Textile Plot: a new LD display of multiple SNP genotype data

Natsuhiko Kumasaka　Naoyuki Kamatani
Center for Genomic Medicine,
RIKEN, JAPAN

Outline

- What is Linkage Disequilibrium (LD)?
- LD statistics and a LD display
- Textile Plot
- Practical examples

Linkage Disequilibrium (LD)

- Non-independence of alleles at different loci

$$\Pr\{X_1 = x_1, \ldots, X_p = x_p\} \neq \prod_{i=1}^{p} \Pr\{X_i = x_i\}$$

- Power of association study (Pritchard et al. 2001, McVean 2007)
- Historical and biological processes such as mutation, recombination and natural selection (Peterson et al. 1995, Hudson 2001)
- Population admixture (Pritchard et al. 1999)

Pairwise LD statistics for biallelic loci

(Devlin et al 1995, Jorde 2000)

- Covariance

$$\begin{aligned} D &= \mathrm{Cov}(X_1, X_2) \\ &= \Pr\{X_1 = 1, X_2 = 1\} - p_1 p_2 \end{aligned}$$

$$(p_i = \Pr\{X_i = 1\},\ i = 1, 2)$$

- Correlation coefficient (Hill and Robertson 1968)

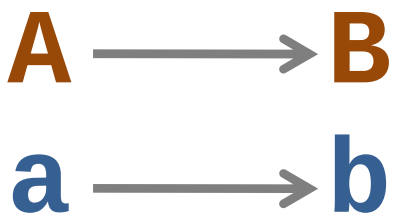
$$r^2 = \mathrm{Corr}(X_1, X_2)^2 = \frac{D^2}{p_1(1-p_1)p_2(1-p_2)}$$

- Lewontin's *D'* (Lewontin 1964)

$$D' = \frac{D}{\min\{p_1(1-p_2), (1-p_1)p_2\}} \quad (D > 0)$$

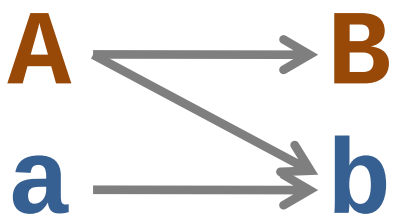
LD Statistics and the Number of Haplotypes

- $r^2 = 1,\ D' = 1$ (Absolute LD)



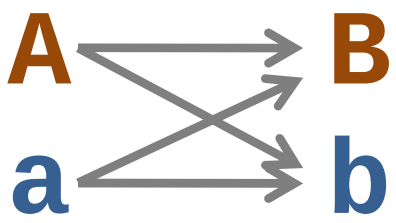
h1: AB
h2: ab

- $r^2 < 1,\ D' = 1$ (Complete LD)



h1: AB
h2: Ab
h3: ab

- $r^2 < 1,\ D' < 1$

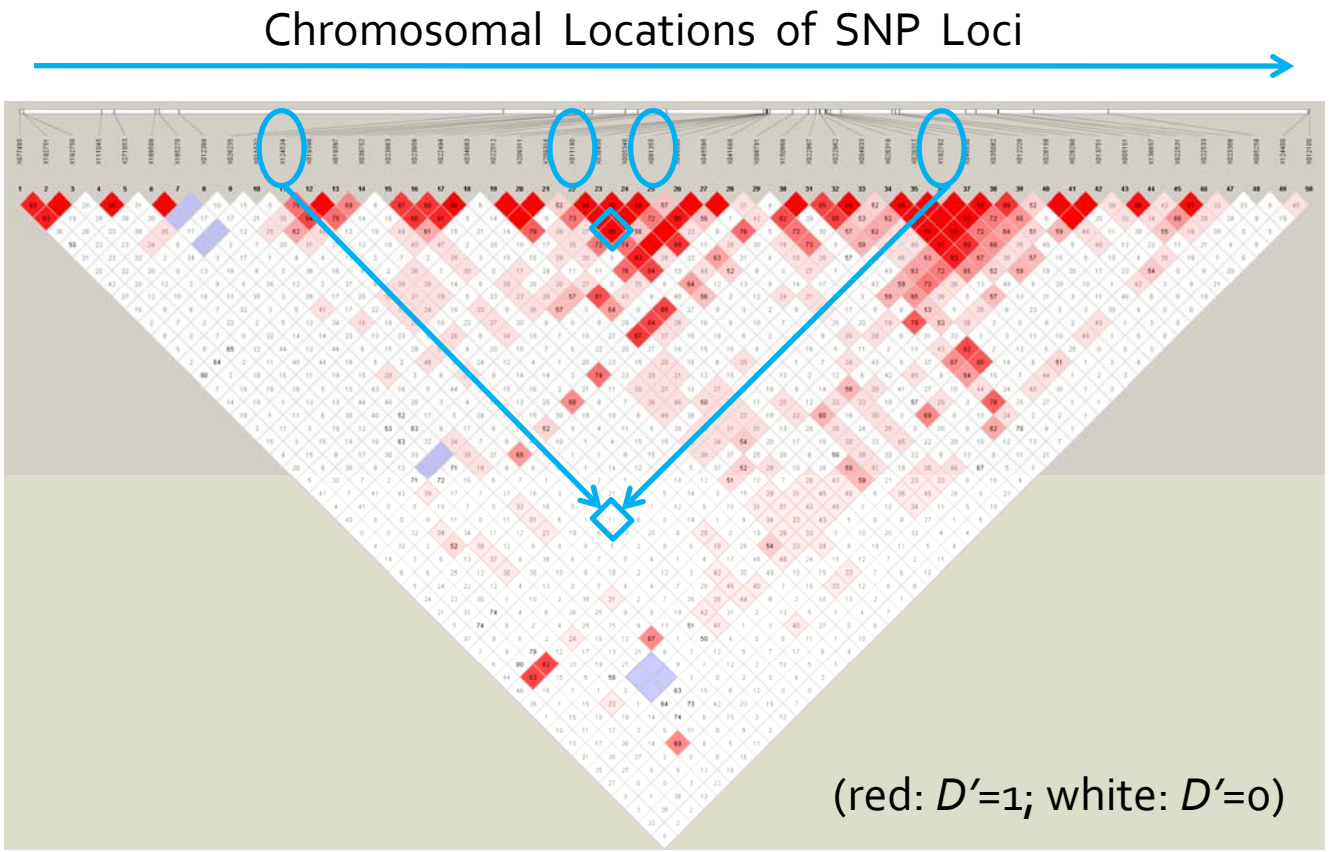


h1: AB
h2: aB
h3: Ab
h4: ab

	L1	L2
Major	A	B
Minor	a	b

HaploView (Barrett et al., 2005)

- Heat map display
- Pairwise LD statistics (r^2 and D')



Inference of Pairwise LD statistics

- SNP genotype of diploid organisms

$$G = X_f + X_m \ (X_f\text{: father's allele, } X_m\text{: mother's allele })$$

- Hardy-Weinberg Equilibrium (HWE)

$$X_f, X_m \overset{i.i.d.}{\sim} \mathrm{Bin}(1, p)$$

- Maximum Likelihood Estimation

$$H_f = (X_f^{(1)}, X_f^{(2)}), H_m = (X_m^{(1)}, X_m^{(2)})$$

$$H_f, H_m \overset{i.i.d.}{\sim} \mathrm{Bin}(1; p_1, p_2, D)$$

$$L(p_1, p_2, D | g_1, \ldots, g_n) = \prod_{i=1}^{n} \Pr(G_i = g_i | p_1, p_2, D) \rightarrow \max$$

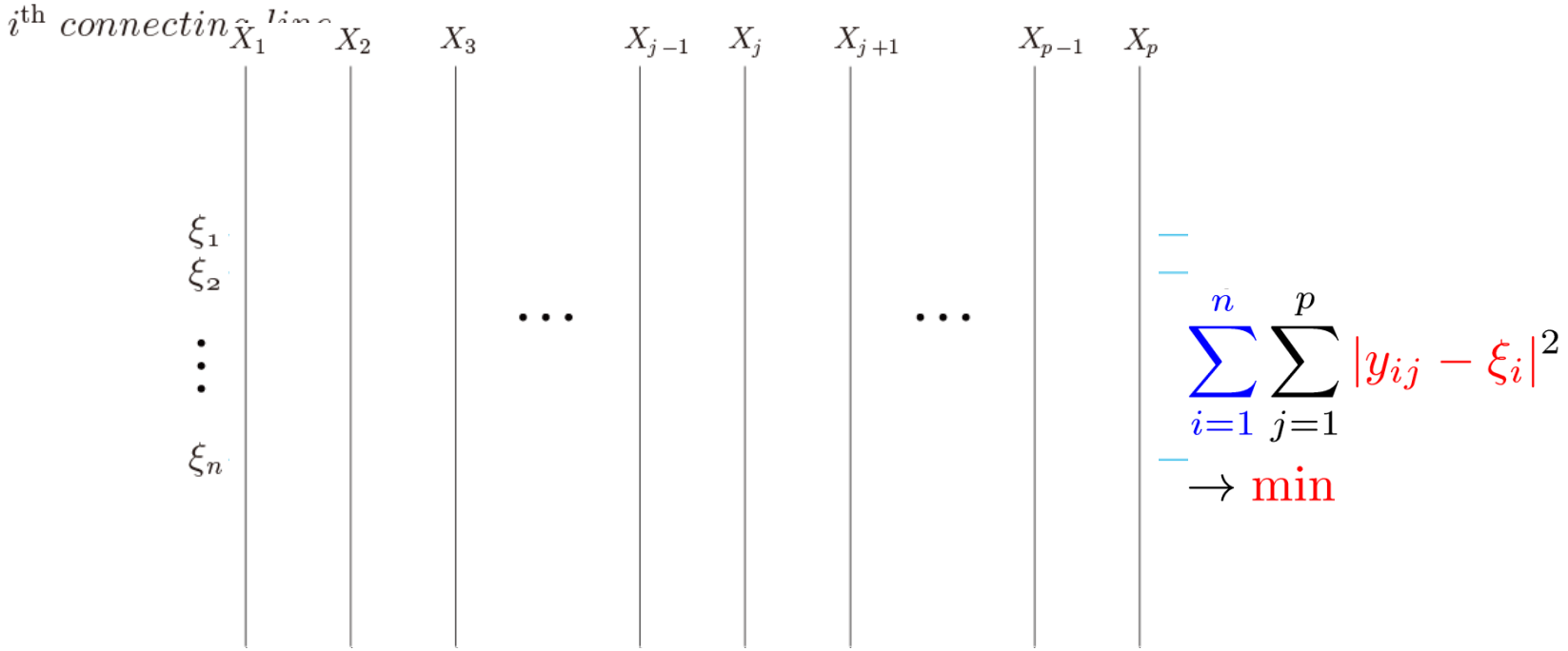
Textile Plot (Kumasaka and Shibata 2008)

- Visualising high dimensional data as it is without any aggregation
- Accepts any data types
- Horizontalisation criterion

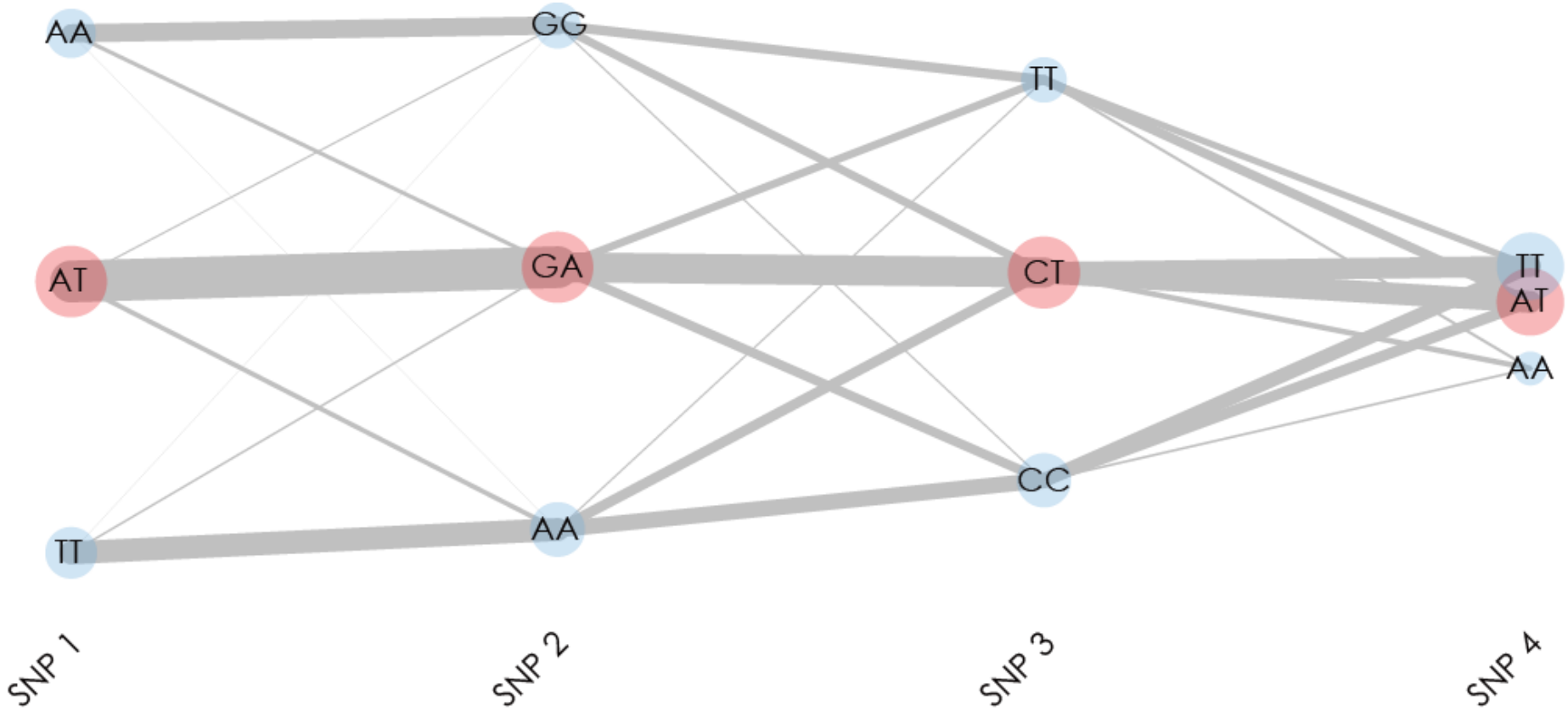
- Genetic point of view:
  - Direct association between SNP genotypes
  - LD statistics and Haplotype structure
  - Approximation of correlation matrix
  - Deviation from Hardy-Weinberg equilibrium

Horizontalisation Criterion

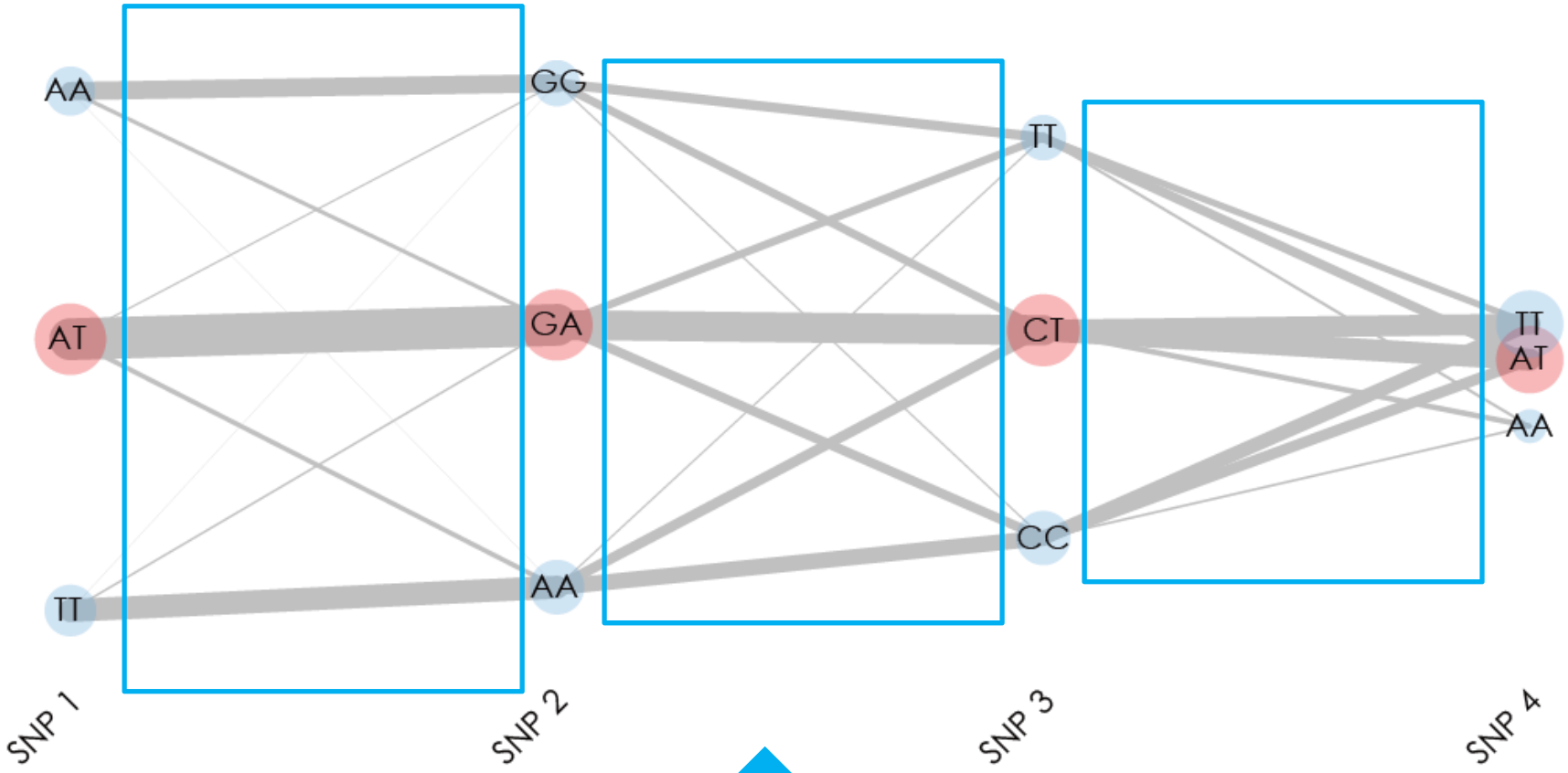
- Degree of horizontalness
- Minimisation problem
- Linearity and orthogonality

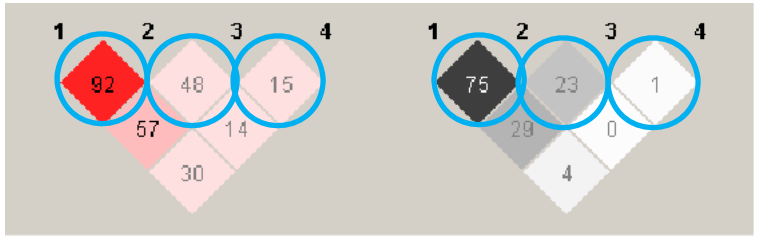


Line Crossings between adjacent SNPs



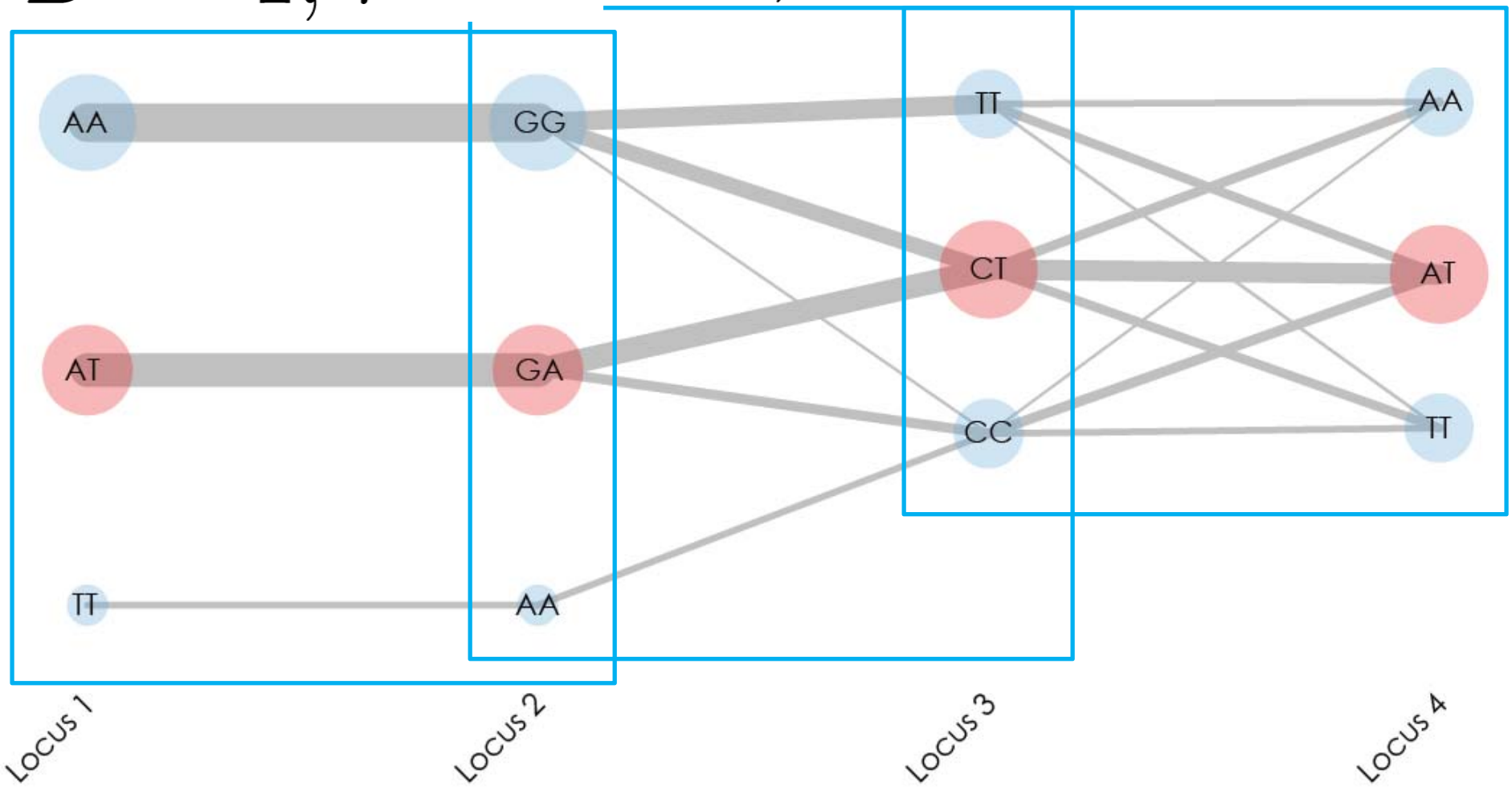
Line Crossings between adjacent SNPs

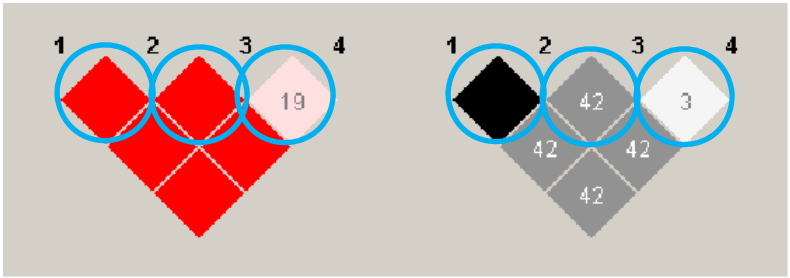




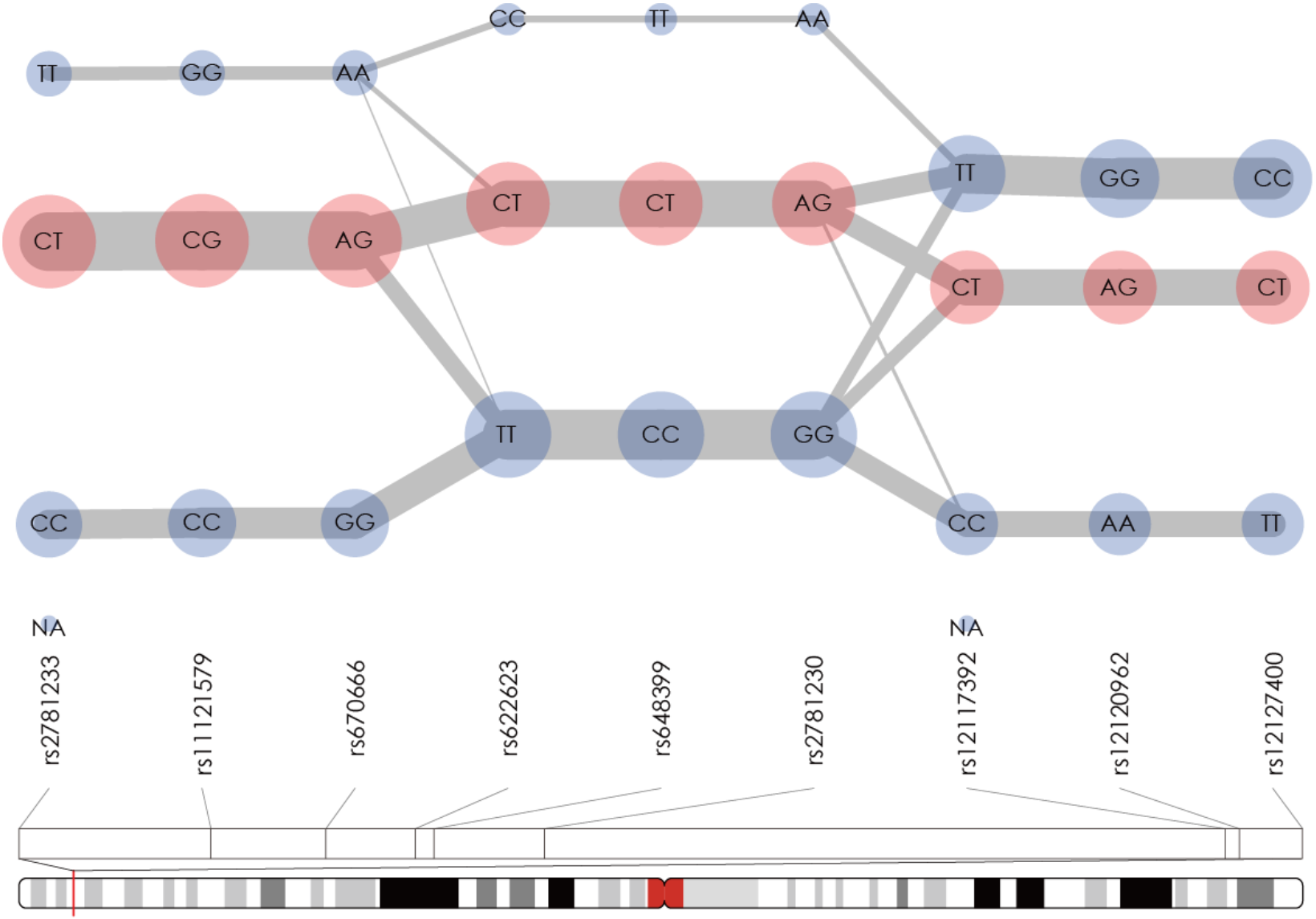
D' r^2

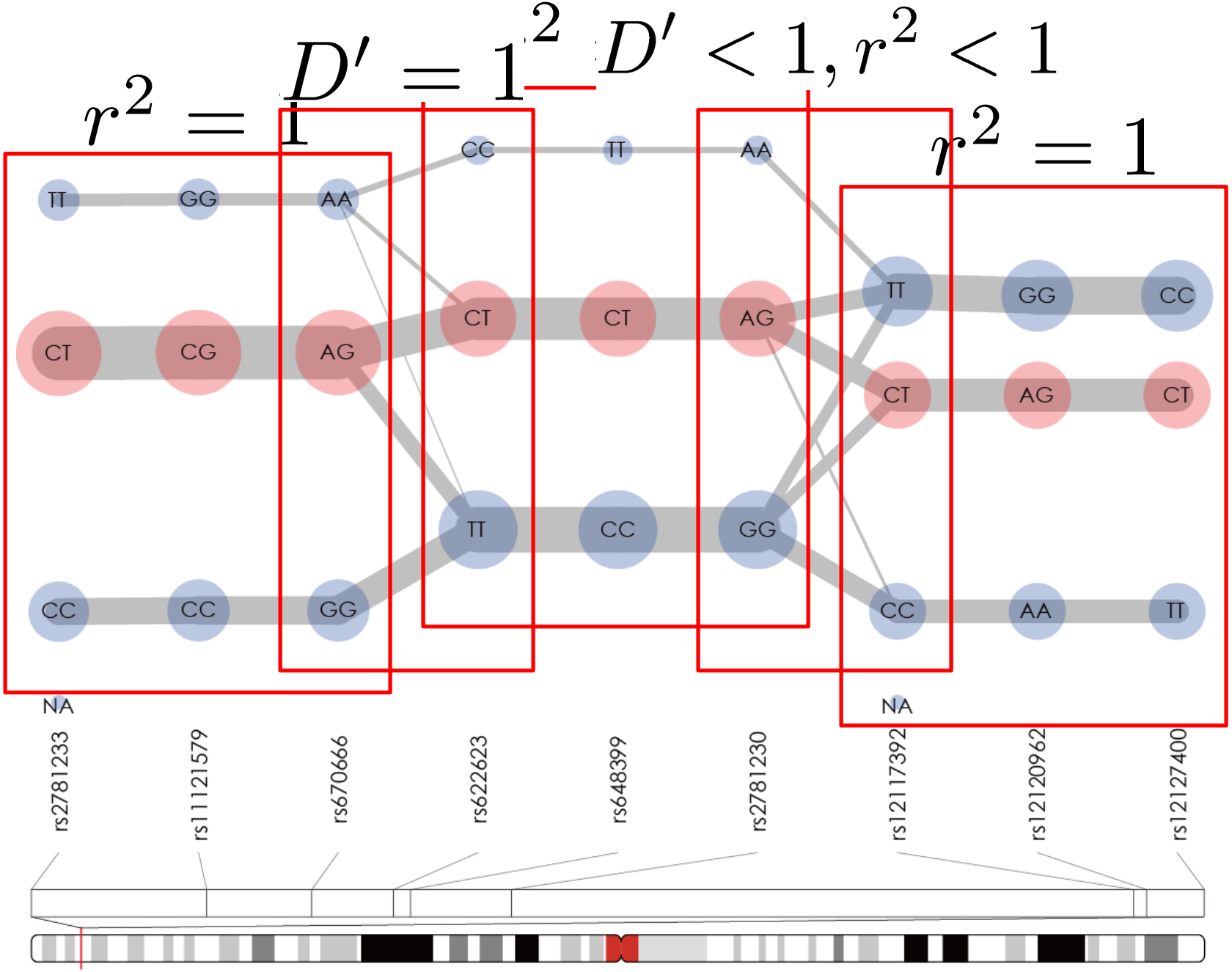
$D' = 1,\ r^2 = D' = 1,\ r^2 < 1 < 1, r^2 < 1$



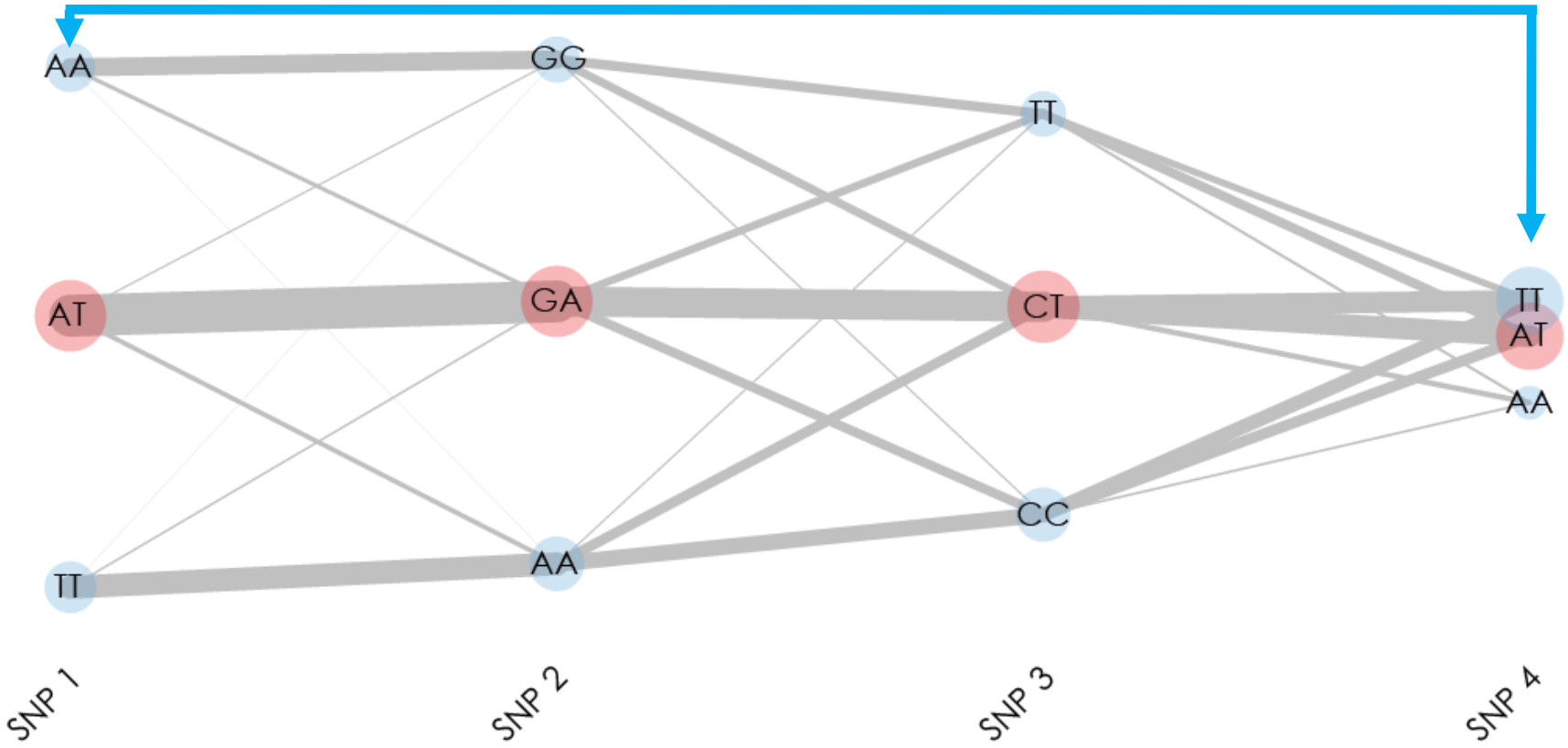


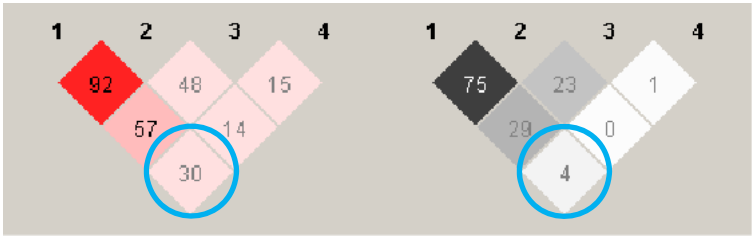
HapMap European-American samples on Chromosome 1




$r^2 = 1$
$D' = 1$
$D' < 1, r^2 < 1$
$r^2 = 1$
TT
GG
AA
CC
TT
AA
CT
CG
AG
CT
CT
AG
TT
GG
CC
CT
AG
CT
TT
CC
GG
CC
CC
GG
CC
AA
TT
NA
NA
rs2781233
rs11121579
rs670666
rs622623
rs648399
rs2781230
rs12117392
rs12120962
rs12127400

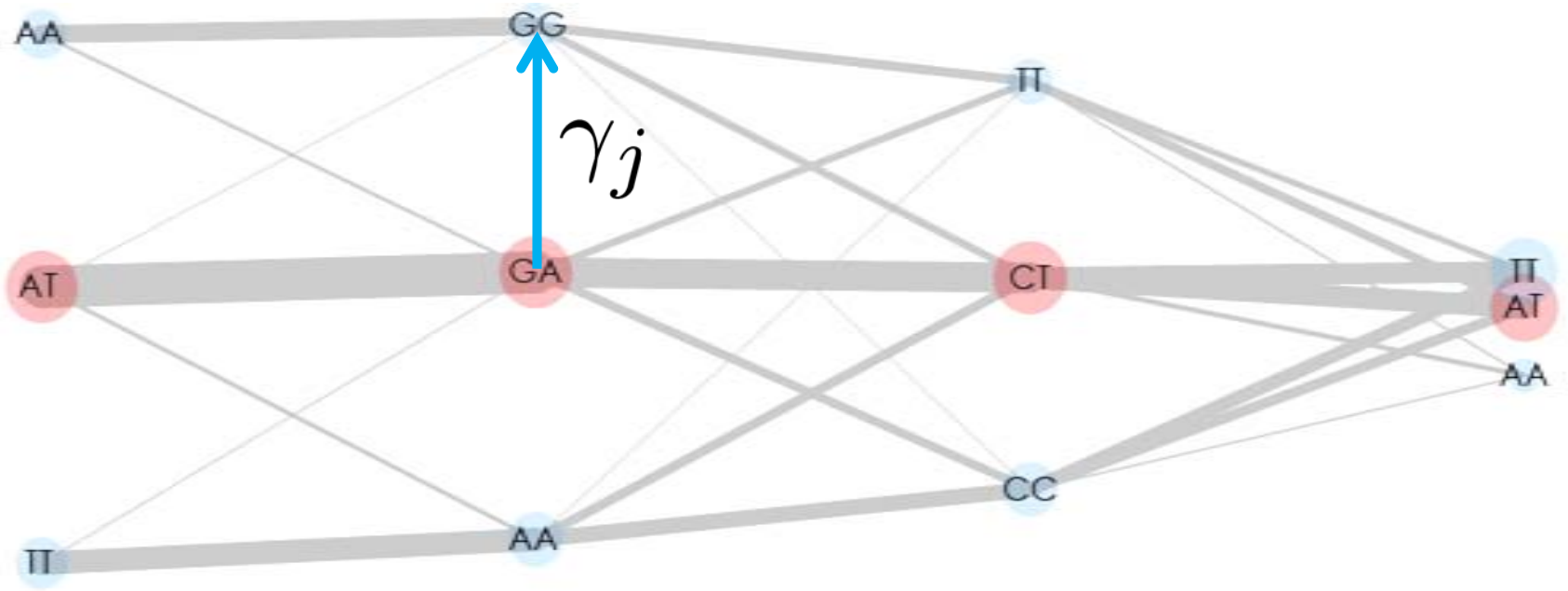
LD between Distant SNPs ?





D' r^2

Dispersion from Heterozygote under HWE

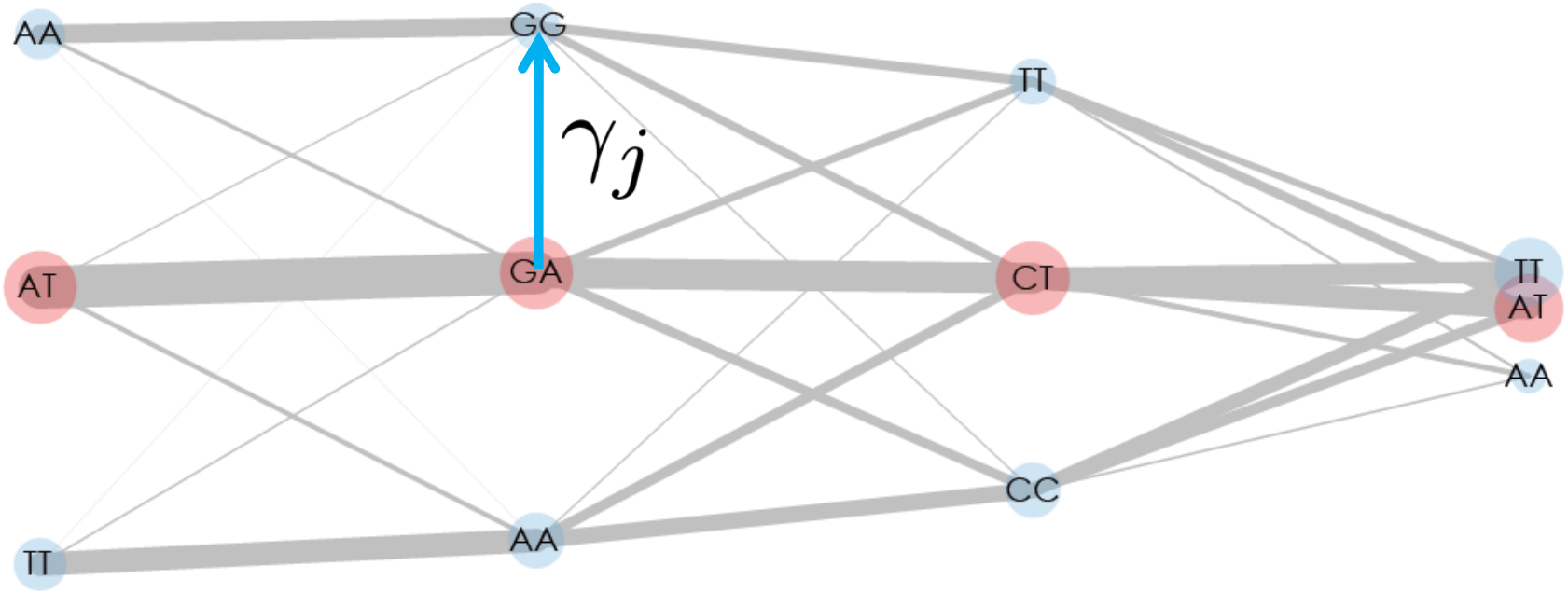


$$\gamma_j \propto \frac{|\eta_j|}{\sqrt{p_j(1-p_j)}};\ j = 1, \ldots, p$$

η_j : jth element of eigenvector for correlation matrix $\mathbf{R}$ associated with the largest eigenvalue λ

p_j : marginal probability of SNP j

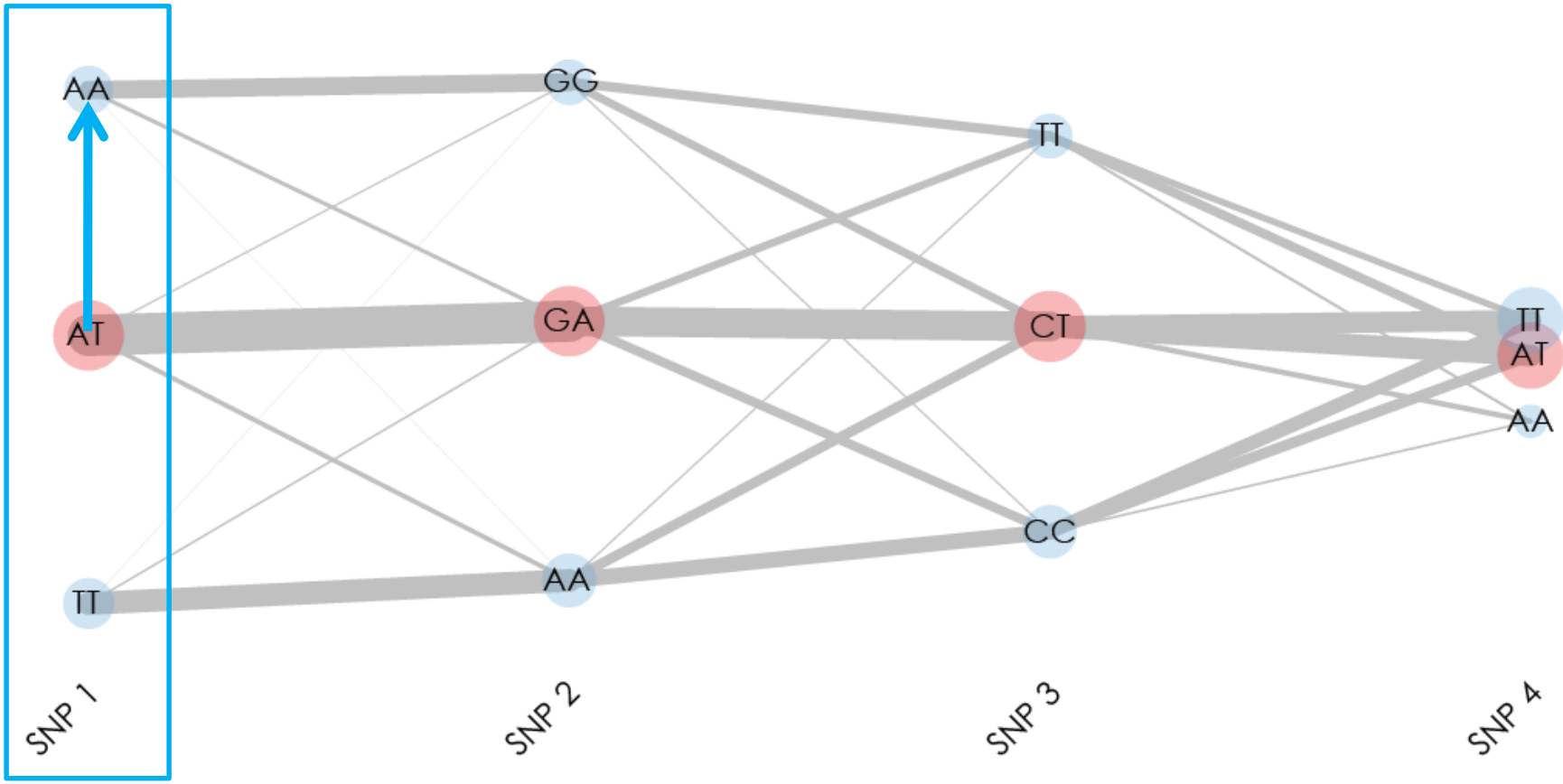
Dispersion from Heterozygote under HWE



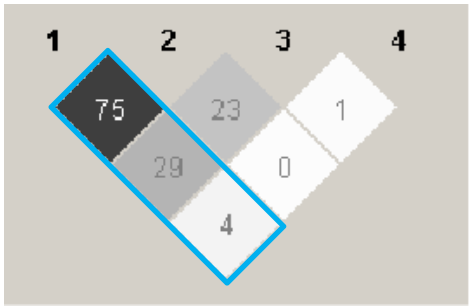
$$\gamma_j \propto \frac{|\eta_j|}{\sqrt{p_j(1-p_j)}} \approx \frac{\lambda}{p} \frac{\sum_{k=1}^{p} |r_{jk}|}{\sqrt{p_j(1-p_j)}}$$

$$(\mathbf{R} \approx \lambda \boldsymbol{\eta} \boldsymbol{\eta}^T)$$

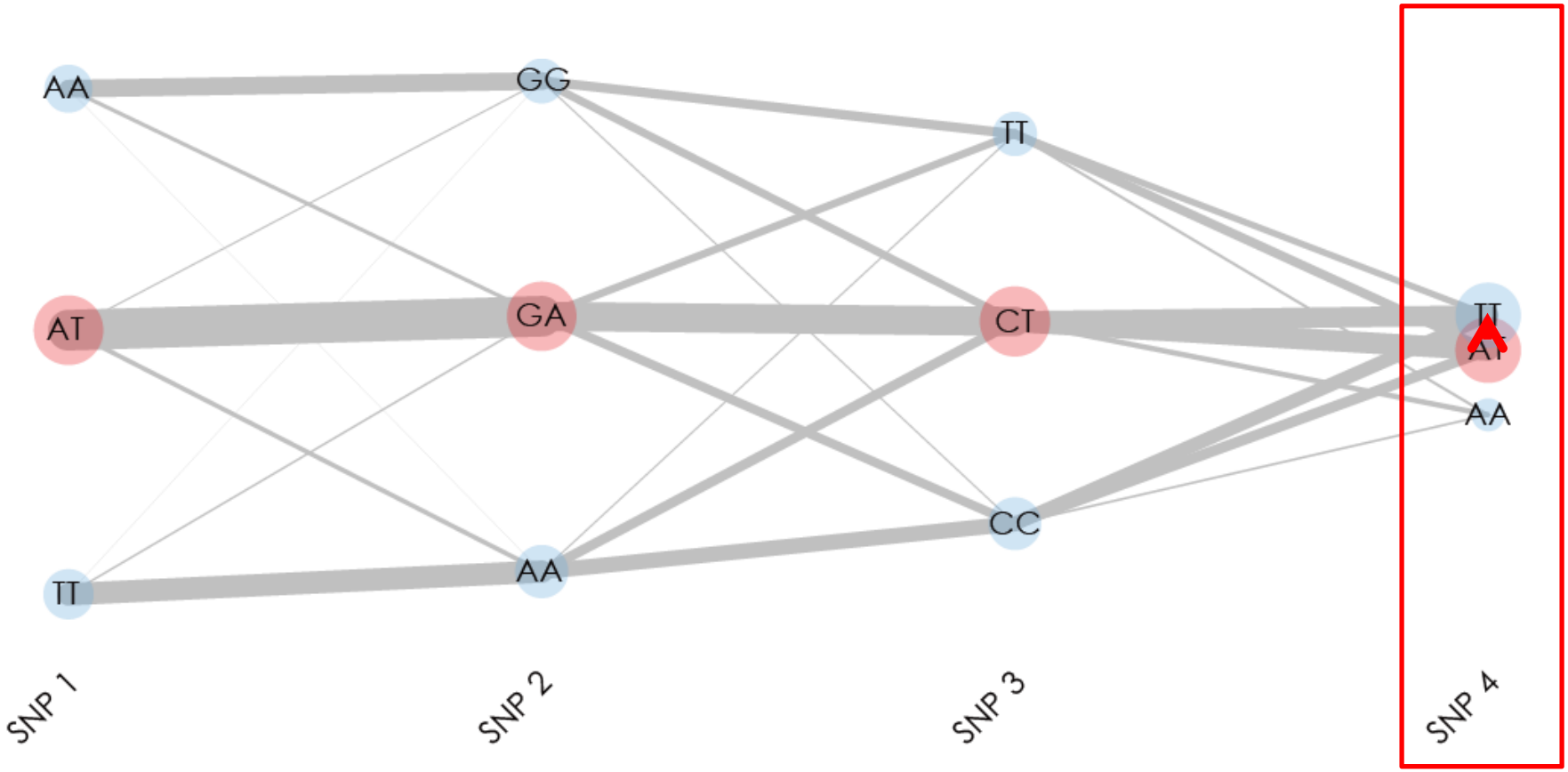
Dispersion from Heterozygote under HWE



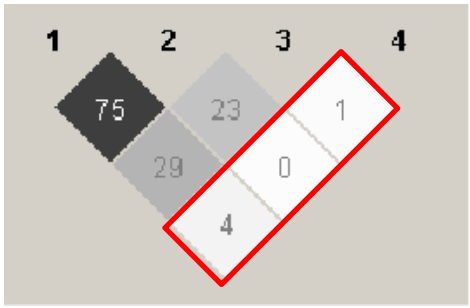
$$\gamma_j \approx \frac{c}{\sqrt{p_j(1-p_j)}} \sum_{k=1}^{p} |r_{jk}|$$



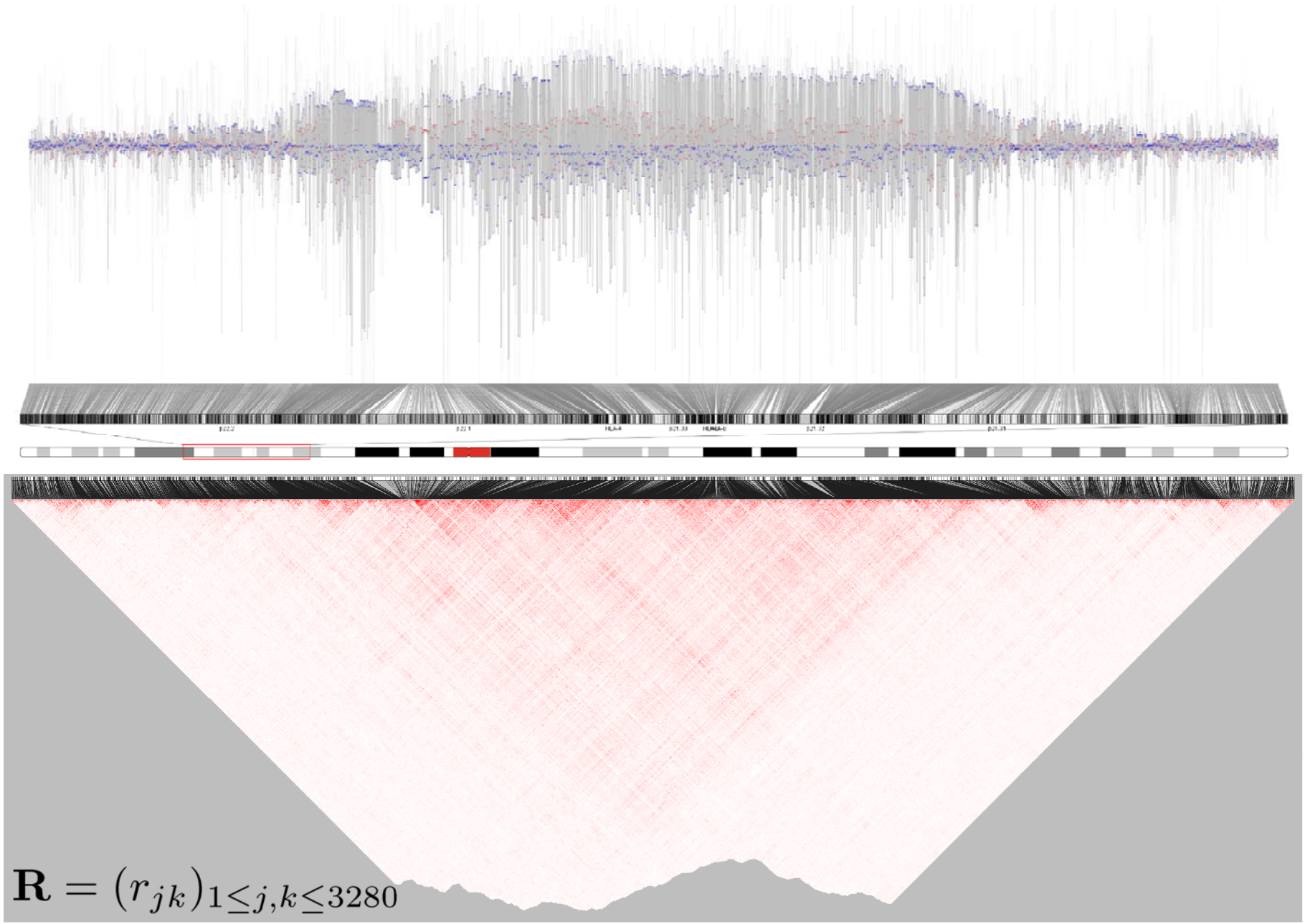
Dispersion from Heterozygote under HWE



$$\gamma_j \approx \frac{c}{\sqrt{p_j(1-p_j)}} \sum_{k=1}^{p} |r_{jk}|$$



3280 SNPs in MHC (major histocompatibility complex) region on Chromosome 6



Inflation of Linkage Disequilibrium

- High density of markers

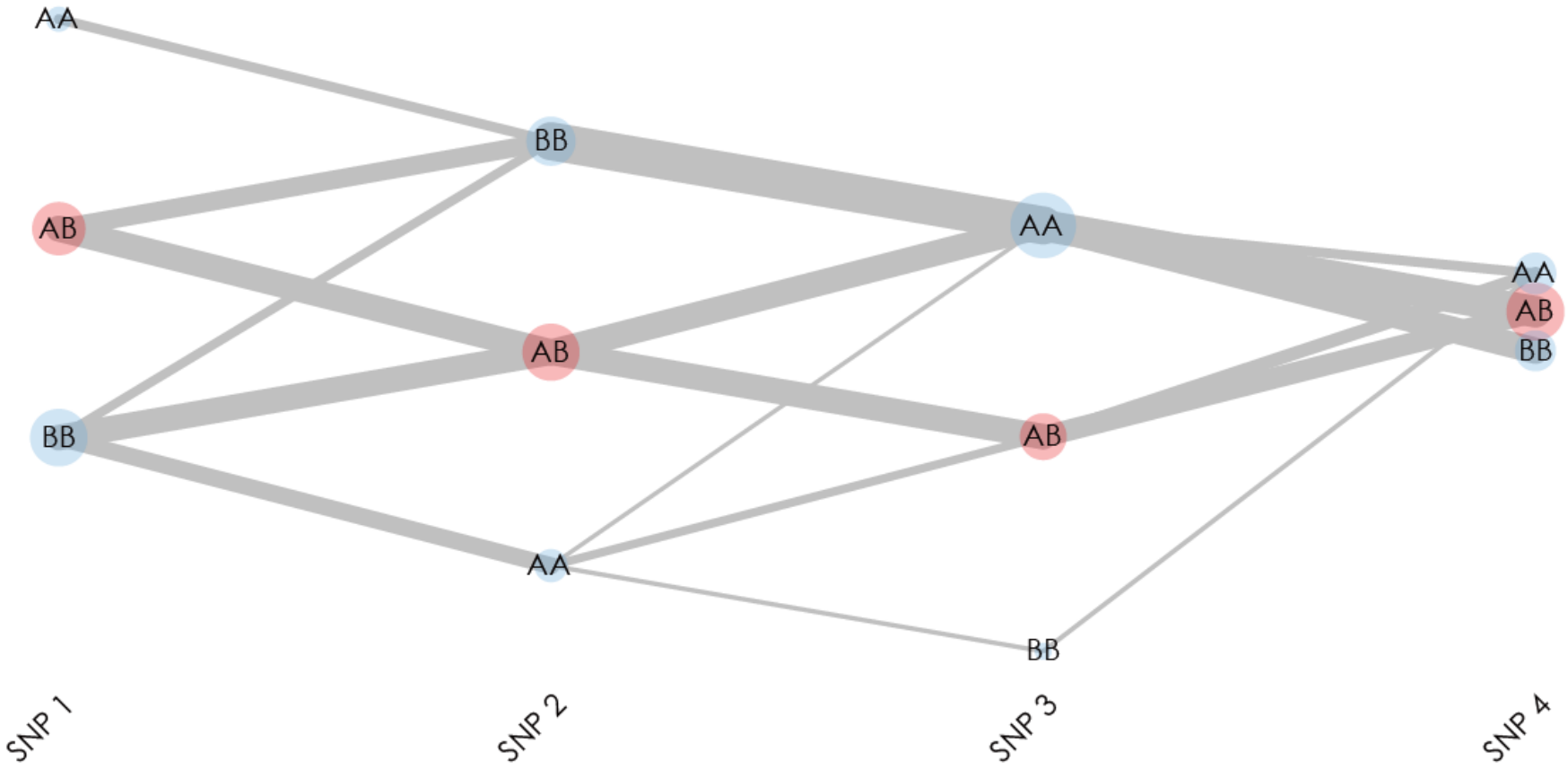
$$\theta_n = \frac{1}{2}\left[1 - (1-2\theta)^n\right] \qquad \left(\theta_1 = \theta,\ \lim_{n\to\infty} \theta_n = \frac{1}{2}\right)$$

$$D_t = (1-\theta_n)^t D_0$$

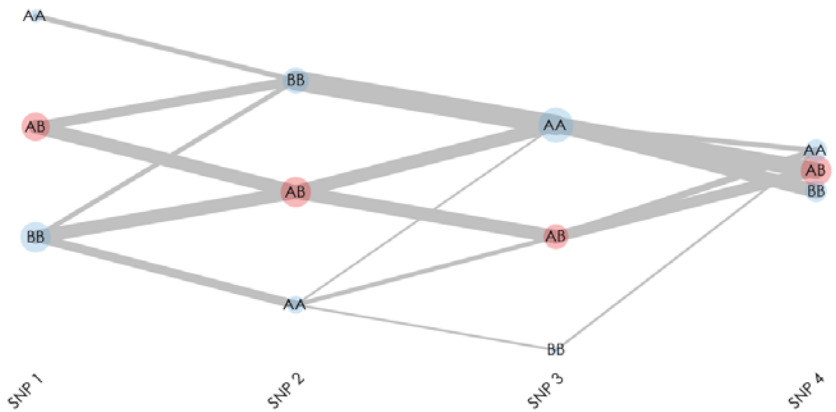
- Natural selection
- Bottleneck and Genetic drift
- Population mixture and admixture

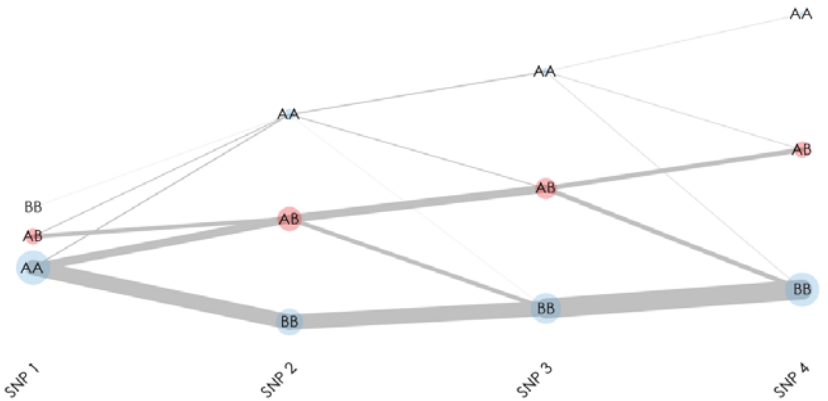
} HWE is violated.

Deviation from HWE

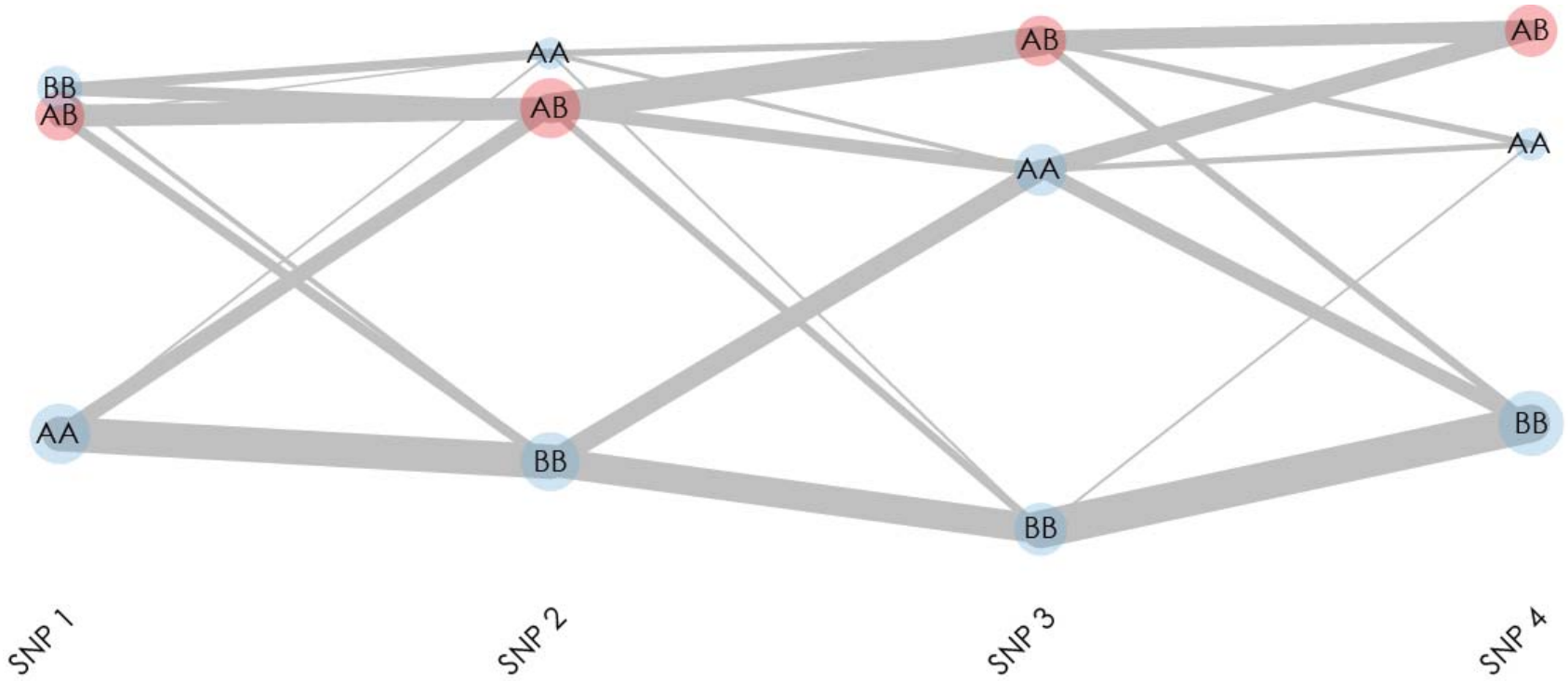


All heterozygotes are located in the middle of homozygotes, if any pair of SNP genotypes are observed under HWE.





Population Mixture



Concluding Remarks

- Direct assciation between SNP genotypes at different loci
- Various types of LD and haplotype structure
- Potential usefulness for disease association study

- Population stratification and admixture
- Coalescent tree or Ancestral Recombination Graph (ARG)
- LD mapping of common complex diseases gene

Bibliography

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29:311-322.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226-231.

Hudson RR (2001) *Handbook of statistical genetics (2nd ed)*, Wiley & Sons, New York.

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. *Gnome Res* 10:1435-1444.

Kumasaka N, Shibata R (2008) High-dimensional data visualisation: the textile plot, *Computational Statistics & Data Analysis*, 52: 3616-3644.

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49-67.

McVean G. (2007) *Handbook of statistical genetics (3rd ed)*, Wiley & Sons, New York.

Peterson AC, et al (1995) The distribution of linkage disequilibrium over anonymous genome regions. *Mum Mol Genet*, 4:887-894.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data, *Am. J. Hum. Genet*, 69: 1-14.

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies, *Am. J. Hum. Genet*, 65:220-228.