Genomics 98 (2011) 422–430

Contents lists available at SciVerse ScienceDirect

Genomics

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

Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm

Thomas J. Hoffmann [a,e,*,1], Yiping Zhan [b,1], Mark N. Kvale [a], Stephanie E. Hesselson [a], Jeremy Gollub [b], Carlos Iribarren [c], Yontao Lu [b], Gangwu Mei [b], Matthew M. Purdy [b], Charles Quesenberry [c], Sarah Rowell [c], Michael H. Shapero [b], David Smethurst [c], Carol P. Somkin [c], Stephen K. Van den Eeden [c], Larry Walter [c], Teresa Webster [b], Rachel A. Whitmer [c], Andrea Finn [b,†], Catherine Schaefer [c,‡], Pui-Yan Kwok [a,d,§], Neil Risch [a,c,e,§]

[a] Institute for Human Genetics, University of California, San Francisco, CA, USA
[b] Affymetrix Incorporated, Santa Clara, CA, USA
[c] Kaiser Permanente Northern California Division of Research, Oakland, CA, USA
[d] Cardiovascular Research Institute, University of California, San Francisco, CA, USA
[e] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

ARTICLE INFO

Article history:
Received 8 August 2011
Accepted 22 August 2011
Available online 28 August 2011

Keywords:
Microarray
Genome-wide association study
Coverage
Imputation
Single nucleotide polymorphism
Throughput

ABSTRACT

Four custom Axiom genotyping arrays were designed for a genome-wide association (GWA) study of 100,000 participants from the Kaiser Permanente Research Program on Genes, Environment and Health. The array optimized for individuals of European race/ethnicity was previously described. Here we detail the development of three additional microarrays optimized for individuals of East Asian, African American, and Latino race/ethnicity. For these arrays, we decreased redundancy of high-performing SNPs to increase SNP capacity. The East Asian array was designed using greedy pairwise SNP selection. However, removing SNPs from the target set based on imputation coverage is more efficient than pairwise tagging. Therefore, we developed a novel hybrid SNP selection method for the African American and Latino arrays utilizing rounds of greedy pairwise SNP selection, followed by removal from the target set of SNPs covered by imputation. The arrays provide excellent genome-wide coverage and are valuable additions for large-scale GWA studies.

1. Introduction

Genome-wide association (GWA) studies have produced a large number of replicated novel genetic variants [1–4] for many diseases for which no variants had been previously found. The success of these studies has been a result of high-throughput genotyping platforms assaying hundreds of thousands to a million SNPs, with large sample sizes leading to an increased number of replicated associations [5,6].

Abbreviations: GWA, genome-wide association; MAF, minor allele frequency; KGP, 1000 Genomes Project; RPGEH, Research Program on Genes, Environment and Health; EUR, European and West Asian; EAS, East Asian; AFR, African American; LAT, Latino; 2-rep, 2 features; 1-rep, 1 feature; ASW, African Ancestry in Southwest USA; CEU, Utah residents with ancestry from Northern and Western Europe from Centre d'Etude du Polymorphisme Humain; CHB, Han Chinese in Beijing; CHS, Han Chinese South; CLM, Colombian in Medellin, Colombia; Fin, Finnish from Finland; GBR, British individuals from England and Scotland; IBS, Iberians in Spain; JPT, Japanese in Tokyo; LWK, Luhya in Webuye Kenya; MXL, Mexican in Los Angeles, CA; PUR, Puerto Rican in Puerto Rico; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria; KGHP, 1000 Genomes High Pass; KPNC, Kaiser Permanente Northern California; AIMs, Ancestry Informative Markers; KG2011, 1000 Genomes interim June 2011 release.

* Correspondence to: T. Hoffmann, Institute for Human Genetics, University of California, San Francisco, 513 Parnassus Ave, Suite S965, Box 0794, San Francisco, CA 94143-0794, USA. Fax: +1 415 476 1356.

† Correspondence to: A. Finn, Affymetrix, 3420 Central Expressway, Santa Clara, CA 95051, USA. Fax: +1 408 731 5380.

‡ Correspondence to: C. Schaefer, Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA. Fax: +1 510 891 3761.

§ Corresponding authors at: Institute for Human Genetics, University of California, San Francisco, 513 Parnassus Ave, Suite S965, Box 0794, San Francisco, CA 94143-0794, USA. Fax: +1 415 476 1127.

E-mail addresses: hoffmannt@humgen.ucsf.edu (T.J. Hoffmann), andrea_finn@affymetrix.com (A. Finn), cathy.schaefer@kp.org (C. Schaefer), pui.kwok@ucsf.edu (P.-Y. Kwok), rischn@humgen.ucsf.edu (N. Risch).

[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.08.007

Many of these have focused on common genetic variation (MAF (minor allele frequency) of 0.10 or greater), based on the HapMap catalog [7]. Sequencing projects, particularly the 1000 Genomes Project (KGP) (http://www.1000genomes.org), are developing larger catalogs which can be leveraged to design arrays that assay lower frequency variants, further enabling discovery of disease-associated genetic variations.

Here we describe the development of three new microarrays for the Axiom Genotyping Solution tailored to individuals of East Asian, African American, and Latino race/ethnicity. These are the remaining three of four custom microarrays developed for the genome-wide genotyping analysis of 100,000 participants in the Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH). A description of the genotyping project and RPGEH cohort is included in [8]. Axiom arrays are limited to approximately 700,000 SNPs when SNPs are tiled with two replicates, which is the standard. Budget constraints for this project allowed for the genotyping of either a single array on 100,000 individuals or two arrays (up to 1.4 million SNPs) on 50,000 individuals. We opted to genotype 100,000 individuals with a single array. As a consequence, however, we chose to design four different arrays to maximize genome-wide coverage, especially for lower frequency variants, in each of the major US race/ethnicity groups (African Americans, East Asians, Latinos and Whites) represented in the RPGEH cohort.

The design of the first array in the series, optimized for US whites (designated EUR), has been described [8]. The East Asian (EAS) array was designed for individuals of East Asian ancestry, although we also included SNPs to provide coverage of European-specific variants to accommodate some RPGEH subjects with mixed East Asian/European ancestry. The target set for the African American (AFR) array included both West African and European variants, recognizing the mixed ancestry of African Americans. Because Latinos have ancestry from three continents, we targeted SNPs common and specific to Europeans, West Africans and Native Americans for the Latino (LAT) array. These arrays were developed to maximize the number of high resolution SNPs for genome-wide coverage; to saturate regions previously identified as disease associated from prior GWA studies for both replication and fine mapping; to improve coverage of both common and uncommon variants by making use of data from the low pass and high pass phases of the KGP; and to incorporate redundant coverage of SNPs with known strong disease associations [8]. For the EAS, AFR and LAT arrays, we used several approaches to enhance the overall genome-wide coverage, including modification to the SNP selection algorithm and reduction of the number of replicates for some SNPs on the array to create more space for additional SNPs.

There have been several methods proposed for SNP selection, starting with a greedy pairwise correlation ("tagging") algorithm [9]. There have also been efforts to extend pairwise tagging to tagging using multi-marker correlations to increase efficiency [10]. However, to our knowledge, less has been done with imputation for tagging, aside from using it to tag singleton SNPs [8].

Imputation has played a major role in the analysis of genome-wide association data [11]; here we explore its use in the design of genotyping microarrays. Imputation of missing SNPs using HapMap reference samples can lead to an overall increase in power of up to 10% [12], and is becoming possible with larger sequenced reference panels, e.g., from the KGP. Simulations show that imputation is potentially the most beneficial for rare variants, which are harder to tag with a single marker [13]. Several papers that imputed all variants in HapMap found significant associations with imputed SNPs that would not have been found by analyzing only the SNPs on the GWA array [14]. Motivated by this analysis strategy of imputing all variants from a reference panel, in this paper, we describe a novel hybrid design method for selection of SNPs for genotype microarrays. The method uses alternating rounds of SNP selection based on pairwise tagging followed by rounds of target set coverage calculations based on imputation r^2 values, which enables removal from the target set of SNPs that can be covered by imputation but were not covered by pairwise tagging. Using this approach, we were able to increase genome-wide coverage with the same fixed number of SNPs on the designed array.

The three new custom arrays described here utilize the Axiom Genotyping Solution (http://media.affymetrix.com/support/technical/datasheets/axiom_genotyping_solution_datasheet.pdf). Briefly, it is a two-color ligation-based assay utilizing 30-mer oligonucleotide probes synthesized in situ on a microarray substrate with automated parallel processing of 96 samples per plate, with a total of ~1.38 million features available for experimental content. In the design of the EUR array, every SNP was represented by at least 2 features (2-rep); some high-value SNPs that had poor resolution were tiled on the array with more than two representations, and hence required more than 2 features (e.g., 4 features or 8 features). As a consequence, the EUR array contains a total of 674,518 SNPs. At the time of design of the EAS, AFR and LAT arrays, it became apparent through analysis of the two representations on the EUR array that the highest resolution SNPs could be tiled on the array with a single feature with only a very small reduction in call rate. We therefore increased the genome-wide coverage of these arrays by tiling some of the highest resolution SNPs with only a single feature (1-rep), enabling greater SNP content on the arrays.

At the time of design of the AFR and LAT arrays, Affymetrix introduced a new reagent kit, Axiom Reagent Kit 2.0. An increased number of SNPs were validated by Affymetrix on the new kit, providing a larger sample of candidate SNPs for the design of these two arrays. The benefits were two-fold: more of the primary, secondary and tertiary SNPs could be directly tiled onto the arrays, and a wider choice of high resolution SNPs were available for selection for genome-wide coverage.

2. Results

2.1. Genome-wide coverage algorithm comparison

Results in Fig. 1 for the HapMap sample African Ancestry in Southwest USA (ASW) and Fig. 2 for the Luhya in Webuye, Kenya (LWK) compare coverage for a hypothetical array designed in the

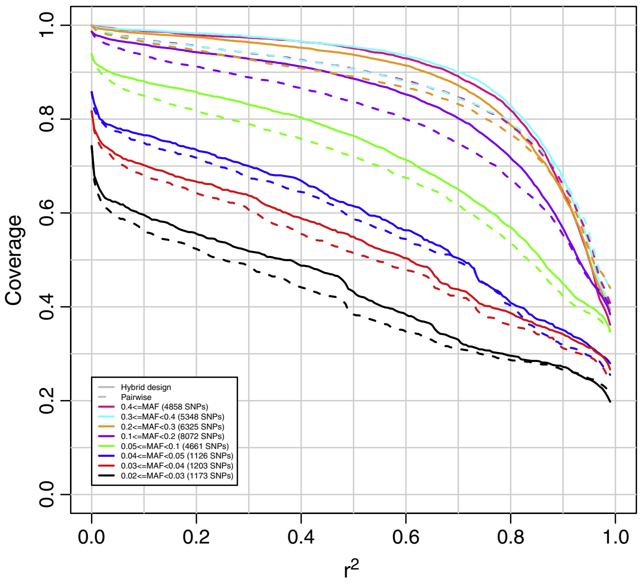

Fig. 1. Chromosome 21 coverage of the African Ancestry in Southwest USA (ASW) population based on two hypothetical arrays, one designed by pairwise tagging and the other by hybrid SNP selection for the Yoruba in Ibadan (YRI) population. Coverage was based on imputation using the YRI population as reference. The numbers in parentheses in the legend are the numbers of markers in the target set in each particular minor allele frequency range.