Discovering Haplotype Blocks in the Human Genome

Alessandro Rinaldo[1], Bernie Devlin[2], Larry Wasserman[1], Kathryn Roeder[1]

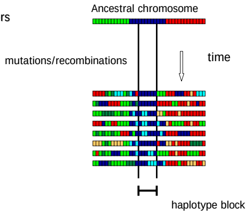
Carnegie Mellon

http://www.stat.cmu.edu/~arinaldo

[1] Department of Statistics, Carnegie Mellon University [2] Department of Psychiatry, University of Pittsburgh

Introduction

A **haplotype block** is a set of closely linked alleles/markers on a chromosome that, over evolutionary time, tend to be inherited together.



Accurate identification of haplotype blocks is important because these data will:

1. produce new insights into the physical and evolutionary dynamics of human chromosomes.
2. prove useful for the genetic analysis of complex diseases.

Goals

❶ Provide a statistical characterization of haplotype blocks

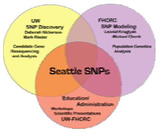
❷ Develop a methodology to identify LONG haplotype blocks, which are apparently regions of evolutionary conservation

Data

Seattle SNPs http://pga.gs.washington.edu/

The data used in the analysis are freely available from :

The UW-FHCRC Variation Discovery Resource (SeattleSNPs).



1. The analysis considered only “fully sequenced” genes (at least 90% of the gene).
2. SNPs with minor allele frequency less than 10% and/or not in Hardy-Weinberg equilibrium were discarded.
3. Reported results are for the **African American sample (24 subjects)**.

The haplotype sequences were estimated from the genotypes using PHASE:

http://www.stat.washington.edu/stephens/phase.html

Statistical Formalization

Each SNP i is a Bernoulli random variable X_i with some probability p > 50% of producing the major allele.

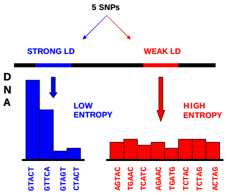
The haplotype corresponding to m SNPs is the binary random vector $\underline{X}=\{X_1,\dots,X_m\}$.

A haplotype block is a subset of $\{X_1,\dots,X_m\}$ formed by contiguous SNPs exhibiting strong linkage disequilibrium.

In statistical terms, those SNPs possess a **high multivariate correlation**.

Entropy as a Measure of LD

Traditional measures of linkage disequilibrium only consider **pair-wise comparisons** between SNPs.



Shannon’s Entropy provides a **multivariate** measure of linkage disequilibrium among SNPs.

Let $\underline{X}$ be a random vector taking on k values with probabilities p1,...,pk. The **entropy of $\underline{X}$** is:

$$H(\underline{X})=-\sum_{i=1}^{k} p_i \log_2(p_i)$$

The *normalized entropy* and model selection procedure

Rather than using the raw entropy, the following entropy-based measure of LD seems to be more interpretable: the ***normalized entropy***

$$0\leq 1-\frac{H(\underline{X})}{\sum_i H(X_i)}<1$$

Values of the normalized entropy close to 0 indicate lack of LD.

Values close to 1 are observed in presence of strong LD.

Regions containing SNPs with high values of the normalized entropy are likely to form a block or be part of a block

Identification of haplotype blocks is achieved through a model selection procedure.

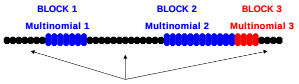
The proposed algorithm allows for partitions (blocking schemes) of the gene into sub-regions that are classified as either in-linkage disequilibrium (blocks) or in-linkage equilibrium.

Blocks are modeled as **independent Multinomial binary vectors** and SNPs not included in blocks are modeled like **independent strings of Bernoullis**.

Each partition of the SNPs into blocked and non-blocked regions is treated as a descriptive model of the genomic region. The optimal blocking scheme is selected as the solution to an **AIC-type model selection procedure.** A penalty is imposed for the model complexity. The loss function to minimize is:

$$-2\lambda(\hat{\vartheta})+2d$$

The parameter d measures the statistical complexity of the model and can be related to the minimal number of mutations leading to the observed blocks.



NON-BLOCKED REGIONS: Independent Bernoullis

The number of possible blocking schemes grows very fast with the number of SNPs. The complexity of the searching procedure was reduced by designing a greedy iterative algorithm that accounts only for partitions corresponding to regions with high values of the *normalized entropy*.

Specifically, each set of contiguous SNPs with *normalized entropies* above a predefined threshold are treated as a unique block. Different values of the threshold correspond to different blocking schemes. Therefore, finding the optimal partition is equivalent to choosing an optimal value of the threshold.

An approximate solution is obtained from the greedy algorithm. This solution is then perturbed locally to refine the estimated block structure

Some results

Haplotypes

🟦 Major allele 🟨 Minor allele

Pictures produced using the software **VH1**, available at the Seattle SNPs website.

Correlation matrix

0 1

Absolute value of the correlation matrix.

Normalized Entropy

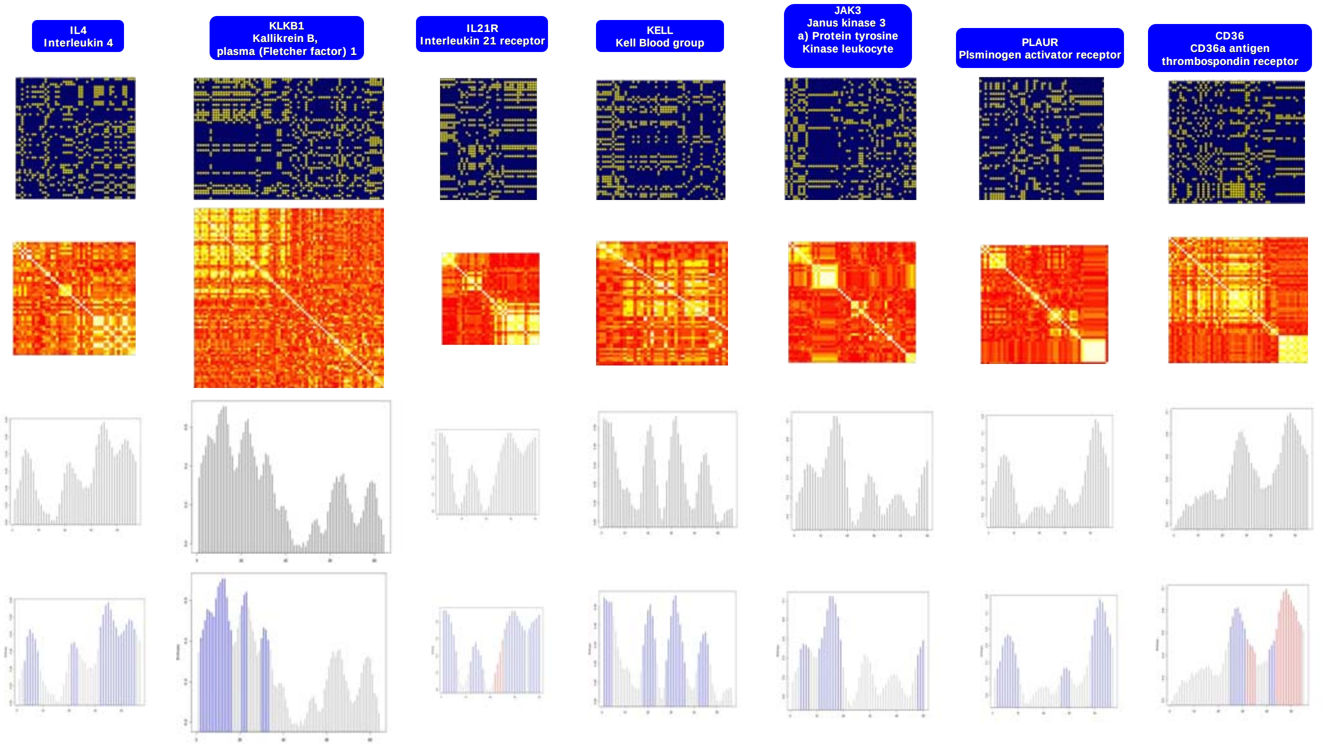
Smoothed normalized entropy over a sliding windows of 5 SNPs.

SNPs are displayed equally spaced

Haplotpye Blocks

Contiguous blocks are displayed in blue and red.

🟦 🟥



IL4 Interleukin 4

KLKB1 Kallikrein B, plasma (Fletcher factor) 1

IL21R Interleukin 21 receptor

KELL Kell Blood group

JAK3 Janus kinase 3 a) Protein tyrosine Kinase leukocyte

PLAUR Plsminogen activator receptor

CD36 CD36a antigen thrombospondin receptor