ch1. LD

ch2. Tagging SNPs

ch3. Haplotype Phasing

GWAS

GWAS ≡ Genome-wide Association studies

LD two measures:

① D, D' = normalized

$-1 \leq D' \leq +1$

$0 \leq |D'| \leq 1$

② r^2, Note $\sqrt{r^2}$ = correlation coefficient

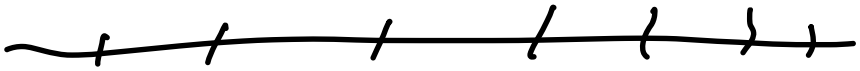
Do not confuse r^2 with $(r)^2$

↑ symbolic historic

not the "r" recomb. fraction

Ch2. Tagging SNPs

Genome — 10 mill



Data Compression Pb — 10 mill → 0.5 Mil

The most informative SNPs

tSNPs — tagging SNPs "captain"

2 Algorithms for SNP selection to put them on a SNP chip

1000 colors for 1000 people in Africa founding population

GWAS ≡ hypothesis-free association

DIRECT and INDIRECT assaying

HAP MAP PROJECT

ALGORITHM: LD-Select

Haplotyper

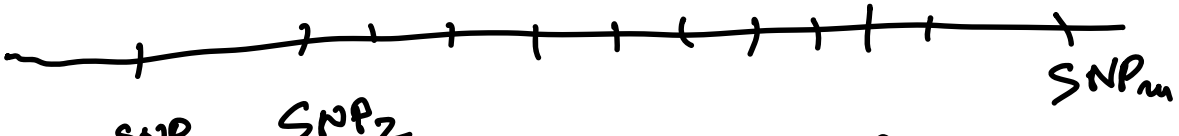
Preferred algorithm of HAPMAP scientists

based on r^2

r^2 = favorite LD measure for gene associations.

LD-Select Algorithm

"Most informative SNPs"



genomic region

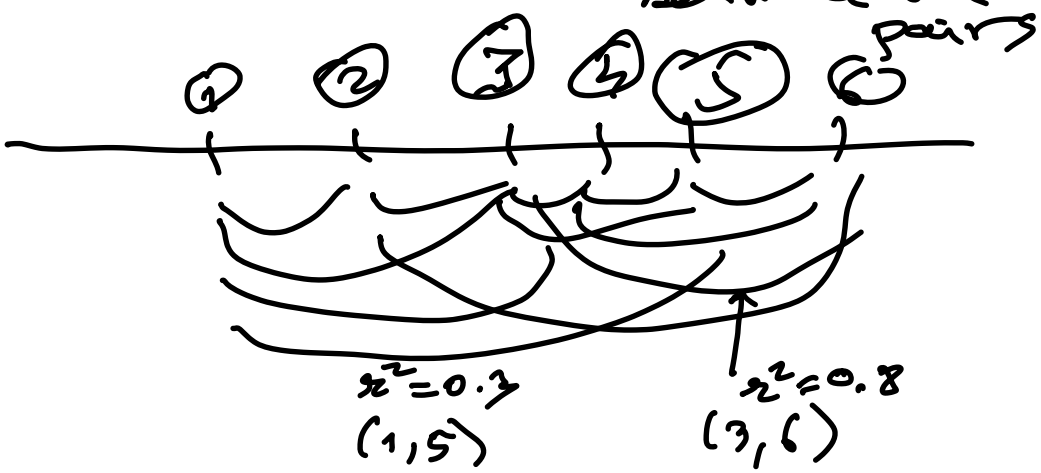
Minimum # of SNPs of max information

"Capturing genomic variation"

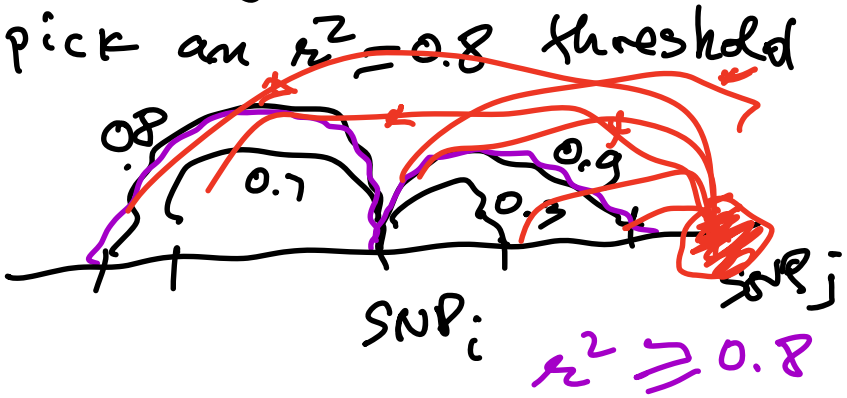
± SNPs: how many haplotypes you "capture" w.r.t a subset of SNPs.

LD-Select intuition

pairwise r^2 between all pairs



Def. "Most informative"
pick an $r^2 = 0.8$ threshold



Most informative ≡ SNP with the !

Largest # of other SNPs in LD of $r^2 \geq 0.8$ with it

Algorithm:

Pick the most informative SNP.

Remove it, and remove all the SNPs at LD $r^2 \geq 0.8$ to it

Pick the most informative SNP from the remaining collection, and remove also all the SNPs in its $r^2 \geq 0.8$ neighborhood

common SNPs = HAP MAP
rare SNPs

There was a hypothesis: Foundational to HapMap

CDCV: Common Disease – Common Variants

NOT VALID

Common & Rare Variant for Complex Disease

Common SNPs:

a SNP is common if its minor allele freq (MAF) is > 5%.

LD-Select: Initially we remove from input the SNPs with MAF ≥ 10%

only Common SNPs.

- Average gene in the human genome ~ 27 Kb
- ~ 50 common SNPs in any gene on average

We would like to select a set of maximally informative set of common SNPs for disease association analysis

tag SNPs or tSNPs

Conditions / Desiderata for this SNP selection?

1) Resolving common haplotypes

"resolve" ≡ "capture"

HAPMAP is about common variation

2) Minimize the number of tagging SNPs

3) Maximize haplotype information content

SNPs at < 10 kb apart tend to be correlated

LD describes this relationship

D' and r^2

$|D'|$

Def

complete LD: $|D'| = 1$

Perfect LD: $r^2 = 1$

$|D'|=1$ complete LD is when the allelic association is as strong as possible given the allele frequencies of the two SNPs.

- if neither site has experienced recurrent mutations or gene conversions and if there was no recombination between the two sites.

- However, genotypes can be perfectly correlated between the two sites only if the MAFs are the same.
  only then $r^2=1$ perfect LD

r^2 as an LD measure satisfies the "Interpretability of Intermediate Values" axiom:

What this means for r^2 is the following:

THEOREM

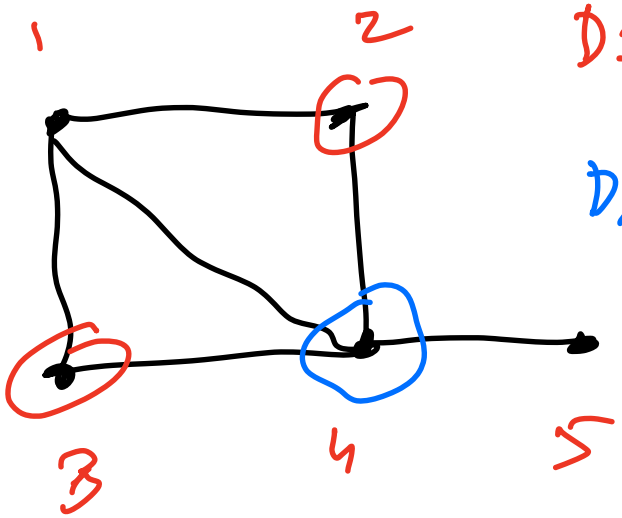
The power to detect the disease associated polymorphism indirectly in N samples is equivalent to the power to detect it directly in Nr^2 samples.

Pritchard & Przeworski 2001

GRAPH THEORY

undirected graphs

Def A **dominating set** for a graph $G=(V,E)$ V = vertices, E = edges is a subset D of V, $D \subseteq V$, such that every vertex not in D has at least one edge to a vertex in D.



$D=\{2,3\}$

$D_1=\{4\}$

Minimum dominating set

Dominating Set is the formal optimization function objective of LD-Select

The set of tSNP = maximum of LD select is a dominating Set for the graph with the SNPs as nodes, and all pairwise edges weighted by the r^2 value.

A greedy algorithm for finding the minimum dominating set