Published in final edited form as:

Genomics. 2005 August ; 86(2): 117–126.

High-density single-nucleotide polymorphism maps of the human genome

Raymond D. Miller^{a,1}, Michael S. Phillips^{b,1}, Inho Jo^{c,1}, Miriam A. Donaldson^b, Joel F. Studebaker^b, Nicholas Addleman^{a,2}, Steven V. Alfisi^b, Wendy M. Ankener^b, Hamid A. Bhatti^b, Chad E. Callahan^b, Benjamin J. Carey^b, Cheryl L. Conley^b, Justin M. Cyr^b, Vram Derohannessian^b, Rachel A. Donaldson^a, Carolina Elosua^b, Stacey E. Ford^b, Angela M. Forman^b, Craig A. Gelfand^b, Nicole M. Grecco^b, Susan M. Gutendorf^b, Cricket R. Hock^b, Mark J. Hozza^b, Soyoung Hur^c, Sun Mi In^d, Diana L. Jackson^b, Sangmee Ahn Jo^c, Sung-Chul Jung^{c,3}, Sook Kim^d, Kuchan Kimm^e, Ellen F. Kloss^a, Daniel C. Koboldt^a, Jennifer M. Kuebler^b, Feng-Shen Kuo^b, Jessica A. Lathrop^b, Jong-Keuk Lee^e, Kathy L. Leis^b, Stephanie A. Livingston^b, Elizabeth G. Lovins^a, Maria L. Lundy^b, Sima Maggan^b, Matthew Minton^a, Michael A. Mockler^b, David W. Morris^b, Eric P. Nachtman^b, Bermseok Oh^e, Chan Park^e, Chang-Wook Park^d, Nicholas Pavelka^a, Adrienne B. Perkins^b, Stephanie L. Restine^b, Ravi Sachidanandam^f, Andrew J. Reinhardt^a, Kathryn E. Scott^b, Gira J. Shah^b, Jatana M. Tate^b, Shobha A. Varde^b, Amy Walters^b, J. Rebecca White^b, Yeon-Kyeong Yoo^d, Jong-Eun Lee^{d,*}, Michael T. Boyce-Jacino^{b,*}, and Pui-Yan Kwok^{a,*},² The SNP Consortium Allele Frequency Project

a Washington University, Division of Dermatology, St. Louis, MO, USA

b Orchid BioSciences, Inc., Princeton, NJ, USA

c Department of Biomedical Sciences, National Institute of Health, Seoul, South Korea

d DNA Link, Inc., Seoul, South Korea

e National Genome Research Institute, National Institute of Health, Seoul, South Korea

f Cold Spring Harbor Laboratories, NY, USA

Abstract

Here we report a large, extensively characterized set of single-nucleotide polymorphisms (SNPs) covering the human genome. We determined the allele frequencies of 55,018 SNPs in African Americans, Asians (Japanese–Chinese), and European Americans as part of The SNP Consortium's Allele Frequency Project. A subset of 8333 SNPs was also characterized in Koreans. Because these SNPs were ascertained in the same way, the data set is particularly useful for modeling. Our results document that much genetic variation is shared among populations. For autosomes, some 44% of these SNPs have a minor allele frequency $\geq 10\%$ in each population, and the average allele frequency differences between populations with different continental origins are less than 19%. However, the

*Corresponding authors. Jong-Eun Lee is to be contacted at fax: +82 2 364 4778. Michael T. Boyce-Jacino, fax: +1 609 818 0054. Pui-Yan Kwok, fax: +1 415 476 2956. E-mail addresses: jonglee@dnalink.com (J.-E. Lee), mbj@bionanomatrix.com (M.T. Boyce-Jacino), kwok@cvmil.ucsf.edu (P.-Y. Kwok).

¹These authors contributed equally to this paper.

²Current address: University of California, San Francisco, Cardiovascular Research Institute, San Francisco, CA, USA.

³Current address: Department of Biochemistry, College of Medicine, Ewha Womans University, Seoul, South Korea.

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.ygeno.2005.04.012.

several percentage point allele frequency differences among the closely related Korean, Japanese, and Chinese populations suggest caution in using mixtures of well-established populations for case-control genetic studies of complex traits. We estimate that ~7% of these SNPs are private SNPs with minor allele frequencies <1%. A useful set of characterized SNPs with large allele frequency differences between populations (>60%) can be used for admixture studies. High-density maps of high-quality, characterized SNPs produced by this project are freely available.

Keywords

SNP; Human variation; The SNP Consortium; Pooled sequencing; Single-base primer extension; Korean population; Complex disease variation search

Since genetic variation plays an important role in many diseases, a major focus of the human genome project has been to identify a large number of uniquely mapped single-nucleotide polymorphisms (SNPs) to serve as tools in genetic studies of complex traits. To date, 10.1 million human reference SNPs have been deposited into the public database dbSNP (build 123, <http://www.ncbi.nlm.nih.gov/SNP/>) [1]. This immense data set provides a framework map of SNP markers that can be exploited for the mapping of genetic factors relevant in complex disease using whole-genome association studies [2–4], for the assembly of dense local SNP maps required in positional cloning projects [5,6], for admixture studies that take advantage of SNPs with large allele frequency differences between populations [7], and for genotyping projects including the International HapMap [8]. To facilitate these studies, however, the SNPs must be characterized in multiple individuals and populations to determine their utility.

The SNPs found in the public domain have been identified by comparing homologous DNA sequences derived from different chromosomes. Two major methods of DNA comparison were utilized in the SNP discovery process: (1) variants identified by the comparison of genomic sequence derived from overlapping bacterial artificial chromosome (BAC) sequences and (2) variants identified by the comparison of “shotgun” genomic sequences overlaid on the “working draft” sequence of the human genome [1]. The SNP Consortium (TSC; <http://snp.cshl.org/>), a coalition of companies and academic institutions and the British charity the Wellcome Trust, was founded for the purpose of advancing SNP research and preventing the privatization of SNP sequences [9].

Analysis of TSC sequence data in the Discovery Resource uncovered varying degrees of heterozygosity, a measure of nucleotide diversity among chromosomes. A striking feature of these estimates was the difference between autosomes (7.6×10^{-4} , one variant every 1300 bp), the X chromosome (4.7×10^{-4}), and the Y chromosome (1.5×10^{-4}). The observed reduced diversities for the X and Y chromosomes compared with autosomes can be best explained by a reduced effective population size for the X and Y chromosomes and an altered proportion of time spent in males, who have a higher mutation rate [1].

Only a limited amount of information about the characterization of a large number of SNPs is found in the literature. Data from a previous study, based on pooled DNA sequencing of a European-derived panel, found that there was a common SNP (minor allele frequency (MAF) $\geq 20\%$) about every 1100 bp on the X chromosome [10]. Surprisingly, the incidence of SNPs was not uniform and long regions without SNPs, called SNP deserts, were found to be largely devoid of common SNPs [10–12]. Other groups have observed that the prevalence of common SNPs within coding sequences was somewhat lower than for adjacent sequences [13–15] and that there was autocorrelation in the local incidence of SNPs [16]. A study of SNPs found by TSC and overlap BAC sequence comparison methods showed that about 76% of SNPs were common SNPs in one or more populations, and about 27% were common SNPs in all three

populations studied [17]. Results from the studies reported here permit a much more extensive view of human genetic variation.

In this paper, we report the results from three large studies by Orchid BioSciences (Princeton, NJ, USA) (Orchid), Washington University (St. Louis, MO, USA) (WU), and the Korean Team (Korea), with members from the Korean National Institute of Health and DNA Link, Inc. Orchid and WU were participants in The SNP Consortium Allele Frequency Project. Using different approaches, the three groups determined the allele frequencies of over 55,000 candidate SNPs in three populations and over 8000 candidate SNPs in the Korean population. The Orchid study analyzed 33,488 SNPs by determining ~4.2 million genotypes using its proprietary single-base extension genotyping technology (SNP-IT), and the Korean study used the same genotyping technology and assay design to type candidate SNPs in the Korean population. The WU study analyzed 21,530 SNPs and estimated the allele frequencies by amplifying and sequencing pooled DNA samples and then analyzing allelic peaks in the sequencing traces. All groups used candidate SNPs that had a known and well-described ascertainment, either from the TSC SNP discovery project or from a comparison of BAC end overlaps from the genome project. In addition to providing an extensive resource of characterized SNPs, these studies provide a detailed view of genetic variation in humans.

Results

Samples

The primary goal of these studies was to characterize SNPs so that useful sets of them could be put together as tools for genetic studies. We anticipated some frequency differences between ethnic groups. The DNA sampling strategy was therefore chosen to maximize the usefulness of the data, given finite resources for genotyping. Three TSC allele frequency DNA panels were assembled, each comprising DNAs from 42 individuals and isolated from established cell lines maintained by the Coriell Institute (see Methods). Both the Orchid and the WU studies used these panels for SNP frequency estimation. Each panel represented a sample from a population of primarily different continental ancestry: African American, Asian (with parents identified as born in Japan or China), and European American, also called Caucasian. While these panels were derived from populations with different primary continental origins, some admixture was expected from other populations [18]. Inclusion of the Korean DNAs (from 43 individuals, see Methods) permitted identification of variation that might be of particular use to the Korean population and provided a comparison of the similarities and differences between regional populations represented by samples from Korea, China, and Japan.

SNPs characterized by two groups

To estimate efficiently the approximate allele frequency of a large number of SNPs in multiple populations, two different approaches were used: (1) genotyping of individuals by Orchid and Korea and (2) sequencing of pooled DNAs by WU. Some 1250 uniquely mapped TSC SNPs were independently characterized by Orchid with genotyping and by WU using sequencing of pooled DNAs, providing the opportunity to compare results. For each population, the correlation between frequencies estimated by the two methods was high (0.82; $p < 0.0001$; Supplementary Fig. A). However, because sequencing of pooled DNA samples is optimized for estimating allele frequencies of common SNPs, the results from the two approaches diverge significantly for SNPs with low MAF. In some cases, SNPs with MAF $< 5\%$ were called as “monomorphic” by the sequencing approach. Nonetheless, the similar frequency estimates by both studies for most SNPs, and the similar distribution of results found by each, validated both approaches and enabled the production of a combined genome-wide SNP map consisting of common SNPs.

Allele frequency distribution within each population

For each of the populations, more than 70% of the SNPs were polymorphic (with MAF >1%), and more than 55 and 45% of SNPs had MAF ≥ 10 and $\geq 20\%$, respectively (Supplementary Table A). Thus, from this large study of SNPs, it was very possible to assemble high-density SNP maps for genetic studies in particular populations (Supplementary Table B).

We analyzed the variation within each population using only the data from the Orchid study (for highest precision) including 20,574 SNPs from the autosomes and 446 from the X chromosome (Fig. 1). For SNPs from autosomes for each population, as found by others [2], the fraction of those in the first bin (0–10% MAF) was elevated compared with other bins, and the fraction of those in other bins is relatively uniform (Fig. 1a). For X chromosome SNPs, the pattern is similar to that found for SNPs on autosomes except the fraction of SNPs with a MAF in the first bin was even larger (Fig. 1b).

The identification of a population-specific SNP (variant allele found only in one of the populations studied) is a function of the number of samples used for characterization in the other populations (42 individuals in this case). With that caveat, we were able to identify thousands of apparent population-specific SNPs in this study. They occurred in fairly low proportions but with a striking pattern. Asian and European American populations had similar distributions of population-specific SNPs, each with 1.0% of total SNPs as population-specific, but the proportion of SNPs specific to the African Americans was 7.1% (Fig. 1c, plotted as 2% bins). As expected, a higher proportion of the SNPs in the 0–2% MAF bin are population-specific compared to other bins. For example, 34, 12, and 18% of the SNPs in the 0–2% MAF bin were population-specific for African Americans, Asians, and European Americans, respectively; whereas 17, 0.9, and 0.8% of the SNPs in the 10–12% bin were population-specific for these populations (data not shown).

Some 7.3% of SNPs were monomorphic in all populations, and for each of the populations, additional SNPs were also found to be monomorphic (Table 1).

Allele frequency variation between populations

By characterizing SNPs in several populations, we were able to identify a collection of common SNPs that could be assembled into genetic maps useful in any population. This was easily accomplished by using the many SNPs that were highly polymorphic in each of the three populations. For example, 44% of SNPs on autosomes had a MAF $\geq 10\%$ in each of the three populations (Fig. 2). We call these SNPs “common-link SNPs” and they are available for download from our Web site, <http://snp.wustl.edu/characterization> (Supplementary Table C). Although the proportion of common-link SNPs on the X chromosome was slightly lower than those on the autosomes, they still represent a significant resource (Fig. 2, Supplementary Table C). There is a linear relationship between the common-link SNPs as a fraction of total SNPs and the minimum MAF (Fig. 2), with over 99.7% of variation explained for both autosomes and the X chromosome by linear regression analysis.

For common-link SNPs with MAFs that are ≥ 10 , ≥ 20 , or $\geq 30\%$, the average spacing is 179, 297, or 747 kb, respectively. The largest gap between common-link SNPs is 20 Mb, and the number of gaps >1.5 Mb is 132, 253, and 583, respectively. When mapped within contigs (dbSNP build 105), the average spacing between SNPs with the minor allele frequencies ≥ 10 , ≥ 20 , and $\geq 30\%$ is 133, 209, and 436 kb, respectively. Within contigs, the largest gap is 6.4 Mb and the number of gaps >1.5 Mb is 24, 66, and 235, respectively. Many of these gaps represent missing sequences in genome assemblies or repeated sequences in the genome (Fig. 3). The blue sections that are indicated in the leftmost column for each of the chromosomes in Fig. 3

correspond to true gaps that exist in the current genome assembly. The largest gaps shown are indicative of heterochromatin present at the centromeres and in the p arms of chromosomes 13, 14, 15, 21, and 22. The sections marked as blue in the center and right columns, but not the left column, are candidate regions for SNP deserts in European Americans. Taking these factors into account, the SNPs in this study provide very good coverage for almost all regions of the sequenced genome.

In addition, we were able to identify SNPs that could be used for admixture mapping studies, i.e., those with markedly different allele frequencies between populations [7]. The distributions of the allele frequency differences between populations of different continental origins are shown for SNPs mapping to autosomes and the X chromosome (Figs. 4a and 4b). A major feature of these data is that although there is statistically significant divergence between all of the populations, the divergence is on average small. Also, SNPs mapping to the X chromosome showed somewhat greater divergence than those mapping to autosomes. For example, for SNPs on autosomes (or X in parentheses), >88.1 (80.2) and 97.8% (93.8%) had frequency differences <40 or <60% between any two populations, respectively. The weighted average of the divergences between African Americans, Asians, and European Americans for SNPs on autosomes is less than 19% (Table 2). Although very few SNPs had frequency differences $\geq 60\%$ between populations these SNPs are very useful for admixture studies (Supplementary Table D).

The curves of the distributions of frequency differences between populations with different continental origins drop sharply for frequency differences $\geq 55\%$ and for differences $\geq 80\%$ the curve is nearly at the zero base line (Fig. 4d). For SNPs scored by Orchid, $\leq 0.023\%$ had differences $\geq 80\%$. There may be no cases of a divergence between populations $\geq 90\%$ and any putative case should surely be independently confirmed.

Genotyping results were also analyzed for allele frequency differences among Chinese, Japanese, and Korean in pair-wise comparisons. Due to smaller sample sizes (20 chromosomes for Chinese and 64 for Japanese) variation in divergence due to sampling was greater and is shown (Fig. 4c, dotted lines). The differences among the Asian populations were small but significant. The Japanese–Korean comparison was the smallest (Fig. 4c, Table 2). For each of the three comparisons, at least 99.0% of SNPs have a divergence of less than 35% (Fig. 4d). For autosomes, the divergence between Chinese and Japanese is 46% of that between Asians and African Americans, and the divergence Japanese and Koreans is 31% of that between Asians and African Americans (Table 2).

Discussion

The high-density maps of characterized SNPs produced by these studies provide very effective tools for various strategies in the search for genetic variants causing disease. For example, the 3945 SNPs with $\geq 30\%$ MAF in all populations are useful for linkage analysis, the 36,202 SNPs with $\geq 10\%$ MAF in at least one population will be very useful in association studies, and the 1410 SNPs with allele frequency difference of $\geq 60\%$ between populations will be extremely useful in admixture mapping studies. The data have already been used to provide characterized SNPs for the International HapMap project [8].

This study provides a very extensive data set of human SNPs with a uniform ascertainment. The collection has been used to model recent human history and estimate fractions of the genome under selection between populations [19,20]. Sampling strategies for ascertainment of SNPs can be characterized as $S(n,k)$, where n is the number of chromosomes examined and k is the minimum number of chromosomes required to carry the minor allele before a site is

called a candidate SNP [21]. The vast majority of SNPs in this study were S(2,1) and a small number were S(3,1). The frequency distributions found (Figs. 1a and 1b) approximately fit S(2,1) predictions for an expanded population [21]. Given that population structure, the prediction for an S(50,1) strategy (e.g., resequencing of 25 individuals) is that about 80% of identified SNPs will be private SNPs. Large numbers of private SNPs have been found in a study resequencing many individuals [22].

A number of general patterns of human SNP variation are evident in this study and should be considered in studies of complex disease. Much, but not all, SNP variation is shared among the three populations with different continental origin. (1) With S(2,1) ascertainment, 44% of SNPs on autosomes are common-link SNPs with a minor allele frequency >10% in each population (Fig. 2). If common-link SNPs are used to construct collections for pedigree studies, the collections will be useful in many populations; conversely, if the SNPs are not common link, the utility of the collections will be more limited. (2) If common SNPs are ascertained in one population, in a second population, some will be common and some will not. (3) Some SNPs appear to be population-specific, particularly for African Americans, but most of these have very low MAF, making their practical utility as population-specific markers doubtful (Fig. 1c). (4) With S(2,1) ascertainment, 2.2% of SNPs on autosomes have frequency difference between populations of different continental origins of $\geq 60\%$ but almost none $\geq 80\%$. Mapping strategies based upon admixture should plan on appropriate differences to have sufficient markers. (5) On average, allele frequencies in populations of different continental origin differ by 16–19%, and in populations within a continent, such as Koreans and Japanese, they differ by several percent (Table 2). These differences are sufficiently large, even from populations within a continent, to cause substructure problems with association studies from two combined populations if the cases and controls are differentially sampled from the populations.

The evolutionary dynamics of SNPs on the X chromosome are clearly different from those on the autosomes; for example, the fraction of SNPs in the lowest minor allele frequency bin is increased (Figs. 1a and 1b), common-link SNPs are reduced (Fig. 2), and divergence between populations is greater (Figs. 4a and 4b). These patterns may be due to the smaller effective population size for the X chromosome compared with autosomes, speeding incorporation of new SNPs and divergence between populations.

Our observations confirm earlier reports that populations derived from Africa harbor a higher number of variations than those from Asia or Europe (e.g., [23]). One cause of this observation is the population-specific SNPs: African Americans have 7.1 times the incidence of population-specific SNPs as Asians or European Americans (Fig. 1c). The other cause is the patterns of SNPs found in two (not three) populations. We detected 12.3% monomorphic SNPs in African Americans (Table 1), consisting of 7.3% private SNPs, 2.0% SNPs that are population-specific in other populations, and 3.0% SNPs monomorphic in this population but not in the other two. For Asians and European Americans, the latter category was 11.0 or 4.1%, respectively. Since the Discovery Resource was used to identify these SNPs, there is no obvious bias as to population source. SNPs shared between populations are most likely to be found in African Americans, followed by European Americans, then Asians. Due to incidences of population-specific SNPs and patterns of allele sharing between populations likely because of population histories, African Americans have a slightly higher incidence of SNP variation than European Americans, and European Americans have a slightly higher incidence of SNP variation than Asians.

SNPs in which different alleles are monomorphic in two groups (diverged SNPs) have been important in evolution when the alternate alleles have functional consequences. An interesting example has been provided in the *FOXP2* gene, in which there are two diverged SNPs, each

causing coding changes in exon 7 of the gene in humans compared with chimpanzees, gorillas, and other primates [24]. The human allele has been shown to be required for speech [25]. We found for humans that the curves of the distributions of frequency differences between populations with different continental origins drop sharply to zero as frequency differences increase. These data support the following hypothesis: there are no diverged SNPs between the populations we examined. An interesting corollary to this hypothesis is that any differences in phenotypes among the populations caused by SNPs, including disease susceptibility, must be due to one or more polymorphic SNPs, not diverged SNPs.

Our studies have not only identified tens of thousands of SNPs assembled into genetic maps useful in a variety of mapping strategies such as linkage analysis, association studies, and admixture studies, but they have also provided basic information useful in searches for complex diseases. These maps are both highly useful genetic tools and tantalizing reflections of the genetic structure of our species.

Methods

Samples

Purified genomic DNA samples comprising TSC allele frequency panels were obtained from selected human diversity panels assembled at The Coriell Institute for Medical Research (Camden, NJ, USA). The three population panels contained 42 Caucasian samples from the Caucasian HD-100 panel, 42 African American samples from the African American HD-100 panel, 10 Japanese samples from the HD07 panel, and 10 Chinese samples from the HD02 panel. An additional 22 Japanese samples were obtained from the American Diabetes Association to bring the total of the number of Asian samples to 42. TSC allele frequency panels are available directly from the Coriell Institute for Medical Research, http://snp.cshl.org/allele_frequency_project/panels.shtml.

The Korean DNA sample was obtained from 43 randomly selected healthy Korean women ages 34.4–62.6 years (53.2 years mean, 6.25 SD) who did not have any pathological symptoms detected during interview and blood test. Blood was drawn into ACD-A tubes and the lymphocytes were isolated and transformed with EB virus. Genomic DNA was isolated from the EB-virus-transformed lymphocytes using a standard method.

SNP selection

For the Orchid and WU studies, SNPs were chosen to be well distributed throughout the genome (very few Y chromosome SNPs were characterized in this study and they have been excluded from this analysis). Initially, the WU group chose a candidate SNP every 25 kb based on the assembled draft genome sequence at that time, and the Orchid group followed a similar procedure. Later, additional SNPs were chosen in the regions where no SNPs with appreciable minor allele frequencies were initially found, and a small number of SNPs around genes of interest were characterized based on requests from outside groups with specific gene-hunting interests. Most of the SNPs characterized had been identified by TSC. However, 3853 SNPs were found to be identified both by TSC and by comparison of overlapping BAC sequences. As expected, the vast majority of the SNPs were found in noncoding regions. In the course of the projects, some SNPs were withdrawn from the database by the database managers due to reconsideration of the identification criteria or because they failed to map to a unique genomic location. The SNPs mapped to more than one genomic location as a result of gap filling leading to identification of genomic duplications [26]. All such problematic SNPs were excluded from the analysis.

Pooled sequencing: Washington University

For the WU study, allele frequencies were estimated from the sequencing of pooled DNAs [27,28]. Briefly, primers were designed using RepeatMasked sequence, the Primer3 program, and postprocessing protocols. This pipeline provided uniform and stringent thermocycling conditions for PCR and maximized the quality of sequence. Each PCR was performed with a hot-start DNA polymerase, 4 ng of DNA from pooled DNA samples or from a reference DNA, other standard reagents, and a 10-fold excess of one primer compared with the second primer. We have found that excess addition of one primer at this stage removed the need for PCR product purification and subsequent addition of a primer for cycle sequencing. The thermocycling protocol comprised an initial step at 95°C for 2 min to activate the polymerase, then 35 cycles of denaturation at 92°C for 10 s, annealing at 58°C for 20 s, and extension at 68°C for 30 s, followed by a final extension at 68°C for 10 min. Cycle DNA sequencing was conducted using BigDye version 3 mix according to the protocol of the manufacturer (Applied Biosystems). Extra dyes were removed from the sequencing reactions using columns in 96- or 384-well plate format. The samples were electrophoresed on a 3700 DNA sequencer (Applied Biosystems). The relative heights of compound bands in the electropherograms compared with control bands from a reference DNA source were analyzed to estimate allele frequencies [29].

Single-base-pair primer extension: Orchid

Using Orchid's proprietary high-throughput single-base primer extension technology, SNP-IT [30], individual genotypes were determined for each of the samples in the respective populations. A minimum of 30 successful genotypes was required to include a SNP in the data set. Approximately half the study was performed on Orchid's 25K SNPstream genotyping platform, while the remaining half of the study was analyzed on Orchid's SNPcode platform. Primers for the study were designed using Orchid's automated primer design software program, Autoprimer. For each SNP, a set of three primers was chosen, two PCR primers were selected to amplify a 100-to 200-bp product under standard conditions and a single-base primer extension (SBE) primer was designed to be approximately 25 bp in length on one side of the SNP site. For the SNPcode platform, tag sequences were assigned to each SBE primer for use in the tag-capture step. These hybrid sequences were then analyzed for secondary structure using an algorithm developed from empirical data [31]. Any tag-primer combination found to be unsatisfactory by this algorithm was assigned a new tag sequence in silico.

Single-base primer extension: SNPstream 25K platform

Orchid's SNPstream 25K is an integrated automation system customized to perform SNP genotyping of DNA samples in 384 well plates using Orchid's proprietary technology, SNP-IT, with a colorimetric readout [30]. SNPstream 25K is an application of the Beckman Sagian Core System. The system consists of a series of hardware modules and an articulated robotic arm with associated programming and control software. The system has been optimized to perform fully automated processing and allele calling.

Automated liquid-handling robotics were used to set up 10- μ l PCRs in 384-well microtiter plates. Each PCR contained 4.0 ng of DNA, 1 $\times$ PCR buffer, 1.0 unit of Platinum *Taq* (Invitrogen), 5.0 mM MgCl₂, 75 μ M dNTPs, 1.2 μ M primers. Reactions were incubated at 95°C for 2 min and then cycled 35 times at 94°C for 30 s; 50, 55, or 60°C for 2 min; 72°C for 30 s. An annealing temperature of 50, 55, or 60°C was appropriately chosen for the required conditions. Prior to genotyping, the primer extension primer, SNP-IT primer, was aliquoted into the proper well and bound to the surface using octyldimethylamine [30].

Reactions on the SNPstream 25K platform consist of automated step-wise additions of reagents to perform hybridization reactions, extension reactions, and colorimetric detection of incorporated labeled nucleotides [32]. PCR products were made single-stranded by the addition of T7 exonuclease (0.45 U/ μ l) followed by incubation at room temperature for 30 min. The single-stranded PCR product was then hybridized to the SNP-IT primer in a 384-well plate format at room temperature. After hybridization of the template strands, SNP-IT primers were extended by 1 base at the polymorphic site of interest. The extension mixes contained two labeled terminating nucleotides (one fluorescein, one biotin) and two unlabeled terminating nucleotides [30]. Extension reactions were performed at room temperature for 30 min using the Klenow fragment of DNA polymerase I. An ELISA-based technique was utilized for detection of the extension product. Anti-fluorescein-alkaline phosphatase (Boehringer Mannheim, Indianapolis, IN, USA) was used with the substrate *p*-nitrophenyl phosphate (Moss, Pasadena, MD, USA) to detect fluoresceinated nucleotides (405-nm wavelength), representing allele 1. An antibiotin-horseradish peroxidase conjugate (Zymed, San Francisco, CA, USA) followed by the substrate tetramethylbenzidine (Moss) was then used to detect biotinylated nucleotides (620 nm), representing allele 2. The raw OD data from the ELISA detection were captured by a standard plate reader and analyzed by an in-house software program, GetGenos, which uses cluster analyses of the raw OD signals to determine sample genotypes. Each genotype call was automatically assigned a confidence measure according to the most likely or probable cluster in which a data point was located. Automated genotype calls were corroborated by visual inspection of the data. Analysis of the European samples was undertaken first and successful assays were “cherry picked” into new plates to be analyzed against the other populations.

Single-base primer extension, SNPcode platform: Orchid

SNPcode is a high-throughput genotyping platform that detects a SNP by the specific incorporation of a fluorescent dye, using a multiplex thermocycled single-base primer extension, followed by solid-phase sorting using a Universal Tag Array or Zip-Code chip prior to readout. The SNPcode platform specifically uses the Affymetrix GenFlex Tag Array chip, which has 2000 unique features. Typical SNPcode reactions routinely assay 1824 SNPs per chip and are performed using 12-plex PCRs.

Automated liquid-handling robotics were used to set up 10- μ l PCRs, which contained 4.0 ng of genomic DNA. The PCR protocol used on this platform is similar to the one previously described [33], with the exception that only 35 cycles were used for the reactions. Prior to commencement the SBE genotyping reactions, excess nucleotides and PCR primers were removed using shrimp alkaline phosphatase and exonuclease I (Custom ExoSap-IT; USB Corp.). A cocktail containing one fluorescein-labeled and one biotin-labeled nucleotide terminator (PE-NEN, Boston, MA, USA), along with the two remaining unlabeled terminators, was combined with a pool of 12 extension primers and a thermostable polymerase such as ThermoSequenase (Amersham Biosciences, Piscataway, NJ, USA) with its appropriate buffer. SBE reactions were then incubated at 96°C for 3 min, followed by 46 cycles of 94°C for 20 s and 40°C for 11 s.

Prior to the solid-phase sorting of the multiplexed reactions for readout, 152 12-plex PCRs were pooled together and precipitated to concentrate the volume of the reaction for hybridization to the Affymetrix Genflex chip. Pellets were resuspended in hybridization buffer (100 mM Mes, pH 6.6, 1 M NaCl, 20 mM EDTA, 0.01% Tween 20) and injected onto the GenFlex chip. Chips were incubated at 45°C for 16 h in the Affymetrix GeneChip system hybridization oven [34]. Arrays were washed with Buffer A (6 $\times$ SSPE/0.01% Tween) at 25°C, followed by Buffer B (3 $\times$ SSPE/0.01% Tween) at 45°C. Chips were then stained for 10 min

at 25°C with streptavidin-conjugated r-phycoerythrin for biotin detection (6× SSPE, 1× Denhardt's solution (Sigma), 0.01% Tween 20, 5 µg/ml streptavidin-conjugated r-phycoerythrin, 5 µg/µl streptavidin), followed by a rinse with Buffer A.

Chips were scanned on the GeneArray scanner (Affymetrix, Santa Clara, CA, USA) at 530 and 570 nm to detect fluorescein and biotin, respectively. Hybridization controls were used to normalize the resulting fluorescence intensity scores for signal bleedthrough between the two channels. Genotyping scores were generated from the ratio of the signal from both channels (fluorescein/(fluorescein + biotin)).

Genotype calling and data analysis

Data were analyzed for each individual SNP separately. Scatter plots were generated with the *x* axis as the genotype score described above and the *y* axis the log of the total signal intensity from both channels. Thresholds were set for each of the three possible genotype clusters. The resulting data for each of the SNPs was initially classified into categories (such as failed, monomorphic, and good), to speed up the data review process and to improve data calling.

Acknowledgements

We are grateful for the contributions of Dr. Patrick K. Bender, Ms. Betsy Messina, and Dr. Lorraine H. Toji, at the Coriell Institute of Medical Research (Camden, NJ, USA), for their guidance and assistance in the assembly of the TSC DNA allele frequency panels. We also acknowledge the assistance of Dr. Mat Petersen from the American Diabetes Association, for allowing us to use samples from the ADA collection to build the TSC panels. We also thank James Marcella, Jack Ball, Felicia Watson, and Robert Tomacelli for their advice and guidance during the development of the project at Orchid. This study was supported in part by IMT-2000 Grant (01-PJ11-PG9-01BT05-0003) from the Korean Ministries of Health and Welfare and Information and Communication. This work is funded in part by The SNP Consortium and by the NHGRI (HG1720 to P.Y.K.).

References

1. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–933. [PubMed: 11237013]
2. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–2229. [PubMed: 12029063]
3. Taillon-Miller P, Bauer-Sardina I, Saccone NL, et al. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* 2000;25:324–328. [PubMed: 10888883]
4. Taillon-Miller P, Saccone SF, Saccone NL, et al. Linkage disequilibrium maps constructed with common SNPs are useful for first-pass disease association screens. *Genomics* 2004;84:899–912. [PubMed: 15533707]
5. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997;278:1580–1581. [PubMed: 9411782]
6. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517. [PubMed: 8801636]
7. McKeigue PM, Carpenter JR, Parra EJ, et al. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* 2000;64:171–186. [PubMed: 11246470]
8. International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–796. [PubMed: 14685227]
9. Holden AL. The SNP Consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques Suppl* 2002;26:22–24.
10. Taillon-Miller P, Kwok PY. A high-density single-nucleotide polymorphism map of Xq25–q28. *Genomics* 2000;65:195–202. [PubMed: 10857743]

11. Miller RD, Taillon-Miller P, Kwok PY. Regions of low single-nucleotide polymorphism incidence in human and orangutan Xq: deserts and recent coalescences. *Genomics* 2001;71:78–88. [PubMed: 11161800]
12. Miller RD, Kwok PY. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* 2001;10:2195–2198. [PubMed: 11673401]
13. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999;22:231–238. [PubMed: 10391209](Published erratum appears in *Nat. Genet.* 23 (1999) 373).
14. Halushka MK, Fan JB, Bentley K, et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 1999;22:239–247. [PubMed: 10391210]
15. Crawford DC, Carlson CS, Rieder MJ, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* 2004;74:610–622. [PubMed: 15015130]
16. Reich DE, Schaffner SF, Daly MJ, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 2002;32:135–142. [PubMed: 12161752]
17. Marth G, Yeh R, Minton M, et al. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* 2001;27:371–372. [PubMed: 11279516]
18. Parra EJ, Kittles RA, Argyropoulos G, et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am J Phys Anthropol* 2001;114:18–29. [PubMed: 11150049]
19. Marth GT, Czabarka E, Murvai J, et al. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004;166:351–372. [PubMed: 15020430]
20. Akey JM, Zhang G, Zhang K, et al. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805–1814. [PubMed: 12466284]
21. Eberle MA, Kruglyak L. An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet Epidemiol* 2000;19(Suppl 1):S29–S35. [PubMed: 11055367]
22. Carlson CS, Eberle MA, Rieder MJ, et al. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003;33:518–521. [PubMed: 12652300]
23. Yu N, Chen FC, Ota S, et al. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* 2002;161:269–274. [PubMed: 12019240]
24. Enard W, Przeworski M, Fisher SE, et al. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 2002;418:869–872. [PubMed: 12192408]
25. Lai CS, Fisher SE, Hurst JA, et al. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 2001;413:519–523. [PubMed: 11586359]
26. Eichler EE. Segmental duplications: what's missing, misassigned, and misassembled—And should we care? *Genome Res* 2001;11:653–656. [PubMed: 11337463]
27. Vieux EF, Kwok PY, Miller RD. Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* 2002;32:S28–S32.
28. Miller RD, Duan S, Lovins EG, et al. Efficient high-throughput resequencing of genomic DNA. *Genome Res* 2003;13:717–720. [PubMed: 12654721]
29. Kwok PY, Carlson C, Yager TD, et al. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* 1994;23:138–144. [PubMed: 7829062]
30. Picoult-Newberg L, Ideker TE, Pohl MG, et al. Mining SNPs from EST databases. *Genome Res* 1999;9:167–174. [PubMed: 10022981]
31. Yuryev A, Huang J, Scott KE, et al. Primer design and marker clustering for multiplex SNP-IT primer extension genotyping assay using statistical modeling. *Bioinformatics* 2004;20:3526–3532. [PubMed: 15284101]

32. Reynolds, JE.; Head, SR.; McIntosh, TC., et al. Genetic bit analysis: a solid-phase method for genotyping single nucleotide polymorphisms. In: Caetano-Anolles, G., editor. DNA Markers: Protocols, Applications, and Overviews. Wiley-Liss; New York: 1997. p. 199-211.
33. Bell PA, Chaturvedi S, Gelfand CA, et al. SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery . Biotechniques 2002;Suppl:70–72. 74, 76–77. [PubMed: 12083401]
34. Fan JB, Chen X, Halushka MK, et al. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. Genome Res 2000;10:853–860. [PubMed: 10854416]

Fig 1.

Distribution of minor allele frequencies. These data were from the Orchid portion of the complete data set and include SNPs for which no variant was detected in the three panels (monomorphic SNPs). SNPs were chosen from TSC database. Very similar distributions were found for SNPs in dbSNP in the WU study. The error bars, present on all points but often too near the points to be visible, represent the standard error of the mean and were calculated by randomly assigning each SNP to one of three subsets. (a) SNPs mapping to autosomes. Results are shown for the three assay panels, African American (Af), Asian (As), and European American (Eu). The first bin contains SNPs with MAF of 0 to <10%, including 7.3% monomorphic SNPs. The last bin contains SNPs with MAF of 40 through 50%. (b) SNPs

mapping to the X chromosome. The first bin contains 14.1% monomorphic SNPs. (c) Population-specific SNPs. Data from (a), except that only SNPs with variation found in a single population are shown (note change in scale).

Fig 2.

Common-link SNPs. SNPs with a high MAF in all three populations. The combined data set was used, and the error bars represent the 95% confidence intervals. Common-link SNPs with an MAF $\geq 30\%$ are also included as SNPs with an MAF $\geq 20\%$, and so forth.

Fig 3.

Display of SNP distributions across the genome. A graphical representation of SNP distributions across each of the autosomes is shown. Gaps containing no SNPs and greater than 800 kb are indicated by a blue bar. The numbering of the chromosomes starts at the p arm and is displayed as units of 100 kb. In the three columns drawn for each chromosome, the left one illustrates the distribution of all the SNPs in TSC database. The middle column and the right column illustrate the distribution of SNPs selected from the European-derived population in this study with MAFs of ≥ 1 and $\geq 10\%$, respectively.

Fig 4.

Allele frequency divergence between populations. As a measure of divergence between two populations, the difference in allele frequencies is shown. SNPs with detected variation (excluding monomorphic SNPs) from the Orchid and Korean portion of the combined complete data set were used for the distributions. The assay panels are as in Fig. 1, with the addition of Japanese (Ja) and Chinese (Ch) as subsets of the Asian panel and Korean (Ko). The error bars represent the 95% confidence intervals. The allele frequency differences are collected into bins. For example, in (a) the 80% bin shows the fraction of total SNPs with allele frequency differences $\geq 60\%$ and less than 80%; these SNPs may be useful for admixture studies. Simulations based upon the appropriate number of chromosomes were conducted to estimate the apparent divergence between populations due to sampling alone; the sampling divergence was restricted to the first bins for (a) and (b) and is not shown. (a) SNPs mapping to autosomes. (b) SNPs mapping to the X chromosome. (c) SNPs mapping to autosomes in Asian populations. Simulated results are shown with dotted lines. (d) The divergence tail for SNPs mapping to autosomes.

Table 1

Monomorphic SNPs

Population	% total	<i>N</i>
All	7.3	765
Af	12.3	975
As	26.4	902
Eu	19.5	1192

SNPs scored monomorphic by Orchid and WU. Population designations are as in Fig. 1.

Table 2

Differences between groups (%)

Groups	Autosomes	X
Af-As	18.82 ± 0.01	26.07 ± 0.40
Af-Eu	16.30 ± 0.09	23.44 ± 0.20
As-Eu	16.09 ± 0.04	20.53 ± 0.45
Ch-Jp	8.61 ± 0.11	9.65 ± 0.39
Jp-Ko	5.86 ± 0.06	4.74 ± 0.59
Ch-Ko	8.39 ± 0.12	7.80 ± 0.69

Values ± SEM are shown. The simulated difference between the continental groups ($2N = 84$) is 3.14%. The simulated differences for Ch-Jp, Jp-Ko, and Ch-Ko are 6.78, 3.61, and 6.92%, respectively.