File S1
Technical details of a SNP array optimized for population genetics

Yontao Lu, Nick Patterson, Yiping Zhan, Swapan Mallick and David Reich

Overview

One of the promises of studies of human genetic variation is to learn about human history and also to learn about natural selection.

Array genotyping of hundreds of thousands of SNPs simultaneously—using a technology that produces high fidelity data with an error rate of ~0.1%—is in theory a powerful tool for these studies. However, a limitation of all SNP arrays that have been available to date is that the SNPs have been chosen in a complicated way for the purpose of medical genetics, biasing their frequencies so that it is challenging to make reliable population genetic inferences. In general, the way that SNPs have been chosen for arrays is so complicated that it has been effectively impossible to model the ascertainment strategy and thus to correct for the bias.

This technical note describes the design, validation, and manufacture of an array consisting of SNPs all ascertained in a clearly documented way. We anticipate that this will provide a useful resource for the community interested in learning about history and natural selection. We hope that this array will be genotyped in many different cohorts, as has been done, for example, in the Marshfield panel where approximately 800 microsatellites have been genotyped in diverse populations[1,2,3,4,5]. By establishing a common set of simply ascertained SNPs that have been genotyped in diverse populations, it should be possible to learn about human history not only in individual studies, but also through meta-analysis.

The array is designed as a union of 13 different SNP panels. In our experience, a few tens of thousands of SNPs is enough to produce powerful inferences about history with regard to summary statistics like measurements of F_{ST}. Thus, it is better for many analyses to have (for example) 13 sets of tens to hundreds of thousands of SNPs each with its own ascertainment strategy than a single set of 600,000 SNPs. We have included a particularly large number of SNPs from particularly interesting ascertainments—discovery in the two chromosomes of a single San Bushman, a single Yoruba West African, a single French, a single Han Chinese, and a single Papuan—as for some analyses like scans of selection it is valuable to have dense data sets of hundreds of thousands of SNPs. All SNPs chosen for the array were selected from sites in the genome that have read coverage from Neandertals, Denisovans, and chimpanzees, allowing users of the array to compare data from modern humans to archaic hominins and apes.

This array is <u>not</u> ideal for gene mapping, since: (i) No attempt has been made to tag common variation genome-wide. (ii) There are gaps in the genome where no homologous sequence is available from chimpanzee. (iii) Unlike many existing arrays, we have not oversampled SNPs in the vicinity of genes, or adjusting SNP density in order to fully tag haplotypes. Instead we simply sampled SNPs in proportion to their genomic density as discovered by sequencing.

The array is being made commercially available by Affymetrix. Importantly, the academic collaborators who have been involved in the design will not benefit from sales of the array (they

will not receive any financial compensation from Affymetrix). The CEPH-Human Genome Diversity Project (CEPH-HGDP) samples that were genotyped during the course of the project will not be used for any commercial purposes. Affymetrix deposited the genotypes of unrelated CEPH-HGDP samples, collected as part of the array development, into the CEPH-HGDP database on August 12, 2011, more than six months before commercial release of the array (in Spring 2012), and this genotyping data is freely available to the public.

Design strategy for the 13 panels

(Panels 1-12) Discovery of heterozygous sites within 12 individuals of known ancestry
The first 12 SNP ascertainment strategies are based on the idea of the Keinan, Mullikin et al. Nature Genetics 2007 paper[6]. That paper takes advantage of the fact that by discovering SNPs in a comparison of two chromosomes from the same individual of known ancestry, and then genotyping in a larger panel of samples from the same population, one can learn about history in a way that is not affected by the frequency of the SNP in human populations. In particular, even though we may miss a substantial proportion of real SNPs in the individual (false-negatives), and even if a substantial proportion of discovered SNP are false-positives, we expect that the inferences about history using SNPs discovered in this way will be as accurate as what would be obtained using SNPs identified from deep sequencing with perfect readout of alleles.

To understand why false-negative SNPs should not bias inferences, we note that if a SNP is truly heterozygous in the individual in whom we are trying to discover it, there is exactly one copy of the ancestral allele and exactly one copy of the derived allele. Thus, conditional on the SNP being heterozygous in the discovery individual, its probability of being discovered is not further affected by whether it has a high or low minor allele frequency in the population. This contrasts with ascertainment strategies that discover SNPs in more than one individual, where there is always a real (and extremely difficult to quantify) bias toward missing rarer variants. By genotyping SNPs discovered in this way, and making a simple $p(1-p)$ correction for discovery in two chromosomes (where p is the minor allele frequency), one can obtain an unbiased reconstruction of the allele frequency distribution in the population.

An important feature of this SNP discovery strategy is that false-positive SNPs (for example, due to sequencing error, mapping error, segmental duplications or copy number variation) are not expected to substantially bias inferences. The reason is that we have validated all candidate SNPs by genotyping them using a different technology, and we have required the genotypes to match the individuals in whom they were discovered. Thus, we expect to have a negligible proportion of false-positive SNPs on the final array.

This procedure has produced 12 panels of uniformly discovered SNPs, which can be used for allele frequency spectrum analysis. There is some overlap of SNPs across panels. Importantly, we have separately determined validation status for the SNPs in each panel, and have only used SNPs that validate in the same sample in which they were discovered. Thus, we have not biased toward SNPs with a high minor allele frequency, or that are polymorphic across multiple populations, which might be expected to have a higher chance of validation if we did not perform the validation in each discovery sample independently.

(Panel 13) SNPs where a randomly chosen San allele is derived relative to an archaic hominin
A 13th ascertainment strategy used alignments of three genomes: chimpanzee, Denisova (an archaic hominin from southern Siberia for whom there is 1.9× genome sequence coverage[7]), and San. We examined sites where we had ≥1-fold coverage of Denisova, and ≥3-fold coverage of San. We made an allele call for each individual by majority rule, randomly selecting an allele when there was a tie (this means that we are effectively sampling one of two haplotypes in the individual, and the allele call is not expected to be being biased if the individual is heterozygous at that site). We placed on the array the subset of sites where San is derived relative to both Denisova and chimpanzee, in this case requiring agreement between the Denisova and chimpanzee allele. These are sites that likely arose due to mutations in the last million years.

We chose to use San rather than another modern human for building this panel because there is evidence that the San are approximately symmetrically related to all other present-day humans[8]. Panel 13 is also the only one with SNPs from chromosome X (all the other panels are based on SNPs discovered in males), and thus this panel permits X-autosome comparisons.

Description of the sequencing data and filtering used in SNP ascertainment

The sequencing data that we use for identifying candidate SNPs has been described in two recent papers: Green et al. 2010[9] and Reich et al. 2010[7]. The data were all generated in the Max Planck Institute in Leipzig using Illumina Genome Analyzer IIx (GAIIx) sequencing instruments via protocols that are described in refs. 9 and 7 (Table 1). Population genetic analyses for ref. 7 were carried out on the very data file that was used to select SNPs for the array.

Table 1: Characteristics of the sequencing data we are using for SNP ascertainment

Name	Identifier	Sequenced by	Genomic coverage*	Cutoff† A (Pr)	Cutoff† C (Pr)	Cutoff† G (Pr)	Cutoff† T (Pr)
Han	HGDP00778	Green 2010	3.8	16 (0.489)	14 (0.239)	17 (0.003)	15 (0.11)
Papuan1	HGDP00542	Green 2010	3.6	13 (0.051)	10 (0.119)	15 (0.434)	13 (0.880)
Yoruba	HGDP00927	Green 2010	4.3	17 (0.692)	14 (0.440)	18 (0.562)	16 (0.985)
San	HGDP01029	Green 2010	5.9	17 (0.830)	15 (0.914)	18 (0.649)	16 (0.877)
French	HGDP00521	Green 2010	4.4	17 (0.317)	16 (0.985)	18 (0.024)	17 (0.515)
Mbuti	HGDP00456	Reich 2010	1.2	17 (0.041)	14 (0.504)	17 (0.704)	16 (0.379)
Karitiana	HGDP00998	Reich 2010	1.1	18 (0.210)	14 (0.126)	17 (0.147)	17 (0.589)
Sardinian	HGDP00665	Reich 2010	1.3	19 (0.789)	15 (0.302)	18 (0.474)	17 (0.200)
Bougainville	HGDP00491	Reich 2010	1.5	18 (0.810)	14 (0.288)	17 (0.445)	16 (0.291)
Cambodian	HGDP00711	Reich 2010	1.7	18 (0.717)	14 (0.303)	17 (0.331)	16 (0.398)
Mongolian	HGDP01224	Reich 2010	1.4	18 (0.371)	15 (0.789)	17 (0.051)	16 (0.090)
Papuan2	HGDP00551	Reich 2010	1.4	17 (0.188)	14 (0.661)	17 (0.932)	16 (0.885)
Neandertal	Vindija.3.bones	Green 2010	1.3	27 (0.428)	26 (0.049)	27 (0.308)	27 (0.579)
Denisova	Phalanx	Reich 2010	1.9	40 (1.000)	40 (1.000)	40 (1.000)	40 (1.000)

* Genomic coverage is calculated for the modern humans as (# of reads mapping to chimpanzee) × (read length which is 76bp for Green et al. 2010 and 101bp for Reich et al. 2010) × (0.95 as we filtered out the 5% of the lowest quality data) / (2.8 Gb). For the archaic hominins we report the coverage from the abstracts of Green et al. 2010 and Reich et al. 2010.

† For each base used in SNP discovery, we give the quality score cutoff and probability of acceptance at that cutoff (parentheses). The cutoffs are chosen to filter out the data of the lowest 5% quality for each nucleotide class (SI 6; Reich et al. 2010).

The 12 modern human samples are all from the CEPH-HGDP panel. A valuable feature of this panel is that DNA for all samples is available on request on a cost-recovery basis for researchers who wish to carry out further sequencing and genotyping analysis on these samples for the purpose of research into human population history[8,10]. Five of the samples (San, Yoruba, Han, French and a Papuan) were sequenced by Green et al. 2010 using Illumina paired-end 76bp reads[9], while the remaining 7 (Mbuti, Sardinian, Karitiana, Mongolian, Cambodian, Bougainville, and a second Papuan) were sequenced by Reich et al. 2010 using Illumina paired-end 101bp reads[7]. All reads from all 12 samples were mapped to chimpanzee (*PanTro2*). To filter the sequence data for analysis, we used a similar procedure as described in Reich et al. 2010[7], removing the lowest quality of 5% of nucleotides on a sample and nucleotide-specific basis to maximize the amount of sequencing data available for analysis. After this procedure, we had 3.6-5.9× coverage for the 5 samples and 1.1-1.7× for the 7 samples (Table 1).

We also used data from 4 ancient DNA samples to aid our choice of SNPs. To represent Neandertals, we used a pool of sequences from 3 bones from Vindija Cave in Croatia (Vi33.16, Vi33.25 and Vi33.26) for which we had 1.3× genome coverage altogether[9]. To represent Denisovans, we used data from a finger bone (fifth distal manual phalanx) from the Altai mountains of southern Siberia, with 1.9× coverage[7].

<u>All reads are mapped to chimpanzee and a chimpanzee allele is available</u>
We mapped sequencing reads from modern and ancient genomes to the chimpanzee reference sequence (*PanTro2*) to avoid biases toward one present-day human group more than another.

<u>We filtered out reads with a substantial probability of poor mapping</u>
Each read that we analyzed had a mapping quality score (MAPQ) that reflects the confidence of its mapping to *PanTro2*. Based on empirical exploration of the usefulness of the scores, which were generated by either the ANFO or BWA software, we only used reads that had MAPQ of at least 90 for Neandertal (ANFO mapping), 37 for Denisova (BWA), and 60 for present-day humans (BWA). We also rejected reads if the alignment to the chimpanzee resulted in any insertion/deletion difference. This filter was applied in addition to the filtering of Table 1.

<u>Filtering of sites with ≥2 alleles not matching chimp across the humans used for SNP discovery.</u>
At a small proportion of sites, we observe more than one non-ancestral allele in the individual sequencing data used for SNP discovery. Such sites cannot be due to a single historical mutation. Instead, the data must reflect at least two mutations or sequencing errors. We filter out such sites.

For a very small fraction of sites, we found that the derived allele is *different* depending on which human is used in SNP discovery (these are potentially triallelic SNPs in the population, although they are not triallelic in the discovery individual). We keep such sites in our list of SNPs for designing, and use multiple probe sets to assay such SNPs.

The raw data file that emerges from this process is available on the "orchestra" Harvard Medical School filesystem at: /groups/reich/CLEAN_SNP_ARRAY/rawsnps and is freely available from David Reich on request (a README file is in the same directory at rawsnps_readme) (Table 2). For brevity, this file only lists the 2,173,116 SNPs where 2 copies of the derived and 1 copy of the ancestral allele are observed a hominin; these are the only SNPs that are candidates for inclusion. Thus, it is an abbreviated version of a larger file used in analyses for ref. 7.

Filtering the nucleotide calls of the lowest reliability

(a) We do not use nucleotides for which there is no valid nucleotide call for chimpanzee.
(b) For Neandertals, we do not use nucleotides within 5 nucleotides of either end of the reads, because of the elevated rate of ancient DNA degradation errors that we empirically observe.
(c) For Denisova, we do not use nucleotides within 1 nucleotide of either end of the read.
(d) For both Neandertals and Denisova, we do not use nucleotides with sequence quality <40.
(e) For present-day humans, we do not use nucleotides with sequence quality $<T_{ij}$, where T_{ij} is a threshold chosen such that half of nucleotides generated from individual i and of allele class j {j = A, C, G, T} are less than this value. For nucleotides that have exactly a quality score of T_{ij}, we randomly choose ones to eliminate such that exactly 5% are dropped (note that this differs from the 50% used in Reich et al. 2010). The cutoffs used are presented in Table 1.
(f) For the "Papuan1" individual from ref. 9 (HGDP00542), the sequencer had a high error rate at position 34 (41 on the reverse strand). We excluded data from position 34 for this individual.

Table 2: Datafiles summarizing the SNP ascertainment for the population genetics array

File name	Readme	Description	Entries
rawsnps	rawsnps_readme	This file contains all sites where there are at least 2 copies of a derived allele and 1 copy of the ancestral allele in 12 present-day humans, 3 Neandertals, and Denisova, and further filtered to be candidates for inclusion in the SNP array.	2, 173,116
ascertained	ascertained_readme	This file contains all SNPs chosen in any ascertainment panel (there are a few hundred that are triallelic and we list them on different lines, so the number of unique SNPs is 1,812,990).	1,813,579
screening	screening_readme	This file contains all probesets we considered for screening array design, as well as the metrics for prioritization and indicator variables indicating whether they were chosen. If chosen, a column indicates the genotyping outcome, and whether the SNP was taken forward to the production array.	3,882,158

Note: These files can be found in the Harvard Medical School orchestra filesystem at /groups/reich/CLEAN_SNP_ARRAY/.

1,353,671 SNPs for testing on an Affymetrix Axiom™ screening array

1,812,990 candidate SNPs discovered in 13 different ascertainment panels

We used the following algorithm to choose candidate SNPs for validating on the array.

(a) We mapped all reads used for SNP discovery to the chimpanzee reference sequence, *PanTro2*, without using data from the human reference sequence at all for read mapping. This was important to avoid biases due to the ancestry of the human reference sequence.

(b) We rediscovered all SNPs *de novo*, blinding ourselves to any prior information about whether the sites were polymorphic in present-day humans.

(c) At all SNPs, we required coverage from at least 1 Neandertal read and at least 1 Denisova read. This is expected to result in bias toward locations of the genome where the ancient DNA tends to be better preserved or the sequencing technology tends to work better. However, there is no reason why it would be expected to result in a bias in allele frequencies toward one

modern human population more than another (as all Neandertal and Denisova reads are mapped to chimpanzee, and no modern human data influences the mapping). The availability of data from archaic hominins from each of the SNPs on our array should be of value for some types of population genetic analysis. (For a handful of sites, the Denisova and Neandertal alleles may not be the same as those seen in present-day humans, but we nevertheless considered these sites to be covered by Denisova and Neandertal as we were concerned that not doing so could introduce bias. Users can treat such sites how they wish.)

(d) All A/T and C/G polymorphisms were excluded, since genotyping these SNPs requires twice the number of probes using the Axiom™ technology. Thus, removing them increases the number of SNPs we can include on a single array. Removing these SNPs has the additional benefit that it eliminates any strand ambiguity. (Illumina arrays do not genotype A/T or C/G SNPs, either.) However, it also had the disadvantage that A/T and C/G SNPS constitute the one class of SNPs that is believed to be immune to biased gene conversion. Thus, in population genetic analyses of the data generated from the array, it will be important to assess whether inferences are potentially explained by biased gene conversion.

(e) For the SNPs for panels 1-12 (candidate heterozygotes in an individual of known ancestry), we required the observation of at least 2 copies of the derived (non-chimpanzee) and at least 1 copy of the ancestral allele in the studied person (Reich et al. 2010; SI 6). We did not include chromosome X SNPs from these panels as the 12 individuals were all male.

(f) For the SNPs in panel 13 (derived in San relative to Denisova), we restricted to sites where we had ≥3-fold read coverage of San and ≥1-fold read coverage of Denisova.

A complication in choosing SNPs discovered in two individuals is that both the San and Denisova individuals are diploid. What we want is to have a panel of SNPs ascertained by comparing a single haploid Denisovan and a single haploid San chromosome, but if we are not careful, we are going to be biased toward the SNPs that are fixed differences. For example, if we accepted only SNPs where all Denisova reads matched chimpanzee and all San reads were derived, then we would bias against SNPs that were truly heterozygous.

To obtain data of the type that would be expected from sampling a single haploid Denisovan and a single haploid San chromosome, we picked the allele that was seen more often in each sample to represent that sample (if there was a tie in terms of the number of reads supporting each allele, we chose one allele at random). In this way, we are picking one of the two chromosomes from each individual (at random), and hence we are effectively sampling a haploid chromosome despite having diploid data. An additional benefit of using the majority rule is that we are also increasing the quality and reliability of the allele call, such that we expect a larger proportion of these SNPs to be real than in panels 1-12.

From the SNPs discovered in this way, we restrict our analysis to sites where Denisova matches the chimpanzee allele and where San is derived (we throw away sites where San is ancestral and Denisova is derived). The reason for this is that this is the only subset of SNPs that we can experimentally validate. To validate these SNPs, we can genotype the San individual and require the observation of an allele that differs from chimpanzee. In contrast, we cannot validate sites where San is ancestral and Denisova is derived, since the Denisova sample is extremely limited and does not provide enough for genotyping assays.

Some of the SNPs from panels 1-13 overlap. Thus, while the sum of the number of SNPs in each panel is 2,581,282, the number of unique SNPs is only 1,812,990. However, the fact that a SNP is

present in more than one panel does not mean that it has a higher likelihood of being validated for the array for a given ascertainment strategy. For SNP identified in more than one panel, we designed a single probe to test the SNP, but we assessed its validation status separately for each panel to avoid bias toward more easily validating more polymorphic SNPs (see below).

The perl script used for choosing SNPs is on the "orchestra" Harvard Medical School filesystem at: /groups/reich/CLEAN_SNP_ARRAY/newformat_affypick.pl (available on request from David Reich). The output file is at /groups/reich/CLEAN_SNP_ARRAY/ascertained (available on request from David Reich). This list contains a single entry for each unique SNP, with the exception of triallelic sites that have multiple designs (thus, there are 1,813,579 entries rather than 1,812,990). A readme file is at /groups/reich/CLEAN_SNP_ARRAY/ascertained_readme (available on request from David Reich) (Table 2). The number of SNPs that we selected using each strategy is summarized in Table 3.

Table 3: Ascertainment of SNPs for panels 1-13

Panel no.	Ascertain-ment	Sample ID	Genomic coverage	# SNPs found	# SNPs placed on screening array	# SNPs that validate on screening array	# SNPs that validate on final array
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	Den-HGDP01029	-	418,841	308,210	166,422	151,435
			Unique SNPs	*1,812,990*	*1,354,003*	*599,175*	*542,399*
			Unique probe designs	*1,941,079*	*1,385,672*	*605,069*	*546,581*

1,941,079 unique flanking sequences corresponding to the 1,812,990 unique SNPs

To ensure clean SNP ascertainment, we followed a rigorous procedure whereby the flanking sequence assay for each SNP were chosen only based on sequencing data from chimpanzee and the modern human sample used in SNP ascertainment. Thus, while some SNPs were discovered in multiple panels, we did not use this information in probe design. We used the simple rules below to pick a probe, and if the optimal design was different depending on the sample in which the SNP was ascertained, we used more than one probe for the SNP.

For each SNP in each of the 13 ascertainment panels, we specified 71 base pair (bp) flanking sequences that would be used for probe designing as follows:

(a) *Ancestral and derived allele are specified based on the individuals used in SNP ascertainment.* For each SNP in each panel, we specified the ancestral and derived alleles based on the two

alleles observed in SNP ascertainment, defining as "ancestral" the allele that matched chimpanzee. SNPs within any ascertainment panel almost always had two observed alleles, since we filtered out sites with three or more. However, for SNPs that were discovered in multiple panels, we performed the specification of the ancestral and derived allele independently, and thus for a small fraction of sites, there was a different derived allele depending on the ascertainment panel (even if flanking sequence were sometimes identical).

(b) *Flanking sequence is specified entirely based on the modern sample used for SNP discovery.* For initial probe design, we provided 35 bp of flanking sequence on either side of the SNP. We started with 71 bp of sequence from the chimpanzee genome, *PanTro2*, centered on the SNP. To decrease the number of mismatches between the flanking sequence and any human that might be analyzed using the array, we "humanized" the flanking sequence based on the modern sample used for SNP discovery (importantly, only the discovery sample is used for the humanization of the sequence, and so the ancestry of other samples cannot bias results).

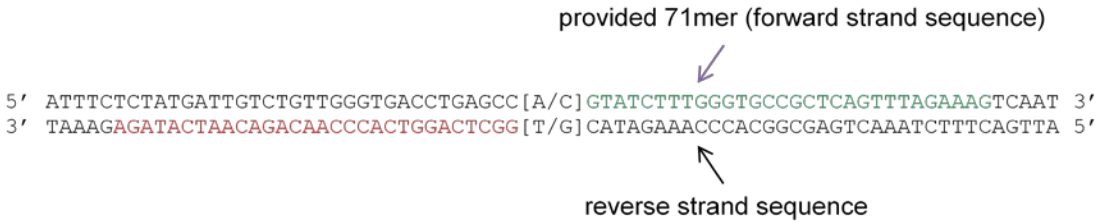
Specifically, for each of panels 1-13, we took all reads from the modern human used in SNP ascertainment that mapped to the flanking nucleotide. Where 100% of reads disagreed with *PanTro2*, we edited the flanking sequence to reflect that in the ascertainment sample. Otherwise, we kept the chimpanzee allele. An example is: "acctggctccagGgccagcagctccgtcaAggtcc[G/A]ctgcatgaaactgatgaaggggagggcaccaggcg". Here, capital [G/A] indicates the [chimp/alternate allele] at the SNP and other capital letters indicate bases edited from the chimpanzee reference to match the ascertainment sample. For ascertainment panel 13 (Denisova ancestral and a randomly chosen San allele derived), we did not use the Denisova genome in primer editing. Instead, we edited the sequence to match San whenever San consistently had a non-chimpanzee allele at all reads overlapping the site.

Because the steps above sometimes result in different flanking sequences for the same nucleotide (depending on the particular sequencing reads from the sample used in SNP ascertainment), we were left with more unique flanking sequences (n=1,941,079) than unique SNPs (n=1,812,991).

<u>Procedure used to choose 1,385,671 oligonucleotide probes for the screening array</u>

With the list of 1,951,079 flanking sequences, we needed to design oligonucleotide probes, or "probesets", for a screening array. We blinded ourselves to prior knowledge about which probes worked in previous assays using the Axiom™ technology, since doing so would expected to lead to a higher validation success rate for probes that have been previously tried on SNP arrays (introducing complex biases). For the same reason, we did not modify probe design based on using information in databases about polymorphism in flanking sequence. The only two types of information that were used in probe design were the physical chemistry considerations of which probes are expected to work well, and mapping information to the *PanTro2* chimp genome. All the metrics used are in a file on the "orchestra" Harvard Medical School filesystem /groups/reich/CLEAN_SNP_ARRAY/probesets, available on request from David Reich (Table 2). Details of the filtering procedure that we applied are as follows:

(a) We first identified 3,882,158 candidate probesets (two 30mers for each flanking sequence) For each of the 1,941,079 flanking sequences, it is possible to design two probesets corresponding to the 30 bp 5' or 3' direction of the SNP. We use the shorthand "red" to designate the 5' probe and "green" to designate the 3' probe, always referenced relative to the positive strand of the chimpanzee genome sequence *PanTro2* (Figure 1).



(b) We next restricted analysis to 2,294,760 probesets predicted to have greater success

Of the 3,882,158 candidate probesets (2 for each of 1,941,079 flanking sequences), we computed metrics that based on past experience were useful for predicting the success of genotyping. The values of the metrics are in /groups/reich/CLEAN_SNP_ARRAY/probesets (see probesets_readme), available on request from David Reich. We applied the following filters to winnow the list to 2,294,760:

(i) *Removing probesets that map to multiple positions in chimpanzee.*

(ii) *Best BLAT hit to PanTro2 is much better than the second-best hit.* We used BLAT to map each 35 bp flanking sequence to *PanTro2*. We required a minimum of 33 bp of alignment, and required the difference between the first and second hits to be >5.

(iii) *16mers within the probeset are relatively unique.* For each candidate 30 bp probeset, we examined each unique 16mer in a sliding window along the sequence (15 in all), and counted the number of exact matches in *PanTro2*. We defined "16mer-max" as the maximum number of exact matches seen for any of these 16-mers. In the experience of Affymetrix scientists who have worked on the Axiom™ technology, non-specific binding is unlikely when 16mer-max is small. We required "16mer-max" <110.

(iv) *No runs of 4 G's.* When more than 4 consecutive Gs stack up into quartets, hybridization tends to be compromised. We filtered out probes that had runs of 4 G's (or 4 C's),

(v) *Terminal 5mer is not complemented elsewhere in the probeset.* We required the 5' terminal 5mer to not have a reverse complement elsewhere in the probeset sequence, to minimize the tendency toward inter/intra probe annealing during hybridization, which in previous experience with the Axiom™ technology could cause a lower success rate.

(vi) *Number of G and C nucleotides is >5.* We required that >5 of the nucleotides were either G or C. Previous experience suggests that probesets with extremely low G or C usually do not work well for hybridization assays.

(c) A list of 1,477,155 probesets after eliminating redundancy

For flanking sequences where both candidate probesets passed the filters above, we chose the probeset that was deemed more likely to succeed based on having a lower value of "16mer-max" metric. When both probesets had the same value of "16mer-max", we used a random number generator to choose. This resulted in 1,525,604 candidate probesets.

Even after representing each flanking sequence by no more than one probeset, the resulting list contained 48,449 duplicative entries. This occurred when the same SNP (and probeset) had been independently selected in more than one of the 13 ascertainment panels. In such cases, the 71bp flanking sequence obtained as described above could be distinct for multiple SNP ascertainments, but sub-strings could be identical, so that it could happen that the 30mer that was selected to represent the SNP was identical. We therefore merged these probes to eliminate redundancy, leaving us with 1,477,155 unique probesets.

Our naming scheme for probesets contains a binary string of 13 characters providing the ascertainment information for that probe. Because we merged some probesets, we created a new ascertainment code called "asc.new". This was generated by applying a bitwise-or operation to the binary strings of 13 characters corresponding to the ascertainment information for the redundant probes.

(d) A final list of 1,385,672 probes that were placed on the screening array

The 1,477,155 probes that passed our filters were more than could fit into the screening array. Thus, we ranked all the probes based on their "16mer-max" score, breaking ties using a random number generator (lower values have a higher rank). After this ranking, all probes had "16mer-max" of no more than 110, and we were left with 1,385,672 probes.

Design, genotyping, and analysis of screening array

Design of the screening array

We designed two arrays to screen these 1.39 million probesets (0.69 million probesets fit onto a single screening array). To minimize bias, we randomized the probes with respect to which one of the 2 screening arrays was used to test them. We also used standard chip design strategies that are applied at Affymetrix for determining probe location in each screen design. The number of SNPs from each panel placed on the screening arrays is presented in Table 3.

The probesets used in the screening array are named like [chr]_[pos]_[alleles]_[asc.new]_[strand], with the 5 data fields indicating *PanTro2* chromosome / *PanTro2* physical position / ancestral-derived alleles, and the 13 bit binary string indicating the ascertainment panels in which the SNP was discovered, and the strand (f=forward or r=reverse compared to *PanTro2*).

Genotyping the screening array

Three 96-well plates of samples were genotyped on the 2 screening arrays in early 2011, with the goals of (a) deciding if each SNP passes quality control criteria and can be taken forward to the production array, and (b) generating useful data for preliminary population genetic analysis.

Validation plate #1: The goal of validation plate #1 was to genotype the same 12 modern human samples that were used in SNP discovery and in which the derived allele was observed, and to validate that we observe an allele at these samples that is distinct from the ancestral allele seen in primates. There was a high level of redundancy on the plate:

- Each of the 12 modern human samples was genotyped 6 times (six different wells)
- The chimpanzee and bonobo were each genotyped 6 times
- The gorilla and orangutan were each genotyped 4 times

The position of each sample on the plate (except for the upper right 4 wells which were left empty for control samples) was assigned using a random number generator.

Validation plates #2 and #3: We also took advantage of the screening array to genotype 2 plates of samples from CEPH-HGDP populations. We genotyped 184 samples from the same populations that were used in SNP discovery, consisting of French (n=28), Han (n=27), Papuan (n=17), San (n=6), Yoruba (n=21), Mbuti (n=13), Karitiana (n=13), Sardinian (n=28), Melanesian (n=11), Cambodian (n=10) and Mongola (n=10). Analysis of the data allowed us to perform further validation of the SNPs on the array, and also to assess whether useful population genetic analyses can be generated from these genotyping data.

Determining which SNPs "validated"

All samples were genotyped using the Axiom™ Assay 2.0 and genotype calls were made using the apt-probeset-genotype program in the Affymetrix Power Tools (APT) package[11] (the apt-probeset-genotype program is integrated in the Genotyping Console (GTC) version 4.1 software[12], which also provides visualization tools). Both programs use the Axiom™ GT1 algorithm to call genotypes. The algorithm adapts pre-positioned clusters to the data using a probability-based method. Clustering is carried out in two dimensions, log ratio ($\log_2(A) - \log_2(B)$) and size ($\log_2(A + B)/2$). The algorithm derives from BRLMM-P[13,14], which clusters in a single signal-contrast dimension, and is tuned to the signal characteristics of the Axiom™ assay.

To avoid ascertainment bias, only the sample used for SNP discovery, chimpanzees and bonobos, were used to assign a validation status to each candidate SNP for each of the 13 ascertainment panels. After an initial inspection of the data from Validation Plate #1, we chose not to use the data from the gorilla and orangutan as part of validation. This is because for a substantial fraction of SNPs, the signal intensities were different for one or both alleles in the apes than in humans, which we hypothesized was due to differences in the flanking DNA sequence under the primers. This occurred most often in gorilla and orangutan, and is expected to confound the genotyping algorithm, and thus we restricted to chimpanzees and bonobos.

We used a separate procedure for deciding whether a SNP was validated for ascertainment panels 1-12 (SNPs discovered as a heterozygote in a single modern human) or in ascertainment panel 13 (SNPs where San was derived and Denisova was ancestral). Table 4 summarizes the number of SNPs that validate in one, two, or all three genotyping runs.

Table 4: Results of genotyping on the screening array

Panel	Ascertainment	Sample ID	Screened SNPs	Validated in 3 runs	Validated in 2 runs	Validated in 1 run
1	French	HGDP00521	241,707	94,139	12,283	17,700
2	Han	HGDP00778	204,841	66,885	8,341	12,780
3	Papuan1	HGDP00542	232,408	43,622	5,308	8,000
4	San	HGDP01029	401,052	139,689	18,266	27,648
5	Yoruba	HGDP00927	302,413	103,670	13,542	20,017
6	Mbuti	HGDP00456	28,532	11,123	1,499	1,950
7	Karitiana	HGDP00998	8,535	2,839	326	511
8	Sardinian	HGDP00665	29,358	11,555	1,630	2,232
9	Melanesian	HGDP00491	36,392	13,626	1,769	2,527
10	Cambodian	HGDP00711	38,399	15,606	1,954	2,772
11	Mongolian	HGDP01224	24,858	9,890	1,312	1,824
12	Papuan2	HGDP00551	29,305	11,256	1,464	2,181
13	Denisova-San	Den-HGDP01029	308,210	107,708	26,280	32,845
		Unique probesets	*1,385,391*	*455,942*	*82,978*	*110,248*

Panels 1-12 (SNPs ascertained as a heterozygote in a single modern human)

We performed the ascertainment three times by carrying out three genotyping runs: once using only the 6 chimpanzee replicates to represent the apes, once using only the 6 bonobo replicate, and once using both chimpanzee and bonobo, a total of 12 *Pan* samples.

a) We required that all 6 human replicates are called heterozygous and all apes homozygous.

b) We required that the homozygous cluster and heterozygous cluster were well resolved in the clustering space, referred to as "A vs. M space". M and A are defined as

$$M = \left[log_2\left(A_{allele_{signal_{intensity}}}\right) - log_2\left(B_{allele_{signal_{intensity}}}\right)\right]$$
$$A = \left[log_2\left(A_{allele_{signal_{intensity}}}\right) + log_2\left(B_{allele_{signal_{intensity}}}\right)\right]/2$$

Based on the experience of Affymetrix scientists with the Axiom™ 2.0 Assay, five conditions were required to be satisfied to ensure that the clusters were well resolved in clustering space. Using the definitions "hetero"=samples called heterozygous, "homo"=samples called homozygous, "std"=standard error, and "abs"=absolute value, the 5 conditions that we required to be met to consider a SNP as validated were:

(i) $mean(M_{hetero}) \in (-1,1)$ and $mean(M_{homo}) \in (-\infty,-1]\ or\ [1,+\infty)$
(ii) $mean(A_{hetero}) - 2 \times std(A_{hetero}) > mean(A_{homo}) - 2 \times std(A_{homo})$
(iii) $mean(A_{hetero}) \geq 8.5$
(iv) $\Delta_{sep} \geq 5$, where Δ_{sep} is computed using the following formula

$$\Delta_{sep} = abs\left(\frac{mean(M_{homo}) - mean(M_{hetero})}{[std(M_{homo}) + std(M_{hetero})]/2}\right)$$

(v) $abs\left(mean(M_{hetero}) - mean(M_{homo})\right) > 1$

c) We required that the chimpanzee and bonobo agree at least partially in their genotype calls, for SNPs where a call was made in at least one of the three genotyping runs. The goal was to exclude SNPs that completely disagreed between chimpanzees and bonobos, which would imply that the ancestral allele determination was unreliable at these sites.

Panel 13 (SNPs where San was derived and Denisova was ancestral)
SNPs were considered as "validated" for panel 13 if they passed the following validation criteria:

a) All six San replicates were called heterozygote or derived homozygotes, and all ape replicates were called ancestral homozygotes.
b) SNPs in chromosome X were not in pseudoautosomal regions (PARs) and were called as homozygous derived in the San individual.
   (i) PARs were determined by converting coordinates of the human PARs (Build36) to *PanTro2* using the liftOver program from the UCSC genome browser.
   (ii) The San sample is a male, so SNPs in this chromosome are expected to be homozygotes.
c) The following three criteria were required to be met to make sure that the clusters were located around expected locations and well separated (that is, they were well resolved)
   (i) $mean\left(M_{ape_{homo}}\right) \in (-\infty,-1]\ or\ [1,+\infty)$
   (ii) $mean\left(A_{ape\_homo}\right) \geq 9.5$
   (iii) $std\left(M_{ape\_homo}\right) < 0.45$
d) For a SNP passing the above criteria in any one of three genotyping runs, we required that the chimpanzee and bonobo genotypes, compared across runs, did not completely disagree.

For autosomal SNPs in Panel 13, the true genotype for San replicates could be either heterozygote or derived homozygote. To avoid potential bias that might cause either heterozygous or derived

homozygous genotypes to be validated at a higher rate, we did not apply any metrics involving measuring the coherence of the heterozygous or derived homozygous clusters. Thus, the criteria used for Panel 13 are looser than the other 12 panels, which we expect will minimize the potential for ascertainment bias at the cost of lowering the validation rate of SNPs.

Filtering of SNPs based on the genotyping of 184 samples on Validation Plates #2 and #3
Up to this point, all decisions about which SNPs were considered to be validated were based entirely on the results of genotyping Validation Plate #1 on the screening array. As these decisions were only based on data from apes and the human sample used in SNP discovery, this is a perfectly clean strategy from the point of view of SNP ascertainment.

In practice on inspection of the genotyping results for Validation Plates #2 and #3, we found that a small fraction of SNPs that passed the validation filters described above were completely heterozygous in modern humans, or nearly so. This is unexpected based on population genetic considerations, and suggests that these SNPs overlap segmental duplications (which we did not screen out from our array in the interests of having a completely unbiased ascertainment procedure). An observation of more than half of individuals being heterozygous is unexpected at a true SNP. In an unstructured population for a SNP of frequency p, the expected proportion of heterozygous genotypes is $2p(1-p)$, which is at most 0.5, and the expected rate of heterozygous genotypes is less than this for a structured population.

We therefore implemented a further filter where for each SNP, we computed its frequency across all of the N modern humans on Validation Plates #2 and #3 that successfully genotyped ($N \geq 184$). We then counted the observed number of heterozygous genotypes het_{obs} versus the conservative expectation of $het_{exp} = Np_{het}$, where $p_{het} = 2p(1-p)$ (here, p is the empirical frequency of the derived allele, (het_{obs}+2(number of homozygous genotypes)/2N)). By dividing the difference between the observed and the expected number of heterozygous genotypes by the binomially distributed standard error, we can compute an approximately normally distributed Z-score:

$$Z = \frac{het_{obs} - het_{exp}}{\sqrt{Np_{het}(1 - p_{het})}}$$

We filtered out SNPs for which $Z > 5$, which is expected to remove at most a fraction 3.0×10^{-7} of true SNPs by chance. This removed 1,932 additional SNPs.

Summary of results of the validation genotyping
A total of 605,069 unique probesets (599,175 unique SNPs) were validated by the screen. The numbers of validated SNPs in each panel is listed in Table 3.

Taking forward SNPs to a final production array
All of the 605,069 probesets that passed the validation criteria after genotyping on the screening array were tiled on the final production array. In addition to those 605,069 "Human Origins" SNPs, a set of 87,044 "Compatibility" SNPs were also tiled on the final production array, choosing from a set of 8.8 million SNPs that had previously been validated using the Axiom 2.0^{TM} genotyping assay. Among those SNPs, there are 2,091 non-PAR chromosome Y SNPs, 259 mitochondrial SNPs, and 84,694 SNPs that overlap between the Affymetrix SNP Array 6.0 and Illumina 650Y array. No A/T or C/G SNPs were selected for the Compatibility SNPs, as they take up more space on the array (two probes for each SNP), so that excluding them thus allowed us to

maximize information from the array. For the 84,694 nuclear SNPs, we increased the value of the SNPs by maximizing the fraction that were also genotyped on the Affymetrix SNP Array 5.0 (78.5%), that were covered by sequencing from Neandertal (53.9%) and Denisova (64.7%), and for which a chimpanzee allele was available (nearly 100%).

Validation of the final SNP array through genotyping of 952 CEPH-HGDP samples
We attempted to genotype 952 CEPH-HGDP samples that were previously determined to be unrelated up to second degree relatives[15]. This genotyping had three goals:

(a) Round 2 validation: Evaluating the performance of every SNP in the final product array
Although all of the SNPs that were tiled on the final product array had previously been validated in screening arrays, there is variability in how an assay performs on a real product. Hence after manufacturing the final SNP array, we genotyped 952 unrelated CEPH-HGDP samples (including the 12 modern human samples used in SNP ascertainment) using the final product array. We used these data to create a list of SNPs that had gone through two rounds of validation and would be robust for genotyping.

(b) Building up prior distributions for SNP calling
The Axiom™ GT1 algorithm makes more accurate genotype calling for a SNP if it has prior distributions for the 3 genotype clusters (AA, AB, and BB) based on data (by default, the Axiom™ GT1 algorithm uses the generic prior distributions of the 3 clusters, which is just a best guess). Because the CEPH-HGDP panel has such a large number of samples from diverse ancestries, we expect to observe clusters from all 3 genotypes for most SNPs. This allows us to construct prior distributions that could be used for SNP calling in other projects.

(c) *Creating a dataset that will be useful for population genetics*
The genotyping of the unrelated CEPH-HGDP samples has the benefit that it creates a dataset that will be widely available to the population genetics community. Users who wish to genotype samples that they are interested in on this array, will be able to merge the data that they collect with data collected on the CEPH-HGDP samples, to enable a richer comparison of genetic variation in one region to worldwide variation.

Table 5. Eighteen HGDP samples that did not pass quality control

Identifier	Population	Reason removed
HGDP00009	Brahui	Failed DQC
HGDP00708	Colombian	<97% genotype call rate
HGDP01266	Mozabite	<97% genotype call rate
HGDP01267	Mozabite	<97% genotype call rate
HGDP01403	Adygei	<97% genotype call rate
HGDP00885	Russian	<97% genotype call rate
HGDP00886	Russian	<97% genotype call rate
HGDP00795	Orcadian	<97% genotype call rate
HGDP00804	Orcadian	<97% genotype call rate
HGDP00746	Palestinian	<99% concordance with Illumina 650Y data
HGDP00326	Kalash	<99% concordance with Illumina 650Y data
HGDP00274	Kalash	<99% concordance with Illumina 650Y data
HGDP00304	Kalash	<99% concordance with Illumina 650Y data
HGDP00309	Kalash	<99% concordance with Illumina 650Y data
HGDP01361	Basque	<99% concordance with Illumina 650Y data
HGDP00710	Colombian	<99% concordance with Illumina 650Y data
HGDP01376	Basque	<99% concordance with Illumina 650Y data
HGDP01009	Karitiana	anomalous ancestry relative to others in group

Filtering out 18 samples that did not genotype reliably
After assaying all 952 samples, we filtered to 934 samples as follows (Table 5):

(a) We filtered out 9 samples that did not pass standard Axiom™ 2.0 Array QC metrics: a "DQC" score (chip-level quality metric) and a call rate score. This suggests problems such as low input DNA amount, contamination of DNA samples, or technical issues with hybridization. These 9 samples were excluded from the genotyping calling.

(b) We excluded an additional 9 samples based on their genotype patterns. Of these, 8 were excluded because there was a greater than 1% genotype discrepancy between our current data and earlier data from the Illumina 650Y array genotyped on the same samples**Error! Bookmark not defined.**. We also excluded HGDP01009, an individual that our data (as well as analyses of previous datasets) suggests is a sample whose ancestry is an outlier relative to others from the Karitiana group, suggesting a history of recent gene flow with other Native American populations.

Special filters applied to chromosome X and Y data
Chromosome X occurs in only a single copy in men but in two copies in women. Chromosome Y occurs only in men. This means that SNPs on these chromosomes need to be treated differently from autosomal SNPs; for chromosome X we genotyped males and females separately, and for chromosome Y we only genotyped males. For males, we required that genotypes on both chromosome X and Y were always homozygous.

Filtering out additional probesets based on the genotyping of the final array
Not all probesets tiled onto the final array performed well enough to produce reliable results. We filtered out a total of 58,488 additional probesets by sequentially applying the seven criteria listed in Table 6. Three of the criteria used in Table 6 require more detailed explanation.

Table 6. Phase 2 validation (determining probesets for which we report genotypes)

Order	Filter	Removed	Definition
1	SNP call rate ≥ 95%	23,476	(no. of called samples) / (no. of genotyped samples = 943)
2	Concordance	31,415	For panels 1-12, the SNP must be heterozygous in the sample used in ascertainment (for panel 13, heterozygous or derived homozygous).
3	het_rate > 5	79	This is the same metric used in SNP validation
4	het_offset > -0.5	892	See below for explanation
5	resolution score ≥ 3.6	2,450	See below for explanation
6	chrX annotation	94	Panel 13 SNPs that are *PanTro2* chrX but not *hg18* chrX are removed.
7	chrX SNPs separate males and females	82	See below for explanation
	Total removed by all filters	*58,488*	

<u>*het_offset:*</u> Using the definition of "A vs. M space" described in the discussion of the screening array filters, we defined a quantity called *het_offset* that measures whether the heterozygous genotype is appropriately intermediate between the homozygous clusters. For a probeset with three observed genotype clusters (AA, AB, and BB), it is defined as

$$het\_offset: mean(M_{AB}) - \frac{mean(M_{AA})+mean(M_{BB})}{2}$$

For a probeset with one observed homozygous and one heterozygous cluster, it is defined as:

$$het\_offset: mean(M_{AB}) - mean(M_{AA|BB})$$

For other situations, *het_offset* is not used as a filter.

resolution score: This is again defined in the M space of the "A vs M space", and it measures how well the heterozygous cluster separates from the homozygous cluster(s). We define:

$$resolution = \frac{abs(mean(M_{homo}) - mean(M_{hetero}))}{sd(M_{homo}) + sd(M_{hetero})} \times 2$$

For a probeset with three observed genotype clusters (AA, AB, and BB), the resolution score is defined as: $\min(resolution_{AA-AB}, resolution_{BB-AB})$. For a probeset with one observed homozygous cluster and one observed heterozygous cluster, the resolution score is the resolution between two clusters. For other situations, the resolution score is NA.

chromosome X SNPs separate males and females: It was found that for some chromosome X SNPs, female samples and male samples formed distinct genotype clusters. Such cases most likely are not real chromosome X SNPs. One possible explanation for this pattern is SNPs derived from fixed differences between homologous chromosome X and chromosome Y sequences[15,16]. We removed chromosome X SNPs that meet all of the following criteria

1. All called male samples have the same genotype call
2. Greater than 85% of called female samples have the same genotype call and there are at most 2 different called genotypes for females
3. The distance between the male genotype cluster center and the major female genotype cluster center is at least 0.8 units in the M genotype clustering space.

The number of final validated SNPs is given in the final column of Table 7, and this is the set of SNPs for which we publically released data for 934 unrelated CEPH-HGDP samples on August 12, 2011. Table 7 summarizes the SNPs on the final product array.

Table 7. Summary of SNPs in the final array

Category	**number of probesets**	**number of SNPs**
Human Origins	546,581	542,399
Chromosome Y	2,091	2,091
Mitochondrial DNA	259	259
Compatibility	84,694	84,694
Total	*633,625*	*629,443*

Upon commercial release of the array, Affymetrix is planning to release user-friendly software that will facilitate SNP calling using each of the ascertainment panels. Users who are interested in any particular ascertainment will open up one of 14 available folders of files (the first 13 corresponding to the SNPs in each ascertainment, and the 14th corresponding to all SNPs together). Users will then be able to use that folder (which will include ascertainment-panel specific priors) to call genotypes relevant to any given ascertainment panel.

The genotyping data on the 934 unrelated CEPH-HGDP samples that we collected as part of this project has been made freely available without restriction to the community by depositing the data into the CEPH-HGDP database on August 12, 2011 (ftp://ftp.cephb.fr/hgdp_supp10/). There are no restrictions on using these data and publishing papers based on these data.

In addition to the dataset of 934 CEPH-HGDP samples that we released on August 12, 2011, we have also carried out further filtering to create a dataset of 828 samples that might be more useful for some population genetic analyses. This dataset, which is the one that we used for the analyses of population history reported in the present paper, is available for downloading from the Reich laboratory website (http://genetics.med.harvard.edu/reich/Reich_Lab/Welcome.html). To generate this dataset, we started with the dataset that was released to the CEPH-HGDP website on August 12, 2011, and then carried out population-specific Principal Component Analysis to identify individual samples that are outliers with respect to their own populations (consistent with admixture with other populations without the last few generation). These individuals were then filtered out of the dataset, allowing us to analyze data from a homogeneous population sample. Table 8 lists the number of samples from each population before and after the filtering.

Table 8. Number of CEPH-HGDP samples in each of the two datasets reported here

Population	Region	Aug. 12 2011	Further filtering
BantuKenya	Africa	11	10
BantuSouthAfrica	Africa	8	6
BiakaPygmy	Africa	23	20
Mandenka	Africa	22	20
Mbuti*	Africa	13	12
Mozabite	Africa	27	25
San*	Africa	6	5
Yoruba*	Africa	22	22
Cambodian*	East Asia	10	10
Dai	East Asia	10	10
Daur	East Asia	9	7
Han*	East Asia	34	33
Han-NChina	East Asia	10	10
Hezhen	East Asia	9	9
Japanese	East Asia	29	28
Lahu	East Asia	8	7
Miao	East Asia	10	10
Mongola*	East Asia	10	8
Naxi	East Asia	9	7
Oroqen	East Asia	9	8
She	East Asia	10	10
Tu	East Asia	10	9
Tujia	East Asia	10	9
Uygur	East Asia	10	9
Xibo	East Asia	9	7
Yakut	East Asia	25	23
Yi	East Asia	10	10
Adygei	West Eurasia	17	15
Basque	West Eurasia	22	20
Bedouin	West Eurasia	46	38
Druze	West Eurasia	42	32
French*	West Eurasia	28	27
Italian	West Eurasia	13	11
Orcadian	West Eurasia	13	13
Palestinian	West Eurasia	45	34
Russian	West Eurasia	23	22
Sardinian*	West Eurasia	28	27
Tuscan	West Eurasia	8	7
Balochi	South Asia	24	21
Brahui	South Asia	24	22
Burusho	South Asia	25	24
Hazara	South Asia	22	17
Kalash	South Asia	19	18
Makrani	South Asia	25	22
Pathan	South Asia	24	22
Sindhi	South Asia	24	22
Colombian	America	5	4
Karitiana*	America	13	8
Maya	America	21	18
Pima	America	14	11
Surui	America	8	6
Melanesian*	Oceania	11	9
Papuan*	Oceania	17	14

* Indicates a population used in SNP ascertainment. Analysis of data from these populations should remove the individual used in SNP discovery, as they have highly biased SNP genotypes (all heterozygotes) relative to others in the same group.

References

[1] Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1,e70.

[2] Wang S et al. (2007) Genetic variation and population structure in native American. PLoS Genet. 3, e185.

[3] Tishkoff SA et al. (2009) The genetic structure and history of Africans and African Americans. Science. 324, 1035-44.

[4] Friedlaender JS et al. (2008) The genetic structure of Pacific Islanders. PLoS Genet. 4, e19.

[5] Rosenberg NA et al. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet. 2, e215.

[6] Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than Europeans. Nature Genetics 39, 1251-1255

[7] Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M & Pääbo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053-1060.

[8] Li JZ et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-4.

[9] Green RE et al. (2010) A draft sequence of the Neandertal genome. Science 328, 710-722.

[10] Cann HM et al. (2002) A human genome diversity cell line panel. Science 296, 261-2.

[11] Affymetrix Power Tools: http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx

[12] Genotyping Console: http://www.affymetrix.com/browse/level_seven_software_products_only.jsp?productId=131535&categoryId=35625#1_1

[13] Single-sample analysis methodology for the DMET™ Plus Premier Pack - this white paper also applies to the Axiom GT1 genotyping algorithm. http://www.affymetrix.com/support/technical/whitepapers/dmet_plus_algorithm_whitepaperv1.pdf (pdf, 292 KB)

[14] BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf (pdf, 163 KB)

[15] Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70, 841-7.

[16] Ross MT et al. (2005) The DNA sequence of the human X chromosome. Nature, 434,325-337