefully applied to these more complex situations.

We observed that the LD-aware genotype refinement algorithm improves genotype accuracy for low-pass sequence data. However, accuracy was still substantially lower than for uncontaminated data when the contamination level α was high. This may be due in part to the fact that our LD-aware genotype refinement algorithm is not aware of the possibility of contamination. With increasing interest in whole genome sequencing studies, accounting for

the contamination in the genotype refinement step has the potential to further improve genotyping and phasing accuracy.

Our contamination modeling methods are implemented in the program cleanCall. Source code for this program is available online. cleanCall requires sequencing data in samtools [Li et al. 2009] pileup format. Extracting pileups only for variant sites allows cleanCall to read data quickly compared to scanning large BAM files. The total runtime for cleanCall is comparable to other simple likelihood-based genotype callers; modest additional time is spent estimating allele frequencies via the EM algorithm, but the average number of iterations at a given site is minimal (2-5) and does not significantly affect overall performance.

We developed methods to correct for DNA contamination in variant calling by extending our likelihood-based framework to detect and estimate contamination. Our correction methods improve genotype calling accuracy and association power compared to ignoring contamination or discarding contaminated samples. Even if the contamination level is small ($\hat{\alpha}$<5%), we observe considerable improvement in genotype accuracy using our correction methods. Our methods are effective both for high-depth and low-pass data, and given the ubiquity of DNA sample contamination, we expect that our methods to be of real benefit to a large number of DNA sequencing studies.

Chapter 3 References

1) The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
2) Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet *91*, 839–848.
3) Ewing, B., Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. Genome Res *8*, 175-185.
4) Brent, R.P. (1973). Algorithms for Minimization without Derivatives. (New Jersey: NJ: Prentice-Hall).
5) Dempster, A.P., Laird N.M., Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B *39*, 1-38.
6) Browning, B.L., Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet *85*, 847–861.
7) Li, Y., Sidore, C., Kang, H.M., Boehnke, M., Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome Res *21*, 940-951.
8) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.R., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

Chapter 3 Appendix

Figure 3-S1 – Overcalling heterozygous genotypes in contaminated data

These boxplots show the relative excess heterozygote genotypes: the average number of heterozygous genotypes from sequence-based analyses of sites genotyped using arrays, divided by the number of heterozygous genotypes in the array data (per sample)

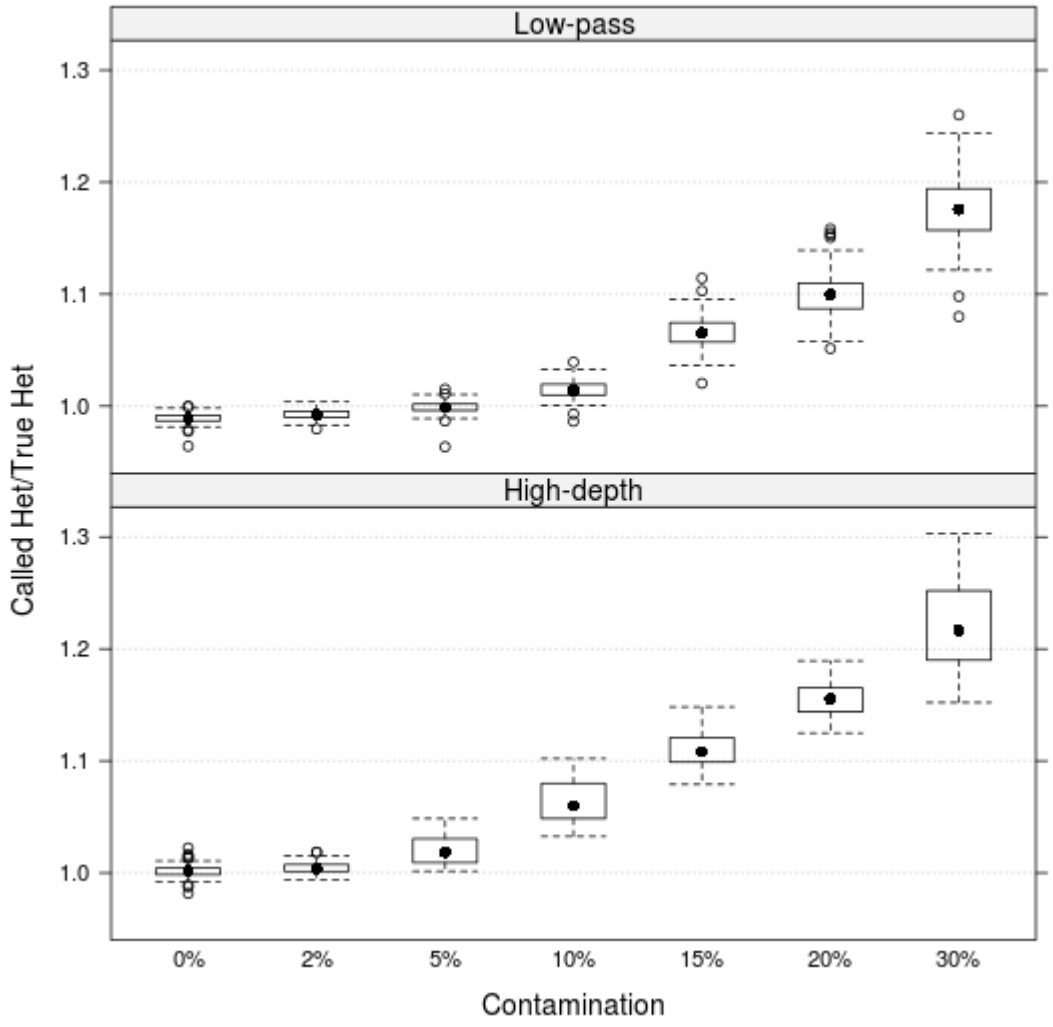

Figure 3-S2 – Effects of LD-refinement on adjusted calls for low-pass data

We saw a modest improvement in genotype calls in the low-pass data prior to LD-refinement; after refinement, the effects were substantial. Each color represents the value for α used to simulate *in-silico* contamination

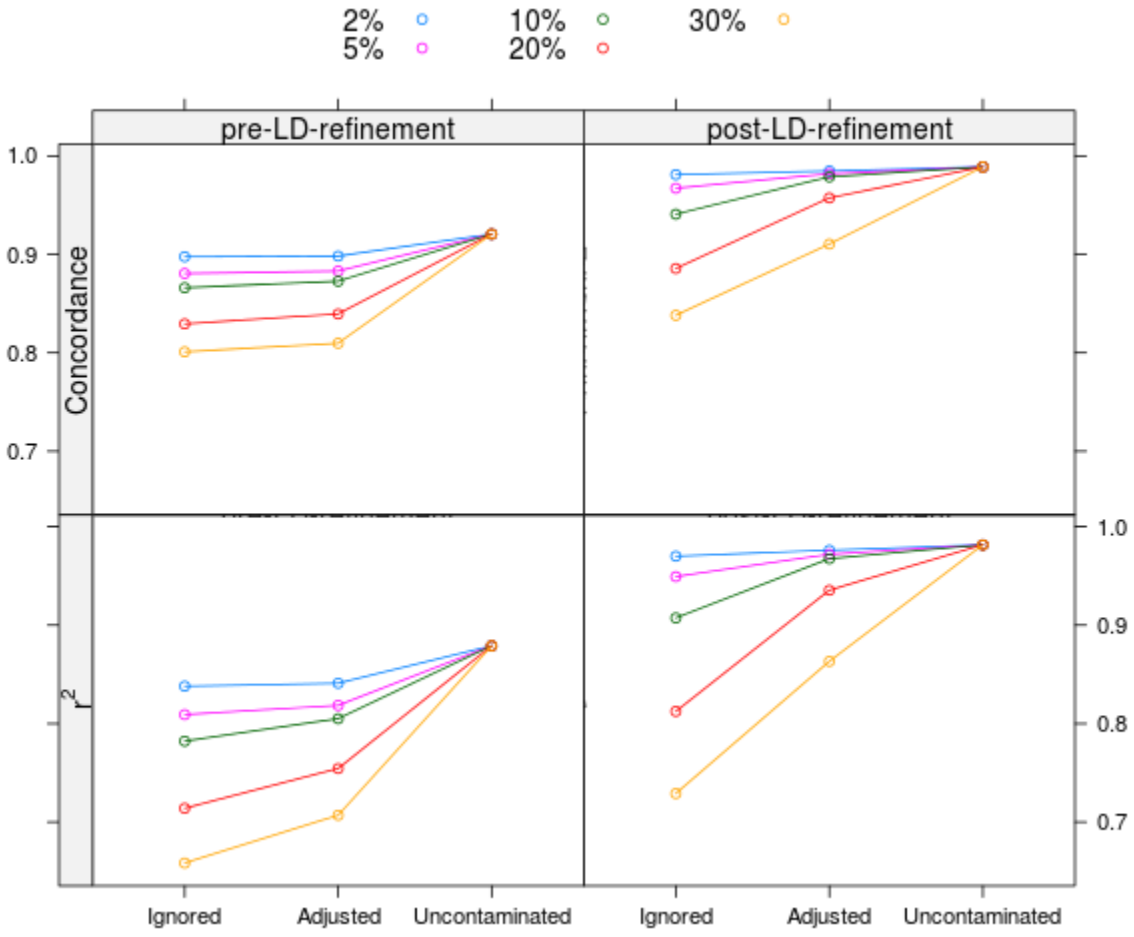

Figure 3-S3 – Effects of incorrect estimation of the contamination level α

Effects of incorrect estimation of α on total genotype concordance and dosage r^2 for contaminated and uncontaminated samples when $\pi = 0.50$. The scaling factor applied to $\hat{\alpha}$ is listed along the x-axis. The values for "UN" are the measures where all samples are uncontaminated and "IG" are the values where contamination is ignored.

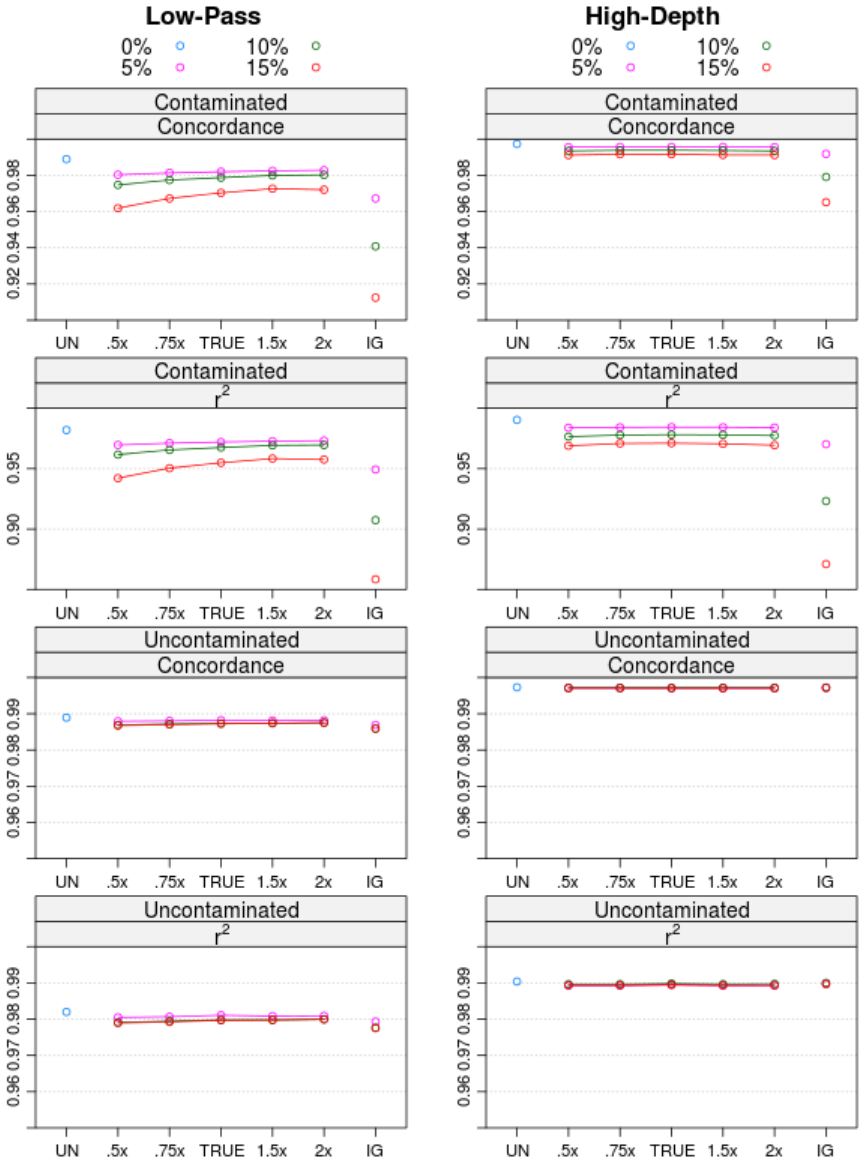

Figure 3-S4 – Allele frequency estimation with contaminated data

This is a qq-plot comparing the distribution of allele frequency estimates ($\hat{f}$) from the array-based genotypes against the distribution of allele frequencies calculated from the sequencing-based genotypes for both the uncontaminated samples and contaminated samples after adjustment ($\alpha = 0.15$). We have scaled the frequencies using a $-\log_{10}$ transformation to focus on the rarer variants.

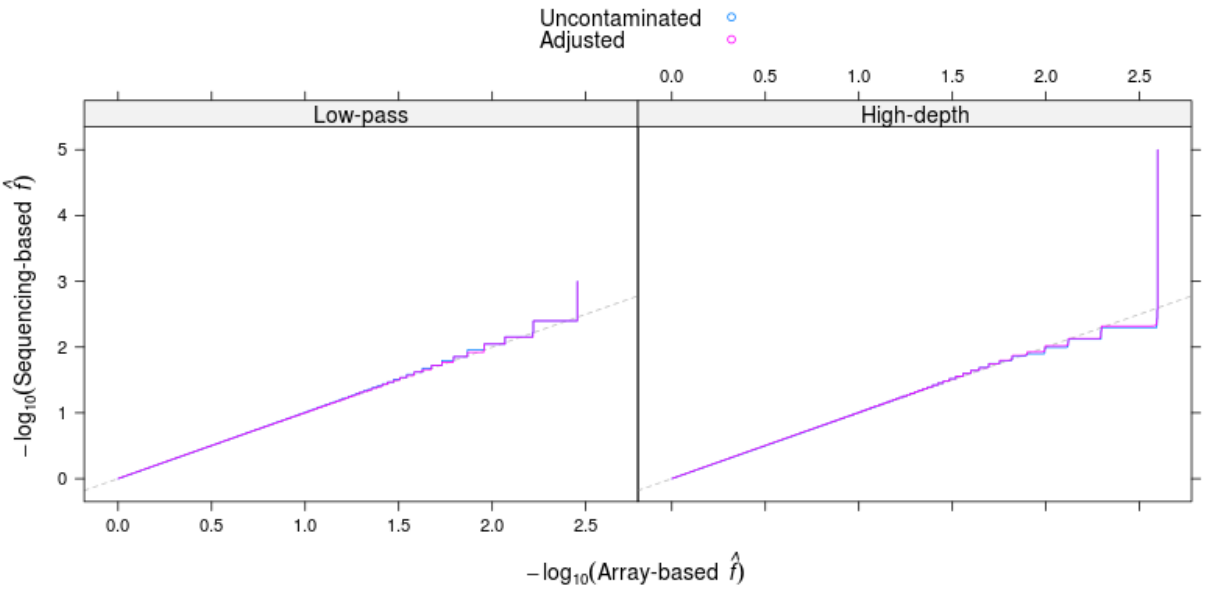

Figure 3-S5 – False positive heterozygote SNPs

Here we have plotted the distribution of the ratios of heterozygous SNPs to non-reference homozygous SNPs for samples with $\alpha = 15\%$. We expect ratios close to 2 based on observations from genotype data in other studies. The high ratio for samples where contamination was ignored likely indicates many false positive heterozygote genotypes.

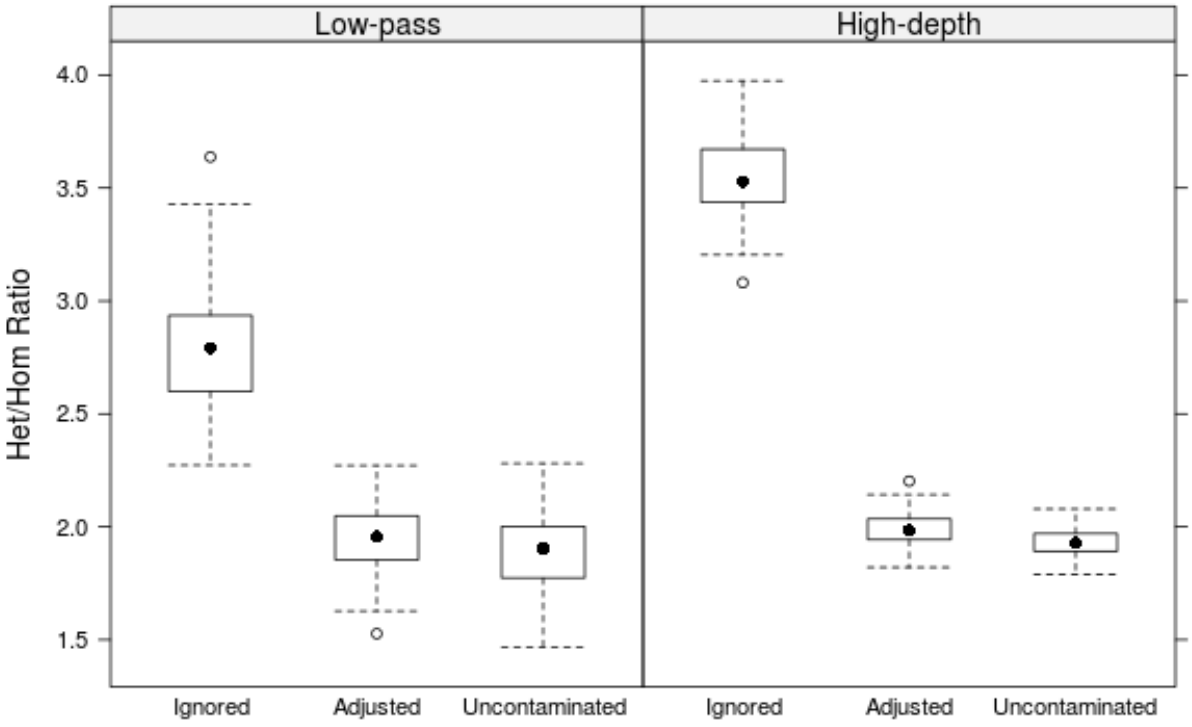

Table 3-S1 – Genotype accuracy for contaminated samples

Total genotype concordance and dosage r^2 for contaminated samples when π=50% of the samples in the call set were contaminated.

		Low-Pass					High-Depth		
	α	Ignored	Adjusted	Paired		α	Ignored	Adjusted	Paired
Total	0%		0.989		Total	0%		0.997	
Concordance	2%	0.981	0.985	0.985	Concordance	2%	0.996	0.996	0.996
	5%	0.967	0.982	0.983		5%	0.992	0.996	0.996
	10%	0.941	0.979	0.981		10%	0.979	0.994	0.994
	15%	0.913	0.970	0.976		15%	0.965	0.992	0.991
	20%	0.886	0.957	0.966		20%	0.950	0.987	0.987
	30%	0.838	0.910	0.908		30%	0.929	0.971	0.971
Non-Ref	0%		0.973		Non-Ref	0%		0.981	
Concordance	2%	0.955	0.963	0.965	Concordance	2%	0.971	0.975	0.975
	5%	0.925	0.957	0.960		5%	0.945	0.971	0.971
	10%	0.869	0.949	0.955		10%	0.868	0.959	0.959
	15%	0.806	0.930	0.942		15%	0.791	0.944	0.944
	20%	0.755	0.900	0.920		20%	0.719	0.913	0.913
	30%	0.669	0.798	0.794		30%	0.638	0.817	0.817
r^2	0%		0.982		r^2	0%		0.990	
	2%	0.970	0.976	0.977		2%	0.984	0.986	0.986
	5%	0.949	0.972	0.974		5%	0.970	0.984	0.984
	10%	0.907	0.967	0.971		10%	0.923	0.978	0.978
	15%	0.858	0.955	0.963		15%	0.871	0.971	0.971
	20%	0.813	0.936	0.949		20%	0.815	0.955	0.955
	30%	0.729	0.863	0.859		30%	0.740	0.907	0.907
Allele Freq <1%	0%		0.997		Allele Freq <1%	0%		1.000	
Concordance	2%	0.996	0.996	0.997	Concordance	2%	1.000	1.000	1.000
	5%	0.994	0.996	0.996		5%	0.999	1.000	1.000
	10%	0.994	0.995	0.996		10%	0.999	0.999	0.999
	15%	0.992	0.993	0.994		15%	0.997	0.999	0.999
	20%	0.990	0.991	0.991		20%	0.995	0.999	0.999
	30%	0.985	0.987	0.985		30%	0.991	0.998	0.998
Allele Freq 1-5%	0%		0.993		Allele Freq 1-5%	0%		0.998	
Concordance	2%	0.990	0.991	0.992	Concordance	2%	0.997	0.997	0.997
	5%	0.986	0.989	0.990		5%	0.996	0.997	0.997
	10%	0.978	0.988	0.990		10%	0.990	0.996	0.996
	15%	0.966	0.983	0.986		15%	0.983	0.995	0.995
	20%	0.955	0.976	0.979		20%	0.971	0.992	0.992
	30%	0.933	0.952	0.942		30%	0.953	0.983	0.983
Allele Freq >5%	0%		0.986		Allele Freq >5%	0%		0.992	
Concordance	2%	0.976	0.981	0.981	Concordance	2%	0.987	0.989	0.989
	5%	0.957	0.977	0.979		5%	0.974	0.987	0.987
	10%	0.922	0.973	0.976		10%	0.934	0.981	0.981
	15%	0.881	0.962	0.969		15%	0.891	0.974	0.974
	20%	0.845	0.946	0.958		20%	0.849	0.960	0.960
	30%	0.781	0.884	0.884		30%	0.793	0.912	0.912

Table 3-S2 – Estimated contamination (mean and SD) for constructed contaminated samples

	Low-Pass		**High-Depth**	
Intended α (%)	Mean $\hat{\alpha}$	SD $\hat{\alpha}$	Mean $\hat{\alpha}$	SD $\hat{\alpha}$
2	1.8	0.5	2.2	0.2
5	4.6	0.7	5.6	0.5
10	9.4	1.0	10.7	0.7
15	14.1	1.4	15.4	1.0
20	18.7	1.7	19.8	1.2
30	27.6	2.4	27.3	2.2

Table 3-S3 – Genotype accuracy for uncontaminated samples

Total genotype concordance and r^2 for the uncontaminated samples when π=50% of the samples in the call set were contaminated.

		Low-Pass					High-Depth		
	α	Ignored	Adjusted	Paired		α	Ignored	Adjusted	Paired
Total Concordance	0%		0.989		Total Concordance	0%		0.997	
	2%	0.988	0.988	0.988		2%	0.997	0.997	0.997
	5%	0.987	0.988	0.988		5%	0.997	0.997	0.997
	10%	0.986	0.987	0.988		10%	0.997	0.997	0.997
	15%	0.986	0.987	0.987		15%	0.997	0.997	0.997
	20%	0.985	0.986	0.987		20%	0.997	0.997	0.997
	30%	0.985	0.986	0.986		30%	0.997	0.997	0.997
Non-Ref Concordance	0%		0.973		Non-Ref Concordance	0%		0.981	
	2%	0.971	0.971	0.971		2%	0.981	0.981	0.981
	5%	0.970	0.971	0.972		5%	0.980	0.980	0.980
	10%	0.967	0.969	0.970		10%	0.981	0.980	0.980
	15%	0.966	0.969	0.970		15%	0.980	0.979	0.979
	20%	0.965	0.967	0.968		20%	0.981	0.980	0.980
	30%	0.964	0.966	0.966		30%	0.980	0.979	0.979
r^2	0%		0.982		r^2	0%		0.990	
	2%	0.980	0.981	0.981		2%	0.990	0.990	0.990
	5%	0.979	0.981	0.981		5%	0.989	0.989	0.989
	10%	0.978	0.980	0.980		10%	0.989	0.989	0.989
	15%	0.977	0.980	0.980		15%	0.989	0.989	0.989
	20%	0.977	0.979	0.979		20%	0.990	0.989	0.989
	30%	0.976	0.978	0.978		30%	0.989	0.989	0.989
Allele Freq <1% Concordance	0%		0.997		Allele Freq <1% Concordance	0%		1.000	
	2%	0.997	0.997	0.997		2%	1.000	1.000	1.000
	5%	0.997	0.997	0.997		5%	1.000	1.000	1.000
	10%	0.996	0.997	0.997		10%	1.000	1.000	1.000
	15%	0.997	0.997	0.997		15%	1.000	1.000	1.000
	20%	0.997	0.997	0.997		20%	1.000	1.000	1.000
	30%	0.997	0.997	0.997		30%	1.000	1.000	1.000
Allele Freq 1-5% Concordance	0%		0.993		Allele Freq 1-5% Concordance	0%		0.998	
	2%	0.993	0.993	0.993		2%	0.998	0.998	0.998
	5%	0.993	0.993	0.993		5%	0.998	0.998	0.998
	10%	0.992	0.993	0.993		10%	0.998	0.998	0.998
	15%	0.992	0.993	0.993		15%	0.998	0.998	0.998
	20%	0.992	0.992	0.992		20%	0.998	0.998	0.998
	30%	0.991	0.992	0.992		30%	0.998	0.998	0.998
Allele Freq >5% Concordance	0%		0.986		Allele Freq >5% Concordance	0%		0.992	
	2%	0.985	0.985	0.985		2%	0.992	0.991	0.991
	5%	0.984	0.985	0.985		5%	0.991	0.991	0.991
	10%	0.982	0.984	0.984		10%	0.991	0.991	0.991
	15%	0.982	0.984	0.984		15%	0.991	0.991	0.991
	20%	0.981	0.983	0.983		20%	0.992	0.991	0.991
	30%	0.981	0.982	0.982		30%	0.991	0.991	0.991

Chapter 4
Detecting Contamination in RNA Sequencing Experiments

Abstract

Until recently, microarrays were the standard way to collect data about differences in gene expression between individuals. However, microarrays can only target specific regions in known transcripts, giving only a partial view of gene expression. The next evolution of gene expression data collection is RNA-Seq, which enables genome-wide studies of both known and novel transcripts. By converting mRNA to cDNA, we can leverage the speed and accuracy of DNA sequencing machines to collect lots of data without much effort. However, as with any genetic data, the possibility exists for a sample to become contaminated with the RNA of a different individual. Here we propose a likelihood based model to detect and quantify inter-sample RNA contamination using data already generated by the sequencer without requiring an additional experiment. Our method produces estimates with an average error of 0.5% for contamination levels from 2%-10%.

Introduction

Just as high-throughput sequencing technologies have revolutionized the collecting of sequence information from genomic DNA, these technologies have created new opportunities to investigate gene expression. RNA-Seq is a method by which RNA is captured from a cell, converted to cDNA, and undergoes sequencing on a sequencing machine[1]. Typical high-

throughput pipelines produce tens of millions of short reads (35-100bp) from the collection of cDNA fragments. Then an aligner or assembler determines the most likely genomic positions from which the fragments originated[2]. Once fragments are placed, tools are available to compare the relative number of observed transcripts for each gene to understand which genes are being expressed and at what abundance.

While current sequencing machines are capable of quickly producing vast amounts of sequencing data with few read errors, there is still a need to thoroughly check the quality of the data prior to analysis. For example, samples sequenced at different times or with different library preparations often have batch effect differences[4]. Differences in gene transcript abundance may simply be due to technical artifacts from library preparation or RNA capture rather than true expression differences. Failure to account for these differences may result in false positive associations or reduced power. Proper quality control is an important step in a RNA-Seq processing pipeline.

One important quality control measure for all NGS data is a screen for sample contamination. We previously developed methods to identify and quantify contamination in DNA sequencing studies in which DNA from two or more individuals are present in a single sample[3]. This method has been used to detect contamination in many large sequencing studies and is recommended as a standard test during DNA sequence quality control[5,6].

There are three basic types of within-species contamination that may occur in RNA-Seq studies: 1) intra-sample DNA-RNA contamination, 2) intra-samples RNA-RNA contamination, and 3) inter-sample RNA contamination. Intra-sample DNA-RNA contamination occurs when

not all the genomic DNA is removed from the cDNA prior to sequencing. This sort of contamination manifests as reads mapping to intergenic regions and thus can be identified during read alignment. Intra-samples RNA-RNA contamination occurs when different tissue types from a single individual are sequenced together. In this case, the observed reads will not have genotype differences between tissues, but they may potentially have expression differences. Unfortunately most tissues are likely a mixture of cell types which makes resolving the source of each RNA fragment difficult. Since differences between cell types are more subtle because transcripts across tissues are likely to have identical genotypes, we will not attempt to model this type of contamination. In this paper we focus on the problem of inter-sample RNA contamination whereby the RNA from two different samples are sequenced together as one.

While the data produced by RNA-Seq and genomic DNA sequencing are similar, there are two characteristics of RNA-Seq data that could require more complex modeling for accurate contamination estimation than DNA sequencing data. First is gene expression variability. For DNA contamination detection, we may reasonably assume that samples contribute sequence reads in proportion to the contamination level at each genomic position. For RNA-Seq, the two samples may be expressing genes at different levels. In the most extreme case, an intended sample does not express a gene but the contaminating sample does, so that all the reads for this gene would come from the contaminating sample and on its own may not appear contaminated. Second is allele specific expression (ASE). In contrast to DNA sequencing in which we expect either allele at a heterozygous site is equally likely observed, ASE alters that expectation by preferentially transcribing from the DNA strand with a particular allele. The deviation from balanced chromosomal expression depends on the strength of ASE for the

particular gene. It has been suggested that ASE may occur for ~20% of human genes[7]. We are interested what effects these phenomena may have on the estimation of contamination for RNA-Seq data.

Materials and Methods

Here we provide a statistical model to detect contamination in RNA-Seq data and describe the experimental datasets used for its validation.

We begin by making a few simplifying assumptions about RNA-Seq data in order to model contamination. First we assume a list of known variant sites within transcripts with known allele frequencies. Second we assume all reads are independent and all observed bases at a particular site are independent. Third we assume that either allele is equally likely to be observed at a heterozygous site; thus this model does not explicitly model deviations as the result of ASE. Fourth we assume that when base read errors occur, all three other bases are equally likely observed. Finally, this formulation assumes gene expression is consistent across individuals.

Following our previous work for DNA sequence contamination[3], we model RNA sample contamination with a mixture model. Let $g_i^{(1)}$ and $g_i^{(2)}$ be the genotypes for the intended and contaminating samples at variant site i ($1 \leq i \leq M$), b_{ij} be the observed base at position i for read j ($1 \leq j \leq R_i$), e_{ij} a latent variable indicating whether a base calling error occurred ($e_{ij} = 1$) or did not ($e_{ij} = 0$), and α the proportion of reads from the contaminating sample.

We model the probability of observing a particular base b_{ij} as

$$P\left(b_{ij}\middle|g_i^{(1)},\ g_i^{(2)};\alpha\right) = (1-\alpha)P\left(b_{ij}\middle|g_i^{(1)}\right) + \alpha P\left(b_{ij}|g_i^{(2)}\right)$$

where

$$P\left(b_{ij}|g_i\right) = P\left(b_{ij}\middle|g_i,\ e_{ij}=1\right)P(e_{ij}=1) + P\left(b_{ij}\middle|g_i,\ e_{ij}=0\right)P(\ e_{ij}=0)$$

We present the read probabilities allowing for error $P\left(b_{ij}\middle|g_i,\ e_{ij}\right)$ in Table 1. We estimate the probability of a read error as $P(e_{ij}=1) = 10^{-Q_{ij}/10}$ and $P(e_{ij}=0) = 1 - P(e_{ij}=1)$, where Q_{ij} is the phred-scaled base quality score for the RNA sequence data[8]. To estimate the genotype probability $P(g_i)$, we use allele frequencies from the population from which the sample was drawn and assume Hardy-Weinberg equilibrium.

Taking expectations over the unknown genotypes and assuming all reads and loci are independent, we write the likelihood for contamination level α in an individual sample as

$$L(\alpha) = P(B|\alpha) = \prod_{i=1}^{M}\sum_{g_i^{(1)}}\sum_{g_i^{(2)}}\left\{P\left(g_i^{(1)}\right)P\left(g_i^{(2)}\right)\prod_{j=1}^{R_i}\left[(1-\alpha)P\left(b_{ij}\middle|g_i^{(1)}\right) + \alpha P(b_{ij}|g_i^{(2)})\right]\right\}$$

For each sample, we first maximize $L(\alpha)$ using a grid search in the interval $0.0 \le \alpha \le 0.5$ and then apply Brent's algorithm[9] to obtain the maximum likelihood estimate of α. By using information across a large number of variants M, we determine if the observed reads are better explained by a single sample or a combination of two samples with mixing proportion $0 < \alpha < 1$.

To validate this method, we constructed contaminated samples *in-silico* using publically available data from the GEUVADIS project[10]. We used RNA-Seq data for 452 samples drawn from the 1000 Genomes Project[11] from 5 different populations: Utah residents with Northern

and Western European ancestry (CEU), Finns in Finland (FIN), British in England and Scotland (GBR), Tuscans in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI). The RNA for each sample was extracted from lymphoblastoid cell lines. These samples also had Illumina HumanOmni2.5 array genotypes available from the 1000 Genomes Project. We only used samples that had estimated levels of RNA contamination <1% (dropped 8 samples). We combined reads from pairs of samples within populations adjusting for differences in overall read depth to simulate contamination levels from 2%-30%. We then estimated contamination using (subsets of) variant sites from the HumanOmni2.5 genotype array.

To evaluate performance in real-world setting, we estimated contamination using 185 samples with RNA-Seq data from an ongoing psoriasis skin RNA-Seq project. Samples were sequenced on an Illumina Genome Analyzer IIx with a read length of 80bp. Alignment was performed with BWA[12] against the NCBI build 37 human reference genome. Samples had genotype array data for the Illumina HumanExome chip which allowed us to verify identity.

In all analyses, we estimated contamination using our cleanCall software[13]. We increased the default setting for DNA studies for the maximum number of reads from 20 to 500 to better accommodate the larger read depth from RNA-Seq data. We used population allele frequency estimates specific to each population calculated from the HumanOmni2.5 genotype array data.

Results

When estimating contamination with DNA data, we used sites across the entire genome, but for RNA-Seq we observed more accurate results when using sites in exons. While

it is often possible to observe reads outside exons due to non-coding or novel transcripts, we focus on exons to avoid the impact of alignment errors enriched in the reads mapped outside the exome. We estimated contamination from the constructed samples (1) using a random set of sites genome-wide from the HumanOmni2.5 genotype array and (2) using a set of sites annotated as being in a gene exon by GENCODE[14] to compare estimates based on the genome and exome. We limited both sets of sites to roughly 100,000 so any differences were not based on simply using a different number of sites. We found that results using sites from all over the genome ignoring exonic annotation produced less accurate estimates of contamination compared to using sites in the exome (Figure 1). For example, when $\alpha = 15\%$, the mean absolute difference in the estimated $\hat{\alpha}$ and intended α was 0.9% for the exonic sites and 1.3% for genomic sites. The average absolute difference between the estimated and intended α was less than 0.5% for the exonic sites for $2\% \leq \alpha < 20\%$. For $\alpha \geq 20\%$, both sets underestimated α on average, but the exonic sites were more accurate. Furthermore, the standard deviation from the genomic estimates was >1.6x compared to exomic estimates under all experimental settings. Much of the difference can be explained by the smaller number of reads present at the off-target sites.

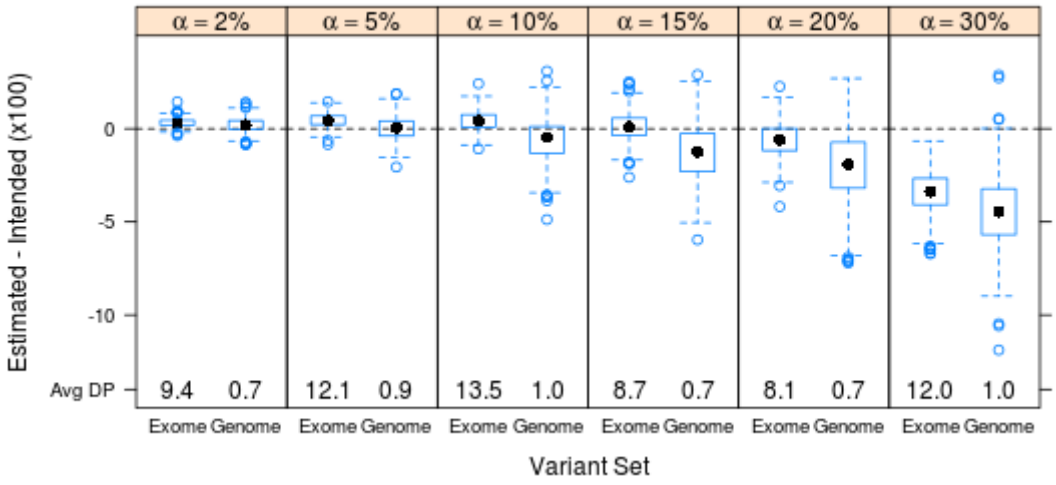

Figure 4-1- ***Estimating Contamination Using Exonic vs Genomic Sites****. We plot the distribution of* $(\hat{\alpha} - \alpha) \times 100$ *from our experimentally constructed contaminated samples. We compare estimates using sites exclusively from the exome to sites from across the genome. "Avg DP" is the average depth (# reads) per variant site.*

Excluding sites inside exons of genes with the highest expression variation among samples did not meaningfully change our contamination estimates. We believed that by dropping the most variably expressed genes, we could reduce some of the noise in the estimation and obtain more precise estimates however we did not observe a practical difference. To test this we used the data provided by the GEUVADIS project which estimated the reads per kilobase of transcript per million mapped reads (RPKM). We then calculated the coefficient of variation (standard deviation/mean) for the RPKM values for each site across all samples and removed the sites from those genes in the >90 percentile and those in the >50 percentile. The total number of sites used for estimation was ~100,000 for all exonic sites compared to ~95,000 and ~80,000 for the <90 and <50 percentile sites respectively. We observed that estimates that avoided the most variable genes were highly concordant with the

original estimates (Figure 2). The Pearson correlation of estimates from all exome sites was >0.999 for both the <50 and <90 percentile sites. We concluded that individual expression differences was not interfering with our ability to estimate contamination in a meaningful way.

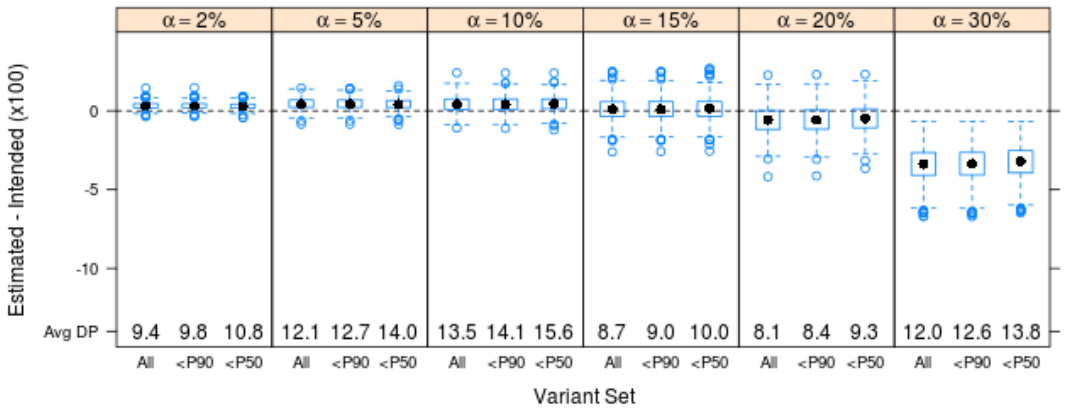

*Figure 4-2- **Estimates of Contamination Ignoring Most Variable Genes** We plot the distribution of $(\hat{\alpha} - \alpha) \times 100$ from our experimentally constructed contaminated samples. We compare estimates using all genes, excluding the >90% most variable genes, and excluding the >50% most variable genes. "Avg DP" is the average depth (# reads) per variant site.*

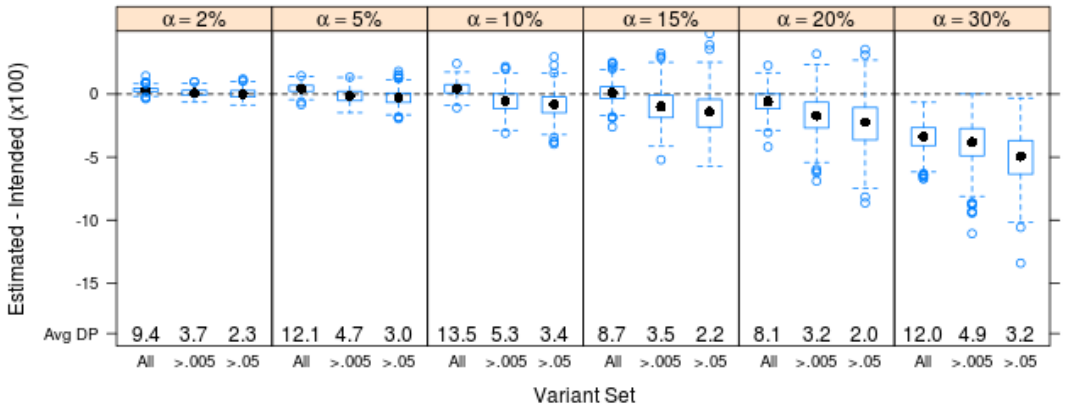

*Figure 4-3 - **Contamination Estimation Dropping Sites with Evidence for ASE**. We plot the distribution of $(\hat{\alpha} - \alpha) \times 100$ from our experimentally constructed contaminated samples. We compare estimates using all sites, sites with no significant evidence of ASE at the p<.005 level, and sites with no significant evidence of ASE at the p<.05 level. . "Avg DP" is the average depth (# reads) per variant site.*

To investigate the impact of allele specific expression estimates, we excluded sites that showed evidence of ASE in any of the samples. We used the estimates of ASE provided by the GEUVADIS project and collected lists of sites where at least one individual had evidence for ASE at either the p=.005 or p=.05 level, which left ~77,500 and ~67,000 sites per individual compared to ~100,000 using all exonic sites. Estimating contamination using sites with no evidence for ASE leads to deflation of the contamination estimate and increased variance >1.4x for all levels of contamination (Figure 3). The difference is primarily driven by the reduction in summed read depth for all included sites; when we exclude sites with significant evidence for ASE, we exclude sites with greater read depth. More reads result in an increased power to detect expression differences and can generate more significant p-values.

When we applied our method to data from the psoriasis sequencing project, we identified a set of samples that had been incorrectly labeled during sequencing. We found 21 samples with estimated levels of contamination $\hat{\alpha}$>90% when using the existing genotype array data (Figure 4). By comparing the read data to the original genotype data, it was possible to correct the labels for 19 of the 21 samples. An additional 6 samples showed contamination between 20% and 90%. We recommend that these samples should be dropped from analyses that intend to make inference at the individual level since reads cannot be easily assigned to a particular sample.

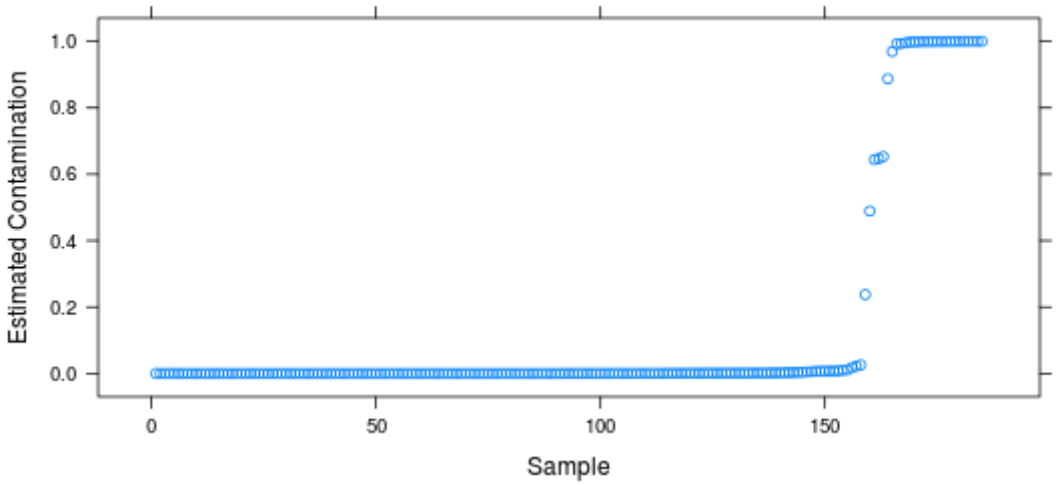

Figure 4-4 - ***Contamination Estimates from Psoraisis Data****. Each point is the value of $\hat{\alpha}$ for a sample using available genotype array data, sorted along the x-axis by $\hat{\alpha}$.*

Discussion

We have shown that modeling contamination in RNA-Seq data the same way that we do with DNA sequencing data can work very well if we focus on variant sites in the exome. On average the difference between $\hat{\alpha}$ and the true α was less than 0.5% for contamination levels up to 10%. We have also shown that this method can be combined with genotype array data to identify mislabeled samples.

In our analyses, we assumed the true sample population allele frequencies for each of the variant sites were known; however, this information may not always be available if you are studying a population that has not previously been characterized. Probably the best alternative would be to use frequency estimates calculated from genotype array data for the same or similar samples. Alternatively, the allele frequencies can be estimated using the proportion of reads that carry each allele across multiple sequenced samples. Finally, one could use allele frequencies from a closely related HapMap or 1000 Genomes Project population. It is important

that the frequencies accurately represent your samples. As with DNA data, misspecification of these frequencies often leads to underestimates of α^3.

Our model assumes that all sites are independent; this is not generally true for real data. For example, some nearby sites are correlated due to linkage disequilibrium (LD). The data for the experiments above used all sites in the exome from the HumanOmni2.5 genotypes array. We also did the experiment after pruning the sites based on LD estimates from the genotype array data. We pruned sites such that no pair in a 50 site-wide, 5 site-sliding window had an r^2 value greater than 0.2, leaving ~40,000 of ~100,000 sites. The correlation between the $\hat{\alpha}$ values for the full exome list and the LD-pruned exome list was $r^2 > 0.996$, demonstrating the impact on estimation from correlated sites was modest.

Alternative gene splicing can make it hard to align reads to a reference, especially at splice junctions at the ends of exons. Splicing results in different sets of exons being merged together into the final RNA message before translation into a protein. This means that sequencing reads at the beginning or ends of exons are more difficult to correctly align because of the uncertainty of the surrounding sequence. We tested if this may have any significant impact on contamination estimation by excluding variant sites located with 10 base pairs of splice sites. This resulted in site list with ~6,000 fewer variants. This filtering produced nearly identical estimates of contamination (r^2>0.999) across all simulation settings.

Ultimately we decided not to directly model expression variability or ASE because our estimates ignoring the phenomena were accurate and the experiments where we dropped potentially troublesome sites did not appear to improve the precision or reduce the variability

of the estimates. The power of this mixture model is that information is combined across a large number of sites so it is difficult for local deviations from the model expectations to overpower the contamination signal. Modeling ASE would most easily be done in a Bayesian framework where we could put a prior distribution on the probability of observing alleles at heterozygous sites rather than assuming it is always 0.5. However, our results suggest the simpler model presented here is entirely sufficient for its purpose.

In summary, we have demonstrated the usefulness of this mixture-model-based method to easily and quickly detect contamination in RNA-Seq data after reads have been aligned. We suggest screening all RNA-Seq data to check for any potential quality problems. Samples may be excluded from a study depending on the level of tolerance for measurement error of the downstream analysis.

Chapter 4 References

1) Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.
2) Engström, P., Steijger, T., Sipos, B. , Grant, G.R., Kahles, A. ,The RGASP Consortium, Rätsch, G., Goldman, N., Hubbard,T.J., Harrow, J., Guigó, R., Bertone, P. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nature Methods *10*, 1185–1191.
3) Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet *91*, 839–848.
4) Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. *11*, 733-739.
5) Do, R., Kathiresan, S., Abecasis, G.R. (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet *21*, R1-R9.
6) Lee, S., Abecasis, G.R, Boehnke, M, Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet *95*, 5-23.
7) Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., Kinzler, K.W., et al. (2002) Allelic variation in human gene expression. Science *297*, 1143-1143.
8) Ewing, B., Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. Genome Res 8, 175-185
9) Brent, R.P. (1973). Algorithms for Minimization without Derivatives. (New Jersey: NJ: Prentice-Hall).
10) Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.
11) The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
12) Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.
13) Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., Kang, H. M. (2015). Correcting for sample contamination in genotype calling of DNA sequence data. Am J Hum Genet *97*, 284-290.
14) Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadizza, A., et. al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.

Chapter 5
Summary

There is no more important lesson that I have learned in my training as a biostatistician than this: real data are messy. Messiness comes in many forms: missing observations, surprising outliers, unusual patterns, etc. This messiness often interferes with the direct application and interpretations of standard statistical methods because of possible assumption violations. We can choose to view messy data as something that interferes with "real" statistics, or we can embrace the irregularities of experimental data and use that to motivate a deeper statistical investigation into of realities of the data. Messy data should make a statistician excited, rather than deterred – it simply means there are lessons to be learned. Those lessons might be that the way of collecting data has errors or can be improved, or that the data has properties we do not yet understand.

This work on contamination was not motivated by a scientific curiosity to intentionally mix samples together; rather it was in response to unusual patterns in our observed data and a strong desire to understand the causes. Contamination was just a hypothesis we had about what might cause those patterns. We then thought about new ways to look at the data and looked for simple statistical models we might be able to use to test this hypothesis. In the end, we developed a new set of tools that made it possible to detect and quantify contamination for a wide variety of genetic data.

In the case of detecting contamination in genotype array data, we were able to use data already generated by the genotyping instrument in a novel way to test this new hypothesis. Our methods look at the probe intensity data which is normally just used to call genotypes. By combining this data with population allele frequencies in a regression model or using this data in multivariate normal mixture model, we can test for contamination without having to run a separate experiment. Our method runs quickly enough to be able to screen large numbers of samples efficiently and can integrate nicely into a standard quality control pipeline for genotype array data.

We learned that contamination usually does not have a large impact on genotype concordance for genotype array data. We compared the concordance for samples in Chapter 2 that were contaminated *in-vitro* at known mixture proportions against their uncontaminated counterparts. Even at 10% contamination, genotype calls were >.999 concordant (Figure 5-1). The only noticeable effect is an increase in the number of missing genotypes. Rather than make an incorrect genotype call, the Illumina software is much more likely to simply not make a call for that variant. Dropping samples with low call-rates has for a long time been a standard quality control filter prior. With the development of these methods to detect contamination, we now can offer a possible answer to why the sample had a low call rate which was not always clear previously.

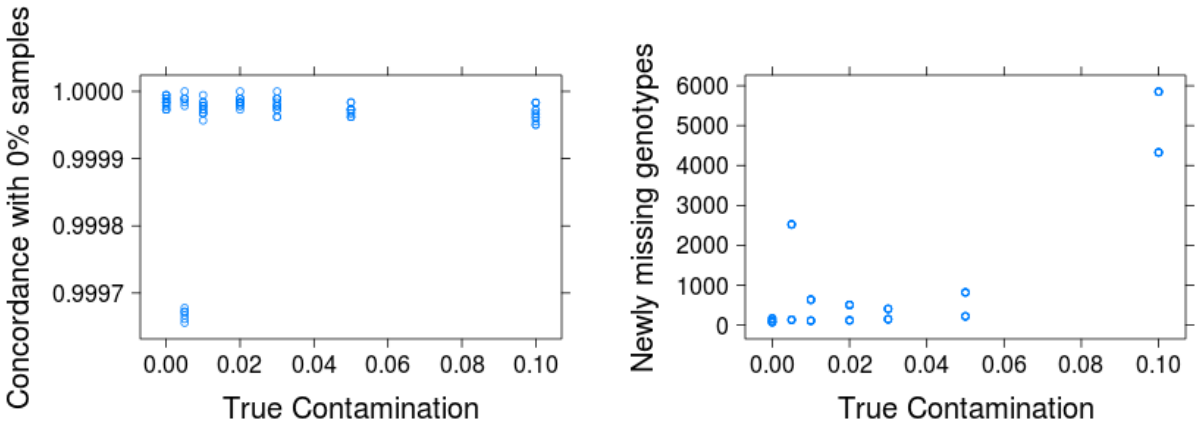

*Figure 5-1 - **Effects of Contamination on Genotype Array Calls** - The left panel shows genotype concordance for each of the contaminated mixtures with the uncontaminated sample. The right panel shows the increase in the number of missing genotypes for each contaminated mixture compared to the uncontaminated sample.*

Even though contamination does not have a large effect on genotype accuracy for array data, the shifts in intensity from contaminated data are still capable of detecting low-levels contamination (2-10%). Since genotype arrays are so much cheaper than DNA sequencing, running data on a genotype array prior to sequencing can be a cost effective way to screen samples prior to sequencing. We saw in Chapter 3 that low levels of contamination have a much greater impact on genotype concordance for sequencing data compared to array data. There is some uncertainty when a sample is identified as contaminated as to when the contamination occurred. It could have happened during preparation for genotyping or it could have happened at the time of collection. If samples are repeatedly tested, it can be possible to determine if the original sample is bad or if something went wrong during genotyping depending on the consistency of the contamination estimate.

While genotype arrays were the go-to source of genetic data for association studies when I first started my studies of statistical genetics, we have clearly started to transition to the

age of sequencing. We are now able to investigate nearly every position in the genome rather than a few select sites. The technology driving this revolution continues to get cheaper, faster, and more accurate; however the problems of contamination have also made their way to this technology.

One additional challenge of working with DNA sequencing data is batch effects. Sequencers may produce systematic differences in the read data for samples depending on when and where they are processed. Ideally, all samples would be sequenced at the same place and time with cases mixed with controls; however this is not always possible. We were reminded of this problem when we created contaminated samples for Chapter 3. When we combined the exome sequencing data and estimated contamination, we observed an odd bimodal distribution of estimates (Figure 5-2A). Further investigation revealed that the samples were sequenced at different sequencing centers. If we stratified our analysis by the sequencing center for each of the samples in the contaminated pair, we can see that the distribution of estimates is much closer to expectation (Figure 5-2B).

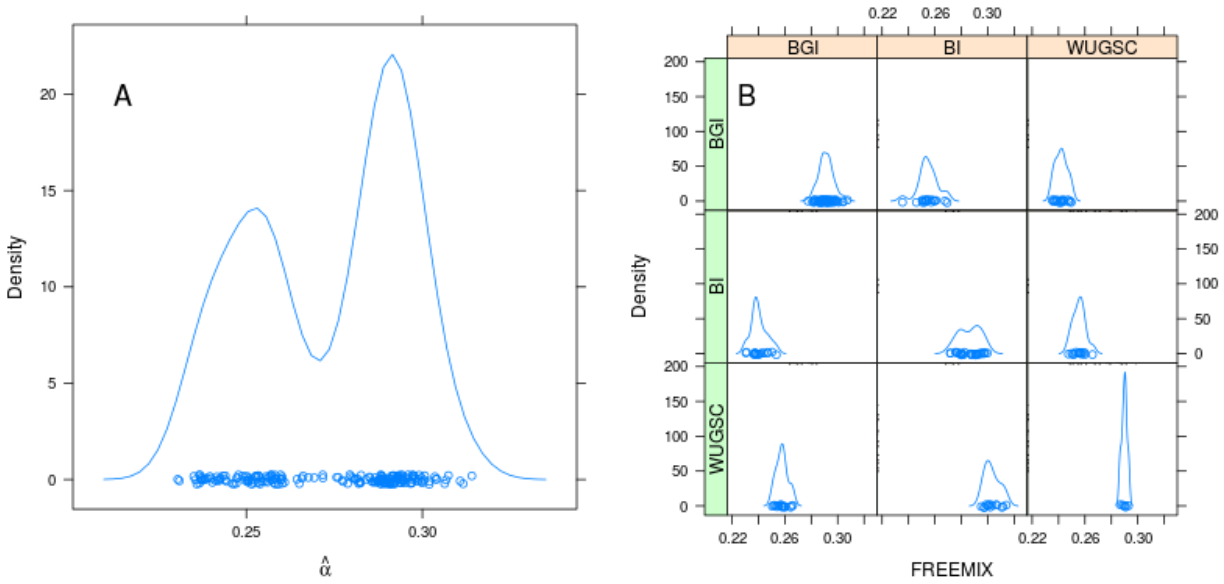

Figure 5-2 - ***Contamination Estimates from Different Sequencing Centers -*** *(A) the distribution of $\hat{\alpha}$ for all exome sequencing samples contaminated at 30%. (B) the distribution of $\hat{\alpha}$ conditioned on the sequencing centers for each of the samples in the contaminating pair.*

It is very useful to view the estimates of contamination in a context that is aware of the experimental conditions under which the samples were created. For example, most of our samples are processed in set of 96-well plates. If we plot our estimate of contamination in a way that reflect the plate and position where a sample came from, we can learn about how contamination may have occurred. An example of this type of plot is given in Figure 5-3. We can see that problems might only affect certain plates and positions rather than indicate a more wide spread problem.

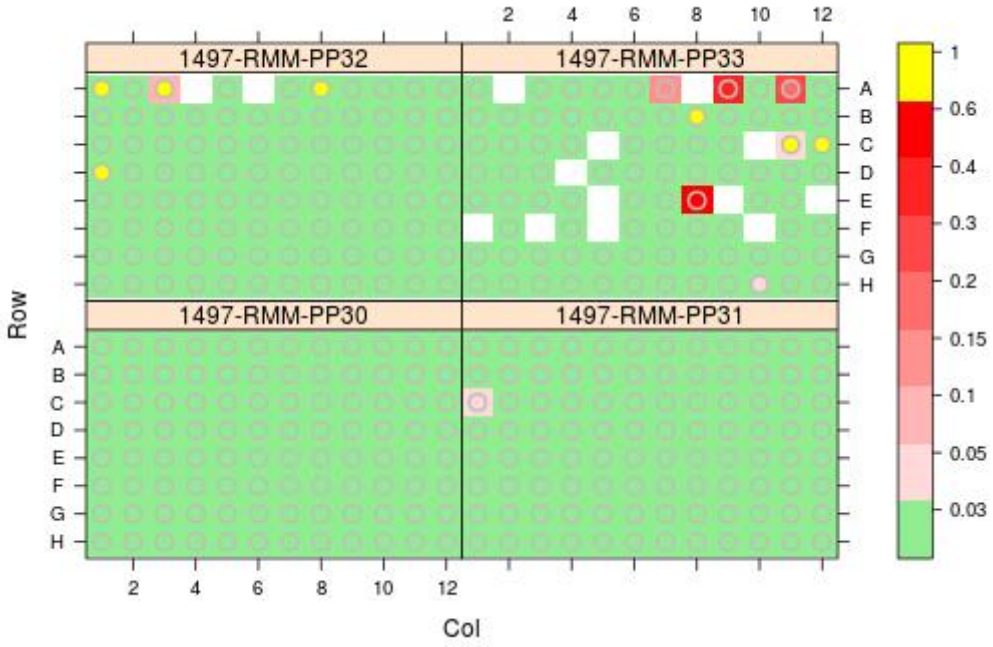

Figure 5-3 – ***Contamination Estimates in 96-Well Plate*** *- The color of each square represents the genotype-free estimate of contamination and the color inside the circle represents the estimate using genotype data. Sample swaps are indicated in yellow.*

Our test for DNA contamination can happen just after the sequence reads are aligned. Previously the effects of contamination typically were not seen until sample genotypes were called. Contaminated samples would have many more heterozygous sites than expected. Genotype calling can be a time consuming process and is normally delayed until a large number of samples are collected. Our method allows for much earlier detection of contamination, possibly allowing time for changes in a protocol or pipeline to correct any errors before all samples are sequenced. To be even more responsive, these methods could be adapted to collect observed bases from fastQ files prior to alignment. If you know the flanking sequencing

around variant sites, you can find relevant reads for those sites without aligning every read. This will save further time because alignment is also a time-consuming process.

While we have worked out a clear method for correcting contaminated DNA samples, the method of correction for RNA-Seq data is not so clear. Currently there is no standard analysis of RNA-Seq data similar to the way DNA sequencing is used for genotyping. Further work is required to understand the impact of contamination on the various RNA pipelines and different corrections may be appropriate for different analyses.

As the manufactures of sequencing machines continue to innovate and new assays are developed, the data we receive from sequencing machines may change. The methods described here have been tested and validated with current generation sequencing machines. It will surely be necessary to adapt these methods in the future for different instruments and protocols because while technology may improve, it is likely that contamination will continue to be a potential problem for genetics studies.

Up until now we have focused on contamination as an important quality control step, however there are natural biological phenomena where some form of "contamination" is expected. Two such examples are genetic chimerism and cell-free fetal DNA. In the case of genetic chimerism, an embryo develops with two or more distinct cell lines when multiple fertilized eggs merge. This means that two cells from the same individual may have distinct DNA sequences. Since both fertilized eggs presumably arose from the same set of parents, there would be a high degree of similarity between the sequences. If we estimated contamination in different windows across the genome, we expect to find $\hat{\alpha} > 0$ for regions where different

cross-over events occurred during meiosis. Cell-free fetal DNA is the phenomenon where DNA from a fetus can be found in the blood of his pregnant mother. This may also look like contamination since the DNA sequences of two individuals would be sequenced as one. But as with chimerism, we expect a high degree of similarity given that the samples are genetically related.

So whether our interests in mixtures of DNA are motivated by genetic hypothesis of natural phenomena or a test for quality from a complicated laboratory protocol, the methods we have outlined here will help with the understanding and interpretation of high-throughput genetic data for years to come.