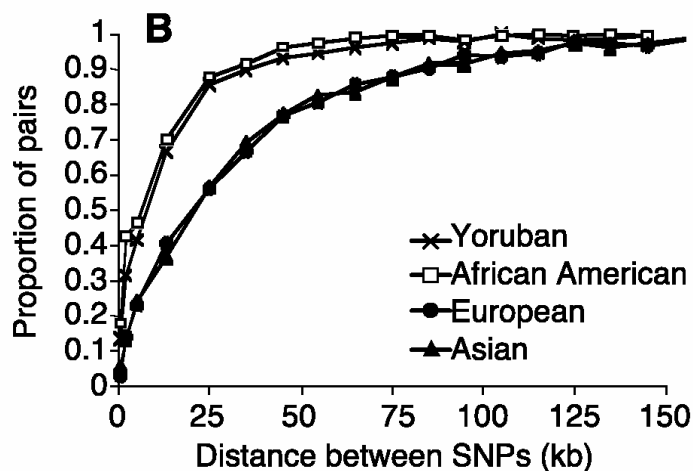


Haplotype Block Detection Using Cira Pattern Discovery

This note reports on a novel method of detection of haplotype blocks in the human genome. It is based upon pattern discovery using technology being patented by Cira Discovery Sciences, Inc. It is intended as a demonstration of our pattern discovery technology in the context of genomic variation (SNP) data. Actual applications in genomic analyses will take advantage of technological advances that are currently under development in Cira.

Although the primary commercial application of pattern discovery will be in the area of association of patterns of genetic markers with phenotype (e.g. disease, drug reaction) it is of interest to demonstrate, with data available today, that (a) pattern discovery is applicable to SNP data, and that (b) pattern discovery can detect SNP associations that replicate known genomic properties. In this case the genomic property of interest is the haplotype structure of the human genome.

Data were obtained from (Gabriel, et al. 2002). This work, done at the Whitehead Institute by the groups of Eric Lander, Mark Daly, and David Altshuler, sets the stage for the Human HapMap project, a follow-on to the SNP Consortium's dbSNP project. Gabriel et. al. demonstrated that the genome can be objectively parsed into blocks of limited diversity, punctuated by spans of significant recombinational diversity. The methodology used was based on analysis of pairwise marker linkage disequilibrium (LD), but required extensive statistical conditioning of the data prior to LD calculations. The following figure, taken from their paper, shows the distribution of haplotype block sizes derived from four populations, 30 parent-offspring trios (90 individuals) from Nigeria (Yoruba), 93 members of 12 multigenerational pedigrees of European ancestry, 42 unrelated individuals of Japanese and Chinese origin, and 50 unrelated African Americans:



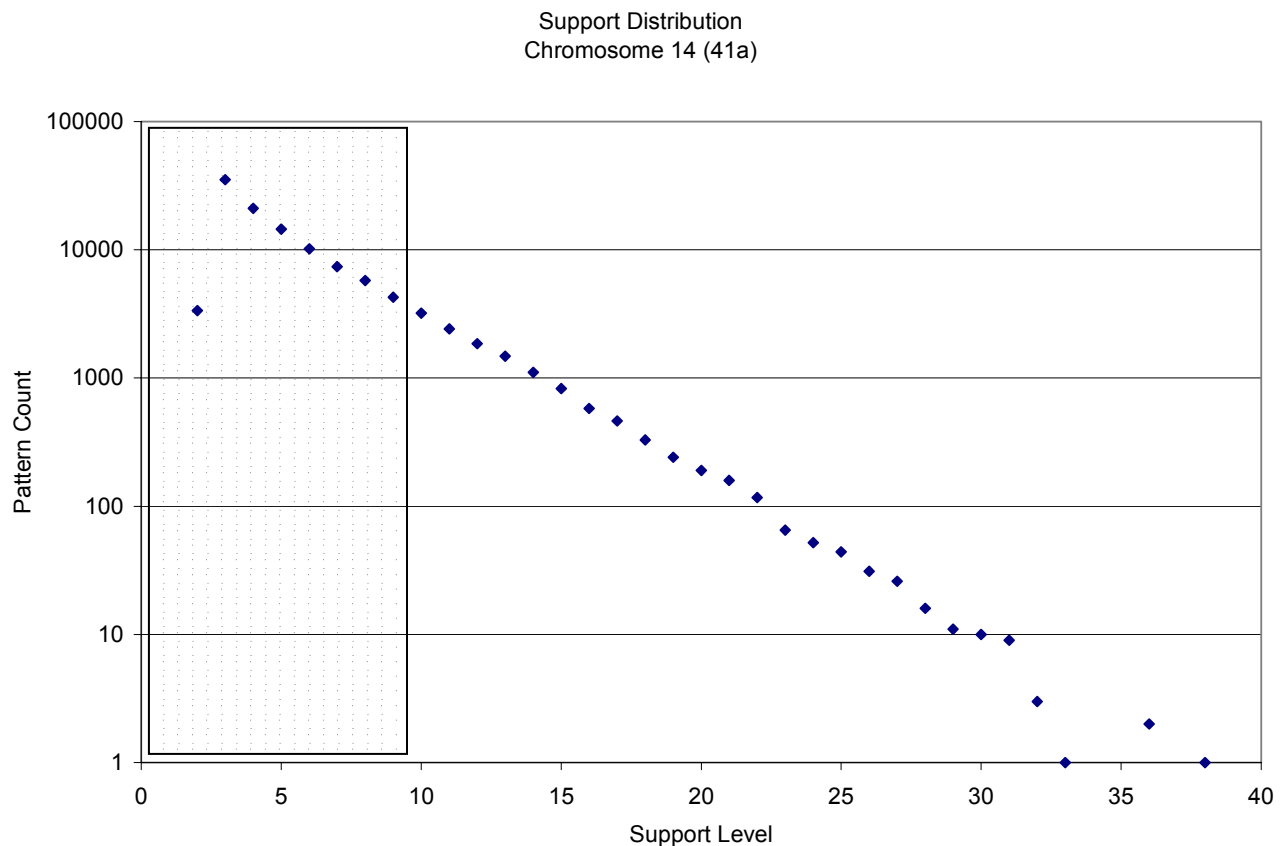
This graph shows that the average haplotype block size is significantly larger in the European and Asian populations as compared with the Yoruban and African American populations.

Cira's Approach

Our approach, outlined here, involved computing patterns of SNPs, and then looking for allele frequencies among patterns that occur in a relatively large percentage of the population (we refer to these as high support patterns). Patterns represent marker linkages. Thus, if there exist blocks in which limited marker diversity exists, these blocks should appear in a histogram as a contiguous grouping of markers that tend to occur in the high-support patterns.

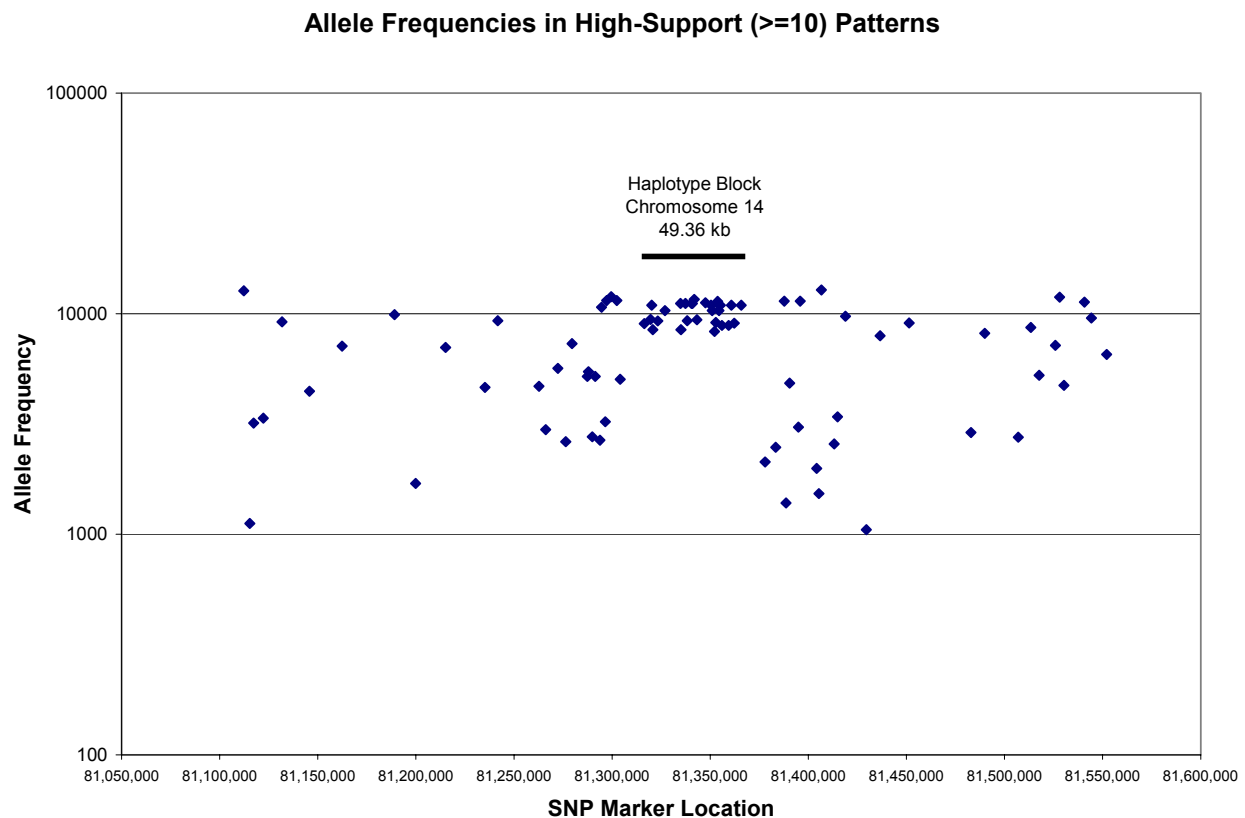
As a test, we arbitrarily chose the largest population, "population A" (corresponding to the European population), and within this chose the largest contiguous grouping of marker data which happens to correspond to chromosome 14. Data corresponding to one of the pair of chromosome 14 was arbitrarily chosen for each of the 93 individuals. These data were then formatted into a form suitable for categorical pattern discovery by means our proprietary methods.

A total of 115,032 patterns were produced up to an initial support level of 3. The total computation time required to discover the patterns was 21 sec on a 2.0 GHz Pentium 4 with 1 GByte RAM, running the Linux operating system. These patterns were then re-matched to the input data to discover their actual support distribution (support of a pattern is the number of individuals in which the pattern occurs), shown in the following figure:



The support of a pattern indicates the degree of correlation amongst the set of markers that comprise the pattern.

A support level of 10 was arbitrarily chosen to limit noise inherent in low-frequency allelic associations. The resulting pool of patterns was examined to determine the frequency with which each marker occurs (we note that in the raw data provided the corresponding “marker frequency” is exactly 1, since every marker appears once in each chromosome, by definition). The next figure shows the distribution of marker frequencies in high-support patterns:

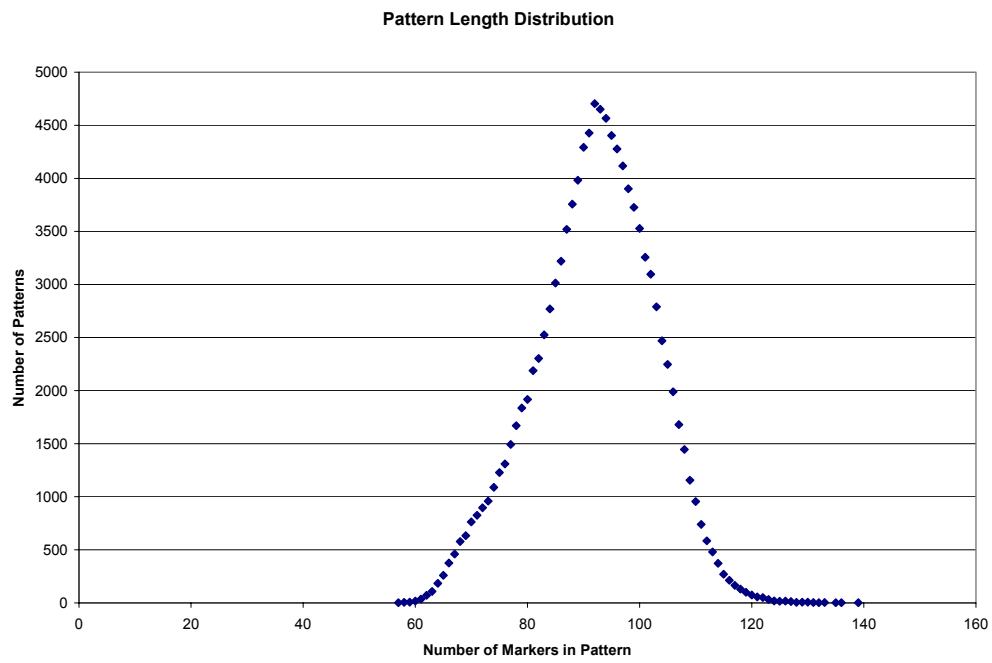


How do patterns of SNPs delineate haplotype blocks? A haplotype block is a set of contiguous markers that are highly inter-correlated across a population. In pattern discovery terms, this translates to spatially dense, high support patterns.

In the graph above a block of markers that frequently occur among high-support patterns is clearly seen as a relatively uniform and high-frequency block imbedded in a much more variable distribution of markers with a much lower average frequency of occurrence. The size of this block, 49.36 kb, is consistent with the distribution of haplotype block sizes shown in the first figure which was reproduced from the Gabriel et. al. paper.

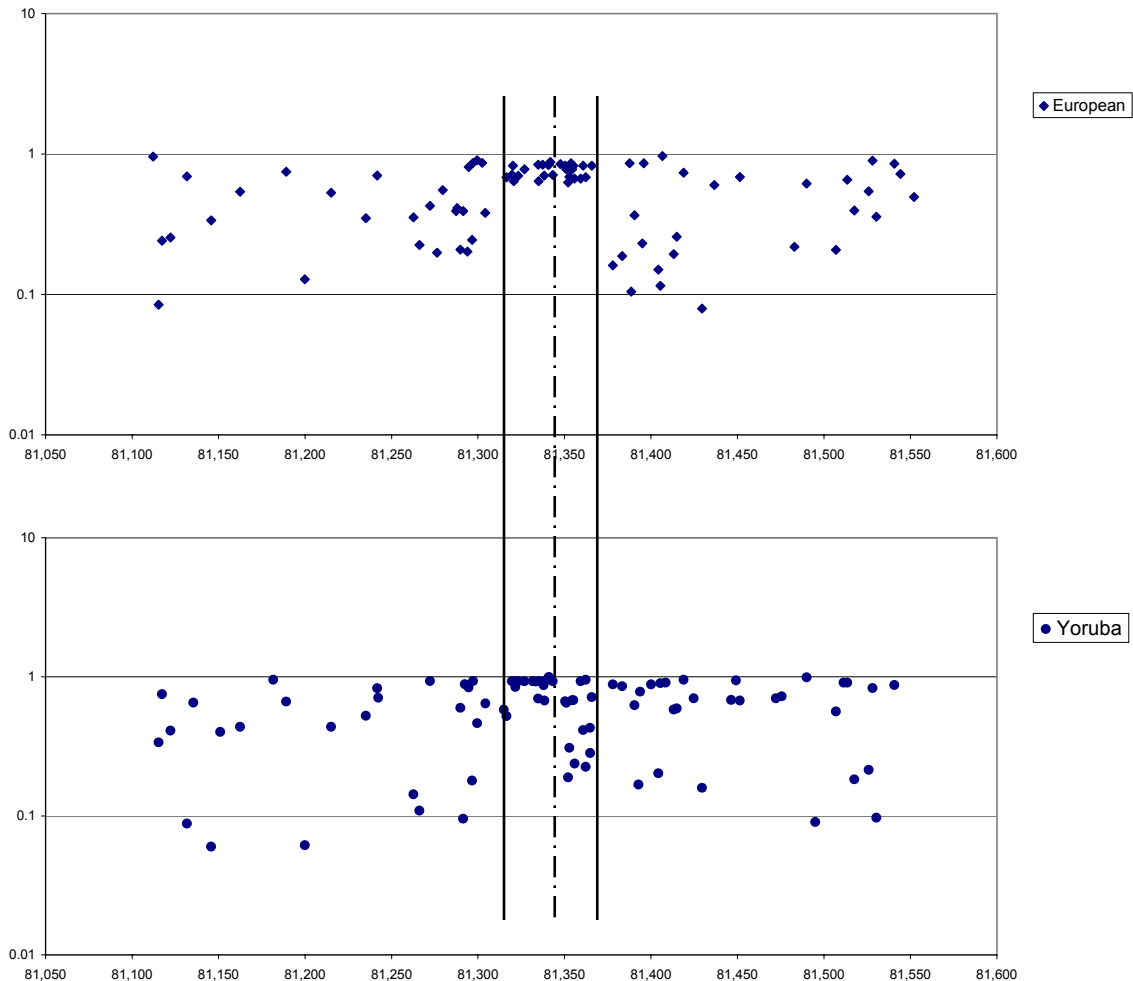
Importantly, our result is based on simultaneous linkages of arbitrary numbers of SNPs, compared with the results of Gabriel et. al., who were able to account only for pairwise correlations (using LD). The following graph shows the distribution of the number of

simultaneous SNPs participating in patterns. The median number of markers per pattern is 96.



Comparison of European and Yoruban Populations

Gabriel et. al. showed that the Yoruban population has haplotype blocks of significantly smaller size. They conclude that the European (and Asian) populations have undergone an evolutionary bottleneck more recently, and thus the haplotype blocks haven't been eroded by recombination to the same extent as the African populations.



We see clear evidence for this effect in the haplotype block in Chromosome 14. The upper graph is a pattern-based marker frequency distribution for the European population. The lower graph is the corresponding one for the Yoruban population. The blocks indeed correspond to each other, but the Yoruban block is about half the size. This size difference is again consistent with the size distributions shown in the first figure taken from Gabriel et. al. Interestingly, erosion seems to have taken place preferentially from the right-hand end of the block.

Conclusion

This work demonstrates that Cira's pattern discovery technology can operate on genomic SNP data, and that it can recover characteristics of genomic variation, namely that the diversity of genetic polymorphisms tends to travel in haplotype blocks. Further, pattern



discovery can accomplish this in a statistically unbiased and computationally efficient way. We again note that this work is a demonstration only. In the interest of time we have analyzed only a portion of the data available in the Gabriel et. al. study. It would be possible however to extend this result straightforwardly to the entire genome.

It is important to reiterate that the methods we have used for this demonstration are far from optimized. We have done a simple re-coding of the genetic polymorphism data for categorical pattern discovery. We have not utilized any of the methods we have under development for parallelization of the discovery step, nor for transformations of dense SNP marker data into a sparser representation. These methodologies will allow far more detailed discovery and representation of patterns of polymorphism. Nevertheless we have shown that even at the level at which both the technology and the data exist today, important results can be obtained.

We expect that clinical association data will become available shortly via collaborations. This will put us in a position to translate this technology demonstration into a novel computational approach with a real commercial endpoint.

References:

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. 2002. The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229.