Gene 540 (2014) 104–109

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Chip-based direct genotyping of coding variants in genome wide association studies: Utility, issues and prospects

Caroline M. Nievergelt [a,*], Nathan E. Wineinger [b], Ondrej Libiger [c], Phillip Pham [d], Guangfa Zhang [e], Dewleen G. Baker [a], Marine Resiliency Study Investigators [1], Nicholas J. Schork [f,**]

[a] Department of Psychiatry, University of California, San Diego; VA Center of Excellence for Stress and Mental Health, VA San Diego
[b] Scripps Genomic Medicine, Scripps Health; The Scripps Translational Science Institute, The Scripps Research Institute
[c] The Scripps Translational Science Institute, The Scripps Research Institute
[d] Cypher Genomics, Inc
[e] Human Longevity, Inc
[f] J. Craig Venter Institute

ARTICLE INFO

*Article history:*
Accepted 23 January 2014
Available online 9 February 2014

*Keywords:*
Illumina HumanExome array
Expanding GWAS
Genotyping rare SNPs
Coding variants

ABSTRACT

There is considerable debate about the most efficient way to interrogate rare coding variants in association studies. The options include direct genotyping of specific known coding variants in genes or, alternatively, sequencing across the entire exome to capture known as well as novel variants. Each strategy has advantages and disadvantages, but the availability of cost-efficient exome arrays has made the former appealing. Here we consider the utility of a direct genotyping chip, the Illumina HumanExome array (HE), by evaluating its content based on: 1. functionality; and 2. amenability to imputation. We explored these issues by genotyping a large, ethnically diverse cohort on the HumanOmniExpressExome array (HOEE) which combines the HE with content from the GWAS array (HOE). We find that the use of the HE is likely to be a cost-effective way of expanding GWAS, but does have some drawbacks that deserve consideration when planning studies.



## 1. Introduction

Methods to extend genome-wide association studies (GWAS) have recently become a topic of high interest. Despite a large number of notable successes in the discovery of genetic variants associated with various traits, including disease via GWAS, the variants identified to date collectively only explain a small fraction of the estimated heritability of most common, chronic diseases (Manolio et al., 2009). Unknown genetic factors, including polymorphisms that have yet to be identified through GWAS studies, likely account for the 'missing heritability' associated with complex traits (Visscher et al., 2012; Yang et al., 2011). One explanation for this missing heritability is that widely-used genotyping platforms for GWAS are designed to directly interrogate only common single nucleotide polymorphisms (SNPs). Therefore, rare coding variants, which have been shown to play a role in the etiology of many diseases, tend to be entirely omitted by most genotyping platforms used in GWAS as they are not in linkage disequilibrium (hence not imputable) with SNPs interrogated on these arrays (Evans et al., 2008; Sun et al., 2011). Thus, the examination of rare coding variants requires either sequencing technology or the direct genotyping of variants which have previously been identified. While the former may lead to a more comprehensive assessment of all forms of variation in coding regions, including the discovery of extremely rare and/or *de novo* variants, the latter provides an efficient, cost-effective alternative for interrogating a subset of known variants in coding regions (Flannick et al., 2012; Pasaniuc et al., 2012).

The value of direct genotyping of previously identified coding variants, as opposed to *de novo* sequencing of coding regions, is dependent on a few key issues. First, if one can identify known functionally relevant variants in coding regions it might be more expedient to focus on them in cost-effective direct genotyping studies than pursuing more costly

*Abbreviations:* HE, HumanExome array; HOEE, HumanOmniExpressExome array; HOE, HumanOmniExpressGWAS array; GWAS, genome-wide association studies; SNPs, single nucleotide polymorphisms; MRS, the Marine Resiliency Study; PTSD, post-traumatic stress disorder; OEF/OIF, Operation Enduring Freedom/Operation Iraqi Freedom; IRB, Institutional Review Board; HGDP, Human Genome Diversity Project; MAF, minor allele frequency; SNVs, single-nucleotide variants.

* Correspondence to: C.M. Nievergelt, Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093–0737, USA. Tel.: +1 858 534 2567.
** Correspondence to: N.J. Schork, J. Craig Venter Institute, 4120 Torrey Pines Road, La Jolla, CA 92037 USA. Tel.: +1 858 200 1813.
*E-mail addresses:* cnievergelt@ucsd.edu (C.M. Nievergelt), nschork@jcvi.org (N.J. Schork).

[1] The Marine Resiliency Study Investigators include William P. Nash (Boston VA Research Institute); Brett T. Litz (Veterans Affairs Boston Healthcare System, Boston, Massachusetts;); Mark A. Geyer (University of California-San Diego and VA Center of Excellence for Stress and Mental Health); Paul S. Hammer (Defense Centers of Excellence for Stress and Mental Health and Traumatic Brain Injury, Arlington, Virginia); Gerald E. Larsen (Naval Health Research Center, San Diego, California); Daniel T. O'Connor (University of California-San Diego); Victoria B. Risbrough (VA Center of Excellence for Stress and Mental Health San Diego and University of California-San Diego); Jennifer J. Vasterling (Veterans Affairs Boston Healthcare System, Boston, Massachusetts and Boston University); and Jennifer A. Webb-Murphy (Naval Center for Combat & Operational Stress Control, San Diego, California).

0378-1119/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2014.01.069

sequencing studies that may identify many likely neutral variants. Second, if coding variants identified via sequencing are easily imputable from variants genotyped on standard GWAS platforms, then the need for directly genotyping these coding regions would be minimized and greater attention could be given to more reliable imputation strategies. Third, many coding variants, whether they are functional or amenable to imputation or not, are very rare and hence likely to be absent in many global populations. Thus, direct genotyping certain coding variants may only be useful for specific populations.

Here we assessed the potential benefits of directly genotyping rare coding variants on the Illumina Human Exome (HE) array by addressing these issues. As such, our assessment includes an examination of the functional content of variants included on the array. We also evaluated the amenability of the HE markers to imputation from the Illumina Human Omni Express (HOE). And lastly, we evaluated the allele frequency spectrum of the variants included on the HE chip. We find that, overall, the HE chip does not suffer severe drawbacks in the context of these issues, but of course is limited to assessments of known (i.e., previously identified) variants. Our analyses and results have important implications for future studies seeking to identify associations with coding variants.

## 2. Material and methods

### 2.1. Subjects and genotyping

Participants were recruited from two southern Californian military personnel cohorts: 1. the Marine Resiliency Study (MRS), a prospective study of post-traumatic stress disorder (PTSD) involving United States Marines bound for deployment to Iraq or Afghanistan (Baker et al., 2012); and 2. a cross-sectional study of active duty service members and veterans of Operation Enduring Freedom/Operation Iraqi Freedom (OEF/OIF) (Pittman et al., 2012). The protocols for these studies were approved by the University of California-San Diego Institutional Review Board (IRB Protocols #110770, #070533, and #080851), and all subjects provided written informed consent to participate.

DNA samples from 2585 study participants were acquired, and genotyping was carried out by Illumina (http://www.illumina.com/) using the HOEE version 12v1.0. Initial allele calling was performed by Illumina in Genome Studio (http://www.illumina.com) and the overall data quality was high: sample success rate was 99.95% (9 samples failed), locus success rate was 99.86%, and genotype call rate was 99.88%. Twenty-eight replicate pairs of samples undergoing genotyping were assessed for consistency and ultimately reproducibility of the assay and agreement of genotyping calls was achieved for >99.99% over all genotypes across these 28 pairs. Additional data cleaning was performed in PLINK v1.07 (Purcell et al., 2007) and included the removal of 224 markers with heterozygous haploid genotypes on the X, Y, or mitochondrial chromosome. The final dataset included 949,469 markers genotyped in 2548 individuals (2538 males and 10 females) with a genotyping rate greater than 99.8%.

### 2.2. Ancestry determination

We estimated each individual's degree of European, African, Native American, Central Asian, East Asian and Oceanic admixture by comparing the individual's genotypes to allele frequencies of 10,079 SNPs in common with a large set of reference individuals (Libiger and Schork, 2013). In short, the reference sample consisted of genotype data for 2513 individuals of known ancestry who originated from 83 populations from around the world. These data were assembled from publicly available sources including the Human Genome Diversity Project (HGDP) (Cann et al., 2002), the Population Reference (POPRES) (Nelson et al., 2008), HapMap3 (Altshuler et al., 2010), and the University of Utah dataset (Xing et al., 2009). Admixture estimates were obtained in two steps using a supervised analysis implemented in the ADMIXTURE software (Alexander et al., 2009). In the first step, we computed initial admixture estimates for all individuals associated with each world population using the entire set of reference individuals and determined the estimates' standard errors via bootstrapping. A subset of reference individuals from populations that exhibited evidence of contributing to an individual's ancestry based on 95% confidence intervals was then used to refine the initial admixture estimates in a subsequent supervised ADMIXTURE analysis.

Final ancestry calling was based first on self-reported race and ethnicity information and second within each of these main population groups. Essentially, subjects were placed into 5 groups: European Americans (subjects with >95% European ancestry; N =1476), Asian Americans (>95% East Asian ancestry; N = 43); African–American (subjects with >5% African ancestry and <5% Native American, Central Asian, East Asian and Oceanic ancestry; N = 109), Hispanic Americans (subjects with >5% Native American and <10% African, Central Asian, East Asian and Oceanic ancestry; N = 321), and Other (all others; N = 599). Thus, our ancestry assignments provide initial assignments consistent with the often-used admixture program except that they have been refined by removing noise and leveraging comparisons to self-reported ancestries.

### 2.3. Genotype imputations

Imputations were conducted using markers available on the HOE platform. Prior to imputation, mitochondrial and unmapped SNPs were removed from each set. Markers that were individually rare (minor allele frequency MAF < 0.0002), showed a large number of missing genotypes (>5%), or failed Hardy–Weinberg equilibrium ($p < 1 \times 10^{-6}$) were also removed (Supplemental Table 1). Imputations were performed using the default parameters in IMPUTE2 v2.2.2, using 1000 Genomes Phase 1 integrated variant set haplotypes for the autosomes and the interim set for the X chromosome (Howie et al., 2009). IMPUTE2 is well suited for imputations on genetically diverse and admixed populations such as that of the present study as the algorithm is robust to ancestral genetic variation within the reference panel and study datasets (Howie et al., 2011). Genomes were divided into approximately 5 Mb segments (minimum 2.5 Mb, maximum 7.5 Mb to avoid chromosome and centromere boundaries), and phasing and imputed genotypes were calculated for each. Imputed markers with low imputation quality values (Info ≤ 0.5) were dropped. GTOOL v0.7.0 was used to convert genotype probabilities into calls. Individual genotype probabilities exceeding 90% were assigned genotype calls and probabilities ≤90% were treated as missing genotypes. Agreement between the imputation results and markers exclusive to HOEE (i.e., HE markers) was examined by calculating the correlation coefficient, $r^2$, between calls on a per marker level. Missing genotypes were assigned an allelic dosage representing the mean genotype at that particular locus for all calculations. Imputation was also performed based on genotype data from the HOEE platform. A comparison of the agreement between the HOE and HOEE to impute markers that were not genotyped on either platform was, likewise, conducted.

### 2.4. Variant functional annotations

We mapped all variants to the closest gene from the UCSC Genome Browser known gene database (Fujita et al., 2011). Full details of our annotation pipeline are described in a previous publication (Torkamani et al., 2012) and the Supplemental Methods. In brief, variants were associated with all transcripts of the nearest gene(s), with functional impact predictions made independently for each transcript. If the variant fell within a known gene, its position within gene elements (e.g. exons, introns, untranslated regions, etc.) was recorded for functional impact predictions depending on the impacted gene element. All variants falling within an exon were analyzed for their impact on the amino acid sequence (e.g. synonymous, nonsynonymous, nonsense, frameshift, in-frame, intercodon etc.).