Development and Evaluation of a Transfusion Medicine Genome Wide SNP Array

Yuelong Guo[1], Grier P Page[2], Mark Seielstad[3,4], Brendan Keating[5], Connie M Westhoff[6], Carolyn Hoppe[7], Aarash Bordbar[8], Bernhard O Palsson[9], Brian Custer[3], Yontao Lu[10], Michael Busch[3]

[1]RTI International, Research Triangle Park, NC, USA; [2]RTI International, Atlanta, GA, USA; [3]Blood Systems Research Institute, San Francisco, CA, USA; [4]Department of Dermatology, University of California San Francisco, San Francisco, CA, USA; [5]The Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [6]New York Blood Center, New York, NY; [7]UCSF Benioff Children's Hospital Oakland, Oakland, CA, USA; [8]Sinopia Biosciences, San Diego, CA, USA; [9]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA; [10]Affymetrix Incorporated, Santa Clara, CA, USA

for the NHLBI Recipient Epidemiology and Donor Evaluation Study-III (REDS-III)

Introduction

The Recipient Epidemiology and Donor Evaluation Study-III (REDS-III) RBC-Omics study has enrolled and characterized fresh (CBC, ferritin) and 42-day stored (spontaneous and stress hemolysis assays) blood from 13,168 blood donors from many races/ethnicities. REDS-III has also developed a cohort of 2,710 sickle cell patients from 6 hemocenters in Brazil (Brazil-SCD) who are being followed longitudinally. High yield DNA was obtained from both cohorts using WBC eluted from LR filters.

We developed an Affymetrix Axiom genome wide SNP array which we call the TM-Array 1) to account for the multi-racial make-up of the RBC-Omics population and 2) to study variation and disease associations of rare variation and transfusion medicine relevant genes in detail.

Objectives

To develop an array for the REDS-III studies, specifically for racially and ethnically diverse cohorts such as RBC-Omics and Brazilian-SCD cohorts

- Genome wide coverage in many ethnic groups
- Transfusion medicine functional contents in depth coverage
- Cost effective

To make the array useful broadly for transfusion medicine community.

Methods

We approached a number of experts in the areas of RBC, platelets, blood groups, transfusion, SCD, Pica, RLS, iron metabolism and other disorders to identify relevant genes and genetic variation that should be queried by the array. We also conducted extensive bioinformatics mining of resources such as PubMed, GTEx, and the GWA SNP catalog. Additionally, we drew upon the Affymetrix catalog of variation and existing arrays such as the UK Biobank array and the transplant array (Li et al 2015).

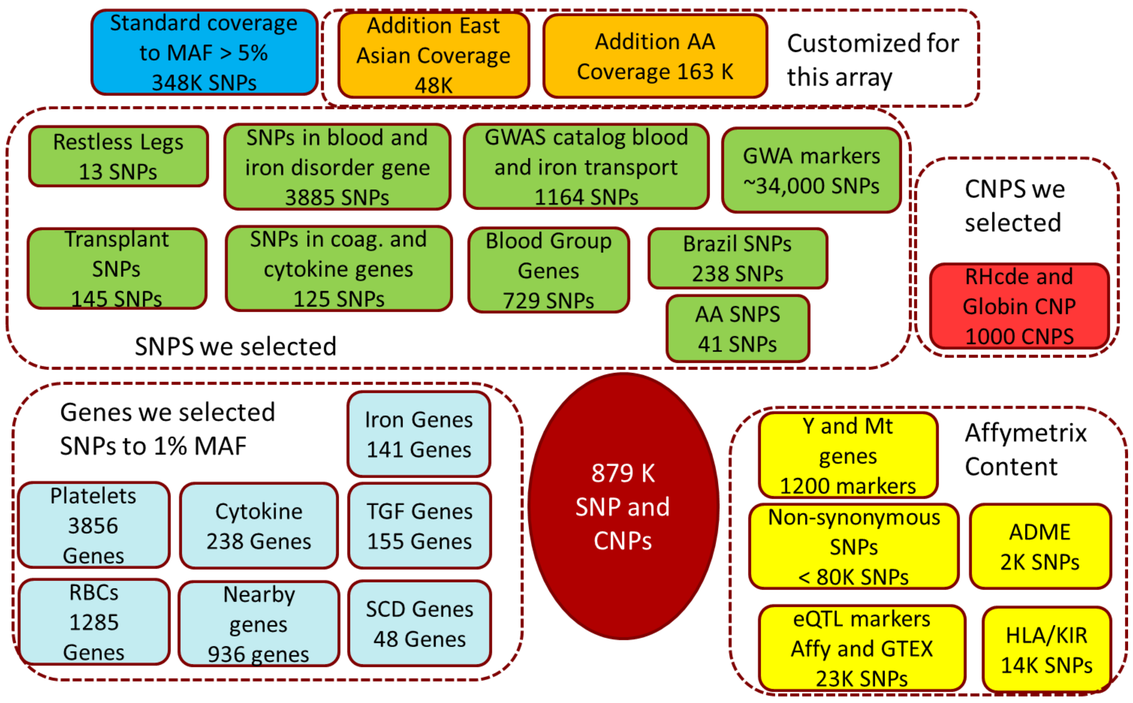
Table 1: Estimated chromosome coverage for imputation by racial group for $r^2 \geq 0.8$

Chromo-some	East Asian	African American		Chrom	East Asian	African American
1	91.93%	91.22%		12	91.19%	91.08%
2	91.45%	91.64%		13	92.03%	91.47%
3	92.13%	92.30%		14	91.47%	90.13%
4	91.70%	91.53%		15	91.46%	91.27%
5	92.02%	91.88%		16	91.51%	91.39%
6	91.84%	91.70%		17	91.49%	90.04%
7	91.84%	90.36%		18	91.80%	90.45%
8	91.96%	92.34%		19	90.77%	89.65%
9	91.22%	89.94%		20	91.30%	92.06%
10	91.63%	91.41%		21	91.77%	91.14%
11	91.83%	91.56%		22	90.58%	90.00%
				X	86.10%	87.00%

Results – Array Design

The TM-Array includes 549,000 SNPs to provide genome coverage down to a minor allele frequency (MAF) of 5% in European, African, and East Asian descended populations (table 1). In addition we included 145 transplant relevant SNPs, 13 Restless Legs Syndrome SNPs, 3,885 SNPs in blood and iron disorder genes, 729 SNPs in blood group genes, all 34,000 SNPs previously associated at a disorder in a GWA study, and 125 SNPs in coagulation and cytokine genes. The array was filled by SNPs down to 1% in gene regions of 141 iron related genes, 3,856 platelet genes, 1,285 RBC genes, 238 cytokines genes, 155 TGF genes, and 48 SCD related genes.

Figure 1: TM-Array Content



We tiled the alpha globin, beta globin, and RHD/RHCE loci with ~1000 copy number polymorphisms (CNPS) to allow the detection of copy number changes. The TM-Array also includes 1200 Y and mitochondrial SNPs, all ~80,000 non-synonymous SNPs with MAF > 1% in European populations, 40 SNPs common in African populations, 14,000 SNPs in the HLA/KIR regions, and 2,000 SNPs in absorption, distribution, metabolism, and excretion (ADME) genes. These lists have overlap with each other and some trimming was performed on loci in high linkage disequilibrium. In total, the TM-Array includes 879,348 SNPs and CNPs (Figure 1).

Table 2 Study	Group ID	Average number of SNPs called (percentage)	Number of Samples < 98% SNPs called	Number of Samples < 95% SNPs called	Number of Samples < 90% SNPs called
RBC Omics	AFR AMRCN	833,339.46 (99.24%)	63 (4.08%)	1 (0.06%)	0 (0%)
	ASIAN	834,021.29 (99.32%)	58 (3.65%)	5 (0.31%)	0 (0%)
	CAUC	833,788.97 (99.29%)	294 (3.59%)	17 (0.21%)	0 (0%)
	HISP	833,547.8 (99.27%)	35 (3.51%)	0 (0%)	0 (0%)
	OTHER	833,684.78 (99.28%)	19 (3%)	0 (0%)	0 (0%)
SCD	SCD	833,740.24 (99.29%)	469 (17.09%)	23 (0.84%)	0 (0%)

Results – Array Performance

Across both REDS-III cohorts and among all racial groups the TM-Array is performing well. In the 16,417 samples genotyped we have generated 14,436,256,116 genotypes. On average, over 99.2% of markers are being called well. Few samples (<0.5%) are calling less than 95% of markers. (table 2)

To assess quality of genotype calling, we called genotypes twice using different permutations of arrays. Comparison of the genotype calls revealed inconsistent genotype calls in 0.08% of high quality markers, primarily (97.6%) no call/call. Inconsistencies were seen in more than 5 people among 5,704 markers. These markers will not be used in analyses.

Examination of genotype calling metrics revealed that the cluster plots are high quality for 833,872 SNP markers (95.2% of markers).

Table 3 shows that the array is generating appropriate minor allele frequencies and can call rare genotypes (<1%) well.

Table 3 Gene	MAF – SCD	MAF – RBC-Omics	SNP rsID	Better known as
ABO	0.1193	0.122	rs7853989	C = likely B blood group;
ABO	0.3014	0.3491	rs8176719	"-" = likely O blood group
ACKR1	0.4606	0.1158	rs2814778	C = Duffy Null/FYB
FUT1	0.02097	0.04386	rs2071699	A = h/Bombay
SLC14A1	0.4706	0.3779	rs1058396	A= JKB/Kidd
KEL	0.00645	0.0051	rs8176059	A=Kel2/Cellano
SLC4A1	0.00443	0.01312	rs2285644	A=Diego

MAF= minor allele frequency
rsID= reference SNP cluster ID

Conclusions

The TM-Array is performing well, with over 99% of SNPs being called across 13,168 in the REDS-III/RBC Omics study and 2,710 in the RED-III Brazil SCD cohort. Individually 99.8% of samples have greater than 95% of SNPs being called. Of the biallelic SNPS, 95.2% (833,872/875,837) are being called with high confidence.

The TM-Array is a powerful tool for studying blood and transfusion related disorders in many ethnic groups. The array is working well across the first 16,000 people genotyped. The number of genotypes per sample called is high and genotype calling is accurate.

We are happy to share the TM-Array design with anyone who would like to use it.

References and Thanks

Thank you to all the REDS-III subjects who selflessly agreed to participate in these studies.

Li et al Concept and design of a genome-wide association genotyping array tailored for transplantation-specific studies. Genome Med. 2015 Oct 1;7:90.

UK Biobank Array http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ .