

SELECTION OF SINGLE NUCLEOTIDE POLYMORPHISMS FOR A WHOLE-GENOME LINKAGE DISEQUILIBRIUM MAPPING SET

Francisco M. De La Vega, Charles Scafe, Yu Wang, Marion Laig-Webster, Xiaoping Su, Ryan Koehler, Hadar Avi-Itzhak, and Eugene Spier. Bioinformatics R&D, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA, 94404, USA

Introduction

Applied Biosystems is developing 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 whole-genome and candidate-gene linkage disequilibrium (LD) mapping studies. The assays are being manufactured, placed in inventory, functionally QC tested, and validated by individually genotyping 90 DNA samples selected from the Coriell Human variation panels in our high-throughput genotyping facility. Our goal is to define a set of about 200,000 assays distributed across the genome for SNPs of high heterozygosity in at least one major population. Allele frequency data in the populations tested will be made available with the assays.

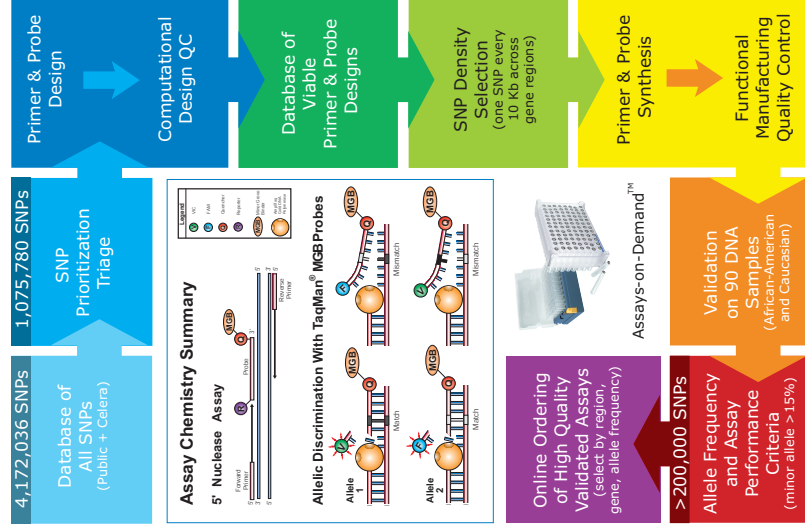
SNP Selection for a Linkage Disequilibrium Marker Set

The extent of LD across a genomic region dictates the SNP density necessary to ensure association between a marker and the causative allele sought. Although this parameter is largely unknown and variable across the genome, empirical studies suggest LD ranges from 5 to more than 200Kb. Common SNPs are the most likely to be useful for LD studies across more than one population since they represent ancient mutations that arose before ethnic group segregation, and simulation studies suggest that they are more likely to be in LD with a given causative allele regardless of whether the allele is present at low or high frequency.

In assembling our SNP set, we are focused on common SNPs in a hybrid gene-based approach. SNPs are considered "common" when the minor allele frequency is >15% in at least one of the populations used for validation. Currently, the gene list we use includes 25,083 gene regions derived by Celera Genomics. We define a "gene region" as bounded by the first and last transcribed base, including untranslated regions, plus 10 kb upstream and downstream to account for uncharacterized exons and regulatory regions. Selecting SNPs within gene regions, at an average density of one per 10 kb, makes the map resemble a gene-focused picket fence. Density for specific regions is adjusted as data on recombination and LD extent emerges. At the close of the project we may elect to complement this set with additional SNPs in intergenic regions, in particular non-coding regions of homology between mouse and human.

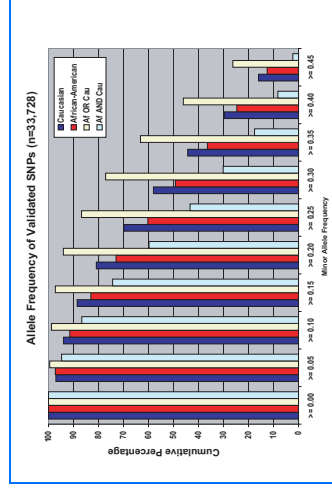
SNP Assay Development

The candidate SNPs are selected from the Celera RefSNP database (version 3.4.1) through a "triage" process that requires evidence of independent discovery of the minor allele. PCR primers and TaqMan® probes are then designed by an algorithm pipeline that picks oligonucleotide sequences and then screens the assays against the genome database as a computational QC step for potential artifacts. As of April 30, we have selected >200,000 SNPs for assay development after selecting for the target density, and orders for primers and probes have been placed on our oligo manufacturing facility. After the primers and probes are synthesized two additional quality-control steps occur. The first tests oligonucleotide integrity, and the second tests assay performance against a panel of 10 DNA samples. Only assays that pass this manufacturing QC are moved on for validation in the population panels, which include DNA from 45 Caucasians and 45 African-Americans from Coriell. Assay validation in population samples ensures 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.

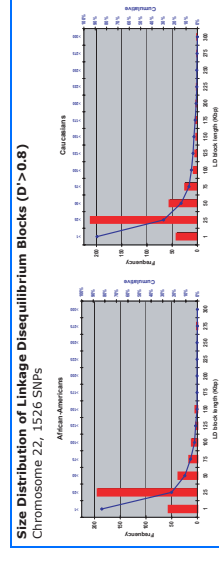


Results

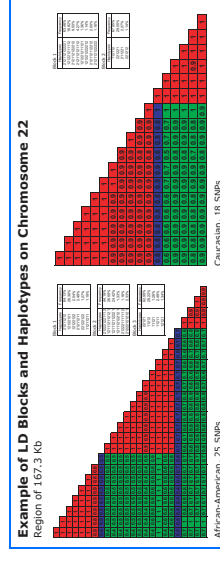
Our preliminary results indicate that 94% of the SNPs tested with the population panels were polymorphic and about 90% of the assays passed our stringent performance criteria. When using a minor allele frequency cut-off of 15% or greater, 85% of the samples from the African-American panel, and 90% from the Caucasian panel meet the criteria (see histogram below). These figures represent an extremely high SNP validation rate, and an unprecedented yield of common SNPs useful in LD mapping.



We have prioritized the validation of the assays in chromosomes 6, 21, and 22 to begin analyzing the extent of LD among these markers, as well as surveying the common haplotype blocks across the gene regions of whole chromosomes. Preliminary analysis of the genotypes of 1,526 SNPs on chromosome 22 shows the existence of large blocks of LD ($D' > 0.8$) in the Caucasian population, that are found more frequently in the African-American individuals. In Caucasians, 77.5% of the typed SNPs were harbored in LD blocks with an average size of 25.8 Kb; ~30% of the blocks found were greater than 25 Kb, the largest spanning 299 kb. However, in African-Americans 67.5% of the SNPs were found in LD blocks with an average size of 18 Kb, and only ~22% of the blocks found were greater than 25 Kb, the largest being 252 Kb long.



One large block found encompasses 11 genes and extends to at least 270 Kb in Caucasians, whereas in African-Americans the block is split in two smaller blocks of 6 and 66 Kb. Another example is shown in the figure below, for a region spanning 167 Kb, showing the consistent finding of smaller blocks in the African-Americans, as compared to Caucasians. These blocks have low haplotype diversity, as computed from the genotypes of the unrelated individuals by the EM haplotype algorithm included in the software package MERLIN (Abecasis et al., Nat Genet 30:97-101, 2002).



Conclusions

Integrating information from both public and private human genome efforts, we are creating a high-quality LD map of validated SNPs. Expertise in assay design and bioinformatics allows us to turn this information resource into a set of over 200,000 validated SNPs, and ready-to-use assay reagents. The individual genotypes being generated will enable us to survey the magnitude of LD and the haplotype diversity across all gene regions of the genome for these populations, to better document the utility of the SNPs selected for the set.

Acknowledgements

We are indebted to Janet Ziegler, Lewis Wogan, and the Genomics Applications genotyping lab team for generating the data presented here, as well as the Global Oligo Operations for providing manufacturing QC data. The authors thank Jinghui Zhang and Frank Liu for performing custom queries and processing the JSNP database through their data integration pipeline, and Emily Winn-Deen for genotyping some samples in her lab, all from Celera Genomics. Thanks are also due to Andrew Clark (U Penn.), Kenneth Klot (Yale Medical School), Kit Lau, and John Shinsky (Celera Diagnostics) for many helpful discussions, and Joanna Curfee for computer systems support.

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.