# Definitions

**Linkage disequilibrium:** The nonrandom association of alleles at two loci in a population

# Gametic phase

“Gametic phase” (=”phase”) describes which alleles are found at two or more sites on a chromosome

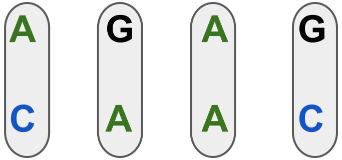
It is common to define “gametes” as the haploid allele combinations at two-loci that could be found in the egg or sperm

For any pair of sites, the haplotypes observed in a diploid individual represent the gametes inherited from the two parents

**Example:** Double heterozygote at two genomic positions in a diploid organism

| | Genotype |
|---|---|
| Position 1 | **A / G** |
| Position 2 | **C / A** |

Possible “gametes”



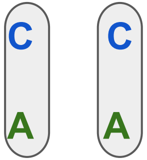
# Gametic phase

When one of two loci are homozygous, only two gametes are possible

Example: Double homozygotes

| | Genotype |
|---|---|
| Position 1 | C / C |
| Position 2 | A / A |

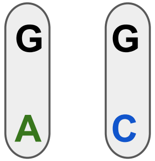
Possible “gametes”



Example: homozygote + heterozygote

| | Genotype |
|---|---|
| Position 1 | G / G |
| Position 2 | A / C |

Possible “gametes”



# Linkage

Two loci on the same chromosome are “linked”

Traditionally, can use a genetic cross to establish if two loci are within 50 cM (implying linkage) or >50 cM (implying free recombination between loci)

# Linkage disequilibrium (LD)

LD is a nonrandom association between genotypes at two loci inferred from population data

LD does not require that the two loci are linked

**Example:** Consider two biallelic loci

A is an allele at locus A with frequency $p_A$

B is an allele at locus B with frequency $p_B$

If locus A and locus B are independent, then expect frequency of AB gamete of $p_A$*$p_B$

**If the observed frequency of AB gamete in a sample from a population differs from expectation, then there is linkage disequilibrium between the two loci**

# Pairwise measures of LD

Consider two bi-allelic loci with allele frequencies $p_1$, $p_2$ and $q_1$ and $q_2$ and observed gamete frequencies of $g_{ij}$, where i is the allele from p and j is the allele from locus j

$g_{11} = p_1q_1 + D$

$g_{12} = p_1q_2 - D$

$g_{21} = p_2q_1 - D$

$g_{22} = p_2q_2 + D$

$g_{11}$ and $g_{22}$ are in “coupling phase”

By convention $g_{11}$ consists of two most common alleles, $g_{22}$ the two rare

# Pairwise measures of LD

Consider two bi-allelic loci with allele frequencies $p_1$, $p_2$, $q_1$ and $q_2$ and observed gamete frequencies of $g_{ij}$, where i is the allele from p and j is the allele from locus q

$g_{11} = p_1q_1 + D$

$g_{12} = p_1q_2 - D$

$g_{21} = p_2q_1 - D$

$g_{22} = p_2q_2 + D$

$g_{11}$ and $g_{22}$ are in “coupling phase”

By convention $g_{11}$ consists of two most common alleles, $g_{22}$ the two rare

$$\boxed{D = g_{ij} - p_iq_j}$$

# Pairwise measures of LD: D' ("D prime")

The range and magnitude of D is dependent on allele frequencies (*undesirable property)

To adjust for this, range of D can be constrained to be between -1 and 1 by defining D' as follows:

$$D' = \frac{D}{\max(-p_1q_1, -p_2q_2)} \quad \text{for } D < 0$$

$$D' = \frac{D}{\min(p_1q_2, p_2q_1)} \quad \text{for } D > 0$$

D' depends on strength of association between loci (i.e., LD) and allele frequencies at each locus

# Pairwise measures of LD: $r^2$

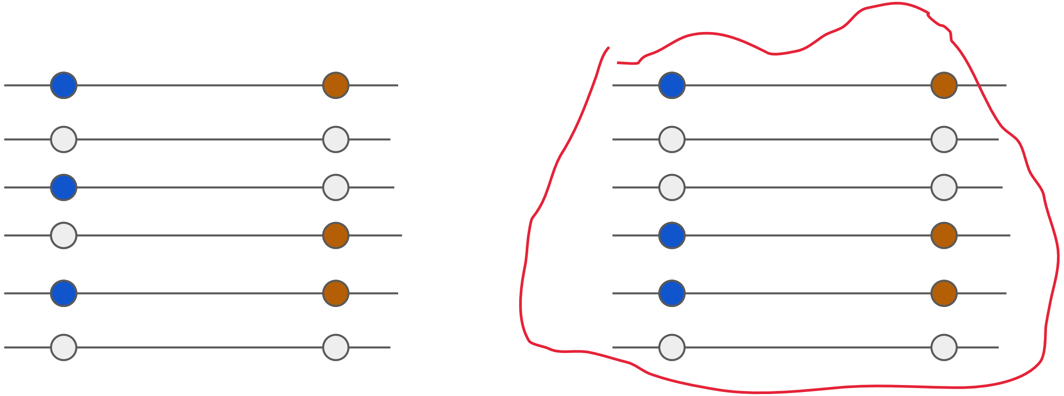
r is the correlation in allelic states at two loci

$r^2$ is the squared correlation of alleles

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

Where D = $g_{ij}$ - $p_i$qj and $p_1$ $p_2$ $q_1$ $q_2$ are allele frequencies

# Which sample of chromosomes has higher LD?



越连锁 LD越高

# Interpretation of measures of LD

D’ is only <  1 if all four gametes are present

$r^2$ of 1 indicates “perfect LD”

Both $r^2$ and D’ depend on allele frequencies

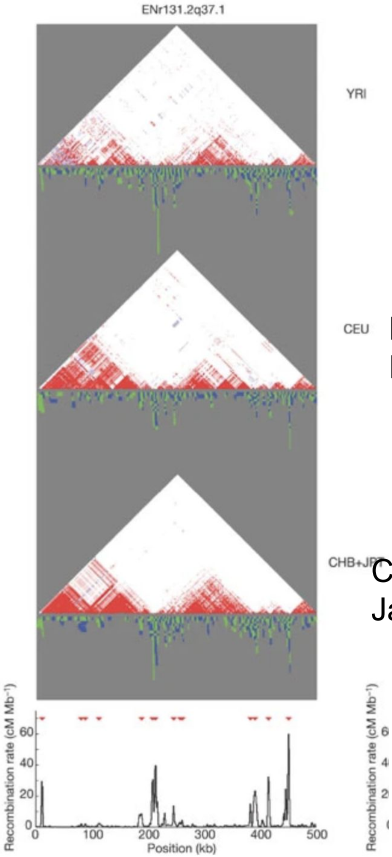
Statistical significance can be tested with Fisher’s Exact Test

# Example: LD heatmap (2q37.1)

D’ of 1 indicates no recombination has occured between two sites

Block-like structure of LD common in humans

Signature suggestive of recombination hotspots



Yoruban

Northern European

Chinese + Japanese

Upper: red points indicate pairwise LD with D’ = 1, white are values with D’ < 1

Lower: recombination rates inferred from SNP data with red triangles showing hotspots of recombination

International HapMap Consortium 2005

# Interpretation of measures of LD

D’ is only < 1 if all four gametes are present

$r^2$ of 1 indicates “perfect LD”

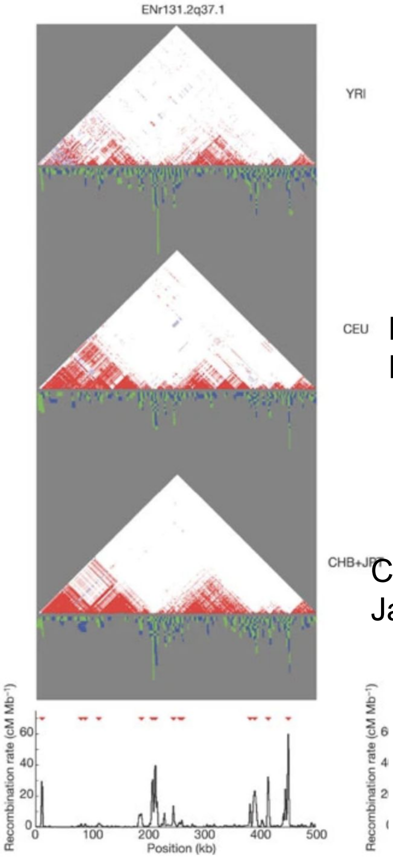
Both $r^2$ and D’ depend on allele frequencies

Statistical significance can be tested with
Fisher’s Exact Test

# Example: LD heatmap (2q37.1)

D’ of 1 indicates no recombination has occured between two sites

Block-like structure of LD common in humans



Yoruban

Northern European

Chinese + Japanese

Upper: red points indicate pairwise LD with D’ = 1, white are values with D’ < 1

Lower: recombination rates inferred from SNP data with red triangles showing hotspots of recombination

International HapMap Consortium 2005

# Example: LPL locus

Pairwise LD shown for the lipoprotein lipase, here shown as a deviation of observed gamete frequencies from expected under linkage equilibrium in a 2 x 2 contingency test

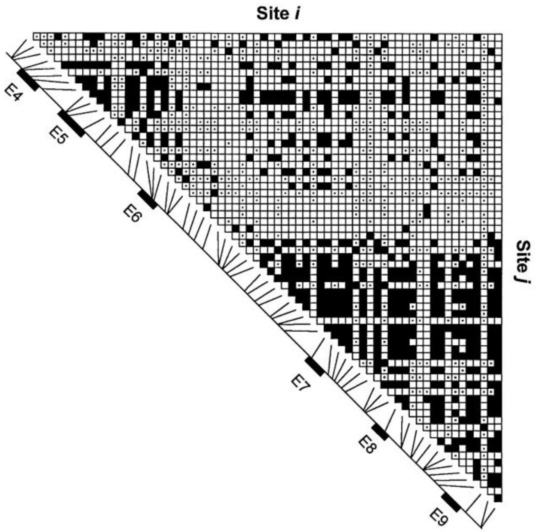
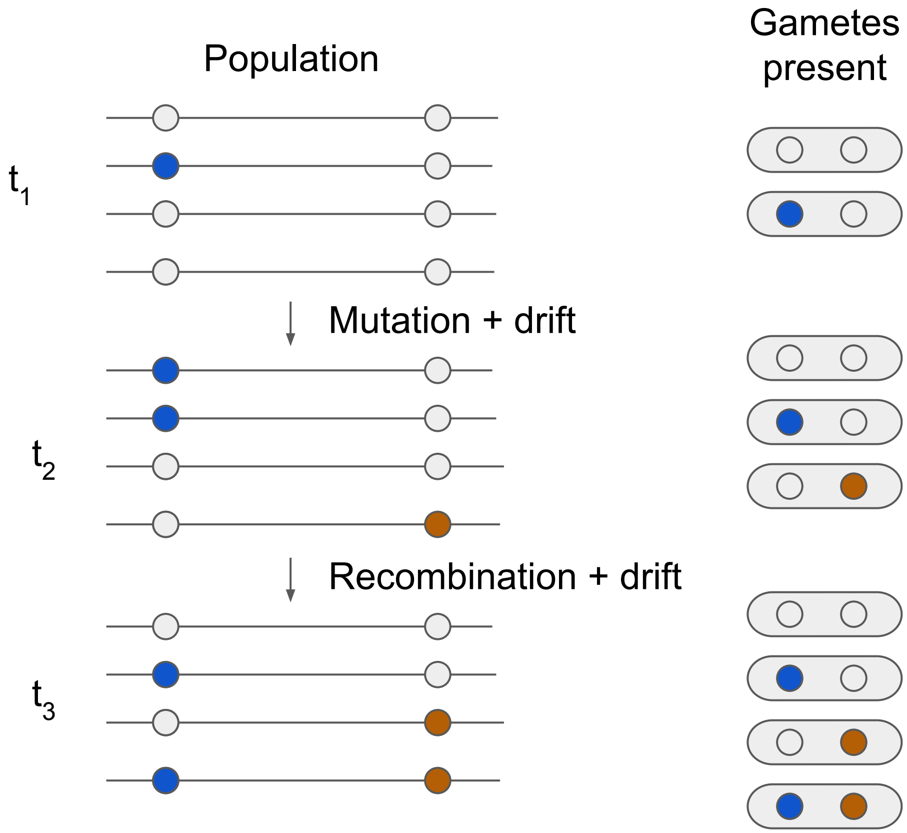


**Figure 5** Plot showing pairwise linkage disequilibrium, indicated by a blackened square, for site pairs with a significant Fisher's exact test ($P < .001$) and no correction for multiple comparisons. In comparisons of site pairs in which both sites have rare nucleotides, there can be complete disequilibrium (one of the four possible gametes having a count of 0), yet the Fisher's exact test can be not significant. Site pairs that lack the power to test a significant association are indicated by a dot in the center of the square. The layout of the figure is the same as that for figure 4.

Clark et al. (1998) Am. J. Human Genetics.

# Recombination generates four gametes in populations

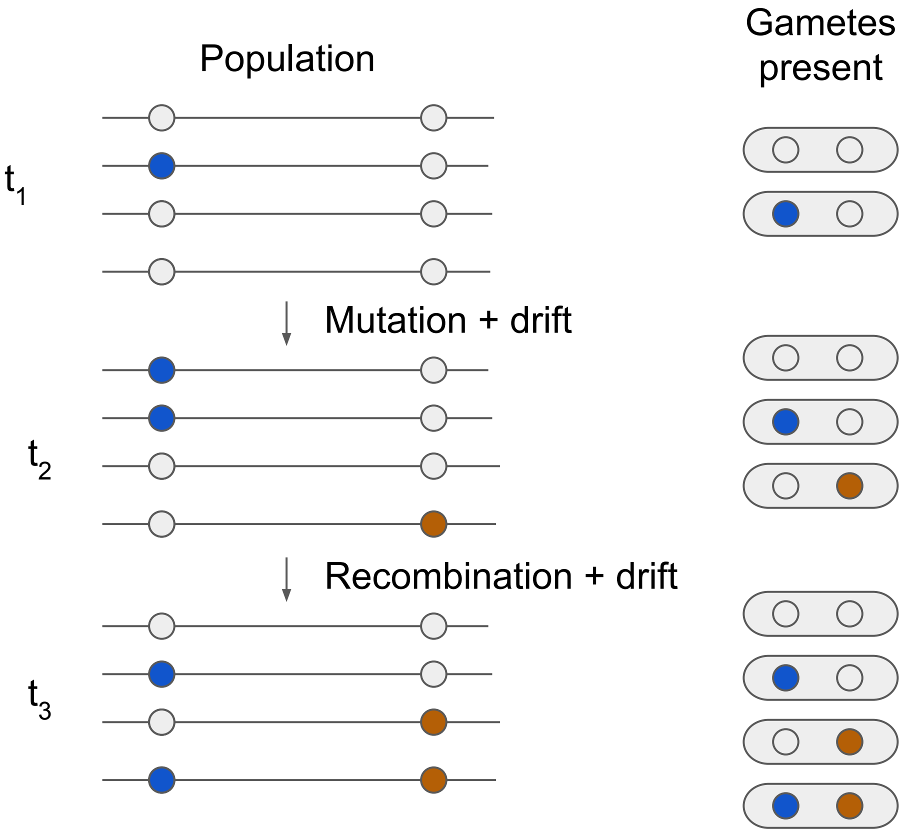
Crossing-over between loci creates four gametes



# The four gamete test

Under infinite sites model, the only way four gametes can be generated is via crossing over

If we assume finite sites, then mutation can also generate four gametes



# LD decay with genetic (cM) and physical (bp) distance

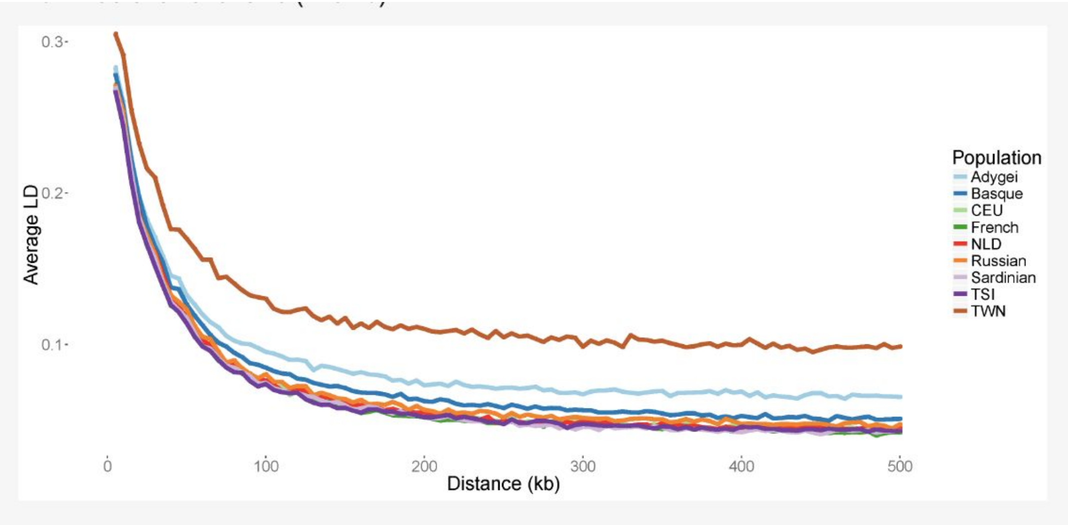
The probability of crossing-over between two loci is a function of the distance between them

Pairwise LD measures should therefore decrease for more distant loci

A common summary of LD decay is the ½ decay distance (= the distance at which average LD is ½ the observed maximum)

Smaller ½ decay distance indicates pairs of sites are correlated for shorter distances

**Example:** Small isolated populations (e.g., TWN of the Netherlands) often have slower rates of LD decay possibly due to increased rates of genetic drift



Somers et al. (2017) Genes

# Approaches to estimating the recombination rate

Rates of recombination can be estimated from genetic crosses (e.g., Drosophila), pedigree-based studies (e.g., human trios), sperm sequencing, or from whole genome re-sequencing (e.g., SNPs)

Ideally, each approach would yield an estimate for each interval of the genome in a common currency (e.g., cM/Mb)

The centimorgan (cM) is a measure of the frequency of recombination where 1 cM corresponds to a 1% frequency of two markers (e.g., SNPs) will recombine during meiosis.

# Approaches to estimating the recombination rate

Population-based estimates of recombination are derived from analysis of haplotype blocks and the pattern of LD in a population sample

Regions of low recombination are expected to have:

- Longer haplotype blocks (regions with long haplotypes shared among sample chromosomes)
- High LD between physically distant pairs of SNPs

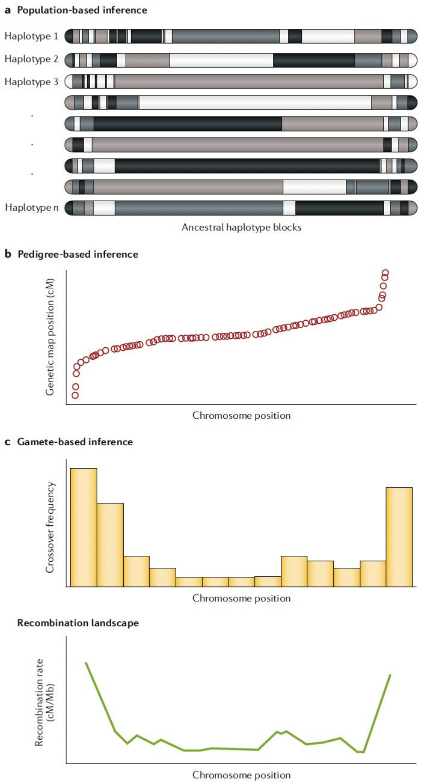


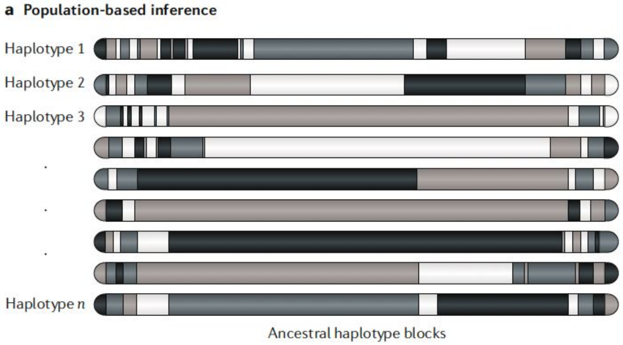
Fig. 4 | **From inference to landscape.** The actual result of each inference method and how it translates to the recombination landscape. **a** | Population-based inference involves direct analysis of haplotype structure along chromosomes. Contemporary haplotypes are composed of ancestral haplotypes (various shades) that arose at different points in the past. The identity and length of ancestral haplotype blocks are a function of the time at which the haplotype arose and recombination. **b** | Pedigree-based inference involves a comparative representation of genetic distance and physical distance where the local recombination rate is the slope at any given location. **c** | Gamete-based inference takes the crossover frequency of a given window and translates it into the recombination landscape. cM, centiMorgans.

Penalba and Wolf (2020) Nature Reviews Genetics

# Recombination in population genomic data

DNA sequences from population samples show signature of historical recombination accumulated from thousands of generations of transmission (i.e., thousands of meioses)

These cross-over events cannot be easily captured in pedigree-based or sperm sequencing studies (owing to cost of sequencing large numbers of trios or sperm to capture similar numbers of crossing-over events)



# The population recombination parameter (“rho” = $\rho$)

How do we quantify how much recombination has occured in the history of a sample of sequences?

In an equilibrium population with infinite sites model of mutation, the only factors contributing to the amount of recombination in a sample are the per site recombination rate (*c*) and the effective population size $N_e$

As can be seen from the $\rho = 4N_e c$, higher effective size and higher rates of crossing over lead to higher $\rho$

The population recombination parameter is defined as:

$\rho = 4N_e c$

# The population recombination parameter (“rho” = $\rho$)

How do we quantify how much recombination has occured in the history of a sample of sequences?

In an equilibrium population with infinite sites model of mutation, the only factors contributing to the amount of recombination in a sample are the per site recombination rate (*c*) and the effective population size $N_e$

As can be seen from the $\rho = 4N_e c$, higher effective size and higher rates of crossing over lead to higher $\rho$

The population recombination parameter is defined as:

$\rho = 4N_e c$

*note: Recall that $\theta$ ($=4N_e\mu$) is a population mutation rate, similarly $\rho$ ($=4N_e c$) is a population recombination rate

# The population recombination parameter (“rho” = $\rho$)

The expected number of recombination events (*R*) in a sample of sequences can be estimated from $\rho$

The parameter $\rho$ is inversely related to linkage disequilibrium ($r^2$) whose exact relationship depends on a set of simplifying assumptions

Therefore, populations with smaller Ne should have higher LD

$E(R) = \rho{*}a$

where a is defined as::

$$\sum_{i=1}^{n-1} \frac{1}{i}$$

$\rho$ is related to $r^2$ by:

$E(r^2) \sim 1 / (1 + \rho)$

Therefore:

$E(r^2) \sim 1 / (1 + 4N_e c)$

# The population recombination parameter (“rho” = $\rho$)

The expected number of recombination events ($R$) in a sample of sequences can be estimated from $\rho$

The parameter $\rho$ is inversely related to linkage disequilibrium ($r^2$) whose exact relationship depends on a set of simplifying assumptions

Therefore, populations with smaller Ne should have higher LD

Recall that the expected number of segregating sites is $E(S) = \theta * a$, so reasoning by analogy $E(R) = \rho * a$ is the expected number of recombination events in a set of sequences

$E(R) = \rho * a$

where a is defined as::

$$\sum_{i=1}^{n-1} \frac{1}{i}$$

$\rho$ is related to $r^2$ by:

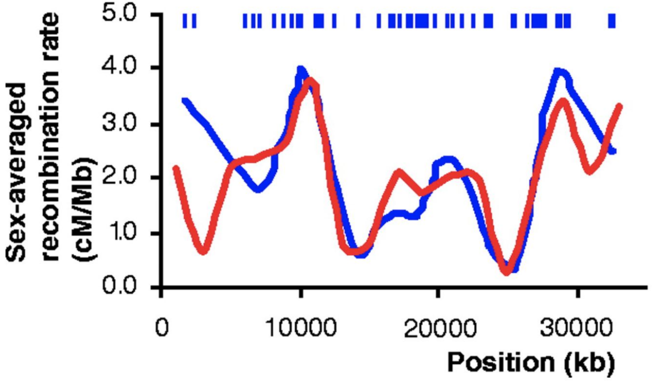
$E(r^2) \sim 1 / (1 + \rho)$

Therefore:

$E(r^2) \sim 1 / (1 + 4N_e c)$

# Pedigree rates versus recombination rates, *c*, from population data

Pedigree-based recombination versus recombination rate inferred from sequence polymorphism data



Blue line represents recombination rate estimates from pedigree data

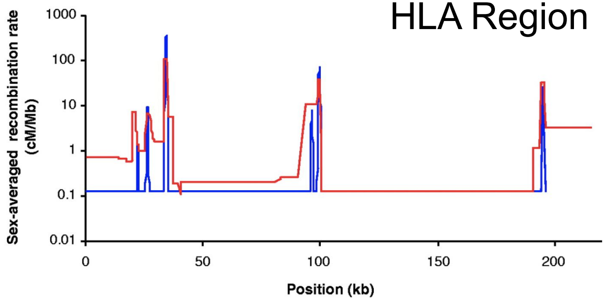
Red line represents recombination rate fromand population genetic data (red)

McVean et al. (2004) Science

# Population genetic data reveal hotspots of recombination

Human recombination landscape is characterized by hotspots

>15,000 hotspots in the human genome



McVean et al. (2004) Science

# Recombination hotspots in humans

Recombination in the human is concentrated in narrow regions

Typically 80% of the recombination occurs in 20% of the DNA sequence

Regions of high recombination are accompanied by megabase-scale recombination desserts

Estimates are that there are 25,000 to 50,000 hotspots in the human genome

**Example:** chromosome 12

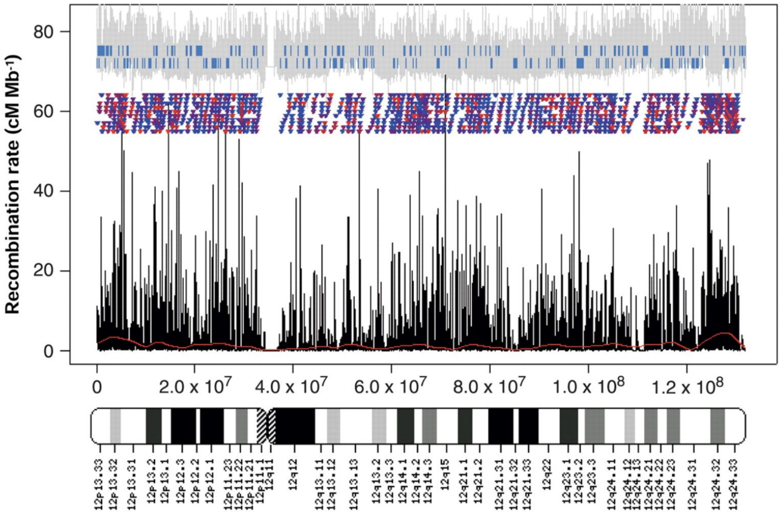


**Fig. 1**. Recombination rate variation along chromosome 12. Shown are estimated recombination rate (black), locations of statistically significant recombination hotspots [triangles; colors indicate relative amount of recombination from low (blue) to high (red)], and estimated recombination rates from the deCODE (6) genetic map (red curve near bottom). Also shown are the location of ENSEMBL genes on the two strands (blue segments), fluctuations in local GC content (gray lines; averages over 1000-bp windows shown on an arbitrary scale), and an ideogram of chromosome banding.

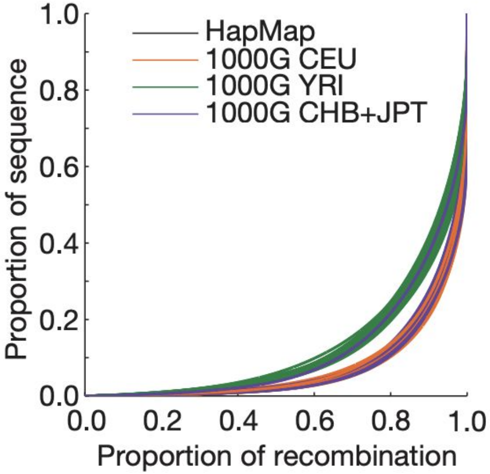
Myers et al. (2005) Science

# Recombination is less concentrated in hotspots in African populations

Curves are generated by rank ordering genome segments based on the E(R) (=expected number of recombination events) and plotting cumulative totals

Uniform recombination would be a diagonal line with slope = 1 and intercept = 0

Suggestion that YRI (Yoruban) have more uniform recombination than CEU (European) and CHB+JPT (Asian) populations



1000 Genomes Consortium (2010) Nature

# Recombination hotspots in humans

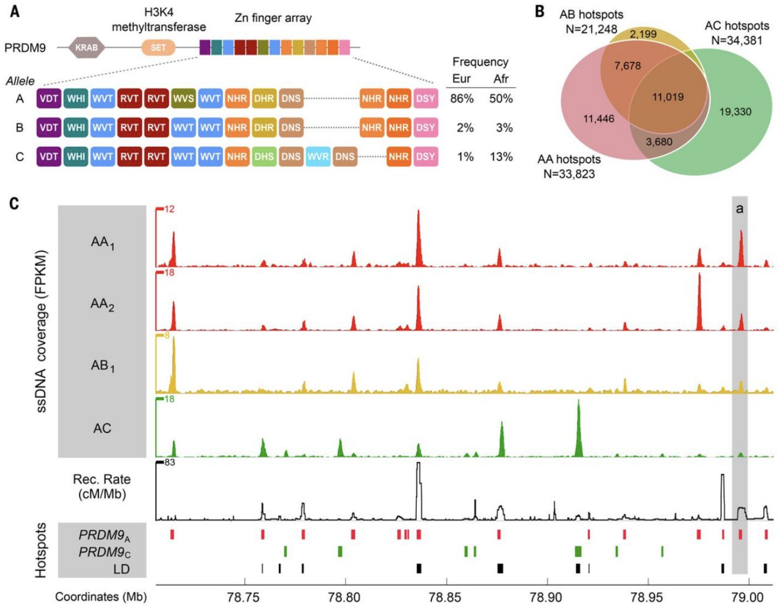
Recombination ('crossing-over') is initiated via the formation of programmed double-stranded breaks (DSB)

PRDM9 is a DNA-binding protein that recognizes short motifs in DNA and initiates DSB formation

PRDM9 alleles are highly polymorphic

Different PRDM9 alleles have different binding preferences for DNA

PRDM9 alleles contribute to variation in meiotic DSB formation and recombination hotspots



Baudat et al. (2010). Science.
Pratto et al. (2014). Science.

# LD and the discovery of the causal basis of complex disease

Genomewide association studies (GWAS) are a primary means of mapping of disease loci and other quantitative traits

In principle, GWAS studies can identify causal SNPs for disease (i.e., statistical association of a trait and a SNP could be because the SNP causes the disease)

In most cases, statistical association is because the SNP is in LD with a mutation that causes the disease (or other trait)

Therefore, the rate of LD decay is critical to GWAS because a trait can only be mapped with sufficiently high SNP density

# LD and the discovery of causes of complex diseas

Whether a trait can be mapped to a gene region depends on LD in the region

In high LD regions, the trait can be mapped, but the mapping interval can be large

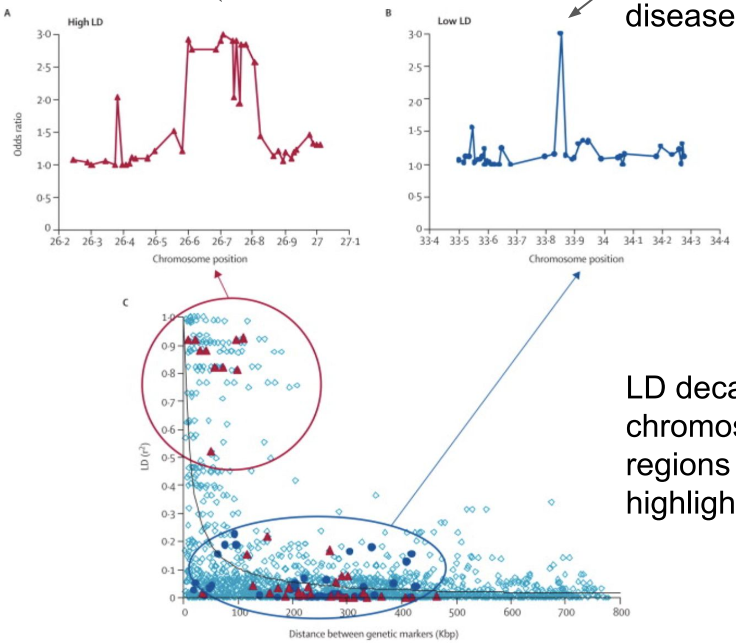
With low LD, its possible the region cannot be mapped unless the causal variant is surveyed or marker density is sufficiently high

*Odds ratio is a measure of the strength of association between marker and trait

Many loci associated with the trait

hitchhiking

Only causal SNP is associated with disease trait



LD decay on chromosome 22 with regions in A and B highlighted

Palmer et al. (2005) The Lancet