nature genetics

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

Bogdan Pasaniuc[1–3], Nadin Rohland[3,4], Paul J McLaren[3,5], Kiran Garimella[3], Noah Zaitlen[1–3], Heng Li[3], Namrata Gupta[3], Benjamin M Neale[3,6], Mark J Daly[3,6], Pamela Sklar[7,8], Patrick F Sullivan[9], Sarah Bergen[3], Jennifer L Moran[3], Christina M Hultman[10], Paul Lichtenstein[10], Patrik Magnusson[10], Shaun M Purcell[3,6], David W Haas[11], Liming Liang[1–3], Shamil Sunyaev[3,5], Nick Patterson[3], Paul I W de Bakker[3,5,12,13], David Reich[3,4,14] & Alkes L Price[1–3,14]

Genome-wide association studies (GWAS) have proven to be a powerful method to identify common genetic variants contributing to susceptibility to common diseases. Here, we show that extremely low-coverage sequencing (0.1–0.5×) captures almost as much of the common (>5%) and low-frequency (1–5%) variation across the genome as SNP arrays. As an empirical demonstration, we show that genome-wide SNP genotypes can be inferred at a mean r^2 of 0.71 using off-target data (0.24× average coverage) in a whole-exome study of 909 samples. Using both simulated and real exome-sequencing data sets, we show that association statistics obtained using extremely low-coverage sequencing data attain similar *P* values at known associated variants as data from genotyping arrays, without an excess of false positives. Within the context of reductions in sample preparation and sequencing costs, funds invested in extremely low-coverage sequencing can yield several times the effective sample size of GWAS based on SNP array data and a commensurate increase in statistical power.

Genome-wide association studies have identified over a thousand SNPs associated with complex traits[1]. To date, these studies have been carried out using SNP arrays that assay up to 2.5 million polymorphisms at a cost of hundreds of dollars per sample, with these data often augmented by imputation of non-genotyped variants using the HapMap or 1000 Genomes Project reference panels[2–5]. At the same time, DNA sequencing has emerged as a powerful new technology[3,6,7], with the first major applications to disease gene discovery arising in the course of exome sequencing[8]. Recent cost reductions raise the question of whether sequencing might be a viable alternative for GWAS, analogous to RNA sequencing (RNA-seq) in gene expression studies[3,9,10]. One limitation to using sequencing for GWAS has been the cost of preparing each DNA sample, which historically has been at least as expensive as SNP array genotyping. However, this is no longer the case; for example, Epicentre offers high-throughput sample preparation for roughly $100 per sample (see URLs), and we have recently shown that sequencing libraries appropriate for whole-genome sequencing can be produced for approximately $15 per sample on a scale of thousands of samples[11]. In this paper, we show that, by sequencing such libraries at extremely low coverage (0.1–0.5×, at an effective sequencing cost of $10–100 per sample) combined with genotype calling using 1000 Genomes Project reference panels[2], the effective sample size per unit cost of this approach is several times greater than for the standard GWAS study design using SNP arrays. This gap will increase if sequencing costs continue to drop more quickly than genotyping costs.

RESULTS

To explore the effectiveness of GWAS based on low-coverage sequencing, we simulated sequencing data at various coverage levels, accounting for sequencing errors, as well as for variation in average coverage across samples and loci. We used the 762 haplotypes inferred from the 381 European samples of the 1000 Genomes Project (Phase 1, June 2011 release) and restricted the analysis to 10 distinct 5-Mb regions (total of 50 Mb, containing 150,261 SNPs) that were randomly chosen to represent the average genome-wide recombination rate and SNP density (**Supplementary Note** and **Supplementary Table 1**). One-half of the haplotypes were used to build simulated data, and the other half were used as an imputation reference panel. Simulated data were used to infer genotype dosages at known SNPs using Beagle[12], an imputation engine appropriate for the analysis of sequencing data. To assess the accuracy of imputation,

[1]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [3]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [5]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [7]Department of Psychiatry, Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York, USA. [8]Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, New York, USA. [9]Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA. [10]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [11]Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [12]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [13]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [14]These authors jointly directed this work. Correspondence should be addressed to B.P. (bpasaniu@hsph.harvard.edu), D.R. (reich@genetics.med.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 13 December 2011; accepted 16 April 2012; published online 20 May 2012; doi:10.1038/ng.2283

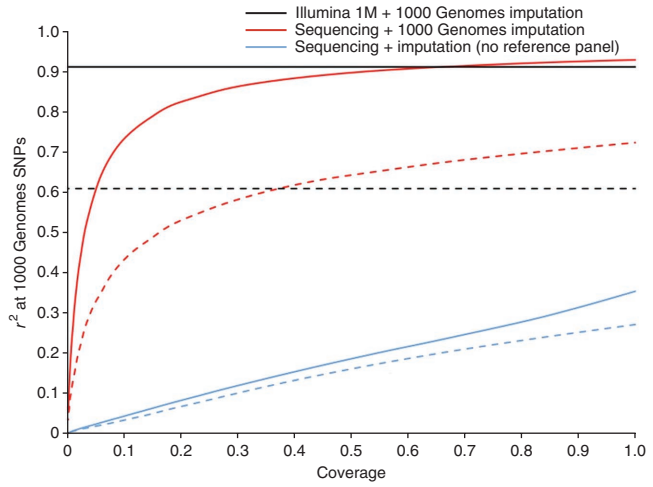

Figure 1 Genotype imputation accuracy as function of coverage in 1000 Genomes Project simulations. Accuracy as function of coverage is shown using solid lines for common SNPs (MAF > 5%) and dashed lines for low-frequency SNPs (MAF < 5%).

we used the squared correlation (r^2) between imputed dosages and true genotypes, which quantifies the reduction in effective sample size in GWAS due to imperfect imputation[13] (Online Methods).

The accuracy of imputation, either using just the sequencing reads to impute genotypes or using the reads coupled with the 1000 Genomes Project reference panels[2] (Online Methods), are shown (**Fig. 1**). We observed high accuracies at extremely low coverage (0.1–0.5×) when reference panels were used (**Fig. 1**, **Supplementary Fig. 1** and **Supplementary Note**). Sequencing at 0.2× coverage yielded more than 90% of the effective sample size that was achieved by Illumina Human-1M-Duo array plus conventional imputation, as assessed by average r^2 to SNPs in the 1000 Genomes Project data set for both common (>5% minor allele frequency) as well as low-frequency (1–5% minor allele frequency) variants (**Fig. 1**). These simulation results suggest that sequencing at 0.1–0.5× coverage with imputation using the 1000 Genomes Project data sets can, in principle, achieve power comparable to high-density SNP arrays. These simulation results are robust to model assumptions and parameter values (**Supplementary Tables 1–3** and **Supplementary Note**).

We investigated whether similar results could be achieved with real data by analyzing whole-exome sequencing data from 909 individuals of European ancestry, combining samples from the International HIV Controllers Study (IHCS) (84), Swedish Schizophrenia Study (SCZ) (503) and Autism National Institute of Mental Health (NIMH) Controls Study (AUT) (322) (Online Methods)[14–18]. Whole-exome studies enrich the sample DNA for genic content before sequencing[3,19,20] and usually discard data from non-exonic regions. However, current DNA capture technologies do not yield perfect enrichment, and off-target data can often be substantial, given the high coverage of many exome-sequencing studies. For example, in the 909 exomes included, the average coverage was 0.24× for non-exonic regions and more than 60× for exons (**Supplementary Fig. 2** and **Supplementary Note**). We explored whether the whole-exome data, coupled with imputation based on the 1000 Genomes Project reference data set, could support a GWAS. We imputed genotypes at all polymorphic sites indentified in the European samples of the 1000 Genomes Project, using sequencing data together with the 762 haplotypes inferred from the European samples of the 1000 Genomes Project Phase 1 data (Online Methods), and quantified accuracy by comparing imputed calls with Illumina array genotyping calls (Online Methods). To remove effects of high coverage at or near exons, we removed data at all SNPs covered at more than 4× (**Supplementary Fig. 2**). At 0.24× coverage, we observed an average $r^2 = 0.71$ (s.d. = 0.15) to the genotype calls assayed by genome-wide SNP arrays, roughly similar in average expected power to a conventional GWAS with 71% of the sample size (**Supplementary Fig. 3**, **Supplementary Table 4**, results averaged by chromosome, minor allele frequency and coverage, and **Supplementary Note**). We also quantified the genome-wide accuracy achieved by using all data from the whole-exome scan (off-target and on-target data); the average r^2 value increased to 0.77 when all data from the whole-exome study were used.

To illustrate how this approach might be used in practice to carry out a GWAS, we used the off-target exome data to compute association statistics at 103,977 SNPs across the genome using simulated phenotypes starting from the genotype calls from the arrays (Online Methods). We observed similar association statistics when imputed dosages were used compared to SNP arrays under both null (phenotype uncorrelated to the genotype) and true nonzero effect sizes (**Fig. 2**, **Supplementary Figs. 4–6** and **Supplementary Table 5**), indicating that our approach is robust to false positives, while accurately recovering the association signal when present. In addition, we also performed a case-control scan in which the AUT samples were treated as controls and the SCZ samples were defined as cases. After adjusting for differences in genetic ancestry between the SCZ and AUT samples, we observed no association at genome-wide significance, thus further emphasizing the robustness of our approach (**Supplementary Fig. 7** and **Supplementary Note**). To assess the power of detecting true positives, in addition to simulated phenotypes, we also carried out a case-control study comparing HIV-1 controllers (61) and progressors (23) from the IHCS data set (Online Methods). The higher off-target coverage (0.5×) in the IHCS data led to an average of $r^2 = 0.82$ to the genotype calls at the 398,098 SNPs assayed by arrays in the IHCS data[14]. A similar λ_{GC} (genomic control)[21] value of 1.05 for imputed

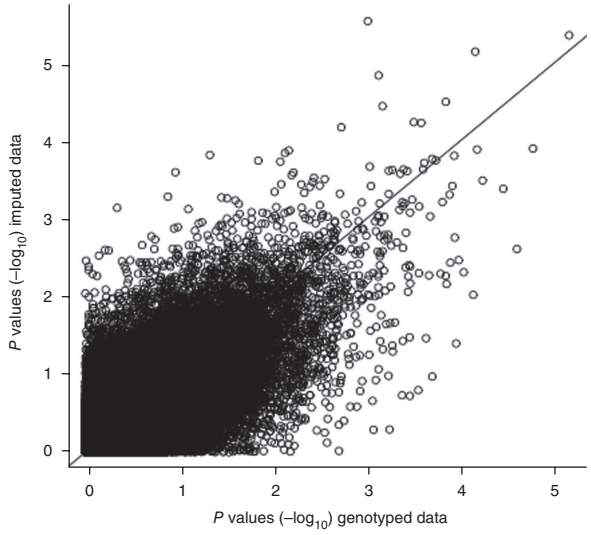

Figure 2 Observed versus expected association $-\log_{10}$ P values at 103,977 SNPs across the genome in simulated null data sets over 909 samples of the combined data set. We observed r^2 of 0.64 between P values computed in genotyped versus imputed data, similar to simulations of association statistics at imputed versus genotyping calls (**Supplementary Note**). Results for alternate hypothesis of association can be found in the **Supplementary Note**.

Table 1 Statistics attained at known associated SNPs in the IHCS

RsID	Chr.	Position	Coverage	r^2	Association *P* value ($-\log_{10}$) for genotyped data	Association *P* value ($-\log_{10}$) for imputed data	Ratio	Effect typed (confidence interval)	Effect imputed (confidence interval)
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)
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)
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)
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)
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)
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)
	Average		0.48	0.90	1.67	1.74	1.04[a]	–	–

Statistics were computed over genotyped or imputed genotypes (only SNPs with nominal *P* value < 0.05 in the genotyped data are shown). Chr., chromosome.
[a]Average ratio is computed as the ratio of the sum of association *P* values. Effect is computed assuming a linear additive model associating genotype to phenotype.

data compared to 1.04 for directly genotyped data was observed (**Supplementary Fig. 4** and **Supplementary Note**). We specifically analyzed SNPs that were previously reported to be significantly associated with HIV-1 controller status[14] and observed similar association statistics and effect sizes compared to SNP arrays, both for the entire set of 47 previously associated SNPs (**Supplementary Table 5** and **Supplementary Note**) and for the subset of 10 SNPs with nominal $P < 0.05$ in the SNP array data (**Table 1**). The association statistics obtained using extremely low-coverage sequencing did not show the 9% drop that might have been expected given the $r^2 = 0.91$ imputation accuracy at these SNPs (ratio between the average $-\log_{10} P$ values at imputed versus genotyped data of 1.04), but this can be explained by statistical fluctuation (**Table 1** and **Supplementary Note**).

We also evaluated empirical results at lower coverage (0.005–0.5×) by subsampling reads with corresponding probability. Because of the large number of experiments and the higher non-exome coverage of the IHCS data compared to all the 909 samples, we restricted this analysis to the 10 distinct 5-Mb regions (total of 50 Mb) in the IHCS data set (84 samples). As coverage decreased, we observed a reduction in accuracy, analogous to our simulations based on the 1000 Genomes Project data set, restricted to the same set of 6,070 SNPs from the array (**Fig. 3**). At 0.5× coverage, we observed a mean r^2 of 0.82, with s.d. of 0.03 and standard error of 0.01 across the ten regions. However, the accuracy of imputation in the IHCS sequencing data was lower than in simulations for any level of coverage (**Fig. 3**). The discrepancy between simulations and real data could be an effect of increased similarity across haplotypes inferred from the 1000 Genomes Project Phase 1 data due to the genotype calling and phasing procedure from 4× sequencing data that aggregated information across samples (**Supplementary Table 6** and **Supplementary Note**). Other possible explanations include nonuniform error rates in base-calling and alignment of reads across the genome or simulation parameters that do not perfectly model aspects of the empirical data, such as variance in coverage across samples and loci, although our experiments suggest that these are unlikely to be the primary explanations (**Supplementary Note**).

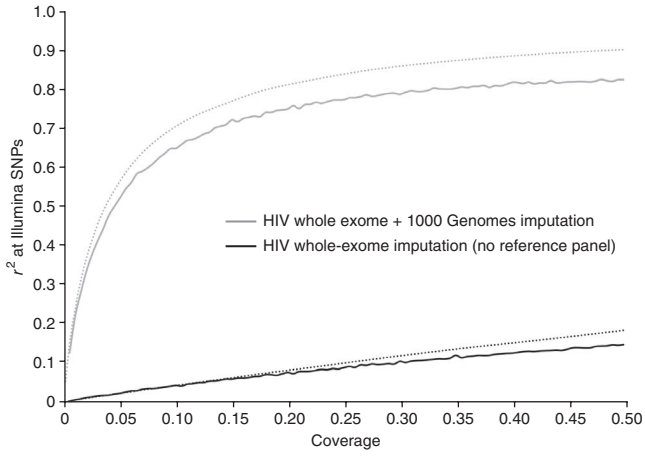

Figure 3 Genotype imputation accuracy in IHCS whole-exome data as a function of coverage (solid lines). Illumina 1M genotype calls were used as the standard, with restriction to 6,070 SNPs in 10 distinct 5-Mb regions (total of 50 Mb) of the genome. Dotted lines denote results attained in 1000 Genomes Project simulations on the same set of SNPs.

DISCUSSION

To explore the economic ramifications of sequencing-based GWAS, we considered the tradeoff between the number of samples sequenced and average coverage (which affects accuracy). We evaluated the expected effective sample size attained with different strategies and compared this with the effective sample size that would be obtained by genotyping using standard genotyping arrays (for example, the Illumina Human-1M-Duo array). We derived all results from empirical accuracies using sequencing data sets subsampled from the IHCS data, so that results did not rely on any simulation assumptions. We compared accuracies only at SNPs genotyped on the array, a conservative computation that ignored the potentially greater benefit at SNPs not present on the array. We assumed a fixed total budget of $300,000, an arbitrarily large number of samples available, a sample preparation cost of $30 (conservatively double the cost that we have recently shown[11]) and DNA sequencing cost of $133 per 1× sequencing (based on the Illumina Network cost of $4,000 for 30× sequencing of 50 samples or more, which scales linearly with lower coverage). We calculated the effective sample size of a sequencing-based GWAS as a function of average coverage, which determines the number of samples sequenced under a fixed budget (Online Methods). Under zero sample preparation cost and ignoring the benefit of imputation, the optimal study design involves sequencing a maximal number of samples at minimal coverage[22,23]. However, when sample preparation cost and imputation are taken into account, there is an optimal number of samples to sequence for any budget. For a fixed budget of $300,000, the highest effective sample size (roughly equivalent to more than 4,600 genotyped individuals) is achieved at an average coverage of 0.1× (6,800 samples sequenced at $45 total cost per

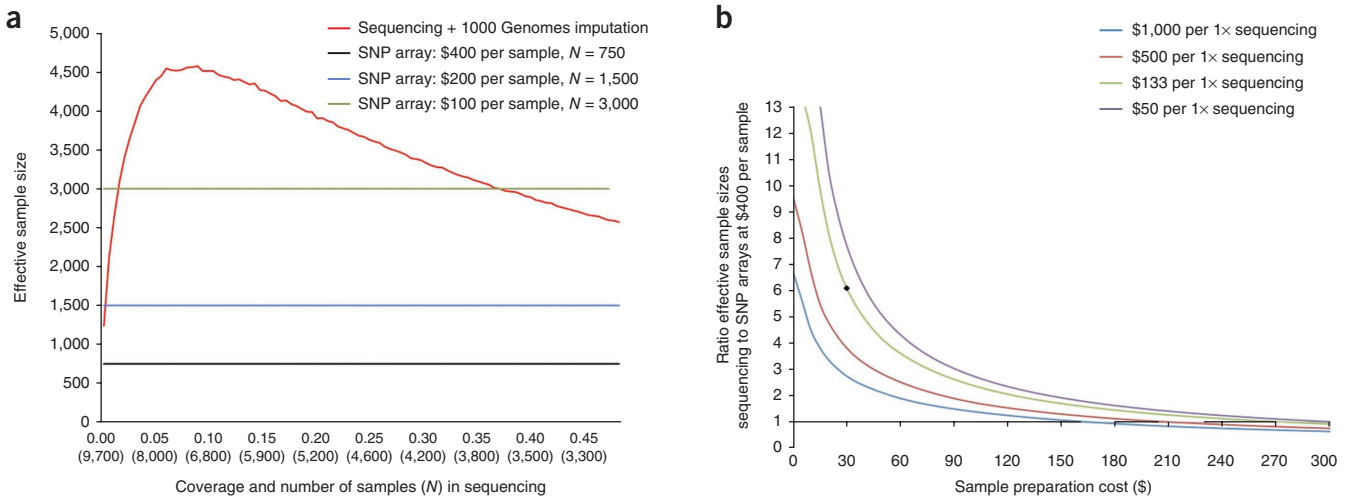

Figure 4 Coverage and corresponding number of samples for a fixed budget of $300,000. (**a**) Effective sample size in sequencing-based GWAS as a function of the number of samples and resulting coverage. Cost assumptions: $30 per sample preparation cost, $133 per 1× coverage sequencing cost. (**b**) Ratio of expected association statistic (effective sample size) in sequencing-based GWAS versus array-based GWAS at $400 per sample, as a function of sample preparation and sequencing costs. Expected association statistics for sequencing-based GWAS are based on optimum coverage and number of samples (assuming arbitrarily large number of samples available) subject to budget constraint. The optimum coverage and number of samples vary at different points on the graph (not shown). The black dot denotes $30 per sample preparation cost and $133 per 1× coverage.

sample, $r^2 = 0.65$) (**Fig. 4a**). The optimal value of average coverage varies as a function of sample preparation and sequencing costs, but we obtained qualitatively similar results for other cost assumptions (**Supplementary Note**). We note that a sequencing-based approach can attain a higher effective sample size than SNP arrays, even when constraints on sample availability limit the space of available study designs (**Fig. 4a**).

A notable finding is that the effective sample size achieved using sequencing-based GWAS with current costs[11] is more than six times higher than SNP-array genotyping at $400 per sample, corresponding to a large increase in power (**Fig. 4b**, **Supplementary Fig. 8** and **Supplementary Note**). Only if SNP array genotyping is less than $70 per sample or if sample preparation and sequencing costs are much higher (for example, greater than $120 per sample for sample preparation or $1,000 for 1× sequencing) does sequencing-based GWAS lose its advantage in terms of statistical power to associate variants. If sequencing technology—both in terms of the efficiency of library preparation and the cost of sequencing—continues to improve more quickly than genotyping technology, the advantage of sequencing-based GWAS will increase. We note that a critical ingredient for attaining high accuracy at extremely low coverage is the availability of large panels of reference haplotypes. As additional reference haplotypes over larger numbers of SNPs become available from the 1000 Genomes Project and other projects, we expect the accuracy attained by extremely low-coverage sequencing to further increase.

We conclude with several caveats. First, computational methods for sequencing-based GWAS are still under development[3,7,22,24], whereas the SNP array–based GWAS is a proven method that produces high-quality data that can be analyzed using readily available computational tools. Second, sequencing data require additional computational resources beyond what is necessary to analyze conventional GWAS data, as the analysis pipeline for sequencing data is typically more demanding than for genotyping data. Third, sequencing-based GWAS of the type described here do not involve sufficient coverage to discover rare variants and to associate them with disease; thus, as with SNP arrays, the power of this approach is limited to common and (to a lesser extent) low-frequency variants. Fourth, although results from our empirical IHCS sequencing data are encouraging, no study to date has used sequencing-based GWAS to identify new disease risk variants. A priority for future work should be to carry out studies that show that this approach can be used to discover new associations between genetic variants and common diseases.

URLs. The 1000 Genomes Project, June 2011 Phase 1 release, http://www.1000genomes.org/node/506; Beagle, http://faculty.washington.edu/browning/beagle/beagle.html; MACH, http://www.sph.umich.edu/csg/abecasis/MACH/index.html; Picard, http://picard.sourceforge.net/index.shtml; GATK, http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit; Epicentre sample preparation, (accessed on 1 November 2011), http://www.epibio.com/; NIMH controls, https://www.nimhgenetics.org/available_data/controls/; Illumina Human1M-Duo array, http://www.illumina.com/products/human1m_duo_dna_analysis_beadchip_kits.ilmn; Illumina Network, http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&id=1561106; The International HIV Controllers Study, http://www.hivcontrollers.org/; sample repository research concept sheet, http://cfar.globalhealth.harvard.edu/fs/docs/icb.topic938249.files/Harvard%20CFAR%20Concept%20Sheet%20Template%20.docx.

METHODS

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

Accession codes. AUT data (phs000298.v1) and SCZ data (phs000473.v1.p1) have been deposited at dbGaP. IHCS data are available by direct request from P. Richtmyer (prichtmyer@partners.org); investigators can submit a concept sheet detailing their study design, research questions and other needs in order to request access to IHCS genetic data. The concept sheet with detailed instructions can be downloaded (see URLs). Requests will be reviewed on the basis

of scientific merit, feasibility and potential overlap with accepted concept sheets or ongoing investigations.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to acknowledge the ARRA Autism Sequencing Consortium (AASC) principal investigators for use of the autism data sets, including E. Boerwinkle, J.D. Buxbaum, E.H. Cook Jr., M.J. Daly (communicating principal investigator), B. Devlin, R. Gibbs, K. Roeder, A. Sabo, G.D. Schellenberg and J.S. Sutcliffe. We thank T. Lehner, A. Felsenfeld and P. Bender for their support and contribution to the AASC project and to the generation of AUT sequencing data. This research was supported by US National Institutes of Health (NIH) grants (R01 HG006399 to B.P., N.P., D.R. and A.L.P. and R01 MH084676 to S.S.). The IHCS acknowledges generous support from the Mark and Lisa Schwartz Foundation and the Collaboration for AIDS Vaccine Discovery of the Bill and Melinda Gates Foundation. The IHCS was also supported in part by NIH grants (P-30-AI060354 to the Harvard University Center for AIDS Research, AI069513, AI34835, AI069432, AI069423, AI069477, AI069501, AI069474, AI069428, AI69467, AI069415, Al32782, AI27661, AI25859, AI28568, AI30914, AI069495, AI069471, AI069532, AI069452, AI069450, AI069556, AI069484, AI069472, AI34853, AI069465, AI069511, AI38844, AI069424, AI069434, AI46370, AI68634, AI069502, AI069419, AI068636 and RR024975 to the AIDS Clinical Trials Group and AI077505 to D.W.H.). Data generation for the NIMH controls was directly supported by NIH grants (R01MH089208, R01 MH089025, R01 MH089004 and R01 MH089482). SCZ data generation was supported by an NIMH grant (5RC2MH089905; P.S. and S.M.P.) and by the Sylvan Herman Foundation and the Stanley Medical Research Institute (a gift to the Stanley Center for Psychiatric Research).

AUTHOR CONTRIBUTIONS

B.P., N.R., N.P., A.L.P. and D.R. conceived and designed the study. B.P. conducted the analyses. L.L., S.S., N.R., P.J.M., N.Z. and H.L. provided bioinformatics and statistical support. P.I.W.d.B., N.G., K.G., B.M.N., M.J.D., P.S., P.F.S., S.B., J.L.M., C.M.H., P.L., P.M., S.M.P. and D.W.H. recruited and provided samples and data for these analyses. B.P., A.L.P. and D.R. wrote the paper. All authors contributed to the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2283.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. 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).
2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
4. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
5. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
6. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
7. Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
8. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
9. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
10. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
11. Rohland, N. & Reich, D. Cost-effective high-throughput DNA sequencing libraries. *Genome Res.* published online, doi:10.1101/gr.128124.111 (20 January 2012).
12. 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 (2009).
13. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
14. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
15. Suarez, B.K. *et al.* Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. *Am. J. Hum. Genet.* **78**, 315–333 (2006).
16. O'Donovan, M. C. *et al.* Analysis of 10 independent samples provides evidence for association between schizophrenia and a SNP flanking fibroblast growth factor receptor 2. *Mol. Psychiatry* **14**, 30–36 (2009).
17. The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.* **39**, 1045–1051 (2007).
18. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
19. Musunuru, K. *et al.* Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).
20. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
21. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
22. Sampson, J., Jacobs, K., Yeager, M., Chanock, S. & Chatterjee, N. Efficient study design for next generation sequencing. *Genet. Epidemiol.* **35**, 269–277 (2011).
23. Kim, S.Y. *et al.* Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.* **34**, 479–491 (2010).
24. Le, S.Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **21**, 952–960 (2011).

ONLINE METHODS

Simulation of sequencing data based on the 1000 Genomes Project data set. For our simulations, we used the 381 diploid European individuals from the Phase 1 release of the 1000 Genomes Project (June 2011)[2]. The 381 individuals included 87 Centre d'Etude du Polymorphisme Humain (CEPH) individuals of Northern European ancestry (CEU), 93 Finnish individuals from Finland (FIN), 89 British individuals from England and Scotland (GBR), 98 Tuscan individuals (TSI) and 14 individuals from the Iberian peninsula (IBS). Genotype calls and haplotype phase were inferred from low-coverage sequencing (4×) using an imputation strategy that borrowed information across samples and loci. The 762 haplotypes were divided at random between two panels of 381 haplotypes; one panel was used to build simulated data, and the other was used as an imputation reference panel. We simulated data for 100 samples by randomly sampling (without replacement) pairs of haplotypes from the simulation panel. All simulation results were generated over ten distinct 5-Mb regions across the genome (for a total of 50 Mb), which were randomly chosen to represent the average genome-wide recombination rate and SNP density (**Supplementary Note**). Reads spanning polymorphic sites identified in the 1000 Genomes Project were simulated assuming a fixed error rate of 1%, per-locus coverage multipliers were drawn from a gamma distribution $\Gamma(\alpha,\beta)$, with shape parameters $\alpha = 4$ and $\beta = 1/\alpha$ and mean = 1 (ref. 25), and per-sample coverage multipliers were drawn from a normal distribution $N(1,0.2)$ (matching the empirical IHCS sequencing data), with negative values set to 0. Reads were sampled assuming a Poisson distribution, with the mean equal to the average coverage × the per-locus multiplier × the per-sample multiplier. Results were generally insensitive to the choice of simulation parameters (with the exception of average coverage per sample) (**Supplementary Note**).

Imputing genotypes from sequencing data. Genotypes can be inferred from sequencing data by either (i) inferring genotypes independently at each SNP in each individual, (ii) making use of allele frequencies inferred from all sequenced individuals, (iii) making use of linkage-disequilibrium (LD) patterns inferred from sequenced individuals or (iv) making use of LD patterns inferred from sequenced individuals as well as reference panels of haplotypes[7,22,24,26]. Here, we focused on methods (iii) and (iv), using a two-step imputation approach (see **Supplementary Note** for details and results of other approaches). In the first step, we computed genotype likelihoods at all polymorphic loci identified in the 1000 Genomes Project data set independently for each individual. We disregarded all observed alleles that did not match either the reference or alternate allele identified in the 1000 Genomes Project data set and computed likelihoods of zero, one or two copies of the 1000 Genomes Project data set reference allele at all SNPs identified in the Phase 1 release of the 1000 Genomes Project. Reads that did not overlap any polymorphic sites were discarded. In the second step, the genotype likelihoods for all loci in all samples (with or without the reference panel of haplotypes, 381 in total for simulations) were passed to the Beagle imputation software[12] with default parameters (with 'like' for the genotype likelihoods and 'phased' for the reference haplotypes).

Imputing genotypes from GWAS arrays. Imputation from the Illumina Human-1M-Duo array was simulated by masking all genotypes at SNPs (in the 50-Mb simulated region) not present on the array and then performing imputation at all polymorphic loci identified in the European samples of the 1000 Genomes Project Phase 1 data set, using the remaining reference panel of haplotypes (381 in total). We used MaCH[27] imputation software with default parameters: –rounds 40 –greedy –mle –mldetails.

Metric for imputation accuracy. Imputation accuracy was measured by the squared Pearson's correlation coefficient (r^2) between imputed dosages and genotypes typed on the arrays.

Simulated phenotypes. Starting from the typed genotype calls (g), we simulated continuous randomly ascertained phenotypes as $Y = g\beta + \varepsilon$, with $\varepsilon = N(0,1)$. $\beta = 0$ represents the null model of no association between genotype and phenotype.

IHCS whole-exome data set. Genome-wide SNP genotyping and whole-exome sequencing data for 84 samples were obtained from the IHCS[14], with 43 samples genotyped on the Illumina HumanHap 650Y array and 41 sequenced on the Human-1M-Duo array. Of the 84 samples, 61 were HIV-1 controllers enrolled by the IHCS, and 23 were enrolled by the AIDS Clinical Trials Group. Only unrelated samples of European ancestry with high genotyping rates (>95%) were included after filtering out SNPs with low MAF (<1%), missing data (>2%) or departure from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$). The SNP sets were intersected to obtain 398,098 SNPs genotyped in all samples. Imputation was performed using all the 762 available 1000 Genomes Project Phase 1 haplotypes as opposed to 381 for simulations using non-overlapping regions of 2.5 Mb in size with 250 kb of flanking sequence on either side.

Combined whole-exome data set. Exome sequencing for the Autism NIMH Controls (AUT, 322 samples), for the SCZ set (503 samples) and for the IHCS set (84 samples) was carried out at the Broad Institute[14–18]. We only used samples ascertained as controls in the AUT and SCZ data (from individuals with no presence of disease). Exons were captured using Agilent 38Mb SureSelect v2 Libraries and were sequenced using either an Illumina HiSeq2000 or Illumina Genome Analyzer II instrument. All samples met the criterion of >90% of targeted bases having >10× coverage and >80% of targeted bases having >20× coverage. Reads were mapped to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA) and were processed with Picard and GATK (see URLs). The SCZ samples were genotyped on the Affymetrix 5.0 or 6.0 platforms. The AUT samples were genotyped on the Affymetrix 500K array. Genotype data across all samples (SCZ, AUT and IHCS, 909 in total) was merged with SNPs filtered by missing data and departure from Hardy-Weinberg equilibrium. Genotype likelihoods obtained using GATK[28] were passed to Beagle in windows of 1 Mb with 250 kb of flanking sequence on either side to impute all SNPs identified as polymorphic in the haplotypes of the European 1000 Genomes Project Phase 1 data. In total, 103,977 genome-wide SNPs, both genotyped and imputed from sequencing across all 909 samples, were used in all experiments over combined data (**Supplementary Note**). To remove effects of high coverage at or near exons, we removed data at all SNPs covered at more than 4×.

Association statistic for GWAS. A standard test for association in GWAS is the Armitage trend test[21,29], equal to N times the squared correlation between genotypes G (0, 1 or 2) and phenotypes Φ (0 or 1), where N is the number of samples. This statistic extends to imputed data by using genotype dosages. The value of the statistic decreases by a factor of r^2 if computed at a genotyped or imputed SNP in partial LD with the causal SNP[13]. To estimate the expected association statistic in a GWAS over a set of N samples sequenced at average coverage c, we first estimated the accuracy $r^2(c)$ attained at coverage c by subsampling IHCS data. We then estimated the expected association statistic as $N\rho^2(G,\Phi)\, r^2(c)$.

25. Prabhu, S. & Pe'er, I. Overlapping pools for high-throughput targeted resequencing. *Genome Res.* **19**, 1254–1261 (2009).
26. Bansal, V. *et al.* Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.* **20**, 537–545 (2010).
27. 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 (2010).
28. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
29. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).

doi:10.1038/ng.2283