American Journal of Human Genetics, Volume *91*

Supplemental Information

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

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

Supplemental Inventory

Supplemental Figures and Tables

Figure S1

Figure S2

Figure S3

Figure S4

Table S1

Table S2

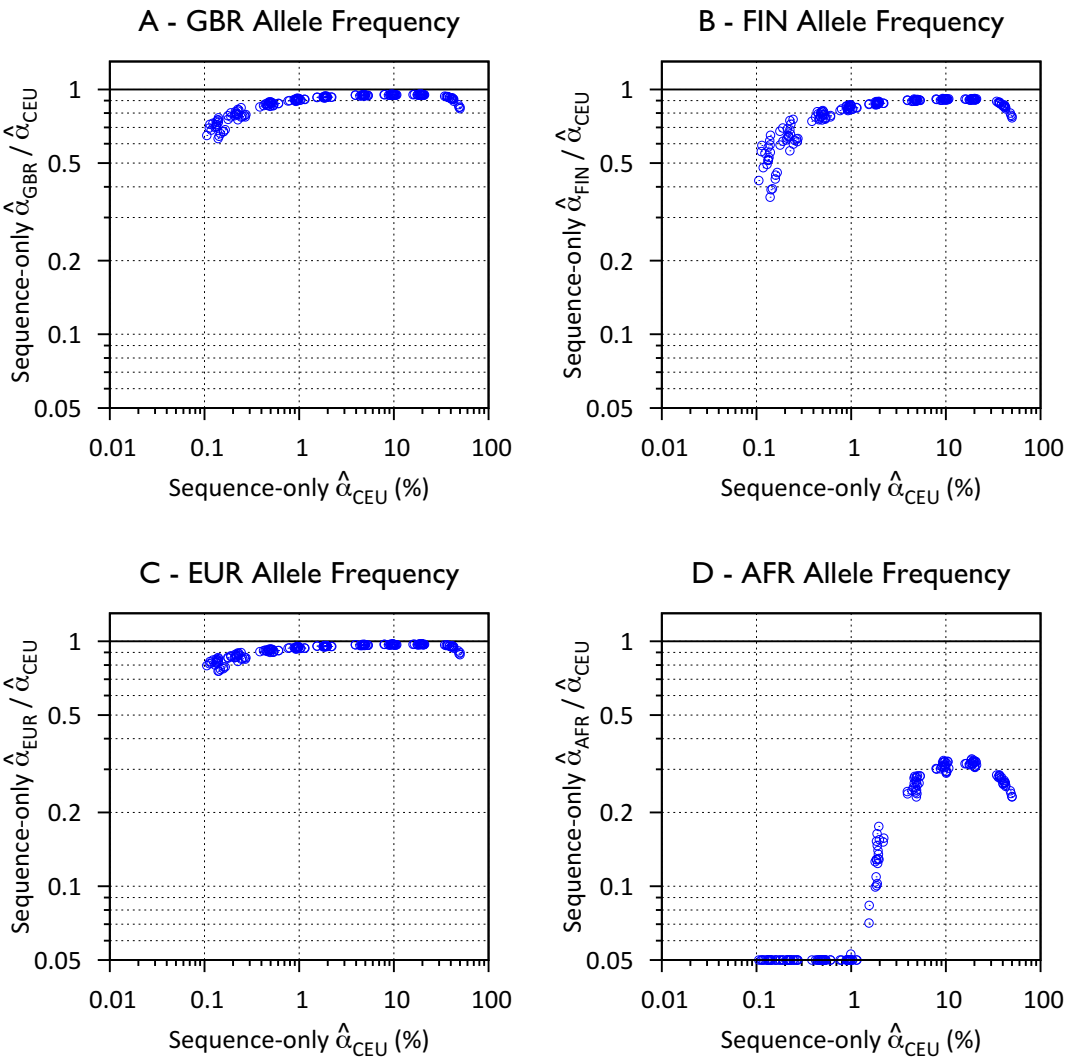

Figure S1. Impact of Population Allele Frequency on Estimated Contamination Levels.

Ratio between estimated contamination levels using different population allele frequencies with the sequence-only method.

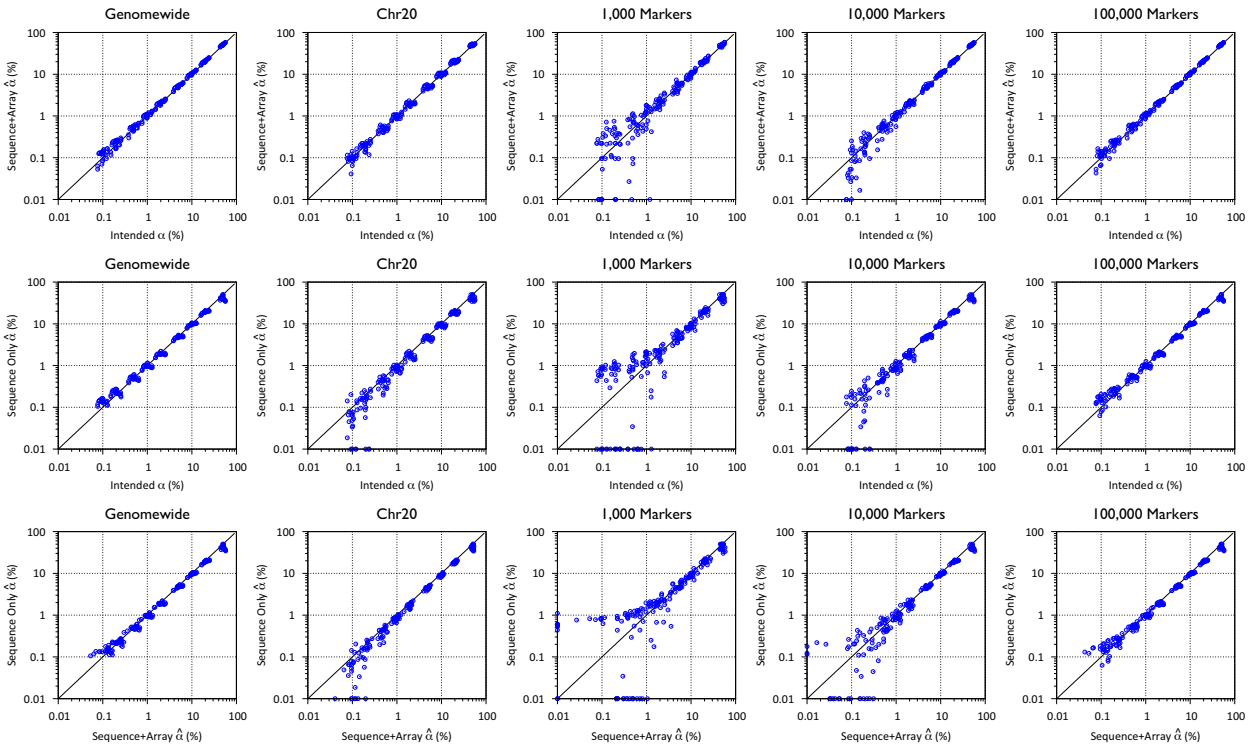

Figure S2. Estimated Contamination Levels Across Different Number of Markers.

Comparison between each pair of intended contamination level, estimated contamination levels $\hat{\alpha}$ using joint sequence and array-based method and $\hat{\alpha}$ using sequence-only method across different number of markers.

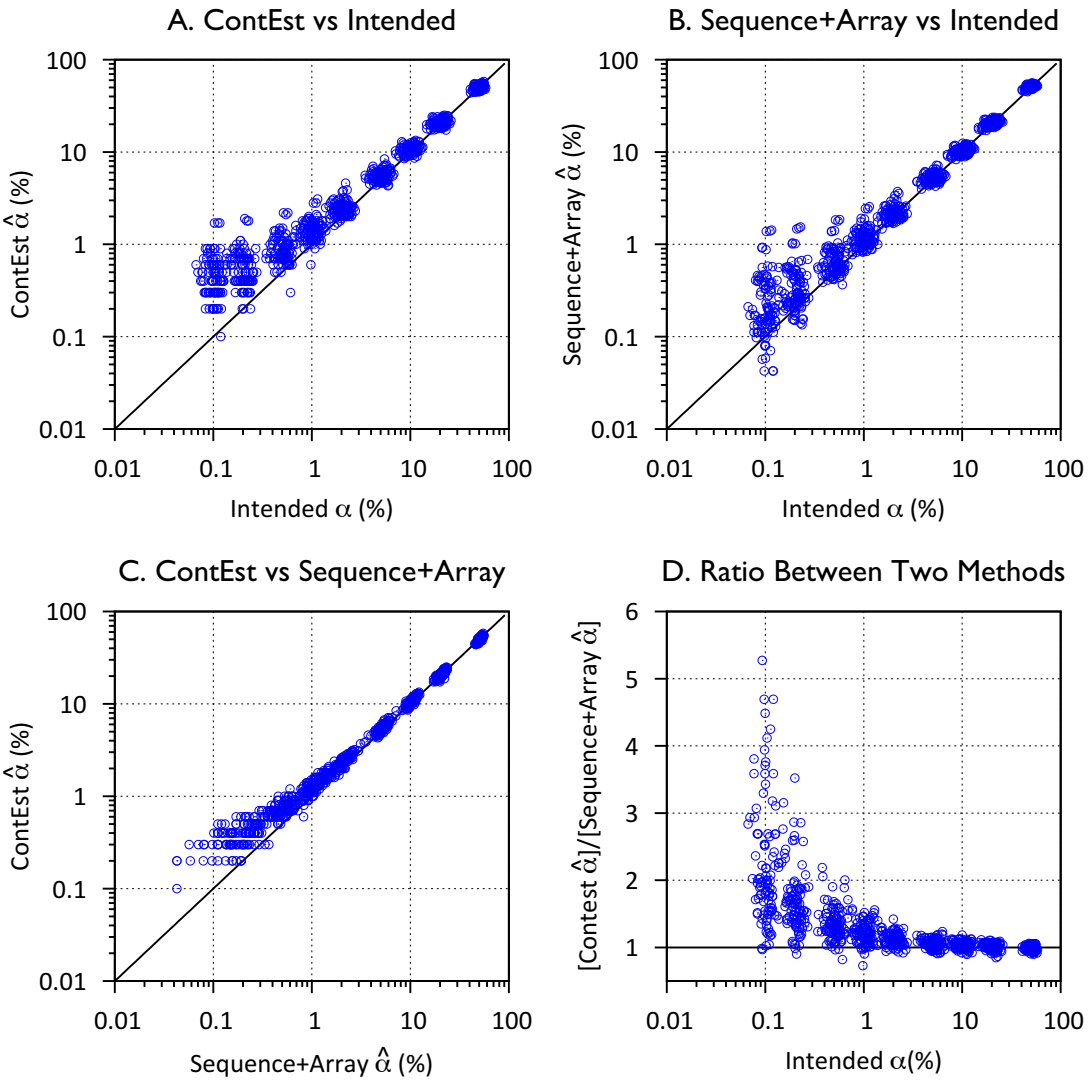

Figure S3. Comparison of Our Methods with ContEst Software
Comparison of estimated contamination levels between joint sequence and array-based method and ContEst on the *in-silico* simulated data for chromosome 20. (A) intended contaminations versus ContEst estimates (B) Our joint sequence and array-based method versus ContEst estimates (C) ratio between the two estimates.

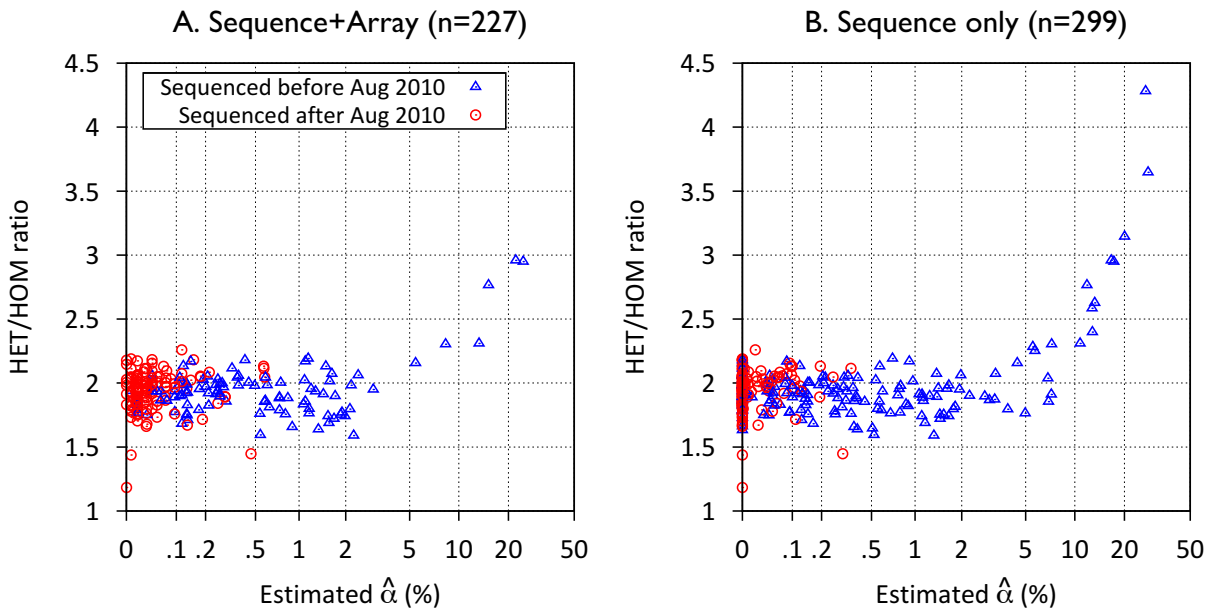

Figure S4. Excess Heterozygosity in relation to Estimated Contamination
Comparison of HET/HOM ratio to estimated contamination level $\hat{\alpha}$ in the type 2 diabetes sequencing study based on analysis of (A) sequence and genotype array data (n=227) and (B) sequence data only (n=299).

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

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

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

Table S2. Impact of multiple contaminating samples on estimated contamination

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

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