Supplementary Information for *Extremely low-coverage sequencing and imputation increases power for genome-wide association studies*

Bogdan Pasaniuc[1,2,3,*], Nadin Rohland[3,4], Paul J. McLaren[3,5], Kiran Garimella[3], Noah Zaitlen[1,2,3], Heng Li[3], Namrata Gupta[3], Benjamin Neale[3], Mark Daly[3], ARRA Autism Sequencing Collaboration[6], Pamela Sklar[7] Patrick F. Sullivan[8], Sarah Bergen[3], Jennifer L. Moran[3], Christina M. Hultman[9], Paul Lichtenstein[9], Patrik Magnusson[9], Shaun M. Purcell[10], David W. Haas[11], Liming Liang[1,2,3], Shamil Sunyaev[3,5], Nick Patterson[3], Paul I.W. de Bakker[3,5,12], David Reich[3,4,*, ±], Alkes L. Price[1,2,3,*, ±]

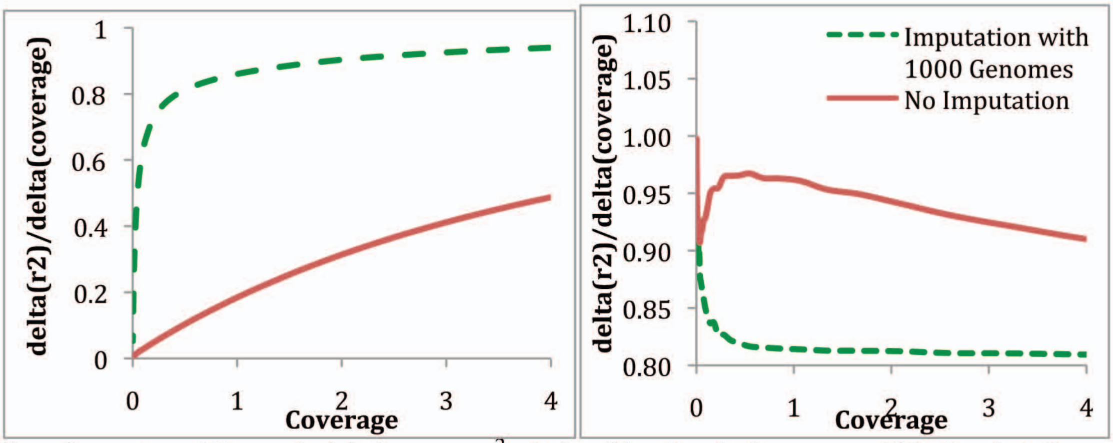
* To whom correspondence should be addressed (bpasaniu@hsph.harvard.edu, reich@genetics.med.harvard.edu, aprice@hsph.harvard.edu)

± Co-senior authors

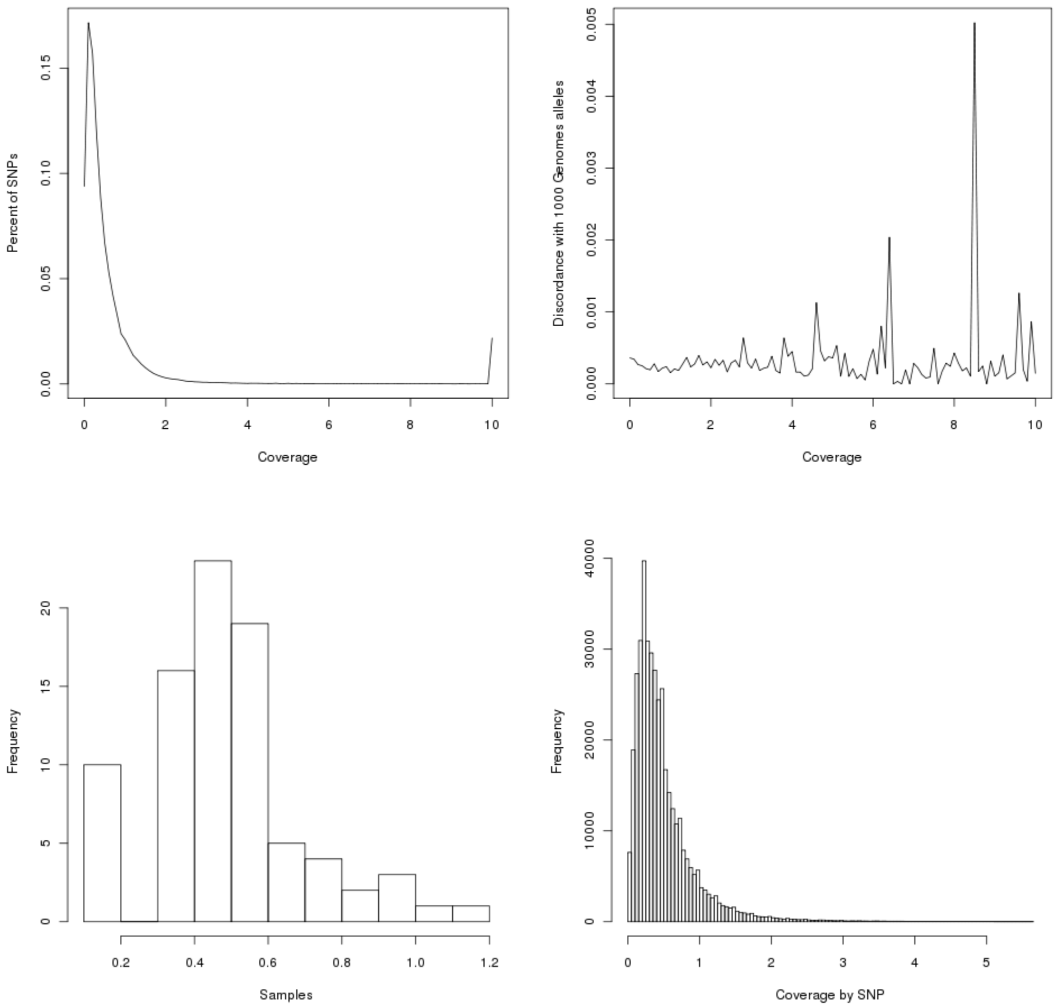
1. Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA, 02115.
2. Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA, 02115.
3. Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
4. Department of Genetics, Harvard Medical School, Boston, MA, USA.
5. Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA, 02115.
6. A complete list of authors can be found in the Supplementary Note
7. Department of Psychiatry, Friedman Brain Institute, & Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, New York, USA
8. Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC 27599
9. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden
10. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
11. Vanderbilt University School of Medicine, Nashville, TN
12. Department of Medical Genetics and Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Nature Genetics: doi:10.1038/ng.2283

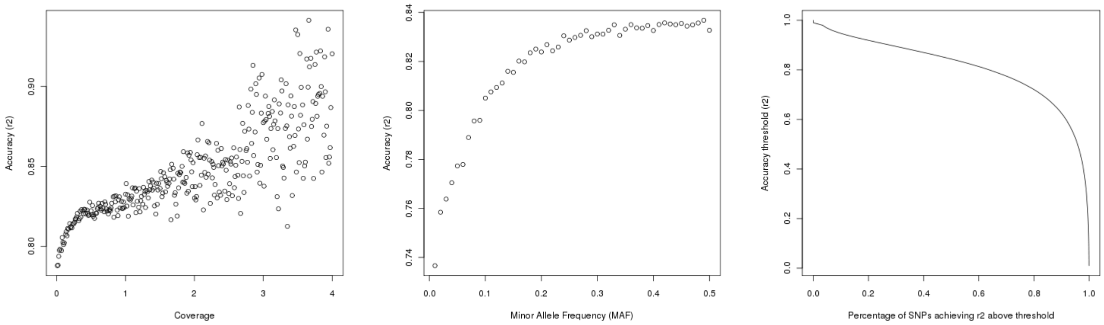
Supplementary Figures:



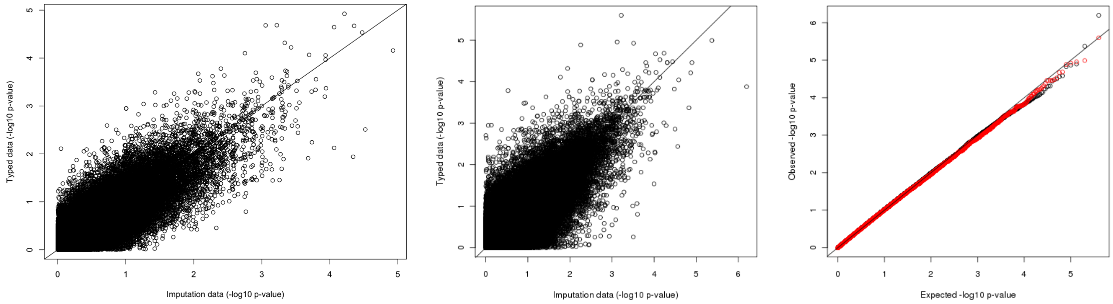
Supplementary Figure 1. (a) Average r^2 attained in simulations over 100 simulated samples when genotype dosages where either inferred based on imputation using 381 reference haplotypes or with no imputation from the reads spanning each SNP independently. (b) Partial derivative of accuracy (r^2) as function of coverage attained at extremely low-coverage, showing the gain in accuracy as function of added coverage when imputation is used at extremely low coverage.



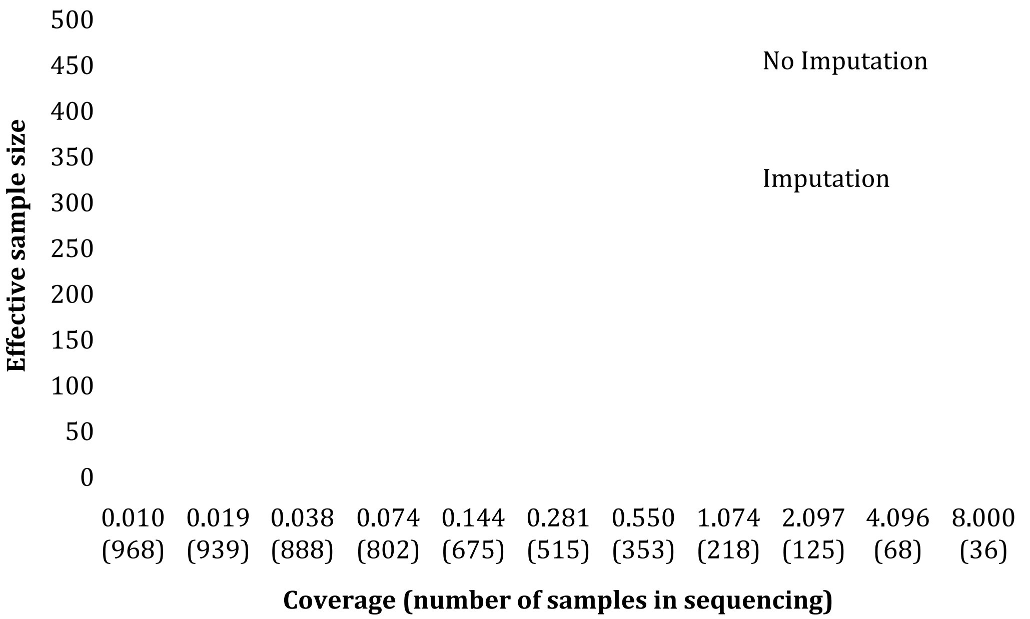
Supplementary Figure 2: (a) Percentage of SNPs at different coverages in IHCS data showing that the great majority of SNPs are covered (albeit at ultra-low coverage) in exome sequencing data in which the exons are coverage at coverage greater than 10x (data plotted for chromosome 20; similar plots are obtained for all chromosomes). (b) Discordance rate (computed as percentage of bases discordant with the reference and alternate allele called at all European polymorphic loci in the 1000 Genomes project on chromosome 20, other chromosomes show similar results) plotted as function of coverage in the IHCS data. No unusual increase in the discordance rate is found at coverage less than 1x. (c)Distribution of coverage by sample (chromosome 20) in the IHCS data set. (d) Distribution of coverage by SNP in the IHCS data set

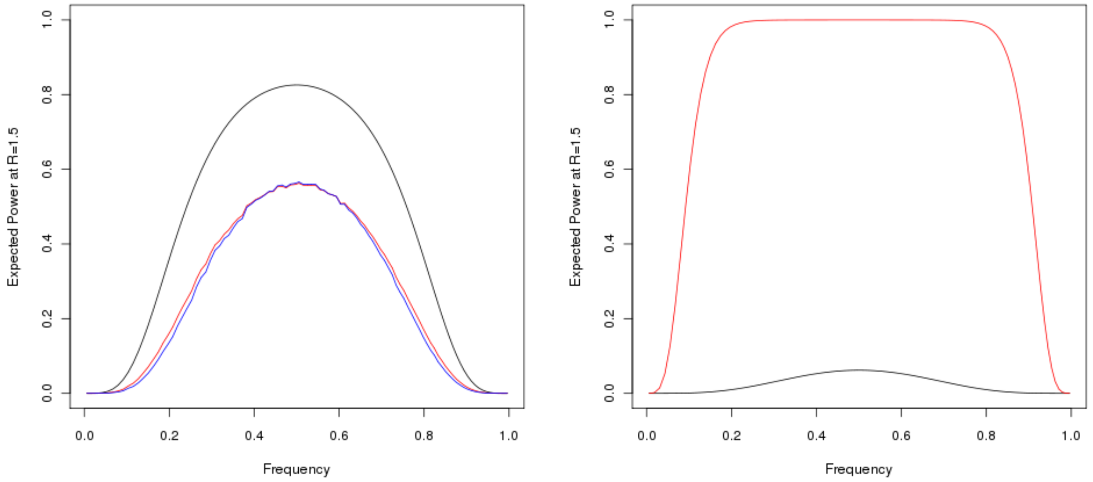


Supplementary Figure 3: (a) Accuracy as function of coverage in IHCS whole-exome data set computed across 398,098 SNPs using Illumina genotype calls as ground truth. (b) Accuracy as function of frequency in IHCS whole-exome data set computed across 398,098 SNPs using Illumina genotype calls as ground truth. (c) Cumulative distribution of SNPs with accuracy above threshold in the IHCS whole-exome data set. Results computed across 398,098 SNPs using Illumina genotype calls as ground truth.

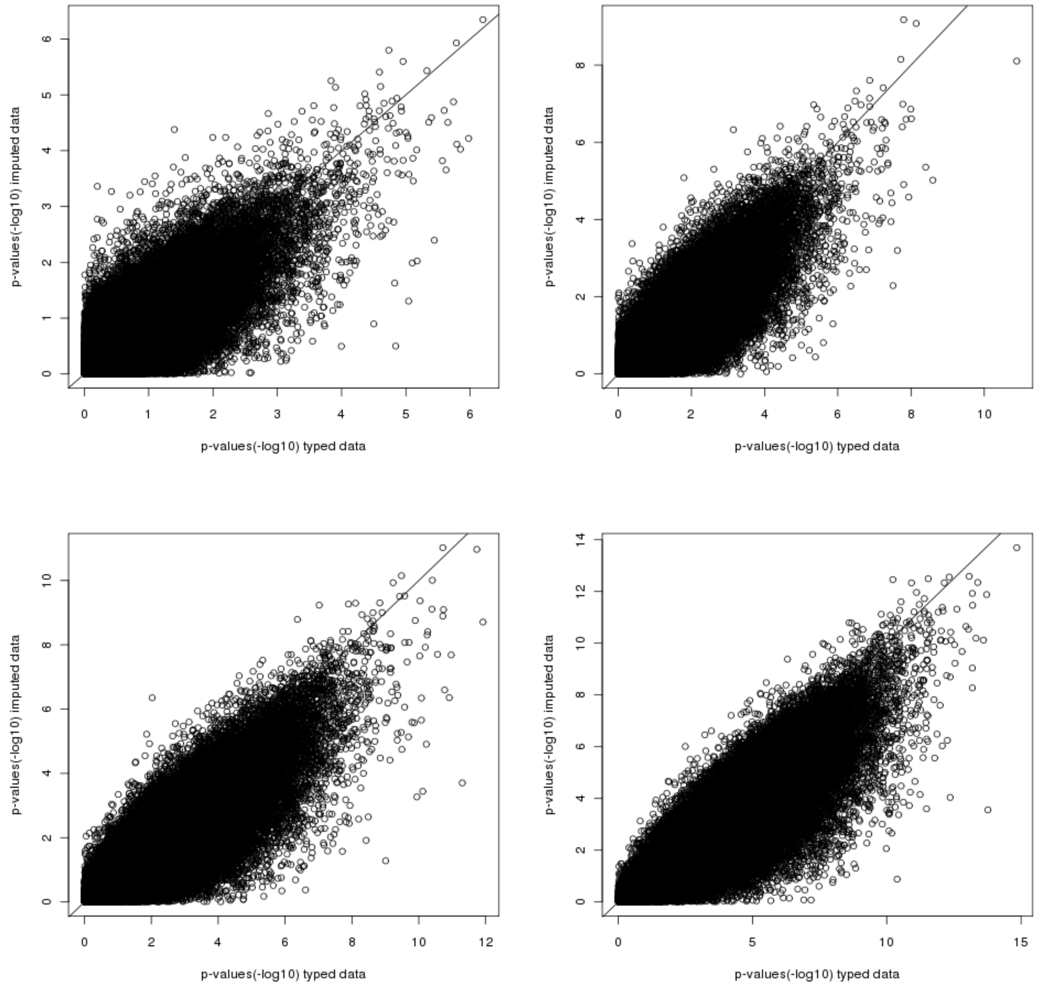


Supplementary Figure 4. (a) Plot of association p-values in 100,000 simulated SNPs computed over genotype data versus simulated imputation genotypes with r2=0.82 to the true data. We simulated imputation with r2=0.82 accuracy by adding random errors with corresponding probability in the genotyping calls. (b) Plot of association p-values computed on typed versus imputed data on IHCS data set (average coverage of 0.5x). We observe a Pearson squared correlation of 0.68 between p-values attained on typed versus imputed data. (c) Observed versus expected association minus log 10 p-values at 398,098 SNPs across the genome in 84 samples (61 cases and 23 controls) ascertained for either HIV Controller (cases) or HIV Progressor (controls) phenotype. Red denotes statistics computed over typed data, while black denotes statistics using imputed data from sequencing reads.

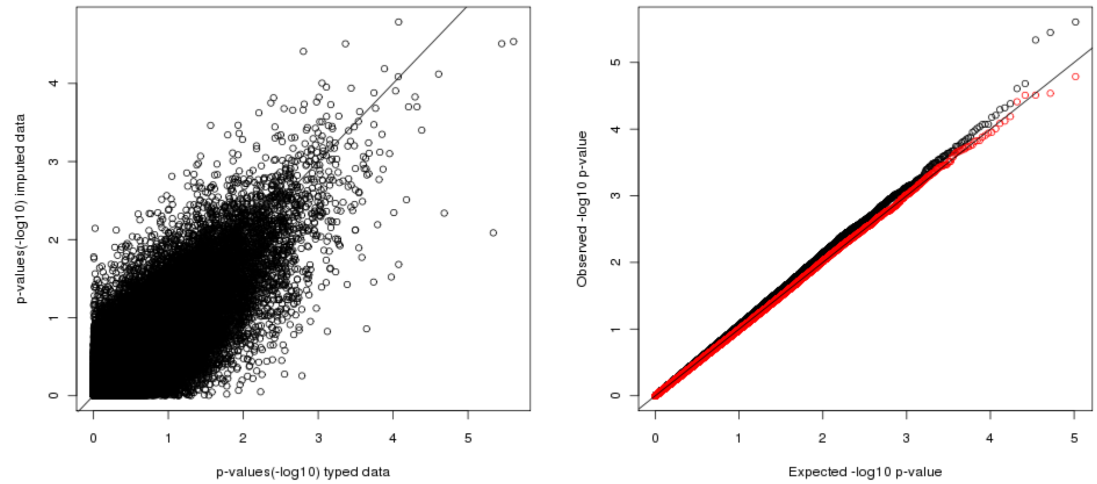




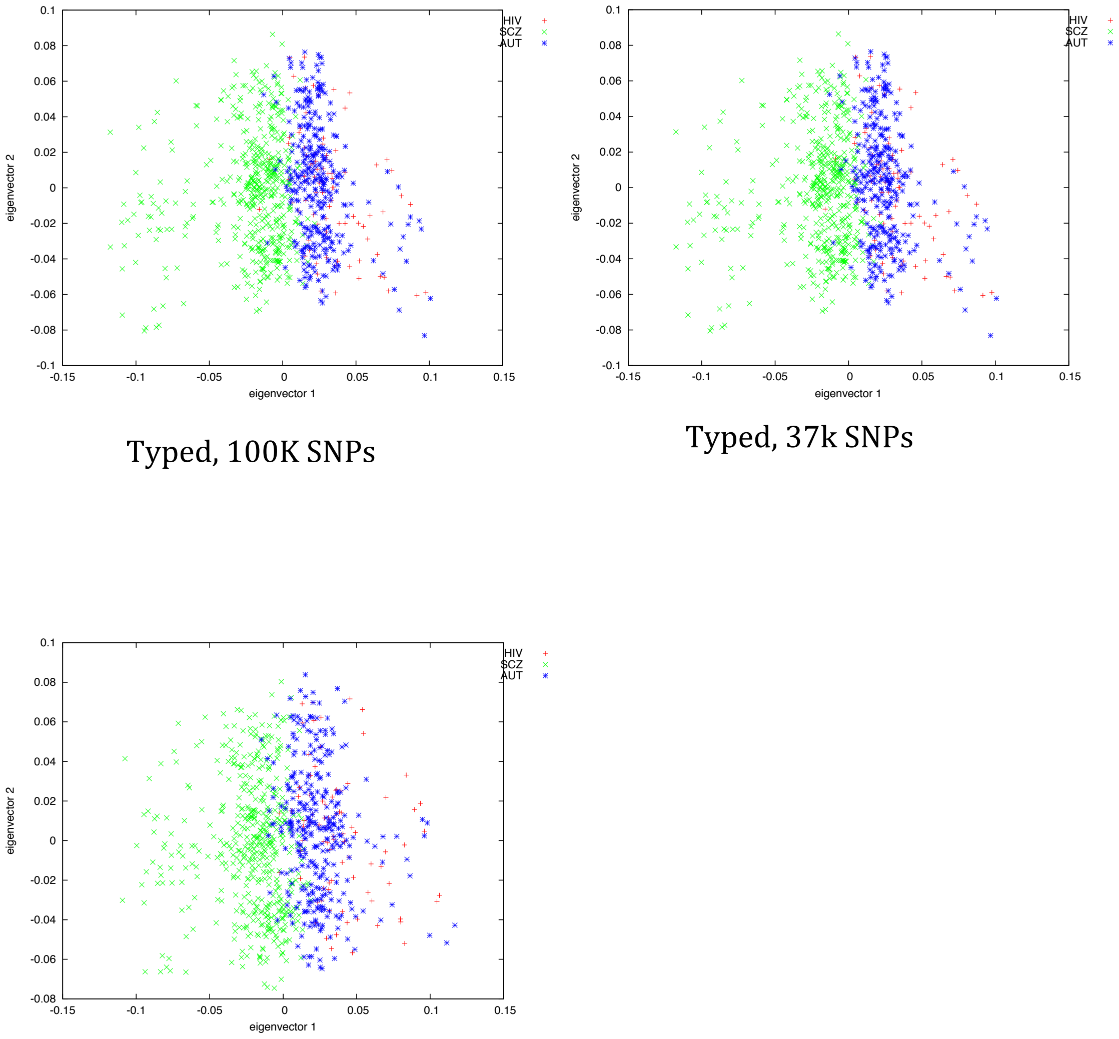
Supplementary Figure 5. (a) Effective Sample size attained within a given budget of \$300,000 with fixed sample preparation cost of \$300 and cost per 1x of \$1,000. We observe the existence of an optimal coverage for maximizing effective sample size. (b)Expected power as function of allele frequency assuming study with 1000 cases and 1000 controls at causal with odds ratio R=1.5 as function of allele frequency. Black line denotes expected average power from genotyping while red line denotes expected power computed assuming ncp of λ* sqrt(0.82*N), while blue line denotes the average across all SNPs (in the MAF bin) of expected power (incorporating the observed variance in r^2 across SNPs of given frequency in the IHCS data). (c)Expected power for a fixed budget of \$300,000 as function of frequency. Sequencing is denoted in red and assumes 6,800 samples sequenced at cov=0.1x (with \$133 per 1x and \$30 per sample prep) yielding an effective sample size of roughly 4,600 (r^2=0.65). Genotyping at \$400 per sample is denoted in black (effective sample size of 750).



Supplementary Figure 6. P-values attained on simulated phenotype data (beta=0.05,0.1,0.15,0.2) using either the typed genotypes versus imputed genotypes from sequencing data.



Supplementary Figure 7. (a) P-values attained on SCZ (503) vs. AUT(322) analysis in which SCZ samples were used as "controls" and AUT samples as "cases" using the typed genotypes versus imputed genotypes from sequencing data. (b) Expected versus observed P-values attained on SCZ (503) vs. AUT (322) analysis in which SCZ samples were used as "controls" and AUT samples as "cases" using either the typed (black) genotypes versus imputed (red) genotypes from sequencing data.



Typed, 100K SNPs

Typed, 37k SNPs

Imputed, 37k SNPs

Supplementary Figure 8. Principal component analysis of the 909 samples using either typed (103,977 SNPs), imputed (37,796 SNPs imputed with accuracy accuracy (r^2) greater than 0.8) or typed (same accurately imputed 37,796 SNPs) genotype data. EIGENSTRAT software was used for principal component analysis.

Supplementary Tables:

Chr	Start (Mb)	End (Mb)	Recombination Rate (cM/Mb)	SNPs typed in IHCS	Total SNPs (1000 Genomes Project, phase1 2011)	SNPs after filtering
3	180	185	1.24	568	25447	20792
17	15	20	1.08	449	28167	24397
11	25	30	0.74	599	28490	25010
1	215	220	1.19	834	28120	24251
15	75	80	0.81	579	26005	21711
4	175	180	1.75	644	29705	25823
11	60	65	1.03	503	25251	20792
17	55	60	1.22	573	22109	18418
12	15	20	0.97	650	26826	22982
1	115	120	1.01	671	25567	21659
Total	**50Mb**		**1.11**	**6070**	**265687(avg 26568.7)**	**150261**

Supplementary Table 1. Summary of randomly selected regions used in simulations.

	0.5x coverage		
	1-3%	**3-5%**	**>5%**
Beagle	0.60	0.79	0.90
MaCH/Thunder	0.47	0.70	0.86
IMPUTE2	0.55	0.75	0.88
	4x coverage		
	1-3%	**3-5%**	**>5%**
Beagle	0.87	0.93	0.97
MaCH/Thunder	0.75	0.89	0.96
IMPUTE2	0.78	0.89	0.96

Supplementary Table 2. Accuracy (average r^2) binned by minor allele frequency of compared methods in simulations of short read data across the 10 considered regions. All methods were provided reference panels of haplotypes. From a runtime perspective Beagle took ~1h for a 5Mb region, Impute2 close to ~1.5h, while MaCH/Thunder performed imputation in ~7h for a given region.

(a)

Sequencing error rate	Imputation + Reference Panel	Imputation	No Imputation
0.000	0.826 (0.915)	0.196 (0.197)	0.107 (0.149)
0.005	0.819 (0.910)	0.186 (0.191)	0.105 (0.146)
0.010	0.812 (0.904)	0.177 (0.185)	0.104 (0.143)
0.015	0.804 (0.898)	0.169 (0.180)	0.102 (0.141)
0.020	0.795 (0.891)	0.162 (0.175)	0.100 (0.138)
0.025	0.787 (0.883)	0.155 (0.170)	0.098 (0.135)
0.030	0.777 (0.874)	0.149 (0.166)	0.096 (0.133)

(b)

σ^2 in average sample coverage (cov=0.5x)	Imputation + Reference Panel	Imputation	No Imputation
0.05	0.831 (0.925)	0.185 (0.193)	0.110 (0.150)
0.10	0.829 (0.923)	0.184 (0.191)	0.108 (0.149)
0.15	0.825 (0.919)	0.180 (0.187)	0.106 (0.145)
0.20	0.812 (0.904)	0.177 (0.185)	0.104 (0.143)
0.25	0.790 (0.879)	0.175 (0.182)	0.101 (0.141)
0.30	0.764 (0.848)	0.173 (0.182)	0.099 (0.139)
0.35	0.723 (0.801)	0.171 (0.179)	0.098 (0.137)
0.40	0.698 (0.773)	0.171 (0.179)	0.098 (0.137)

(c)

Shape of Gamma distr(α)	Imputation + Reference Panel (r^2)	Imputation (r^2)	No Imputation (r^2)
2.0	0.812 (0.904)	0.173 (0.184)	0.103 (0.144)
3.0	0.813 (0.905)	0.176 (0.185)	0.103 (0.144)
4.0	0.812 (0.904)	0.177 (0.185)	0.104 (0.143)
5.0	0.812 (0.904)	0.179 (0.190)	0.104 (0.147)
6.0	0.814 (0.904)	0.180 (0.191)	0.104 (0.147)
7.0	0.813 (0.904)	0.180 (0.189)	0.104 (0.145)
8.0	0.813 (0.905)	0.180 (0.191)	0.105 (0.146)

(d)

Sample size (N)	Imputation + Reference Panel (r^2)	Imputation (r^2)	No Imputation (r^2)
50	0.807 (0.910)	0.181 (0.185)	0.108 (0.147)
100	0.812 (0.904)	0.177 (0.185)	0.104 (0.143)
150	0.825 (0.909)	0.179 (0.189)	0.105 (0.144)
190	0.828 (0.907)	0.178 (0.187)	0.104 (0.142)

Supplementary Table 3. Accuracy (measured as average r^2 across SNPs) as function of (a) sequencing error rate (b) variance across samples (c) distribution of coverage across loci (d) sample size. When not varying, N=100, error rate is set to 0.01, cov=0.5x, σ^2 =0.2 and α = 4. Results in parenthesis denote averages over only the 6070 SNPs genotyped in IHCS data set.

Chr.	Avg. coverage (Illumina SNPs)	σ^2 avg. sample coverage	σ^2 avg. locus coverage	SNPs typed on Illumina arrays	Accuracy (r^2)
1	0.54	0.16	0.48	31068	0.83
2	0.48	0.15	0.43	32975	0.83
3	0.47	0.13	0.43	27837	0.82
4	0.41	0.12	0.38	23457	0.82
5	0.47	0.13	0.43	25314	0.82
6	0.47	0.13	0.43	26147	0.82
7	0.47	0.14	0.44	21736	0.82
8	0.48	0.15	0.42	22873	0.82
9	0.49	0.15	0.46	19602	0.81
10	0.51	0.16	0.44	21566	0.82
11	0.52	0.15	0.47	20034	0.83
12	0.50	0.14	0.46	19881	0.82
13	0.43	0.12	0.39	14966	0.80
14	0.50	0.14	0.45	13401	0.81
15	0.52	0.16	0.47	12293	0.81
16	0.58	0.20	0.49	12687	0.81
17	0.59	0.20	0.52	10701	0.80
18	0.47	0.15	0.43	12304	0.80
19	0.63	0.23	0.57	6646	0.82
20	0.58	0.21	0.50	10602	0.81
21	0.46	0.15	0.42	5984	0.79
22	0.63	0.25	0.53	6024	0.81
All	0.50	0.14	0.45	398098	0.82

Supplementary Table 4. Average coverage by Chromosome and accuracy attained by genotype imputation from read data at SNPs also typed on Illumina platforms in the 84 IHCS samples. The Illumina genotyped SNPs were used as gold standard.

RsID	Chr	Pos	Coverage	r2	p-value typed (-log10)	p-value imputed (-log10)	Ratio	Effect typed [conf. int]	Effect imputed [conf. int]
rs6905949	6	30140525	0.29	0.89	0.23	0.35	1.53	0.04 [-0.11 0.19]	0.06 [-0.10 0.21]
rs17475879	6	30364508	0.3	0.81	0.42	0.36	0.87	0.10 [-0.13 0.33]	0.09 [-0.14 0.33]
rs13201769	6	30756066	0.62	0.86	0.03	0.09	3.63	0.00 [-0.13 0.14]	0.02 [-0.13 0.16]
rs4713380	6	30785273	0.32	0.93	0.33	0.21	0.63	0.07 [-0.12 0.26]	0.05 [-0.15 0.25]
rs4713385	6	30787593	1.12	0.95	0.33	0.36	1.08	0.07 [-0.12 0.26]	0.08 [-0.12 0.27]
rs9295928	6	30823630	1.32	0.94	0.6	0.71	1.17	0.12 [-0.09 0.33]	0.14 [-0.07 0.36]
rs7756521	6	30848253	0.33	0.96	1.38	1.12	0.81	0.19[0.01 0.38]	0.17 [-0.02 0.36]
rs3873332	6	30895990	0.79	0.94	0.9	1.04	1.16	0.16 [-0.05 0.37]	0.18 [-0.03 0.39]
rs3871466	6	30983683	0.07	0.88	0.94	0.86	0.92	0.16 [-0.04 0.36]	0.16 [-0.05 0.37]
rs13210132	6	31001143	0.19	0.8	1.01	1.08	1.07	0.18 [-0.03 0.40]	0.21 [-0.03 0.44]
rs3130981	6	31083813	0.4	0.98	0.28	0.24	0.85	0.05[-0.11 0.22]	0.05 [-0.12 0.22]
rs1062470	6	31084435	0.36	0.97	0.4	0.42	1.05	0.05[-0.07 0.18]	0.06 [-0.07 0.19]
rs3094212	6	31085770	0.73	0.96	1.37	1.42	1.03	0.15 [0.01 0.29]	0.15 [0.01 0.29]
rs3095320	6	31087934	0.55	0.98	0.28	0.24	0.85	0.05 [-0.11 0.22]	0.05 [-0.12 0.22]
rs3094205	6	31091862	0.96	0.97	0.4	0.41	1.05	0.05 [-0.07 0.18]	0.06 [-0.07 0.19]
rs9263715	6	31095801	0.85	0.93	0.67	0.68	1	0.10 [-0.06 0.25]	0.09 [-0.05 0.23]
rs3823418	6	31100942	1.12	0.96	0.23	0.4	1.69	0.05 [-0.13 0.23]	0.07 [-0.10 0.25]
rs3130453	6	31124849	0.35	0.97	0.07	0.03	0.38	0.01 [-0.12 0.14]	0.01 [-0.13 0.14]
rs720465	6	31125777	0.4	0.97	0.09	0.01	0.17	0.02 [-0.13 0.16]	0.00 [-0.15 0.15]
rs1419881	6	31130593	0.44	0.99	0.69	0.76	1.1	0.09 [-0.05 0.22]	0.09 [-0.04 0.23]

rs3130932	6	31133943	0.42	0.98	0.71	0.75	1.06	0.10 [-0.05 0.25]	0.10 [-0.05 0.25]
rs9263870	6	31170514	0.43	0.99	0.16	0.07	0.45	0.05 [-0.20 0.29]	0.02 [-0.23 0.27]
rs9263871	6	31170528	0.37	0.87	0.12	0.09	0.75	0.03 [-0.13 0.18]	0.02 [-0.14 0.18]
rs2395471	6	31240692	0.27	0.96	1.38	1.41	1.02	0.14 [0.01 0.28]	0.14 [0.01 0.27]
rs5010528	6	31241032	0.37	0.76	0.29	0.16	0.54	0.07 [-0.14 0.29]	0.05 [-0.19 0.29]
rs9366778	6	31269173	0.43	0.84	1.34	1.87	1.4	0.15 [0.00 0.29]	0.18 [0.04 0.31]
rs9264942	6	31274380	0.26	0.69	1.77	2.35	1.33	0.19 [0.04 0.34]	0.24 [0.08 0.40]
rs2156875	6	31317347	0.31	0.94	1.56	1.16	0.74	0.17 [0.02 0.31]	0.13 [-0.01 0.28]
rs2442719	6	31320538	1.12	0.8	0.66	0.35	0.52	0.10 [-0.06 0.26]	0.06 [-0.10 0.22]
rs2523554	6	31331829	0.63	0.95	0.79	0.81	1.02	0.12 [-0.05 0.28]	0.12 [-0.04 0.28]
rs9266409	6	31336568	0.77	0.9	0.08	0.11	1.52	0.02 [-0.15 0.18]	0.02 [-0.14 0.18]
rs2844529	6	31353593	0.94	0.92	2.53	3.02	1.19	0.21 [0.08 0.35]	0.23 [0.10 0.37]
rs2523467	6	31362930	0.63	0.93	2.53	2.39	0.94	0.21 [0.08 0.35]	0.21 [0.07 0.34]
rs2596531	6	31387557	0.55	0.94	1.31	1.38	1.05	0.15 [0.00 0.30]	0.16 [0.01 0.31]
rs2844513	6	31388214	0.17	0.97	1.23	1.35	1.1	0.14 [-0.00 0.28]	0.15 [0.01 0.29]
rs2516513	6	31447588	0.36	0.86	1.53	1.25	0.82	0.18 [0.02 0.34]	0.15 [-0.00 0.31]
rs3093662	6	31544189	0.5	0.95	0.92	0.65	0.7	0.15 [-0.04 0.34]	0.12 [-0.07 0.31]
rs2844480	6	31564821	0.6	0.83	0.3	0.5	1.65	0.05 [-0.09 0.19]	0.07 [-0.07 0.22]
rs9378200	6	31572927	0.12	0.85	0.31	0.64	2.06	0.07 [-0.13 0.28]	0.13 [-0.08 0.34]
rs9348876	6	31575276	0.33	0.86	0.31	0.64	2.07	0.07 [-0.13 0.28]	0.13 [-0.08 0.34]
rs4151664	6	31920873	2.8	0.98	0.19	0.2	1.03	0.05 [-0.16 0.26]	0.05 [-0.16 0.26]
rs12198173	6	32026808	0.19	0.85	0.04	0.33	8.24	0.01 [-0.19 0.21]	0.08 [-0.13 0.28]
rs13199524	6	32066765	0.6	0.96	0.01	0.04	3.28	0.00 [-0.20 0.21]	0.01 [-0.19 0.21]

rs12153855	6	32074804	1.31	0.98	0.04	0.02	0.55	0.01 [-0.19 0.21]	0.01 [-0.19 0.20]
rs12663103	6	32161324	0.31	0.81	0.71	0.28	0.39	0.13 [-0.07 0.32]	0.07 [-0.14 0.28]
rs6906662	6	32266506	0.25	0.95	0.46	0.71	1.54	0.09 [-0.10 0.29]	0.13 [-0.07 0.32]
rs7356880	6	32401327	0.07	0.93	0.46	0.4	0.88	0.10 [-0.11 0.32]	0.09 [-0.12 0.31]
Average			**0.57**	**0.91**	**0.69**	**0.72**	**1.04**		

Supplementary Table 5. Association statistics computed at known variants associated to the HIV Progressor/Controller phenotype[5] using typed or sequencing-based imputation. *Average ratio is computed as the ratio of the average of association p-values. To compute the statistical significance of the ratio being different from 1 we randomly flipped the typed and imputed p-value in the computation of the ratio to observe that in more than 18% of the 1,000 permutations the ratio to exceed 1.04.

Window Size	Haplotype length in #SNPs	Number of Distinct Haplotypes (Hapmap 3)	Haplotype Similarity (Hapmap 3)	Difference in number of distinct haplotypes (Hapmap3 – 1000 Genomes)	Difference in haplotype similarity metric (Hapmap3 – 1000 Genomes)
10 Kb	5.74 (0.05)	5.37 (0.06)	0.4324 (0.0029)	0.32 (0.01)	-0.0026 (0.0001)
50 Kb	26.35 (0.17)	20.89 (0.25)	0.2087 (0.0022)	1.59 (0.03)	-0.0050 (0.0002)
100 Kb	51.97 (0.47)	37.56 (0.58)	0.1244 (0.0030)	2.50 (0.06)	-0.0046 (0.0004)

Supplementary Table 6. Difference in haplotype similarity statistics across the European samples part of both Hapmap 3 and 1000 Genomes project. Numbers in parenthesis show standard errors of the mean as computed across randomly sampled windows of given size across Chromosome 1 phased data in both projects. Both metrics show an increase in the number of distinct haplotypes in HapMap 3 as compared to 1000 Genomes, presumably due to the more accurate phasing using trio families or due to joint phasing of all samples that missed rare haplotypes in the 4x sequencing coverage data.

Supplementary Note:

1. Selection of genomic regions used in simulation

We used the following procedure to select regions representative of the whole genome in terms of SNP density and recombination rate. First, we divided the genome into non-overlapping 5Mb windows and computed the average recombination rate in cM per Mb (https://mathgen.stats.ox.ac.uk/wtccc-software/recombination_rates/) as well as the number of SNPs identified in the 1000 Genomes project (June 2011 phase 1 release). Across all windows we found the average recombination rate to be 1.285 cM/Mb (sd 0.706) and 28,159 (sd 5,471) polymorphic sites in the European data. We randomly chose 10 regions totaling 50Mb that are within 1 standard deviation to the average recombination rate and number of SNPs (Supplementary Table 1). We assigned the 762 European haplotypes (174 CEU, 186 FIN, 178 GBR, 196 TSI, 28 IBS) of the 1000 Genomes project to two non-overlapping panels of 381 haplotypes each, with one panel serving as the "reference" panel in all our imputation runs and one panel for simulating sequencing data. We filtered out all SNPs monomorphic in the reference simulation panel.

2. Comparison of imputation methods from short read data

Our procedure for simulating short read sequencing data sets relies on assumptions regarding distribution of coverage across samples and loci as well as the sequencing error rate. Following standard approaches[1], we use three steps to simulate data sets at average coverage *cov*. First, we draw from the normal distribution with mean equal to *cov* and σ^2 to obtain *cov* at each sample $\phi(j)$. Second, we use a draw from a Gamma distribution $\Gamma(\alpha,1/\alpha)$ to obtain the shape of coverage at each loci $\gamma(i)$. Finally, the number of reads at locus i in sample j is drawn from a Poisson distribution with mean $\phi(j)\gamma(i)$. Each read is generated by randomly copying one of the 2 alleles for sample j and locus i, with miscopying errors inserted at a rate of ε.

Nature Genetics: doi:10.1038/ng.2283

We compared three approaches (Beagle[2], Impute2[3] and MaCH/Thunder[4]) to incorporate LD in the calling of genotypes from short read data. For each considered region, we simulated short read data sets assuming 0.5x and 4x coverage followed by imputation. Each method was provided the genotype likelihood for all calls, defined as the probability of the set of reads given a genotype value under a simple error model with ε=0.01 assuming uncorrelated errors across reads[1]. All methods were provided the reference panel of haplotypes as input either as an external haplotype panel for Beagle and Impute (default settings) or included in the sample for MaCH/Thunder using the following version and parameters:

- Impute2 (version 2.1.2): -prob_g -pgs_prob –Ne 11418. We used the genetic map provided on impute2 website for June 2011 phase 1 release of 1000 Genomes.
- Beagle (version 3.3.1): no parameters
- MaCH/Thunder_Glf (version 1.1.0) --shotgun -r 50 --states 100 --dosage –phase

Results in Supplementary Table 2 suggest that Beagle and Impute2 attain increased accuracy over MaCH/Thunder per unit of runtime when default parameters are used. We also attempted running MaCH/Thunder starting from results of Beagle but the increase in accuracy was marginal. Our results do not represent an exhaustive comparison among methods for imputation from sequencing (such a comparison is beyond the scope of this manuscript); different parameter settings could change the relative performance of compared methods and we did not assess all possible settings here. In this work we used Beagle for all our experiments as it provides a good balance between runtime and accuracy. Improved imputation from sequencing data will only improve the accuracy we observe in our experiments from short read sequencing. Thus, our results can be viewed as a lower bound on accuracy that can be attained in imputation from sequencing data.

3. Effect of simulation parameters on accuracy

The main parameters of our simulations are the standard deviation in average coverage across samples σ^2, the shape of the Gamma distribution α and the error rate ε. Intuitively, the accuracy is increased as σ^2 decreases, α increases and ε decreases. We conducted simulation experiments at various values for these parameters to assess their effect on the results using the 762 European haplotypes of the 1000 Genomes project phase 1 June 2011 release. The 762 haplotypes were split at random between two panels of haplotypes each of size 381; one panel was used as reference and another one to simulate sequencing data. We simulated data over 100 samples (number chosen to roughly match the number of samples in the IHCS whole-exome sequencing data) by randomly pairing with no replacement haplotypes from the simulation panel. We compared 3 approaches for inferring genotypes from reads: imputation with reference panel (381 haplotypes not used in simulation of sequencing data), imputation with no reference panel and no imputation. The procedure for no imputation sets the genotype dosage independently for each SNP genotype as 2*P(reads | genotype is 2) + 1* P(reads | genotype is 1). The probability of the set of reads given a genotype value is computed under a simple error model with ε=0.01 assuming uncorrelated errors across reads[1].

Supplementary Table 3(a) displays the accuracy when the error rate is increased showing that, as expected, all approaches yield poorer estimates of genotypes with increase in error rate. However, we note that the approach that uses a reference panel of haplotypes shows the smallest decrease in accuracy demonstrating the robustness of proposed approach to increased sequencing error rates. Supplementary Table 3(b) displays the accuracy as a function of variation in average coverage across samples. All methods show large decreases in performance as the variability in coverage across samples increases. This emphasizes the importance of reducing the variation in coverage across samples. We also quantified the robustness of our approach to the distribution of coverage across loci. We varied the shape of the Gamma distribution α with results displayed in Supplementary Table 3(c). As expected, as the variance in coverage across loci decreases (as α increases) we notice an increase in performance across all methods. Supplementary Table 3(d) shows that the accuracy marginally increases with sample size suggesting that the performance of our approach is bound to increase with larger samples. In our approach, we impute all SNPs that are polymorphic in the reference panel. To quantify the effect of SNPs not polymorphic in the reference panel but at considerable frequency in the simulation panel, we computed the number of SNPs filtered out from our simulations due to the fact that they were non-polymorphic in the reference panel but attained over 1% minor allele frequency (maf) in the simulation panel. 930 SNPs out of the original set of 265,687 SNPs were filtered out from our simulations due to this reason and would attain an accuracy r^2 of 0 in our approach (none of these SNPs attained over 5% maf in the simulation panel). Therefore, when adjusting for the SNPs not polymorphic in the reference panel, the overall average r^2 at 0.5x coverage (including the 930 SNPs on top of the 150,261 used in simulations) decreases from 0.812 to 0.807 (σ^2 =0.2 and $\alpha = 4$). As the 1000 Genomes project catalogues a larger proportion of low frequency variation across the genome, the number of such SNPs will decrease.

4. Imputation using reference panels of haplotypes boosts accuracy at ultra low-coverage

We carried out simulations over 100 samples sequenced at various extremely low-coverage to estimate the accuracy attained by imputation with reference panels of haplotypes as compared with no imputation (independent estimation of genotype calls at each sample and each SNP based on the set of reads overlapping that SNP). To quantify the potential benefit of performing imputation, we plotted the gain in accuracy as function of gain in coverage (derivative of accuracy as coverage) (Supplementary Figure 1(a)). For a coverage *c* we estimated the accuracy at coverage *1.25c* and plotted the ratio $r^2(1.25c)/r^2(c)$ normalized by the gain in coverage of *0.25*. As coverage increases, reads are sampled from the same LD blocks and the amount of new information present in each read decreases. As expected, we observed a much faster rate of decrease for imputation (in imputation a read contains information about the whole LD block and therefore smaller amount of sampled reads are required to overlap in information content) rather than no imputation. In the absence of imputation, the derivative is always smaller than 1 showing, similar to other studies[1], that more samples with less coverage are preferred (Supplementary Figure 1(b)).

5. Reduced haplotype diversity in the 1000 Genomes phase 1 haplotypes

Our simulations rely on using half of the 1000 Genomes European haplotypes for simulating sequencing data sets and the other half as the reference panel. In such a set up, a critical assumption being made is that the simulation and reference panels of haplotypes comprise random samples from the population (Europeans in this case). An artificial increase in the haplotype similarity between the reference and simulation panel as compared to two random samples from the same population will yield increased accuracy in simulations as opposed to real data. This is a direct effect of the imputation methodology that uses reference haplotypes to fill in missing data in the target sample. An artificial increase in haplotype similarity between reference and simulation panel, as opposed to random samples from the populations, can be caused by under representation of rare haplotypes in the data. Intuitively, 1000 Genomes haplotypes have been generated from sequencing data at 4x coverage with rare haplotypes being more likely to be missed in the joint calling across all samples; such a deficiency in the haplotypic diversity in the 1000 Genomes haplotype data will cause non-overlapping random subsets of haplotypes to be more similar than in real data. To search for such an effect, we compared the haplotypic diversity in 1000 Genomes phase 1 European haplotypes with the haplotypic diversity in the HapMap 3 phase 2 data (www.hapmap.org), which used a trio-aware phasing methodology for increased haplotype inference accuracy. For consistency, we restricted to only the samples present in both data sets. We randomly selected regions of 10, 50 and 100 Kb across Chromosome 1, and we compared the number of distinct haplotypes across each window in the HapMap 3 phase 2 data as opposed to the 1000 Genomes phase 1 data. In addition, we computed the average haplotype similarity, defined as the percent of all pairs of haplotypes that are identical across all pairs of haplotypes in the data (we excluded windows with only one haplotype across all samples). Supplementary Table 6 shows that, 1000 Genomes phase 1 data contains a smaller number of haplotypes as compared to HapMap 3 data, presumably due to joint calling across all samples that may miss rare haplotypes in the calling from 4x coverage.

6. IHCS data set

Genome-wide SNP genotype and whole-exome sequence data on 84 HIV-positive individuals of European descent were obtained by the International HIV Controllers Study[5]. 43 of these samples have been genotyped on the Illumina HumanHap 650Y, and 41 on the Human-1M-duo array.

Investigators can submit a concept sheet detailing their study design, research questions and other needs in order to request access to the genetic data presented here. The concept sheet with detailed instructions can be downloaded from:
http://cfar.globalhealth.harvard.edu/fs/docs/icb.topic938249.files/Harvard%20CFAR%20Concept%20Sheet%20Template%20.docx
Please e-mail completed forms to Pamela Richtmyer (prichtmyer@partners.org). Requests will be reviewed on the basis of scientific merit, feasibility and potential overlap with accepted concept sheets or ongoing investigations.

We only used the intersection of SNPs for all analyses described. Only unrelated samples with high genotyping rates (>95%) of European ancestry were included, after filtering out SNPs with low frequency (MAF < 1%), high missingness (>2%), and departure from Hardy-Weinberg equilibrium ($P < 10^{-6}$).

Starting with 3 ug of genomic DNA (gDNA), sample preparation and in-solution hybridization were preformed as described by Fisher et al[6]. The quantified libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates was then performed according to manufacturer's protocol (Illumina) using V4 Chemistry and V4 Flowcells. Sybr Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were sequenced on Genome Analyzer II's, using V4 Sequencing-by-Synthesis kits and analyzed with the Illumina RTA v1.8.67 pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis.

A subset of samples was prepared using the protocol previously mentioned with some slight modifications. Initial genomic DNA input into shearing was reduced from 3 ug to 100 ng of total gDNA in 50 uL of solution. Illumina paired end adapters were replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter. These samples were then processed as described by Fisher et al[6]. The quantified libraries were normalized to 3nM and then denatured. Cluster amplification was then performed according to the manufacturer's protocol using the HiSeq V2 Cluster Chemistry and HiSeq V2 Flowcells. Cluster density was checked using the previously described Sybr Green dye assay. Flowcells were then sequenced on Illumina HiSeq 2000 using the HiSeq V2 Sequencing-by-Synthesis kits and analyzed using the Illumina RTA v1.10.15. Because of the indexed adapters, it was necessary to use the Illumina Multiplexing Sequencing Primer Kit; however the indexed read was not performed.

Read data was processed using the GATK/Picard software package[7] for next generation sequencing data. Alignment to the human reference genome hg19 was performed using the bwa aligner, using parameters "-q 5 -l 32 -k 2 -t 4 -o 1". Reads with low mapping quality as well as reads mapping to multiple locations were removed. The Genome Analysis toolkit GATK[7] was used to re-calibrate the mapping quality as part of the default pre-processing pipeline of next generation sequencing data at the Broad institute.

Although whole-exome sequencing approach increases the amount of reads generated from the exonic regions, due to imperfect capture technologies, a significant amount of reads falls outside of the exome. For example, using the latest exonic annotation for the human genome (http://www.ncbi.nlm.nih.gov/CCDS, April 22, 2011), including 100 bases around each exon, we observe that only 72% of reads align to exonic regions on chromosome 20 (similar numbers are obtained at other chromosomes) with remainder of the reads falling in outside of exonic regions (at an average coverage of 0.5x). Importantly, we do not observe any increase in discordance with 1000 Genomes reference and variant allele calls at polymorphic loci identified by 1000 Genomes in the European panel (Supplementary Figure 2).

To remove the effect of highly covered loci (e.g. exonic regions and other potential artifacts of the capture technology) on genotype imputation from sequencing data, we sub-sampled all loci with more than 4x coverage to a mean 0.5x. The threshold of 4x was chosen to achieve a near-complete separation between exonic (and near exonic) and non-exonic loci (Supplementary Figure 2). The average distribution in the 100 samples after re-sampling is 0.5x with σ^2=0.14 in average coverage across samples and σ^2=0.45 in average coverage across loci for chromosome 20 (see Supplementary Figure 2 and Supplementary Table 5 for data for each chromosome). The distribution of average coverage by locus can be approximated using a Gamma distribution with shape parameter α = 4 (see Supplementary Figure 2).

7. Genotype imputation from sequencing data

Following standard approaches for genotype imputation from short read data[1,8] we computed genotype likelihoods from reads overlapping any polymorphic site identified in the 1000 Genomes project[9] independently at each sample and locus as follows. Given a SNP locus i in individual j, a set of observed reads R(i,j) overlapping this locus (given as counts of the reference allele; reads not matching both reference and alternate are discarded) and the reference and alternate alleles called by 1000 Genomes at this SNP, the genotype likelihood of the genotype g having x=0,1,2 copies of the reference allele is computed as: $P(R(i,j) \mid g = x) = \prod_{r \in R(i,j)} P(r \mid g = x)$. The probability of observing a read given a genotype is computable assuming an error model (we used a fixed an error rate ε =0.01); e.g. if r=1, P(r|g=0)=ε, P(r|g=2)=1-ε, P(r|g=1)=0.5.

The genotype likelihoods, together with the reference European haplotypes from Phase 1 release of 1,000 Genomes project, are provided to the Beagle imputation engine that computes dosages using an LD-aware approach. To quantify accuracy, we used the squared Pearson correlation coefficient r^2, as this metric quantifies the loss in effective sample size due to errors in imputation. As “gold standard” data, we used genotype data inferred using genotyping arrays on the same 84 samples. Therefore, we compared the imputed genotypes from extremely low-coverage sequencing data with genotype calls obtained from Illumina genotyping arrays.

Accuracy across all chromosomes is displayed in Supplementary Table 4. We observe similar performance across all chromosomes (average r^2 from 0.79 for chromosome 21 to 0.83 for chromosome 2) showing the robustness of our proposed approach. As expected we see accuracy increasing as function of coverage and minor allele frequency (Supplementary Figure 3). We also observe a high percentage of all SNPs to achieve high accuracy (Supplementary Figure 3) showing that our approach is appropriate for genome-wide scans.

8. Computing association statistics at imputed genotypes from extremely low-coverage short read sequencing

Using the 61 HIV controllers and 23 HIV progressors, we computed case-control association statistics at every SNP across the genome using either imputed data from sequencing or data from SNP arrays. As an association statistic we computed the standard Armitage trend test[10] defined as $N*\rho^2(G,\Phi)$, where N is equal to the number of samples, G is the vector of genotypes at given SNP and Φ is the phenotype vector (defined as 0 if sample is a Controller and 1 if Progressor[5]). This statistic has a χ^2 distribution with 1 degree of freedom and accounts for uncertainty in the imputation data if computed over dosages.

Supplementary Table 5 shows that imputation-based statistics recover the same signal (as measured by the value of the –log10 p-values) when compared to typed data. We observe instances in which statistics over imputed data attain greater significance than typed data, an effect likely due to statistical noise. To quantify whether the ratio of average –log 10 p-values in imputed versus typed data is significantly different from 1, we performed 1,000 permutations in which we randomly flipped the typed and imputed p-value at each SNP in the computation of the ratio. We observed that in more than 18% of the permutations, the ratio attained a value greater than 1.04 thus showing that the increase in significance of imputed data is most likely due to chance. Although our permutation approach makes the assumption that all SNPs are independent, the statistical noise is only increased when LD among SNPs is considered (only 4 SNPs were deemed as showing an independent signal of association in this region[5]) thus further reducing the significance of the p-value.

In addition, we computed the correlation between statistics computed over typed genotypes as compared to imputed data (Supplementary Figure 4) and we observed a Pearson squared correlation of 0.68 which as expected from imputed genotypes that are correlated with 0.83 to the typed data (Supplementary Figure 4). Supplementary Figure 4 shows the QQ-plot of typed versus imputed p-values attained on the Progressor phenotype in the IHCS exome data set.

9. Optimal coverage for sequencing-based GWAS under a fixed budget assumption

Previous work[1] has shown that under simplistic cost assumptions (e.g. no sample preparation costs) and in the absence of imputation, the optimal design for maximizing expected association power is attained at arbitrarily large sample sizes with arbitrarily low coverages per sample. Here we show that when realistic cost assumptions are taken into consideration, there exists an optimal coverage in short read sequencing for maximizing expected association power (effective sample size).

Using simulation experiments at various ultra-low coverages, we estimated the accuracy (average squared correlation across SNPs between typed and imputed genotyping calls) as a function of average coverage depth in sequencing $r^2(cov)$. For a given sample size N, sequenced at average coverage *cov*, we estimated the effective sample size as $Nr^2(cov)$. We note that by averaging correlations across all markers, we are giving every SNP equal weight in our estimation of effective sample size; various assumptions on distribution of causal variants may lead to different weights per SNP arising in different computations.

Assuming a fixed budged of $300,000, a sample preparation cost of $300 and cost per generating 1x of genome-wide coverage of $1,000 we computed the effective sample size when genotypes were estimated through imputation from reference panels (e.g. 1,000 Genomes) versus when no imputation is performed (see Supplementary Figure 5). We observe that there exists an optimal coverage for which the expected power (effective sample size) is maximized.

We explored the expected power of sequencing versus genotyping at different minor allele frequencies and odds ratio. Consider a case-control study in which we assess statistics at causal SNP using the standard z-score statistic[11]:

$$z = \frac{f^+ - f^-}{\sqrt{2/N}\sqrt{f(1-f)}} \sim \mathrm{N}(\frac{(f^+ - f^-)\sqrt{N}}{\sqrt{2f(1-f)}},1) = \mathrm{N}(\lambda\sqrt{N},1)$$

f is the frequency in controls, f^+ is the frequency in cases, f is the mean frequency, N in the total number of samples in the study (although we assumed balanced study with half cases and half controls, the statistic extends easily when the number of cases is different from number of controls[12]). Power at significance level α is then calculated from the non-centrality parameter (ncp) $\lambda\sqrt{N}$ as:

$$P(\alpha,\lambda\sqrt{N}) = \Phi(\Phi^{-1}(\alpha/2) + \lambda\sqrt{N}) + 1 - \Phi(\Phi^{-1}(1-\alpha/2) + \lambda\sqrt{N}) \qquad (1)$$

Φ denotes the standard normal cumulative distribution function and Φ^{-1} is the standard normal quantile function. The expected non-centrality parameter at imputed data with r^2 to the typed genotypes is estimated as $\lambda\sqrt{r^2N}$, which allows us to estimate the expected power in imputed data from sequencing.

Although we estimate the effective sample size using the average across loci of imputation accuracy (as measured by r^2), it is important to quantify the effect that the variance in r^2 (e.g. SNP accuracy varies by allele frequency, Supplementary Figure 5(a)) has on expected power. Therefore, we compared expected power using either the average accuracy across all SNPs of 0.82 plugged into equation (1) vs. the average power computed across all SNPs from the IHCS data (with their respective accuracies) thus incorporating variance in r^2 across SNPs. We observe (Supplementary Figure 5(b)) minor differences between the two strategies of computing the average power at any MAF considered.

We also computed expected theoretical power as function of allele frequency of a study with budget of $300,000 assuming genotyping array cost of $400 per sample as compared to sequencing cost of $133 per 1x and $30 per sample in DNA preparation (Supplementary Figure 5(c)).

10. Combined data set of 909 samples from IHCS, SCZ and AUT whole exome studies

Exome data was generated using the data processing and variant calling protocol described previously[13]. Reads were aligned to the reference genome using Burrows-Wheeler Aligner (BWA)[14], PCR duplicate reads were removed using Picard (see Main text, Web Resources), base quality scores were recalibrated using Genome Analysis Toolkit (GATK[7]), and alignments near putative indels were refined using GATK. Similar to the IHCS data, we observed significant amount of off-target data. Average coverage of 0.16x (0.30x) in SCZ (AUT) data across the 97% of the SNPs covered at coverage less than 4x and over 60x coverage at the remaining 3% of SNPs (threshold of 4x was chosen to attain near perfect separation between exome and non-exome SNPs, see above). We observe smaller off-target coverage in SCZ and AUT data as compared to ICHS data, most likely due to improvements in exome sequencing capture technologies. To remove effects from high coverage at or near exons, we removed all data at SNPs covered at 4x or more coverage. Our procedure for imputing genotypes from sequencing data follows three steps. First reads are aligned to hg19 reference human genome using the BWA aligner, using parameters “-q 5 -l 32 -k 2 -t 4 -o 1”. Reads with low mapping quality as well as reads mapping to multiple locations were removed. Second, starting from the bam aligned files, we computed genotype likelihoods at all loci identified as polymorphic in the 1,000 Genomes phase 1 project, using GATK[7] software (“GenomeAnalysisTK.jar -T UGCalcLikelihoods -out_mode EMIT_ALL_SITES”) for each sample independently in batches of 1M loci. Third, genotype likelihoods across all of the 909 samples were provided to the Beagle[2] imputation software, together with the 762 European haplotypes of 1000 Genomes phase 1 data to compute dosages at each sample and each site polymorphic in 1000 Genomes phase 1 (to improve Beagle runtime with no effect on accuracy, we restricted Beagle imputation to only the 15,709,633 sites found polymorphic in the 762 European haplotypes of the1000 Genomes phase 1 data). Imputation using Beagle was performed in windows of 1Mb in size with 250Kb flanking regions across all samples. 9 out of the 2764 Beagle window runs crashed due to very large CPU and memory requirements.

The 909 samples were also genotyped on a variety of genome-wide SNP arrays (see Main Text). We intersected all genotype array data (including the IHCS data) using plink[15] software, to achieve a data set of 909 samples typed on 104,454 genome-wide SNPs, with each SNP containing no more than 10% missing calls. After filtering for Hardy-Weinberg at a nominal threshold of 0.001, only 104,318 SNPs were retained for all subsequent analyses. 103,977 SNPs were successfully imputed using Beagle (the remaining 341 SNPs belonged to the 9 windows where Beagle imputation crashed). All results below were derived using the 103,977 genome-wide SNPs typed using arrays and imputed from sequencing data. We observed an r^2 of 0.71 between imputed genotyping calls and typed data across the 103,977 SNPs across all the 909 samples, with average r^2 of 0.69 for SCZ samples, 0.71 for AUT data and 0.83 for ICHS data, consistent with the different amounts of off-target reads in these data sets and with simulation results (see Main Text).

Starting from the typed genotype calls, we simulated phenotypes using an additive model Φ=g*beta+N(0,1), for various values of beta for each SNP independently to simulate 103,997 data sets. Statistics were computed at each SNP independently using the standard Armitage trend test defined as $N\rho^2(G, \Phi)$. Supplementary Figure 6 shows that association statistics over imputed dosages recover the same signal as association statistics computed at typed genotyping calls.

To search for potential batch effects of DNA collection and different sequencing approaches, we also performed a case-control analysis in which the AUT samples were treated as “controls” and SCZ samples as “cases”. Since all samples were not ascertained for any phenotype, this analysis creates a null data set with “real” phenotype, where the phenotype is defined as the cohort label. In this setup, population stratification due to genetic differences between SCZ and AUT data sets is a major concern[16]. Indeed, the standard metric of differentiation F_{st} between the AUT and SCZ samples has a value of 0.001272 (standard deviation of 2x10-5). Such a non-zero value of differentiation among the cases and controls leads to an expected inflation factor $\lambda_{GC} = 1+N*F_{st}$ of 2.02 (N=800 samples)[17] in association statistics computed when the cases and controls are sampled from AUT versus SCZ. As expected, we observed an inflation factor λ_{GC} of 1.96 (1.81) in association statistics computed over typed (imputed) data, λ_{GC} likely explained by differences in genetic ancestry between the AUT and SCZ populations. The standard approach for correcting for population substructure due to ancestry is to use PCA[16]; however, the approach of correcting using PCA is inappropriate in this scenario because all cases are sampled from one population and all controls are sampled from the other population (AUT vs. SCZ). To obtain properly distributed association statistics, we applied λ_{GC} to both typed or imputed association statistics. This would not be appropriate to rigorously correct for stratification[18], but if no false positives remain after applying this approach, we can conclude that the approach we propose is not susceptible to false positives. Supplementary Figure 7(a) shows the association statistics computed over typed versus imputed data demonstrating the high correlation between p-values (r^2=0.63), as expected from the imputation accuracy. Supplementary Figure 7(b) shows the QQ-plot of the association statistics, demonstrating that both imputed and typed association statistics attain expected distributions. Most importantly, we do not observe any genome-wide significant association between AUT vs. SCZ phenotype and any SNP in either typed or imputed genotypes.

Finally, we show that standard approaches from the genotyping arrays GWAS analysis toolkit extend to sequencing based GWAS. Using the EIGENSTRAT software, we performed principal component analysis on the 909 samples using either the genotypes typed in arrays (103,977 in total) or imputed from ultra low-coverage. We performed PCA on imputed data by rounding dosages at well-imputed SNPs. As metric for imputation accuracy that does not rely on typed data we used r^2hat[19] (threshold of 0.8, 37,796 in total). Supplementary Figure 8 shows that PCA over imputed data recovers the same principal axis of variation as data, separating the AUT from the SCZ samples. We observed a square correlation of 0.91 across the first Eigenvector. Supplementary Figure 8(bottom) suggests that the variation between PCA on typed versus imputed data most likely comes from smaller SNP set used for PCA over imputed data. Although in this simple experiment we assessed the capacity of imputed data to recover the principal components as inferred in typed data, we note that in standard GWAS over more than a million SNPs, a reduction of 70% to 300k SNPs for performing PCA over imputed data is likely going to recover similar axes of variation. Although our analysis shows that PCA can be performed using rounded dosages at accurately imputed SNPs, we caution that such an approach may lead to biases, especially when there is a sequencing depth difference between cases and controls and therefore special care should be taken in this scenario.

References

1 Sampson, J., Jacobs, K., Yeager, M., Chanock, S. & Chatterjee, N. Efficient study design for next generation sequencing. *Genetic Epidemiology* **35**, 269-277 (2011).

2 Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85**, 847-861, doi:S0002-9297(09)00519-9 10.1016/j.ajhg.2009.11.004 (2009).

3 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

4 Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-834, doi:10.1002/gepi.20533 (2010).

5 Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551-1557, doi:science.1195271 10.1126/science.1195271 (2010).

6 Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* **12**, R1, doi:10.1186/gb-2011-12-1-r1 (2011).

7 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

8 Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. Low coverage sequencing: Implications for the design of complex trait association studies. *Genome Res*, doi:10.1101/gr.117259.110 (2011).

9 Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

10 Armitage, P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11**, 375-386 (1955).

11 Zaitlen, N., Kang, H. M. & Eskin, E. Linkage effects and analysis of finite sample errors in the HapMap. *Hum Hered* **68**, 73-86, doi:10.1159/000212500 (2009).

12 Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**, 1-14, doi:10.1086/321275 (2001).

13 Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

14 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:btp324 10.1093/bioinformatics/btp324 (2009).

15 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:S0002-9297(07)61352-4 10.1086/519795 (2007).

16 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:ng1847 10.1038/ng1847 (2006).

17 Price, A. L. *et al.* The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet* **5**, e1000505, doi:10.1371/journal.pgen.1000505 (2009).

18 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, doi:nrg2813 10.1038/nrg2813 (2010).

19 Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**, 235-250, doi:S0002-9297(09)00020-2 10.1016/j.ajhg.2009.01.013 (2009).