Published in final edited form as:
Nat Genet. 2014 April ; 46(4): 409–415. doi:10.1038/ng.2924.

Ancestry Estimation and Control of Population Stratification for Sequence-based Association Studies

Chaolong Wang[1,2,§,*], **Xiaowei Zhan**[2,*], **Jennifer Bragg-Gresham**[2], **Hyun Min Kang**[2], **Dwight Stambolian**[3], **Emily Y. Chew**[4], **Kari E. Branham**[5], **John Heckenlively**[5], **The FUSION Study**[6], **Robert Fulton**[7], **Richard K. Wilson**[7], **Elaine R. Mardis**[7], **Xihong Lin**[1], **Anand Swaroop**[8], **Sebastian Zöllner**[2,9], and **Gonçalo R. Abecasis**[2,§]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

[2]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109

[3]Department of Ophthalmology, University of Pennsylvania Medical School, Philadelphia, PA 19104

[4]Division of Epidemiology and Clinical Research, National Eye Institute, Bethesda, MD 20892

[5]Department of Ophthalmology, University of Michigan Kellogg Eye Center, Ann Arbor, MI 48105

[7]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63108

[8]Neurobiology-Neurodegeneration & Repair Laboratory, National Eye Institute, Bethesda, MD 20892

[9]Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI 48109

Abstract

Knowledge of individual ancestry is important for genetic association studies where population structure leads to false positive signals. Estimating individual ancestry with targeted sequence data, which constitutes the bulk of current sequence datasets, is challenging. Here, we propose a new method for accurate estimation of genetic ancestry. Our method skips genotype calling and directly analyzes sequence reads. We validate the method using simulated and empirical data and show that the method can accurately infer worldwide continental ancestry with whole genome shotgun coverage as low as 0.001X. For estimates of fine-scale ancestry within Europe, the method performs well with coverage of 0.1X. At an even finer-scale, the method improves discrimination between exome-sequenced participants originating from different provinces within

[§]Correspondence: cwang@hsph.harvard.edu (C.W.), goncalo@umich.edu (G.R.A.).
[6]Full lists of members and affiliations appear in the Supplementary Note.
[*]C.W. and X.Z. are joint first authors.

AUTHOR CONTRIBUTIONS
C.W., X.Z., S.Z. and G.R.A. conceived and implemented the approach. X.L. provided critical feedback on methodology and simulations. J.B.G., D.S., E.Y.C., K.E.B., J.H., R.F., R.K.W., E.R.M. and A.S. contributed the macular degeneration targeted sequencing data. H.M.K and FUSION collaborators contributed the Finnish exome sequence data. C.W. and G.R.A wrote the first draft of the paper. All authors reviewed, revised and contributed critical feedback to the manuscript and presentation.

COMPUTER SOFTWARE
LASER software and source code are available from our website at http://genome.sph.umich.edu/wiki/LASER.

Finland. Finally, we show that our method can be used to improve case-control matching in genetic association studies and reduce the risk of spurious findings due to population structure.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified thousands of common complex trait associated variants[1–4], but translating these discoveries into mechanistic insights has been challenging. In order to dissect the genetic architecture of complex traits, efforts are shifting to rare functional variants that can be detected with next generation sequencing. Building on advances in sequencing technologies and large sample sets obtained through collaboration, targeted sequencing studies can now interrogate abundant rare variants in samples of >10,000 individuals[5–9]. Early successes from these studies include type 1 diabetes[10], inflammatory bowel disease[11], and age-related macular degeneration (AMD)[12].

A key challenge in genetic association studies is to avoid spurious association signals caused by differences in ancestral background[13–16]. The identification of population structure is challenging for studies with targeted sequencing data. One reason is that targeted regions are typically short, account for only a fraction of the genome and do not contain sufficient genetic variation to infer global individual ancestry. Furthermore, targeted regions around disease-susceptibility loci are likely to harbor variants associated with the traits of interest so that corrections for stratification based on only these loci could mask true association signals.

Fortunately, targeted sequencing experiments also produce many reads that map outside target regions[6,17]. These off-target reads, resulting from limitations in capture technology, are often discarded and excluded from analysis. Still, when average off-target depth reaches >1–2X these reads can be used to discover and genotype SNPs across the genome[18,19], and with off-target depth >0.2–0.5X these reads can genotype common variants, albeit with high error rates[20]. Nevertheless, most targeted sequencing studies produce few off-target reads and off-target coverage is decreasing as capture technologies improve. In most targeted sequencing experiments it is thus difficult to accurately call off-target genotypes. In addition, the off-target sequence reads are distributed sparsely and randomly across each genome, so that the number of covered sites in any pair of samples is typically small. Methods for estimating ancestry that rely on high quality genotype data across a shared set of markers, such as principal components analysis (PCA)[21,22], do not produce good results when applied to targeted sequencing experiments – whether they are applied to targeted regions (which typically do not include enough information to estimate global ancestry) or to off-target regions (which typically do not produce high quality genotypes and where most pairs of samples will share few high-quality genotypes).

With high-quality genotype data, each principal component is defined as the product of a weight vector and a genotype vector, with weights reflecting the marginal information about ancestry provided by each site. With off-target sequence reads, entries in the genotype vector are often missing and can only be estimated with varying and often high error rates depending, for example, on the number of reads covering each locus. Intuitively, we might wish to adjust for missing data patterns and high error rates by adjusting the weight vector – for example, to ignore the contributions of loci with no data and to up-weigh the contributions of loci that have higher coverage.

Here, we propose a novel statistical method that addresses these challenges by estimating individual ancestry directly from off-target sequence reads without calling genotypes. We compare each sequenced sample to a set of reference individuals whose ancestral information is known and whose genome-wide SNP data are available[23,24]. Our method first constructs a reference coordinate system by applying PCA to SNP genotypes of the reference individuals, and then uses off-target reads to place study samples in this reference PCA space, one at a time. With an appropriate reference panel, the estimated coordinates of the study samples identify their ancestral background and can be directly used to correct for population structure in association studies or to ensure adequate matching of cases and controls.

To place each sample, we proceed as follows: First, we simulate sequence data for each reference individual, exactly matching the coverage pattern of the sample being studied (in this way, each reference individual will have the same number of reads covering each locus as the study sample). Then, we build a PCA ancestry map based on these simulated sequence reads for the reference individuals together with the real sequence reads for the study sample. Finally, we project this new ancestry map into the original PCA space using Procrustes analysis[25,26]. The transformation obtained from this analysis of the reference samples is then used to place the study sample in the original PCA space, appropriately up and down weighing sites according to their coverage and the information they contain about ancestry. The process is illustrated in Figure 1 and is described in detail in the Online Methods.

We validate the method using simulated low-coverage sequence data for a worldwide sample set[23] and a European sample set[24] and empirical targeted sequencing data from the 1000 Genomes exon project[27] and a case-control study of the macular degeneration[28]. Our results show that our method can accurately infer worldwide continental ancestry or even the fine-scale ancestry within Europe with extremely low off-target coverage (~0.001X for worldwide ancestry and ~0.10X for European ancestry). We have implemented our method in the publicly available LASER software (Locating Ancestry from SEquence Reads).

Overview of Simulations

To evaluate the performance of LASER, we first simulated sequence data for two sets of samples whose array genotype data are publicly available. One is the Human Genome Diversity Panel (HGDP), consisting of 938 individuals from 53 populations worldwide[23]; and the other is a subset of the Population Reference Sample (POPRES), consisting of 1,385 individuals from 37 European populations[24]. We split each sample set into one test set of individuals for whom we would simulate low coverage sequence data, and one reference set of individuals whose high quality genotypes would be used to construct the reference PCA space.

Inference of Worldwide Ancestry

For the worldwide sample set, we randomly selected 238 individuals from the HGDP[23], and used their array genotypes at 632,958 loci as templates to simulate sequence data (Online Methods). We simulated multiple sequence datasets with mean coverage ranging from 0.001X to 0.25X. The remaining 700 HGDP samples were used to construct the reference PCA space. We examined the first four principal components. These can be used to separate major continental groups in the HGDP (see Figure 2): PC1 and PC2 separate major continental groups in the Old World, while PC3 and PC4 further separate Native American and Oceanian populations, respectively. We applied LASER to each simulated sequence dataset to estimate the ancestry coordinates of the test individuals in the reference PCA space. We assessed the accuracy by comparing ancestry estimates derived from LASER to PCA coordinates of the test individuals based on their original SNP genotypes using the squared Pearson correlation r^2 along each PC and the Procrustes similarity t_0 (Online Methods). Our results show consistently high accuracy across all simulated datasets (Figure 2, Supplementary Table 1). When the simulated coverage is 0.001X (corresponding to ~630 loci covered with ≥1 reads), r^2 ranges from 0.7396 for PC4 to 0.9506 for PC1 and the Procrustes similarity is $t_0 = 0.9508$. Figure 2B shows that although the patterns are a bit fuzzy, major continental groups are well separated at 0.001X coverage. Accuracy increases with coverage; when the coverage is 0.10X, the estimated coordinates are almost identical to coordinates estimated using a GWAS SNP panel with $t_0 = 0.9993$ (Figure 2D, Supplementary Table 1). Thus, our method should be able to reconstruct worldwide ancestry with even very modest amounts of sequence data.

Inference of Ancestry Within Europe

Similarly, for estimates of fine-scale ancestry within Europe, we used genotypes at 318,682 loci and 385 randomly selected POPRES individuals[24] as templates to simulate low coverage sequence data (from 0.01X to 0.40X). The remaining 1,000 POPRES European ancestry samples were used to construct the reference PCA space. We focused on the top two PCs of the POPRES reference panel, which mirror the geographic map of Europe[24] (Figure 3A). Compared to the estimates of worldwide continental ancestry, much higher coverage is required to reveal the more subtle differences in population structure within Europe (Figure 3, Supplementary Table 2). With an average coverage of 0.01X, samples clump in the center of the reference PCA space (Figure 3B, $r^2 = 0.5687$ for PC1 and 0.0108 for PC2, $t_0 = 0.4786$). As coverage increases to 0.05X (Figure 3C), we become able to observe population structure along PC1 ($r^2 = 0.8851$), which separates Northern and Southern Europeans, but still no structure along PC2 ($r^2 = 0.2516$). Clear population structure within Europe is revealed when coverage is >0.10X (Figure 3D–F), with t_0 increasing from 0.9126 (0.10X coverage) to 0.9764 (0.40X coverage) (Supplementary Table 2). Thus, reconstructing ancestry within Europe requires substantially more data than reconstructing continental ancestry in a worldwide sample.

Evaluation with 1000 Genomes Project Data

We then evaluated LASER using empirical data from the 1000 Genomes exon pilot[27], which produced deep sequence data for the exons of 906 genes in a subset of the samples studied by the International HapMap Consortium[29]. We examined 410 samples passing quality control from seven worldwide populations (see Supplementary Table 3 and Online Methods). We used all 938 HGDP individuals to construct the reference PCA space. The average off-target sequencing coverage for the 410 samples was ~0.096X at the 633K SNP loci genotyped in the HGDP (Supplementary Figure 1). In this comparison, we generated ancestry estimates for each sample first using HapMap Consortium genotypes, and then using off-target sequence reads from the 1000 Genomes exon sequencing project. As shown in Supplementary Figure 2, coordinates estimated from off-target sequence reads are highly consistent with those based on SNP genotypes ($t_0 = 0.9955$, $r^2 = 0.9950$, 0.9871, 0.9439, and 0.7747 for PC1 to PC4). Even when focusing on 103 samples whose off-target coverage is below 0.06X, we still obtained $t_0 = 0.9938$ ($r^2 = 0.9930$, 0.9884, 0.9012, and 0.6811 for PC1 to PC4, Supplementary Table 4). Surprisingly, t_0 for the 103 samples with highest off-target coverage (from 0.10X to 0.55X) was slightly lower than t_0 for the lower coverage groups (Supplementary Table 4). This might be explained by different ancestry representation of samples in different coverage groups and by possible DNA contamination of some samples.

Evaluation Using Targeted Sequencing Data

We next applied LASER to 3,159 samples sequenced around eight macular degeneration susceptibility loci and two candidate regions[28]. The samples include 2,362 macular degeneration cases, 789 controls, two samples with unknown phenotype, and one European (CEU) and one Yoruba (YRI) nuclear family selected among the HapMap Project samples (each nuclear family included mother, father and a child). Macular degeneration cases and controls were recruited in Ophthalmology clinics across the United States. In these samples, off-target coverage was 0.224X across the 633K loci in HGDP, and 0.241X across the 319K loci in POPRES (Supplementary Figure 3). When using the HGDP as the reference panel, the two trios were placed to the correct positions: the CEU trio clustered with the HGDP Europeans, and the YRI trio clustered with the HGDP Africans. Diverse ancestral background was observed among the 3,153 case-control samples: 3,069 clustered with Europeans/Middle Eastern ancestry individuals; 73 aligned between Africans and Europeans (likely corresponding to African American samples); five aligned between Europeans and Native Americans; three clustered with Central/South Asians; and three clustered with East Asians (Supplementary Figure 4A–B). We then used the POPRES reference panel to dissect the population structure among samples in the cluster with European/Middle Eastern ancestry. Our results show that although most of these samples had northern European ancestry, many other samples formed a small cluster around southern Europe (Supplementary Figure 4C–D). For 931 of the sequenced AMD cases and controls, GWAS array genotype data are also available[30]. For these samples, results based on the off-target reads match well with the coordinates estimated using SNP genotypes, in both the HGDP PCA space ($t_0 = 0.9068$, Supplementary Figure 5) and the POPRES PCA space ($t_0 = 0.9209$, Supplementary Figure 6). The accuracy increased for samples with higher off-target coverage (Supplementary Table 5).

Evaluation Using Exome Sequence Data

The previous experiments examined situations where targeted regions were relatively small. A large number of modern sequencing studies target entire exomes. To explore whether our method might be useful in this setting, we examined ancestry estimates derived from exome sequence data. For this analysis, we used 941 Finnish individuals from the FUSION[31] (Supplementary Table 6) that have been extensively characterized as part of the GoT2D Study of Type 2 Diabetes Genetics, by genotyping on the Omni 2.5M array, by deep exome sequencing (~96X depth, 0.69 million variants) and by low pass whole genome sequencing (~5X depth, 27 million variants). We constructed a reference PCA space using 470 individuals and genotypes at ~8.4 million SNPs with minor allele frequency (MAF) ≥0.01. We then placed the remaining 471 individuals into this reference map, using ancestry estimates derived using whole genome sequencing data as a gold standard. Figure 4 shows that ancestry estimates derived using our method are much more similar to this gold standard (t_0 = 0.9763; r^2 = 0.9778 for PC1 and 0.9259 for PC2) than results based on exome genotypes alone (t_0 = 0.8263; r^2 = 0.9411 for PC1 and 0.4373 for PC2) and better separate individuals born in the different provinces of Finland. This improved separation of individuals originating from different parts of Finland is highlighted when variance in PCA coordinates is decomposed into within-province and between-providence components: between-province variation in coordinates increases from 48% when using exome genotypes to 64% using our method (see Online Methods).

In contrast to our Finnish example, many contemporary analyses will rely on reference panels where array-based genotypes (rather than whole genome sequence data) are available. In this setting, the advantages of our method are even more dramatic, as illustrated by an analysis of simulated exome sequence data for samples with diverse European ancestries[24] (Online Methods). For each simulated sample, we used the empirical coverage pattern from a randomly selected exome sequencing project sample[32], with overall average on-target and off-target depths of ~88.9X and ~1.0X, respectively. In this setting, ancestry placements within a PCA ancestry map of Europe were inaccurate when based on genotypes for deeply sequenced regions (Procrustes similarity t_0 = 0.5031, r^2 = 0.7589 for PC1 and 0.0007 for PC2, Supplementary Figure 7A). In contrast, using off-target reads, our method provided accurate estimates of individual ancestry (t_0 = 0.9467, r^2 = 0.9744 for PC1 and 0.7640 for PC2, Supplementary Figure 7B). Incorporating both on-target and off-target reads, our ancestry estimates improve further (t_0 = 0.9669, r^2 = 0.9804 for PC1 and 0.8610 for PC2, Supplementary Figure 7C). We also note that, compared to the simulations in Supplementary Table 2, ancestry estimates appear less accurate in this setting, for two reasons: first, because empirical coverage patterns in the exome sequencing project data are more uneven than in our original simulation, second (and more importantly), because there is great variation in per individual off-target coverage in the exome sequencing project samples (ranging from 0.49X to 4.70X in our simulated samples). As reference panels of sequenced individuals become commonplace, we expect that ancestry estimates using exome genotypes or using our method will both improve substantially.

Controlling for Population Structure in Association Studies

Our final set of simulations explored whether ancestry coordinates estimated using our method could help control for population stratification[21,22]. To mimic population structure within Europe, we simulated individuals distributed along a 20 × 20 lattice, as suggested by Mathieson and McVean (2012)[16]. We then preferentially sampled 1,500 cases from one half of the lattice. When these cases were matched to 1,500 controls sampled at random across the whole lattice, we observed strong inflation in association test statistics with genomic control inflation factor of $\lambda_{\text{common}} = 1.326$ for common variants (MAF ≥ 0.05) and $\lambda_{\text{lowfreq}} = 1.267$ for low-frequency variants ($0.01 \leq \text{MAF} < 0.05$) (Table 1). When our estimated principal components were used as covariates in association analysis, evidence for stratification was much reduced, resulting in $\lambda_{\text{common}} = 0.992$ and $\lambda_{\text{lowfreq}} = 0.996$ at 0.10X coverage ($t_0 = 0.9993$; $r^2 = 0.9986$ for PC1 and 0.9985 for PC2); and in $\lambda_{\text{common}} = 0.991$ and $\lambda_{\text{lowfreq}} = 0.998$ with more modest 0.005X coverage ($t_0 = 0.9853$; $r^2 = 0.9711$ for PC1 and 0.9706 for PC2) (Table 1). In a second analysis, we simulated sequence data for 10,800 potential controls and used estimated ancestry coordinates to select 1,500 controls matching our cases[33]. In this second analysis, we again successfully controlled for stratification with $\lambda_{\text{common}} = 1.011$ and $\lambda_{\text{lowfreq}} = 1.013$ at 0.10X coverage and to $\lambda_{\text{common}} = 1.041$ and $\lambda_{\text{lowfreq}} = 1.045$ at 0.005X coverage (Table 1). We next explored more challenging sampling strategies where all cases were sampled from one or two 8 × 8 grids (Supplementary Figure 8). In these more challenging settings, using estimated PCA as covariates did not adequately control for stratification (Supplementary Table 7). In comparison, matching-based analyses were more robust, and were able to control for stratification in all scenarios, provided off-target coverage was greater than 0.10X (Supplementary Table 7). This observation is important, since it suggests that while using PCA as covariates will be adequate in situations where mild stratification is expected, matching based strategies will be robust in a wider variety of settings.

DISCUSSION

We show that the genetic ancestry of an individual can be accurately estimated using off-target sequence reads that are a by-product of most targeted sequencing studies. With off-target reads corresponding to 0.001X coverage of the genome, worldwide continental ancestry can be reconstructed; and with off-target reads corresponding to 0.10X coverage, ancestry can be estimated within Europe. Since Europe is the continent with the most homogeneous genetic variation[34], we expect LASER can be used to infer fine-scale structure within other continents when appropriate reference panels are available. A key ingredient for successful application of our method is the availability of appropriate reference samples that can be used to define the PCA space. We used HGDP samples[23] to construct a worldwide continental ancestry map and POPRES samples[24] to construct a genetic ancestry map of Europe. Both HGDP and POPRES samples were genotyped with standard GWAS arrays; if these reference samples were genotyped at higher density or whole genome sequenced, we would expect our method to perform even better as it would increase the number of overlapping sites between sequenced samples and these reference panels, making it easier to discern subtle population structure[34,35]. We also note that one should be extremely careful in interpreting PCA ancestry maps when the reference panel does not include ancestries in the study sample. For this reason, we always recommend starting with a worldwide ancestry map and gradually focusing on more regional maps.

Our simulations used several simplifying assumptions. For example, we used a Poisson distribution to simulate coverage and assumed a uniform sequencing error rate of 1% per base. In practice, we expect these assumptions will have only a minor impact on our results.

For example, although less uniform distribution of coverage might require slight increases in depth for accurate estimates of ancestry, this could be counter-acted by improved genotyping of reference samples. In addition, simulations showed that our method is relatively robust to misspecification of sequencing error rates (Supplementary Tables 8–9).

We foresee several potential enhancements to our approach. For example, since different runs of our method will show small stochastic variation in the placement of each individual, we expect that repeated analysis of the same sample can improve results (Supplementary Figure 9) – particularly when coverage is very low or when trying to place samples in a European ancestry map (or another map where differences between populations are small). Our simulations show that averaging results over 10 repeated runs for a sample sequenced with 0.10X coverage produces ancestry placements within the map of Europe almost as accurate as generating a single placement based on a sample sequenced with 0.20X coverage. Another interesting challenge is the development of methods that can be used with other ancestry spaces, such as those derived from multidimensional scaling approaches[36,37] or direct modeling of allele frequency gradients[38].

As targeted sequencing technologies improve, there has been a constant drive to reduce off-target sequencing coverage. In principle, reducing off-target coverage can decrease sequencing costs, by minimizing the amount of sequencing effort expended on low priority areas of the genome. Our work shows that, even in the context of disease association studies, reads that map to low priority areas of the genome can be of high value – for example, because they enable sequencing studies to access large pools of sequenced controls. Often, PCA has been used to model experimental artifacts, such as batch effects, in addition to population structure. Our approach, which places one sample at a time in a pre-defined reference ancestry space, does not capture artifacts due to experimental batch effects or close relatedness between samples. This allows us to separate genetic ancestry from other contributors to sample structure. In practice, when artifacts due to batch effects are a concern, ancestry estimates derived using our method can be combined with key summaries of sequence data (for example, by summarizing sequencing depth, read length, or even locus-by-locus coverage information in an additional set of PCs)[39,40]. When relatedness is a concern, our method can robustly estimate individual ancestry but will not identify cryptic relatedness. If pedigree information is available, the ancestry information provided by our method can be combined with mixed models for association analysis[41–43]. In other cases, further methodological developments may be needed to accurately identify related individuals using off-target sequencing reads.

Computationally, our method examines one sample at a time. Thus, computational costs increase linearly with the number of samples to be analyzed and analyses can easily be run in parallel. The cost for analysis of each sample depends on the number of individuals, N, and markers, L, in the reference panel and the fraction of loci with nonzero coverage, λ, in the study sample. Roughly, we expect computational cost for each sample to be $O(N^2L\lambda + N^3)$, which is the time required to compute the pairwise similarity matrix of the sample specific reference panel and the corresponding eigen decomposition. In our simulations, analysis typically required no more than a few minutes per sample (*e.g.*, ~1.3 minutes when $N = 1000$, $L \approx 319K$, and $\lambda \approx 0.2$).

Our simulations show that using estimated ancestry coordinates as covariates is expected to reduce modest inflation in test statistics due to population structure and imperfect matching of case-control samples. However, our simulations also show that when stratification is more severe, matching based strategies can control for stratification in a wider variety of settings. Alternative solutions might be to estimate higher order PCs[21] or to use nonlinear techniques, such as the kernel smoothing methods, to correct for structure based on our estimated PCs[44]. The diverse ancestry observed among sequenced AMD samples further illustrates the importance and utility of estimating the ancestry of study samples in genetic association studies. Using off-target reads to estimate ancestry enabled us to match cases to previously sequenced controls and increase sample size and statistical power in a targeted sequencing study of macular degeneration. In this way, we were able to ancestry match potential control samples from public resources with sequenced cases, enabling the discovery of a rare variant, p.Lys155Gln, in the *C3* gene that is significantly associated with increased risk of macular degeneration[28]. This sort of matching of study samples to public resources illustrates how accurate reconstructions of ancestry enable new and interesting study designs and analytical possibilities.

ONLINE METHODS

All experiments relied on pre-existing data. The original collection of DNA, genotypes and sequence data was carried out with informed consent of human participants. Experiments described here were approved by the University of Michigan Institutional Review Board.

THE LASER METHOD

The LASER method consists of (1) principal component analysis (PCA) on reference genotypes to define a reference ancestry space; (2) simulation of sequence data for reference individuals, matching the coverage of each study sample; (3) PCA on combined sequence data; and (4) Procrustes analysis to transform coordinates from step 3 into the reference ancestry space. Step 1 is performed once, and later steps are repeated for each sample.

PCA on reference genotypes

We code reference genotypes in matrix $\boldsymbol{G}$. Each $\boldsymbol{G}_{ij} = 0, 1, 2$ represents the number reference alleles at locus $j = 1 \dots L$ for individual $i = 1 \dots N$. Let μ_j and σ_j represent column means and standard deviations for this matrix. A standardized genotypic matrix $\boldsymbol{Q}$ is defined by $\boldsymbol{Q}_{ij} = (\boldsymbol{G}_{ij} - \mu_j)/\sigma_j$. Missing entries and invariant columns ($\sigma_j = 0$) in $\boldsymbol{G}$ are set to 0 in $\boldsymbol{Q}$. After eigen decomposition of the $N \times N$ matrix $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{Q}^T$, the kth PCA is given by $\lambda_k^{1/2}\overrightarrow{v}_k$, where λ_k is the k^{th} eigen value of $\boldsymbol{M}$ and $\overrightarrow{v_k}$ is the corresponding eigen vector. Coordinates of the top K PCAs for reference individuals are stored in $N \times K$ matrix $\boldsymbol{Y}$.

Simulating sequence data for the reference individuals

We simulate sequence data for reference individuals matching the coverage pattern of study samples. Suppose we are analyzing study sample h. For locus j, let $\boldsymbol{C}_{hj}$ tally the total number of overlapping reads and $\boldsymbol{S}_{hj}$ tally the subset that match the reference allele. We store simulated sequence data in matrices $\boldsymbol{C}'$ and $\boldsymbol{S}'$. We fix simulated coverage $\boldsymbol{C}'_{ij}=\boldsymbol{C}_{hj}$ for all i

and j, exactly matching the sample being analyzed. We draw the count of reference alleles as:

$$\boldsymbol{S}'_{ij}|\boldsymbol{G}_{ij},\boldsymbol{C}'_{ij}\sim\begin{cases}\text{Binomial}(\boldsymbol{C}'_{ij},\varepsilon),\text{ if }\boldsymbol{G}_{ij}=0\\ \text{Binomial}(\boldsymbol{C}'_{ij},0.5),\text{ if }\boldsymbol{G}_{ij}=1\\ \text{Binomial}(\boldsymbol{C}'_{ij},1-\varepsilon),\text{ if }\boldsymbol{G}_{ij}=2\end{cases}\quad(1)$$

Here, ε is the estimated sequencing error rate per base (ε =0.01 unless noted). If $\boldsymbol{G}_{ij}$ is missing, we set $\boldsymbol{S}'_{ij}$ to missing.

PCA on combined sequence data

To perform PCA on the reference individuals together with the study sample h, we next stack matrix $\boldsymbol{S}'$ and row vector $\boldsymbol{S}_h$. To reduce computational complexity, we remove columns where all elements are zero and obtain matrix $\tilde{\boldsymbol{S}}$. We then perform PCA on matrix $\tilde{\boldsymbol{S}}$ and store the top K PCs for reference individuals in $N \times K$ matrix $\boldsymbol{X}$ and for the study sample in K -element vector $\boldsymbol{Z}_h$.

Procrustes analysis

To place the study sample into the reference PCA space, we apply Procrustes analysis[25,26] to find a transformation f (including translation, scaling, rotation, and reflection) that maximizes the similarity between $f(\boldsymbol{X})$ and $\boldsymbol{Y}$ while preserving the relative pairwise distances among points within $\boldsymbol{X}$. We then obtain $\boldsymbol{Z}_h^*=f(\boldsymbol{Z}_h)$, the coordinates of the study sample in the reference coordinate space. Success can be quantified by a Procrustes similarity statistic $t(\boldsymbol{X},\boldsymbol{Y})=\sqrt{1-D}$, where D is the scaled minimum sum of squared Euclidean distances between $f(\boldsymbol{X})$ and $\boldsymbol{Y}$ across all possible transformations, ranging from 0 to 1 (ref. [26]). Lower Procrustes similarity corresponds to greater uncertainty and a less reliable $\boldsymbol{Z}_h^*$.

GENETIC DATA

Genotype Data

We used Human Genome Diversity Panel (HGDP)[23] and Population Reference Sample (POPRES)[24,45] genotypes to define reference coordinate spaces. The HGDP dataset includes 632,958 autosomal SNPs and 938 unrelated individuals from 53 worldwide populations[23]. Our POPRES subset contains 318,682 autosomal SNPs and 1,385 individuals from 37 European populations[24]. For both datasets, we pre-processed data as summarized in Supplementary Figure 10, excluding SNPs with different alleles in 1000 Genomes data and dbSNP, >2 alleles, ambiguous strand or missing from dbSNP (version 135).

We also analyzed genotypes from HapMap Project[29] and AMD GWAS[30]. In the HapMap dataset we focused on 410 individuals that overlap with the 1000 Genomes pilot exon project (1,294,658 SNPs). In the AMD GWAS we focused on 931 individuals also in our targeted sequencing study (316,475 SNPs; Supplementary Figure 11).

Targeted Sequencing Data

The 1000 Genomes pilot exon project sequenced exons of 906 randomly selected genes at >50X average depth[27]. We analyzed 410 individuals from 7 populations overlapping with HapMap and estimated contamination rates < 10% (Supplementary Table 3)[46]. The AMD targeted sequencing dataset included 6 HapMap individuals (CEU trio: NA12878, NA12891, and NA12892; YRI trio: NA19238, NA19239, and NA19240), 2,362 cases and 789 controls recruited in ophthalmology clinics across the United States[28]. These were sequenced for 0.97 megabases across 10 regions to 127.5X average depth.

Exome Sequence Data

The GoT2D Study of Type 2 Diabetes characterized 941 Finnish individuals from the Finland-United States Investigation of NIDDM Genetics (FUSION): exome sequencing at mean depth ~96X, whole genome sequencing at mean depth ~5X, genotyping on the Illumina Omni2.5 BeadChip. Our analyses focused on autosomal biallelic SNPs with missingness <5%, Hardy-Weinberg equilibrium $p > 10^{-6}$, and MAF > 0.01. After QC[47], whole genome analyses included 8,447,085 SNPs and exome analyses included 95,741 SNPs (of which 94,423 overlapped).

Pre-processing of Sequence Data

We started with BAM files and used the “mpileup” command in SAMtools[48] to extract bases overlapping loci genotyped in the reference panel. Sequence reads with Phred mapping quality <30 and bases with Phred quality score <20 were discarded. Unless noted, we only analyzed reads *outside* targeted regions.

HGDP and POPRES Simulations

We simulated sequence data for 238 randomly selected HGDP and 385 randomly selected POPRES samples. The remaining 700 HGDP and 1,000 POPRES individuals defined reference coordinate spaces. We first simulated Poisson coverage with mean between 0.001 and 0.40, then sampled reference alleles using Equation 1 (Supplementary Tables 1 and 2). We next repeated the simulation using coverage patterns from the NHLBI Exome Sequencing Project[32]. Among randomly selected NHLBI samples, mean exome coverage was ~88.9X and mean off-target coverage was ~1.0X.

COMPARISON WITH SNP-BASED PCA

When analyzing SNP genotypes, we combined genotypes for one study sample and N reference individuals and performed PCA on the shared set of SNPs. Then, we used Procrustes analysis to project the study sample into the reference PCA space[49]. When estimating SNP-based coordinates for samples in the 1000 Genomes pilot exon project, we used 581,686 SNPs that overlap between HapMap and HGDP. For the AMD samples, we used 45,700 SNPs shared between the HGDP, POPRES, and AMD datasets.

We used the squared Pearson correlation r^2 to measure concordance between sequence and SNP-based coordinates along each PCA. We also report overall similarity between the two sets of coordinates using the Procrustes similarity statistic t_0, obtained by using Procrustes analysis[26] to translate between sequence and SNP-derived coordinates for test samples.

FINE-SCALE POPULATION STRUCTURE

Each FUSION sample can be assigned to one of 12 sub-populations according to birth province. We split each sub-population into two groups, resulting in 470 reference individuals and 471 test individuals (Supplementary Table 6). We constructed a reference PCA map based on whole genome sequence results. We placed test individuals into this map using (1) whole genome genotypes, (2) genotypes across loci overlapping between exome and whole genome data, and (3) off-target reads generated during exome sequencing (~0.89X off-target depth).

To evaluate how well the three analyses capture population structure, we define statistic ψ as the proportion of between-population variance in PCA coordinates. We use K-dimensional vectors $\vec{x}_{ij}$, $\vec{\mu}_i$, and $\vec{v}$ to represent the coordinates of sample j from population i, the centroid of population i, and the overall centroid. For m populations, each with n_i sampled individuals, the proportion of between-population variance in the PCA is defined as

$$\psi=1-\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\vec{x}_{ij}-\vec{\mu}_i)(\vec{x}_{ij}-\vec{\mu}_i)^T}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\vec{x}_{ij}-\vec{v})(\vec{x}_{ij}-\vec{v})^T}. \qquad (2)$$

This statistic ranges from 0 to 1, and is similar in spirit to the F_{ST} statistic, which estimates between-population variance in allelic states[50]. Larger values of ψ indicate population structure is better captured.

SIMULATED CASE/CONTROL STUDIES

We simulated[51] 20,000 diploid individuals evenly distributed along a 20 × 20 lattice, each genotyped at 1M independent biallelic SNPs with MAF ≥ 0.01. The scaled migration rate between neighboring lattice points was $M = 10$, as suggested by ref.[16] to mimic population structure within Europe. In each lattice-point, we assigned 3 individuals to a reference set and, among the remainder, marked 20 as potential cases and the rest as potential controls. In total, this resulted in 1,200 reference individuals, 8,000 potential cases, and 10,800 potential controls. We first created stratified case/control data by preferentially sampling cases from the right half of the lattice (900 vs. 600 elsewhere) and sampling 1,500 controls randomly from the entire lattice. We explored more extreme settings by sampling cases from smaller regions of the lattice (Supplementary Figure 8). In these additional scenarios, we sampled 1,280 cases and 1,280 controls.

We then simulated sequence coverage between 0.001 and 0.20X and used LASER to place cases and controls in the 2-dimensional ancestry space defined by reference individuals. In association tests[52], we first used logistic regression or Cochran-Armitage trend tests and without correcting for stratification. To correct for stratification, we either incorporated estimated PCs as covariates in the logistic regression model or used a heuristic algorithm to identify one matched control for each case based on proximity in the reference ancestry space[28] and applied the Cochran-Mantel-Haenszel tests on the matched case/control pairs. Genomic inflation was calculated as ref.[16].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank investigators from the Finland United States Investigation of NIDDM Genetics (FUSION) and the Grand Opportunity Type 2 Diabetes Sequencing Project (GoT2D) for generously sharing whole genome and deep exome sequence data for 941 individuals prior to publication, and the D2D, Finrisk 2002, Health 2000, Action LADA, and Saviatipale studies for providing some of the FUSION-sequenced DNAs. We thank Jun Li for his assistance with the HGDP dataset, Heather Stringham and Adam Locke for assistance with the FUSION dataset and Matthew Brooks for organizing the macular degeneration samples. C.W. acknowledges funding support from a Howard Hughes Medical Institute International Student Research Fellowship. This study is supported by National Institute of Health (DK062370, HG000376, HG005552, HG006513, EY022005, HG007022, HG005855, HG003079, CA076404, CA134294) and the National Eye Institute Intramural Research Program.

References

1. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–8. [PubMed: 18988837]
2. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–69. [PubMed: 18398418]
3. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009; 10:241–51. [PubMed: 19293820]
4. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–7. [PubMed: 19474294]
5. Coventry A, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010; 1:131. [PubMed: 21119644]
6. Mamanova L, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010; 7:111–8. [PubMed: 20111037]
7. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011; 12:745–55. [PubMed: 21946919]
8. Shen P, et al. High-quality DNA sequence capture of 524 disease candidate genes. Proc Natl Acad Sci U S A. 2011; 108:6549–54. [PubMed: 21467225]
9. Nelson MR, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337:100–4. [PubMed: 22604722]
10. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–9. [PubMed: 19264985]
11. Rivas MA, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011; 43:1066–73. [PubMed: 21983784]
12. Raychaudhuri S, et al. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nat Genet. 2011; 43:1232–6. [PubMed: 22019782]
13. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003; 361:598–604. [PubMed: 12598158]
14. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004; 36:512–7. [PubMed: 15052271]
15. Clayton DG, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005; 37:1243–6. [PubMed: 16228001]
16. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012; 44:243–6. [PubMed: 22306651]

17. Clark MJ, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011; 29:908–14. [PubMed: 21947028]

18. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011; 21:940–51. [PubMed: 21460063]

19. Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 2011; 21:952–60. [PubMed: 20980557]

20. Pasaniuc B, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012; 44:631–5. [PubMed: 22610117]

21. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]

22. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11:459–63. [PubMed: 20548291]

23. Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319:1100–4. [PubMed: 18292342]

24. Novembre J, et al. Genes mirror geography within Europe. Nature. 2008; 456:98–101. [PubMed: 18758442]

25. Schönemann PH, Carroll RM. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika. 1970; 35:245–255.

26. Wang C, et al. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol. 2010; 9:Article 13. [PubMed: 20196748]

27. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]

28. Zhan X, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nat Genet. 2013; 45:1375–1379. [PubMed: 24036949]

29. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–8. [PubMed: 20811451]

30. Chen W, et al. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. Proc Natl Acad Sci U S A. 2010; 107:7401–6. [PubMed: 20385819]

31. Valle T, et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care. 1998; 21:949–58. [PubMed: 9614613]

32. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2012

33. Guan W, Liang L, Boehnke M, Abecasis GR. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. Genet Epidemiol. 2009; 33:508–17. [PubMed: 19170134]

34. Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012; 8:e1002886. [PubMed: 22927824]

35. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]

36. Miclaus K, Wolfinger R, Czika W. SNP selection and multidimensional scaling to quantify population structure. Genet Epidemiol. 2009; 33:488–96. [PubMed: 19194989]

37. Zhu C, Yu J. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics. 2009; 182:875–88. [PubMed: 19414565]

38. Yang WY, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. Nat Genet. 2012; 44:725–31. [PubMed: 22610118]

39. Fromer M, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012; 91:597–607. [PubMed: 23040492]

40. Krumm N, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012; 22:1525–32. [PubMed: 22585873]

41. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–54. [PubMed: 20208533]

42. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8:833–5. [PubMed: 21892150]

43. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44:821–4. [PubMed: 22706312]

44. Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol. 2003; 24:44–56. [PubMed: 12508255]

45. Nelson MR, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. 2008; 83:347–58. [PubMed: 18760391]

46. Jun G, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am J Hum Genet. 2012; 91:839–48. [PubMed: 23103226]

47. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–8. [PubMed: 21653522]

48. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]

49. Skoglund P, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science. 2012; 336:466–9. [PubMed: 22539720]

50. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009; 10:639–50. [PubMed: 19687804]

51. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–8. [PubMed: 11847089]

52. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. [PubMed: 17701901]

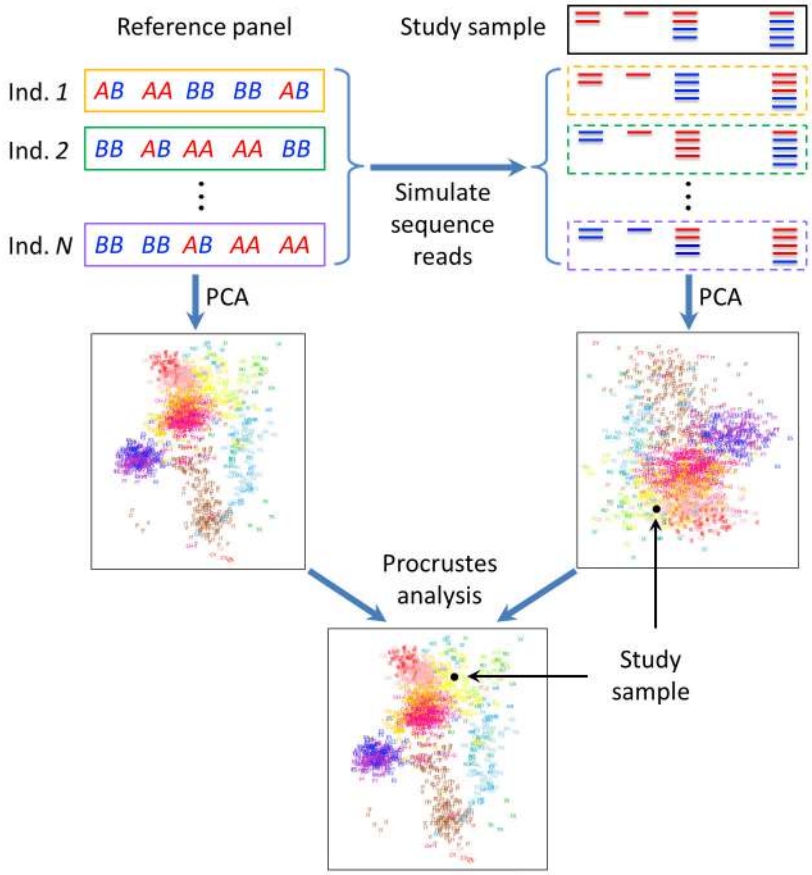

Figure 1.

Graphical illustration of the LASER method.

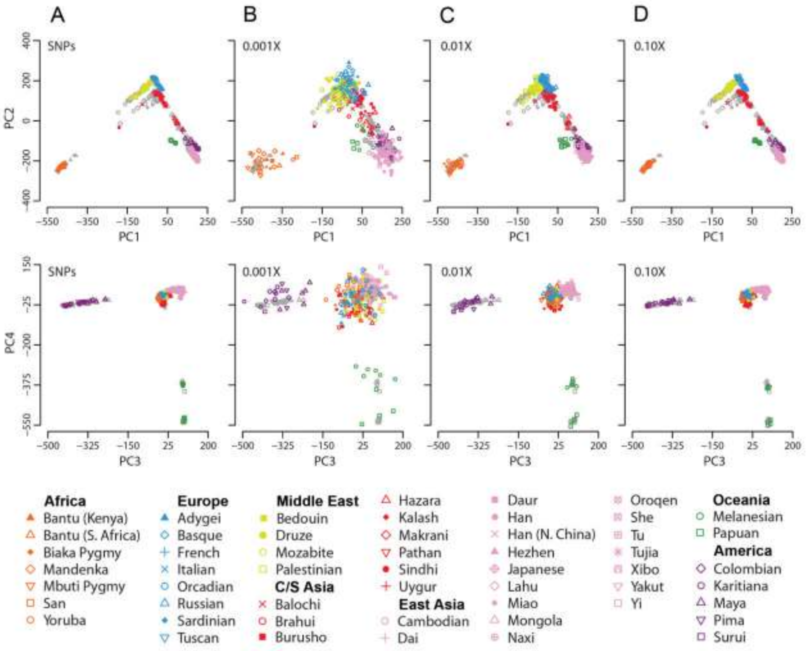

Figure 2. Estimation of worldwide continental ancestry

238 individuals were randomly selected from the HGDP as the testing set (colored symbols), and the remaining 700 HGDP individuals were used as the reference panel (gray symbols). The upper row shows PC1 and PC2, and the lower row shows PC3 and PC4. (A) Results based on SNP genotypes. (B) Results based on simulated sequence data at 0.001X coverage. The Procrustes similarity to the SNP-based coordinates is $t_0 = 0.9508$. (C) Results at 0.01X coverage ($t_0 = 0.9949$). (D) Results at 0.10X coverage ($t_0 = 0.9993$).

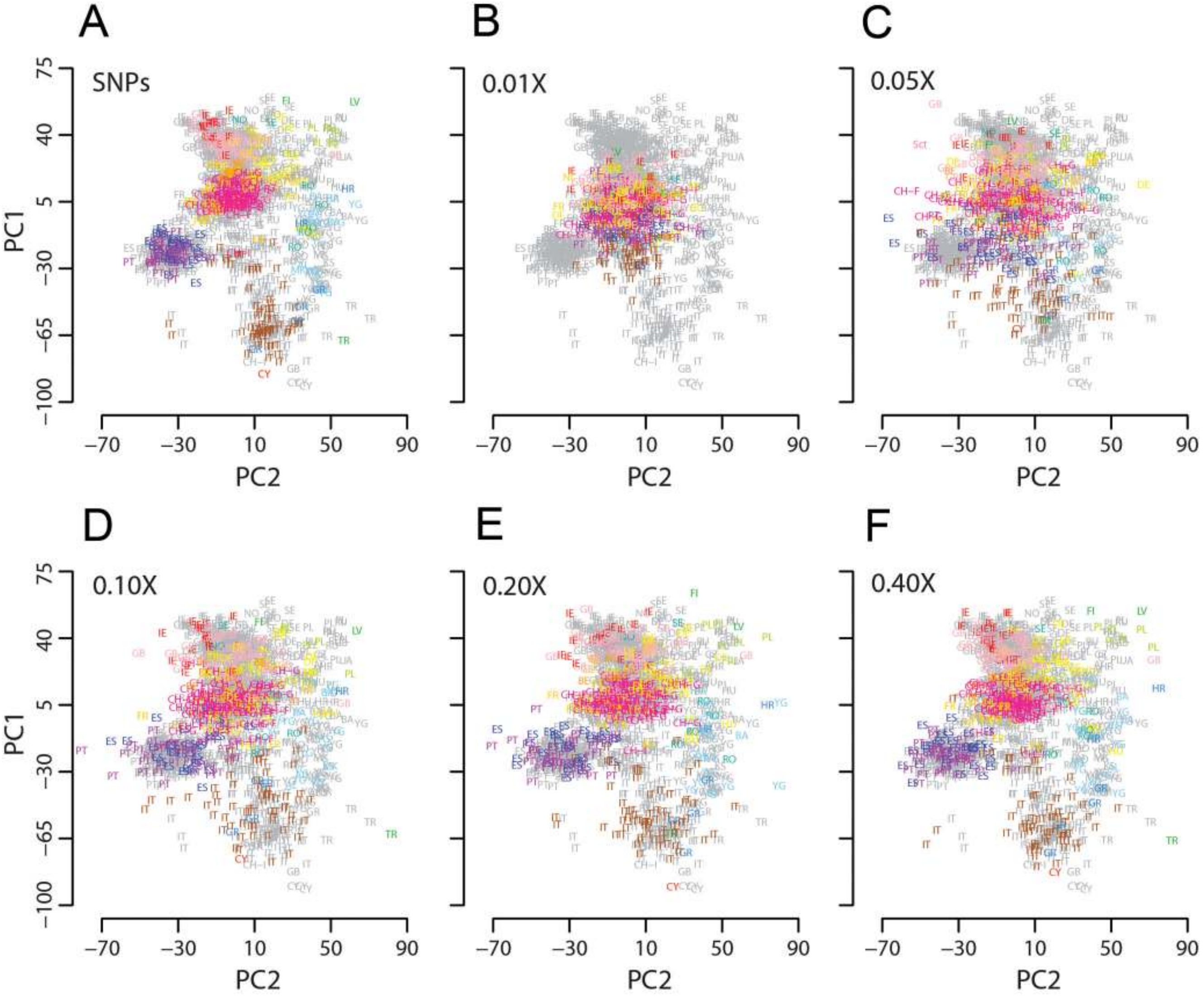

Figure 3. Estimation of ancestry within Europe

385 individuals were randomly selected from the POPRES as the testing set (colored symbols), and the remaining 1,000 POPRES individuals were used as the reference panel (gray symbols). (A) Results based on SNP genotypes. (B) Results based on simulated sequence data at 0.01X coverage. The Procrustes similarity to the SNP-based coordinates is $t_0 = 0.4786$. (C) Results at 0.05X coverage ($t_0 = 0.7720$). (D) Results at 0.10X coverage ($t_0 = 0.9137$). (E) Results at 0.20X coverage ($t_0 = 0.9495$). (F) Results at 0.40X coverage ($t_0 = 0.9764$). Population labels follow the color scheme of ref. [24]. Abbreviations are as follows: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH-F, Swiss-French; CH-G, Swiss-German; CH-I, Swiss-Italian; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NL, Netherlands; NO, Norway; PL, Poland; PT, Portugal; RO,

Romania; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Serbia and Montenegro.

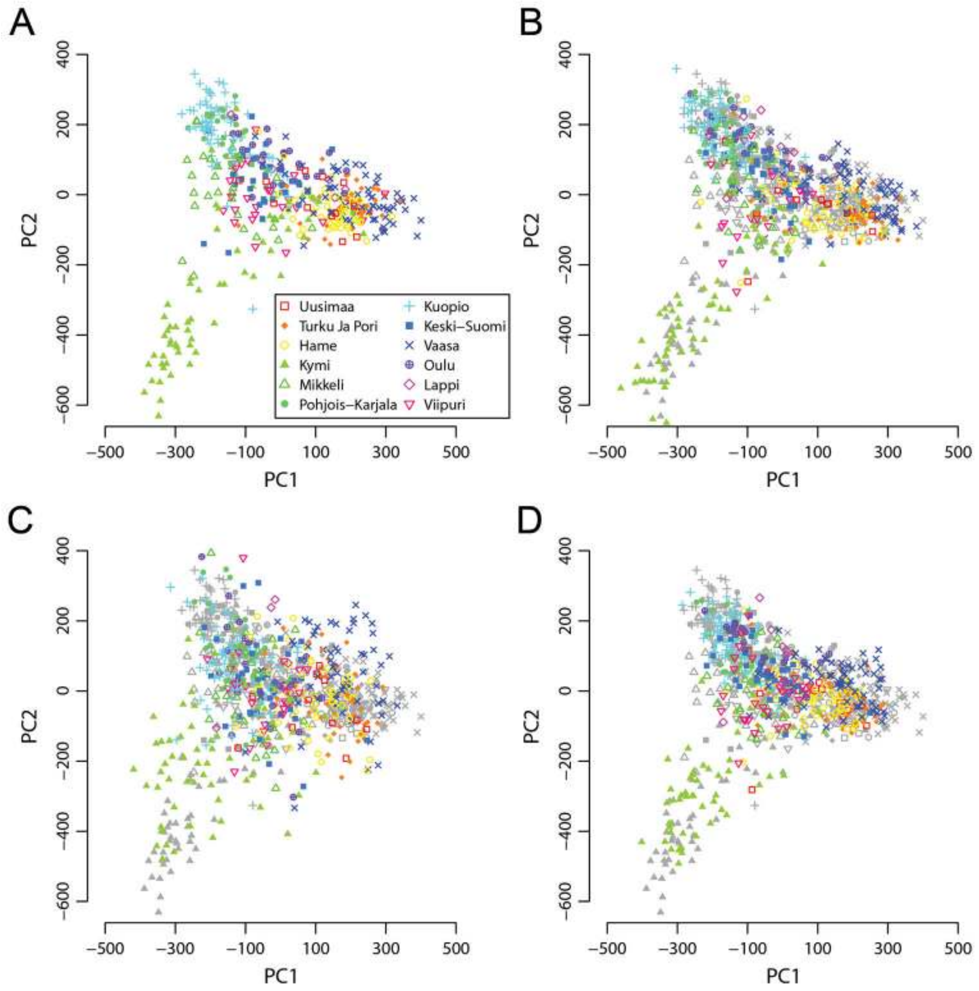

Figure 4. Estimation of fine-scale ancestry within Finland

(A) Reference PCA map based on integrated whole genome genotypes of 470 reference individuals. Proportion of among-population variance in the PCA map is $\psi = 0.6623$. (B) Estimation of ancestry for 471 test individuals based on integrated whole genome genotypes ($\psi = 0.6685$). Reference individuals are indicated by gray points. (C) Estimation of ancestry for test individuals based on exome sequencing genotypes ($\psi = 0.4849$). Compared to panel B, Procrustes similarity $t_0 = 0.8263$, and $r^2 = 0.9411$ and 0.4373 for PC1 and PC2 respectively. (D) Estimation of ancestry for test individuals based on genome-wide off-target reads from exome sequencing experiments ($\psi = 0.6385$). Compared to panel B, $t_0 = 0.9763$, and $r^2 = 0.9778$ and 0.9259 for PC1 and PC2 respectively. The mean coverage is ~96X and ~0.89X for on-target and off-target regions, respectively.

Table 1

Evaluation of corrections for stratification in simulated case/control data with 900 and 600 cases sampled respectively from two halves of the simulated 20 × 20 lattice.

Sequencing coverage	Similarity to SNP-based PCs			Regression based analyses		Matching based analyses	
	t_0	r^2 (PC1)	r^2 (PC2)	λ_{common}	$\lambda_{lowfreq}$	λ_{common}	$\lambda_{lowfreq}$
Uncorrected	-	-	-	1.326	1.264	1.326	1.267
SNP-based PCs	1	1	1	0.991	0.995	0.996	0.998
0.20X	0.9996	0.9993	0.9993	0.992	0.996	1.009	1.019
0.15X	0.9995	0.9990	0.9991	0.992	0.995	1.007	1.005
0.10X	0.9993	0.9986	0.9985	0.992	0.996	1.011	1.013
0.05X	0.9985	0.9972	0.9968	0.992	0.995	1.018	1.015
0.01X	0.9925	0.9851	0.9851	0.991	0.995	1.036	1.034
0.005X	0.9853	0.9711	0.9706	0.991	0.998	1.041	1.045
0.001X	0.9317	0.8635	0.8723	0.994	1.000	1.076	1.084

In this table, λ_{common} is the genomic inflation factor calculated based on 625,481 common variants (MAF ≥0.05), and $\lambda_{lowfreq}$ is the inflation factor based on 374,519 low frequency variants (0.01 ≤ MAF < 0.05). The Procrustes similarity score and squared correlations were calculated by comparing sequenced-based PCs to SNP-based PCs of the 1,500 cases. For uncorrected results, we used logistic regression (under regression based analyses) and Cochran-Armitage trend tests (under matching based analyses). Two approaches to correct for stratification were examined: (1) including estimated ancestry coordinates as covariates in logistic regression; (2) identifying one-to-one ancestry-matched case/control pairs for Cochran-Mantel-Haenszel tests (treating each matched pair as a stratum).