Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data

Goo Jun,[1,3] Matthew Flickinger,[1,3] Kurt N. Hetrick,[2] Jane M. Romm,[2] Kimberly F. Doheny,[2] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Hyun Min Kang[1,*]

DNA sample contamination is a serious problem in DNA sequencing studies and may result in systematic genotype misclassification and false positive associations. Although 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.

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. Although they are 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 nonreference allele homozygous genotypes (HET/HOM ratio) (see Figures 1A and 1B available online). 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.

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 prescreened 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).

Material 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 (1) sequence data and array-based genotype data, (2) sequence data alone, and (3) array-based genotype data alone. We also present a regression-based method that uses array-based genotype data alone. For each of these

[1]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA; [2]Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD 21224, USA
[3]These authors contributed equally to this work
*Correspondence: hmkang@umich.edu
http://dx.doi.org/10.1016/j.ajhg.2012.09.004.

Table 1. Conditional Probability $P(b_{ij}|e_{ij}\, g_i)$ of Read b_{ij} Given True Genotype g_i, and Read Error e_{ij}

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

[a]AA, AB, BB: A allele homozygote, heterozygote, and B allele homozygote
[b]E: alleles other than A or B; assumes four possible alleles (bases)

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 by 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 i, let g_i be the true genotype, b_{ij} $(1 \le j \le R_i)$ be the base call for the jth 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 1). The conditional likelihood of a single overlapping base can then be written as the two-sample mixture model

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

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 that 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.

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

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

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 $2 \le k \le N$. 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 by 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

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

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, because $\mathcal{L}(\alpha) = \mathcal{L}(1-\alpha)$ here we restrict attention to $0 \le \alpha \le 1/2$.

Detecting Sample Contamination by 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 that 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, and 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, 1/2, or 1, for genotypes AA, AB, and BB, respectively. We describe two types of contamination detection, and estimation methods in this setting: two

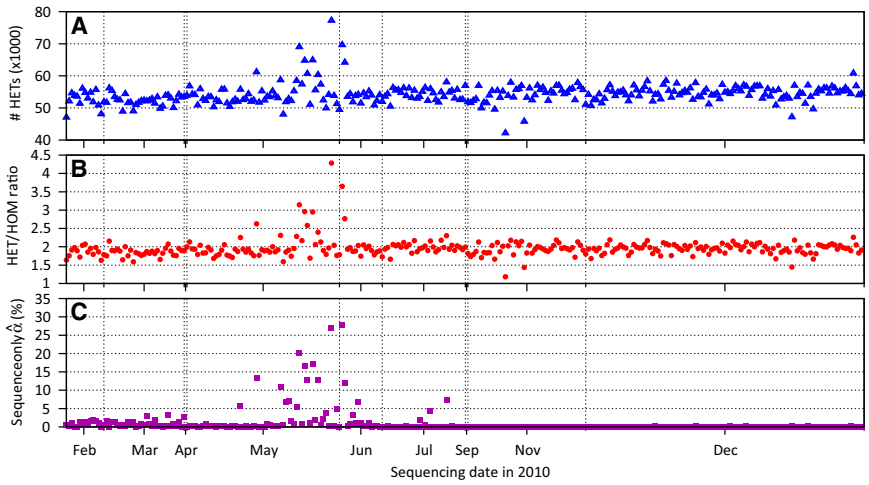

Figure 1. SNP Genotype Calling and Estimation of Contamination from 299 European Sequenced Samples Across chromosome 20
(A) Numbers of heterozygous genotypes.
(B) Ratio of the numbers of nonreference homozygous genotypes to heterozygous genotypes (HET/HOM ratio).
(C) Estimated level of DNA sample contamination estimated from sequence data only.

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

Detecting Sample Contamination by 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 multisample 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:

$$p_i(\mathbf{x}_i \mid g_i) \sim \mathcal{N}\left(\boldsymbol{\mu}_i^{g_i}, \sum\nolimits_i^{g_i}\right), \quad g_i = \{AA, AB, BB\}, 1 \le i \le M$$

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 $\sum_i^{g_i}$ is the covariance matrix of the A and B allele intensities. We estimate $\boldsymbol{\mu}_i^{g_i}$ and $\sum_i^{g_i}$ 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 that the observed DNA sample is a mixture of two unrelated DNA samples, we can model the intensity values as a bivariate Gaussian mixture:

$$p_i(\mathbf{x}_i \mid g_i^1, g_i^2; \alpha) \sim \mathcal{N}\Big(\alpha\boldsymbol{\mu}_i^{g_i^1} + \alpha\boldsymbol{\mu}_i^{g_i^2}, \alpha^2\sum\nolimits_i^{g_i^1} + (1-\alpha)^2\sum\nolimits_i^{g_i^2}\Big) \quad 1 \le i \le M$$

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

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

Genotype probabilities $P(g_i^k)$ 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, 1/2] followed by Brent's algorithm. With genotype array data alone, we cannot detect sample swaps.

The single-sample implementation is analogous to the multisample implementation. In the multisample implementation, the bivariate Gaussian parameters for p_i at each marker are estimated across all N samples, whereas 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

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

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 multisample 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 by 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). In the case of no contamination, we expect BAF values close to 0, 1/2, 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

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

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

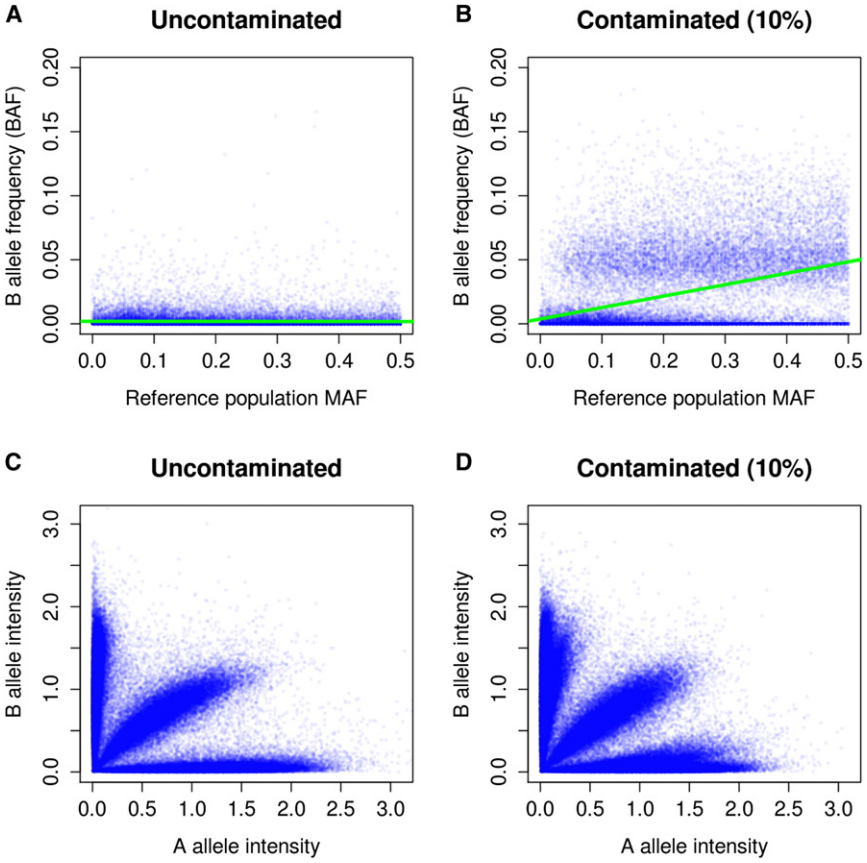

Figure 2. Distribution of Array Intensity for Contaminated and Uncontaminated Samples

BAF versus population MAF for (A) uncontaminated ($\alpha = 0$) and (B) contaminated ($\alpha = 10\%$) samples. Normalized intensity plots for (C) uncontaminated ($\alpha = 0$) and (D) contaminated ($\alpha = 10\%$) samples.

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 $\alpha > 0$.

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

$$\Pr(B \text{ is minor allele} \mid g = AA; f) = \frac{\Pr(B \text{ is minor allele},\ g = AA; f)}{\Pr(B \text{ is minor allele},\ g = AA; f) + \Pr(A \text{ is minor allele},\ g = AA; f)} = \frac{(1-f)^2}{(1-f)^2 + f^2}$$

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: (1) trimming overlapping ends of reads generated from the same template before analysis, (2) 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 (3) 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 ~4× coverage on an Illumina platform as part of the 1000 Genomes Project. We retained reads from the targeted sample with probability 1-α, 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

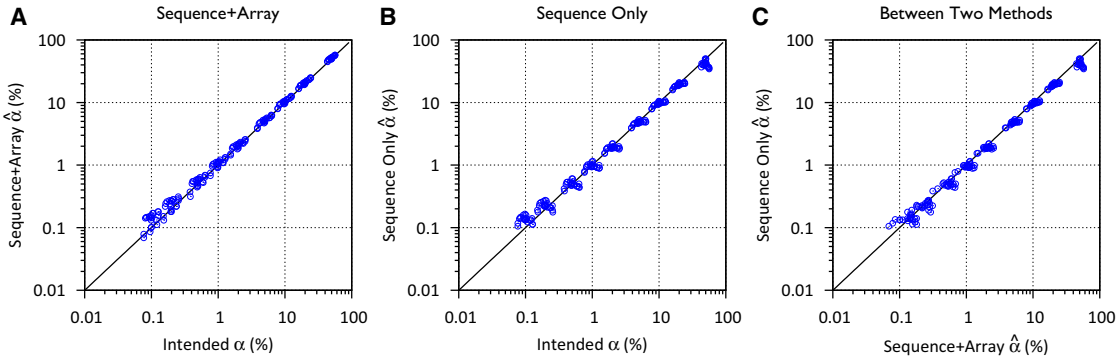

Figure 3. Estimated Contamination Levels for In Silico Contaminated Samples
(A) Joint sequence and array-based method, (B) sequence-only method, and (C) between these two methods.

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 $\alpha = 0$ 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 cardiometabolic 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 ~4× average coverage in a study of type 2 diabetes. One hundred and fifty samples were sequenced before a change in the sample handling process in August 2010; the remaining 149 samples were sequenced after the change. Among 299 sequenced samples, 227 were also 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 nonreference 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 and proper informed consent was obtained.

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 $r = 0.9996$ for the joint method and $r = 0.9840$ 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 3). In contrast, the relative error of the estimated contamination was 0.380 ± 0.257 (mean ± SD) for the joint method and 0.390 ± 0.241 for the sequence-only method when $\alpha \approx 0.1\%$, but it was reduced to 0.044 ± 0.035 and 0.056 ± 0.055 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%.

We evaluated the impact of estimated population allele frequencies on accuracy of contamination estimates (Figure S1). Compared to the original sequence-only estimates of $\hat{\alpha}$ that used CEU allele frequencies, using allele frequencies from the GBR samples resulted in reduced estimates of $\hat{\alpha}$ (mean ratio ± SD for $\hat{\alpha}_{GBR}/\hat{\alpha}_{CEU} = 0.884 \pm 0.083$). Allele frequencies from the more distantly related FIN samples resulted in further reduced contamination estimates (mean ratio ± SD for $\hat{\alpha}_{FIN}/\hat{\alpha}_{CEU} = 0.804 \pm 0.135$). Allele frequencies from the broader European (EUR) continental population (CEU, GBR, FIN, IBS, and TSI) performed better (mean ratio ± SD for $\hat{\alpha}_{EUR}/\hat{\alpha}_{CEU} = 0.926 \pm 0.054$), whereas allele frequencies from the very different African (AFR) samples (YRI, LWK, and ASW) resulted in severe reduction in contamination estimates (mean ratio ± SD for $\hat{\alpha}_{AFR}/\hat{\alpha}_{CEU} = 0.160 \pm 0.121$).

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 (Figure S2); for example, when $\alpha = 1\%$, the mean relative errors

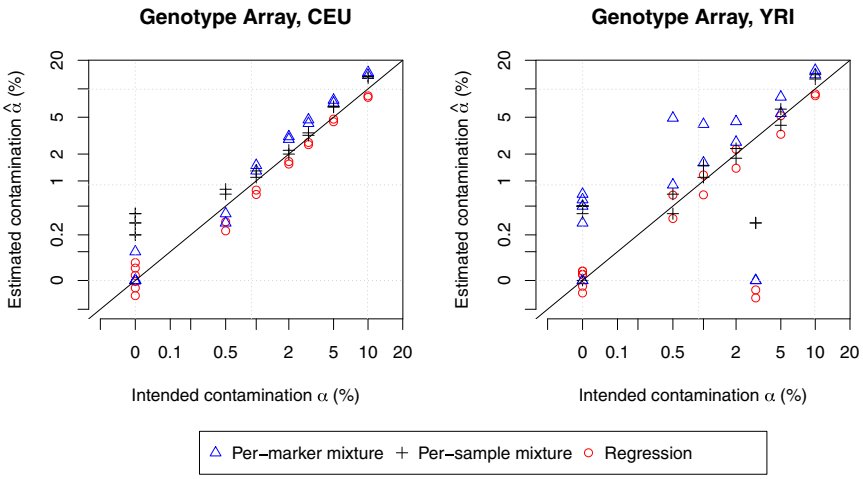

Figure 4. Estimated Versus Intended Contamination Levels from the Experimentally Contaminated Array Intensity Data

Three methods—regression-based method, multisample mixture model method, and single-sample mixture model method—were compared in two populations (CEU and YRI).

$|\hat{\alpha} - \alpha|/\alpha$ for the joint method, were 0.414, 0.135, 0.103, and 0.099 for 1,000, 10,000, 100,000, and all 1.2M sites, and 0.112 when using the 30,471 chromosome 20 sites. Because 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 $\alpha > 1\%$; when $\alpha < 1\%$ ContEst tended to overestimate contamination levels to a larger degree than ours (Figure 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 multisample mixture model methods produced contamination level estimates that matched our constructs, except for two YRI samples with 3% intended contamination (Figure 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, whereas the regression-based method tends to underestimate α.

Using the mixture-model methods, 0 of the 6 uncontaminated CEU samples were identified as contaminated, whereas 3 of 6 uncontaminated YRI samples were identified as slightly ($0 < \hat{\alpha} < 1\%$) contaminated. We suspect that 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 $0.05/36 = 0.0013$; the results correctly identified the contamination state of 34 of the 36 experimental samples except for the two YRI samples with intended $\alpha = 3\%$. Given our consistent results across our three different methods, we suspect that 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 (Figures 1A and 1B). We applied our sequence-based and sequence-only methods to these samples. Because HumanOmni2.5 genotype data were available on only 227 of these 299 individuals, we display results for the sequence-only method (Figure 1C); contamination level estimates for the sequence and array data jointly were very similar, particularly for individuals with higher contamination levels (Figure 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).

Comparison of results (Figure 1; Table 2; Figure S3) suggests that our contamination estimates were more

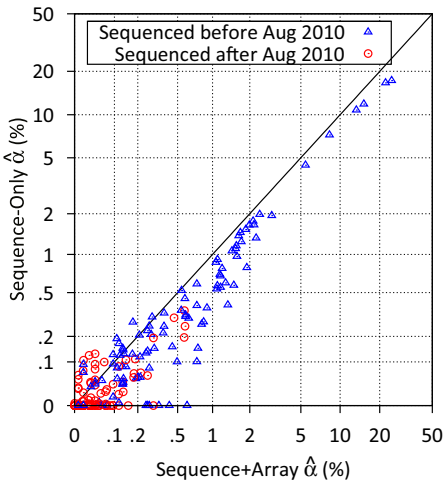

Figure 5. Estimated Contamination Levels between Sequence-Based Methods

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

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 1C).

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 1; Figure 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.

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 (4×) 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 3). Both of 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 nonreference homozygous genotypes (HET/HOM ratio)—both of which can only detect contamination rates of >5%–10% (Figure 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 (4×) 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 hr 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 (Figure S2).

Although 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

Table 2. Summary of Estimated Contamination Levels $\hat{\alpha}$ Ratio of the Numbers of Heterozygous to Nonreference Allele Homozygous Genotypes, and Genotype Discordance with Array Data for 299 Samples from Type 2 Diabetes Study Using Sequence Data Only

		$\hat{\alpha}$ (sequence only)			
Array Genotypes?	**Measure**	**<1%**	**1%–2%**	**2%–5%**	**≥5%**
Yes (n = 227)	Number of samples	208	13	1	5
	– Before August 2010	81	13	1	5
	– After August 2010	127	0	0	0
	RR discordance[a]	0.0021	0.0030	0.0071	0.0492
	RA discordance[b]	0.0154	0.0157	0.0172	0.0300
	AA discordance[c]	0.0085	0.0143	0.0377	0.176
	HET/HOM ratio[d]	1.92	1.84	2.16	2.66
No (n = 72)	Number of samples	46	8	7	11
	– Before August 2010	24	8	7	11
	– After August 2010	22	0	0	0
	HET/HOM ratio[d]	1.87	1.88	1.88	2.64

[a]RR discordance: Genotype discordance when array-based genotype is homozygous reference
[b]RA discordance: Genotype discordance when array-based genotype is heterozygous
[c]AA discordance: Genotype discordance when array-based genotype in homozygous nonreference
[d]HET/HOM ratio: Ratio of number of heterozygous genotypes to homozygous nonreference genotypes

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 nonzero 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 ${\chi_0}^2$ and ${\chi_1}^2$ 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 $\hat{\alpha} > 1\%$, power to detect contamination using the regression method based on 1,000 random subsets of 50, 100, 500, and 1,000 homozygous SNPs was 37.3%, 59.6%, 99.0%, and 100%,

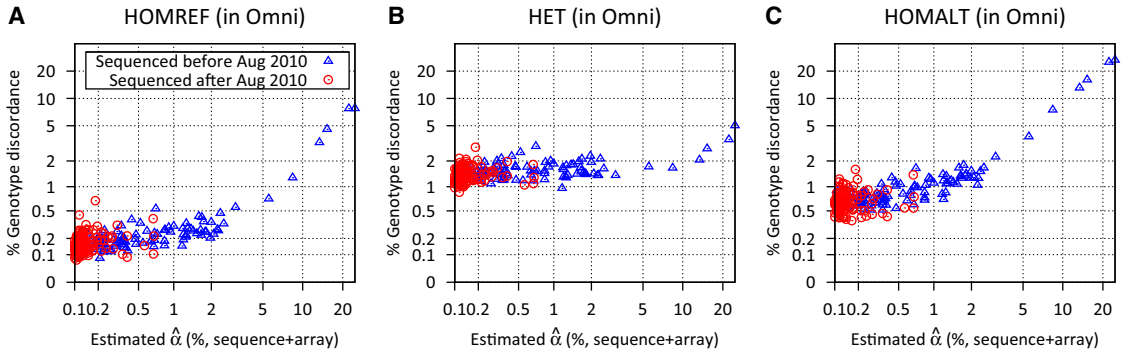

Figure 6. Genotype Discordance between Sequence-Based and Array-Based Genotypes
A function of estimated contamination level $\hat{\alpha}$ in the type 2 diabetes sequencing study; contamination level estimates based on the combined sequence and genotype array data, stratified by genotypes from HumanOmni2.5 array data.
(A) Homozygous reference genotypes, (B) heterozygous genotypes, and (C) homozygous nonreference genotypes.

respectively (Table 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%, and 8%–14% relative increases in the estimated contamination levels compared to actual contamination (Table 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 $(1\text{-}f)\alpha$, 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 thousands 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 prescreening 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 $\widehat{\alpha} > 2\%$.[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.

Supplemental Data

Supplemental Data includes four figures and two tables, and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

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.

Received: May 6, 2012
Revised: August 10, 2012
Accepted: September 7, 2012
Published online: October 25, 2012

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

References

1. Schmieder, R.A.E., and Edwards, R. (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., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601–2602.
3. Brent, R.P. (2002). Algorithms for Minimization without Derivatives (New York: Dover Publications).
4. Gordon, D., Yang, Y., Haynes, C., Finch, S.J., Mendell, N.R., Brown, A.M., and Haroutunian, V. (2004). Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Stat. Appl. Genet. Mol. Biol. *3*, e26.
5. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (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, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. *21*, 940–951.
7. Browning, B.L., and 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, S.G., and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. *82*, 605–610.
9. Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., and Holmes, C.C. (2008). GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics *24*, 2209–2214.
10. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (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, S.F., Hakonarson, H., and 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.