nature genetics

Ancestry estimation and control of population stratification for sequence-based association studies

Chaolong Wang[1,2,10], Xiaowei Zhan[2,10], 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] & Gonçalo R Abecasis[2]

Estimating individual ancestry is important in genetic association studies where population structure leads to false positive signals, although assigning ancestry remains challenging with targeted sequence data. We propose a new method for the accurate estimation of individual genetic ancestry, based on direct analysis of off-target sequence reads, and implement our method in the publicly available LASER software. We validate the method using simulated and empirical data and show that the method can accurately infer worldwide continental ancestry when used with sequencing data sets with whole-genome shotgun coverage as low as 0.001×. For estimates of fine-scale ancestry within Europe, the method performs well with coverage of 0.1×. On an even finer scale, the method improves discrimination between exome-sequenced study participants originating from different provinces within Finland. Finally, we show that our method can be used to improve case-control matching in genetic association studies and to reduce the risk of spurious findings due to population structure.

Genome-wide association studies (GWAS) have successfully identified thousands of common variants associated with complex traits[1–4], but translating these discoveries into mechanistic insights has been challenging. 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, such 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 the 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 more than 1–2×, these reads can be used to discover and genotype SNPs across the genome[18,19], and, with off-target depth of more than 0.2–0.5×, 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. It is thus difficult in most targeted sequencing experiments to accurately call off-target genotypes. In addition, off-target sequence reads are distributed sparsely and randomly across each genome, such 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

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [2]Department of Biostatistics, Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [3]Department of Ophthalmology, University of Pennsylvania Medical School, Philadelphia, Pennsylvania, USA. [4]Division of Epidemiology and Clinical Research, National Eye Institute, Bethesda, Maryland, USA. [5]Department of Ophthalmology, University of Michigan Kellogg Eye Center, Ann Arbor, Michigan, USA. [6]Full lists of members and affiliations appear in the **Supplementary Note**. [7]Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA. [8]Neurobiology-Neurodegeneration & Repair Laboratory, National Eye Institute, Bethesda, Maryland, USA. [9]Department of Psychiatry, University of Michigan Medical School, Ann Arbor, Michigan, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to C.W. (chaolong@umich.edu) or G.R.A. (goncalo@umich.edu).

Received 8 August 2013; accepted 21 February 2014; published online 16 March 2014; doi:10.1038/ng.2924

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, by ignoring loci with no data and increasing the weights of loci that have higher coverage.

Here we propose a new 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 the 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 backgrounds 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 increasing or decreasing the weighting of sites according to their coverage and the information they contain about ancestry. The process is illustrated in **Figure 1** and is described 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 macular degeneration[28]. Our results show that our method can accurately infer worldwide continental ancestry or even fine-scale ancestry within Europe with extremely low off-target coverage (~0.001× for worldwide ancestry and ~0.10× for European ancestry). We have implemented our method in the publicly available LASER (Locating Ancestry from SEquence Reads) software.

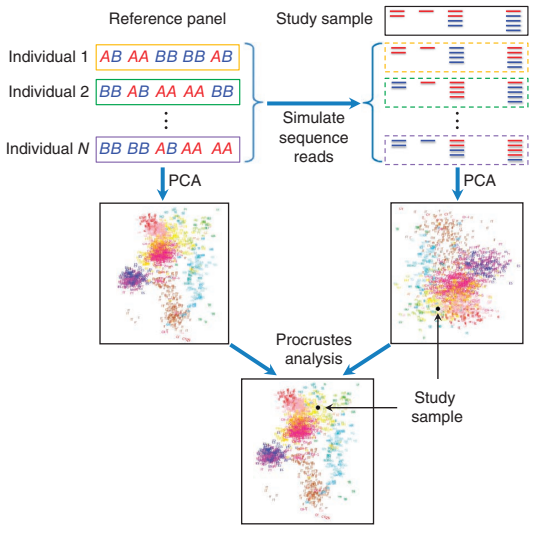


Figure 1 Graphic illustration of the LASER method.

RESULTS

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 HGDP[23] and used their array genotypes at 632,958 loci as templates to simulate sequence data (Online Methods). We simulated multiple sequence data sets with mean coverage ranging from 0.001 to 0.25×. 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 HGDP (**Fig. 2**): PC1 and PC2 separate major continental groups in the Old World, whereas PC3 and PC4 further separate Native American and Oceanian populations, respectively. We applied LASER to each simulated sequence data set to estimate the ancestry coordinates of the test individuals in the reference PCA space. We assessed accuracy by comparing the ancestry estimates derived with LASER to the PCA coordinates of the test individuals based on their original SNP genotypes using the squared Pearson's correlation r^2 along each principal component and the Procrustes similarity score t_0 (Online Methods). Our results show consistently high accuracy across all simulated data sets (**Fig. 2** and **Supplementary Table 1**). When the simulated coverage was 0.001× (corresponding to ~630 loci covered with ≥ 1 read), r^2 ranged from 0.7396 for PC4 to 0.9506 for PC1, and the Procrustes similarity score t_0 was 0.9508. Although the patterns were a bit fuzzy, major continental groups were well separated at 0.001× coverage (**Fig. 2b**). Accuracy increased with coverage: when the coverage was 0.10×, the estimated coordinates were almost identical to the coordinates estimated using a GWAS SNP panel with $t_0 = 0.9993$ (**Fig. 2d** and **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 0.01 to 0.40× coverage). The remaining 1,000 POPRES samples of European ancestry were used to construct the reference PCA space. We focused on the top two principal components of the POPRES reference panel, which mirror the geographic map of Europe[24] (**Fig. 3a**). Compared to estimates of worldwide continental ancestry, much higher coverage was required to identify the more subtle differences in population structure within Europe (**Fig. 3** and **Supplementary Table 2**). With an average coverage of 0.01×, samples clumped in the center of the reference PCA space (r^2 = 0.5687 for PC1 and 0.0108

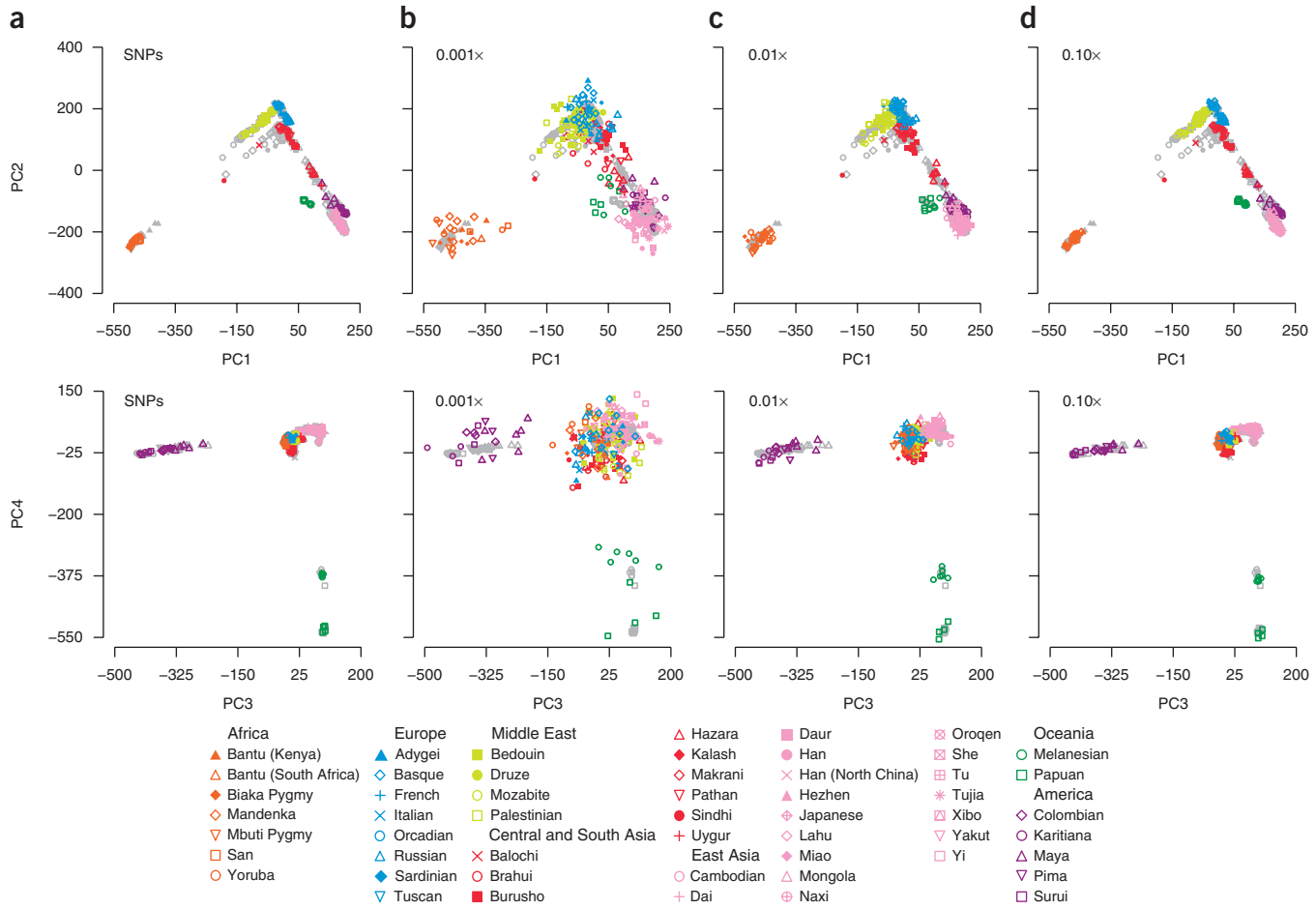


Figure 2 Estimation of worldwide continental ancestry. We randomly selected 238 individuals from HGDP as the testing set (colored symbols), and the remaining 700 HGDP individuals were used as the reference panel (gray symbols). The upper row shows PC1 and PC2, and the lower row shows PC3 and PC4. (**a**) Results based on SNP genotypes. (**b**) Results based on simulated sequence data at 0.001× coverage. Procrustes similarity to the SNP-based coordinates $t_0 = 0.9508$. (**c**) Results at 0.01× coverage ($t_0 = 0.9949$). (**d**) Results at 0.10× coverage ($t_0 = 0.9993$).

for PC2; t_0 = 0.4786; **Fig. 3b**). As coverage increased to 0.05× (**Fig. 3c**), we were 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 was evident when coverage was >0.10× (**Fig. 3d–f**), with t_0 increasing from 0.9126 (0.10× coverage) to 0.9764 (0.40× 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 next 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 7 worldwide populations (Online Methods and **Supplementary Table 3**). We used all 938 HGDP individuals to construct the reference PCA space. Average off-target sequencing coverage for the 410 samples was ~0.096× at the 632,958 SNP loci genotyped for HGDP (**Supplementary Fig. 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 Project exon sequencing project. Coordinates estimated from off-target sequence reads were 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, respectively; **Supplementary Fig. 2**). Even when focusing on 103 samples whose off-target coverage was below 0.06×, we still obtained t_0 = 0.9938 (r^2 = 0.9930, 0.9884, 0.9012 and 0.6811 for PC1 to PC4, respectively; **Supplementary Table 4**). Surprisingly, t_0 for the 103 samples with the highest off-target coverage (from 0.10 to 0.55×) was slightly lower than t_0 for the groups with lower coverage (**Supplementary Table 4**). This finding 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 8 susceptibility loci and 2 candidate regions for macular degeneration[28]. Samples included 2,362 macular degeneration cases, 789 controls, 2 samples with unknown phenotype, and 1 European (CEU) and 1 Yoruba (YRI) nuclear family selected among the HapMap Project samples (each nuclear family included a mother, a 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 0.224× across the 632,958 loci in HGDP and 0.241× across the 318,682 loci in POPRES (**Supplementary Fig. 3**).

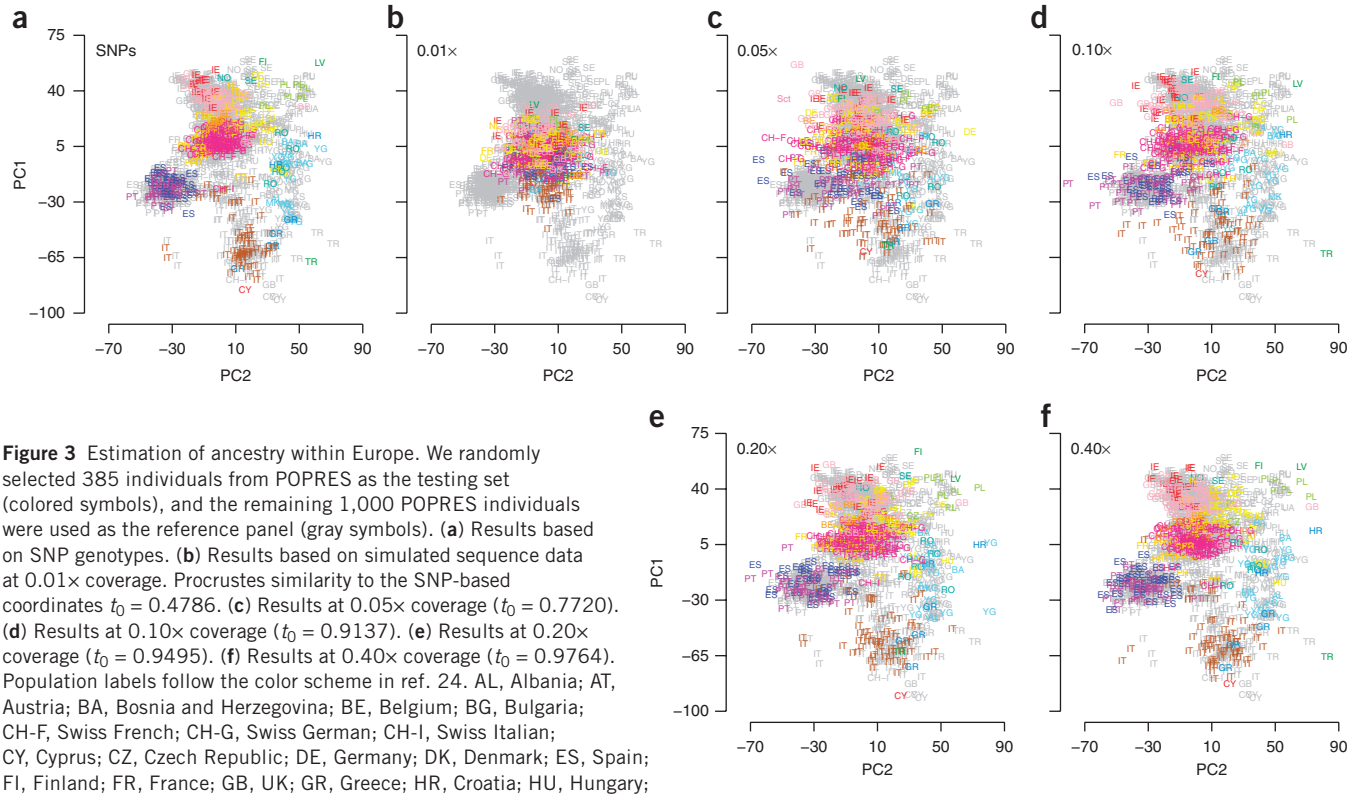


Figure 3 Estimation of ancestry within Europe. We randomly selected 385 individuals from POPRES as the testing set (colored symbols), and the remaining 1,000 POPRES individuals were used as the reference panel (gray symbols). (**a**) Results based on SNP genotypes. (**b**) Results based on simulated sequence data at 0.01× coverage. Procrustes similarity to the SNP-based coordinates $t_0 = 0.4786$. (**c**) Results at 0.05× coverage ($t_0 = 0.7720$). (**d**) Results at 0.10× coverage ($t_0 = 0.9137$). (**e**) Results at 0.20× coverage ($t_0 = 0.9495$). (**f**) Results at 0.40× coverage ($t_0 = 0.9764$). Population labels follow the color scheme in ref. 24. AL, Albania; AT, Austria; BA, Bosnia and 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, UK; 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.

When using HGDP as the reference panel, the two trios were placed in 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 individuals of European or Middle Eastern ancestry, 73 aligned between Africans and Europeans (likely corresponding to African-American samples), 5 aligned between Europeans and Native Americans, 3 clustered with Central and South Asians, and 3 clustered with East Asians (**Supplementary Fig. 4a**,**b**). We then used the POPRES reference panel to dissect the population structure of the samples in the cluster with European and Middle Eastern ancestry. Our results showed that, although most of these samples had northern European ancestry, many other samples formed a small cluster around southern Europe (**Supplementary Fig. 4c**,**d**). For 931 of the sequenced AMD cases and controls, GWAS array genotype data were also available[30]. For these samples, results based on the off-target reads were well matched with the coordinates estimated using SNP genotypes, in both the HGDP PCA space ($t_0 = 0.9068$; **Supplementary Fig. 5**) and the POPRES PCA space ($t_0 = 0.9209$; **Supplementary Fig. 6**). Accuracy was greater for samples with higher off-target coverage (**Supplementary Table 5**).

Evaluation using exome sequence data

The previous experiments examined situations in which 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 Finland–United States Investigation of NIDDM Genetics (FUSION)[31] (**Supplementary Table 6**) who have been extensively characterized as part of the Genetics of Type 2 Diabetes (GoT2D) Study by genotyping on the Omni 2.5M array, by deep exome sequencing (~96× depth, 0.69 million variants) and by low-pass whole-genome sequencing (~5× 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 on this reference map, using ancestry estimates derived from whole-genome sequencing data as a gold standard. Ancestry estimates derived using our method were 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 separated individuals born in the different provinces of Finland (**Fig. 4**). This improved separation of individuals originating from different parts of Finland was highlighted when variance in PCA coordinates was decomposed into within-province and between-province components: between-province variation in coordinates increased from 48% when using exome genotypes to 64% using our method (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 evident, 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 sequencing depths of ~88.9× and ~1.0×, 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 Fig. 7a**). In contrast, using off-target reads, our method provided accurate estimates of

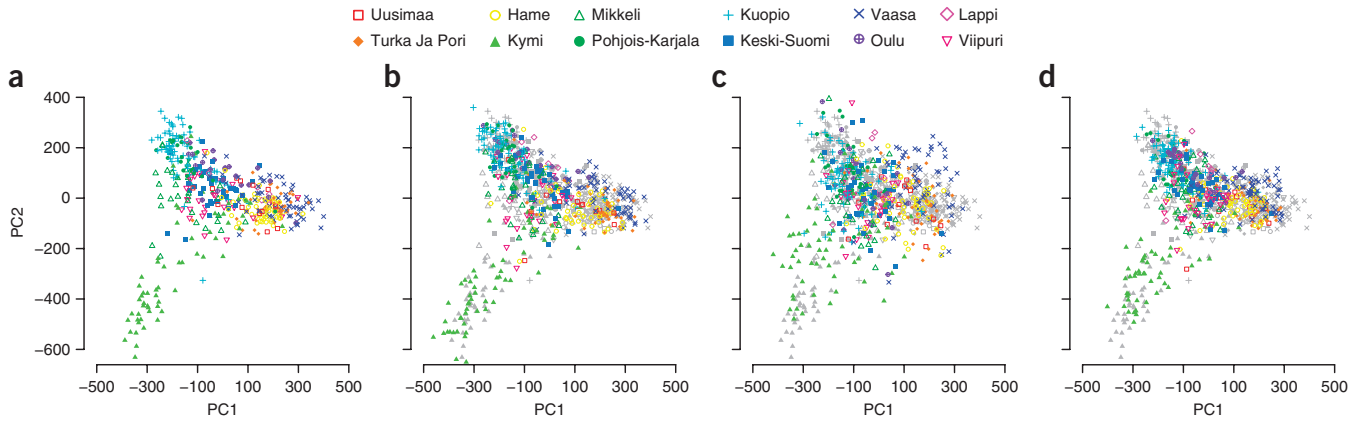


Figure 4 Estimation of fine-scale ancestry within Finland. (**a**) Reference PCA map based on the integrated whole-genome genotypes of 470 reference individuals. The proportion of among-population variance in the PCA map $\psi = 0.6623$. (**b**) Estimation of ancestry for 471 test individuals based on integrated whole-genome genotypes ($\psi = 0.6685$). Reference individuals are indicated by gray symbols. (**c**) Estimation of ancestry for test individuals based on exome sequencing genotypes ($\psi = 0.4849$). Compared to **b**, Procrustes similarity $t_0 = 0.8263$, $r^2 = 0.9411$ 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 **b**, $t_0 = 0.9763$, $r^2 = 0.9778$ and 0.9259 for PC1 and PC2, respectively. Mean coverage is ~96× and ~0.89× for on-target and off-target regions, respectively.

individual ancestry ($t_0 = 0.9467$; $r^2 = 0.9744$ for PC1 and 0.7640 for PC2; **Supplementary Fig. 7b**). Incorporating both on-target and off-target reads, our ancestry estimates improved further ($t_0 = 0.9669$; $r^2 = 0.9804$ for PC1 and 0.8610 for PC2; **Supplementary Fig. 7c**). We also note that, compared to the simulations in **Supplementary Table 2**, ancestry estimates appeared 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 and, second (and more notably), because there is great variation in per-individual off-target coverage in the exome sequencing project samples (ranging from 0.49 to 4.70× 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[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 $\lambda_{common} = 1.326$ for common variants (MAF ≥ 0.05) and $\lambda_{low\text{-}freq} = 1.267$ for low-frequency variants (0.01 ≤ MAF < 0.05) (**Table 1**). When our estimated principal components were used as covariates in association analysis, evidence for stratification was greatly reduced, resulting in $\lambda_{common} = 0.992$ and $\lambda_{low\text{-}freq} = 0.996$ at 0.10× coverage ($t_0 = 0.9993$; $r^2 = 0.9986$ for PC1 and 0.9985 for PC2) and in $\lambda_{common} = 0.991$ and $\lambda_{low\text{-}freq} = 0.998$ with more modest 0.005× 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 $\lambda_{common} = 1.011$ and $\lambda_{low\text{-}freq} = 1.013$ at 0.10× coverage and with $\lambda_{common} = 1.041$ and $\lambda_{low\text{-}freq} = 1.045$ at 0.005× coverage (**Table 1**). We next explored more challenging sampling strategies in which all cases were sampled from one or two 8 × 8 grids (**Supplementary Fig. 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 that off-target coverage was greater than 0.10× (**Supplementary Table 7**). This observation is noteworthy, as it suggests that, although 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.

Table 1 Evaluation of corrections for stratification in simulated case-control data with 900 and 600 cases sampled, respectively, from 2 halves of a 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}	$\lambda_{low\text{-}freq}$	λ_{common}	$\lambda_{low\text{-}freq}$
Uncorrected	–	–	–	1.326	1.264	1.326	1.267
SNP-based PCs	1	1	1	0.991	0.995	0.996	0.998
0.20×	0.9996	0.9993	0.9993	0.992	0.996	1.009	1.019
0.15×	0.9995	0.9990	0.9991	0.992	0.995	1.007	1.005
0.10×	0.9993	0.9986	0.9985	0.992	0.996	1.011	1.013
0.05×	0.9985	0.9972	0.9968	0.992	0.995	1.018	1.015
0.01×	0.9925	0.9851	0.9851	0.991	0.995	1.036	1.034
0.005×	0.9853	0.9711	0.9706	0.991	0.998	1.041	1.045
0.001×	0.9317	0.8635	0.8723	0.994	1.000	1.076	1.084

Here λ_{common} is the genomic inflation factor calculated on the basis of 625,481 common variants (MAF ≥ 0.05), and $\lambda_{low\text{-}freq}$ is the inflation factor calculated on the basis of 374,519 low-frequency variants (0.01 ≤ MAF < 0.05). Procrustes similarity scores and squared correlations were calculated by comparing sequence-based principal components to SNP-based principal components for the 1,500 cases. For uncorrected results, we used logistic regression (regression-based analyses) and Cochran-Armitage trend tests (matching-based analyses). Two approaches to correct for stratification were examined: (i) including estimated ancestry coordinates as covariates in logistic regression and (ii) identifying one-to-one ancestry-matched case-control pairs for Cochran-Mantel-Haenszel tests (treating each matched pair as a stratum).

DISCUSSION

We show that the genetic ancestry of an individual can be accurately estimated using

the off-target sequence reads that are a byproduct of most targeted sequencing studies. With off-target reads corresponding to 0.001× coverage of the genome, worldwide continental ancestry can be reconstructed, and, with off-target reads corresponding to 0.10× coverage, ancestry can be estimated within Europe. Because Europe is the continent with the most homogeneous genetic variation[34], we expect that LASER can be used to infer fine-scale structure within other continents when appropriate reference panels are available. A key ingredient for the 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 this would increase the number of overlapping sites between sequenced samples and the 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 recommend always 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 that these assumptions will have only a minor impact on our results. For example, although less uniform distributions of coverage might require slight increases in depth for accurate estimation of ancestry, such effects could be counteracted 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** and **9**).

We foresee several potential enhancements to our approach. For example, as different runs of our method show small stochastic variation in the placement of each individual, we expect that repeated analysis of the same sample can improve results (**Supplementary Fig. 9**), particularly when coverage is very low or when trying to place samples on a European ancestry map (or another map where differences between populations are small). Our simulations show that averaging results over ten repeated runs for a sample sequenced with 0.10× coverage produces ancestry placements within the map of Europe that are almost as accurate as when a single placement is generated on the basis of a sample sequenced with 0.20× 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 predefined reference ancestry space, does not capture artifacts due to experimental batch effects or close relatedness of samples, thus allowing 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, summaries of sequencing depth, read length or even locus-by-locus coverage information in an additional set of principal components)[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 of the analysis for 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 (for example, ~1.3 min when $N = 1{,}000$, $L = 318{,}682$ 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 and 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 principal components[21] or to use nonlinear techniques, such as the kernel smoothing methods, to correct for structure based on our estimated principal components[44]. The diverse ancestry observed among sequenced AMD samples further illustrates the importance and usefulness 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 to increase sample size and statistical power in a targeted sequencing study of macular degeneration. In this way, we were able to match by ancestry potential control samples from public resources with sequenced cases, enabling the discovery of a rare variant, encoding 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.

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

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank investigators from the FUSION study and the GoT2D Sequencing Project for generously sharing whole-genome and deep exome sequence data for 941 individuals before publication and the D2D, Finrisk 2002, Health 2000, Action LADA and Saviatipale studies for providing some of the FUSION-sequenced DNA. We thank J.Z. Li for his assistance with the HGDP data set, H. Stringham and A. Locke for assistance with the FUSION data set and

M. 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 the US National Institutes of Health (DK062370, HG000376, HG005552, HG006513, EY022005, HG007022, HG005855, HG003079, CA076404 and CA134294) and by the National Eye Institute Intramural Research Program.

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 manuscript. All authors reviewed, revised and contributed critical feedback to the manuscript and presentation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
2. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
3. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
4. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
5. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010).
6. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
7. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
8. Shen, P. *et al.* High-quality DNA sequence capture of 524 disease candidate genes. *Proc. Natl. Acad. Sci. USA* **108**, 6549–6554 (2011).
9. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
10. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
11. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).
12. Raychaudhuri, S. *et al.* A rare penetrant mutation in *CFH* confers high risk of age-related macular degeneration. *Nat. Genet.* **43**, 1232–1236 (2011).
13. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
14. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
15. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
16. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
17. Clark, M.J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
18. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
19. Le, S.Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **21**, 952–960 (2011).
20. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).
21. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
22. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
23. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
24. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
25. Schönemann, P.H. & Carroll, R.M. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* **35**, 245–255 (1970).
26. Wang, C. *et al.* Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat. Appl. Genet. Mol. Biol.* **9**, 13 (2010).
27. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
28. Zhan, X. *et al.* Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat. Genet.* **45**, 1375–1379 (2013).
29. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
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. USA* **107**, 7401–7406 (2010).
31. Valle, T. *et al.* Mapping genes for NIDDM. Design of the Finland–United States Investigation of NIDDM Genetics (FUSION) Study. *Diabetes Care* **21**, 949–958 (1998).
32. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
33. Guan, W., Liang, L., Boehnke, M. & Abecasis, G.R. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet. Epidemiol.* **33**, 508–517 (2009).
34. Wang, C., Zöllner, S. & Rosenberg, N.A. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet.* **8**, e1002886 (2012).
35. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
36. Miclaus, K., Wolfinger, R. & Czika, W. SNP selection and multidimensional scaling to quantify population structure. *Genet. Epidemiol.* **33**, 488–496 (2009).
37. Zhu, C. & Yu, J. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* **182**, 875–888 (2009).
38. Yang, W.Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.* **44**, 725–731 (2012).
39. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
40. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22**, 1525–1532 (2012).
41. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
42. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
43. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
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.* **24**, 44–56 (2003).

ONLINE METHODS

All experiments relied on preexisting 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 four steps: (step 1) PCA on reference genotypes to define a reference ancestry space; (step 2) simulation of sequence data for reference individuals, matching the coverage of each study sample; (step 3) PCA on combined sequence data; and (step 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 matrix $\boldsymbol{G}_{ij} = 0, 1, 2$ represents the number of reference alleles at locus $j = 1, \ldots, L$ for individual $i = 1, \ldots, N$. We let μ_j and σ_j represent column means and s.d., respectively, 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 principal component is given by $\lambda_k^{1/2}\vec{v}_k$, where λ_k is the kth eigen value of $\boldsymbol{M}$ and $\vec{v}_k$ is the corresponding eigen vector. Coordinates of the top K principal components for reference individuals are stored in the $N \times K$ matrix $\boldsymbol{Y}$.

Simulating sequence data for reference individuals. We simulate sequence data for reference individuals, matching the coverage pattern of study samples. Suppose that 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 follows:

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

Here ε is the estimated sequencing error rate per base ($\varepsilon = 0.01$ 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 principal components for reference individuals in the $N \times K$ matrix $\boldsymbol{X}$ and for the study sample in the 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 less reliable $\boldsymbol{Z}^*_h$.

Genetic data. *Genotype data*. We used HGDP[23] and POPRES[24,45] genotypes to define reference coordinate spaces. The HGDP data set included 632,958 autosomal SNPs and 938 unrelated individuals from 53 worldwide populations[23]. Our POPRES subset contained 318,682 autosomal SNPs and 1,385 individuals from 37 European populations[24]. For both data sets, we preprocessed data as summarized in **Supplementary Figure 10**, excluding SNPs that had different alleles in 1000 Genomes Project data and dbSNP, had >2 alleles, had ambiguous strand or were missing from dbSNP (version 135).

We also analyzed genotypes from the HapMap Project[29] and AMD GWAS[30]. In the HapMap data set, we focused on 410 individuals who 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 Fig. 11**).

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

Exome sequence data. The GoT2D Study of Type 2 Diabetes characterized 941 Finnish individuals from FUSION, with exome sequencing at a mean depth of ~96×, whole-genome sequencing at a mean depth of ~5× and genotyping on the Illumina Omni2.5 BeadChip. Our analyses focused on autosomal biallelic SNPs with missingness of <5%, Hardy-Weinberg equilibrium $P > 1 \times 10^{-6}$ and MAF > 0.01. After quality control[47], whole-genome analyses included 8,447,085 SNPs and exome analyses included 95,741 SNPs (of which 94,423 overlapped).

Preprocessing 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 of <30 and bases with Phred quality of <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 were used to define reference coordinate spaces. We first simulated Poisson coverage with means between 0.001 and 0.40 and then sampled reference alleles using equation (1) (**Supplementary Tables 1** and **2**). We next repeated the simulation using coverage patterns from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project[32]. Among randomly selected NHLBI samples, mean exome coverage was ~88.9× and mean off-target coverage was ~1.0×.

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 onto 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 in HapMap and HGDP. For the AMD samples, we used 45,700 SNPs shared by the HGDP, POPRES and AMD data sets.

We used the squared Pearson's 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 using Procrustes analysis[26] to translate between sequence- and SNP-derived coordinates for test samples.

Fine-scale population structure. Each FUSION sample could be assigned to 1 of 12 subpopulations according to birth province. We split each subpopulation into 2 groups, resulting in 470 reference individuals and 471 test individuals (**Supplementary Table 6**). We constructed a reference PCA map on the basis of whole-genome sequence results. We placed test individuals onto this map using (i) whole-genome genotypes, (ii) genotypes across loci overlapping in exome and whole-genome data and (iii) off-target reads generated during exome sequencing (~0.89× off-target depth).

To evaluate how well the three analyses captured population structure, we defined statistic ψ as the proportion of between-population variance in PCA coordinates. We used 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 was defined as follows:

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

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 that population structure is better captured.

doi:10.1038/ng.2924

Simulated case-control studies. We simulated[51] 20,000 diploid individuals evenly distributed along a 20×20 lattice, each genotyped at 1 million independent biallelic SNPs with MAF ≥ 0.01. The scaled migration rate between neighboring lattice points was $M = 10$, as suggested in ref. 16, to mimic population structure within Europe. In each lattice point, we assigned 3 individuals to a reference set and, among the remaining individuals, 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 versus 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 Fig. 8**). In these additional scenarios, we sampled 1,280 cases and 1,280 controls.

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

45. Nelson, M.R. *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* **83**, 347–358 (2008).
46. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
47. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469 (2012).
50. Holsinger, K.E. & Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* **10**, 639–650 (2009).
51. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
52. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

SUPPLEMENTARY MATERIALS

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] & 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

[6] Full lists of members and affiliations appear in the **Supplementary Note.**

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

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

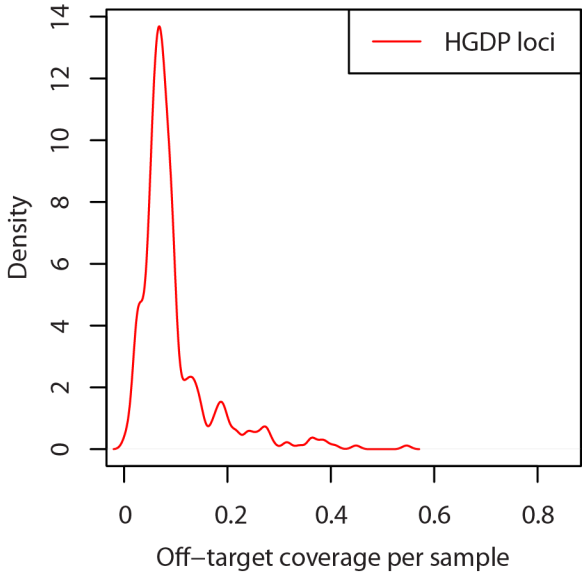
[9] Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI 48109

[*] C.W. and X.Z. are joint first authors.

[§] Correspondence: chaolong@umich.edu (C.W.), goncalo@umich.edu (G.R.A.)

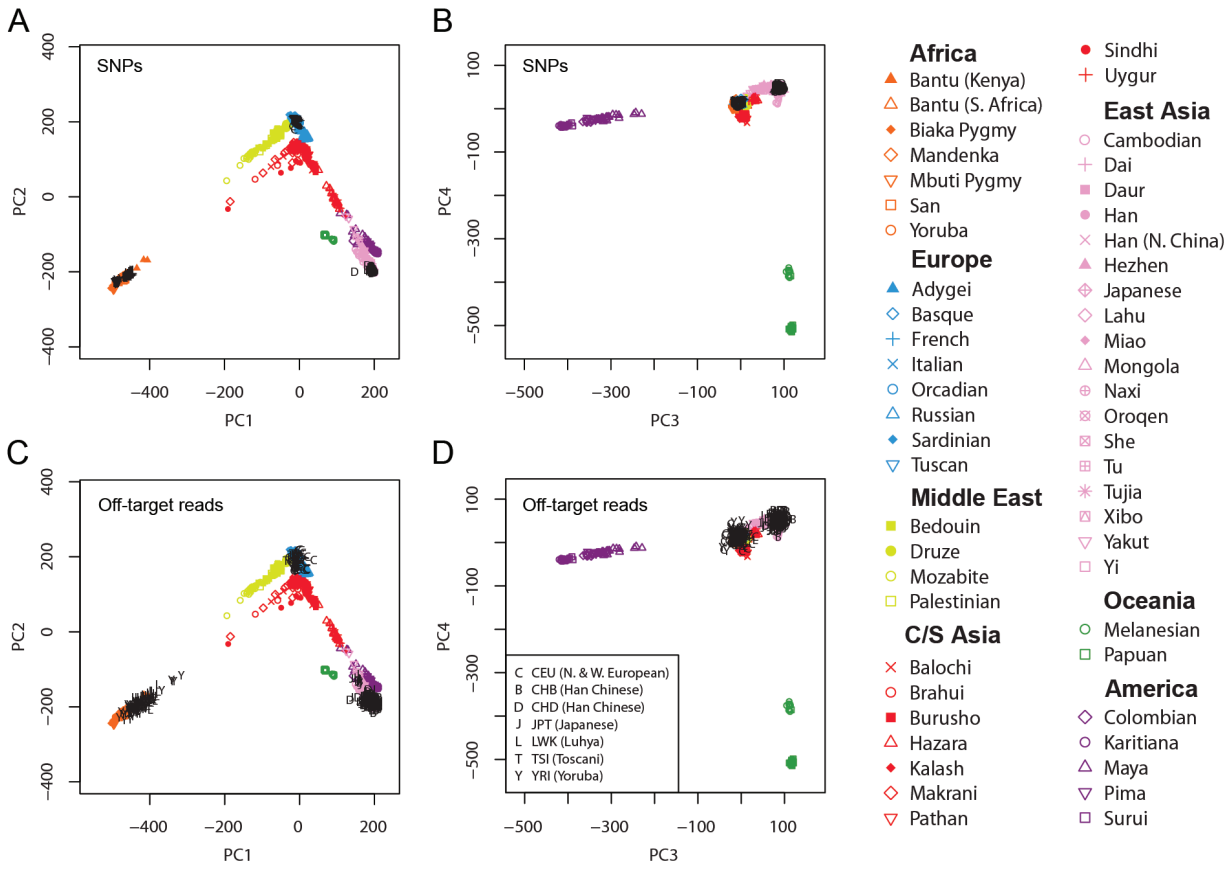
Nature Genetics: doi:10.1038/ng.2924

SUPPLEMENTARY FIGURES



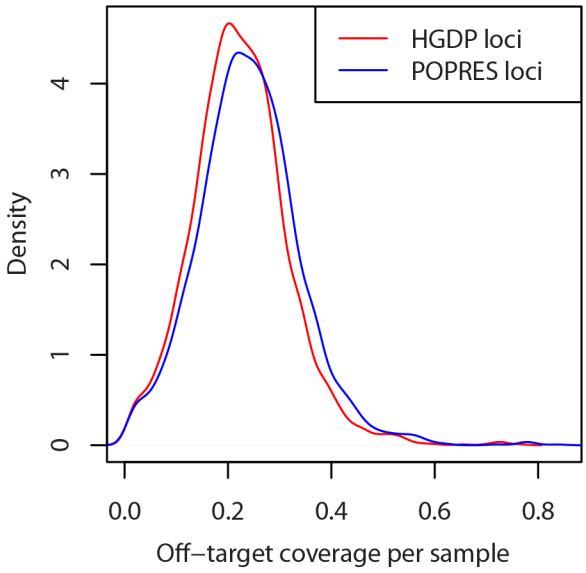
Supplementary Figure 1. Off-target coverage for 410 samples from the 1000 Genomes exon project. The off-target coverage for each sample is calculated by averaging across 632,958 loci in the HGDP. For 270 loci that appear in the targeted regions, we set the coverage at these loci to 0 for all samples. Mean off-target coverage is 0.096X across the HGDP loci.

Nature Genetics: doi:10.1038/ng.2924



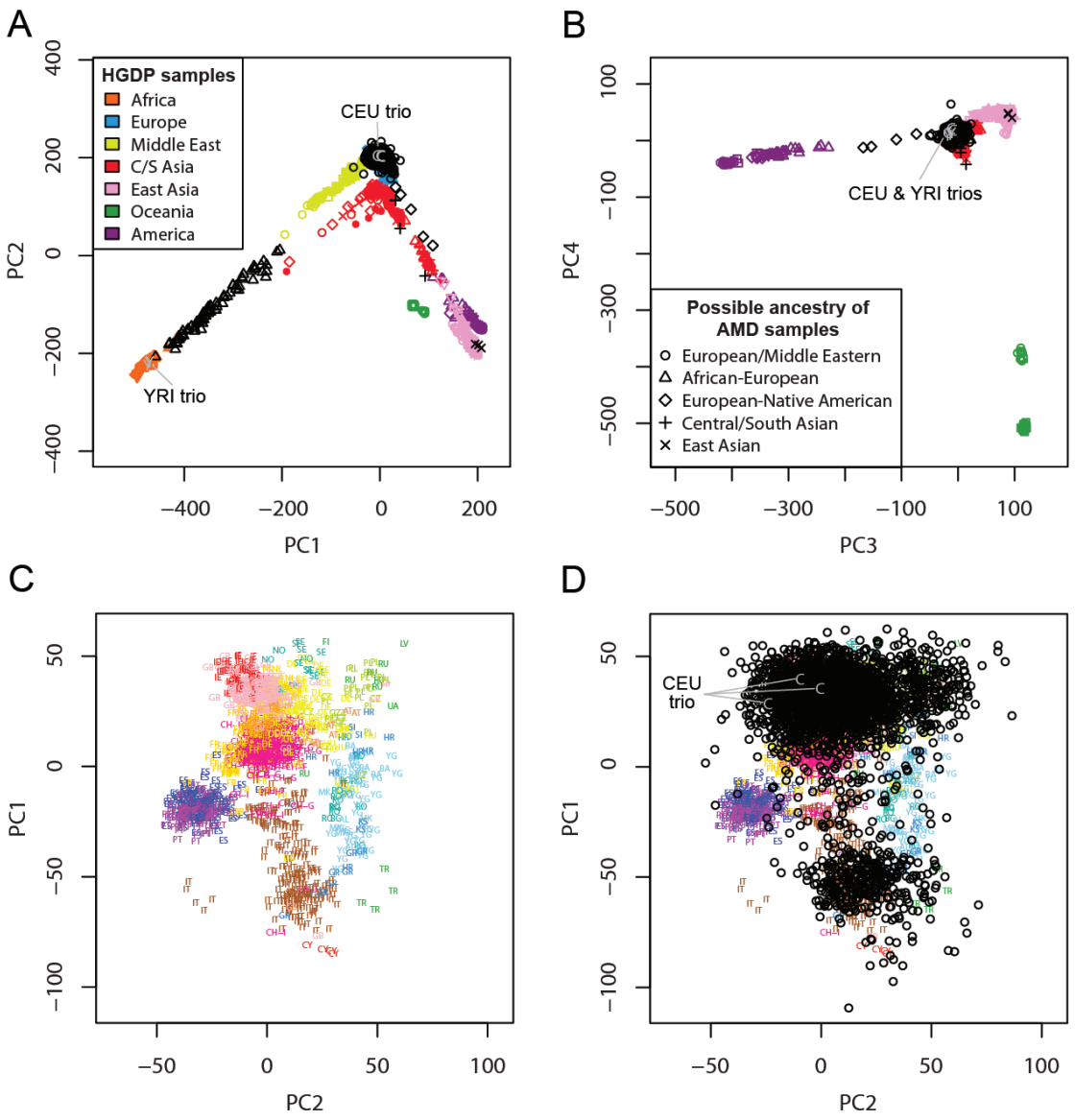
Supplementary Figure 2. Estimation of worldwide ancestry for 410 samples in the 1000 Genomes exon project. The SNP genotypes of these samples are from the HapMap Project. We used all HGDP individuals as the reference panel, as labeled by colored points. (A,B) Results based on SNPs that were genotyped in both HapMap 3 and HGDP. (C,D) Results based on off-target sequence data. The Procrustes similarity to the SNP-based coordinates is $t_0 = 0.9955$. $r^2 = 0.9950, 0.9871, 0.9439,$ and 0.7747 for PC1, PC2, PC3, and PC4, respectively.

Nature Genetics: doi:10.1038/ng.2924



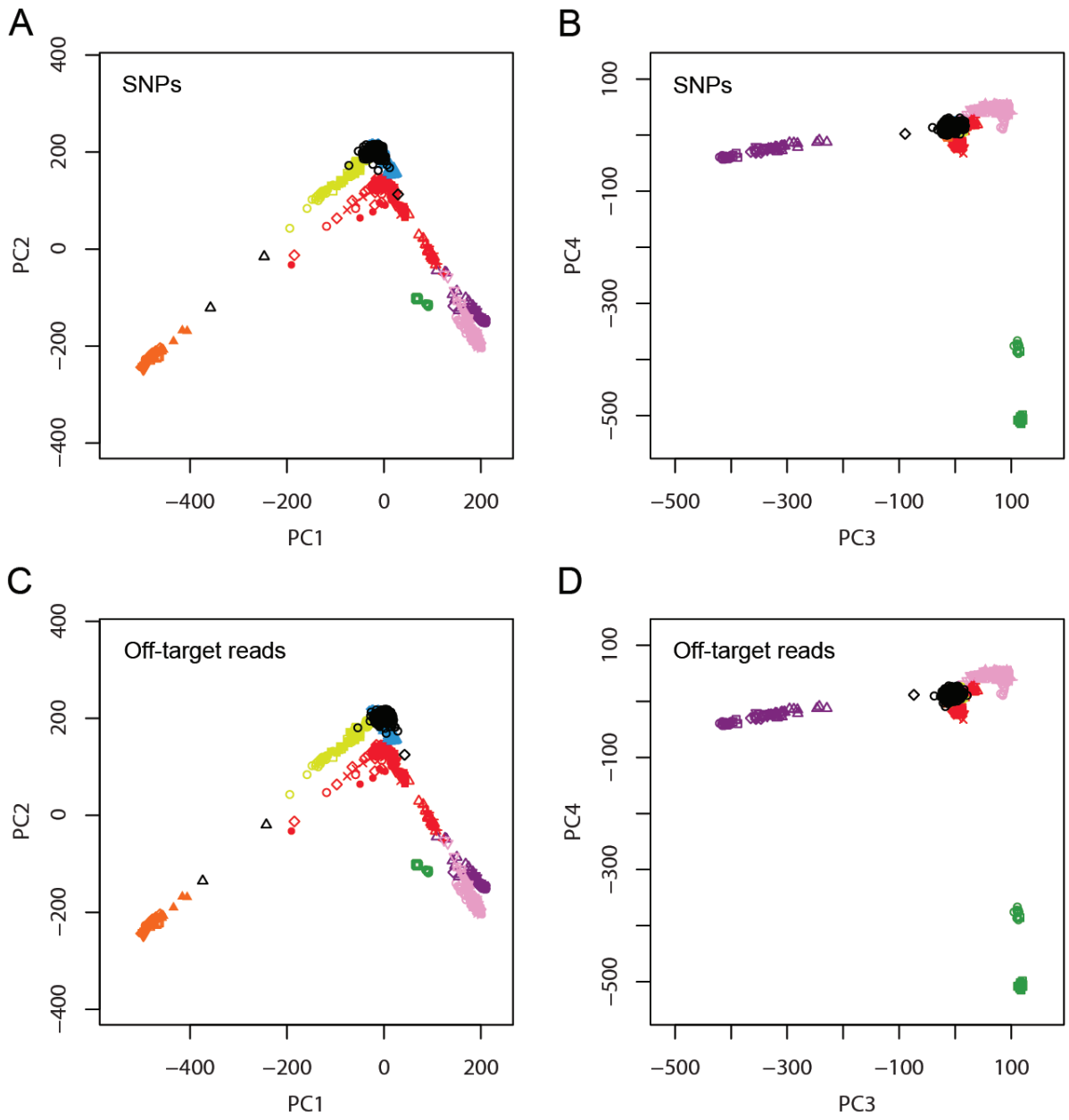
Supplementary Figure 3. Off-target coverage for 3,159 samples from the AMD study. The red line indicates off-target coverage averaged across 632,958 loci included in HGDP. The blue line indicates off-target coverage averaged across 318,682 loci that are included in POPRES. For loci that appear in the targeted regions, we set the coverage at these loci to 0 for all samples, including 215 loci in HGDP and 113 loci in POPRES. Mean off-target coverage is 0.224X across the HGDP loci and 0.241X across the POPRES loci.

Nature Genetics: doi:10.1038/ng.2924



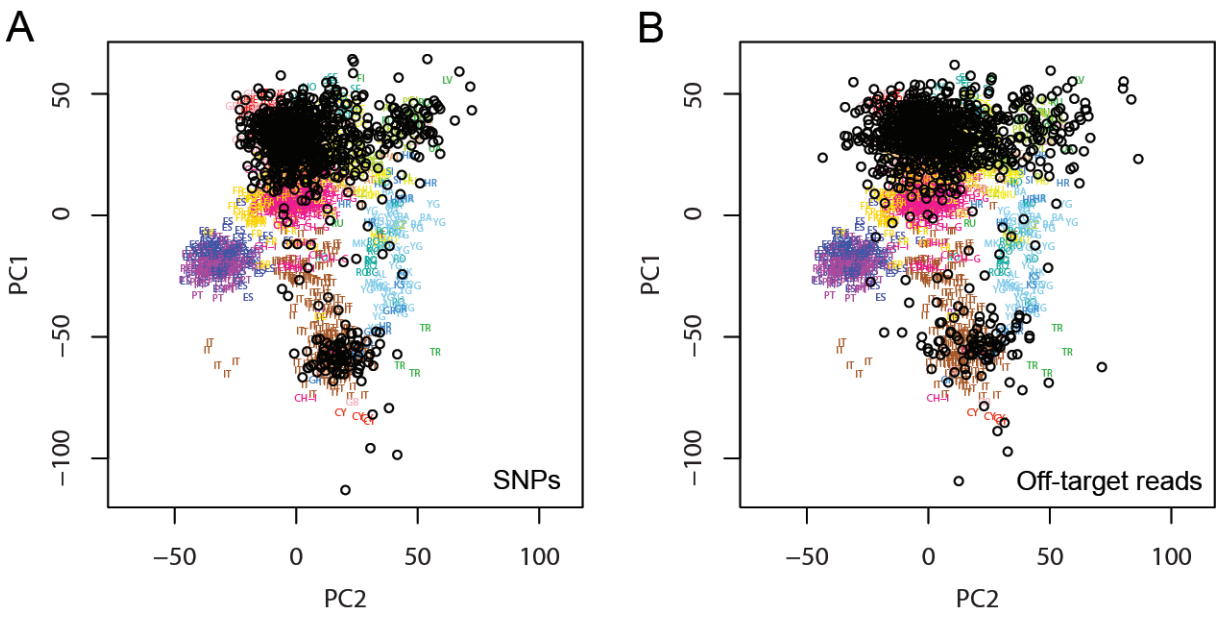
Supplementary Figure 4. Estimation of ancestry for 3,159 samples in the AMD targeted sequencing dataset. (A,B) Results based on the HGDP reference panel, whose colors and symbols follow **Supplementary Figure 2**. AMD samples are displayed in black, with different symbols representing possible ancestries based on their estimated PC coordinates. Two HapMap trios are labeled in gray. (C,D) Results based on the POPRES reference panel. Panel C displays PC1 and PC2 of POPRES; panel D displays 3,072 AMD samples on top of the POPRES samples. These samples are possibly Europeans or Middle Eastern as indicated in panels A and B. Population labels for the POPRES samples 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.

Nature Genetics: doi:10.1038/ng.2924



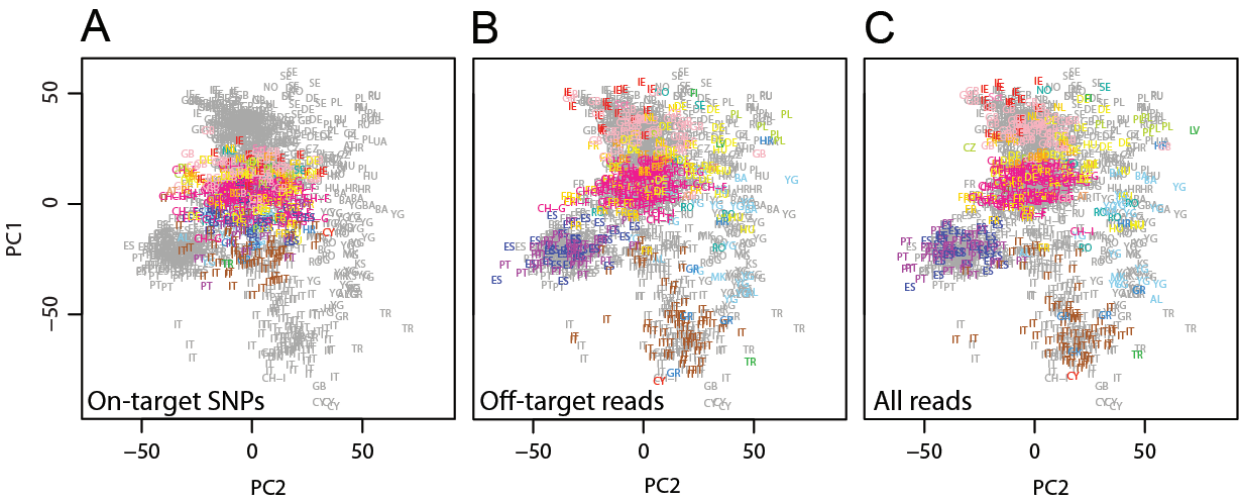
Supplementary Figure 5. Sequence-based coordinates and SNP-based coordinates for 931 AMD samples when using the HGDP reference panel. Colors and symbols for HGDP and AMD samples follow **Supplementary Figure 2**. (A,B) Results based on 45,700 SNPs that are shared by HGDP, POPRES and AMD SNP datasets. (C,D) Results based on off-target sequence data. The Procrustes similarity between SNP- and sequence-based coordinates is $t_0 = 0.9068$. $r^2 = 0.9104$, 0.8881, 0.6031, and 0.1828 for PC1, PC2, PC3, and PC4, respectively.

Nature Genetics: doi:10.1038/ng.2924



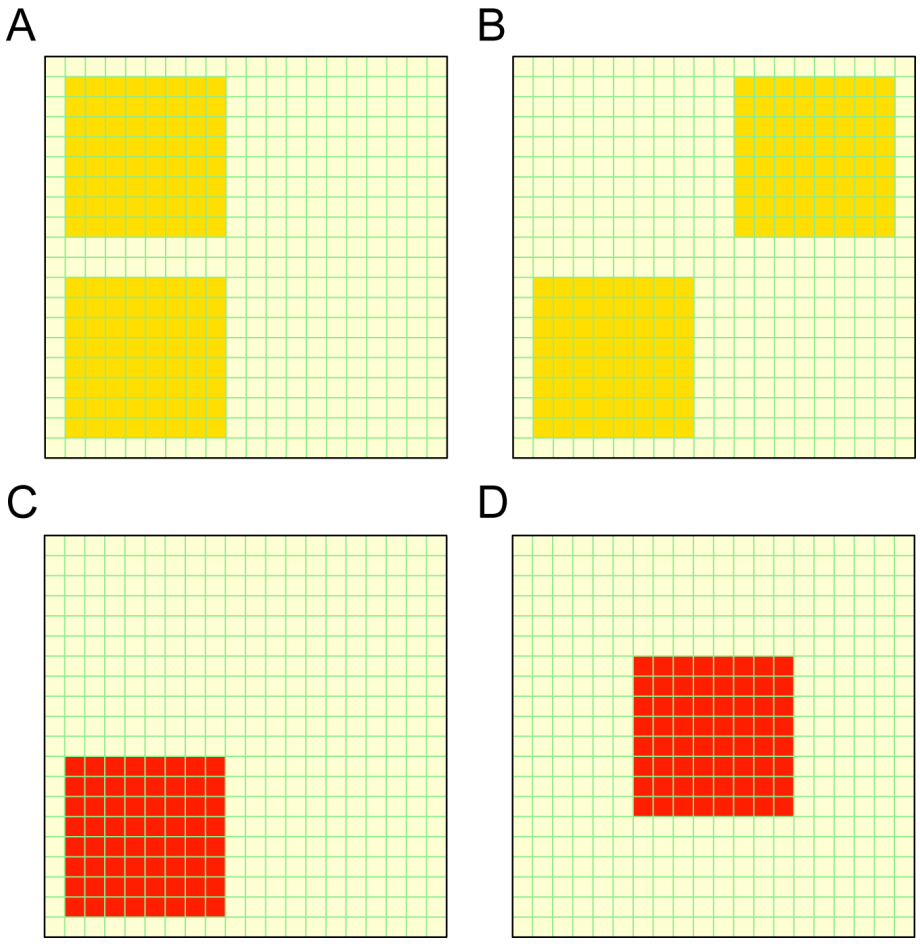
Supplementary Figure 6. Sequence-based coordinates and SNP-based coordinates for AMD samples when using the POPRES reference panel. We only included 928 AMD samples whose genotype data are available and who might be Europeans or Middle Eastern according to results in **Supplementary Figure 5**. (A) Results based on 45,700 SNPs that are shared by HGDP, POPRES, and AMD SNP datasets. (B) Results based on off-target sequence data. The Procrustes similarity between results in panels A and B is $t_0 = 0.9209$. $r^2 = 0.9557$ and 0.6389 for PC1 and PC2, respectively.

Nature Genetics: doi:10.1038/ng.2924



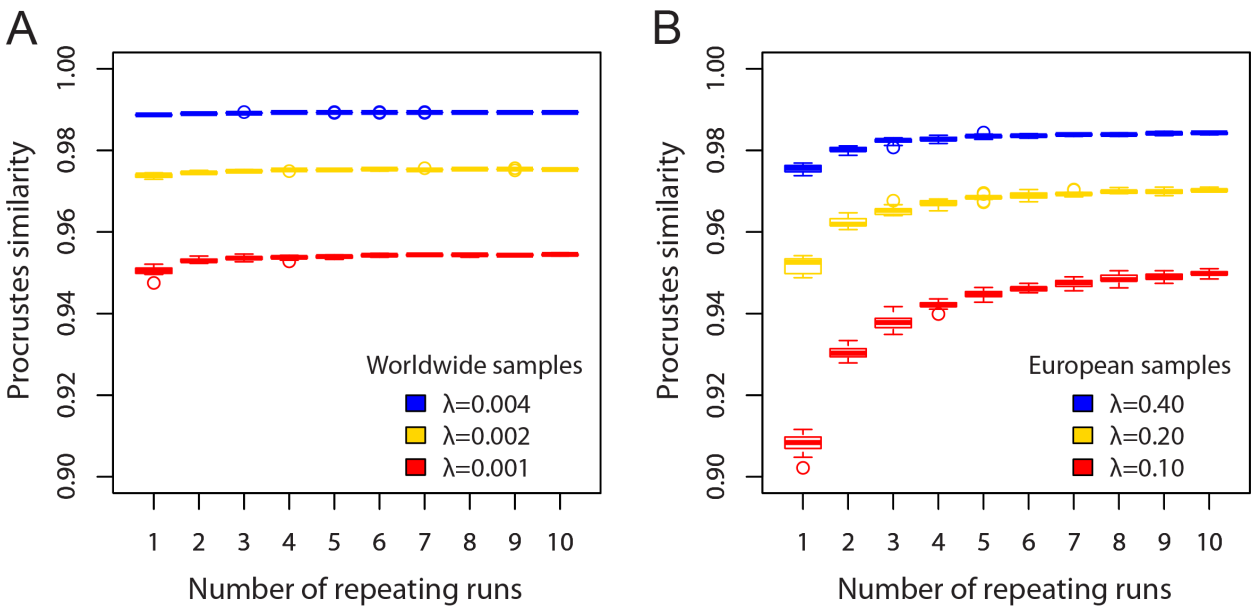
Supplementary Figure 7. **Results for simulated exome sequencing data for 385 POPRES samples.** (A) Coordinates estimated from SNP genotypes at 2,547 on-target loci. The Procrustes similarity to the SNP-based coordinates in **Figure 3A** is $t_0 = 0.5031$. (B) Coordinates estimated based on off-target sequence reads ($t_0 = 0.9467$). (C) Coordinates estimated based on sequence reads from both off-target and on-target regions ($t_0 = 0.9669$). The mean coverage is ~88.9X and ~1.0X for on-target and off-target regions.

Nature Genetics: doi:10.1038/ng.2924



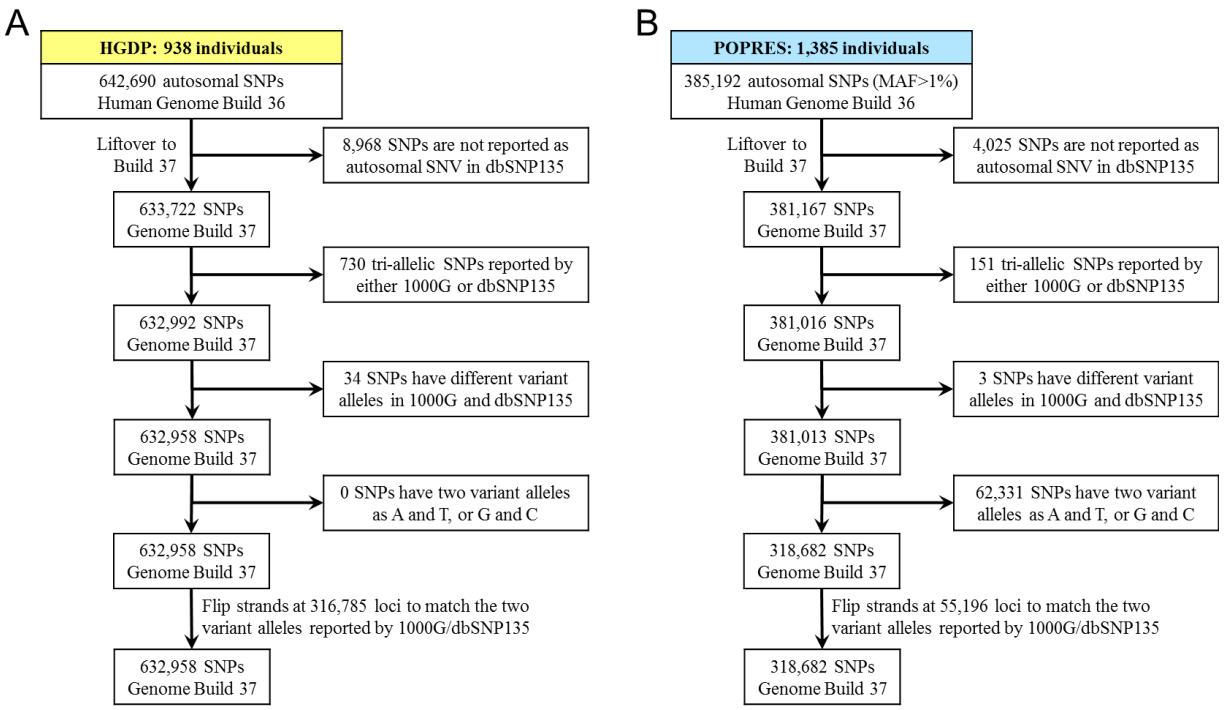
Supplementary Figure 8. Different strategies for sampling 1,280 cases. (A) Sampling from two 8×8 grids along one side, with ten cases from each grid point. (B) Sampling from two 8×8 grids along the diagonal, with ten cases from each grid point. (C) Sampling from one 8×8 grid at the corner, with 20 cases from each grid point. (D) Sampling from one 8×8 grid at the center, with 20 cases from each grid point.

Nature Genetics: doi:10.1038/ng.2924



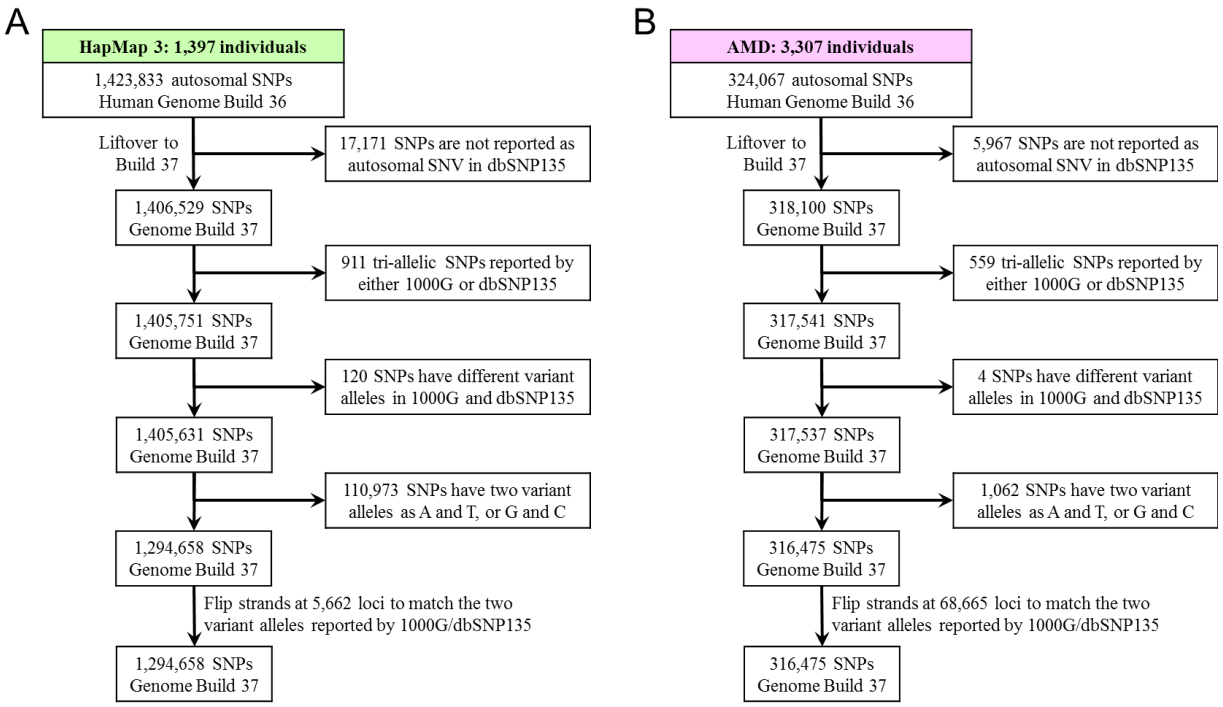
Supplementary Figure 9. Improvement of estimation by using coordinates averaged across multiple runs of LASER on the same data set. The x-axis indicates the number of runs used in calculating the mean PC coordinates. The y-axis indicates the Procrustes similarity t_0 between the mean coordinates and the SNP-based coordinates. Each box represents the distribution of t_0 obtained from 15 repeating runs. (A) Results on sequence data of worldwide samples simulated from genotypes of 238 HGDP individuals, using the other 700 HGDP individuals as the reference panel. We tested on three simulated datasets with coverage of 0.001X, 0.002X, and 0.004X. (B) Results on sequence data of European samples simulated from genotypes of 385 POPRES individuals, using the other 1,000 POPRES individuals as the reference panel. We tested on three simulated datasets with coverage of 0.10X, 0.20X, and 0.40X. We only used one iteration in our examples of the 1000 Genomes and AMD targeted sequencing data, because most samples have relatively high off-target coverage, such that improvement by using multiple iterations is small.

Nature Genetics: doi:10.1038/ng.2924



Supplementary Figure 10. Data processing procedures for the HGDP and the POPRES data sets. (A) The HGDP data set. (B) The POPRES data set.

Nature Genetics: doi:10.1038/ng.2924



Supplementary Figure 11. Data processing procedures for the HapMap 3 and the AMD SNP data sets. (A) The HapMap 3 data set. (B) The AMD SNP data set.

Nature Genetics: doi:10.1038/ng.2924

SUPPLEMENTARY TABLES

Supplementary Table 1. Results on simulated worldwide samples with different sequencing coverage.

Simulated mean coverage λ	Expected number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates				
		Squared correlation of PC1	Squared correlation of PC2	Squared correlation of PC3	Squared correlation of PC4	Procrustes similarity t_0
0.25	140,010	0.9996	0.9996	0.9992	0.9988	0.9997
0.20	114,736	0.9996	0.9996	0.9992	0.9986	0.9996
0.15	88,166	0.9994	0.9996	0.9988	0.9978	0.9995
0.10	60,234	0.9992	0.9992	0.9982	0.9974	0.9993
0.05	30,870	0.9988	0.9986	0.9964	0.9946	0.9989
0.01	6,298	0.9948	0.9932	0.9819	0.9716	0.9949
0.008	5,043	0.9940	0.9920	0.9783	0.9663	0.9940
0.006	3,786	0.9896	0.9882	0.9671	0.9586	0.9911
0.004	2,527	0.9894	0.9882	0.9536	0.9347	0.9887
0.002	1,265	0.9756	0.9706	0.8964	0.8356	0.9729
0.001	633	0.9506	0.9388	0.8350	0.7396	0.9508

Sequence data were simulated for 238 individuals randomly selected from the HGDP dataset and the remaining 700 individuals in the HGDP dataset were used as the reference panel. For each simulated dataset, we compared the estimated ancestry coordinates of the 238 testing individuals to their SNP-based coordinates in **Figure 2A**.

Supplementary Table 2. Results on simulated European samples with different sequencing coverage.

Simulated mean coverage λ	Expected number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates		
		Squared correlation of PC1	Squared correlation of PC2	Procrustes similarity t_0
0.40	105,063	0.9855	0.9078	0.9764
0.35	94,111	0.9866	0.8945	0.9737
0.30	82,597	0.9813	0.8725	0.9671
0.25	70,492	0.9797	0.8540	0.9636
0.20	57,767	0.9738	0.7973	0.9495
0.15	44,390	0.9653	0.7763	0.9428
0.10	30,327	0.9510	0.6647	0.9126
0.05	15,542	0.8851	0.2516	0.7720
0.01	3,171	0.5687	0.0108	0.4786

Sequence data were simulated for 385 individuals randomly selected from the POPRES dataset and the remaining 1000 individuals in the POPRES dataset were used as the reference panel. For each simulated dataset, we compared the estimated ancestry coordinates of the 385 testing individuals to their SNP-based coordinates in **Figure 3A**.

Nature Genetics: doi:10.1038/ng.2924

Supplementary Table 3. Targeted sequencing samples from the 1000 Genomes pilot exon project.

Population label	Ancestral group	Sampling location	Sample size
CEU	N. & W. Europeans	Utah, U.S.A.	56
CHB	Han Chinese	Beijing, China	66
CHD	Han Chinese	Denver, Colorado, U.S.A.	58
JPT	Japanese	Tokyo, Japan	69
LWK	Luhya	Webuya, Kenya	59
TSI	Toscani	Italy	28
YRI	Yoruba	Ibadan, Nigeria	74

Supplementary Table 4. Comparison between sequence-based and SNP-based coordinates for samples from the 1000 Genomes exon project.

Range of coverage per sample	Number of samples	Mean coverage per sample	Average number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates: Squared correlation of PC1	Squared correlation of PC2	Squared correlation of PC3	Squared correlation of PC4	Procrustes similarity t_0
[0.00, 0.06)	103	0.04	8,728	0.9930	0.9884	0.9012	0.6811	0.9938
[0.06, 0.07)	102	0.07	13,431	0.9974	0.9920	0.9204	0.7403	0.9969
[0.07, 0.10)	102	0.09	20,952	0.9982	0.9902	0.9639	0.8503	0.9980
[0.10, 0.55]	103	0.19	46,098	0.9900	0.9805	0.9761	0.8866	0.9931

This table is based on results in **Supplementary Figure 2**, which includes 410 samples analyzed with the HGDP reference panel.

Supplementary Table 5. Comparison between sequence-based and SNP-based coordinates for a subset of the AMD samples.

Range of coverage per sample	Number of samples	Mean coverage per sample	Average number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates: Squared correlation of PC1	Squared correlation of PC2	Procrustes similarity t_0
[0.05, 0.20)	232	0.16	34,114	0.9299	0.5460	0.8770
[0.20, 0.25)	232	0.22	45,603	0.9588	0.6655	0.9285
[0.25, 0.30)	232	0.27	54,837	0.9616	0.6821	0.9254
[0.30, 0.79]	232	0.37	71,102	0.9690	0.6783	0.9480

This table is based on results in **Supplementary Figure 6**, which includes 928 samples analyzed with the POPRES reference panel.

Supplementary Table 6. Distribution of FUSION study samples by birth place.

Place of birth	Reference set	Test set	Total size
Uusimaa	14	14	28
Turku Ja Pori	47	47	94
Hame	56	56	112
Kymi	61	62	123
Mikkeli	31	31	62
Pohjois-Karjala	27	28	55
Kuopio	75	76	151
Keski-Suomi	39	38	77
Vaasa	65	65	130
Oulu	21	21	42
Lappi	7	7	14
Viipuri*	27	26	53

* Viipuri was formally part of Finland and is now part of Russia.

Nature Genetics: doi:10.1038/ng.2924

Supplementary Table 7. Evaluation of corrections for stratification in simulated case/control data with different sampling strategies.

Sampling strategy	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}	$\lambda_{lowfreq}$
Strategy A (All cases from two 8×8 grids along one side)	Uncorrected	-	-	-	11.289	10.515	11.323	12.099
	SNP-based PCs	1	1	1	2.254	2.031	1.003	1.015
	0.20X	0.9993	0.9991	0.9978	2.250	2.031	1.041	1.064
	0.15X	0.9991	0.9988	0.997	2.259	2.033	1.040	1.057
	0.10X	0.9987	0.9982	0.9956	2.251	2.030	1.051	1.078
	0.05X	0.9974	0.9963	0.991	2.247	2.033	1.079	1.099
	0.01X	0.9873	0.9826	0.9556	2.196	2.021	1.181	1.201
	0.005X	0.9737	0.9625	0.9146	2.171	2.019	1.171	1.199
	0.001X	0.8849	0.8329	0.6888	2.409	2.327	1.514	1.670
Strategy B (All cases from two 8×8 grids along the diagonal)	Uncorrected	-	-	-	6.265	6.381	6.276	6.624
	SNP-based PCs	1	1	1	6.463	6.555	1.004	1.011
	0.20X	0.9996	0.9975	0.9995	6.461	6.553	1.034	1.039
	0.15X	0.9995	0.9963	0.9994	6.461	6.555	1.046	1.053
	0.10X	0.9993	0.9951	0.9991	6.461	6.552	1.051	1.058
	0.05X	0.9985	0.9897	0.9982	6.462	6.554	1.084	1.088
	0.01X	0.9926	0.9483	0.991	6.456	6.547	1.197	1.200
	0.005X	0.985	0.8972	0.9822	6.455	6.550	1.202	1.211
	0.001X	0.9311	0.6313	0.9138	6.418	6.516	1.598	1.674
Strategy C (All cases from one 8×8 grid at the corner)	Uncorrected	-	-	-	28.765	20.353	29.057	33.239
	SNP-based PCs	1	1	1	3.445	2.427	0.997	1.042
	0.20X	0.9970	0.9949	0.9934	3.438	2.426	1.065	1.096
	0.15X	0.9959	0.9926	0.9911	3.445	2.427	1.079	1.103
	0.10X	0.9943	0.9898	0.9873	3.439	2.428	1.103	1.120
	0.05X	0.9879	0.9787	0.9728	3.430	2.429	1.147	1.159
	0.01X	0.9451	0.9030	0.8829	3.362	2.432	1.361	1.380
	0.005X	0.8955	0.8123	0.7917	3.469	2.566	1.380	1.466
	0.001X	0.6647	0.4627	0.4225	4.432	3.399	2.271	2.617
Strategy D (All cases from one 8×8 grid at the center)	Uncorrected	-	-	-	10.125	10.349	10.154	11.052
	SNP-based PCs	1	1	1	10.359	10.574	0.999	1.013
	0.20X	0.9986	0.9972	0.9972	10.359	10.568	1.002	1.011
	0.15X	0.9981	0.9963	0.9962	10.360	10.572	1.007	1.011
	0.10X	0.9971	0.9945	0.9942	10.358	10.567	1.005	1.015
	0.05X	0.9944	0.9885	0.989	10.357	10.570	1.016	1.037
	0.01X	0.9715	0.9441	0.9435	10.357	10.572	1.128	1.165
	0.005X	0.9436	0.8904	0.8903	10.348	10.562	1.285	1.357
	0.001X	0.7881	0.6342	0.6082	10.324	10.543	3.591	3.957

The Procrustes similarity score and squared correlations were calculated by comparing sequenced-based PCs to SNP-based PCs of the 1,280 cases sampled from selected regions.

Nature Genetics: doi:10.1038/ng.2924

Supplementary Table 8. Results on simulated worldwide samples with different sequencing error rates specified in LASER.

Specified sequencing error rate in LASER	Sequence-based coordinates vs. SNP-based coordinates				
	Squared correlation of PC1	Squared correlation of PC2	Squared correlation of PC3	Squared correlation of PC4	Procrustes similarity t_0
0	0.9489	0.9368	0.8392	0.7338	0.9504
0.005	0.9500	0.9372	0.8352	0.7365	0.9501
0.010	0.9506	0.9388	0.8350	0.7396	0.9508
0.015	0.9516	0.9370	0.8400	0.7427	0.9516
0.020	0.9489	0.9353	0.8367	0.7539	0.9509

Results in this table are all based on the same simulated sequence dataset of 238 HGDP samples, which were simulated with $\lambda = 0.001$ and $\epsilon = 0.01$.

Supplementary Table 9. Results on simulated European samples with different sequencing error rates specified in LASER.

Specified sequencing error rate in LASER	Sequence-based coordinates vs. SNP-based coordinates		
	Squared correlation of PC1	Squared correlation of PC2	Procrustes similarity t_0
0	0.9522	0.6915	0.9089
0.005	0.9498	0.6537	0.9078
0.010	0.9510	0.6647	0.9126
0.015	0.9526	0.6265	0.9064
0.020	0.9502	0.5937	0.9011

Results in this table are all based on the same set of simulated sequence data of 385 POPRES samples, which were simulated with $\lambda = 0.10$ and $\epsilon = 0.01$.

Nature Genetics: doi:10.1038/ng.2924

SUPPLEMENTARY NOTE

The FUSION Study

University of Michigan, Ann Arbor, Michigan: Goncalo Abecasis, Tom Blackwell, Michael Boehnke, Jeroen Huyghe, Anne Jackson, Hui Jiang, Goo Jun, Hyun Min Kang, Yeji Lee, Adam Locke, Clement Ma, Randy Pruim, Mark Reppell, Cassie Robertson, Laura Scott, Xueling Sim, Heather Stringham, Tanya Teslovich, Ryan Welch, William Wen, Cristen Willer, Pranav Yajnik

National Human Genome Research Institute, Bethesda, Maryland: Lori Bonnycastle, Peter Chines, Mike Erdos, Anthony Kirilusha, Narisu Narisu, Steve Parker, Michael Stitzel, Amy Swift, Leland Taylor, Brooke Wolford

National Institutes of Health, Bethesda, Maryland: Francis Collins

University of North Carolina, Chapel Hill, North Carolina: Maren Cannon, Jennifer Kulzer, Karen Mohlke, Ying Wu

University of Helsinki, Helsinki, Finland: Jaakko Tuomilehto

University of Kuopio, Kuopio, Finland: Markku Laakso

University of Southern California, Los Angeles, California: Richard Watanabe, Tom Buchanan

Cedars-Sinai Medical Center, West Hollywood, California: Richard Bergman

Nature Genetics: doi:10.1038/ng.2924