RESEARCH ARTICLE

Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium

WENQIAN ZHANG[1], HUI WEN NG[1], MAO SHU[1], HENG LUO[1], ZHENQIANG SU[2], WEIGONG GE[1], ROGER PERKINS[1], WEIDA TONG[1] and HUIXIAO HONG[1]*

[1]*National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA*

[2]*Thomson Reuters, IP and Science, 22 Thomson Place, Boston, MA 02210, USA*

Abstract

Single-nucleotide polymorphisms (SNPs) determined based on SNP arrays from the international HapMap consortium (HapMap) and the genetic variants detected in the 1000 genomes project (1KGP) can serve as two references for genomewide association studies (GWAS). We conducted comparative analyses to provide a means for assessing concerns regarding SNP array-based GWAS findings as well as for realistically bounding expectations for next generation sequencing (NGS)-based GWAS. We calculated and compared base composition, transitions to transversions ratio, minor allele frequency and heterozygous rate for SNPs from HapMap and 1KGP for the 622 common individuals. We analysed the genotype discordance between HapMap and 1KGP to assess consistency in the SNPs from the two references. In 1KGP, 90.58% of 36,817,799 SNPs detected were not measured in HapMap. More SNPs with minor allele frequencies less than 0.01 were found in 1KGP than HapMap. The two references have low discordance (generally smaller than 0.02) in genotypes of common SNPs, with most discordance from heterozygous SNPs. Our study demonstrated that SNP array-based GWAS findings were reliable and useful, although only a small portion of genetic variances were explained. NGS can detect not only common but also rare variants, supporting the expectation that NGS-based GWAS will be able to incorporate a much larger portion of genetic variance than SNP arrays-based GWAS.

[Zhang W., Ng H. W., Shu M., Luo H., Su Z., Ge W., Perkins R., Tong W. and Hong H. 2015 Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. *J. Genet.* **94**, 731–740]

Introduction

Personalized medicine is an alternative healthcare approach relying on an individual's genetic profile. Decisions pertaining to prevention, diagnosis, or treatment are those best suited to a cohort of patients with similar genetic profiles (Langreth and Waldholz 1999; Ginsburg and McCarthy 2001; Obama 2007). According to the statement of personalized medicine coalition (PMC) (http://www.personalizedmedicinecoalition.org), personalized medicine promises three key benefits: better diagnoses and earlier interventions, more efficient drug development and more effective therapies. Whether in choosing an individually efficacious drug, or in avoiding an adverse reaction in a susceptible individual, formidable obstacles exist to realize the personalized medicine. The essential prerequisite is identifying the molecular markers that can, for e.g., differentiate those who benefit from a treatment from those who will not, diagnose a disease at an earlier stage, or categorize the stage of a disease.

The first version of human genome was completed in 2001 (Lander *et al.* 2001; Venter *et al.* 2001), and paved the way for identifying genes and their roles in normal human development and physiology, as well as enabling identification of genetic variations among humans. The international HapMap project (hereafter simplified as HapMap) determined genotypes of more than 3.1 million common SNPs in human populations (The International HapMap Consortium 2003; International HapMap *et al.* 2007). Soon, high-throughput SNP genotyping enabled simultaneous genotyping of millions of SNPs. These advances together made genomewide association studies (GWAS) a feasible and a promising research field in human genetics and personalized medicine. Based on the hypothesis of 'common disease – common

*For correspondence. E-mail: huixiao.hong@fda.hhs.gov.

The findings and conclusions in this paper have not been formally disseminated by the US Food and Drug Administration (FDA) and should not be construed to represent the FDA determination or policy.

Keywords. heterozygous rate; minor allele frequency; transition; transversion; genotype discordance; genomewide association studies.

variant' (Lander 1996; Pritchard and Cox 2002; Wang *et al.* 2005), the first GWAS study was published in 2005 (Klein *et al.* 2005). Since then, some 1769 GWAS studies have been published and some 12,042 SNPs have been identified as associated with numerous phenotypes, according to statistics from the National Human Genome Research Institute's GWAS catalogue (Hindorff *et al.* 2009). Despite GWAS being widely applied to identify common genetic variants associated with risk across more than 200 diseases and human phenotypic traits, the SNP array-based GWAS findings were found to explain only a small fraction of the total variances, even with very large sample sizes (Manolio *et al.* 2009; Eichler *et al.* 2010; Hong *et al.* 2012b; Evangelou and Ioannidis 2013; Sharma *et al.* 2014). Generally, concerns arose regarding the usefulness of findings from SNP array-based GWAS (Hirschhorn 2009; Kraft and Hunter 2009).

The enormity of the problem is evident in light of 3.5 million SNPs in a human genome, of which some 10,000 SNPs are nonsynonymous SNPs (Marian 2012). Moreover, both synonymous and nonsynonymous variants have similar effects on odds ratios (Chen *et al.* 2010).

Recently, rare genetic variants having minor allele frequency (MAF), i.e. less than 1%, were believed to account for part of the missing genetic variance in an SNP array-based GWAS, and have been explored in genetic studies (Cirulli and Goldstein 2010; Wagner 2013). For example, rare variants of A1708E, G1738R and R1699Q in gene *BRCA1* lead to amino acid changes and were found associated with breast cancer susceptibility (Lovelock *et al.* 2007). Next generation sequencing (NGS) enables identification of both common and rare genetic variants (Hong 2012; Hong *et al.* 2013; Su *et al.* 2014; Zhang *et al.* 2014). Therefore, it is anticipated that NGS will make significant contributions in understanding the genetics of complex diseases and phenotypic traits, and explaining the 'missing variance' (Marian 2012; Londin *et al.* 2013; Wagner 2013). The 1000 genomes project (hereafter termed 1KGP) applied NGS for identification of SNP spectra of multiple human population groups and the identified SNPs provided a reference resource for genetic studies (Abecasis *et al.* 2012).

The reliability and robustness of GWAS findings rely on the accurate determination of SNPs genotypes as well as other factors such as case–control misclassification (Pearson and Manolio 2008) and nongenetic covariates (Frayling *et al.* 2007). However, bias and errors in SNP determination can be introduced by both technology and data analyses with both microarray (Hong *et al.* 2010a, b, 2012a) and NGS (O'Rawe *et al.* 2013; Ratan *et al.* 2013; Zhang *et al.* 2015). Therefore, comparing SNPs detected in 1KGP (the reference for NGS-based GWAS) with SNPs determined from HapMap (the reference for SNP array-based GWAS) can help to assess the usefulness of the previously reported GWAS findings based on SNP genotyping microarrays. The comparison should also enable measurement of additional benefits of NGS-based GWAS in elucidating that how rare genetic variants add to the explanation of heritability of complex diseases and phenotypic traits (Gibson 2011; Sharma *et al.* 2014). Although some previous studies (Buchanan *et al.* 2012; Rosenfeld *et al.* 2012) performed comparisons on the two references using pilot results from 1KGP, with the availability of more SNPs from the completion of phase 1 of 1KGP, here we conducted a more systematic comparison of SNPs from the two vital references. Comparisons were made on 622 subjects included in both HapMap and 1KGP. We found that 1KGP had many more rare SNPs than HapMap, and that SNPs detected solely in 1KGP had lower heterozygous rate than SNPs contained in both references, confirming NGS-based GWAS's ability to identify additional rare SNPs also associated with diseases and phenotypic traits. We also found high SNP genotype concordance across references, indicating that legacy SNP array-based GWAS findings were reliable and useful, although rare SNPs were not identified.

Materials and methods

Study design and workflow

The study design and workflow are depicted in figure 1. Briefly, HapMap and 1KGP data were downloaded from the public databases. SNPs were removed when their dbSNP IDs could not be found. Insertions and deletions (indels) were not included in the comparative analyses. HapMap SNPs were converted to the forward sequence forms using the human reference genome. After SNP preprocessing, individuals common in both datasets were identified by name, and used in the subsequent comparative analyses. SNPs of common individuals were then grouped into three categories: SNPs common to both 1KGP and HapMap, SNPs only found in HapMap (hereafter termed as HapMap only SNPs) and SNPs only found in 1KGP (hereafter termed as 1KGP only SNPs). Comparative analyses were performed for the three SNP categories for 622 individuals common to both datasets.

Data

The SNPs genotypes of 1417 individuals from HapMap were downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phaseII+III/ (12 February 2013). The genotypes of SNPs detected for 1092 individuals in 1KGP were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ (24 February 2013). The human genome assembly hg18/ncbi b36.3 was obtained from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/Assembled_chromosomes/ (22 March 2013) and the human genome assembly hg19/GRCh37 was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz (3 April 2013).

Data preprocessing

We wrote perl scripts to preprocess the genotype files downloaded from HapMap and 1KGP. For the HapMap data, the SNPs that had dbSNP IDs were first selected. Next, for each

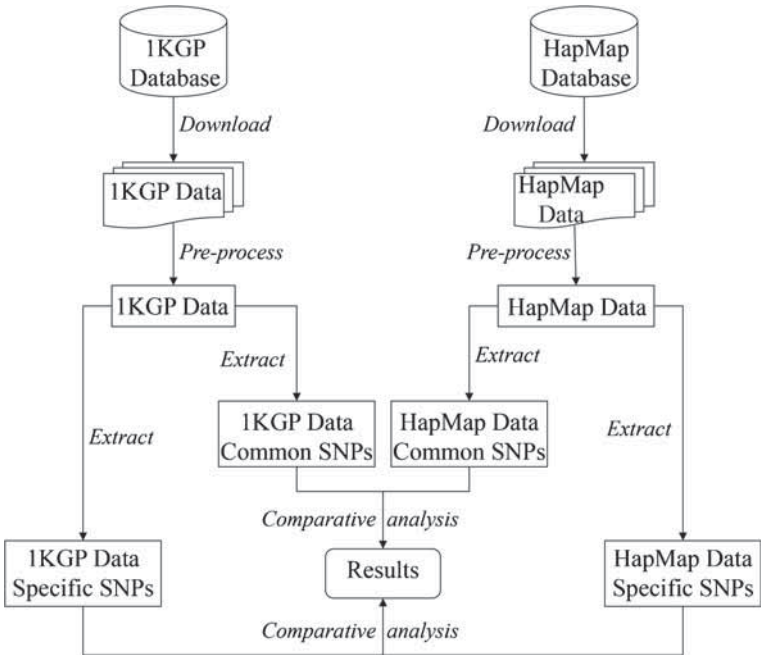

Figure 1. Study design and workflow of comparative analyses. HapMap and 1KGP data were first downloaded and then preprocessed, including merging SNPs from different individuals to generate two comprehensive SNP tables, one for HapMap and the other for 1KGP. After comparing samples covered by HapMap and 1KGP, we only kept the SNPs of 622 common samples for the following analyses. They were then categorized into common SNPs if they were common between two datasets or unique SNPs if they were found only in one dataset.

SNP, the major allele selected was the allele with the higher frequency among the 1417 individuals, and the other allele set as the minor allele. The human genome assembly was then used to convert all the SNP alleles to the forward strand alleles. Thereafter, the SNPs from different populations were merged together. Missing genotypes were labelled as 'NN'. For 1KGP data, the SNPs having string 'VT = SNP' in the INFO column of genotypes files (in vcf format) were first selected. The same criteria and process to determine the major and minor alleles for the HapMap data was repeated for the 1KGP data also.

Data extraction

Individuals common to both HapMap and 1KGP were selected by sample names and used for subsequent comparative analyses. SNPs not detected in common individuals were not used. The selected SNPs were classed into three categories: SNPs common to both HapMap and 1KGP samples (determined using the rs#), SNPs in HapMap data only and SNPs in 1KGP data only. Four datasets were generated as depicted in figure 1, common SNPs from HapMap, HapMap only SNPs, common SNPs from 1KGP and 1KGP only SNPs.

Comparative analyses

Frequencies of A, T, G and C bases were calculated for alleles of the three SNP categories: HapMap only, 1KGP only and common SNPs. The base frequencies of human genome (hg19) were also calculated for comparison.

MAF, heterozygous rate (HR) and ratio of transition mutations (Ts) to transversion mutations (Tv) were calculated for each SNP and for each individual using the four datasets given in figure 1. Perl programs were written to compare MAF, HR, Ts and Tv between HapMap and 1KGP.

Genotype discordance between HapMap and 1KGP for an individual was defined as the percentage of SNPs that had discordant genotypes. There are three possible genotypes: homozygote, homozygote-variant and heterozygote according to the predefined major and minor alleles. Therefore, the genotypes discrepancy for each SNP between HapMap and 1KGP can be one of six types: homozygote in HapMap but homozygote-variant in 1KGP, homozygote in HapMap but heterozygote in 1KGP, homozygote-variant in HapMap but homozygote in 1KGP, homozygote-variant in HapMap but heterozygote in 1KGP, heterozygote in HapMap but homozygote in 1KGP, and heterozygote in HapMap but homozygote-variant in 1KGP. Perl programs were written to classify all discordance SNPs by type.

Hardy–Weinberg equilibrium test

Hardy–Weinberg equilibrium (HWE) model describes an idealistic condition where allelic and genotype frequencies are static and not subject to change by evolutionary mechanisms. Therefore, in HWE, heterozygous frequency solely depends on the homozygous dominant and homozygous

recession fractions that remain static from generation to generation. In a real population, evolutionary factors such as natural selection, gene flow, genetic drift and others may be working to a greater or lesser degree, though departure from HWE should not be drastic. Indeed, large deviation from HWE in a population could be an indicator of either poor or biased population sampling, or poor quality of the genetic data itself. Therefore, we tested HWE for both 1KGP and HapMap as a quality control measure. We downloaded the script from http://csg.sph.umich.edu/abecasis/Exact/snp_hwe.pl for HWE test. The script implemented the exact test algorithm (Wigginton *et al.* 2005).

Statistical analysis

Matlab was used for the statistical analyses and for plotting all the figures.

Results

Datasets generated for comparative analyses

HapMap contained SNPs of 1417 individuals from nine population groups, while 1KGP covered SNPs of 1092 individuals from 14 population groups, as listed in table 1. Of all the individuals, 622 were common in HapMap and 1KGP (the third column in table 1). After filtering SNPs that are not detected in any of the 622 individuals, 4,041,602 SNPs in HapMap and 36,817,799 SNPs in 1KGP were available. According to the positions of the SNPs, we determined 3,466,573 (85.8% for HapMap and 9.4% for 1KGP) were common between the two datasets, 575,029 (14.2%) were only in HapMap and 33,351,226 (90.6%) were only in 1KGP (figure 1).

Base composition analysis

To examine if there was any base bias in HapMap SNPs or 1KGP SNPs, we calculated the base frequencies for 3,466,573 common SNPs, 575,029 HapMap only SNPs and 33,351,226 1KGP only SNPs. The results showed high similarities of base compositions in HapMap only and 1KGP only SNPs (figure 2), no obvious base bias in either dataset. Further, the frequencies of the four bases A, C, G and T were almost equal for the common SNPs, HapMap only SNPs and 1KGP only SNPs. In contrast, the ratio of G and C base frequencies to A and T base frequencies were markedly higher for both HapMap and 1KGP than for the reference genome, indicative of persistent mutations from bases A and T to bases C and G, being more prevalent than in the opposite direction.

HR and MAF of individuals

Heterozygotes or minor alleles are usually responsible for risk of diseases or undesirable phenotypic traits. Therefore, after determining minor alleles, HR and MAF for SNPs of the 622 individuals were calculated for both HapMap and 1KGP. Figure 3a is a plot of HR (y-axis) versus MAF (x-axis) for HapMap only SNPs, while figure 3b is the same HR versus MAF plot for 1KGP only SNPs. As expected, HR and MAF were well correlated with both HapMap and 1KGP SNPs, indicating that the SNPs detected in both HapMap and 1KGP are not far from their HWE and, thus, no conspicuously large

Table 1. Number of individuals from different populations included in 1KGP and HapMap.

Population	HapMap	1KGP	Common
ASW (people with African ancestry in Southwest United States)	87	61	51
CEU (Utah residents with ancestry from northern and western Europe)	174	85	81
CHB (Han Chinese in Beijing, China)	139	97	91
JPT (Japanese in Tokyo, Japan)	116	89	85
LWK (Luhya in Webuye, Kenya)	110	97	87
TSI (Toscani in Italia)	102	98	87
YRI (Yoruba in Ibadan, Nigeria)	209	88	87
CHS (Han Chinese South, China)	0	100	
CLM (Colombians in Medellin, Colombia)	0	60	
FIN (Finnish in Finland)	0	93	
GBR (British from England and Scotland, UK)	0	89	
IBS (Iberian populations in Spain)	0	14	
MEX/MXL (people with Mexican ancestry in Los Angeles, California)*	86	66	53
PUR (Puerto Ricans in Puerto Rico)	0	55	
CHD (Chinese in Metropolitan Denver, Colorado)	109	0	
GIH (Gujarati Indians in Houston, Texas)	101	0	
MKK (Maasai in Kinyawa, Kenya)	184	0	
Total	1417	1092	622

*To represent the people with Mexican ancestry in Los Angeles, two different labels of MEX and MXL were used in HapMap and 1KGP, respectively.

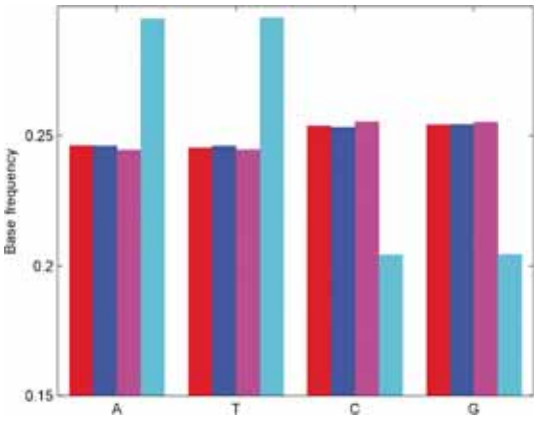

Figure 2. Base composition analyses on the SNPs of 622 common individuals. Frequencies of the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T) were calculated for the SNPs in three categories: 1KGP only SNPs (red), HapMap only SNPs (blue) and SNPs common between HapMap and 1KGP (magenta). The base frequencies of human genome (hg19 from UCSC) are given in the cyan bars.

sequencing or genotyping errors are present. Interestingly, the HR and MAF values of 1KGP only SNPs are considerably smaller than those of HapMap only SNPs. The smaller NGS-based HR and MAF is consistent with an expectation that NGS-based GWAS capability will be able to detect better variance contributions of rare SNPs in genetic architectures of diseases and phenotypic traits.

Ts and Tv mutations

According to the changes in chemical structures, SNPs can be classed into two types: Ts and Tv mutations. Tv mutations result in dramatic changes in chemical structure between a purine and a pyrimidine caused by ionizing radiation and alkylating agents, while Ts mutations are the changes between two purines or between two pyrimidines caused by oxidative deamination and tautomerization. There are more Ts SNPs than Tv SNPs, though there are twice as many possible Tv as Ts in genomes (Collins and Jukes 1994). To analyse Ts and Tv distributions in HapMap and 1KGP datasets, we calculated Ts/Tv ratios for the 622 individuals separately using HapMap only SNPs and 1KGP only SNPs. Results are plotted as a bar chart in figure 4a. The Ts/Tv ratios for HapMap only SNPs (the blue bars in figure 4a) cluster in several bins over a large range, some near 2.2 and some near 3.0. In contrast, the Ts/Tv ratios for 1KGP only SNPs cluster closely around 2.1 (the red bar in figure 4a). The results are consistent with the reports that Ts/Tv is around 2 in the whole human genome, but around 3.0 for exonic regions only (Marth *et al.* 2011). Additionally, SNPs from HapMap were measured using different generations of genotyping arrays, with newer generations interrogating more SNPs in exonic regions of the human genome than nonexonic regions, apparently explaining the wide dispersion seen in Ts/Tv for HapMap.

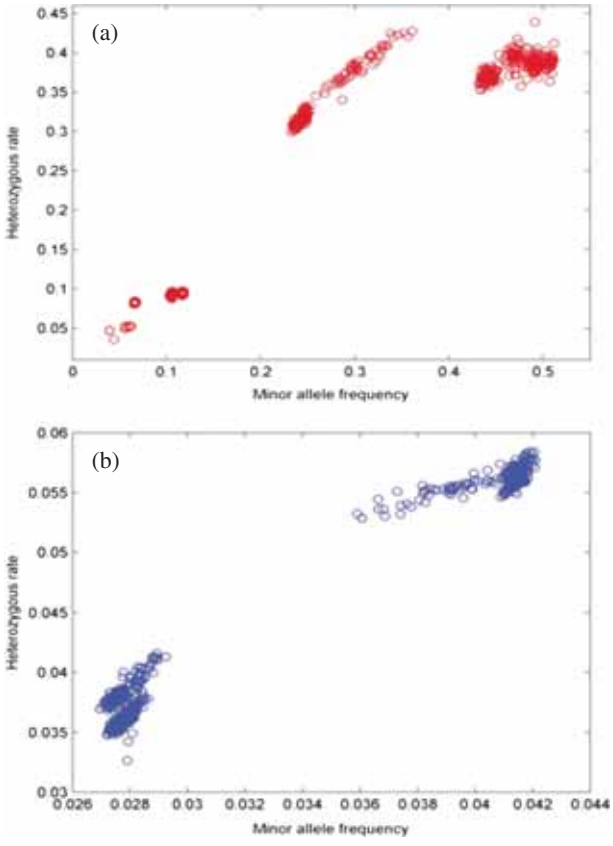

Figure 3. Correlation between minor allele frequency and heterozygous rate for HapMap only SNPs and 1KGP only SNPs in 622 common individuals. The correlations in (a) HapMap only SNPs and (b) 1KGP only SNPs are 0.897 and 0.990, respectively. Each point represents the correlation in each sample.

Differences in Ts/Tv ratio, HR and MAF among the population groups were also calculated for both HapMap only SNPs and 1KGP only SNPs. Figure 4b shows a 3D plot of the results for 1KGP. Individuals from the same population groups are seen to cluster together in accordance with Ts/Tv ratio, HR and MAF, with some groups significantly separated. Further, genetically similar population groups formed larger groups of clusters: people with African ancestry in Southwest United States (ASW), Luhya in Webuye, Kenya (LWK) and Yoruba in Ibadan, Nigeria (YRI) formed a larger group; Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT) clustered; Utah residents with ancestry from northern and western Europe (CEU), Toscani in Italia (TSI) and people with Mexican ancestry in Los Angeles, California (MEX) formed another large group. For HapMap data, the individuals in the same population group were separated in Ts/Tv ratios as an artifact for the use of different array generations. Despite the artifact, figure 4c shows that individuals with high Ts/Tv ratios (>2.8), while figure 4d shows that individuals with low Ts/Tv ratios (<2.3). Further, the same population clustering and genetically similar population clustering as seen for 1KGP were observed in the individuals shown in both figure 4, c and d.

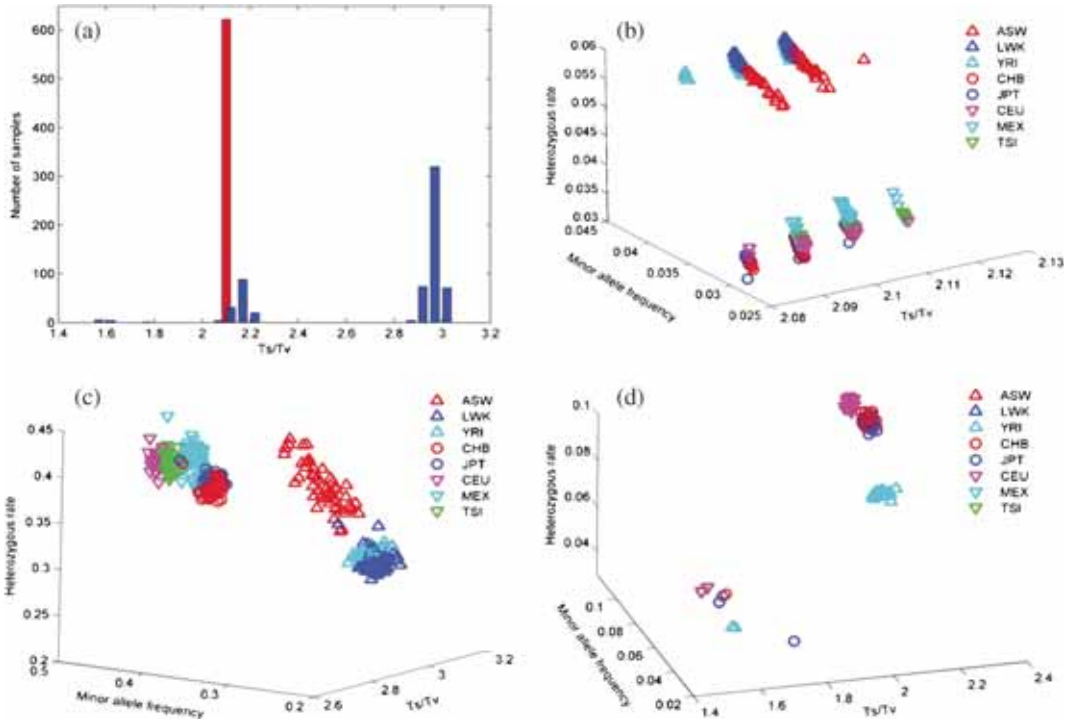

Figure 4. Distributions of transition/transversion ratio (Ts/Tv) for HapMap only SNPs and 1KGP only SNPs in 622 common individuals. (a) Ts/Tv ratios were plotted for HapMap only SNPs (blue bars) and 1KGP only SNPs (red bar). Results show similar Ts/Tv ratios (in the range of 2.08–2.12) for 1KGP only SNPs, while several bins of ratios (some near 2.2 and some near 3.0) for HapMap only SNPs. 3D plots of Ts/Tv ratio versus minor allele frequency versus heterozygous rate were generated for (b) 1KGP only SNPs and (c and d) HapMap only SNPs.

HR and MAF of SNPs

HR and MAF values of common SNPs, HapMap only SNPs and 1KGP only SNPs were calculated and compared in 622 individuals. Figure 5a shows the HR distributions and figure 5b shows the MAF distributions of the three SNPs categories. A much higher proportion of 1KGP only SNPs had very low HR (the red line in figure 5a) and MAF (the red line in figure 5b) than the HapMap only SNPs (the blue lines in figure 5) and the common SNPs (the magenta lines for HapMap and the cyan lines for 1KGP in figure 5), consistent with rare SNPs detection by NGS.

Discordance analysis

Discordance in genotypes for the 3,466,573 common SNPs was calculated for each of the 622 individuals. The discordance distribution is shown in figure 6a bar chart where the average and maximum discordance is 0.0055 and 0.018, respectively, indicating good agreement between genotyping microarray and NGS. The discordance in genotypes between HapMap and 1KGP was further characterized according to discordance type. Figure 6b shows SNP rate versus discordance between heterozygous, homozygous and variant homozygous SNPs. Differences between homozygous genotypes determined in HapMap and heterozygous genotypes detected in 1KGP are the major contributions to the overall discordance.

Discussion

Personalized medicine holds the promise for achieving the best tailored healthcare for individuals, but its realization ultimately pivots on valid identification of the genetic variants associated with diseases and drug responses. Identifying variants, in turn, depends on the validity of reference SNPs and genotyping technology and related data interpretation methods to more completely explain causatively-related genetic variances. GWAS from the first wave of genotyping microarrays fell short of explaining enough variance. Now, entry of whole genome sequencing as a powerful technology for interrogating all possible genetic variants, including both common and rare SNPs. NGS-based GWAS fosters high expectations for realizing personalized medicine through identifying the more complete constellations of genetic variants responsible for disease susceptibility, drug efficacy, or adverse drug responses, and so forth.

Two important related questions await answers: (i) are array-based GWAS findings still useful? and (ii) are NGS-based GWAS approaches able to identify missing rare SNPs, in addition to common SNPs? We conducted comparative analyses on the reference SNPs from HapMap and 1KGP to help answer the questions. Reference SNPs drive probe selection for genotyping arrays and influence the validation of SNPs detected by NGS. We observed a small discordance in genotypes of SNPs included in both HapMap and 1KGP,

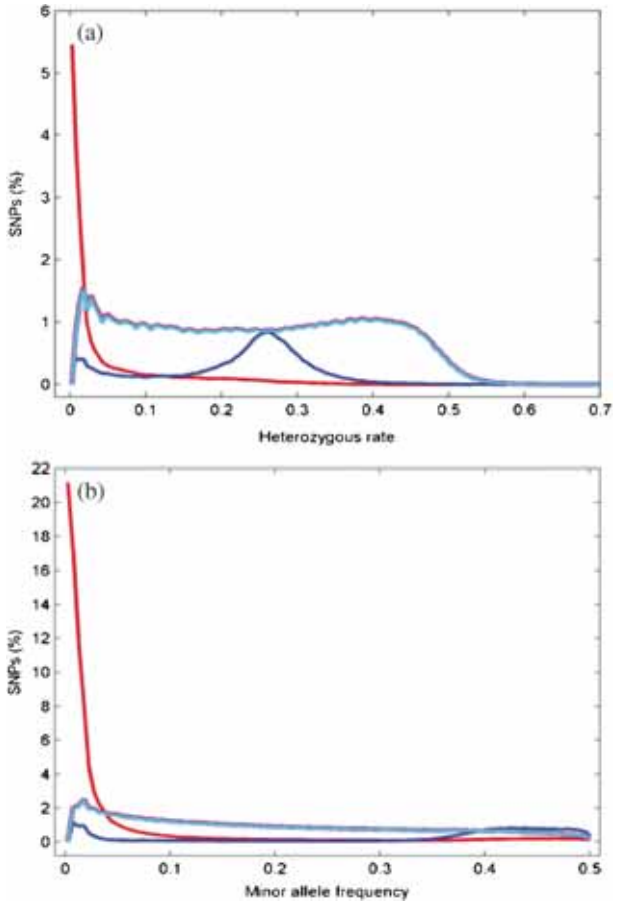

Figure 5. Distributions of heterozygous rate and minor allele frequency for all the SNPs in 622 common individuals. (a) Percentage of the SNPs without heterozygous in the unique SNPs (megenta) and the common SNPs (cyan). (b) Percentage of the SNPs without minor alleles in the unique SNPs (megenta) and the common SNPs (cyan). Distributions of (c) heterozygous rate and (d) minor allele frequency are shown for HapMap only SNPs (cyan), 1KGP only SNPs (magenta), common SNPs from HapMap (blue), common SNPs from 1KGP (red).

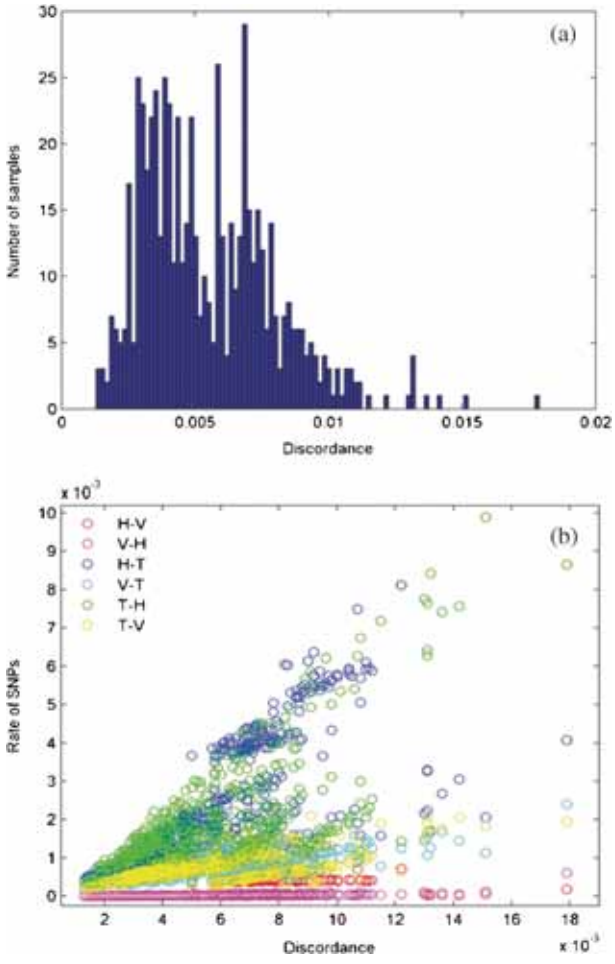

Figure 6. Distributions of genotype discordances in common SNPs between HapMap and 1KGP. Number of samples (*y*-axis) having different discordances in genotypes (*x*-axis) is shown in (a). The overall discordance per sample is illustrated in (b) with six colours representing different discordance sources. The letter on left (b) indicates genotype from HapMap and on right are genotype from 1KGP. H, homozygous; V, variant homozygous; T, heterozygous.

indicating that the genotypes of SNPs detected by genotyping arrays can be confirmed by NGS and, thus, genotyping array-based GWAS findings are still useful. We also found that a large fraction of SNPs from 1KGP had very low MAF, demonstrating the ability of NGS-based GWAS to identify rare SNPs associated with diseases and drug responses.

The nucleotide base compositions of the three SNPs categories (only in 1KGP, only in HapMap and common to both) were very similar (figure 2), suggesting that no overt bias affected SNP detection in either 1KGP or HapMap. Our comparative analysis confirmed that the SNPs provided by both 1KGP and HapMap are sufficiently reliable to serve as references for genetic studies. Interestingly, the base composition of the human genome hg19 was different from the base composition of SNPs detected from both 1KGP and HapMap (figure 2). We believe that the higher base frequencies for cytosine and guanine in SNPs compared to the reference human genome hg19 might be caused by the difference in the interactions between different bases in the double helix structure of human DNA. The human DNA is a normally double-stranded macromolecule in which two polynucleotide chains form the double helix structure that is held together by the weak thermodynamic forces, mainly hydrogen bonds, between the two chains. Within the DNA double helix, two hydrogen bonds are formed between adenine and thymine while three hydrogen bonds link cytosine and guanine. Thermodynamically, mutating bases adenine and thymine (breaking two hydrogen bonds) is easier than mutating bases cytosine and guanine (breaking three hydrogen bonds). Accordingly, the frequencies of cytosine and guanine in SNPs were higher when compared to the reference human genome.

SNPs are mutation products that can be classed into two types: Ts and Tv mutations. Ts mutations are those having interchanges between adenine and guanine that have purine

with two rings, or between cytosine and thymine that contain one ring pyrimidine. Tv mutations are those having interchanges between purine bases (A and G) and pyrimidine bases (C and T), with consequentially large structure effects as one-ring and two-ring structures are exchanged. Although statistically there are twice as many possible Tv mutations, more Ts mutations have been detected because of different chemical reactions involved (Ebersberger *et al.* 2002). However, the ratio of Ts to Tv (Ts/Tv) varies in different genomic regions. Although Ts/Tv is around 2.0 for SNPs from a whole genome, the ratio increases to 2.8–3.0 or higher for SNPs detected in human exomes (Marth *et al.* 2011). One explanation for higher Ts/Tv in exonic regions is the higher prevalence of methylated cytosine in CpG dinucleotides, as deamination of 5-methylated cytosine yields thymine and, thus, creates a Ts mutation of cytosine to thymine (Lynch 2010; Conrad *et al.* 2011). The HapMap SNPs were measured using a variety of SNP genotyping arrays, with most probes in the exonic regions in the human genome. Not surprisingly, then, the Ts/Tv ratios of the SNPs from HapMap for the 622 subjects were different from the Ts/Tv values of the SNPs from 1KGP that were detected across whole genomes. Figure 4a gives SNPs Ts/Tv ratios for the 622 subjects from HapMap grouped into three groups and shows that the newer SNP genotyping arrays had higher Ts/Tv ratios.

MAF and HR are frequently used in association analysis in genetic studies. Our study showed substantial differences in MAF and HR for the 622 subjects common to HapMap and 1KGP, with consistently lower MAF and HR for 1KGP only SNPs than HapMap only SNPs (figure 3). As expected, MAF was correlated with HR for SNPs common to both HapMap (figure 3a) and 1KGP (figure 3b). To ensure the MAF and HR differences between SNPs from 1KGP and HapMap were not caused from quality of the SNPs used in the comparative study, we conducted HWE testing on the SNPs. The distributions of *P*-values from HWE testing are plotted in figure 7 for the SNPs common between HapMap and 1KGP (figure 7a), HapMap only SNPs (figure 7b) and 1KGP only SNPs (figure 7c). The distributions are very similar. There are 11.7 and 9.2% SNVs with a *P*-value less than 0.05 for HapMap only SNPs and 1KGP only SNPs, respectively. Therefore, no obvious SNPs quality bias in terms of the deviations from HWE was observed for HapMap and 1KGP. Both MAF and HR of 1KGP SNPs were much lower than HapMap SNPs. Figure 3a shows that the 662 samples cluster in three groups in accordance with MAF and HR based on HapMap only SNPs. Figure 3b shows that the samples cluster in two groups in accordance with MAF and HR based on 1KGP only SNPs. The reason is apparent from 3D plots of MAF versus HR versus Ts/Tv for the samples, where separate clusters correspond to different genetically related population groups (figure 4). More specifically, the individuals from the three population groups with African ancestry (ASW, LWK and YRI) were markedly separated from other population groups, while the differences between Asians (CHB and JPT) and subjects with European ancestry (CEU, MEX and TSI) were conspicuous

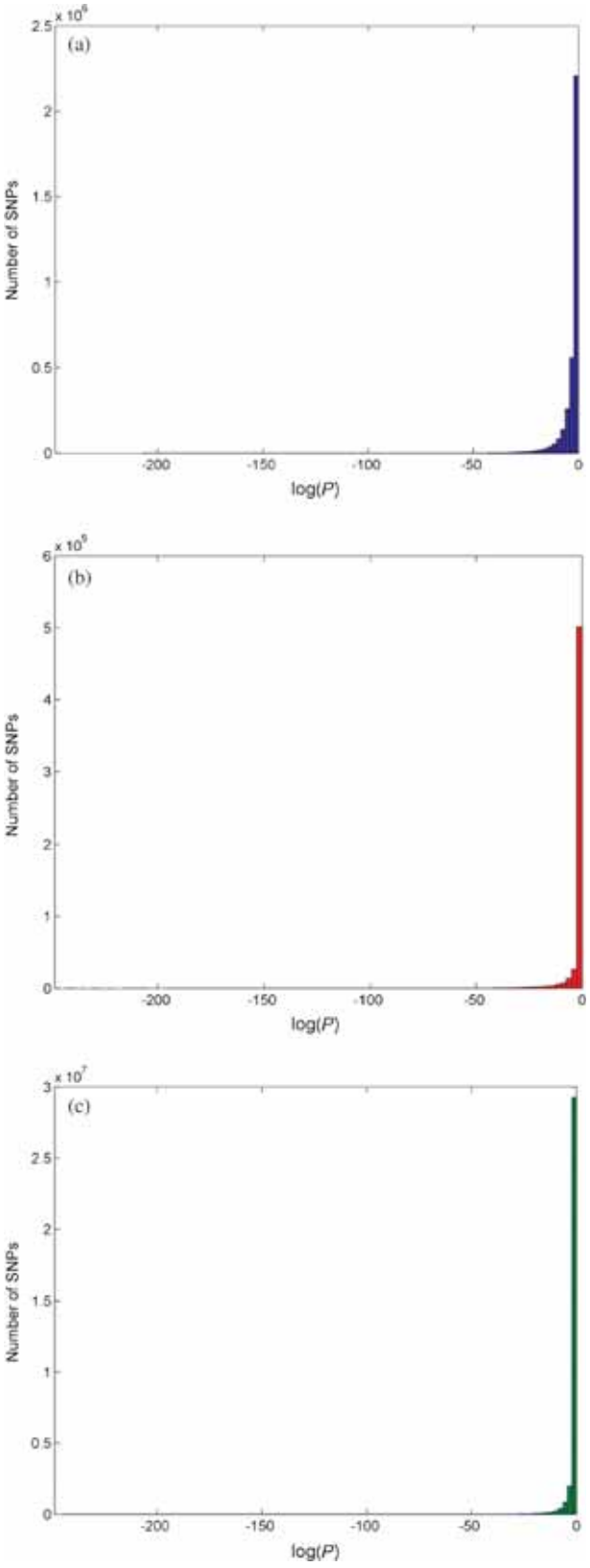

Figure 7. Distributions of *P*-values of HWE tests on SNPs. Number of SNPs (*y*-axis) within a *P*-value range from HWE testing (*x*-axis, logarithmic transform) is plotted as a bar for common SNPs between HapMap and 1KGP (a), HapMap only SNPs (b) and 1KGP only SNPs (c).

(figure 4b), though relatively smaller than African ancestry, using the 1KGP only SNPs (figure 4b) and HapMap only SNPs (figure 4(c and d)).

SNP genotyping array-based GWAS findings have broadened our understanding of the genetics of many complex diseases and phenotypic traits. But the more complete genetic architectures needed for personalized medicine remain elusive. SNP arrays interrogate common SNPs that usually have MAF >5% in a population. The unexplained portion of genetic susceptibility seen in array-based GWAS might be explained by rare SNPs at least in large part. Future GWAS interrogation for rare and common SNPs from whole genome sequencing is widely expected to explain a much larger portion of genetic susceptibility of complex diseases and phenotypic traits than has been previously accessible.

Recently, using imputation based on 1KGP, Wood *et al.* (2013) detected an association between a low frequency variant and phenotype that was previously missed by HapMap-based imputation approaches, demonstrating that imputation using 1KGP will detect novel, low frequency-large effect associations. Our analysis of MAF of SNPs demonstrated that NGS-based 1KGP detected significantly more rare SNPs than arrays-based HapMap (figure 5), further confirming well-placed optimism that NGS-based GWAS will be a major advance in identifying causative associations between individual genetics and complex diseases, and phenotypic traits.

Our comparison of the SNPs common to HapMap and 1KGP showed very small discordances among 622 samples (figure 6a), which mitigated concerns on the reliability and usefulness of the SNP arrays-based GWAS findings; though they appear reliable, the explained genetic variances are incomplete for most personalized medicine applications.

Conclusion

In summary, our comparative analyses demonstrated that the common SNPs from HapMap and 1KGP have high genotype concordances, and that a large portion of SNPs found only in 1KGP were rare genetic variants. We conclude that SNP genotyping array-based GWAS findings are reliable and remain useful, and that the future NGS-based GWAS will completely explain genetic risks of complex diseases as effects of rare SNPs will be ascertained, improving prospects for realizing personalized medicine.

Acknowledgements

This research was supported in part by an appointment to the research participation programme at the National Center for Toxicological Research (Wenqian Zhang, Hui Wen Ng and Heng Luo) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Food and Drugs Administration.

References

Abecasis G. R., Auton A., Brooks L. D., DePristo M. A., Durbin R. M., Handsaker R. E. *et al.* 2012 An integrated map of genetic variation from 1092 human genomes. *Nature* **491**, 56–65.

Buchanan C. C., Torstenson E. S., Bush W. S. and Ritchie M. D. 2012 A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J. Am. Med. Inform. Assoc.* **19**, 289–294.

Chen R., Davydov E. V., Sirota M. and Butte A. J. 2010 Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One* **5**, e13574.

Cirulli E. T. and Goldstein D. B. 2010 Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425.

Collins D. W. and Jukes T. H. 1994 Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386–396.

Conrad D. F., Keebler J. E., DePristo M. A., Lindsay S. J., Zhang Y., Casals F. *et al.* 2011 Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714.

Ebersberger I., Metzler D., Schwarz C. and Paabo S. 2002 Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490–1497.

Eichler E. E., Flint J., Gibson G., Kong A., Leal S. M., Moore J. H. *et al.* 2010 Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450.

Evangelou E. and Ioannidis J. P. 2013 Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389.

Frayling T. M., Timpson N. J., Weedon M. N., Zeggini E., Freathy R. M., Lindgren C. M. *et al.* 2007 A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894.

Gibson G. 2011 Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145.

Ginsburg G. S. and McCarthy J. J. 2001 Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol.* **19**, 491–496.

Hindorff L. A., Sethupathy P., Junkins H. A., Ramos E. M., Mehta J. P., Collins F. S. *et al.* 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367.

Hirschhorn J. N. 2009 Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701.

Hong H. 2012 Next-generation sequencing and its impact on pharmacogenetics. *J. Pharmacogenomics Pharmacoproteomics* **3**, e119.

Hong H., Shi L., Su Z., Ge W., Jones W. D., Czika W. *et al.* 2010a Assessing sources of inconsistencies in genotypes and their effects on genome-wide association studies with HapMap samples. *Pharmacogenom. J.* **10**, 364–374.

Hong H., Su Z., Ge W., Shi L., Perkins R., Fang H. *et al.* 2010b Evaluating variations of genotype calling: a potential source of spurious associations in genome-wide association studies. *J. Genet.* **89**, 55–64.

Hong H., Xu L., Liu J., Jones W. D., Su Z., Ning B. *et al.* 2012a Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS One* **7**, e44483.

Hong H., Xu L., Su Z., Liu J., Ge W., Shen J. *et al.* 2012b Pitfall of genome-wide association studies: sources of inconsistency in genotypes and their effects. *J. Biomed. Sci. Eng.* **5**, 557–573.

Hong H., Zhang W., Shen J., Su Z., Ning B., Han T. *et al.* 2013 Critical role of bioinformatics in translating huge amounts of

next-generation sequencing data into personalized medicine. *Sci. China Life Sci.* **56**, 110–118.

International HapMap C., Frazer K. A., Ballinger D. G., Cox D. R., Hinds D. A., Stuve L. L. *et al.* 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

Klein R. J., Zeiss C., Chew E. Y., Tsai J. Y., Sackler R. S., Haynes C. *et al.* 2005 Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

Kraft P. and Hunter D. J. 2009 Genetic risk prediction—are we there yet? *N. Engl. J. Med.* **360**, 1701–1703.

Lander E. S. 1996 The new genomics: global views of biology. *Science* **274**, 536–539.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J. *et al.* 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Langreth R. and Waldholz M. 1999 New era of personalized medicine: targeting drugs for each unique genetic profile. *Oncologist* **4**, 426–427.

Londin E., Yadav P., Surrey S., Kricka L. J. and Fortina P. 2013 Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations. *Methods Mol. Biol.* **1015**, 127–146.

Lovelock P. K., Spurdle A. B., Mok M. T., Farrugia D. J., Lakhani S. R., Healey S. *et al.* 2007 Identification of BRCA1 missense substitutions that confer partial functional activity: potential moderate risk variants? *Breast Cancer Res.* **9**, R82.

Lynch M. 2010 Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968.

Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J. *et al.* 2009 Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Marian A. J. 2012 Molecular genetic studies of complex phenotypes. *Transl. Res.* **159**, 64–79.

Marth G. T., Yu F., Indap A. R., Garimella K., Gravel S., Leong W. F. *et al.* 2011 The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84.

O'Rawe J., Jiang T., Sun G., Wu Y., Wang W., Hu J. *et al.* 2013 Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28.

Obama B. 2007 The genomics and personalized medicine act of 2006. *Clin. Adv. Hematol. Oncol.* **5**, 39–40.

Pearson T. A. and Manolio T. A. 2008 How to interpret a genome-wide association study. *JAMA* **299**, 1335–1344.

Pritchard J. K. and Cox N. J. 2002 The allelic architecture of human disease genes: common disease-common variant ... or not? *Hum. Mol. Genet.* **11**, 2417–2423.

Ratan A., Miller W., Guillory J., Stinson J., Seshagiri S. and Schuster S. C. 2013 Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One* **8**, e55089.

Rosenfeld J. A., Mason C. E. and Smith T. M. 2012 Limitations of the human reference genome for personalized genomics. *PLoS One* **7**, e40294.

Sharma M., Kruger R. and Gasser T. 2014 From genome-wide association studies to next-generation sequencing: lessons from the past and planning for the future. *JAMA Neurol.* **71**, 5–6.

Su Z., Fang H., Hong H., Shi L., Zhang W., Zhang W. *et al.* 2014 Legacy microarray data in the RNA-seq era—a biomarker investigation. *Genome Biol.* **15**, 523.

The International HapMap Consortium 2003 The international HapMap project. *Nature* **426**, 789–796.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G. *et al.* 2001 The sequence of the human genome. *Science* **291**, 1304–1351.

Wagner M. J. 2013 Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics* **14**, 413–424.

Wang W. Y., Barratt B. J., Clayton D. G. and Todd J. A. 2005 Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118.

Wigginton J. E., Cutler D. J. and Abecasis G. R. 2005 A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893.

Wood A. R., Perry J. R., Tanaka T., Hernandez D. G., Zheng H. F., Melzer D. *et al.* 2013 Imputation of variants from the 1000 genomes project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. *PLoS One* **8**, e64343.

Zhang W., Meehan J., Su Z., Ng H. W., Shu M., Luo H. *et al.* 2014 Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. *BMC Bioinformatics* **15**, S6.

Zhang W., Soika V., Meehan J., Su Z., Ge W., Ng H. W. *et al.* 2015 Quality control metrics improve repeatability and reproducibility of single-nucleotide variants derived from whole genome sequencing. *Pharmacogenomics J.* **15**, 298–309.

Received 19 September 2014, in revised form 13 April 2015; accepted 2 June 2015
Unedited version published online: 12 June 2015
Final version published online: 11 December 2015