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

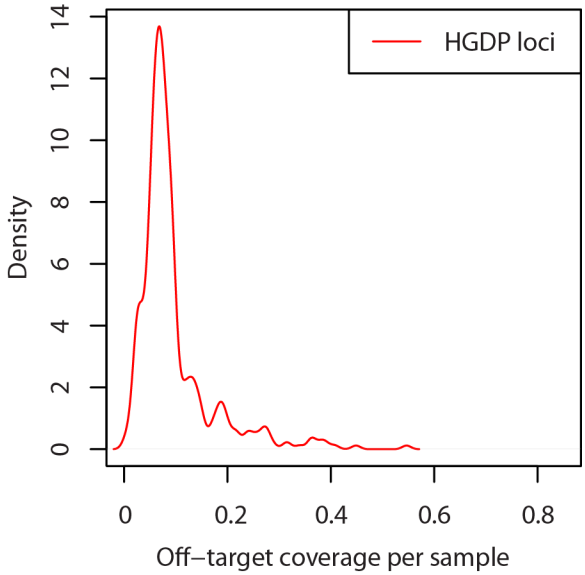
[9] Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI 48109

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

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

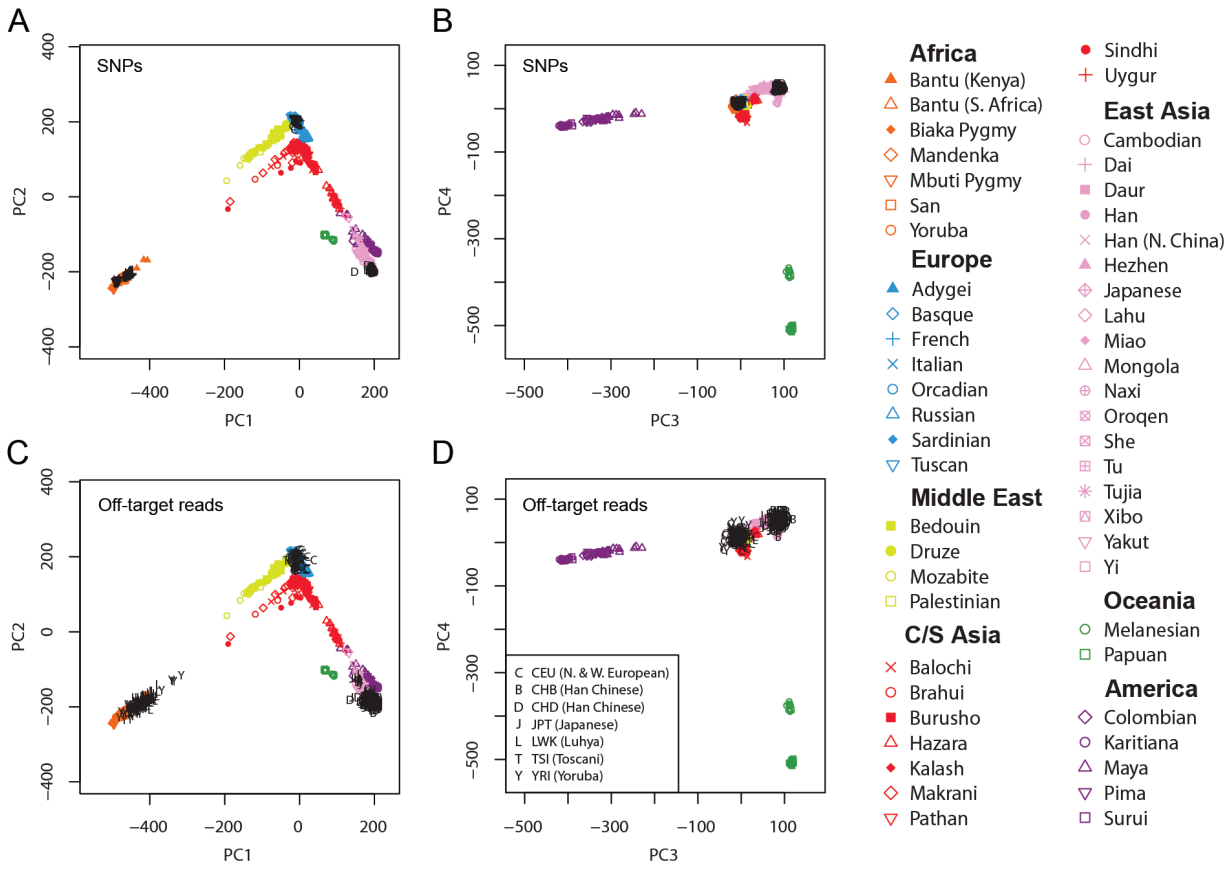
Nature Genetics: doi:10.1038/ng.2924

SUPPLEMENTARY FIGURES



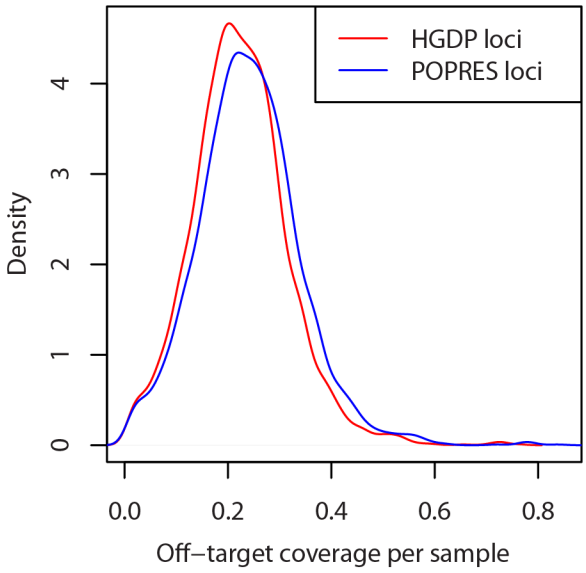
Supplementary Figure 1. Off-target coverage for 410 samples from the 1000 Genomes exon project. The off-target coverage for each sample is calculated by averaging across 632,958 loci in the HGDP. For 270 loci that appear in the targeted regions, we set the coverage at these loci to 0 for all samples.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

Nature Genetics: doi:10.1038/ng.2924



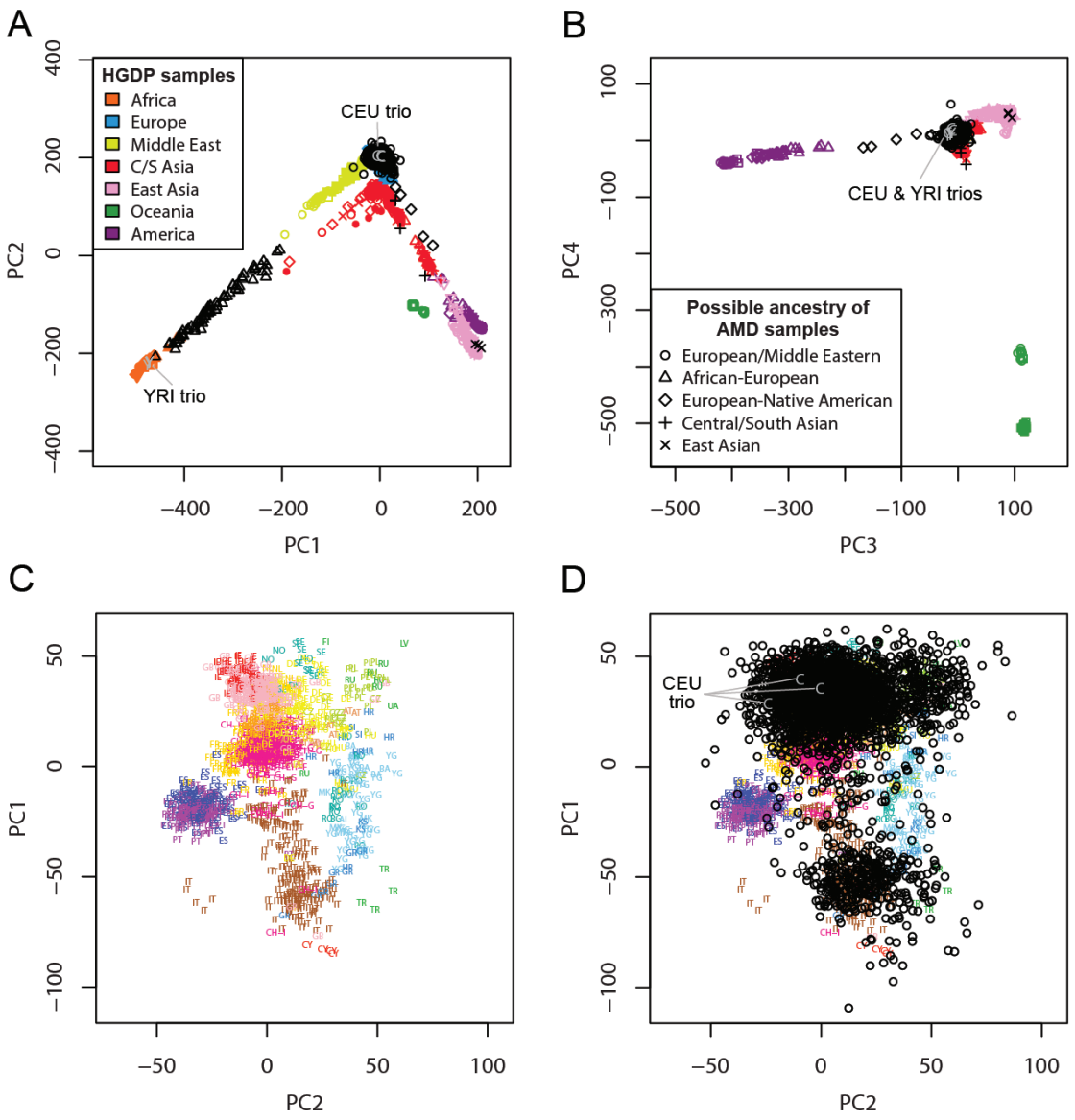
Supplementary Figure 2. Estimation of worldwide ancestry for 410 samples in the 1000 Genomes exon project. The SNP genotypes of these samples are from the HapMap Project. We used all HGDP individuals as the reference panel, as labeled by colored points. (A,B) Results based on SNPs that were genotyped 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



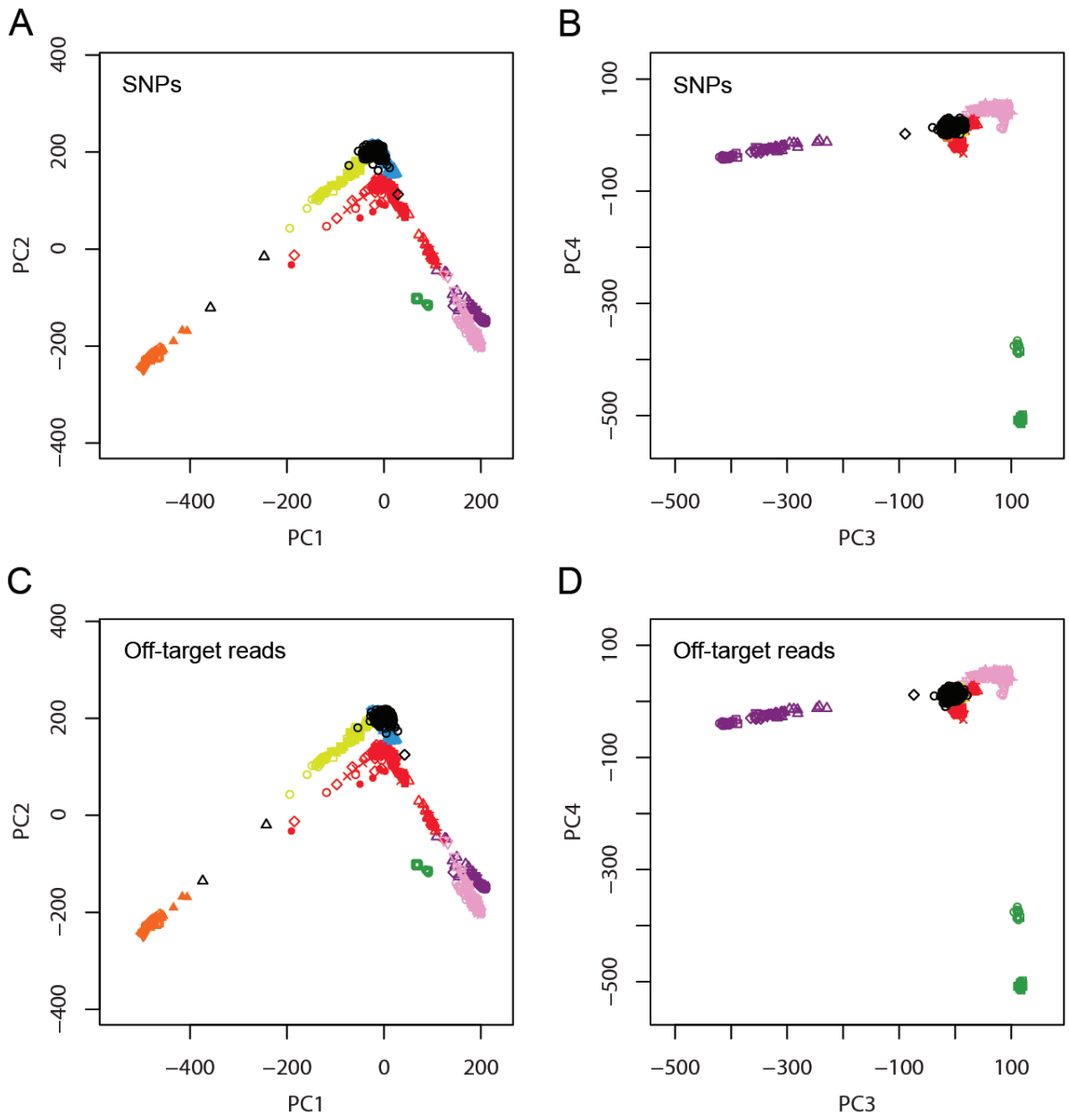
Supplementary Figure 3. Off-target coverage for 3,159 samples from the AMD study. The red line indicates off-target coverage averaged across 632,958 loci included in HGDP. The blue line indicates off-target coverage averaged across 318,682 loci that are included in POPRES. For loci that appear in the 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



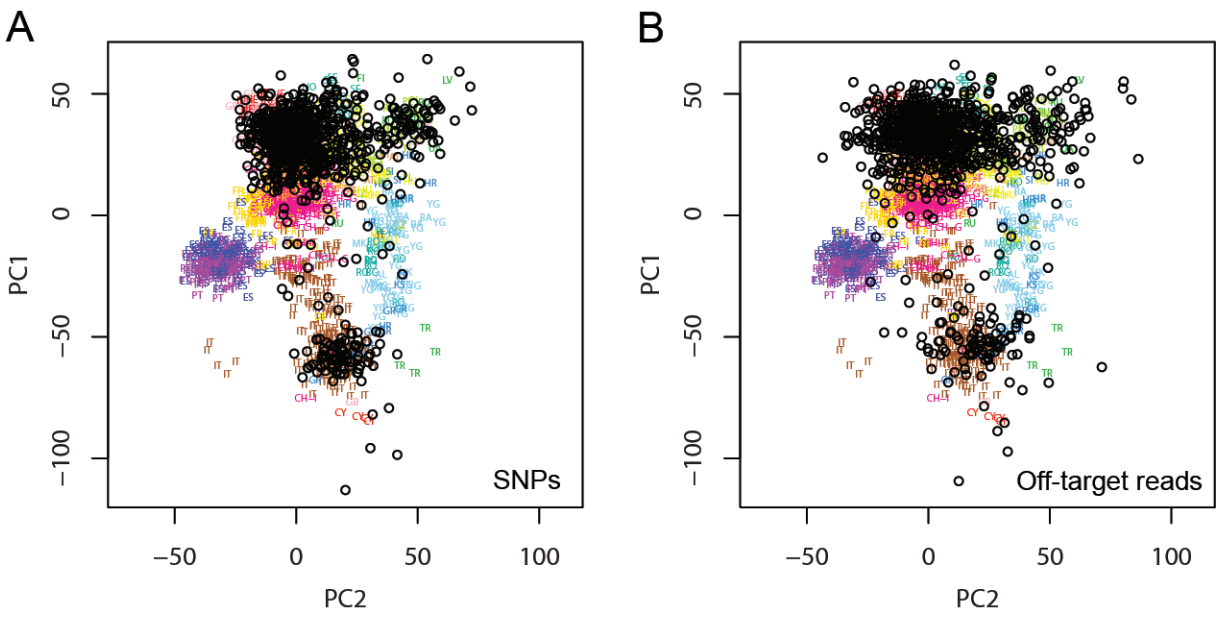
Supplementary Figure 4. Estimation of ancestry for 3,159 samples in the AMD targeted sequencing dataset. (A,B) Results based on the HGDP reference panel, whose colors and symbols follow **Supplementary Figure 2**. AMD samples are displayed in black, with different symbols representing possible ancestries 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



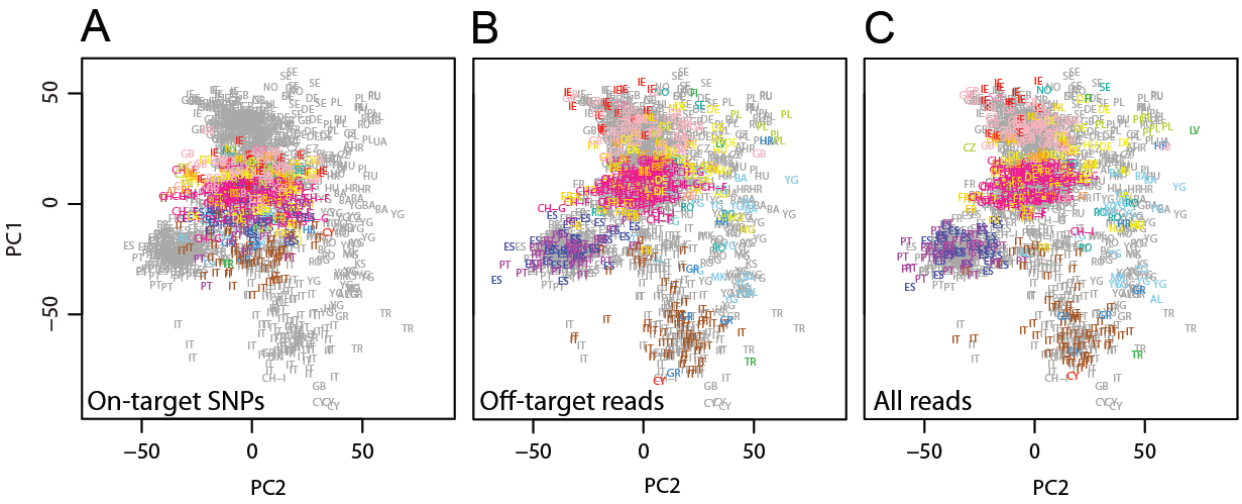
Supplementary Figure 5. Sequence-based coordinates and SNP-based coordinates for 931 AMD samples when using the HGDP reference panel. Colors and symbols for HGDP and AMD samples follow **Supplementary Figure 2**. (A,B) Results based on 45,700 SNPs that are shared by HGDP, POPRES and AMD SNP datasets. (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



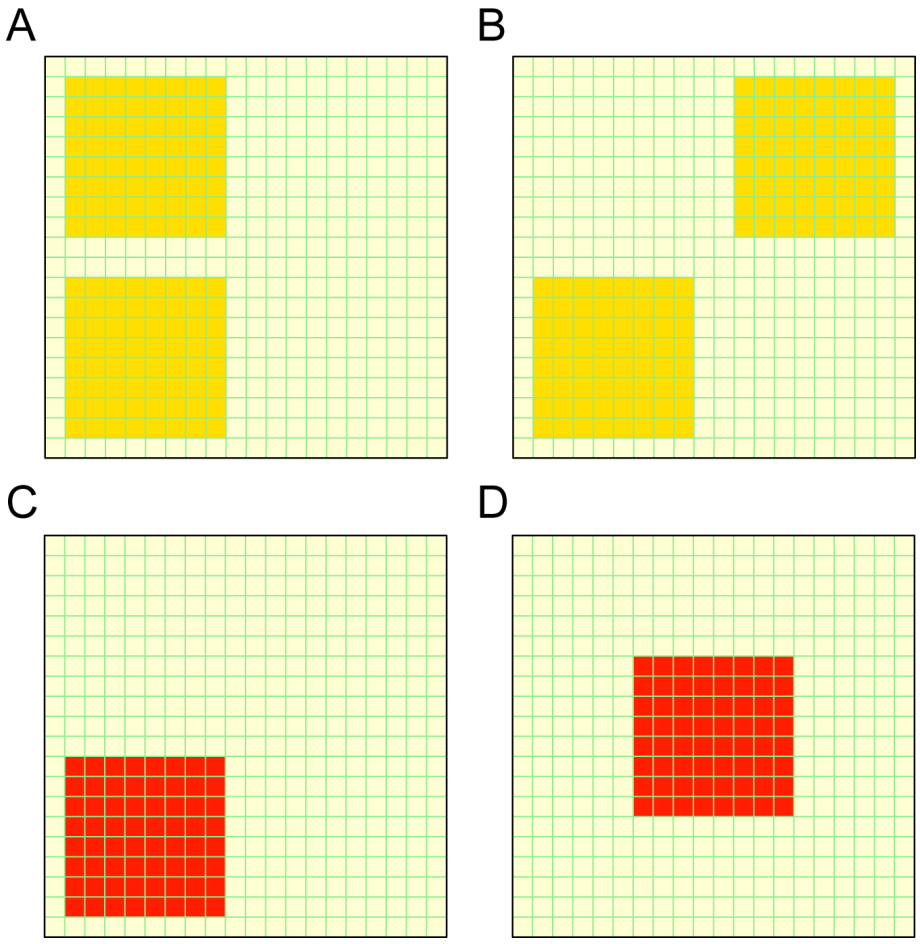
Supplementary Figure 6. Sequence-based coordinates and SNP-based coordinates for AMD samples when using the POPRES reference panel. We only included 928 AMD samples whose genotype data are available and who might be Europeans or Middle Eastern according to results in **Supplementary Figure 5**. (A) Results 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



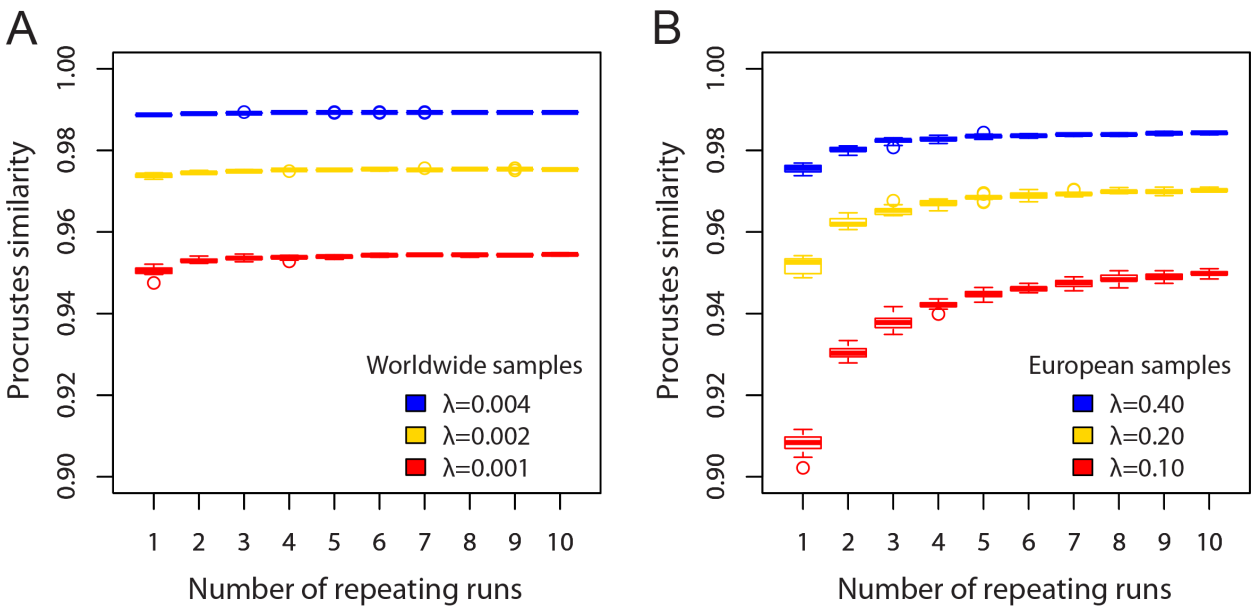
Supplementary Figure 7. **Results for simulated exome sequencing data for 385 POPRES samples.** (A) Coordinates estimated from SNP genotypes at 2,547 on-target loci. The Procrustes similarity to the SNP-based coordinates in **Figure 3A** is $t_0 = 0.5031$. (B) Coordinates estimated based on off-target 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



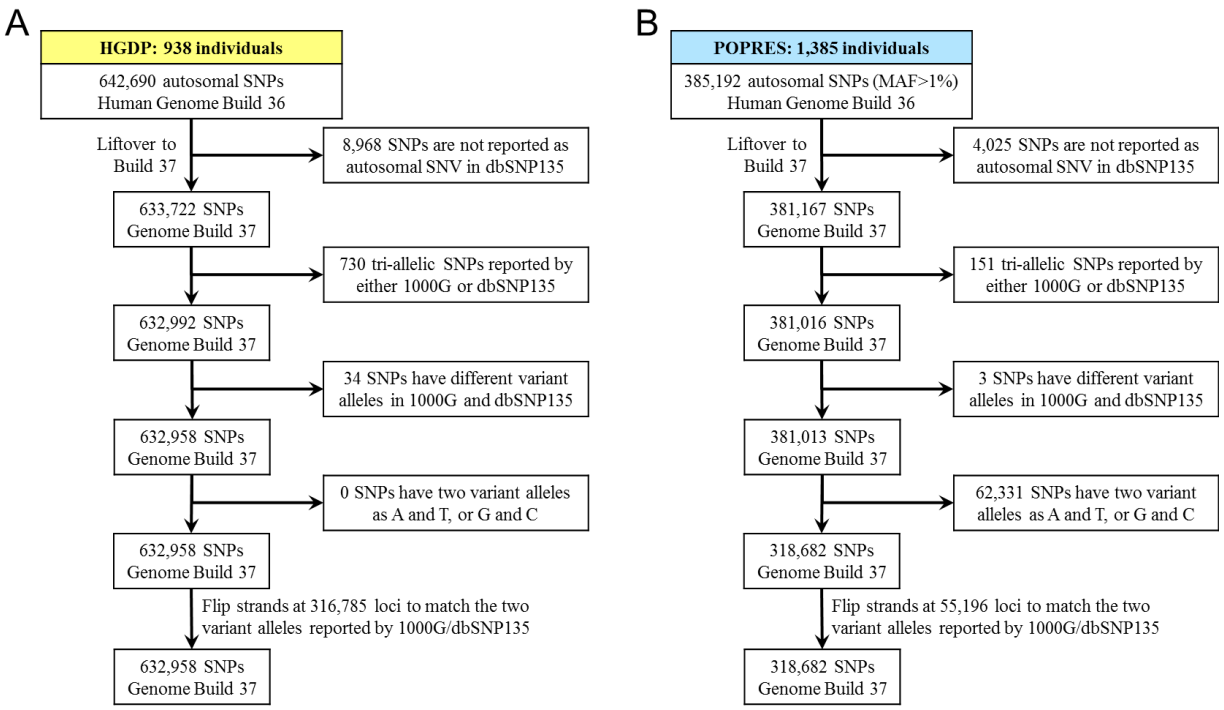
Supplementary Figure 8. Different strategies for sampling 1,280 cases. (A) Sampling from two 8×8 grids along one side, with ten cases from each grid point. (B) Sampling from two 8×8 grids along the diagonal, with ten cases from each grid point. (C) Sampling from one 8×8 grid at the corner, with 20 cases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



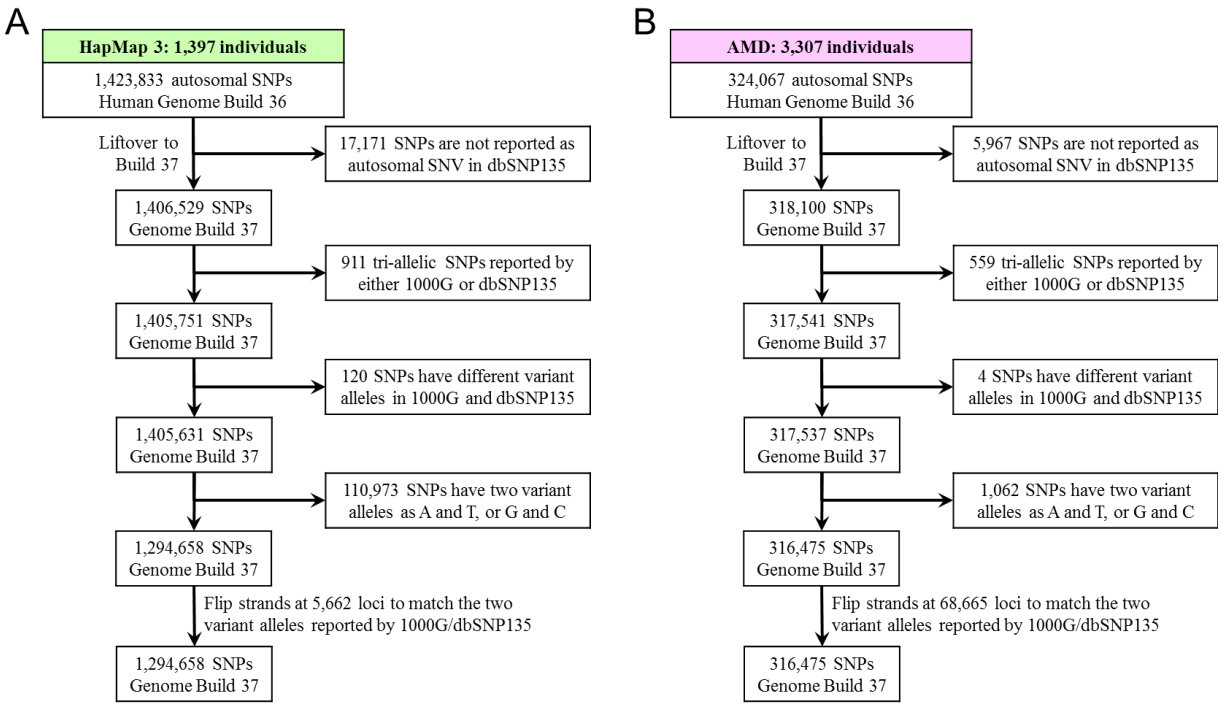
Supplementary Figure 9. Improvement of estimation by using coordinates averaged across multiple runs of LASER on the same data set. The x-axis indicates the number of runs used in calculating the mean PC coordinates. The y-axis indicates the Procrustes similarity t_0 between the mean coordinates and 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