ARTICLE

Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases

Chuanhua Xing^{*,1}, Jie Huang², Yi-Hsiang Hsu^{3,4,5}, Anita L DeStefano^{1,6,7}, Nancy L Heard-Costa^{6,7}, Philip A Wolf^{6,7}, Sudha Seshadri^{6,7}, Douglas P Kiel^{3,4}, L Adrienne Cupples^{1,7} and Josée Dupuis^{*,1,7}

Although emerging sequencing technologies can characterize all genetic variants, the cost is still high. Illumina released the HumanOmni5M-4v1 (Omni5) genotype array with ~4.3M assayed SNPs, a much denser array compared with other available arrays. The Omni5 balances both cost and array density. In this article, we illustrate the power of Omni5 to detect genetic associations. The Omni5 includes variants with a wide range of minor allele frequencies down to <1%. The theoretical power calculation examples indicate the increased power of the Omni5 array compared with other arrays with lower density when evaluating associations with some known loci, although there are exceptions. We further evaluate the genetic associations between known loci and several quantitative traits in the Framingham Heart Study: femoral neck bone mineral density, lumbar spine bone mineral density and hippocampal volume. Finally, we search genome wide for novel associations using the Omni5 genotypes. We compare our association results from Affymetrix 500K+MIPS 50K arrays and two imputed data sets: (1) HapMap Phase II and (2) 1000 Genomes reference panel. We observed increased evidence for genotype–phenotype associations with smaller *P*-values for selected known loci using the Omni5 genotypes. With limited sample sizes, we identify novel variants with genome-wide significant *P*-values. Our observations support the notion that dense genotyping using the Omni5 can be powerful in detecting novel associated variants. Comparison with imputed data with higher density also suggests that imputation helps but cannot replace genotyping, especially when imputation quality is low.

European Journal of Human Genetics (2016) **24**, 1029–1034; doi:10.1038/ejhg.2015.244; published online 18 November 2015

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of variants associated with complex disease traits, greatly improving our understanding of the role genetics has in disease etiology. However, causal variants were rarely genotyped, and a large portion of heritability has remained unexplained by the identified loci. GWAS usually include a limited number of genome-wide markers consisting primarily of common variants with minor allele frequencies (MAF) >0.05. The role of rare variants with MAF <0.05 has not been comprehensively explored in GWAS, whereas rare variant associations are believed to be a possible reason for the unexplained heritability. Emerging sequencing technologies have the potential to detect virtually all rare variants, but the cost is still high. Illumina (San Diego, CA, USA) recently released the HumanOmni5M-4v1 (Omni5) BeadChip that includes ~4.3M variants with MAF ranging from 0.0002 to 0.5. The cost for Omni5 is lower than next-generation sequencing. Nevertheless it is not clear how much information can be obtained from the additional markers, and how much of the extra information is already captured by imputation. Genotype imputation has been widely used in GWAS analysis to increase power and to fine

map associations.¹ But imputation is limited by the information available in the reference panels and the underlying algorithms.

The goal of this paper was to investigate the performance of Omni5 compared with arrays with sparser markers, with and without imputation. We first computed the theoretical power to detect susceptibility variants from Omni5, and compared the power with that of several arrays with lower density. Next, as an illustration, we assessed association between OMNI5 SNPs and three quantitative traits from the Framingham Heart Study: femoral neck bone mineral density (FNBMD), lumbar spine bone mineral density (LSBMD) and hippocampal volume (HV). We evaluated the ability to detect these selected known variant associations with these traits using the Omni5 compared with other data sets: (1) Affymetrix 500K (250K Sty Array +250K Nsp Array)+MIPS 50K, (2) the imputed Affymetrix 500K +MIPS 50K data set from MACH^{2,3} using HapMap Phase II as reference and (3) the imputed Affymetrix 500K+MIPS 50k data set from MACH using the 1000G Phase I reference panel (v3 EUR, 2012-03-14) as reference.⁴ Finally we performed a genome-wide analysis to identify possible novel variants associated with the three traits using the Omni5 genotypes.

¹Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; ²Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK; ³Hebrew SeniorLife, Institute for Aging Research, Boston, MA, USA; ⁴Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA; ⁵Molecular and Integrative Physiological Sciences Program, Harvard School of Public Health, Boston, MA, USA; ⁶Department of Neurology, Boston University School of Medicine, Boston, MA, USA; ⁷Framingham Heart Study, National Heart, Lung and Blood Institute, Framingham, MA, USA

*Correspondence: Dr C Xing, Department of Biostatistics, Boston University, P.O. BOX 152, Gaithersburg, MD 20884, USA. Tel: +1 877 879 5255; E-mail: chuanhua.xing@gmail.com or Professor J Dupuis, Department of Biostatistics, Boston University School of Public Health, The third floor, 801 Massachusetts Avenue, Boston, MA 02118, USA. Tel: +1 617 638 5880; Fax: +617 638 6484; E-mail: dupuis@bu.edu

Received 22 August 2014; revised 9 October 2015; accepted 14 October 2015; published online 18 November 2015

MATERIALS AND METHODS

Data analysis

A total of 4 271 233 SNPs in 2474 individuals were successfully genotyped using the Illumina Omni5 (phs000342.v13.p9). After removing 4445 non-SNP variants, 4 266 788 SNPs remained. Considering the 22 autosomes and the X chromosome, a total of 4 264 060 SNPs were included in our analyses. A frequency histogram of MAF is presented in Figure 1. A large portion (~25%) were very rare variants, with $MAF < 0.01$, many seen in only one or two FHS participants. Approximately 152 745 SNPs (~4%) were singletons and 90 002 SNPs (~2%) were doubletons. Approximately 20% were rare variants with MAF varying between 0.01 and 0.05, and about 20% were common variants having $MAF > 0.2$. In addition, ~16% of the variants, or 677 551 SNPs, were monomorphic in the subset of participants genotyped. Further analysis indicated that Omni5 genotypes had a low rate of missing calls, with almost all variants having call rates $> 95\%$, and ~96% of polymorphic variants having Hardy–Weinberg Equilibrium P -values > 0.05 . Please see Supplementary Tables 1–6 for more details.

Annotations of Omni5 variants indicated that ~70% of variants fell within 60 kb from a gene using RefGene and this increased slightly when using both RefGene and KnownGene (see Web Resources for the links to find more information) as references. RefGene and KnownGene are two gene-annotation databases from NCBI and UCSC genome browser with KnownGene having a larger number of genes because gene predictions are based on data from more databases. Around 36% of variants were in introns, whereas only 2.7% were in exons. Approximately 0.76% of the variants were synonymous, whereas 1.94% were non-synonymous. These percentages changed slightly when restricting the analysis to high-quality SNPs with call rate $> 95\%$ and Hardy–Weinberg Equilibrium P -value > 0.000001 . When we compared Omni5 data with two imputed data sets using HapMap Phase II and 1000 genomes⁴ as references, 72% of the variants in the Omni5 were not present in the HapMap imputed data, whereas 24% of the variants were not present in the 1000 Genomes imputation. Similar percentages of data were absent in HapMap and 1000 imputed data when restricting the analysis to high-quality SNPs. Please see Supplementary Tables 7–9 for more details.

Theoretical power calculation

We evaluated power to detect variant-trait associations using the Omni5 array. The principle is based on the relationship between linkage disequilibrium (LD) and power for genetic association studies.^{5–9} Suppose that genotypes are available in a case–control sample with N_1 individuals at a (true) disease susceptibility locus and that genotypes are available with N_2 individuals at a nearby marker locus. One can show that the distributions of test statistics at two loci, X_1^2 and X_2^2 are approximately the same if $N_2 = N_1/r^2$, where r^2 is the square of the correlation between the two variants.⁸ Alternatively, the power at the marker locus can be calculated using the relationship between the

non-central parameters (NCP) of the distribution of the test statistics at two loci, $NCP_2 = r^2 NCP_1$ or $X_2^2 = r^2 X_1^2$ assuming $N_1 \approx N_2$ where $0 < r^2 < 1$. The details for the power calculation are given in the Supplementary File.

Power and genomic coverage are relevant measures to evaluate arrays. Genomic coverage measures the proportion of genomic variation captured by an array,¹⁰ and is usually evaluated by comparing r^2 with a threshold such as 0.8. It has been evaluated over various arrays.^{11–13} Recently Illumina reported that Omni5 has improved coverage compared with arrays with lower density (http://support.illumina.com/array/array_kits/humanomni5-quad_beadchip_kit/documentation.ilmn). Nelson *et al.*¹⁴ also compared the coverage and power of Omni5 with both 1000 Genomes and arrays with lower density. Specifically, up to 90% of common variants and $> 50\%$ of less common variants are covered in Omni5 and it has consistently higher power than arrays with lower density of variants but has slightly lower power than sequencing over various sample sizes, MAFs and effect sizes.

Previous studies have used all genome-wide variants, including the ones that have no effect on disease risk. And, not every variant is a disease susceptibility locus.¹⁴ We start by selecting variants that are known to increase disease susceptibility from published data. High-risk variants from real variant-trait associations can provide a more accurate illustration of the power for a given array, and avoid bias that results from including non-disease relevant variants.

Genetic association tests

The GEPOS Consortium recently reported 56 genetic variants influencing BMD variation at genome-wide significance levels, after meta-analysis of 17 GWAS with FNBMD and LSBMD including 32 961 individuals.¹⁵ Bis *et al.*¹⁶ recently explored genetic influences on HV by conducting a cross-sectional genome-wide association analysis, using 9232 participants from eight community-based studies and replicated these findings in an additional 9341 participants. For the analysis of known loci, we selected all the significant SNPs for each of the three traits (FNBMD, LSBMD and HV) from the publications listed above^{15,16} that used HapMap II-based imputed data. Standardized BMD residuals adjusted for age,² weight, sex and population substructure were tested for association with SNP allele or dosage using mixed effects linear regression models that account for family relationships. Similarly, standardized HV residuals adjusted for age and sex were tested for association with SNP allele or dosage using mixed linear regression models. For each top association, we compared the smallest P -value within 25 kb of the given variant on the Omni5 to the smallest P -value within 25 kb using HapMap II and 1000 Genomes imputed data sets. We assessed the imputation quality by the imputation ratio, the empirically observed variance of the allele dosage to the expected variance under a binomial model. We selected imputed variants with imputation ratios > 0.3 . Note that the target SNP was not always available on the Omni5 array. For identification of novel variants genome wide, we conducted genome-wide association analyses and picked the top associations for comparison across arrays.

Figure 1 Proportion of SNPs by minor allele frequency.

RESULTS

Illustration of power calculation and comparison

Here, we illustrate the power calculation and comparison over arrays. We selected non-synonymous exonic SNPs that are known to be significantly associated with human disease traits (we call them 'risk SNPs/variants'). We chose such SNPs to increase the odds of their being risk variants, although most current discoveries are likely to be in LD with causal SNPs and not causal. We selected high-risk variants from previously published GWAS from the National Human Genome Research Institute's (NHGRI) Catalog of Published GWAS (<http://www.genome.gov/gwastudies>). The selected variants are non-synonymous (missense) exonic and have P -value less than the genome-wide significance level 5×10^{-8} . To ensure enough power and reduce the odds of picking false positives, we selected variants from GWAS with number of cases > 500 and total sample size > 5000 , and only chose variants with MAF > 0.05 . We also restricted our selection to variants that were replicated by an independent data set.

Given a sample with 80% power at the selected SNPs, we calculated the power for tagging a SNP present in a specific array, selecting the tag SNP that has the highest LD with the given risk SNP. LD between SNPs is calculated using SNAP (<http://www.broadinstitute.org/mpg/snap/ldsearch.php>) for the different arrays. We chose the 1000 Genomes Pilot 1 as the reference data set, the population panel is CEU (CEPH, Utah residents with ancestry from northern and western Europe), and distance limit was 500 kb. The arrays compared include Affymetrix 5.0 (A5), Affymetrix 6.0 (A6), Illumina HumanHap300 (I3), Illumina HumanHap550 (I5), Illumina Human1M single (IM), Illumina OmniQuad (OQ) and Illumina Omni5 (O5).

From NHGRI Catalog of Published GWAS, we obtained 143 high-risk variants, using the selection standards described above. Of these, 84 variants had proxies across arrays using SNAP Proxy Search to obtain the variants with the highest LD with the high-risk variants. The power for the 84 selected variants was calculated and compared between Omni5 and other arrays, as given in Figure 2. The overall power for Omni5 tends to be higher than other arrays with lower marker density. In general, as the marker density increases, power increases especially for arrays from the same manufacturer.

The power values and the r^2 between the selected variants and the proxy available on the array for the 84 variants are given in the Supplementary Table 10a and b. From the Supplementary Table 10a, power can decrease locally with denser arrays for certain SNPs. One example is that the power for Omni5 to detect the

proxy to SNP rs2233434 is 25%, which is smaller than the 80% power for the OmniQuad, an array with a lower number of markers (N SNPs = 1 051 295). The difference in power is reflected by the decreased r^2 in the Omni5. The power for some SNPs, such as rs10490924, are $\sim 80\%$ over all arrays, because the susceptibility SNP is well tagged in all the arrays. The overall trend is consistent with the observation by Nelson *et al.*¹⁴ when genome-wide variants were investigated.

In addition to the limited pool of high-risk variants to select, as discussed earlier in the Materials and methods section, other factors may influence power trend. When some variants are included in arrays with a lower density or are located closer to SNPs in high LD with the susceptibility variant, they tend to have higher power. The power is also higher in arrays that include the risk variants. The different designs from different manufacturers also influence the power, which is reflected by the fact that the power for Affymetrix 5.0 and Affymetrix 6.0 is not always higher than Illumina HumanHap300 and Illumina HumanHap550 for a given number of variants included. This is particularly true for the comparison between Affymetrix 5.0 and Illumina HumanHap300, two arrays with a similar number of markers. Recombination rate and its variation can also influence r^2 and therefore the power in genetic association studies. Particularly, the selected SNPs are discovered from the earlier arrays with lower density than Omni5. This fact makes the power for some SNPs higher in the arrays that were used for discovery. Despite of all the potential influences, Omni5 generally has higher power than other arrays with lower density.

The price for Omni5 is, however, close to the arrays with lower density, but is much lower than sequencing arrays. We provide the list of the price in the Supplementary Table 11. Because Illumina arrays other than Omni5 are no longer available, we provide the cost for the arrays that are currently available.

Association using known loci

As a demonstration of the advantages of the Omni5 array to detect reported risk variants, we evaluated genetic associations between known loci and three quantitative traits from the Framingham Study. The three selected traits were FNBMD,¹⁵ LSBMD¹⁵ and HV.¹⁶ The sample sizes were 2093 for both FNBMD and LSBMD, and 1520 for HV, after restricting analyses to participants with measured phenotypes and available genotypes from both Omni5 and Affymetrix 500K+MIPS 50K arrays. Whereas all significant associations from prior reports^{15,16} reached the genome-wide level of significance ($P < 5 \times 10^{-8}$) in meta-analyses, we expected the associations to be less statistically significant in this smaller FHS subset.

The $-\log_{10}$ transformed adjusted P -value for Omni5 was compared with two imputed data sets using LSBMD, as shown in Figure 3. The adjusted P -value is calculated by the P -value multiplied by the number of independent SNPs within 25 kb. The average $-\log_{10}$ transformed adjusted P -value for Omni5 is larger than the ones from the two imputed data sets, indicating that the adjusted P -value from Omni5 tended to be smaller. The FNBMD $-\log_{10}$ transformed adjusted P -value boxplot for Omni5 is similar to the ones from HapMap II Imputation data, because most of the adjusted p values are close to 1, as seen in Supplementary Figure 1a. In contrast, the $-\log_{10}$ transformed adjusted P -value boxplot for HV suggests that the adjusted P -values from Omni5 tend to be smaller than the ones from the two imputed data sets, as seen in Supplementary Figure 1b. These observations demonstrate that Omni5 tends to have a higher power in detecting genetic associations. The detailed results for all three traits are summarized in Supplementary Table 12a. Omni5 with genotyped

Boxplot of power across arrays for 84 high risk variants

Figure 2 Power comparison for the 84 selected variants between Omni5 and other arrays.

variants is also better at discovering the variants when the imputation ratios are low in the imputed data sets.

Identification of significant variants genome wide

Finally, we evaluated genome-wide associations for the three traits over four genotype data sets: (1) Affymetrix 500K+MIPS 50K genotypes; (2) Affymetrix 500K+MIPS 50K with imputed data using 1000 Genomes as reference; (3) Affymetrix 500K+MIPS 50K with imputed data using Hapmap phase II as reference; and (4) Omni5. We compared the most significant associations from the Omni5 with $MAF > 0.05$ to the others, and summarized our findings in Table 1 for FNBMD and in Supplementary Table 13a and b for LSBMD and HV. We used chromosome:position to represent a SNP, as most new variants from the Omni5 have not been assigned rs numbers (mapping between chromosome:position and rs number for some SNPs is available in the Supplementary Table 14).

Some top hits from the Omni5 analysis were novel variants that were not covered by any other data sets. This suggests the additional variants on the Omni5 may harbor novel risk variants. Notably, the association between chr2:216657553 and FNBMD in Table 1 reached a genome-wide significant level with P -value of 2.39×10^{-10} , despite the relatively small sample size. Replication is needed to validate the discovery. P -values for other associations were also low. SNPs imputed with high quality had similar evidence for association as SNPs available on the Omni5 array. However, SNPs with lower imputation quality

did not display the same level of association evidence as genotyped Omni5 SNPs. Some examples of lower quality imputed SNPs included chr2:216657553 and chr16:11758731 for FNBMD in Table 1, chr1:116019110 for LSBMD in Supplementary Table 13a, and chr1:37814302 and chr1:37814302 for HV in Supplementary Table 13b. Meta-analysis with larger sample sizes and independent replication are needed to further validate these associations and may also strengthen other signals with modest P -values. None of the most significantly associated BMD variants were present in the Affymetrix 500K+MIPS 50K but chr1:60458452 from the HV GWAS was present (not listed in the Supplementary Table 13b). Although the imputed data using 1000 Genomes has a greater number of variants, with $\sim 17M$ SNPs, Omni5 has the potential to identify novel risk variants; for example, chr2:216657553 in Table 1 for FNBMD due to the greater accuracy of direct genotyping compared to imputation for many variants.

To look more closely at the detection of risk variants among four data sets with different marker density (Affymetrix 500K+MIPS 50K, its imputed data by HapMap II and 1000 genomes, and Omni5), we examined the regional association plots. There were three SNPs in the top 10 associations for LSBMD that were close to position 120 000 (kb) on chromosome 8. Figure 4 shows plots of $-\log_{10} P$ -values for nearby common variants with $MAF > 0.05$. We further filtered out SNPs with imputation ratios < 0.3 . Figure 4 shows that SNPs in this region had similar evidence for association. The similar P -values can be the result of a common causal variant that has high LD with the other SNPs. Some associations from two imputed data sets ('square' symbol for 1000 Genomes imputed data, and 'diamond' symbol for HapMap II imputed data) and a few SNPs from Affymetrix arrays (in triangle) around chromosome position 120 000 (kb) tended to have small P -values. The P -values for SNPs from Omni5 ('circle' symbol) however tended to be smaller. The greater number of significant associations coming from 1000 Genomes imputed data are due to the additional variants that are not present on the Omni5 array. However, the observation in Figure 4 shows that Omni5 revealed strong signals that were not even found using 1000 Genomes imputation. This observation supports our previous observation that imputed data helped to detect novel susceptibility variants, but Omni5 with genotyped variants had the potential to detect novel signals that could not be detected by arrays with lower density, even with imputed genotypes. Additional regional plots for trait FNBMD and HV are available in Supplementary Figures 2a and b.

We also compiled a list of less common variants ($0.01 < MAF < 0.05$) with small P -values for all three traits (Supplementary Table 15). Less

Figure 3 Power comparison for trait LSBMD between Omni5 and the imputed data sets.

Table 1 Ten top hits for Omni5 in comparison to the imputed data sets using FNBMD

chr:pos (hg19)	RefGene	Min/Maj	maf	Omni5		1000 Genomes imputation			HapMap II imputation		
				Beta	P-value	Beta	P-value	Ratio	Beta	P-value	Ratio
chr2:216657553	<i>FN1</i>	G/T	0.22	0.03	2.4E-10	0.03	0.0016	0.27			
chr10:116948146	<i>ATRNL1</i>	T/C	0.29	0.02	6.7E-7	0.02	2.7E-6	0.98	0.02	1.4E-6	0.90
chr16:11758731		G/A	0.36	0.02	8.1E-7	0.02	6.3E-5	0.41			
chr10:116885750	<i>ATRNL1</i>	T/G	0.33	0.02	8.2E-7	0.02	3.5E-7	0.97			
chr10:116974167	<i>ATRNL1</i>	C/T	0.30	0.02	9.0E-7	0.02	2.8E-6	0.97	0.02	1.7E-6	0.90
chr10:116966058	<i>ATRNL1</i>	C/T	0.33	0.02	9.1E-7	0.02	2.7E-6	0.96	0.02	1.7E-6	0.98
chr10:116984320	<i>ATRNL1</i>	G/A	0.30	0.02	9.5E-7	0.02	2.7E-6	0.97	0.02	1.8E-6	0.90
chr10:116948174		G/A	0.33	0.02	9.7E-7	0.02	2.5E-6	0.95			
chr10:116939803		C/T	0.34	0.02	1.0E-6	0.02	2.1E-6	0.97			
chr10:116904963	<i>ATRNL1</i>	A/G	0.29	0.02	1.0E-6	0.02	1.5E-6	0.99			

Bold P -values are for SNPs either from Omni5 or having relatively high imputation ratios > 0.9 .

Figure 4 Regional associations plot around position 120 000 kb on chromosome 8 over four genotype data sets using the trait of LSBMD. Circles are from Omni5, squares are from imputed data by 1000 Genomes, diamonds are from imputed data by HapMap II and triangles are from Affymetrix 500 k+MIPS 50 k.

common and rare variants may harbor some risk variants, which may not be observed from earlier arrays with predominantly common variants. We focused on common variants because we could use them to examine the incremental power of Omni5 in detecting new common variants within limited sample sizes. Further investigation is needed to better understand the roles of less common variants.

To gain further insight, we compared Omni5 variants, especially less common ($0.01 < \text{MAFs} < 0.05$) and rare ($\text{MAFs} < 0.01$) variants, with the SNPs imputed from 1000 Genomes project data. We first obtained the overlapping SNPs between Omni5 and 1000 Genomes. We removed the genotyped SNPs from Affymetrix 500K+MIPS 50K so that we could evaluate the quality of imputation for 1000 Genomes for the novel variants on Omni5 that were not present in the genotyped 550k SNPs. We generated the distribution of imputation ratios for SNPs in different ranges of MAF: < 0.01 , $0.01-0.05$, $0.05-0.1$, $0.1-0.2$, $0.2-0.3$, $0.3-0.4$ and $0.4-0.5$. The plots are provided in Supplementary Figure 3. The imputation ratio for common variants is high and mostly lie close to a value of 1. In contrast, the imputation ratio for less common variants with $0.01 < \text{MAF} < 0.05$ and rare variants with $\text{MAF} < 0.01$ imputed to the 1000G were much lower and close to 0. Hence, Omni5 contributes not only novel common variants but also less common and rare variants that are not available on arrays with lower density.

DISCUSSION

In this study, we investigated the potential power of Omni5 to detect novel disease risk variants. The search to identify variants underlying the currently unexplained heritability has prompted new technologies for discovering additional susceptibility genes for human diseases. Omni5 offers the advantage of having denser markers than earlier arrays, whereas having modest cost compared with sequencing arrays. We observed that Omni5 includes a large spectrum of SNPs with MAF varying from 0.0002 to 0.5. After quality control, almost all OMNI5 variants have call rates $> 95\%$, and $\sim 96\%$ of polymorphic variants have Hardy–Weinberg Equilibrium P -values > 0.05 . A large percentage of variants from Omni5 were novel and not present in the

Affymetrix 500 K+MIPS 50 K or reliably imputed using 1000 Genomes and HapMap Phase II as references. We further analyzed the theoretical power in tagging disease risk variants and then evaluated association studies using three human traits in FHS. Both theoretical analysis and our evaluation using FHS traits demonstrated that Omni5 can be powerful in detecting novel variants. This observation is consistent with the reported observations that arrays with denser markers have higher power in detecting novel variants. If arrays with higher density such as Affymetrix 6.0 or Illumina OmniQuad arrays are used as the basis, the resulted imputation data would have a higher imputation quality on average, but the conclusions of the study would stay the same. That is, imputation helps but cannot replace genotyping especially when imputation quality is low.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to acknowledge the work of Larry Atwood for his early efforts in connecting Illumina and Boston University. Genotyping of $\sim 4.3\text{M}$ SNPs was conducted in subset of 2500 Framingham Offspring Cohort participants using the Illumina HumanOmni5M-4v1 array designed to target variation down to 1% minor allele frequency. This genotyping was produced at no charge by Illumina under an agreement between Illumina and Boston University. This work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract no. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract no. N02-HL-6-4278). A portion of this research utilized the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center. This study was also supported by grants from the NINDS (NS17950), the NHLBI (HL102419, HL096917) and the NIA (AG08122, AG16495, AG033193). Additional support for the skeletal phenotypes was obtained from NIAMS (R01 AR/AG 41398).

WEB RESOURCES

RefGene and KnownGene can be found from <http://genome.ucsc.edu/> using GRCh37or hg19. SNAP Proxy Search can be found from <http://www.broadinstitute.org/mpg/snap/ldsearch.php>

DATA ACCESS NUMBER

The dbGaP accession number for the Illumina HumanOmni5M-4v1 data reported in this paper is phs000342.v13.p9.

- Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
- Li Y, Willer CJ, Ding J, *et al*: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
- Li Y, Willer CJ, Sanna S, Abecasis G: Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
- Abecasis GR, Auton A, Brooks LD *et al*: 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
- Kruglyak L: Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999; **22**: 139–144.
- Dunning AM, Durocher F, Healey CS *et al*: The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* 2000; **67**: 1544–1554.
- Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
- Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.

- 9 Sham PC, Purcell SM; Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014; **15**: 335–346.
- 10 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
- 11 Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ: Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet.* 2006; **38**: 663–667.
- 12 Li M, Li C, Guan W: Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet* 2008; **16**: 635–643.
- 13 Bhangale TR, Rieder MJ, Nickerson DA: Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* 2008; **40**: 841–843.
- 14 Nelson SC, Doherty KF, Pugh EW *et al*: Imputation-based genomic coverage assessments of current human genotyping arrays. *G3 (Bethesda)* 2013; **3**: 1795–1807.
- 15 Estrada K, Styrkarsdottir U, Evangelou E *et al*: Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* 2012; **44**: 491–501.
- 16 Bis JC, DeCarli C, Smith AV *et al*: Common variants at 12q14 and 12q24 are associated with hippocampal volume. *Nat Genet.* 2012; **44**: 545–551.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)