Method

Ancestry-agnostic estimation of DNA sample contamination from sequence reads

Fan Zhang,[1,2] Matthew Flickinger,[1,3] Sarah A. Gagliano Taliun,[1,3] InPSYght Psychiatric Genetics Consortium, Gonçalo R. Abecasis,[1,3] Laura J. Scott,[1,3] Steven A. McCaroll,[4,5] Carlos N. Pato,[6] Michael Boehnke,[1,3] and Hyun Min Kang[1,3]

[1]Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA; [2]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan 48109-2218, USA; [3]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan 48109-2029, USA; [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [6]SUNY Downstate Medical Center, Brooklyn, New York 11203, USA

Detecting and estimating DNA sample contamination are important steps to ensure high-quality genotype calls and reliable downstream analysis. Existing methods rely on population allele frequency information for accurate estimation of contamination rates. Correctly specifying population allele frequencies for each individual in early stage of sequence analysis is impractical or even impossible for large-scale sequencing centers that simultaneously process samples from multiple studies across diverse populations. On the other hand, incorrectly specified allele frequencies may result in substantial bias in estimated contamination rates. For example, we observed that existing methods often fail to identify 10% contaminated samples at a typical 3% contamination exclusion threshold when genetic ancestry is misspecified. Such an incomplete screening of contaminated samples substantially inflates the estimated rate of genotyping errors even in deeply sequenced genomes and exomes. We propose a robust statistical method that accurately estimates DNA contamination and is agnostic to genetic ancestry of the intended or contaminating sample. Our method integrates the estimation of genetic ancestry and DNA contamination in a unified likelihood framework by leveraging individual-specific allele frequencies projected from reference genotypes onto principal component coordinates. Our method can also be used for estimating genetic ancestries, similar to LASER or *TRACE*, but simultaneously accounting for potential contamination. We demonstrate that our method robustly estimates contamination rates and genetic ancestries across populations and contamination scenarios. We further demonstrate that, in the presence of contamination, genetic ancestry inference can be substantially biased with existing methods that ignore contamination, while our method corrects for such biases.

[Supplemental material is available for this article.]

Sample contamination is a common problem in DNA sequencing studies. Contamination may occur during sample shipment (due to spillage across wells, pipetting errors, or insufficient dry ice), library preparation (due to gel cut-through in fragment size selection or unexpected switch between barcoded adaptors in vitro), in silico demultiplexing from a sequenced lane into barcoded samples, or on many other unexpected occasions. Even modest levels of contamination (e.g., 2%–5%) within a species substantially increase genotyping error, even for deeply sequenced genomes (Flickinger et al. 2015). Accurate estimation of DNA contamination rates allows us to identify and exclude contaminated samples from downstream analysis, and genotypes of moderately contaminated samples (e.g., <10%) can be improved by accounting for contamination in genotype calling (Flickinger et al. 2015).

Previously, we developed methods and a software tool, *verifyBamID* (Jun et al. 2012), to estimate DNA contamination from sequence reads given known population allele frequencies of common variants. Many investigators and most major sequencing centers use *verifyBamID* as a part of their standard sequence processing pipeline. However, we have shown that *verifyBamID* can underestimate DNA contamination rates if the assumed population allele frequencies are inaccurate (Jun et al. 2012). Such an underestimation can be avoided if correct population allele frequencies are provided in ideal circumstances. However, in early stages of sequence analysis, performing a tailored customization of quality control (QC) steps for each sequenced genome based on their ancestry is not always feasible or is sometimes impossible. Such a tailored customization requires planned coordination between sequencing centers and study investigators prior to sequencing to share the self-reported ancestry (which is not always accurate) or estimated ancestry from external genotypes (which is not always available). Modifying the QC pipeline to accommodate study-specific or sample-specific parameters may not be an option for large sequencing centers. Even if such a tailored customization of the QC pipeline is possible, preparing per-sample ancestry prior to QC may delay time-sensitive issues in the sequencing procedure. If contamination rates can be accurately estimated without having to know the ancestry or allele frequencies a priori, this will simplify the sequence analysis pipeline and expedite the QC.

Corresponding author: hmkang@umich.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.246934.118.

Here, we describe a novel method to robustly detect and estimate DNA contamination by modeling the probability of observed sequence reads as a function of "individual-specific allele frequencies" that account for genetic ancestry of a sample. Instead of assuming that the population allele frequencies are known, we represent individual-specific allele frequencies as a function of genetic ancestry using principal component coordinates and the reference genotypes from a diverse population—for example, the Human Genome Diversity Project (HGDP) (Cavalli-Sforza 2005) or 1000 Genomes (The 1000 Genomes Project Consortium 2015). We then jointly estimate genetic ancestry and contamination rates of a sequenced individual based on a mixture model, without requiring the assumption that population allele frequencies are known. As a result, our method enables robust estimation of DNA sample contamination without relying on externally provided genetic ancestry information. Instead, our method simultaneously estimates the genetic ancestry accurately from sequence reads through a unified likelihood framework.

Results

Our previous method (*verifyBamID*) can estimate sample contamination rate with external genotypes or with population allele frequencies only. Because both methods accurately estimate contamination rates, the latter approach, which only requires allele frequencies, has dominated its practical use (Fig. 1A). However, if allele frequencies are misspecified or unknown, the estimated contamination rates can be severely biased.

Our new method (*verifyBamID2*) avoids such a bias due to misspecified allele frequencies by modeling individual-specific allele frequencies as a function of genetic ancestry and by jointly estimating genetic ancestry and contamination rates to maximize the likelihood of sequence reads. The genetic ancestry can be represented as coordinates of principal components from a cosmopolitan reference panel, such as 1000 Genomes or HGDP (Fig. 1B). In addition, our new method can also be used for genetic ancestry estimation, similar to LASER (Wang et al. 2014, 2015) or *TRACE* (Wang et al. 2015), but accounting for potential sequence contamination together. We show that our method provides (1) comparable or more accurate estimates of genetic ancestry than existing methods such as *TRACE*/LASER even in the absence of contamination, and (2) reduced bias in contamination rate estimates compared to our previous method requiring known population allele frequencies using in silico–contaminated data sets and sequenced genomes from the InPSYght psychiatric genetics sequencing study (Sanders et al. 2017).

We assessed our new methods in the following steps. First, in the absence of contamination, we demonstrate that our estimation of genetic ancestry provides comparably accurate estimates of genetic ancestry as other state-of-the-art methods. Second, in the presence of contamination, we demonstrate that joint estimation of genetic ancestry and contamination substantially improves the estimation accuracy of both parameters. Third, using in silico–contaminated samples, we demonstrate that our methods robustly provide more accurate estimates than previous methods across various combinations of genetic ancestries and contamination rates. Fourth, from the analysis of deeply sequenced genomes in the InPSYght study, we demonstrate that our new methods deliver more accurate contamination estimates than the previous methods.

New model-based methods accurately estimate genetic ancestry

In the absence of contamination, widely used methods such as LASER and *TRACE* are known to estimate genetic ancestry accurately. Because we propose using a new model-based approach to estimate the genetic ancestry (jointly with contamination rates), we first compared the accuracy of our new method, in the absence of contamination, with LASER and *TRACE*. We randomly chose 500 ethnically diverse samples from the 1000 Genomes Project low-coverage (4×) genomes and 500 African-American samples from the deeply sequenced (32×) genomes from the InPSYght project. We estimated their genetic ancestries using 100,000 SNPs from the HGDP reference panel (see Methods for details) and compared their genetic ancestry estimates obtained by LASER (using the same sequence data) and *TRACE* (using the hard-call genotypes). As illustrated in Figure 2, A, C, and E, the estimated PC coordinates of the 1000 Genomes individuals are located close to

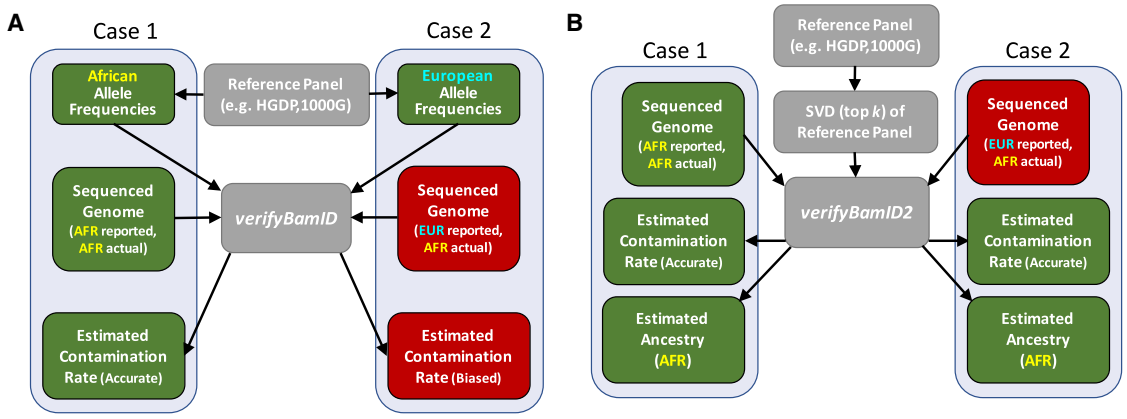

Figure 1. Overview of *verifyBamID* and *verifyBamID2* software tools. (*A*) *verifyBamID* takes aligned sequence reads (in BAM format) and known variant sites annotated with population allele frequencies (in VCF format) to estimate DNA contamination rates. When allele frequencies are correctly specified, the estimated DNA contamination rates are expected to be accurate (green boxes). However, when the allele frequencies are misspecified (e.g., due to incorrect self-reported ancestry), the estimates of DNA contamination rates may be biased (red boxes). (*B*) *verifyBamID2* takes aligned sequence reads (in BAM/CRAM format) and top *k* singular value decomposition (i.e., PCs and SNP loadings) to estimate the genetic ancestries and contamination rates together. Because *verifyBamID2* does not rely on self-reported ancestry, even if ancestry of sample is misspecified or unknown (red box), the estimated contamination rates will be unbiased (green box). In addition, genetic ancestries are also estimated in PC coordinates, adjusting for potential contamination.

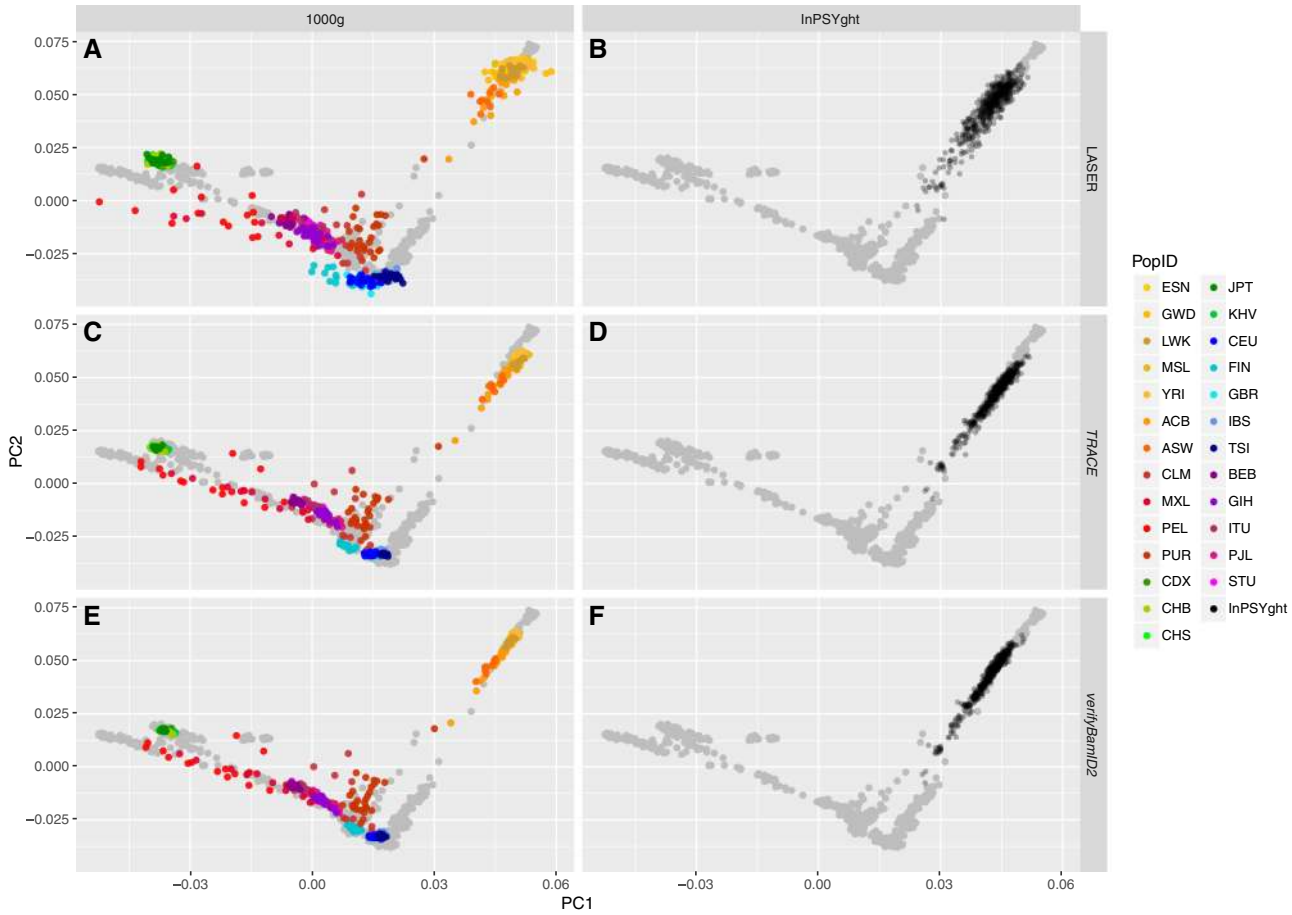

Figure 2. Evaluation of estimated genetic ancestry coordinates, in the absence of contamination, between *TRACE*, LASER, and *verifyBamID2* on samples from the 1000 Genomes low-coverage genome ($n = 500$, diverse ancestry) sequence data (*A*,*C*,*E*) and from the InPSYght deep genome ($n = 500$, African-Americans) sequence data (*B*,*D*,*F*). Panels *A* and *B* show results from *TRACE*, *C* and *D* from LASER, and *E* and *F* from *verifyBamID2* (assuming no contamination). Each point represents a sample and each color represents a population ancestry, with the exception that gray points represent PCA coordinates of reference (HGDP) samples.

their corresponding HGDP populations across all three methods. Compared to *TRACE* and LASER, we observed that the estimated genetic coordinates from *verifyBamID2* were the closest to the centroid of the corresponding HGDP population (Table 1) in four of the five populations (all except TSI). These results suggest that our method provides estimates at least as precise compared to those for other state-of-the-art methods.

Genetic ancestry estimates may be confounded by DNA contamination

Next, we constructed in silico–contaminated sequenced data from the 1000 Genomes Project and estimated contamination parameters and genetic ancestries jointly. We observed that, when sequences are contaminated between different continental populations, the genetic ancestry estimates in PC coordinates drift toward the contaminating population when contamination is ignored (Fig. 3A) or when assuming that intended and contaminating samples originated from the same population (Fig. 3C). As the contamination rate increases, drift increases (Fig. 3A,C,E).

However, when we accounted for possible differences in genetic ancestries between the two intended and contaminating samples using our new methods, PC coordinates remained similar to those for uncontaminated samples (Fig. 3E) and contaminated samples constructed from individuals that belong to the same population (Fig. 3B,D,F).

Table 1. Distance between estimated PCA coordinates of HGDP and 1000 Genomes populations

Population label		Distance between PCA coordinates ($\times 10^{-3}$)		
1000G	**HGDP**	***TRACE***	**LASER**	***verifyBamID2***
CHB	Han-NChina	1.89	3.01	**0.82**
CHS	Han	1.76	1.81	**1.25**
TSI	Tuscan	**1.62**	2.78	1.86
YRI	Yoruba	2.35	2.62	**0.59**
JPT	Japanese	1.66	1.99	**1.29**

Mean distances were measured between the mean PCA coordinates across the population in HGDP (estimated from the array data of Wang et al. [2015]) and the PCA coordinates estimated from each of the 1000 Genomes low-coverage sequence data of the corresponding population, projected onto the same PCA coordinates using *TRACE,* LASER, or *verifyBamID2* (assuming no contamination). Boldface represents the smallest distance among the three methods for each population.

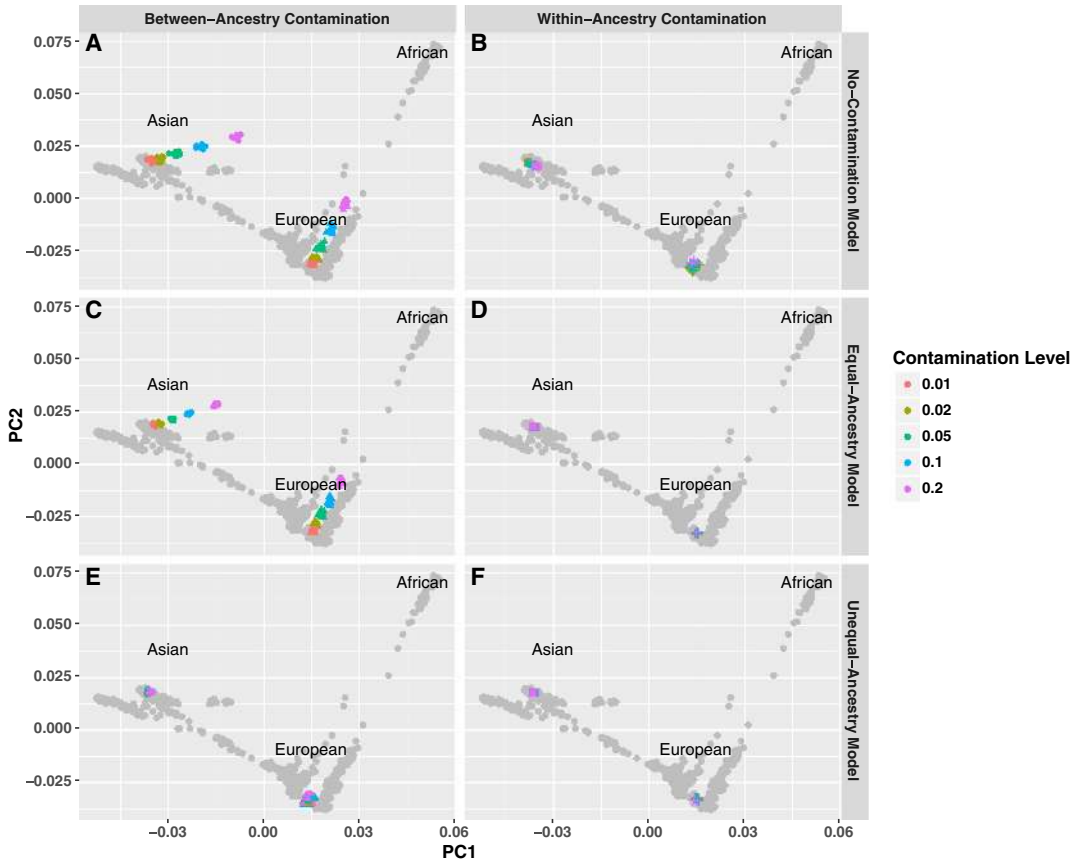

Figure 3. Impact of DNA sample contamination on the estimation of genetic ancestry. Each point represents a sample. Each gray point represents reference (HGDP) sample and its PCA coordinates, similar to Figure 2. Each colored point represents in silico–contaminated samples across various contamination rates and populations. In panels *A*, *C*, and *E*, European (GBR) and East Asian (CHS) samples are contaminated with African (YRI) samples at different contamination rates (i.e., between-ancestry contamination). In panels *B*, *D*, and *F*, European (GBR) and East Asian (CHS) samples are contaminated with another sample in the same population (i.e., within-ancestry contamination). Different colors represent different contamination rates ranging from 1% to 20%. *Upper* panels (*A*,*B*) show *verifyBamID2* estimates without modeling contamination; *middle* panels (*C*,*D*), *verifyBamID2* estimates under the assumption that intended and contaminating populations are identical (i.e., equal-ancestry model); *lower* panels (*E*,*F*), *verifyBamID2* estimates under the assumption that intended and contaminating populations can be different (i.e., unequal-ancestry model).

Robust, accurate, ancestry-agnostic estimation of DNA contamination

Next, we evaluated the effect of genetic ancestry misspecification in estimating DNA contamination rates. We constructed contaminated samples between various combinations of populations and compared the accuracy of estimated contamination rates using both the original methods which assume known allele frequencies and the new methods which estimate contamination rate and genetic ancestry jointly.

When contamination happens within the same population, running original methods with correct continental population allele frequencies specified provided accurate contamination estimates (Fig. 4A,E,I). However, using pooled allele frequencies, which would be a default option when it is infeasible to specify population information a priori before sequencing, consistently underestimated contamination rates. Bias was particularly large when intended individuals were of African ancestry.

Specifying incorrect population allele frequencies results in even larger contamination estimation bias. For example, using African allele frequencies on East Asian samples resulted in an average estimate of 2.9% contamination for samples with contamination 10% (Supplemental Table S1), implying that a large fraction of 10% contaminated samples within East Asian ancestry would not have been flagged for contamination-based exclusion at the contamination-exclusion threshold of 1%–3% used by many studies—for example, the Trans-Omics Precision Medicine (TOPMed) study (Natarajan et al. 2018).

Our results consistently demonstrated that the ancestry-agnostic method provides as accurate estimates as the original methods specified with correct population labels (Fig. 4A,E,I; Supplemental Table S1), and the estimates are substantially better than those from pooled allele frequencies or incorrectly specified allele frequencies (Table 2).

When the intended and contaminating populations are different, we observed that contamination is sometimes overestimated due to an increased fraction of heterozygous genotypes than expected by a given contamination rate under the single population model. Our method based on an unequal-ancestry model outperforms all the other methods in terms of overall bias and mean squared error (MSE) (Fig. 4; Supplemental Table S4), correcting for both upward and downward biases in various ancestry combinations. For example, the relative deviation of estimated to intended contamination rate (i.e., $|\hat{\alpha}/\alpha - 1|$) is reduced by

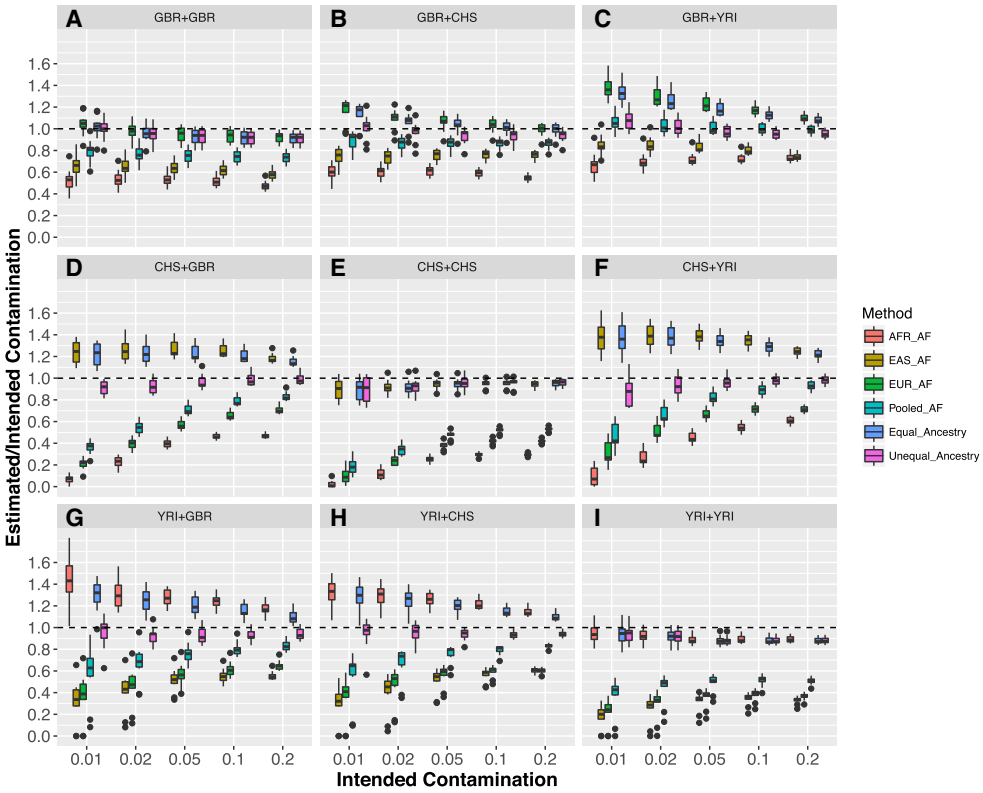

Figure 4. Comparison of different models to estimate contamination rates. Horizontal (*x*) axis shows intended contamination rate, vertical (*y*) axis shows the ratio of estimated to intended contamination rates. Each color represents different models to estimate contamination rates. EUR_AF, EAS_AF, and AFR_AF represent original *verifyBamID* using European, East Asian, and African allele frequencies across the continental population using the 1000 Genomes data. Pooled_AF represents the original *verifyBamID* using aggregated allele frequencies across all 2504 individuals in the 1000 Genomes Project. Equal_Ancestry represents the *verifyBamID2* assuming that intended and contaminating samples belong to the same population. Unequal_Ancestry represents *verifyBamID2* allowing different genetic ancestry between intended and contaminating sample (recommended setting). Each panel (*A–I*) represents different combinations of intended (row) and contaminating (column) populations, in the order of GBR, CHS, and YRI.

80% (73%–88%) compared to the original *verifyBamID* with various population allele frequencies, suggesting reduced bias. MSE is also reduced by 92% (86%–97%). This robustness reflects the ability to incorporate differences in population allele frequencies between intended and contaminating individuals (Fig. 4B–D,F–H; Supplemental Table S1).

We also examined the accuracy of our methods for admixed populations by performing a similar experiment using the Mexican population (MXL) and obtained consistent results (Supplemental Table S2).

Results with deep whole-genome sequence data from the InPSYght study

Next, we applied our methods to 500 African-American samples from the InPSYght study (see Methods). Consistent with the

Table 2. Average contamination estimates for 5% contaminated samples (size $n = 10$)

Sample population		Original model (fixed allele frequencies)					
Intended	Contaminating	European	East Asian	African	Pooled	Equal-ancestry model	Unequal-ancestry model
GBR	GBR	**4.73%**	3.19%	2.67%	3.76%	4.63%	4.63%
CHS	CHS	1.90%	4.73%	1.25%	2.38%	4.73%	**4.76%**
YRI	YRI	1.78%	1.58%	**4.44%**	2.45%	4.40%	4.40%
CHS	YRI	3.33%	6.91%	2.27%	4.10%	6.71%	**4.81%**
YRI	CHS	2.79%	2.55%	6.29%	3.76%	5.99%	**4.67%**
GBR	YRI	6.13%	4.16%	3.60%	5.04%	5.90%	**4.83%**
YRI	GBR	2.81%	2.57%	6.38%	3.80%	6.01%	**4.63%**
CHS	GBR	2.87%	6.33%	1.98%	3.55%	6.13%	**4.83%**
GBR	CHS	5.32%	3.78%	3.05%	4.32%	**5.16%**	4.67%

Average contamination estimates of in silico–contaminated samples when the true contamination rate is 5%. Each mixing configuration (e.g., GBR + CHS) contains 10 samples that are constructed with 95% reads coming from the intended sample and 5% reads from the contaminating sample. The estimated contamination rates are obtained using the original version *verifyBamID* by specifying prior allele frequencies as European, East Asian, African, and Pooled, respectively. Boldface represents the closest estimate to the true value of 5%.

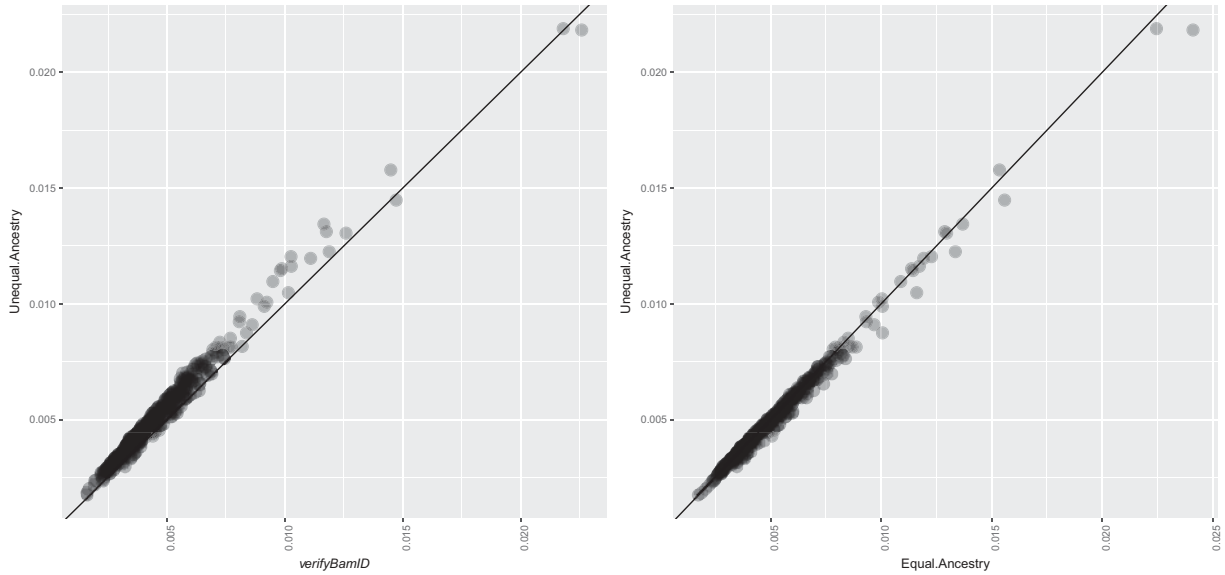

Figure 5. Comparison of contamination estimation between using *verifyBamID* and *verifyBamID2* on 500 InPSYght samples. All subjects are African-Americans. Each dot represents the pair of contamination rate estimates using different methods. The *left* panel shows the estimated contamination rates of the original *verifyBamID* with pooled allele frequencies, which is the default setting of *verifyBamID* on the *x*-axis. The *y*-axis shows *verifyBamID2* with unequal-ancestry model. Each point represents a sequenced subject. The *right* panel compares the estimated contamination rates between two models (unequal-ancestry vs. equal-ancestry) of *verifyBamID2* on the same data set.

results from our in silico contamination studies, we observed that the average contamination rate was 1.1-fold higher with the newer method (0.36% for unequal-ancestry, 0.37% for equal-ancestry) compared to the original method with pooled allele frequency (0.33%) (Fig. 5). The number of samples with an estimated contamination rate >1% increased from 16 (original method with pooled allele frequency) to 21 (unequal-ancestry method) or 23 (equal-ancestry method), suggesting our new method more rigorously screens for contaminated samples.

All 500 deeply sequenced genomes in InPSYght study are reported to be African-Americans, and indeed the estimated PC coordinates for all 500 individuals under all three methods lie between European and African samples. Compared to other methods to estimate genetic ancestry, our estimates resulted in tighter clustering along the European-African segment than LASER and similarly tight clustering to *TRACE* (Fig. 2B,D,F). For example, the correlation coefficient between the PC1 and PC2 coordinates were 0.927 for LASER, 0.981 for *TRACE*, and 0.985 for *verifyBamID2*, corroborating that *verifyBamID2* results in a more precise estimate of African ancestry along the European-African segment in PC coordinates.

Impact of number of markers on accuracy, computational cost, and memory requirements

As we have shown previously (Jun et al. 2012), there are trade-offs between computation cost and accuracy of contamination estimates. Using as many as 100,000 variants results in an accurately estimated intended contamination rate. For example, the MSE of relative deviation (i.e., $|\hat{\alpha}/\alpha - 1|$) was 0.02, 0.01, and 0.01 when the intended contamination was 1%, 2%, and 5%, respectively. When we use 10,000 variants, the MSEs modestly increased to 0.11, 0.04, and 0.01, respectively. When we use only 1000 variants, MSEs further increased to 0.69, 0.25, and 0.11, suggesting that the estimates may not be precise for the low contamination rate when using only 1000 variants (Supplemental Table S3).

We also evaluated the computational cost and memory consumption of *verifyBamID2* on whole-genome sequence data with various coverages. For the BAM files from the 1000 Genomes whole-genome sequence data (4.3–5.1× coverage), the average wall-clock running time was 5.5 min with a single thread and peak memory consumption was 505 MB when using 10,000 markers in a server with a Xeon 2.27 GHz processor. When using 100,000 markers, the average wall-clock running time was 20.5 min with a single thread and 8.0 min with four threads, and peak memory consumption was 528 MB.

For deep genome data from the InPSYght study (31× coverage) stored in CRAM format, the average wall-clock time was 17.3 min and peak memory consumption was 514 MB when using 10,000 markers. For 100,000 markers, the average wall-clock time was 155.6 min (single thread) or 96 min (four threads) and peak memory consumption was 548 MB.

Discussion

Contamination detection is an essential step in the sequence analysis process that has important effects on the following downstream analyses. Early and accurate estimation of DNA contamination can prevent wasted effort, time, and money by identifying the problems early on before too many samples are sequenced using contamination-prone protocols. Our previous method enabled such a timely contamination detection from sequence data and population allele frequencies at known variant sites, without requiring independent SNP genotype data. Our new method maintains these advantages and, in addition, provides three more. First, because our joint analysis method is agnostic to genetic ancestry, it eliminates sample-to-sample variation in the parameter settings for the contamination checking procedure, simplifying the sequence analysis pipeline. Second, it provides more robust contamination estimates against potentially misspecified population allele frequency of the intended (or contaminating) samples

when relying on the reported ancestry information. Third, it provides accurate estimates of genetic ancestries for both intended and contaminating samples. This enables additional sanity checking of the sequence data, such as determining whether a sequenced sample matches its expected (participant-reported) ancestry. It also facilitates incorporating ancestry information in the variant calling and downstream analysis and allows us to track the source of contamination more precisely when contamination occurs.

Our method can be used not only to detect and estimate contamination but also to estimate genetic ancestry from sequence data. Relatively few methods, such as LASER (Wang et al. 2014, 2015) and *bammds* (Malaspinas et al. 2014), exist for estimating genetic ancestry from sequence data, while several methods have been developed for array-based genotypes, such as EIGENSOFT (Price et al. 2006), FRAPPE (Tang et al. 2005), ADMIXTURE (Alexander et al. 2009), and *TRACE* (Wang et al. 2015). We have demonstrated that our method provides ancestry estimates as or more accurate than LASER, particularly when the sequenced samples are contaminated between different ancestries.

By jointly estimating genetic ancestry and contamination, we are able to accurately estimate contamination without requiring ancestry information a priori. Since obtaining population allele frequency information may be infeasible or even impossible at the time of sequencing, it is important to highlight that our ancestry-agnostic approach provides more timely and accurate feedback to the sequencing facilities. Our ancestry-agnostic approach also simplifies the sequence analysis pipeline, because the same input arguments can be applied across all samples regardless of their genetic ancestry. In the case where self-reported ancestries are available, our method can identify errors in the self-reported ancestries while estimating contamination.

The key idea of using individual-specific allele frequencies (ISAF) to account for population structure in genetic analysis has been suggested previously in the context of characterizing population structure or identifying highly differentiated variants across populations (Hao et al. 2015; Conomos et al. 2016). To the best of our knowledge, our method describes the first likelihood-based model utilizing ISAF to represent high-throughput sequence reads under population structure and/or contamination. While previous studies proposed logistic models as an alternative to linear models (Hao et al. 2015; Conomos et al. 2016), we used linear models (bounded by minimum and maximum value) between allele frequencies and population structure represented by singular value decomposition (SVD) on the genotype matrix. We made this choice because the logistic model is computationally more intensive, and the linear model is accurate for the common variants we use, as demonstrated by the previous studies (Hao et al. 2015).

Even though our method substantially improves the accuracy of contamination estimates compared to the original *verifyBamID*, we do see slight underestimation of contamination rates, especially when the intended contamination rate is high. Our method overestimates contamination if there are more heterozygous genotypes than expected by allele frequencies under Hardy-Weinberg Equilibrium (HWE) and underestimates contamination if there are less heterozygous genotypes than expected. We believe that slightly inaccurate allele frequency estimates (even with ISAF) and violation of HWE (due to population structure or copy number variants) are contributing to the slight underestimation of contamination rates, but we have not validated the conjecture experimentally yet.

Because we use Nelder-Mead optimization for maximum likelihood estimation, it is possible that the estimates do not converge to the global maximum, especially when many principal components are used. We observed that estimating the full unequal-ancestry model parameters sometimes does fail to converge, especially when there is little or no contamination, due to the limited identifiability of the genetic ancestry of contaminating samples in this situation. Starting by estimating the contamination rate and shared genetic ancestry parameters using the equal-ancestry model and using those estimates as starting values for the unequal-ancestry model to allow different ancestries between the intended and contaminating samples dramatically improved convergence; in fact, the method converged to consistent estimates across multiple starting points within 1000 iterations in all our benchmark cases, in both real and in silico–contaminated data. When the contamination rate is extremely small (e.g., <0.1%), estimation of genetic ancestry of contaminating samples can still be challenging, but its impact on genotyping accuracy is likely small as demonstrated previously (Jun et al. 2012). We allow unequal ancestries between intended and contaminating samples only when the likelihood substantially improves beyond the Akaike Information Criterion (AIC) (Akaike 1974) threshold between equal-ancestry and unequal-ancestry models. This procedure effectively removed all outlier estimates of genetic ancestries of contaminating samples in our experiments.

There are other possible useful extensions to our joint contamination and estimation method. We are extending these methods to detect and estimate contamination for RNA-seq and other epigenomic sequence data. The method can also be extended to handle contamination in cancer genomic data. The same model has potential utility in other areas, such as single-cell transcriptomics (Kang et al. 2018). As our method leverages excess heterozygosity to estimate contamination rates, it is important that the sequence reads have many variant sites with read depth 2 or greater to have sufficient power to estimate contamination in the extended models.

We expect that our new *verifyBamID2* software will facilitate more accurate, convenient, and timely quality control of sequence genomes. Our software tool is publicly available at http://github.com/Griffan/verifyBamID. Our GitHub repository provides reference files that can be used as test input for our methods. These files contain key input files required for *verifyBamID2*, including variant loadings, supporting various genome builds (GRCh37 and GRCh38), and various numbers of variants.

Methods

We aim to jointly estimate sample contamination rates and genetic ancestry from sequence reads without specifying population allele frequencies. First, we describe our previous mixture model to estimate contamination rates assuming population allele frequencies are known. Second, we introduce a model for sequence reads using population allele frequencies as a function of genetic ancestry represented in principal component coordinates. Third, we extend the model to enable joint estimation of contamination rates and genetic ancestry. Fourth, we evaluate our methods using in silico–contaminated samples and whole-genome sequence data from the InPSYght study.

Likelihood-based mixture model for DNA sequence contamination

In our previous contamination detection methods (Jun et al. 2012), we assumed that the DNA sequence reads from an intended sample are contaminated by sequence reads from, at most, one contaminating sample from the same population and that the

population allele frequencies of all analyzed genetic variants are known. For each bi-allelic variant i ($1 \le i \le m$), let $b_{ij} \in \{R,A,O\}$ ($1 \le j \le D_i$) be the observed base call representing the reference allele (R), alternate allele (A), or other allele (O) for the j-th read that overlaps the variant; D_i is the observed sequence depth at variant i. Let $e_{ij} \in \{0,1\}$ be a random variable indicating whether a sequencing error did (1) or did not (0) occur for observed base b_{ij}; we assume e_{ij} follows a Bernoulli distribution with success probability $10^{-Q_{ij}/10}$ where Q_{ij} is a Phred-scale base quality score of b_{ij}. In the absence of contamination, if the true genotype $g_i^s \in \{0,1,2\}$ represents the count of alternate alleles of the sequenced sample $s \in \{1,2\}$, then $\Pr(b_{ij}|g_i^s,e_{ij})$ can be easily represented as in Table 3, making the simplifying assumption of equally likely errors across four possible nucleotides.

We assume that the observed sequence reads are a $(1-\alpha)$:α mixture of intended and contaminating reads given a contamination rate $0 \le \alpha \le 1$. Let g_i^1 and g_i^2 represent the true genotypes of the intended and contaminating samples at variant i, respectively. Then, the mixture model likelihood of each observed base becomes

$$\Pr(b_{ij}|g_i^1,g_i^2,e_{ij};\alpha) = (1-\alpha)\Pr(b_{ij}|g_i^1,e_{ij}) + \alpha\Pr(b_{ij}|g_i^2,e_{ij}). \quad (1)$$

Assuming a homogenous population with known population allele frequency f_i and Hardy-Weinberg Equilibrium, $\Pr(g_i^2; f_i)$ follows a Binomial(2, f_i) distribution. Under the simplifying assumption of independent variants, the likelihood of the contamination rate becomes

$$L(\alpha) = \prod_{i=1}^{m} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{D_i} \sum_{e_{ij}} \Pr(b_{ij}|g_i^1,g_i^2,e_{ij};\alpha)\Pr(e_{ij}) \right\} \Pr(g_i^2;f_i)\Pr(g_i^1;f_i). \quad (2)$$

The maximum likelihood estimate (MLE) of contamination rate $\hat{\alpha}$ can be obtained using Brent's algorithm (Brent 1973).

As we previously reported (Jun et al. 2012), this model assumes correctly specified population allele frequencies f_i.

Likelihood-based estimation of genetic ancestry (in the absence of contamination)

We extend this model to incorporate genetic ancestry. The key idea of this extension is to use the individual-specific allele frequency (Hao et al. 2015; Conomos et al. 2016) to model the likelihood of the sequence reads. Several methods, including spatial ancestry analysis (SPA) (Yang et al. 2012) and logistic factor analysis (LFA) (Hao et al. 2015), previously proposed modeling allele frequency as a function of genetic ancestry via principal component (PC) coordinates.

Table 3. Conditional probability P(b_{ij} g_i, e_{ij}) of read b_{ij} given true genotype g_i and the variable representing the event of base calling error e_{ij}

True genotype g_i	Base calling error event e_{ij}	Pr (b_{ij} = R)	Pr (b_{ij} = A)	Pr (b_{ij} = O)[b]
g_i = RR[a]	e_{ij} = 0	1	0	0
	e_{ij} = 1	0	1/3	2/3
g_i = RA[a]	e_{ij} = 0	1/2	1/2	0
	e_{ij} = 1	1/6	1/6	2/3
g_i = AA[a]	e_{ij} = 0	0	1	0
	e_{ij} = 1	1/3	0	2/3

As described in Jun et al. (2012).
[a]RR, RA, AA: homozygous reference, heterozygous, and homozygous nonreference genotypes.
[b]O: alleles other than R or A; assumes four possible alleles (bases).

Let G be an $m \times n$ genotype matrix (where G_{ir} = 0, 1, or 2 is the number of nonreference alleles at variant i in individual r) of a genetically diverse reference panel of size n, such as 1000 Genomes or HGDP. We define ISAF $f_i (0 \le f_i \le 1)$ for variant i as a weighted average of genotypes from the reference panel ($f_i = \sum_{r=1}^{n} w_r G_{ir}$), where $0 \le w_r \le 1$ and $G_{ir} \in \{0,1,2\}$ for individual r. For a homogeneous population, $w_r = 1/2n$ results in a *pooled allele frequency* across all individuals in the reference panel. If each individual can be categorically represented as a one of k mutually exclusive subpopulations, the *population-specific allele frequency* for the subpopulation $s \in \{1,2,\cdots,k\}$ can be represented as $w_r = \frac{I(s_r = s)}{2n_s}$, where $s_r \in \{1,2,\cdots,k\}$ represents the subpopulation that individual r belongs to, and n_s represents the size of subpopulation s. More generally, if individuals' genetic ancestry is represented as continuous variables (such as PCs, SPAs, or LFAs), the individual-specific allele frequency can be represented as a function of the continuously represented genetic ancestry (Wang et al. 2014; Hao et al. 2015).

The estimated ISAF can be viewed as one-half times the genotype dosages approximated from a fixed number(=K) of factors, such as PCs, SPAs, or LFAs. In our method, we used a linear model to estimate ISAF from PCs, similar to previous studies (Hao et al. 2015; Conomos et al. 2016). Given the reference panel genotype matrix G, let $1/2\hat{G}$ be the *ISAF matrix* as a function of top K factors. ISAF matrix $1/2\hat{G}$ should well approximate $1/2G$. For example, under a linear model, typical principal component analysis takes the singular value decomposition of the mean-centered genotype matrix $\bar{G} = G - 2\boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}} = UDV^T$, where $\boldsymbol{\mu} = 1/2nG\mathbf{1}_{\boldsymbol{n}}$ is the pooled allele frequencies and $\mathbf{1}_{\boldsymbol{n}}$ is the column-vector of ones. Using the top K eigenvalues and corresponding eigenvectors $U^{(K)},D^{(K)},V^{(K)}$ from the SVD, it is known that $\hat{G} = 1/2U^{(K)}D^{(K)}[V^{(K)}]^T + \boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}$ minimizes $G - \hat{G}_2 = \sum_{i,r}(G_{ir} - \hat{G}_{ir})^2$ among all possible rank K matrices (Pearson 1901), making it a good proxy for the ISAF matrix.

For a new individual s with genetic ancestry represented as $\boldsymbol{x_s} \in \mathrm{R}^K$ in the PC (eigenvector) space of the reference panel, the ISAF for i-th variant can be modeled as $f_i(\boldsymbol{x_s}) = \frac{1}{2}\boldsymbol{u_i^{(K)}}D^{(K)}\boldsymbol{x_s^T} + \mu_i$, where $\boldsymbol{u_i^{(K)}}$ is i-th row of $U^{(K)}$ and μ_i is the i-th element of $\boldsymbol{\mu}$. To avoid a boundary condition, we constrain $\varepsilon/2n \le f_i(\boldsymbol{x_s}) \le 1 - \varepsilon/2n$ for a fixed ε (we used $\varepsilon = 0.5$ in our experiments). Then, the overall likelihood of an individual's genetic ancestry $\boldsymbol{x}$ is

$$L(\boldsymbol{x_s}) = \prod_{i=1}^{m} \sum_{g_i} \left\{ \prod_{j=1}^{D_i} \sum_{e_{ij}} \Pr(b_{ij}|g_i,e_{ij})\Pr(e_{ij}) \right\} \Pr(g_i;f_i(\boldsymbol{x_s})), \quad (3)$$

where g_i represents the unobserved genotype of the sequenced sample at variant i. The maximum-likelihood genetic ancestry coordinates can be estimated as $\hat{\boldsymbol{x}}_{\boldsymbol{s}} = \mathrm{argmax}_{\boldsymbol{x_s} \in \mathrm{R}^k} L(\boldsymbol{x_s})$ using the Nelder-Mead (Nelder and Mead 1965) algorithm, starting with PC coordinates of a randomly selected individual from the reference panel. In all our experiments, we always obtained consistent estimates of $\hat{\boldsymbol{x}}_{\boldsymbol{s}}$ regardless of start values with $K = 4$, which is the default parameter of our implementation. Using $K = 4$ gave us noticeably more precise estimates of contamination rates and genetic ancestry than smaller K (Supplemental Fig. S1). Using larger values of K (e.g., $K = 8$) substantially increased the computational time of the Nelder-Mead algorithm and failed to converge occasionally.

Joint estimation of genetic ancestry and DNA contamination

Because our goal is to obtain unbiased estimates of the DNA contamination rate α agonistic to prior knowledge of the genetic

ancestry, we propose to jointly estimate α and ancestry by combining the models described in the previous sections. Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in R^K$ be the genetic ancestries of the intended and contaminating samples. Then, the likelihood under the combined model is

$$L(\alpha, \boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{i=1}^{m} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{D_i} \sum_{e_{ij}} \Pr(b_{ij}|g_i^1, g_i^2, e_{ij}; \alpha) \Pr(e_{ij}) \right\} \times \Pr(g_i^1; f_i(\boldsymbol{x}_1)) \Pr(g_i^2; f_i(\boldsymbol{x}_2)).$$

When the contamination rate $\alpha \approx 0$, the parameters corresponding to $\boldsymbol{x}_2$ do not contribute (much) to the likelihood, and the estimates of $\boldsymbol{x}_2$ may be unstable. To address this problem, we initially assume that the intended and contaminating samples are from the same population $\boldsymbol{x}_1 = \boldsymbol{x}_2$ ('equal-ancestry' model) and then repeat the analysis allowing for $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ ('unequal-ancestry' model). The dimension of parameter space for the unequal-ancestry model is $2k+1$. We choose final parameter estimates between the two models based on the Akaike Information Criterion (Akaike 1974).

Evaluation on in silico–contaminated data based on 1000 Genomes Project samples

We constructed in silico–contaminated DNA sequence reads using aligned low-coverage whole-genome sequence reads from the 1000 Genomes phase 3 project (The 1000 Genomes Project Consortium 2015). We filtered out unmapped and mark-duplicated reads and then randomly sampled aligned sequence reads proportional to the intended contamination rates $\alpha \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. To match the mixing proportion of sequence reads originated from intended and contaminating to be $(1-\alpha):\alpha$, each read was sampled with probability $(1-\alpha)$ and $B_1/B_2\alpha$ from each sample, where B_1 and B_2 are number of aligned bases from unique reads from intended and contaminating samples. We selected four populations, CHS (Han Chinese South), GBR (British in England and Scotland), MXL (Mexican Ancestry from Los Angeles, CA, USA), YRI (Yoruba in Ibadan, Nigeria), and arbitrarily selected 10 pairs of individuals with similar sequencing depths within the same population and across populations. To estimate genetic ancestry and/or contamination rate for these in silico–contaminated sequence reads, we used a reference panel of 938 HGDP (Cavalli-Sforza 2005) individuals across 1000, 10,000, and 100,000 randomly chosen SNPs (pooled MAF > 0.5%), avoiding variants masked by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). When we compared estimated genetic ancestry with LASER, we used the same set of selected SNPs and sequence reads as input. For *TRACE*, we used genotypes from the phase 3 release (for 1000 Genomes) or an interim call set from the GotCloud software tool (Jun et al. 2015) (for InPSYght, see the next section for details) on the same SNP set.

Experiment with real sequence data from the InPSYght study

Next, we applied our method to 500 deeply sequenced (mean depth 32×) genomes from the first two batches of the InPSYght study. For each sample, we evaluated the results from the six models: (1) the original *verifyBamID* using pooled allele frequencies; the original *verifyBamID* using (2) African, (3) East Asian, and (4) European allele frequencies; (5) the new *verifyBamID2* under the equal-ancestry model; and (6) *verifyBamID2* under the unequal-ancestry model. To calculate pooled, population-specific, and individual-specific allele frequencies, we used the 1000 Genomes phase 3 reference panel ($n = 2504$), randomly selecting 100,000 SNPs among the sites also polymorphic in Illumina Human Omni 2.5 array, with the same filtering criteria (MAF > 5% and 1000 Genomes mask) as above.

Data access

The sequence data from this study have been submitted to the NCBI database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under accession number phs001020.v2.p1. The software is published under the MIT license. The source code of *verifyBamID2* is available in the Supplemental Material as well as at https://github.com/Griffan/VerifyBamID.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by NIH grants HG009976 (from NHGRI, to M.F. and M.B.), HL137182 (from NHLBI, to H.M.K. and F.Z.), HG007022 (from NHGRI, to G.R.A.), and MH105653 (from NIMH, to M.B., S.A.G.T., L.J.S., H.M.K., and InPSYght Consortium).

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr* **19:** 716–723. doi:10.1109/TAC.1974.1100705

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19:** 1655–1664. doi:10.1101/gr.094052.109

Brent RP. 1973. *Algorithms for minimization without derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Cavalli-Sforza LL. 2005. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* **6:** 333–340. doi:10.1038/nrg1596

Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016. Model-free estimation of recent genetic relatedness. *Am J Hum Genet* **98:** 127–148. doi:10.1016/j.ajhg.2015.11.022

Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM. 2015. Correcting for sample contamination in genotype calling of DNA sequence data. *Am J Hum Genet* **97:** 284–290. doi:10.1016/j.ajhg.2015.07.002

Hao W, Song M, Storey JD. 2015. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics* **32:** 713–721. doi:10.1093/bioinformatics/btv641

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91:** 839–848. doi:10.1016/j.ajhg.2012.09.004

Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25:** 918–925. doi:10.1101/gr.176552.114

Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, et al. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36:** 89–94. doi:10.1038/nbt.4042

Malaspinas AS, Tange O, Moreno-Mayar JV, Rasmussen M, DeGiorgio M, Wang Y, Valdiosera CE, Politis G, Willerslev E, Nielsen R. 2014. *bammds*: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics* **30:** 2962–2964. doi:10.1093/bioinformatics/btu410

Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, Khera A V, Zhou W, Bloom JM, Engreitz JM, et al. 2018. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun* **9:** 3391. doi:10.1038/s41467-018-05747-8

Nelder JA, Mead R. 1965. A simplex method for function minimization. *Comput J* **7:** 308–313. doi:10.1093/comjnl/7.4.308

Pearson K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *Philos Mag Ser* **2:** 559–572. doi:10.1080/14786440109462720

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909. doi:10.1038/ng1847

Sanders SJ, Neale BM, Huang H, Werling DM, An JY, Dong S; Whole Genome Sequencing for Psychiatric Disorders (WGSPD), Abecasis G, Arguello PA, Blangero J, et al. 2017. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat Neurosci* **20:** 1661–1668. doi:10.1038/s41593-017-0017-9

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* **28:** 289–301. doi:10.1002/gepi.20064

Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, Branham KE, Heckenlively J, Fulton R, Wilson RK, et al. 2014. Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* **46:** 409–415. doi:10.1038/ng.2924

Wang C, Zhan X, Liang L, Abecasis GR, Lin X. 2015. Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *Am J Hum Genet* **96:** 926–937. doi:10.1016/j.ajhg.2015.04.018

Yang WW-Y, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* **44:** 725–731. doi:10.1038/ng.2285

Received November 28, 2018; accepted in revised form March 11, 2019.

Ancestry-agnostic estimation of DNA sample contamination from sequence reads

Fan Zhang, Matthew Flickinger, Sarah A. Gagliano Taliun, et al.

Genome Res. 2020 30: 185-194 originally published online January 24, 2020
Access the most recent version at doi:10.1101/gr.246934.118