

A Whole-Genome Linkage Disequilibrium SNP Map and Validated Assay Resource

Leila Smith, Charles Scafe, Yu Wang, Marion Laig-Webster, Xiaoping Su, Ryan Koehler, Hadar Avi-Itzhak, Janet Ziegler, Lewis Wogan, Eugene Spier, Dennis A. Gilbert, and Francisco M. De La Vega

Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404, USA

Abstract

We developed a set of 5' nuclease allelic discrimination assays to score single nucleotide polymorphisms (SNPs) with the aim of creating a reference map for use in candidate-gene, candidate region and whole-genome linkage disequilibrium (LD) mapping studies. The assays were validated by individually genotyping 90 DNA samples, 45 from African-American and 45 from Caucasian individuals, selected from the Coriell Human variation collection. Our goal is to define a set of >150,000 assays distributed across all the genes in the genome for SNPs of high heterozygosity in at least one population. Candidate SNPs were prioritized from the Celera RefSNP database which contains 4 million unique SNPs from combined Celera and Public SNP databases through a triage process that requires evidence of independent discovery of the minor allele. We selected SNPs on 27,007 Celera gene predictions, in a gene focused picket-fence with an average density of one SNP per 10 kb of gene length, including 10 kb upstream and downstream of the predicted gene boundaries. PCR primers and TaqMan[®] probes for the 5' nuclease assays were then designed by a software pipeline that picks oligonucleotide sequences and then screens the assays against the genome database for potential artifacts. Following genotyping 90 individuals, the performance of each assay is benchmarked against stringent criteria for background signal, adequate signal generation, and specificity. Our validation results showed that 94% of the SNPs tested in the population panels were polymorphic and about 90% of the assays passed our stringent performance criteria. Of those, 87% have minor allele frequencies ≥ 0.05 in Caucasian panel and 88% in African-American samples. These figures represent an extremely high SNP validation rate, and an unprecedented yield of common SNPs useful in LD mapping. Allele frequency data in the populations tested will be made available with the assays. The individual genotypes being generated have enabled us to identify blocks of LD and the haplotype diversity across all gene regions of the genome for these populations. This information is being used to refine the SNP set coverage.

Presented at the Eighth International Human Genome Meeting

Cancún, México

April 27 – 30, 2003

Poster #86

A WHOLE-GENOME LINKAGE DISEQUILIBRIUM SNP MAP AND VALIDATED ASSAY RESOURCE



Leila Smith, Charles Scafe, Yu Wang, Marion Laig-Webster, Xiaoping Su, Ryan Koehler, Hadar Avi-Itzhak, Janet Ziegler, Lewis Wogan, Eugene Spier, Dennis A. Gilbert, and Francisco M. De La Vega
Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404, USA

Introduction

Applied Biosystems has developed a set of TaqMan[®] probe-based (5' nuclease) assays to score single nucleotide polymorphisms (SNPs) with the aim of creating a reference map for use in candidate region, candidate-gene, and eventually whole-genome association studies by linkage disequilibrium (LD) mapping. This set of ready to use assays provides high-density coverage of all known gene regions to enable easier and more affordable genetic studies, yielding genotyping answers in a matter of days. The assays are manufactured, functionally QC tested, and validated by individually genotyping 180 DNA samples selected from four major populations in our high-throughput genotyping services facility before being put in inventory. The resulting allele frequency data is made available on the web to help in the selection of the assays. Our goal is to define a set of about 150,000 assays distributed across all genes of the human genome for SNPs of high heterozygosity in at least one of the four populations tested: African-American, Caucasian, Chinese, and Japanese.

RESULTS

Assay Validation Yield

Our results have confirmed that the SNP selection "triage" procedure was effective in prioritizing SNPs with higher likelihood of being highly heterozygous in multiple populations. From the 258,260 assays validated so far on African-American and Caucasian populations, approximately 95% of the 122,287 SNP assays that passed our stringent performance criteria were indeed polymorphic. As shown in Figure 2, 88% of the polymorphisms have a minor allele frequency above 5% in the African-American or Caucasian panels, our required product release criteria. To date, we have also obtained allele frequency information on over 67,000 assays on both Chinese and Japanese population samples, showing that for one or the other population 90% of SNPs have a minor allele frequency above 5%, and a very considerable overlap of common SNPs between all four populations tested. We anticipate that these statistics will not change significantly when genotyping of all assays in the Asian population panels is concluded. These figures represent an extremely high SNP validation rate, and an unprecedented yield of common SNPs useful in LD mapping.

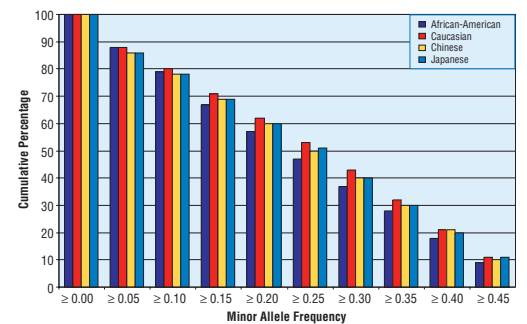


Figure 2. Distribution of the minor allele frequency of validated SNPs in each population studied.

Methods

SNP Selection for a Linkage Disequilibrium Marker Set

Selecting SNPs within gene regions, at an average density of one per 10 kb, makes our map resemble a gene-focused picket fence. Currently, the gene list we use includes 27,007 gene regions derived by Celera Genomics, their boundaries expanded by 10 kb up- and downstream to account for regulatory regions and undiscovered exons and UTRs. The candidate SNPs were selected from the Celera Human RefSNP database (version 3.6) through a "triage" process that requires evidence of independent discovery of the minor allele. First, we culled over 1 million SNPs with increased likelihood of having high heterozygosity from a starting set of more than 4.1 million genomically mapped public and Celera-discovered SNPs. This initial selection required multiple independent observations of a SNP's minor allele. We devised custom queries to the RefSNP database to identify SNPs discovered both by Celera and by the public SNP discovery efforts. In addition, we selected SNPs whose minor alleles were observed in at least two distinct donors of the Celera shotgun sequencing of the human genome. Finally, we compared single-donor Celera SNPs to the public genomic assembly to find cases where the Celera minor allele was confirmed in the public consensus sequence.

SNP Assay Development

In the second major step of our strategy, PCR primers and TaqMan[®] probes were designed by an algorithm pipeline which selects oligonucleotide sequences. These primer and probe designs are then screened against the genome database as a computational QC step for potential artifacts. We subjected 5' nuclease assays that passed the previous step to further selection criteria:

- Being in or within 10kb of a gene region, and
- Being optimally spaced to provide at least 3 SNPs per gene with a maximal inter-SNP physical distance of 10kb.

Finally, we filled remaining gaps in gene regions with 2 unscreened SNPs per 10kb to take into account an expected 50% rate of validation of these lower confidence candidate SNPs.

After the primers and probes are synthesized in our high-throughput manufacturing facility two quality-control steps occur: The first tests oligonucleotide integrity and the second tests assay performance against a panel of 10 individual genomic DNA samples. Only assays that pass this manufacturing QC are moved on for validation in the population panels, which include DNA samples from 45 African-American, 45 Caucasian (from the Coriell Institute/NIGMS Human Variation panels), 45 Chinese, and 45 Japanese individuals (obtained through collaborations from properly consented cell-line repositories). Assay validation in population samples shows that the locus is polymorphic and that the allele frequency will be adequate for association studies in a variety of populations. The performance of each assay is benchmarked against stringent criteria for background signal, adequate signal generation, and specificity. Assays that meet our performance criteria and a minimum minor allele frequency of 5% in either of the populations tested are annotated and released for sale at the Applied Biosystems on-line store.

Analysis of genotype data from reference samples

The individual genotypes of the 180 DNA samples generated during validation have enabled us to study the profile of LD across all gene regions of the genome for these populations. We have applied methods to identify haplotype blocks, regions of strong LD and low haplotype diversity, and where the statistical power for finding association should be high. In addition, we intend to construct metric maps scaled to the strength of LD that can guide the selection of SNPs for association studies independent of block boundaries (cf. Maniatis et al., PNAS 99: 2228-33, 2002). Ultimately, the metric of greatest practical utility will relate to the power of detecting an association between a disease or disease-risk phenotype and SNPs marker(s) in that region. Our empirical data provides a unique opportunity to estimate the power of a LD SNP map for a large number of known genes. These power estimations can be used to properly design a genetic study, selecting the adequate number of markers and sample size.

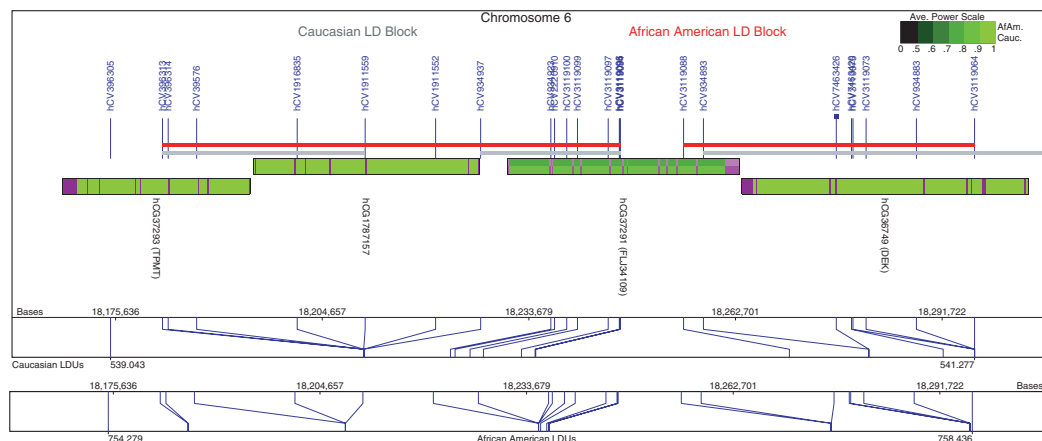


Figure 3. Example of the distribution of Assays-on-Demand[™] SNP Genotyping Products across a region of chromosome 6. Validated SNPs are indicated by vertical lines with Celera identifiers, and gene regions as horizontal rectangles, with Celera identifiers and HUGO names indicated below, and exons colored in purple. Horizontal bars represent haplotype blocks calculated for the African-American (Red) and Caucasian (Gray) populations. Gene regions are displayed in a color scale representing the results of power calculations for a fixed sample size of 500 cases and 500 controls, an assumed disease allele frequency of 0.2, and a multiplicative gene model typical of the common variant/common disease hypothesis. The two axes shown below indicate the physical scale in base-pairs, and the metric linkage disequilibrium units scale calculated with the LDMAP software of Maniatis et al. (PNAS 99: 2228-33, 2002) for the Caucasian and African-American populations.

Using our empirical data, we are in the process of identifying minimum subsets of SNPs ("tagging" SNPs) that would have adequate power in disease association studies, greatly reducing the study time and cost. Furthermore, the data allows the identification of regions where, due to the low LD, additional and complementary SNPs currently not in our validated set would be needed. These custom assays can be ordered through our Assays-by-Design[™] service which utilizes the same design algorithm. Currently, we are developing interfaces to allow researchers to access the analyses of the reference data we have obtained to help them select SNPs for their studies (see Figure 3). We anticipate that with this information, association studies can be designed more rationally according to the specific population and region of the genome under study, knowing in advance which genes would require more SNP coverage and/or larger sample size.

Assay Availability

As of April 15th we have released 122,287 assays, and we expect to reach a total of approximately 150,000 assays later this year. Through the Applied Biosystems on-line store (<http://store.appliedbiosystems.com>), the assay resource can be searched by a number of annotations. For example, researchers who know the exact SNPs they want can search using the appropriate identifiers (e.g., Celera variation ID, dbSNP rs or ss ID). Users can also research SNPs by gene name (e.g., HUGO gene symbol, RefSeq ID, Celera transcript ID), or by location within a particular chromosomal interval (using coordinates from either the public or the Celera genome assembly) or reference marker range (e.g., microsatellite, cytoband) they are interested in. Within these regions, the user can specify filtering criteria based on population allele frequency, SNP type (e.g., intronic, coding), a user-specified flanking region, or gene overlap. Once selected, the assays can be easily ordered on-line. Together with their assay order, researchers receive a CD-ROM containing the assay information file, enabling them to set-up the assay, automatically create the instrument sample sheet, and fully integrate the SNP annotation into their studies (e.g., context sequence, allele-dye key, SNP type, allele frequency).

Figure 4. On-line catalog, search, and ordering interface for the Assays-on-Demand[™] SNP Genotyping Products available at the Applied Biosystems on-line store. (May 2003 release)

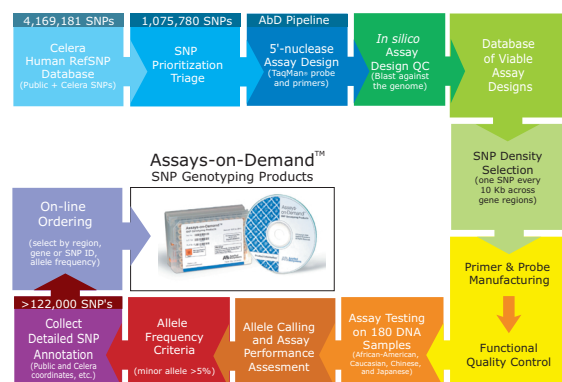


Figure 1. Assays-on-Demand[™] SNP Genotyping Products development and validation workflow.

Acknowledgements

We are indebted to Helen Belcastro, Annie Titus, Joanna Curlee, the production genotyping, LIMS, and IT teams of Services Development and Delivery, and the Global Oligo Operations teams of Applied Biosystems for their support in the generation of the data used in this work.

For Research Use Only. Not for use in diagnostic procedures.

The PCR process and 5' nuclease process are covered by patents owned by Roche Molecular Systems, Inc. and F. Hoffmann-La Roche Ltd.

Applied Biosystems is a registered trademark AB (Design), Applied Biosystems, Assays-by-Design, Assays-on-Demand, Celera and Celera Genomics are trademarks of Applied Biosystems or its subsidiaries in the US and/or certain other countries.

TaqMan is a registered trademark of Roche Molecular Systems, Inc.

© 2003 Applied Biosystems. All rights reserved.

Conclusions

Integrating information from both public and private human genome efforts, we have created a high-quality LD map of validated SNPs. Expertise in assay design and bioinformatics has allowed us to develop a set of validated SNPs and ready-to-use assay reagents for use with an easy workflow. The individual genotypes being generated has already enabled us to begin to survey the magnitude of LD and the haplotype diversity across all gene regions of the genome for these

populations. This will allow us to identify regions that will require higher or lower SNP density to further optimize the map. As of April 15th, we have designed and manufactured assays for 258,260 SNPs, commercialized 122,287 validated and annotated assays, and we expect to reach a total of approximately 150,000 assays later this year.

For more information please visit: <http://www.allsnps.com>