Application Brief

A SNP array for human population genetics studies

Yontao Lu[1], Teri Genschoreck[1], Swapan Mallick[2,3], Amy Ollmann[1], Nick Patterson[3], Yiping Zhan[1], Teresa Webster[1], David Reich[2,3]

1. Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051
2. Harvard Medical School Department of Genetics, Boston, MA 02115
3. Broad Institute of Harvard and MIT, Cambridge, MA 02142

applied biosystems
by **Thermo Fisher Scientific**

Now a part of our Microarray solutions portfolio
Visit us at **thermofisher.com/microarray**

Abstract

High-density arrays that simultaneously genotype hundreds of thousands of single nucleotide polymorphisms (SNPs) can theoretically be a powerful tool for population genetics studies, whether for learning about human history or natural selection. However, SNP arrays designed for medical genetics have been limited in their utility for population genetics because the polymorphisms on the arrays were discovered and selected for inclusion in a complicated manner that is difficult to model. This "ascertainment bias" has severely limited the types of population genetic analyses that can be carried out with SNP arrays. Here we report on the first SNP array developed specifically for human population genetics studies.

A total of 1.81 million candidate SNPs, all from genome locations covered by sequencing reads from Neandertals, Denisovans, and chimpanzees, were ascertained using a simple SNP discovery procedure first described by Keinan, *et al.*, 2007[1]. The most important ascertainment involved using whole-genome shotgun sequencing data to discover differences between the two chromosomes carried by individuals from 11 populations (San, Yoruba, Mbuti, French, Sardinian, Han, Cambodian, Mongolian, Karitiana, Papuan, and Bougainville). This resulted in 13 panels of SNPs ascertained in simple ways that are appropriate for population genetic analysis (Figure 1). An Axiom® genotyping screen was carried to validate 1.35 million SNPs on the Axiom platform (Table 2). A total of 542,399 SNPs were selected that produced genotypes that passed rigorous quality thresholds appropriate for inclusion in a commercial array design. To facilitate joint analysis with other data sets that have been collected on diverse populations, the final array contains an additional approximately 87,044 SNPs that overlap between the Affymetrix® Genome-Wide Human SNP Array 6.0 and Illumina® 650Y Array.

We genotyped 943 unrelated samples from 53 populations in the CEPH-Human Genome Diversity Panel (Cann, *et al.*, 2002[2]), and have made the data freely available in the CEPH-HGDP database (Table 1). In addition, we report on the genotyping of approximately 400 samples from 70 other diverse worldwide populations. Preliminary analyses demonstrate the promise of this SNP array for population genetics. As examples, we use the data to provide a new line of evidence for gene flow from Neandertals into modern humans (Figure 2).

Figure 1: From candidate SNPs ascertained from 13 panels to the Axiom® Human Origins Array

A. Select 1.35M of 1.81M candidate SNPs for validation

B. Phase I SNP validation

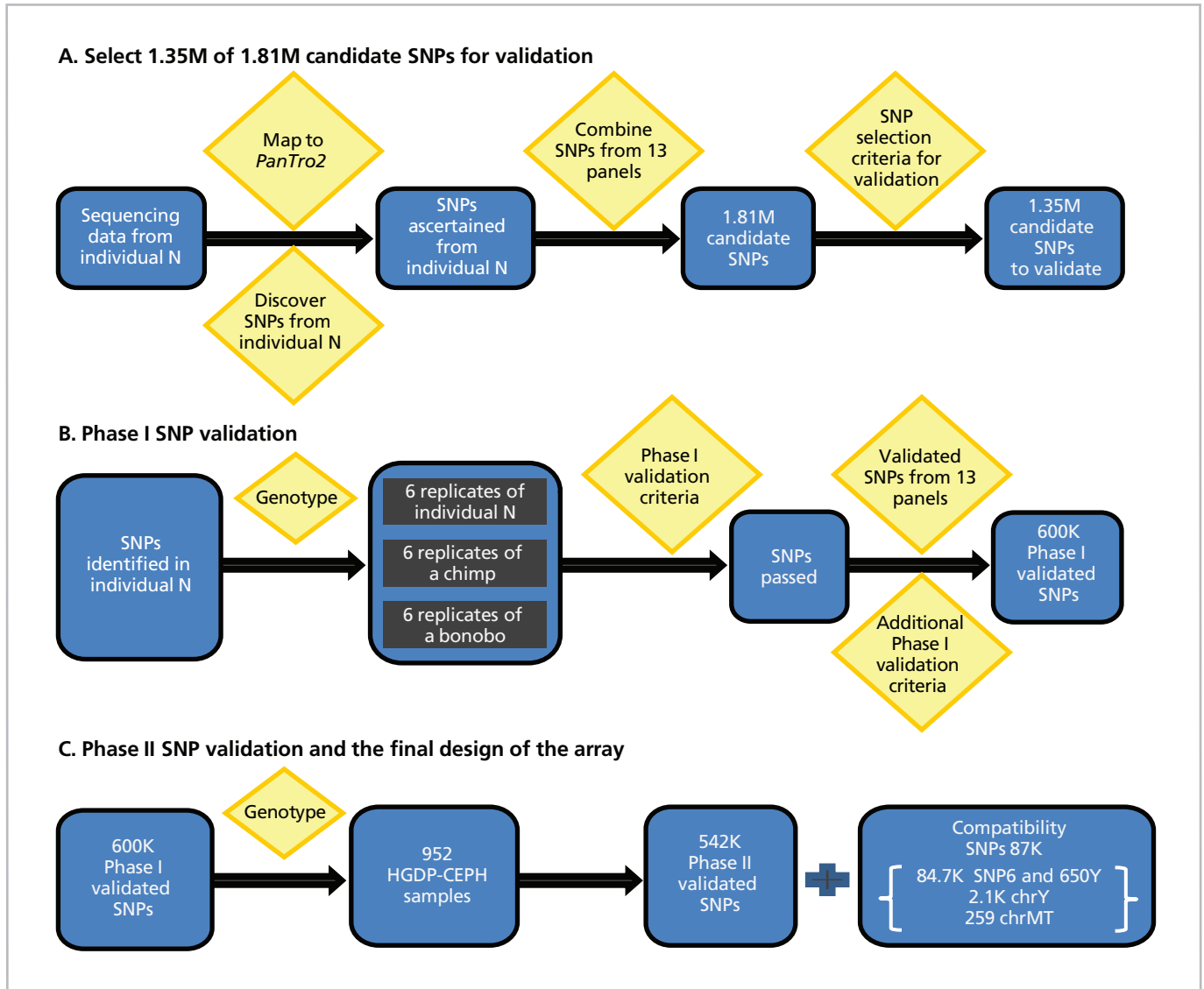### C. Phase II SNP validation and the final design of the array

Figure 1: An overview of the Axiom Human Origins Array development.

A. Sequencing data were mapped to the *PanTro2* reference genome and heterozygous sites were identified. SNPs in Panels 1-12 were discovered using the strategy described by Keinan, *et al*[1]. For Panel 13, SNPs were ascertained by considering one allele from a San individual and one allele from Denisova. SNPs whose ancestral and derived alleles were from Denisova and the San individual, respectively, were kept.

B. Phase I SNP validation: Candidate SNPs discovered in Panel N (N:1-13) were tested by genotyping 18 samples (6 x chimpanzee, 6 x bonobo, and 6 x individual N). During Phase I validation, 1.35M of 1.81M unique candidate SNPs were tested. A total of 599K SNPs were validated in at least one of the 13 panels.

C. Phase II SNP validation: SNPs that passed Phase I validation were further tested against 952 unrelated HGDP-CEPH samples to evaluate the performance. 943 samples and 542K SNPs passed the validation criteria. An additional 87K SNPs were added to the Axiom Human Origins Array to allow haplotype inference at mitochondrial DNA and the Y chromosome, and to provide overlap with previous Affymetrix and Illumina genotyping arrays. The final design contains 629K SNPs.

Details about SNP ascertainment, array design, and SNP validation are described in the technical design document[3].

Table 1: Performance metrics for the Axiom Human Origins Array

Performance metric Results	Overlap with other arrays
Sample pass rate	98.90%
Sample call rate	99.56%
Sample concordance	99.86%
Reproducibility	99.86%

Table 1: Performance metrics for the Axiom Human Origins Array. Numbers are derived from genotyping 952 HGDP-CEPH samples and four HapMap samples (replicated 11 times each).

Table 2: SNPs available in the Axiom Human Origins Array

Ascertainment panel	Sample ID	Sequencing depth	# Candidate SNPs found	# SNPs placed on screening arrays	# Phase I validated SNPs	# Phase II validated SNPs
1 – French	HGDP00521	4.4	333,492	241,707	123,574	111,970
2 – Han	HGDP00778	3.8	281,819	204,841	87,515	78,253
3 – Papuan1	HGDP00542	3.6	312,941	232,408	56,518	48,531
4 – San	HGDP01029	5.9	548,189	401,052	185,066	163,313
5 – Yoruba	HGDP00927	4.3	412,685	302,413	136,759	124,115
6 – Mbuti	HGDP00456	1.2	39,178	28,532	14,435	12,162
7 – Karitiana	HGDP00998	1.1	12,449	8,535	3,619	2,635
8 – Sardinian	HGDP00665	1.3	40,826	29,358	15,260	12,922
9 – Melanesian	HGDP00491	1.5	51,237	36,392	17,723	14,988
10 – Cambodian	HGDP00711	1.7	53,542	38,399	20,129	16,987
11 – Mongolian	HGDP01224	1.4	35,087	24,858	12,872	10,757
12 – Papuan2	HGDP00551	1.4	40,996	29,305	14,739	12,117
13 – Denisova-San	Denisova-HGDP01029	–	418,841	308,210	166,422	151,435
Unique SNPs from 13 panels			1,812,990	1,354,003	599,175	542,399
Compatibility SNPs			NA	NA	NA	87,044

Table 1: Numbers of SNPs at different stages for different panels. Genotypes of 629,443 SNPs across 943 unrelated HGDP-CEPH individuals (Harvard HGDP-CEPH Genotypes) were available at http://www.cephb.fr/en/hgdp/

Figure 2: Preliminary results

A. Example allele frequency spectra

B. Han spectrum if Denisova is derived, Neandertal ancestral

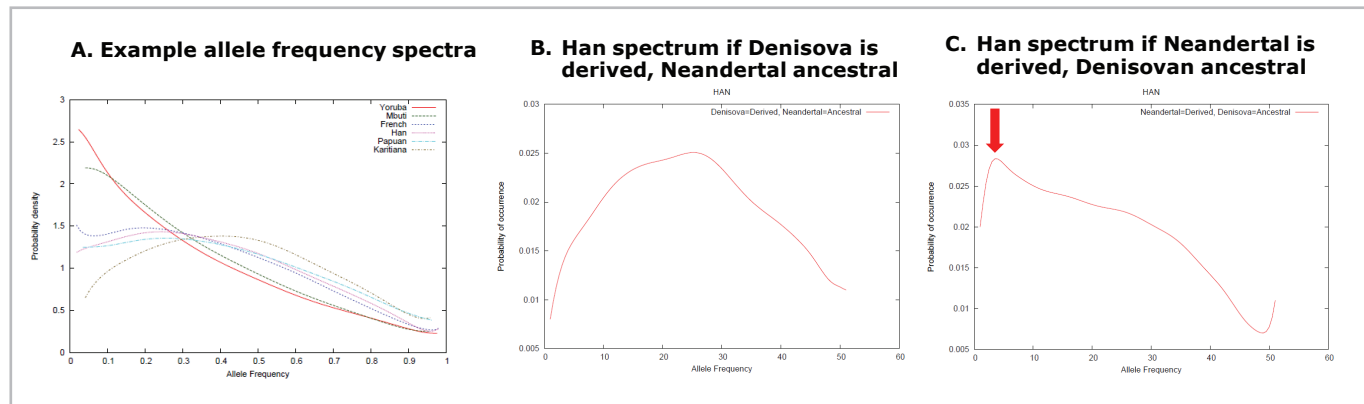**C. Han spectrum if Neandertal is derived, Denisovan ancestral**

Figure 2: Allele frequency spectra

A. Allele frequency spectra for six populations, based on ascertaining in two chromosomes from an individual in that population. The highest rate of rare alleles is in Yoruba, reflecting large population sizes in this group, whereas Karitiana have the lowest rate of rare alleles, indicating strong bottlenecks.

B. Han allele frequency spectrum at sites where Denisova carries the derived allele and Neandertal is ancestral. We observe the upside-down U expected if these were polymorphic in the ancestral population.

C. Han allele frequency spectrum where Neandertal carries the derived allele and Denisova is ancestral. We see an excess of low frequency alleles compared with B, which can only be explained by recent gene flow from Neandertals into modern humans

References

1. Keinan A., Mullikin J. C., Patterson N., Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics* **39**:1251-5 (2007).
2. Cann H., *et al.* A human genome diversity cell line panel. *Science* **296**:261-2 (2002).
3. Lu Y., Patterson N., Zhan Y., Mallick S., Reich D. Technical design document for a SNP array that is optimized for population genetics. ftp://ftp.cephb.fr/hgdp_supp10/8_12_2011_Technical_Array_Design_Document.pdf