Imputación de variantes genéticas no genotipadas

Carla Lluís-Ganella, MSc PhD
Grupo de Epidemiología y Genética Cardiovascular
Parque de investigación biomédica de Barcelona (PRBB)
clluis@imim.es | carla.lluis@gmail.com

Índice

Qué es?

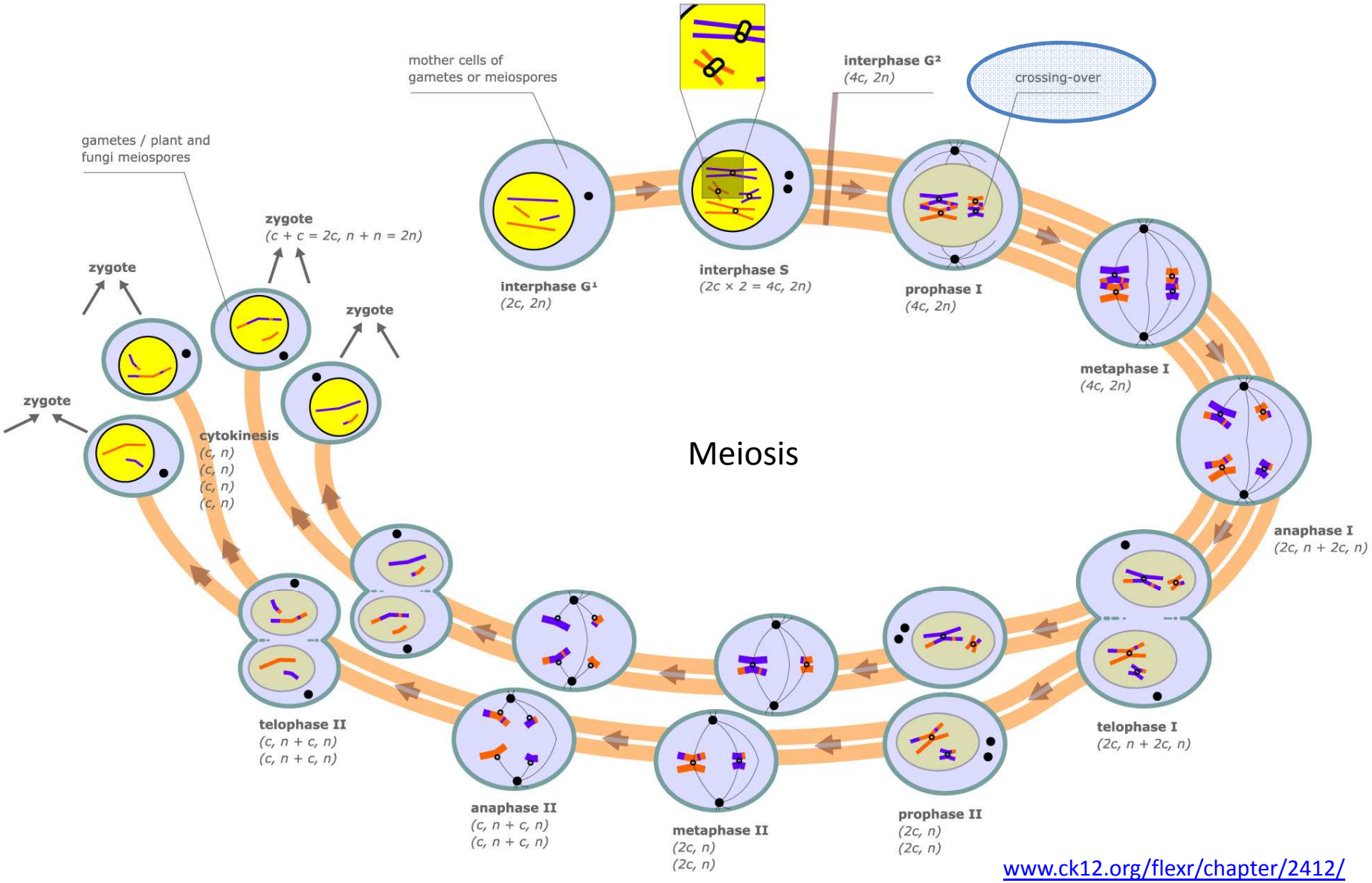
La imputación es un método estadístico que sirve para lidiar con datos no disponibles a partir de la información de otros datos si disponibles.

Por qué la necesitamos en genética?

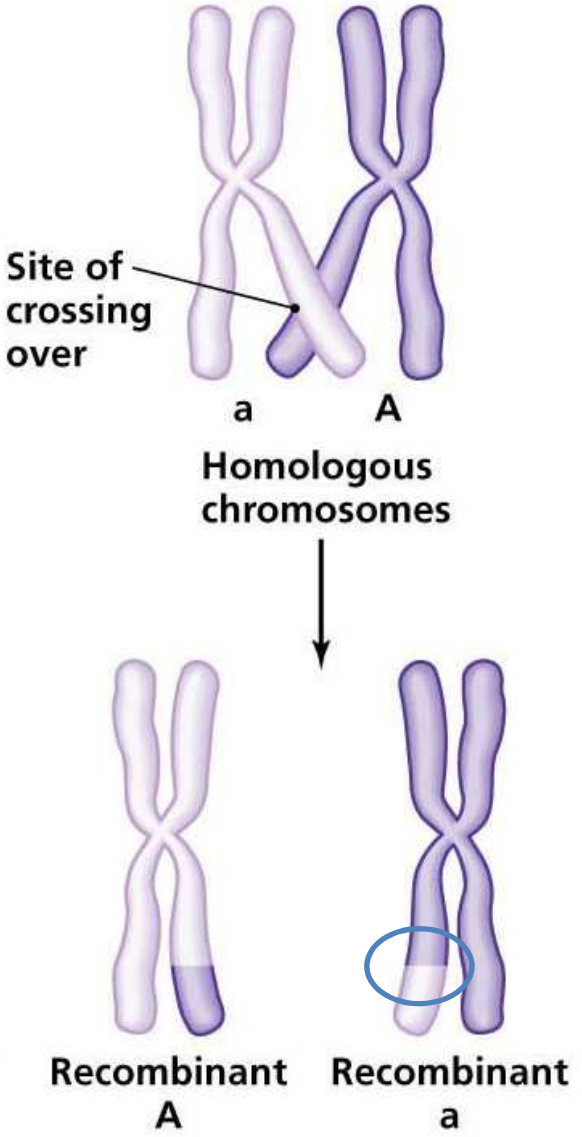
Genotipar un gran número de variantes no es habitualmente posible debido a cuestiones económicas.

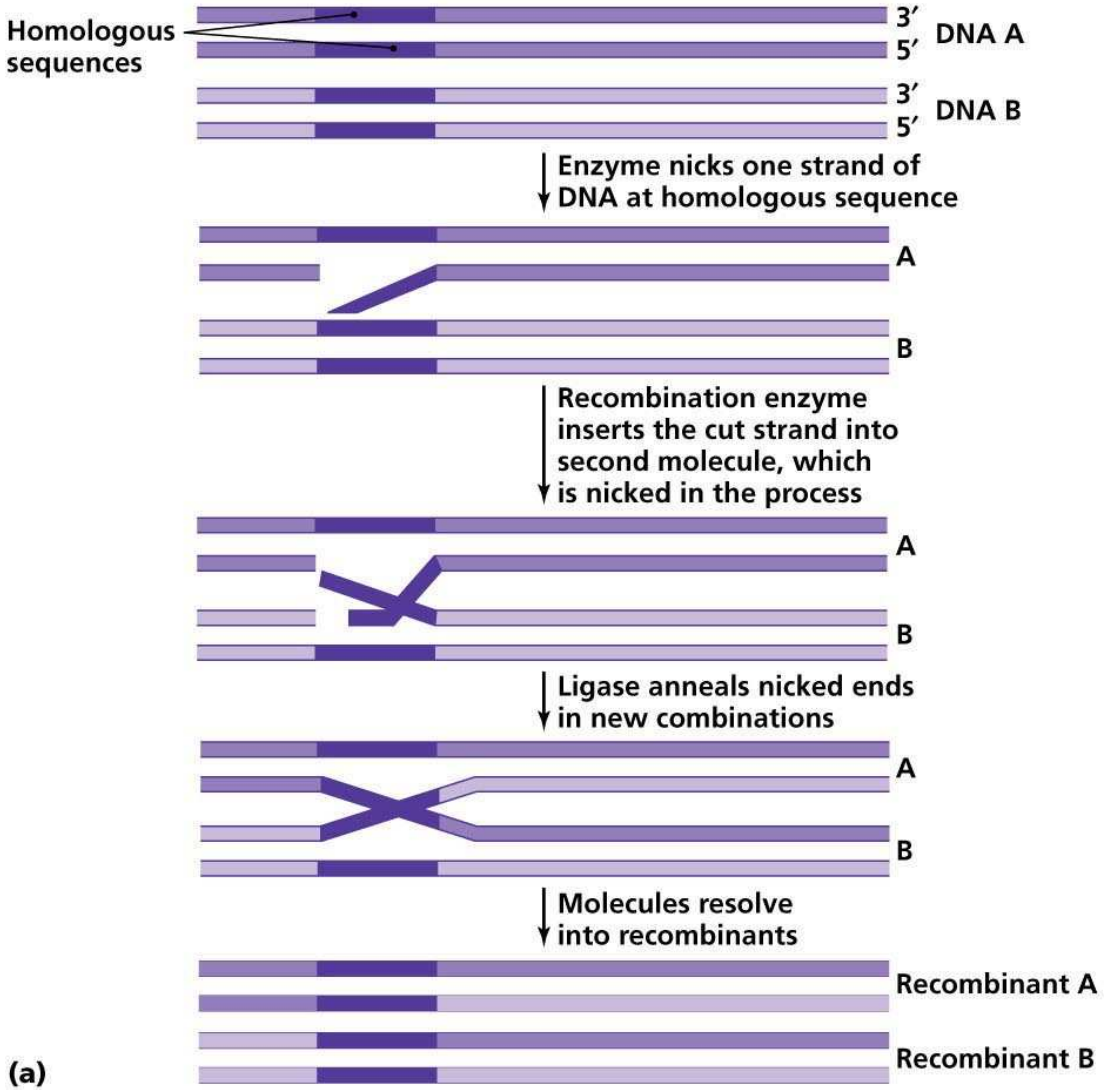
Permite recuperar algunas variantes genéticas con genotipos inciertos en algunos individuos, permitiendo un aumento del número de variantes y muestras disponibles.

La imputación aprovecha un proceso biológico llamado “recombinación cromosómica” para calcular la probabilidad de que un individuo tenga un genotipo determinado.

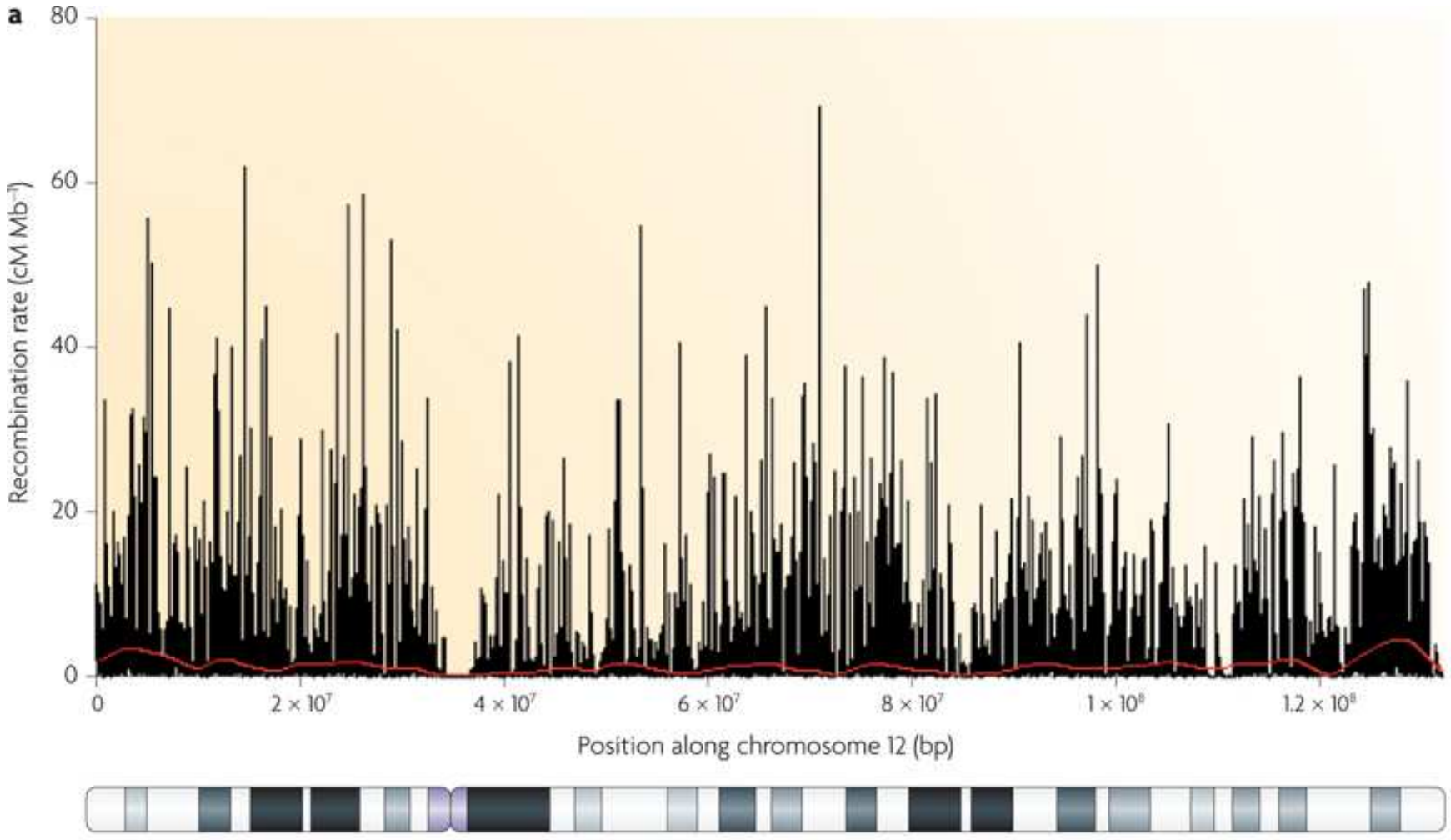


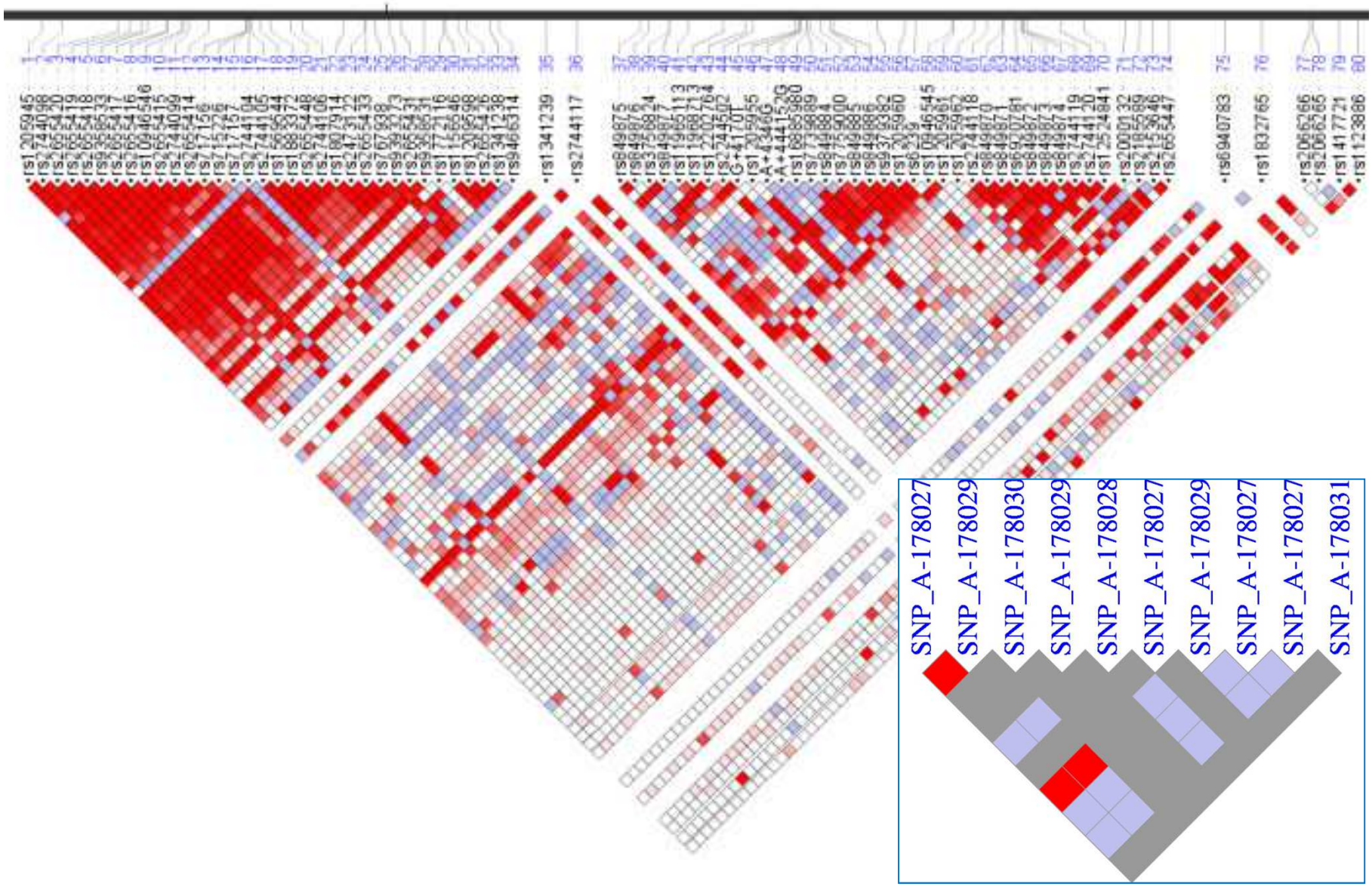
www.ck12.org/flexr/chapter/2412/

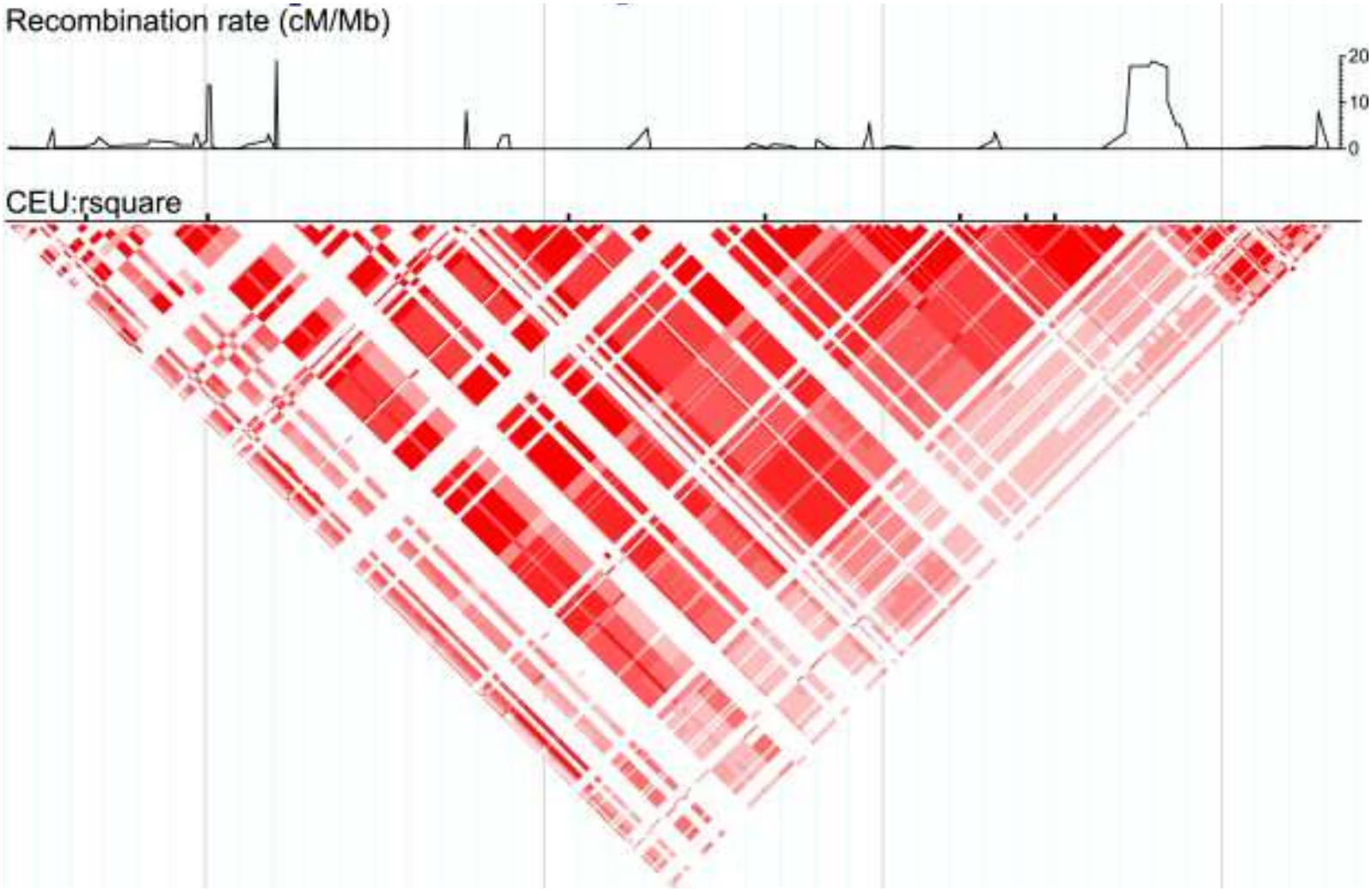




http://academic.pgcc.edu


a
80
60
40
20
0
Recombination rate (cM Mb⁻¹)
0
2 × 10⁷
4 × 10⁷
6 × 10⁷
8 × 10⁷
1 × 10⁸
1.2 × 10⁸
Position along chromosome 12 (bp)


SNP_A-178027
SNP_A-178029
SNP_A-178030
SNP_A-178029
SNP_A-178028
SNP_A-178027
SNP_A-178029
SNP_A-178027
SNP_A-178027
SNP_A-178031


Recombination rate (cM/Mb)
20
10
0
CEU:rsquare

IMPUTE
Marchini *et al.* Nat Genetics 2007

MACH
Li *et al.* Genetic Epidemiology 2010

BEAGLE
Browning and Browning. Hum Genet 2008

PLINK
Purcell *et al.* Am J Hum Genet 2007

at may show
, microsatel-
fferent study
l mapping by
of statistical

requires care.
ensity as well
sed and the
mpare head-

at will not be
ıe population
ıy features of
ır context is
differences in
sample used
ond, whether
ure within a
ı the analysis

nent $G = \{G_O, G_M\}$. To impute the missing genotypes, we require the joint distribution of observed and missing genotype data, and we make the modeling assumption that each individual's genotype vector can be considered independently of the others. That is,

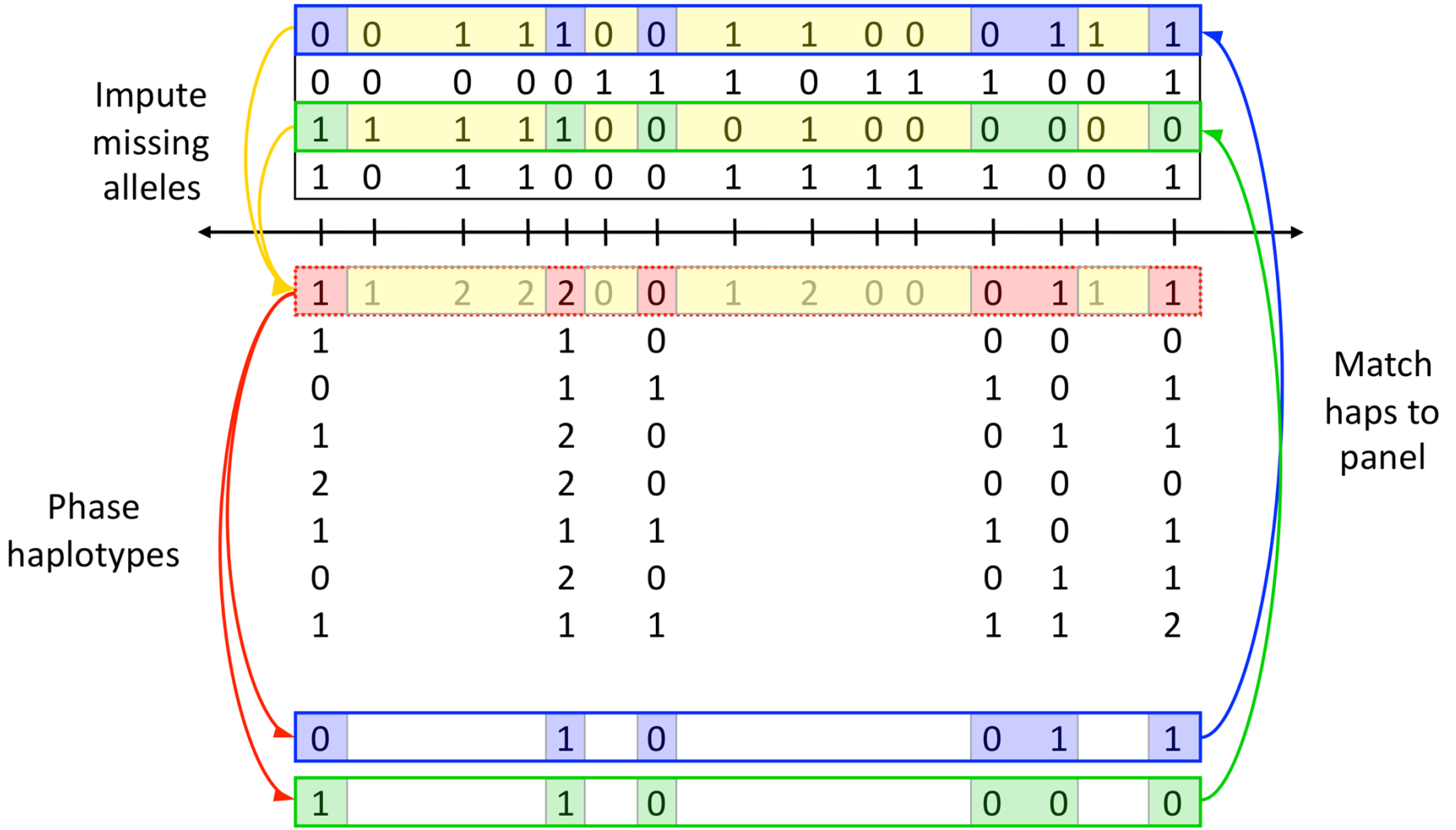
$$\Pr(G_M|G_O, H) \propto \Pr(G_M, G_O|H) = \Pr(G|H) = \prod_{i=1}^{K} \Pr(G_i|H)$$

Our model for each individual's genotype vector, $\Pr(G_i|H)$, is a Hidden Markov Model in which the hidden states are a sequence of pairs of the N known haplotypes in the set H. That is,

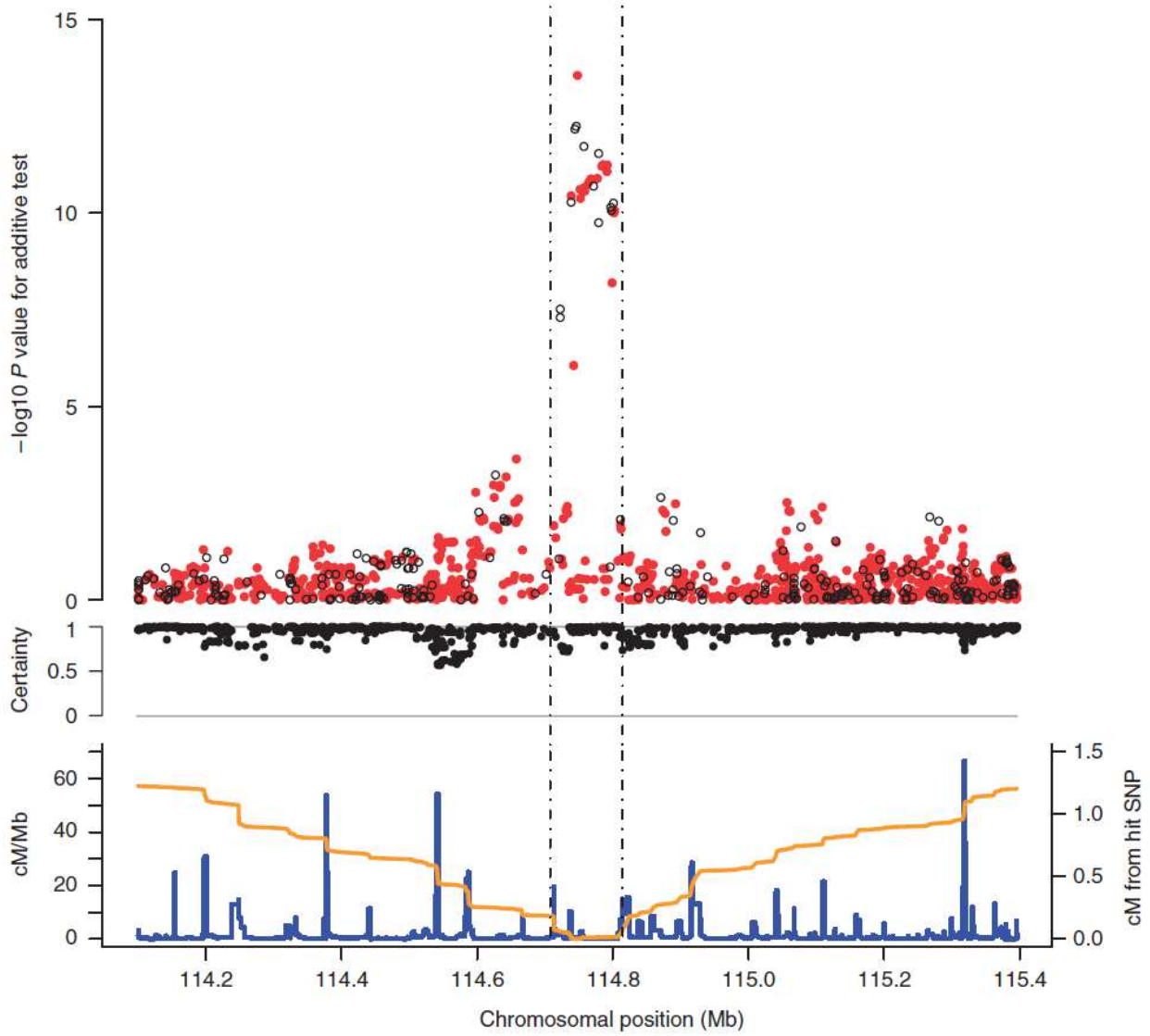
$$\Pr(G_i|H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} \Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H) \Pr(Z_i^{(1)}, Z_i^{(2)}|H)$$

where $Z_i^{(1)} = \{Z_{i1}^{(1)}, \ldots, Z_{iL}^{(1)}\}$ and $Z_i^{(2)} = \{Z_{i1}^{(2)}, \ldots, Z_{iL}^{(2)}\}$ are the two sequences of hidden states at the L sites and $Z_{il}^{(j)} \in \{1, \ldots, N\}$. These hidden states can be thought of as the pair of haplotypes in the set H that are being copied to form the genotype vector G_i. The term $\Pr(Z_i^{(1)}, Z_i^{(2)}|H)$ defines our prior probability on how sequences of hidden states change along the sequence, and $\Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$ models how the observed genotypes will be close to but not exactly the same as the haplotypes being copied. This model extends a

Marchini et al. Nat Genetics 2007

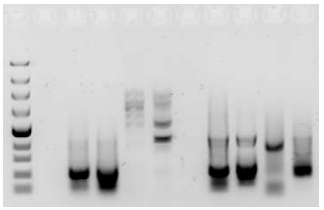


Extracted from Bryan Howie website

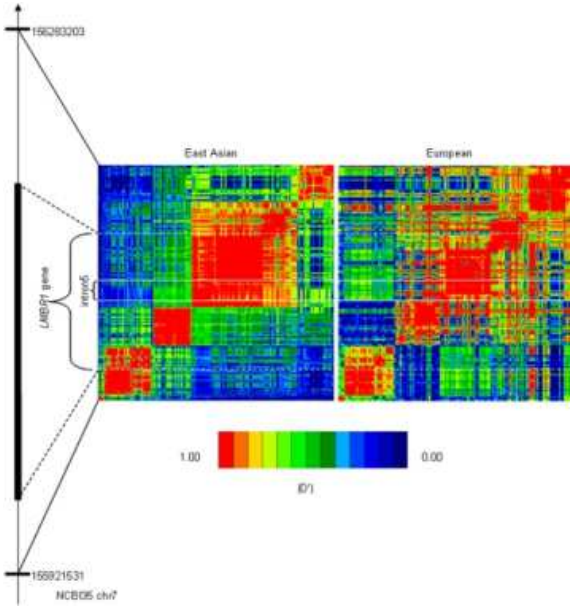


Marchini et al. Nat Genetics 2007

Una visión histórica



EcoRI
XbaI
PvuII
BamII
...



80's — *90's* — *00's* — *10's*

PCR — Human Genome Project — HapMap Project — 1000 Genomes Project

1 ---------- $4x10^7$

Candidate gene studies — **Hypothesis free GWAS** — **Fine mapping**

Si queréis más información:


clluis@imim.es | carla.lluis@gmail.com


www.linkedin.com/in/carlalluis


Lluis-Ganella C[Author]

Fi de la presentació