As a library, NLM provides access to scientific literature. Inclusion in an NLM database does not imply endorsement of, or agreement with, the contents by NLM or the National Institutes of Health.
Learn more: PMC Disclaimer | PMC Copyright Notice

HHS Public Access
Author manuscript
Peer-reviewed and accepted for publication
About author manuscripts | Submit a manuscript

Nat Genet. Author manuscript; available in PMC 2012 Dec 1.
Published in final edited form as:
Nat Genet. 2012 May 20; 44(6): 631–635.
Published online 2012 May 20. doi: 10.1038/ng.2283

PMCID: PMC3400344
NIHMSID: NIHMS390286
PMID: 22610117

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] Pamela Sklar,[6] Patrick F. Sullivan,[7] Sarah Bergen,[3] Jennifer L. Moran,[3] Christina M. Hultman,[8] Paul Lichtenstein,[8] Patrik Magnusson,[8] Shaun M. Purcell,[9] David W. Haas,[10] Liming Liang,[1,2,3] Shamil Sunyaev,[3,5] Nick Patterson,[3] Paul I.W. de Bakker,[3,5,11] David Reich,[3,4,*±] and Alkes L. Price[1,2,3,*±]

Abstract

Genome wide association studies (GWAS) have proven a powerful method to identify common genetic variants contributing to susceptibility to common diseases. Here we show that extremely low-coverage sequencing (0.1–0.5x) 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 (0.24x average coverage) in a whole-exome study of 909 samples. Using both simulated and real exome sequencing datasets we show that association statistics obtained using ultra low-coverage sequencing data attain similar P-values at known associated variants as genotyping arrays, without an excess of false positives. Within the context of reductions in sample preparation and sequencing costs, funds invested in ultra low-coverage sequencing can yield several times the effective sample size of SNP-array GWAS, and a commensurate increase in statistical power.

Genome-wide association studies (GWAS) have identified over a thousand SNPs associated
plex traits[1]. To date, these studies have been carried out using SNP arrays that assay up to
lion polymorphisms at a cost of hundreds of dollars per sample, often augmented by imput
of untyped variants using HapMap or 1000 Genomes reference panels[2–5]. At the same time

Feedback
Back to Top

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 large as the cost of SNP array genotyping. However, this is no longer the case; for example Epicentre offers a high throughput sample preparation for roughly $100 per sample (see URLs), and we have recently demonstrated that sequencing libraries appropriate for whole-genome sequencing can be produced for approximately $15 per sample on a scale of thousands of samples[11]. Below, we show that by sequencing such libraries at ultra-low coverage (0.1–0.5x, at an effective sequencing cost of $10–$100 per sample) followed by 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 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 5Mb 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, 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 analysis of sequencing data. To assess the accuracy of imputation, 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).

Figure 1 shows 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). We observe high accuracies at ultra low-coverage (0.1–0.5x) when reference panels are used (Figure 1, Supplementary Note, Supplementary Figure 1). Sequencing at 0.2x coverage yields more than 90% of the effective sample size than is achieved by Illumina Human-1M-duo array plus conventional imputation, as assessed by average r^2 to SNPs in the 1000 Genomes Project dataset for both common (>5% minor allele frequency) as well as low-frequency variants (1 to 5% minor allele frequency) (Figure 1). These simulations results suggest that sequencing at 0.1–0.5x coverage with imputation using the 1000 Genomes Project datasets can in principle achieve power comparable to high-density SNP arrays. These simulation results are robust to model assumptions and parameter values (Supplementary Note, Supplementary Tables 1,2,3).

Foresight EX1213
Foresight v Personalis
IPR2023-00317

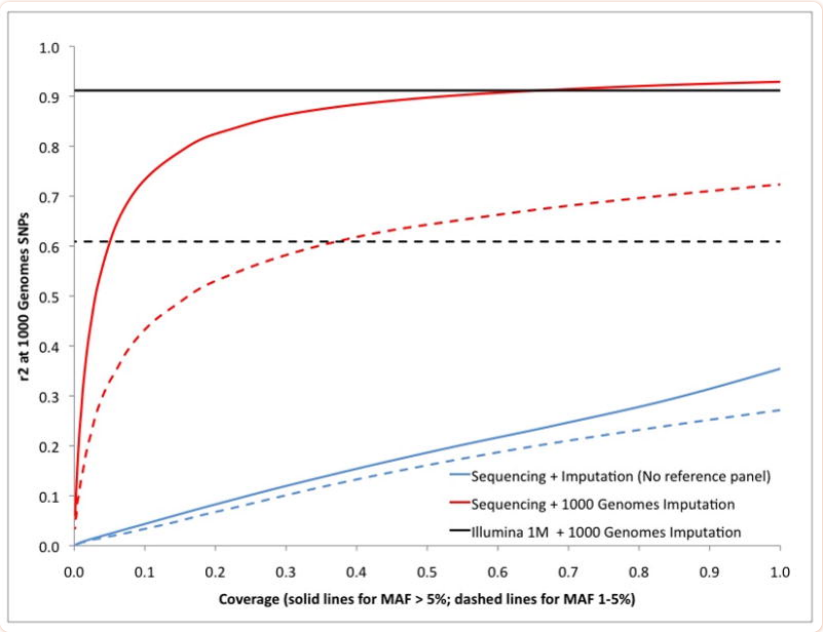

Figure 1

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

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 NIHM Controls Study (AUT) (322) (Online Methods)[14–18]. Whole-exome studies enrich the sample DNA for genic content prior to sequencing[3,19,20] and usually discard data from non-exonic regions. However, current DNA capture technologies do not yield perfect enrichment and the "off-target" data can often be substantial given the high coverage of many exome-sequencing studies. For example, in the 909 exomes, the average coverage is 0.24x for non-exonic regions and more than 60x for exons (Supplementary Note, Supplementary Figure 2). We explored whether the whole-exome data, coupled with imputation based on the 1000 Genomes Project reference dataset, 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 (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 4x (Supplementary Figure 2). At 0.24x coverage we observe 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 (see Supplementary Note, Supplementary Figure 3, Supplementary Table 4 for results averaged by

chromosome, minor allele frequency and coverage). We also quantified the genome-wide accuracy achieved by using all data from the whole exome scan (off-target and on-target); the average r^2 increased to 0.77 when all data from the whole-exome study was 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 as compared to SNP arrays under both null (phenotype uncorrelated to the genotype) and true nonzero effect sizes (Figure 2, Supplementary Figure 4,5,6, 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 SCZ as "cases". After adjusting for differences in genetic ancestry between SCZ and AUT samples, we observed no genome-wide significant association, thus further emphasizing the robustness of our approach (Supplementary Note, Supplementary Figure 7). 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 (0.5x) in the IHCS data leads 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 data as compared to 1.04 for typed data was observed (Supplementary Note, Supplementary Figure 4). We specifically looked at SNPs previously reported to be significantly associated with HIV-1 controller status[14] and observed similar association statistics and effect sizes as compared to SNP arrays, both for the entire set of 47 previously associated SNPs (Supplementary Note, Supplementary Table 5) 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 exhibit the 9% drop that might have been expected given r^2=0.91 imputation accuracy at these SNPs (ratio between the average −log 10 p-values at imputed versus typed data of 1.04), but this can be explained by statistical fluctuation (Table 1 and Supplementary Note).

Foresight EX1213
Foresight v Personalis
IPR2023-00317

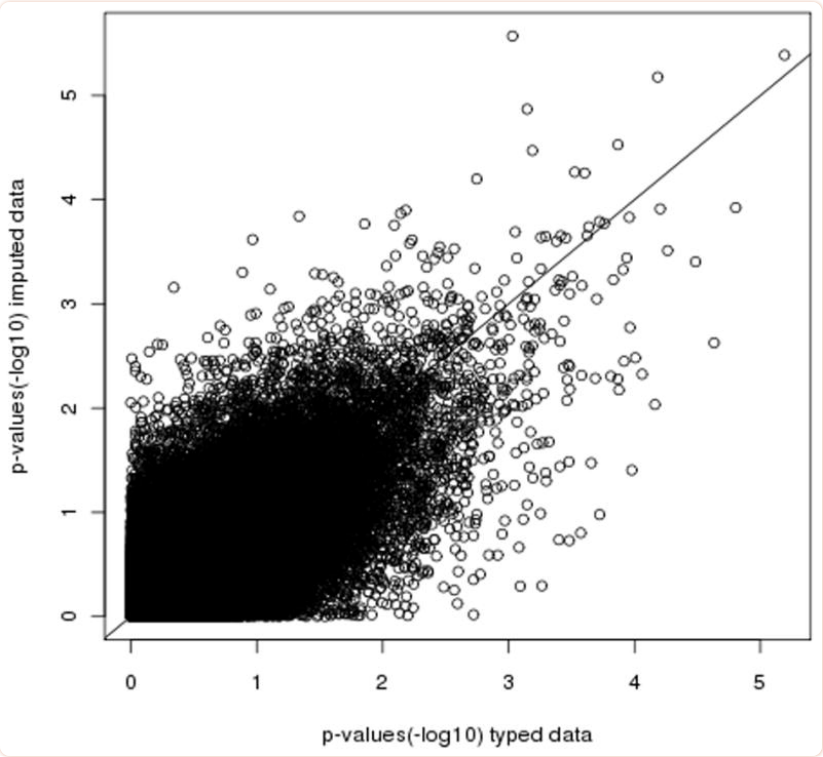

Figure 2

Observed versus expected association minus 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 observe r^2 of 0.64 between p-values computed in typed 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 Supplementary Note.

Table 1

Statistics attained at known associated SNPs in the International HIV Controllers Study computed over typed or imputed genotypes (only SNPs with nominal p-value < 0.05 in the typed data are shown).

RsID	Chr	Position	Coverage	r^2	Association p-value (−log10) typed data	Association p-value (−log10) imputed data	Ratio	Effect typed [confidence interval]	Effe imp [cor inte
rs7756521	6	30848253	0.33	0.96	1.38	1.12	0.81	0.19 [0.01 0.38]	0.1 (
rs3094212	6	31085770	0.73	0.96	1.37	1.42	1.03	0.15 [0.01 0.29]	0.1 (
rs2395471	6	31240692	0.27	0.96	1.38	1.41	1.02	0.14 [0.01 0.28]	0.1 (
rs9366778	6	31269173	0.43	0.84	1.34	1.87	1.4	0.15 [0.00 0.29]	0.1 (
rs9264942	6	31274380	0.26	0.69	1.77	2.35	1.33	0.19 [0.04 0.34]	0.2 (
rs2156875	6	31317347	0.31	0.94	1.56	1.16	0.74	0.17 [0.02 0.31]	0.1 (
rs2844529	6	31353593	0.94	0.92	2.53	3.02	1.19	0.21 [0.08 0.35]	0.2 (
rs2523467	6	31362930	0.63	0.93	2.53	2.39	0.94	0.21 [0.08 0.35]	0.2 (
rs2596531	6	31387557	0.55	0.94	1.31	1.38	1.05	0.15 [0.00 0.30]	0.1 (
rs2516513	6	31447588	0.36	0.86	1.53	1.25	0.82	0.18 [0.02 0.34]	0.1 (
Average			**0.48**	**0.90**	**1.67**	**1.74**	**1.04***	-	

*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.

We also evaluated empirical results at lower coverage (0.005x to 0.5x) by sub-sampling reads with corresponding probability. Due to the large number of experiments and the higher non-exome coverage of the IHCS data as compared to all the 909 samples, we restricted this analysis to the 10 distinct 5Mb regions (total of 50Mb) described above in the IHCS data set (84 samples). As cover-

Foresight EX1213
Foresight v Personalis
IPR2023-00317

age decreases, we observe a reduction in accuracy, analogous to our simulations based on the 1000 Genomes Project dataset, restricted to the same set of 6,070 SNPs from the array (Figure 3). At 0.5x coverage we observe a mean r^2 of 0.82, standard deviation of 0.03 and standard error of 0.01 across the 10 regions. However, the accuracy of imputation in the IHCS sequencing data is lower than in simulations for any level of coverage (Figure 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 4x sequencing data that aggregated information across samples (Supplementary Note, Supplementary Table 6). 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 explanation (Supplementary Note).

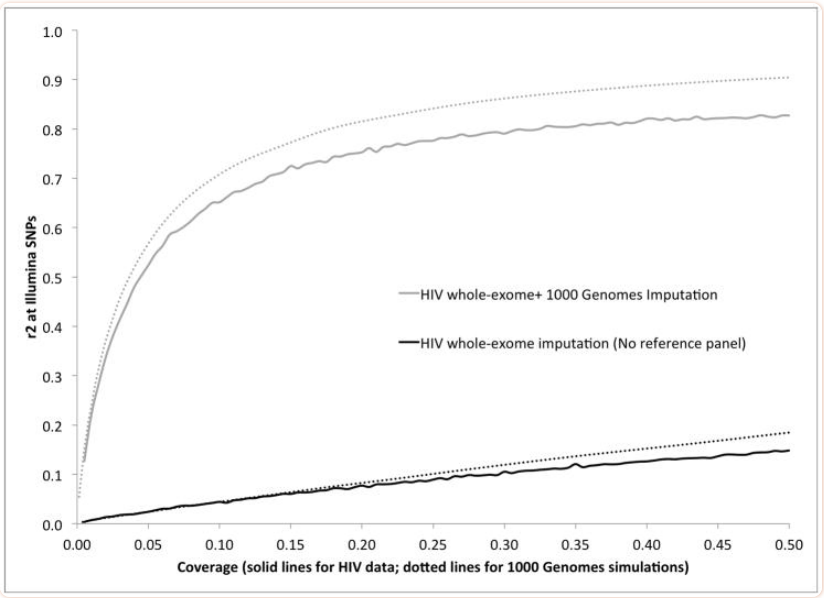

Figure 3

Genotype imputation accuracy in IHCS whole-exome data as a function of coverage. Illumina 1M genotype calls were used as a gold standard, restricting to 6070 SNPs in 10 distinct 5Mb regions (total of 50Mb) of the genome (see main text). Dotted lines denote results attained in 1000 Genomes simulations on the same SNP set.

Discussion

To explore the economic ramifications of sequencing-based GWAS, we considered the trade-off between the number of samples sequenced and the 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 (e.g. Illumina Human-1M-duo). We derived all results from empirical accuracies using sequencing data sets sub-sampled from the IHCS data, so that results do not rely on any simulation

assumptions. We compared accuracies only at SNPs typed on the array, a conservative computation that ignores 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 demonstrated[11]), and DNA sequencing cost of $133 per 1x sequencing (based on the Illumina Network cost of $4,000 for 30x 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 exists 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 typed individuals) is achieved at an average coverage of 0.1x (6,800 samples sequenced at $45 total cost per sample, r^2=0.65) (Figure 4a). The optimal value of average coverage varies as 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 (Figure 4a).

Foresight EX1213
Foresight v Personalis
IPR2023-00317

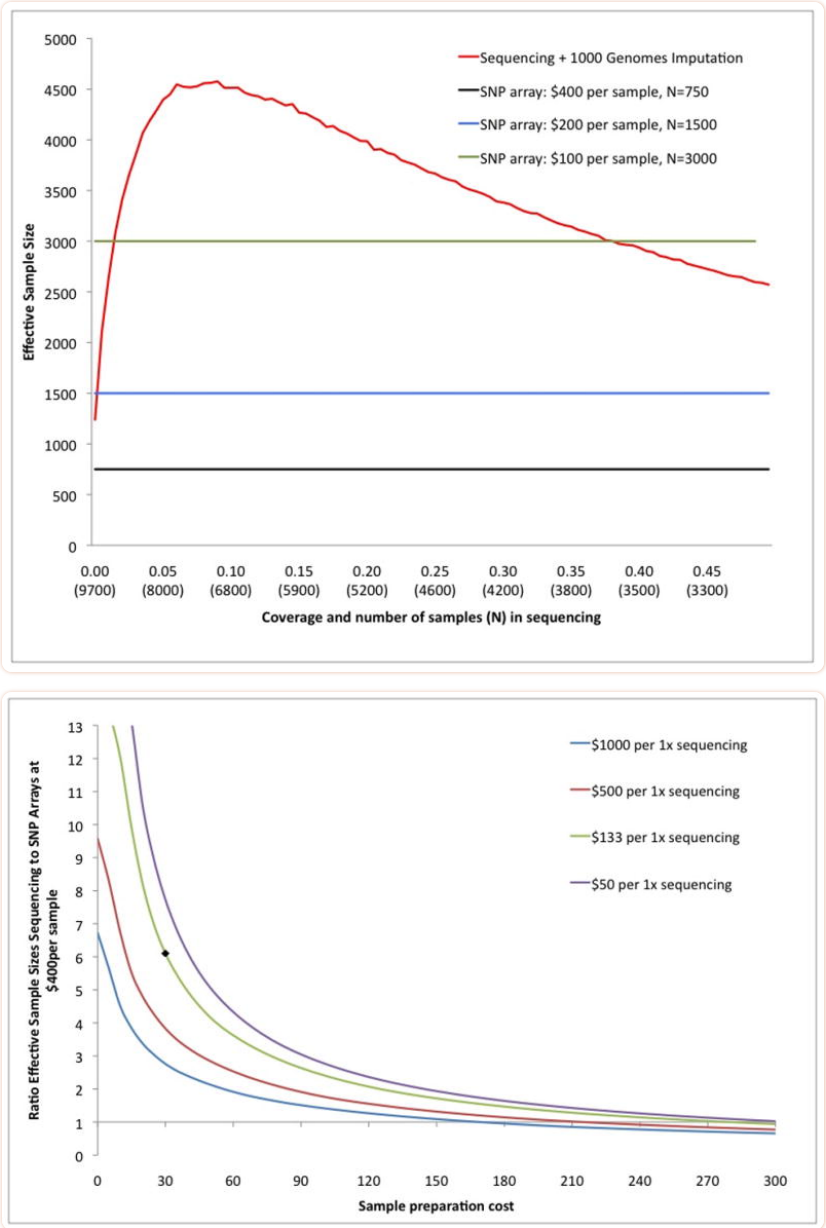

Figure 4

Coverage (and corresponding number of samples) for fixed budget of $300,000. (a) Effective sample size in sequencing-based GWAS as function of number of samples and resulting coverage. Cost assumptions: $30 per sample preparation cost, $133 per 1x sequencing cost (see main text).

(b) Ratio of expected association statistic (effective sample size) in sequencing-based GWAS vs. array-based GWAS at $400/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 varies at different points on the graph (not shown). Black dot denotes $30 sample preparation cost and $133 per 1x.

Foresight EX1213
Foresight v Personalis
IPR2023-00317

A striking 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/sample, corresponding to a large increase in power (Figure 4b, Supplementary Note, Supplementary Figure 8). Only if SNP array typing is less than $70 per sample, or if sample preparation and sequencing costs are much higher (e.g. greater than $120 per sample for sample preparation or $1,000 for 1x sequencing) does sequencing-based GWAS lose its advantage in terms of statistical power to associate variants. If sequencing technology—both in 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 ultra low-coverage is the availability of large panels of reference haplotypes. As additional reference haplotypes over larger numbers of SNPs become available from 1000 Genomes Project and other projects, we expect the accuracy attained by ultra 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 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 requires additional computational resources beyond what is necessary to analyze conventional GWAS as the analysis pipeline of sequencing data is typically more demanding than for genotyping data. Third, sequencing-based GWAS of the type described here does 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 demonstrate that this approach can discover new associations between genetic variants and common diseases.

Online Methods

Simulation of sequencing data based on 1000 Genomes Project dataset

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 include 87 CEPH individuals of North 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 haplotypic phase was inferred from low-coverage sequencing (4x) using an imputation strategy that borrowed information across samples and loci. The 762 haplotypes were split 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 10 distinct 5Mb regions (total of 50Mb) across the genome, 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

Foresight EX1213
Foresight v Personalis
IPR2023-00317

mean 1[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 mean equal to the average coverage times per-locus multiplier times 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 (1) inferring genotypes independently at each SNP in each individual, (2) making use of allele frequencies inferred from all sequenced individuals, (3) making use of linkage disequilibrium (LD) patterns inferred from sequenced individuals, or (4) making use of LD patterns inferred from sequenced individuals as well as reference panels of haplotypes[7,22,24,26]. Here we focus on (3) and (4), 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 dataset 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 dataset and computed likelihoods of 0,1,2 copies of the 1000 Genomes Project dataset "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 (i.e. "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 50Mb simulated region) not present on the array followed by imputation at all polymorphic loci identified in the European samples of the 1000 Genomes Project phase 1 dataset using the remaining reference panel of haplotypes (381 in total). We used the MaCH[27] imputation software with default parameters "--rounds 40 --greedy --mle --mldetails".

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

Simulated phenotypes Starting from the typed genotype calls, we simulated continuous randomly ascertained phenotypes $Y \sim g\beta + \varepsilon$, with $\varepsilon \sim N(0,1)$. $\beta = 0$ 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 International HIV Controllers Study[14], of which 43 were genotyped on the Illumina HumanHap 650Y and 41 on the Human-1M-duo array. Of the 84 samples, 61 are HIV-1 controllers enrolled by the IHCS and 23 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 minor allele frequency (MAF < 1%), >2% missing data, or departure from Hardy-Weinberg equilibrium ($P < 10^{-6}$). The SNP sets were intersected to obtain 398,098

Foresight EX1213
Foresight v Personalis IPR2023-00317

SNPs genotyped in all samples. Imputation was performed using all the 762 available 1000 Genomes phase 1 haplotypes as opposed to 381 for simulations using non-overlapping regions of size 2.5Mb with 250Kb flanking regions on either side.

Combined whole-exome dataset Exome sequencing of the Autism NIMH Controls (AUT, 322 samples), for the Swedish Schizophrenia control data (SCZ, 503 samples) and IHCS data (IHCS, 84 samples) was carried out at the Broad Institute[14–18]. We only used samples ascertained as controls in the AUT and SCZ data (i.e. no presence of disease). Exons were captured using the Agilent 38Mb SureSelect v2 Libraries and sequenced using either an Illumina HiSeq2000 or Illumina GenomeAnalyzerII instrument. All samples met the criterion of >90% of targeted bases having >10x coverage and >80% of targeted bases having >20x coverage. Reads were mapped to hg19 using BWA and processed with Picard and GATK (see URLs). The SCZ samples were genotyped on the Affy 5.0 or 6.0 platforms. 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] software were passed to Beagle in windows of 1Mb with 250Kb to impute all SNPs identified as polymorphic in the haplotypes of the European 1000 Genomes Project phase 1 data. 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 4x.

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 estimate the accuracy $r^2(c)$ attained at coverage c by sub-sampling IHCS data. We then estimate the expected association statistic as $N\rho^2(\mathrm{G}, \Phi)\, r^2(c)$.

Data access The analyses presented here make use of genetic data from Autism NIMH Controls (AUT, 322 samples), the Swedish Schizophrenia control data (SCZ, 503 samples) and the International HIV Controllers Study (IHCS, 84 samples). AUT and SCZ datasets are available from dbGaP under accession numbers phs000298.v1.p1 and phs000473.v1.p1. The IHCS data is available by direct request from Pamela 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 from: http://cfar.globalhealth.harvard.edu/fs/docs/icb.topic938249.files/Harvard%20CFAR%20Concept%20Sheet%20Template%20.docx. Requests will be reviewed on the basis of scientific merit, feasibility and potential overlap with accepted concept sheets or ongoing investigations.

Supplementary Material

Foresight EX1213
Foresight v Personalis
IPR2023-00317

Supplementary information

Click here to view.(966K, docx)

Acknowledgments

This research was supported by NIH grants R01 HG006399 (B.P., N.P., D.R. and A.L.P) and R01 MH084676 (S.S.) The International HIV Controllers Study 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 International HIV Controllers Study was also supported in part by NIH grants P-30-AI060354 (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, RR024975 (AIDS Clinical Trials Group), and AI077505 (D.W.H.). The data generation for the NIMH controls was directly supported by NIH grants R01MH089208, R01 MH089025, R01 MH089004, and R01 MH089482. We thank Thomas Lehner (NIMH), Adam Felsenfeld (NHGRI), and Patrick Bender (NIMH) for their support and contribution to the generation of the AUT sequencing data. SCZ data generation was supported by NIMH grant 5RC2MH089905 (PIs, Sklar and Purcell), by the Sylvan Herman Foundation and the Stanley Medical Research Institute (gift to Stanley Center for Psychiatric Research). We would like to acknowledge the ARRA Autism Sequencing Consortium (AASC) principle investigators (PIs) for the use of the autism datasets: Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Jr., Mark J Daly (communicating PI), Bernie Devlin, Richard Gibbs, Kathryn Roeder, Aniko Sabo, Gerard D Schellenberg, and James S Sutcliffe. We thank Thomas Lehner (NIMH), Adam Felsenfeld (NHGRI), and Patrick Bender (NIMH) for their support and contribution to the AASC project.

Footnotes

URLs.

The 1000 Genomes project, June 2011 phase 1 release:http://www.1000genomes.org/node/506

Beagle software: http://faculty.washington.edu/browning/beagle/beagle.html

MaCH software: http://www.sph.umich.edu/csg/abecasis/MACH/index.html

Picard utilities: http://picard.sourceforge.net/index.shtml

GATK suite: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

Foresight EX1213
Foresight v Personalis
IPR2023-00317

Epicentre sample preparation: http://www.epibio.com/item.asp?ID=566

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/

Accession numbers:

dbGaP accession site for the AUTdata: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1

dbGaP accession site for the SCZdata: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473.v1.p1

Author Contributions:

BP, NR, NP, ALP and DR conceived and designed the study. BP conducted the analyses. LL, SS, NR, PJM, NZ and HL provided bioinformatics and statistical support. PIWdB, NG, KG, BN, MD, ARRA Autism Sequencing Collaboration, PS, PFS, SB, JLM, CMH, PL, PM, SMP and DWH recruited and provided samples and data for these analyses. BP, ALP and DR wrote the paper. All authors contributed to the final version of the manuscript.

References

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 2009;106:9362–9367. doi: 10.1073/pnas.0903103106. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

2. Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467:1061–1073. doi: 10.1038/nature09534. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

3. Depristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–498. doi: 10.1038/ng.806. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

4. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010;11:499–511. doi: 10.1038/nrg2796. [PubMed] [CrossRef] [Google Scholar]

5. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010;467:52–58. doi: 10.1038/nature09298. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

6. Metzker ML. Sequencing technologies -the next generation. *Nat Rev Genet.* 2010;11:31–46. doi: 10.1038/nrg2626. [PubMed] [CrossRef] [Google Scholar]

7. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 2011;12:443–451. doi: 10.1038/nrg2986. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Foresight EX1213
Foresight v Personalis
IPR2023-00317

8. Li Y, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet.* 2010;42:969–972. doi: 10.1038/ng.680. [PubMed] [CrossRef] [Google Scholar]

9. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature.* 2010;464:768–772. doi: 10.1038/nature08872. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

10. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature.* 2010;464:773–777. doi: 10.1038/nature08903. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

11. Rohland N, Reich D. Cost-effective high-throughput DNA sequencing libraries. *Genome Research.* 2012 doi: 10.1101/gr.128124.111. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

12. Browning BL, 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.* 2009;85:847–861. doi: 10.1016/j.ajhg.2009.11.004. S0002-9297(09)00519-9. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

13. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet.* 2001;69:1–14. doi: 10.1086/321275. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

14. Pereyra F, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science.* 2010;330:1551–1557. doi: 10.1126/science.1195271. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

15. Suarez BK, 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.* 2006;78:315–333. doi: 10.1086/500272. S0002-9297(07)62362-3. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

16. O'Donovan MC, et al. Analysis of 10 independent samples provides evidence for association between schizophrenia and a SNP flanking fibroblast growth factor receptor 2. *Mol Psychiatry.* 2009;14:30–36. doi: 10.1038/mp.2008.108. mp2008108. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

17. Manolio TA, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet.* 2007;39:1045–1051. doi: 10.1038/ng2127. ng2127. [PubMed] [CrossRef] [Google Scholar]

18. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009;460:748–752. doi: 10.1038/nature08185. nature08185. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

19. Musunuru K, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med.* 2010;363:2220–2227. doi: 10.1056/NEJMoa1002926. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

20. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 2010;11:685–696. doi: 10.1038/nrg2841. [PubMed] [CrossRef] [Google Scholar]

21. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55:997–1004. [PubMed] [Google Scholar]

22. Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. Efficient study design for next generation sequencing. *Genetic Epidemiology.* 2011;35:269–277. [PMC free article] [PubMed] [Google Scholar]

23. Kim SY, et al. Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet Epidemiol.* 2010;34:479–491. doi: 10.1002/gepi.20501. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

24. Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 2011;21:952–960. doi: 10.1101/gr.113084.110. gr.113084.110. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Foresight EX1213
Foresight v Personalis
IPR2023-00317

25. Prabhu S, Pe'er I. Overlapping pools for high-throughput targeted resequencing. *Genome Res.* 2009;19:1254–1261. doi: 10.1101/gr.088559.108. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

26. Bansal V, et al. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.* 2010;20:537–545. doi: 10.1101/gr.100040.109. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

27. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010;34:816–834. doi: 10.1002/gepi.20533. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

28. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–1303. doi: 10.1101/gr.107524.110. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

29. Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics.* 1955;11:375–386. [Google Scholar]