AFFYMETRIX

The Way Ahead.™

LdCompare: rapid computation of single- and multiple-marker r^2 and genetic coverage

Ke Hao
Algorithm and Data Analysis
Affymetrix, Inc.

AFFYMETRIX

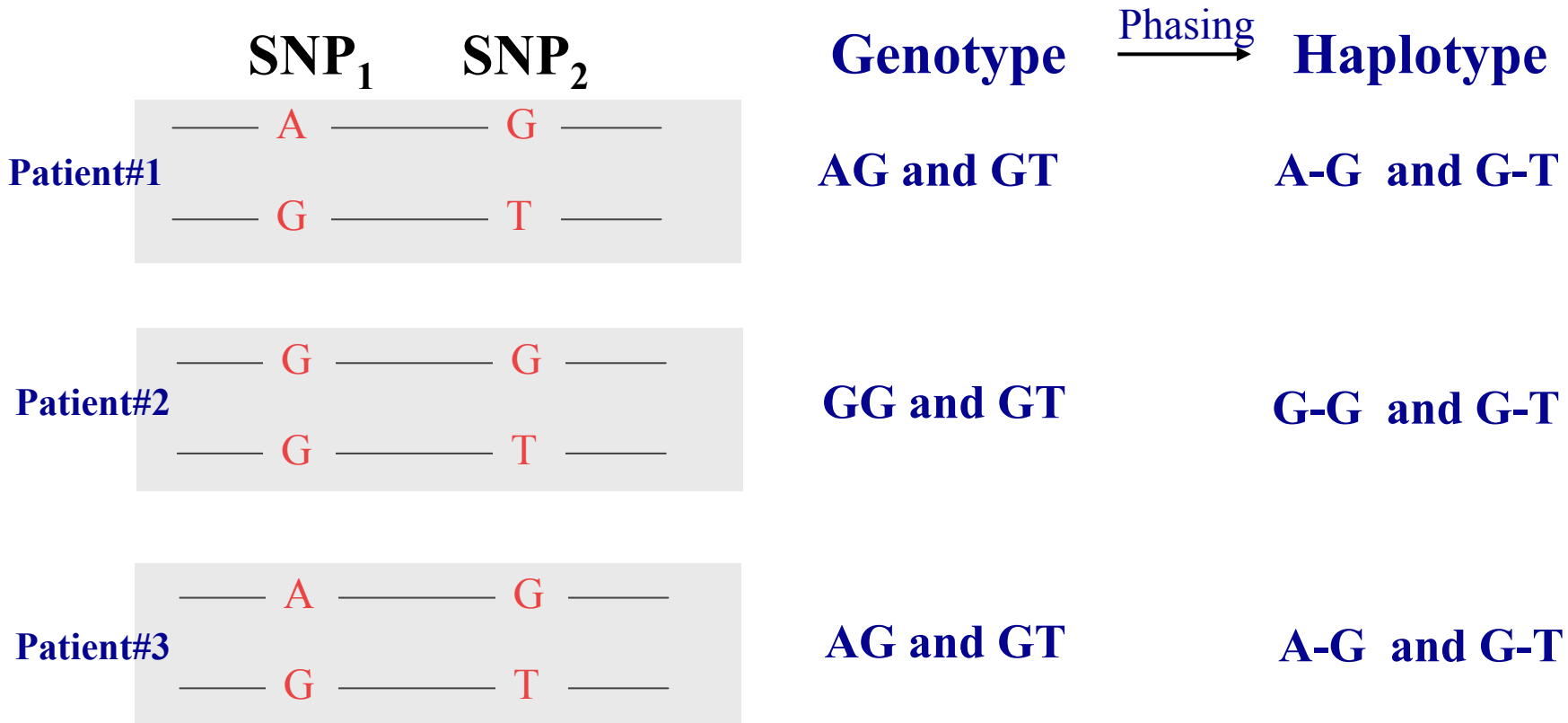
The Way Ahead.™

Outline

- **Concepts and Background**
  - Linkage Disequilibrium (LD) and Measurements
- **Genetic Coverage of SNP Panels**
  - Single- and Multiple- Marker Coverage
  - Association Studies and Common Disease Common Variant Theory (CDCV)
  - Tag SNP Selection
  - Software: LDCompare

AFFYMETRIX
The Way Ahead.™

Genotypes and Haplotypes

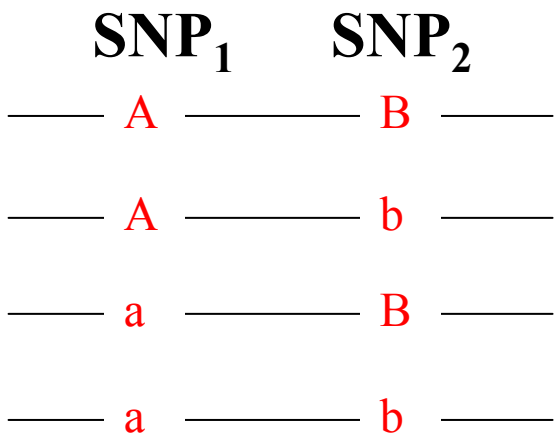


AFFYMETRIX
The Way Ahead.™

Linkage Disequilibrium (LD)

- **The nonrandom correlation between genetic markers (e.g. SNP).**
- **Certain haplotype has higher (or lower) frequency in human population than random chance.**

Four Possible Combinations (Haplotypes)

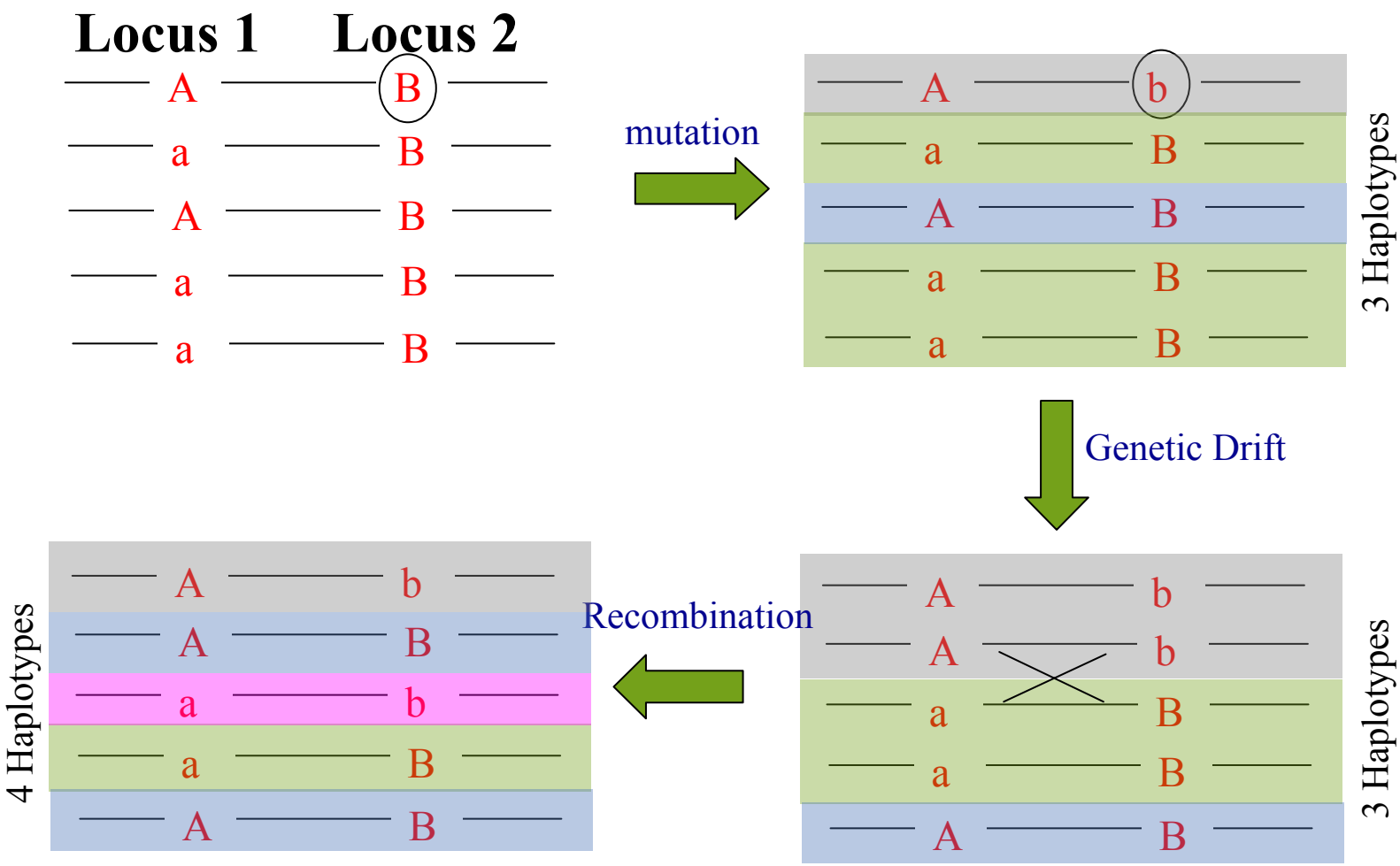


$P_{AB} \neq P_A \times P_B \Leftrightarrow$ Linkage Disequilibrium

$P_{AB} = P_A \times P_B \Leftrightarrow$ Linkage Equilibrium

AFFYMETRIX
The Way Ahead.™

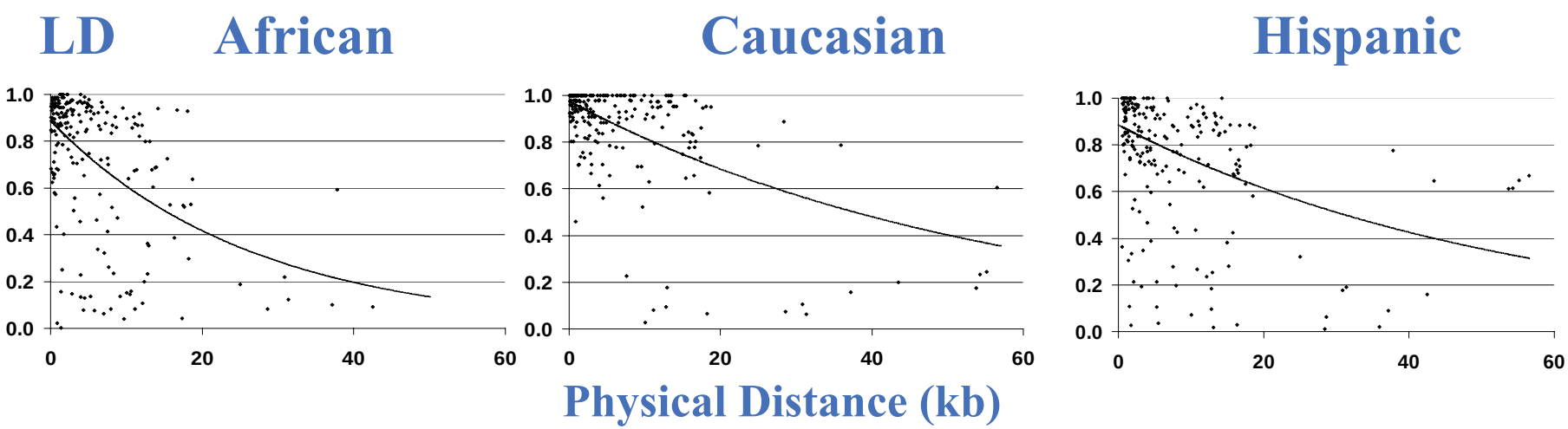
Origin of LD



AFFYMETRIX
The Way Ahead.™

Origin of LD

- **Small number of historical recombination events**
  - **Tightly linked markers**
  - **Short history of human population**



Hao K. et al, Hum Mol Genet. 2004 Apr 1;13(7):683-91

AFFYMETRIX
The Way Ahead.™

Measurements of LD

- **LD:** $D = P_{AB} - P_A \times P_B$

  Positive LD vs. Negative LD

  Measurement affected by P_A and P_B

- **LD'** (used in population genetics studies): $D' = |D| / D_{max}$

  D' =1 <=> "Complete LD"

  <=> Only ≤3 haplotypes exist

- **r^2,** often used in association studies. (1) sample size effect and (2) direct link to statistical power

$$r^2 = \frac{(P_{AB}P_{ab} - P_{aB}P_{Ab})^2}{P_A P_a P_B P_b}$$

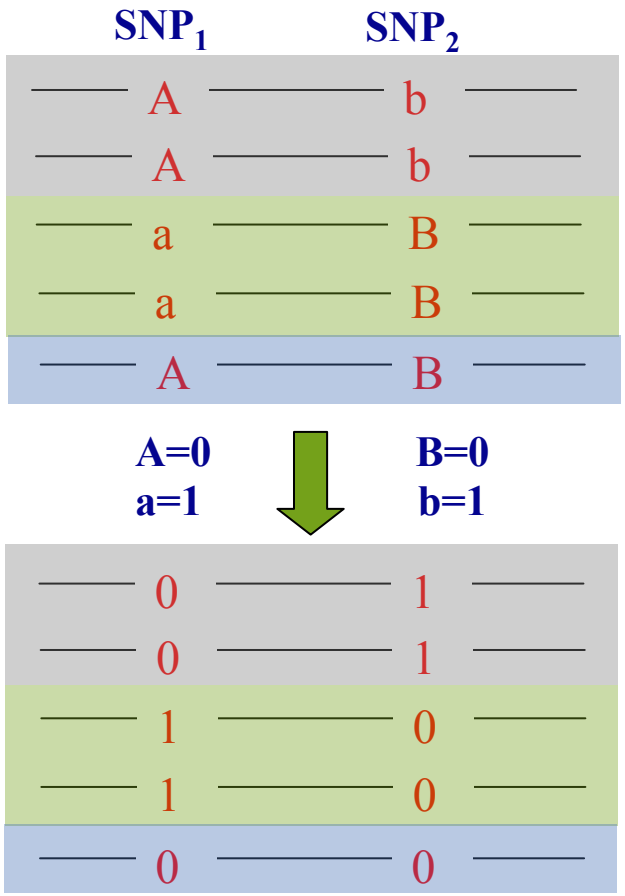
r^2 =1 <=> "Perfect LD"

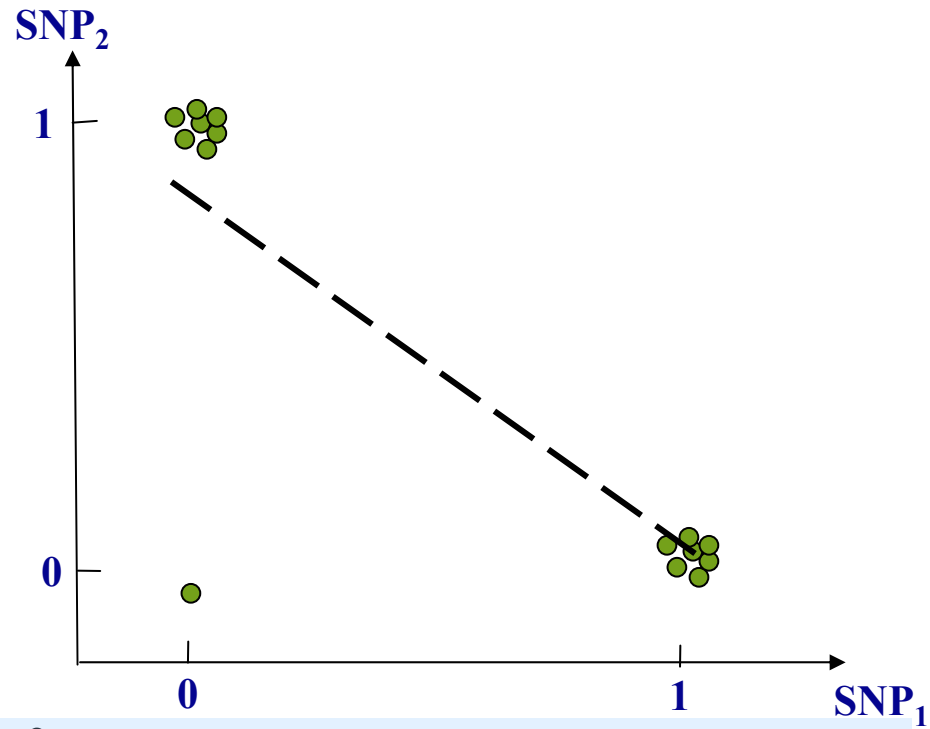
<=> Only 2 haplotypes exist

=> $P_A = P_B$

AFFYMETRIX
The Way Ahead.™

Measurement of LD





$r^2 \rightarrow 1$ $\Rightarrow$ SNP_1 covers SNP_2

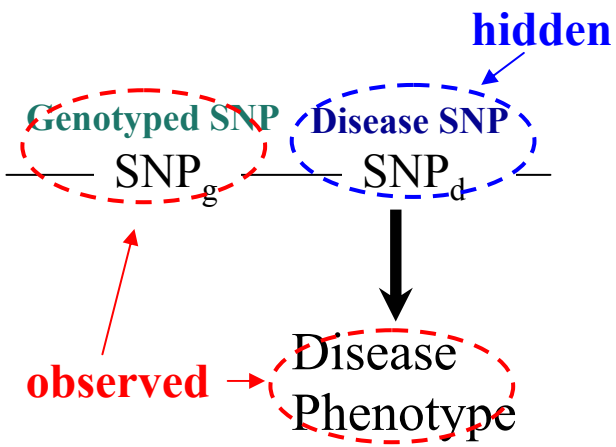
$\Rightarrow$ SNP_2 covers SNP_1

$\Rightarrow$ $MAF_{SNP1} \approx MAF_{SNP2}$

AFFYMETRIX
The Way Ahead.™

Genetic Association Study

- Goal: identify genes underlying human diseases
- Method: detect correlation between phenotypic trait and marker genotypes



Detecting Corr(SNP$_g$, Disease)

⇔ Detecting Corr(SNP$_g$, SNP$_d$)

LD

AFFYMETRIX
The Way Ahead.™

Power of Association Test

Statistical Power

Nature of Disease and Study Design

- Disease prevalence
- Sample size
- Effect size (penetrance)
- Disease allele frequency

Characteristics of Genotyping Panel

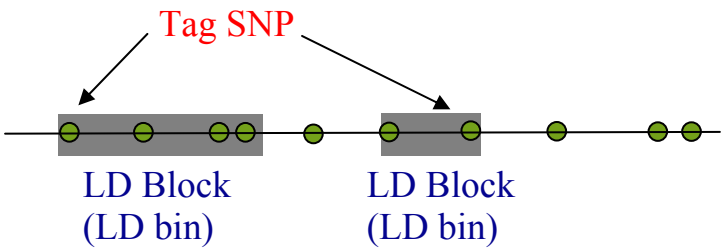
- Genetic coverage
- Genotyping call rate
- Genotyping accuracy

AFFYMETRIX
The Way Ahead.™

Outline

- **Concepts and Background**
  - Linkage Disequilibrium (LD) and Measurements
- **Genetic Coverage of SNP Panels**
  - Single- and Multiple- Marker Coverage
  - Association Studies and Common Disease Common Variant Theory (CDCV)
  - Tag SNP Selection
  - Software: LDCompare

LD Block and Genetic Coverage



LD block: a group of nearby markers with strong pair-wise LD

SNPs Directly Typed: 4

SNPs Covered by LD: 4

Coverage = 8/10

SNPs Directly Typed: 4

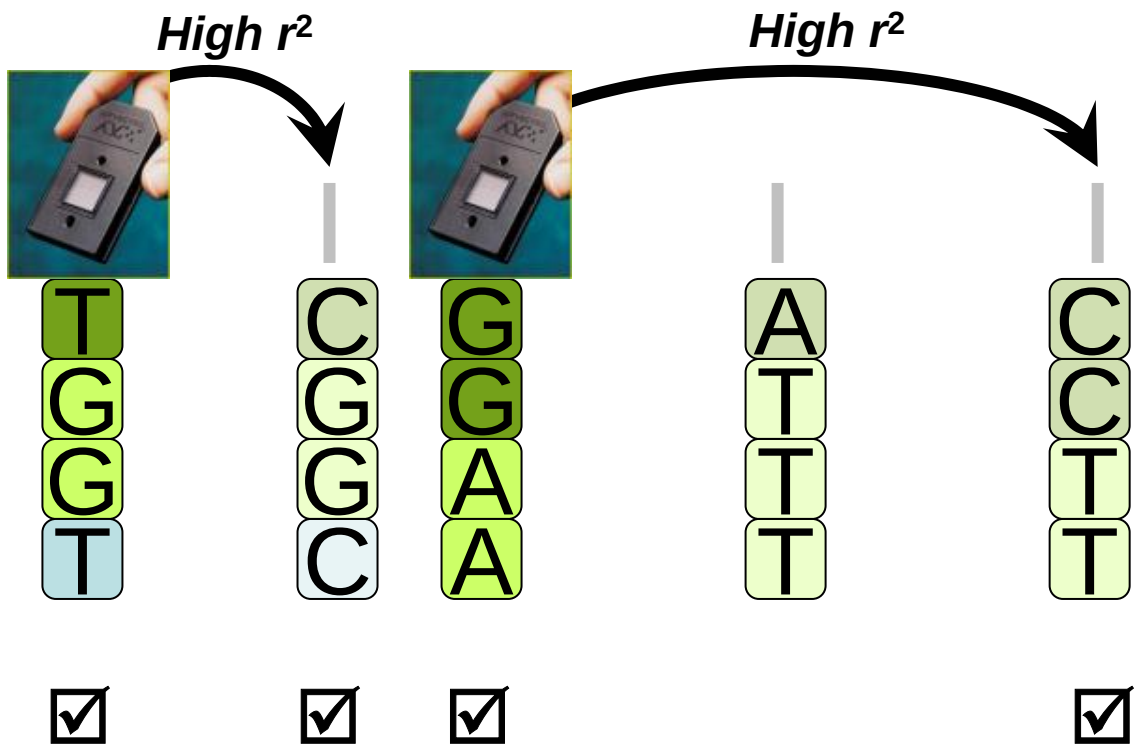
SNPs Covered by LD: 1

Coverage = 5/10

AFFYMETRIX
The Way Ahead.™

Association tests with fixed markers:

Typing a subset of SNPs captures many



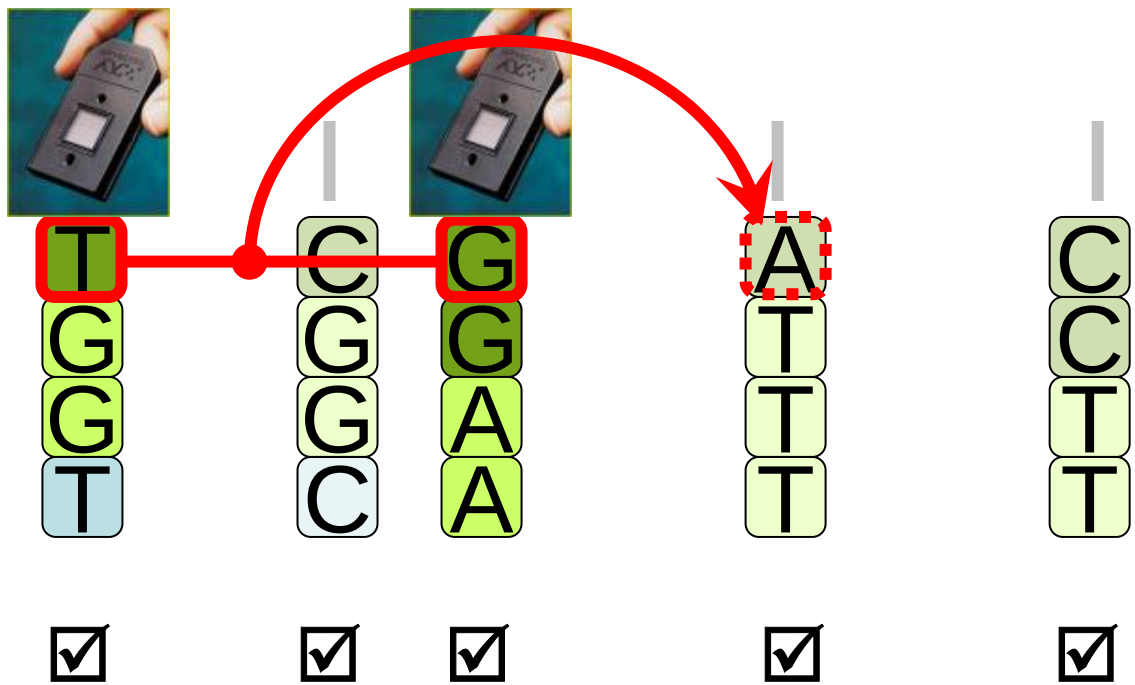
SNPs captured: ☑ ☑ ☑ ☑

Courtesy of Itsik Pe'er, Broad Institute

AFFYMETRIX

The Way Ahead.™

Improving coverage using HapMap data
Multi-marker approaches

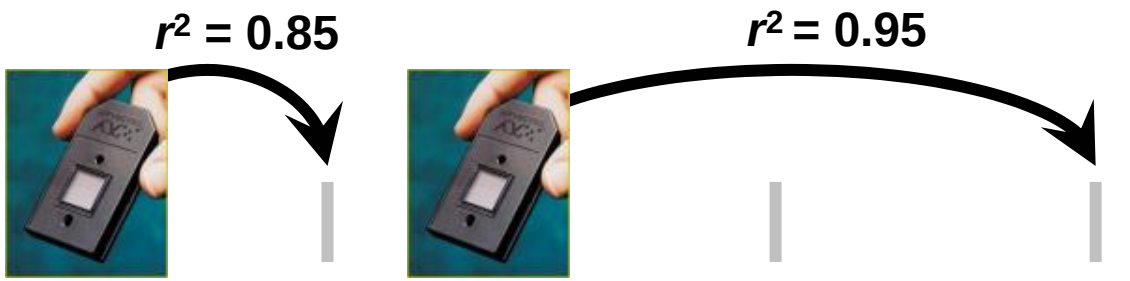


SNPs captured: ☑ ☑ ☑ ☑ ☑

Courtesy of Itsik Pe'er, Broad Institute

AFFYMETRIX
The Way Ahead.™

r^2 Cutoff

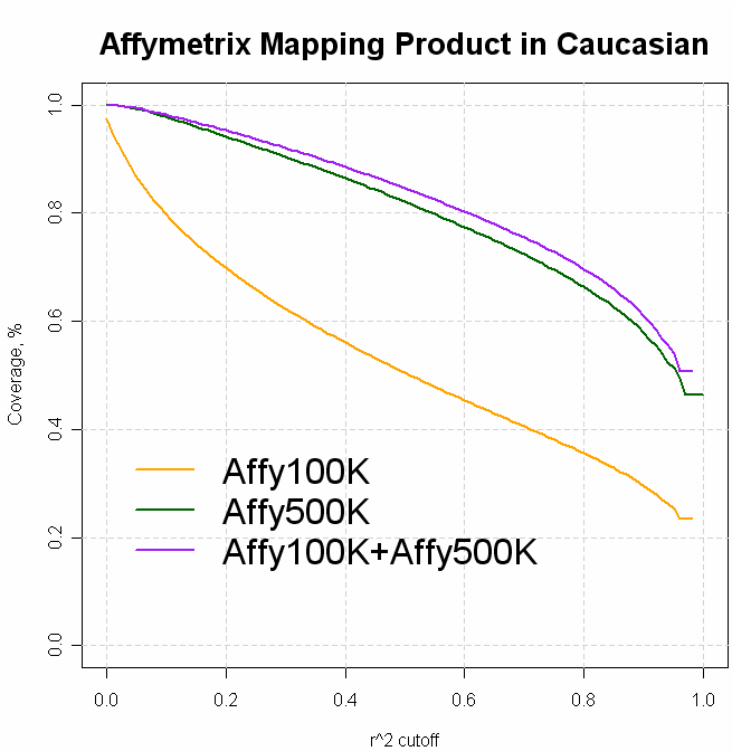


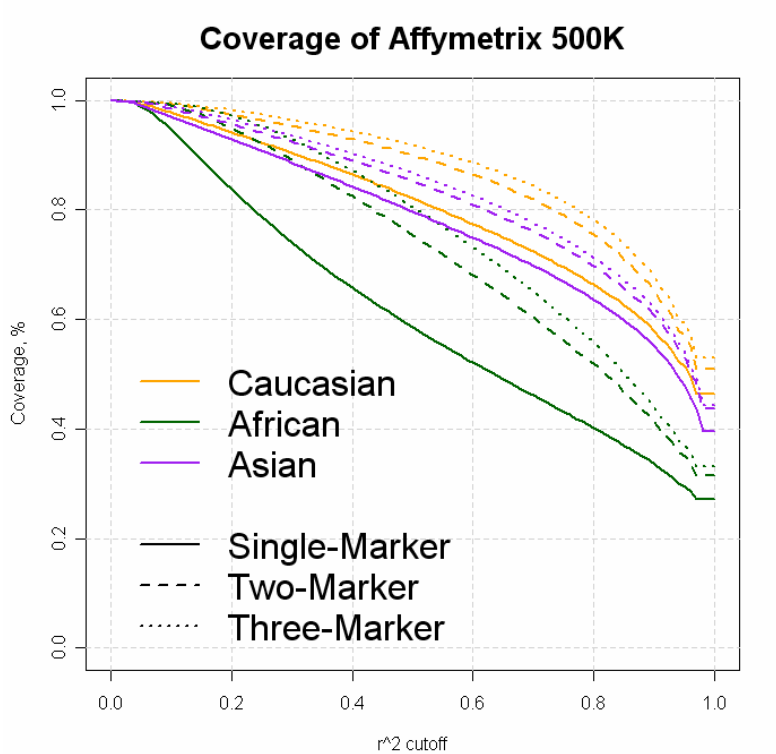
r^2_{cutoff}=0.8	☑	☑	☑	☑
r^2_{cutoff}=0.9	☑		☑	☑

Courtesy of Itsik Pe'er, Broad Institute

AFFYMETRIX
The Way Ahead.™

Genetic Coverage





Focus on Common SNPs (MAF≥5%)

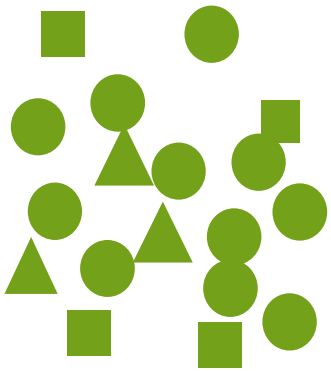
AFFYMETRIX
The Way Ahead.™

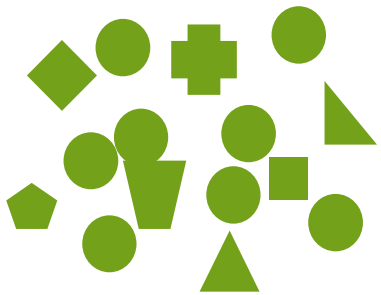
Common Disease Common Variant (CDCV) Theory

CDCV

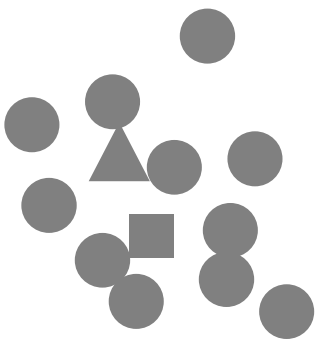
Common Disease Rare Variants

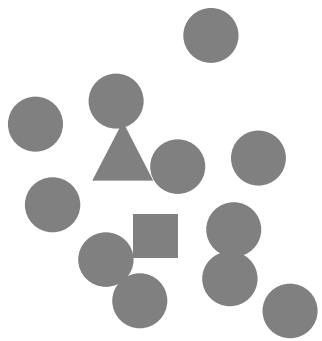
Cases



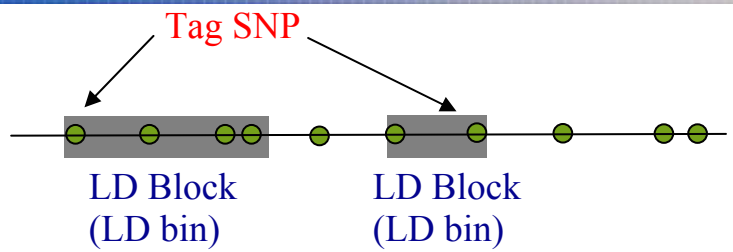


Controls





Tag SNP Selection

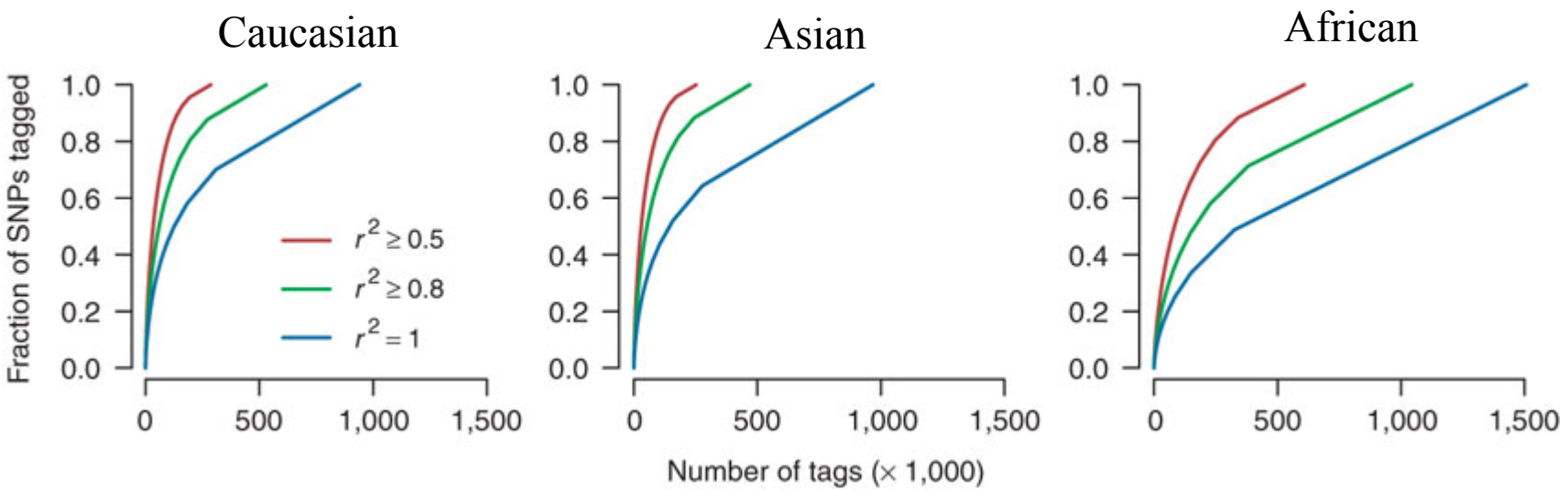


Caucasian & Asian; $r^2_{cutoff} = 0.8$

- Rapid coverage gain at early stage
- Singleton point at ~300K tag SNPs
- 500~600K tag SNPs => 100% coverage

African; $r^2_{cutoff} = 0.8$

- Slow gain (compare to Caucasian and Asian) at early stage
- Singleton point at ~450K tag SNPs
- ~1M tag SNPs => 100% coverage



Barrett and Cardon, Nature Genetics. 2006 May; Advance Online Publication

AFFYMETRIX
The Way Ahead.™

LDCompare: Motivations

- The scale of genetic-variation datasets has increased enormously
- Efficiently characterizing the LD structure of large number of SNPs remains a challenge
- Multiple-marker coverage may further enhance power of association studies, however, no program can compute three-marker coverage at a scale of HapMap II
- Single-marker r^2 ~ 10^8 evaluations (100K Sliding Windows)
- Multiple-marker r^2 ~ 10^{13} evaluations in three-marker scenario (100K sliding windows). Computation burden grows exponentially along with increasing sliding window size

AFFYMETRIX
The Way Ahead.™

LDCompare: Features (1)

- The program has a uniform framework for single- and multiple-marker modes.
- Automatically detect the running mode according to parameter file setting.
- Both diploid genotypes and phased haplotypes can be accommodated.
- Pairwise r^2 and single-marker coverage can be computed in either case; multiple-marker r^2 and coverage requires phased haplotype data.
- Standard linkage format input files are used for diploid data. Haplotype data is also accommodated in a straightforward input file format.

AFFYMETRIX
The Way Ahead.™

LDCompare: Features (2)

- LdCompare is written in ANSI C++ and is usable on most operating systems.
- Running on a 2.8GHz Intel Xeon workstation with 1G RAM, it computes the CEU HapMap II pairwise r^2 (±100K bp sliding window size) within 2 hours.
- Two marker r^2 and coverage take about 12 hours on a single CPU.
- We have run three-marker coverage is run on a Linux cluster splitting jobs by chromosome, the largest finishes within 24 hours.
- The number of evaluations rise exponentially when we increase the sliding window size. Using ±200K bp sliding window, it takes a week for our Linux cluster to compute three-marker genomic coverage.
- We have compared the single- and two-marker coverage results from LdCompare with those from Haploview and found them to be identical, though LdCompare is significantly faster.

AFFYMETRIX
The Way Ahead.™

LDCompare: Features (3)

- Programs output single- and multiple-marker r^2 and coverage cumulative distribution function (CDF). In addition, users can choose to output tables of all pairwise r^2 values for downstream tag SNP selection, etc.
- A list of user-defined parameters, such as minor allele frequency (MAF) filter, is provided for maximum flexibility.
- Program is written in C++. Users are welcome to modify and redistribute the program under GNU.
- The code is optimized for rapid computation and memory efficiency.
- The program has been developed and tested on Microsoft Windows, Linux and Sun Microsystems Solaris.
- ***http://www.affymetrix.com/support/developer/tools/devnettools.affx***

AFFYMETRIX
The Way Ahead.™

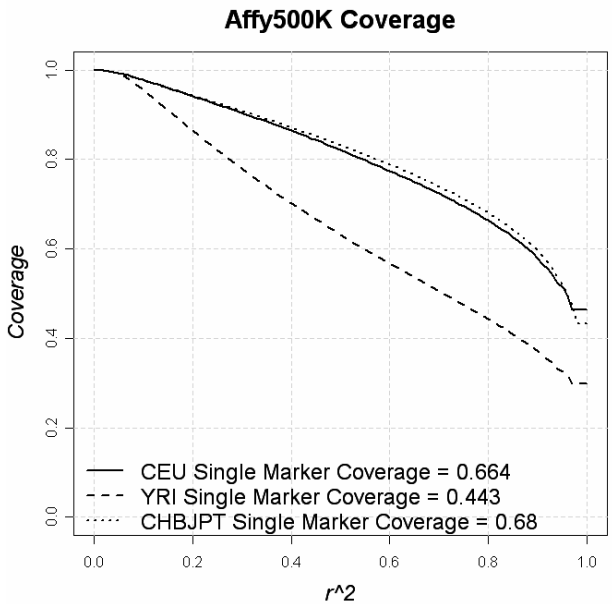
Summary

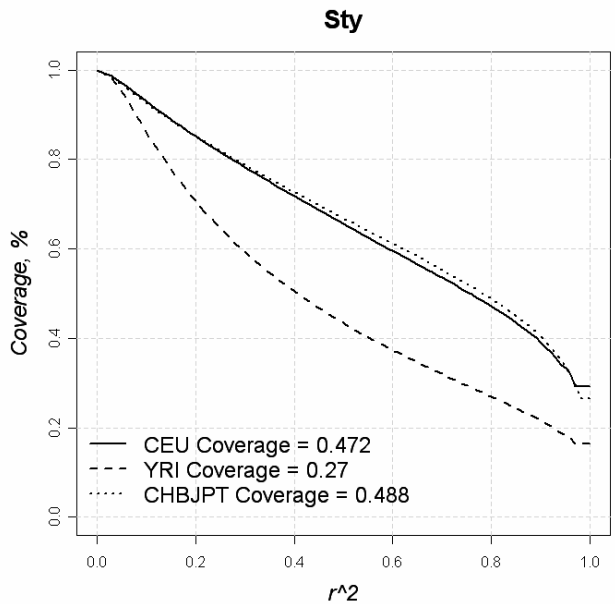
- Linkage disequilibrium is the foundation of genetic association studies
- At the current stage, we focus on common variants
- Single- and multiple- marker coverage is often used in benchmarking SNP genotyping panels
- Computation speed becomes bottleneck with ever increasing SNP number
- We develop LDCompare, a free software, which features fast computation and versatile output.

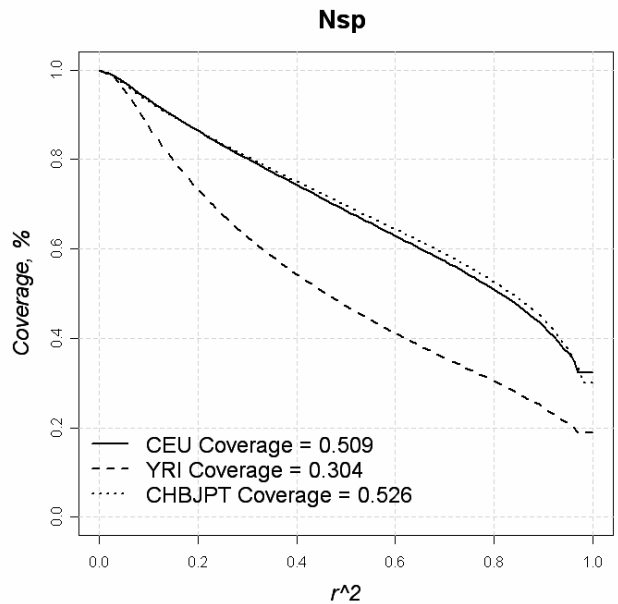
AFFYMETRIX

The Way Ahead.™

Affymetrix Mapping 500K Coverage



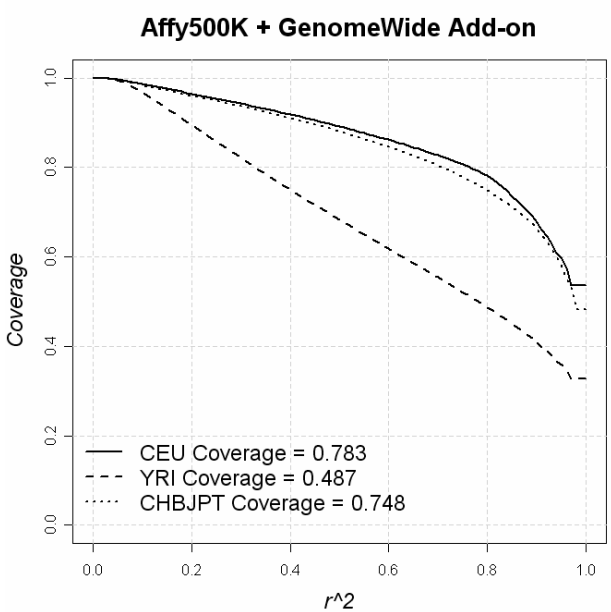


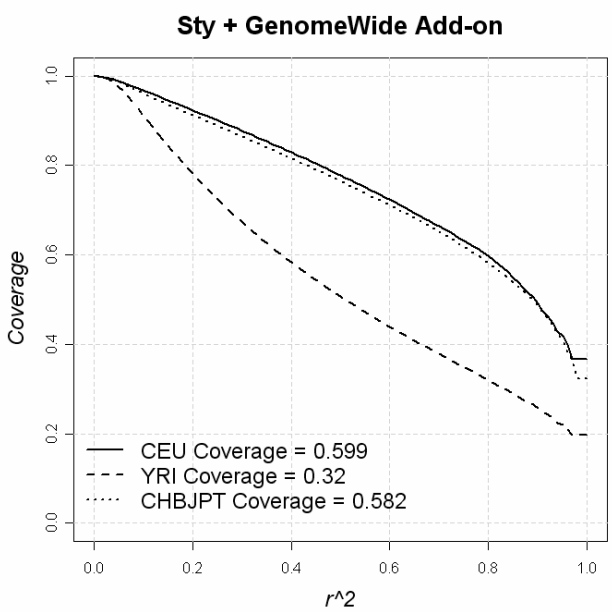


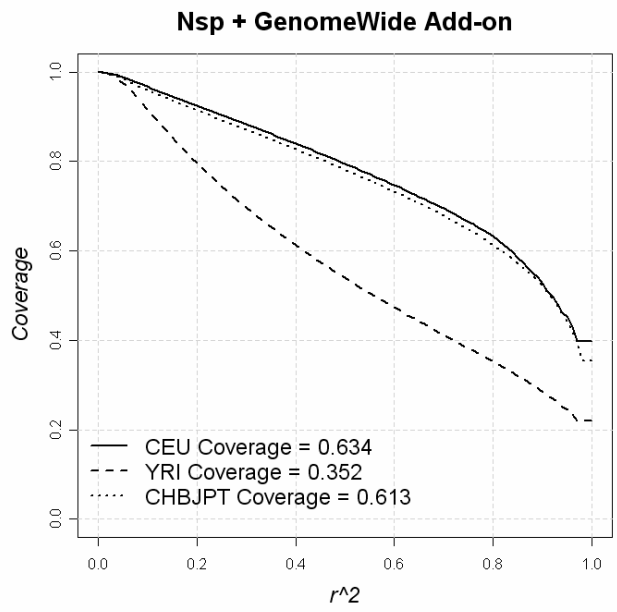
AFFYMETRIX

The Way Ahead.™

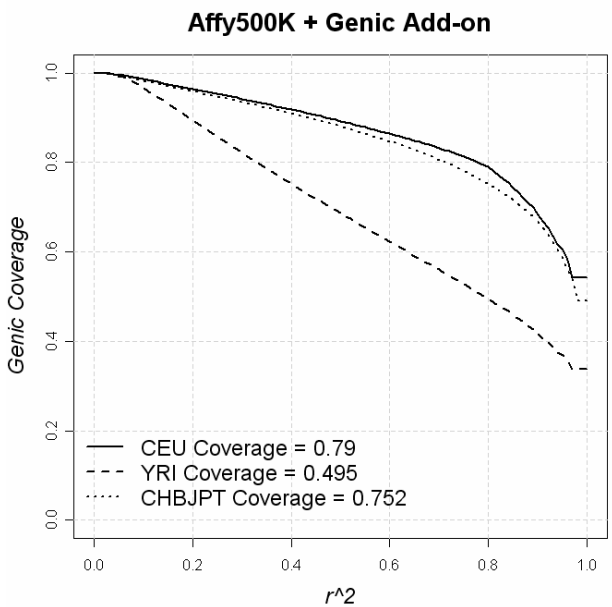
Affymetrix500K + GenomeWide Addon

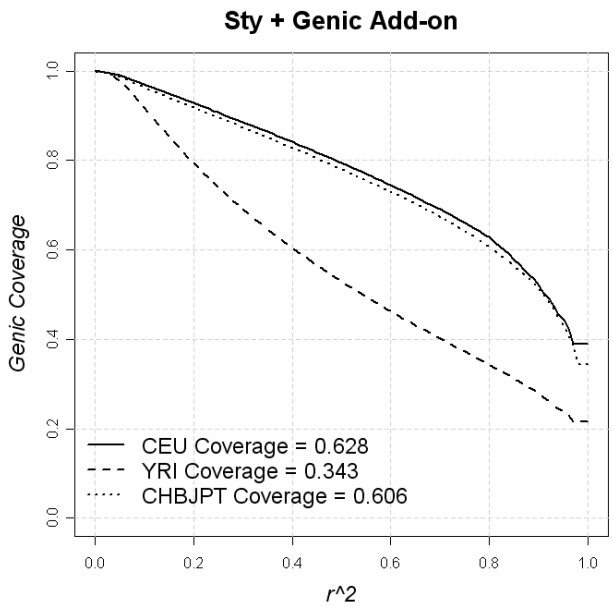


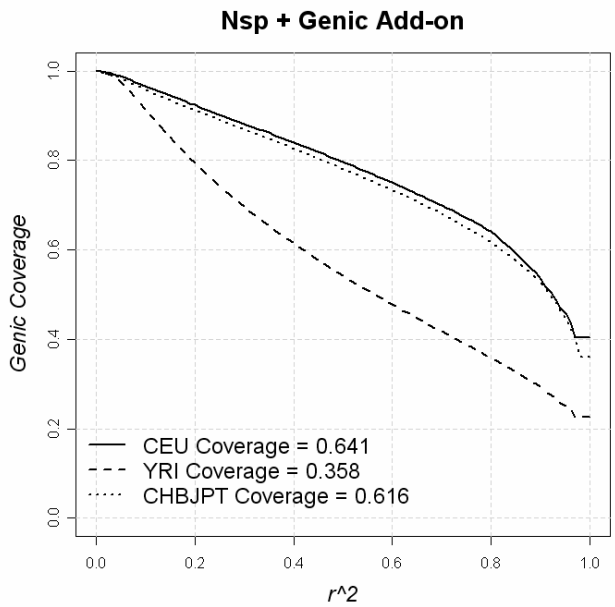




Genic Region Coverage: Affymetrix500K + Genic Addon







A genic region of a gene is defined as the region within 20 Kb upstream of its most 5' boundary, the transcribed regions and the region within 10 Kb of its most 3' boundary (boundaries are indicated in the combined gene databases of Entrez/NCBI and Ensembl). Regions spanning less than 10 Kb between two genic regions are also included in genic regions.