The International HapMap Project

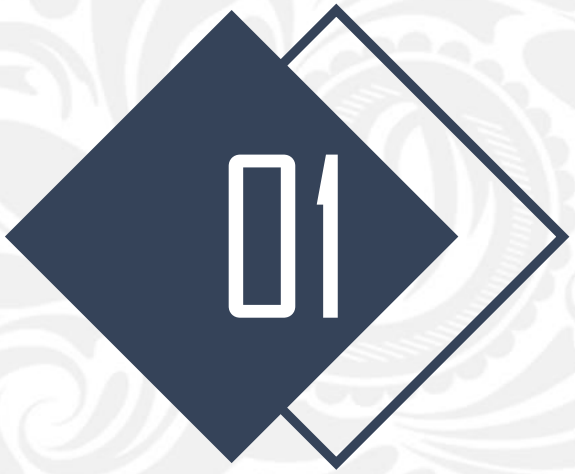
国际单倍体图谱项目

期刊：Nature
IF：49.926
发表时间：2005.10

汇报组：6组
组员： 宋方媛 岳琪桢 田李美丽
刘雪萌 李润础

目录

CONTENTS



研究背景及目的

了解HapMap

01 研究背景

1. 现在疾病研究的现状

2. 现有方法的弊端

了解HapMap

02 国际单倍体图谱创建的目的

1. 目的是建立一个公共的、全基因组的人类常见序列变异数据库，为临床表型的遗传研究提供必要的信息。

2. HapMap是人类基因组中常见遗传多态位点的目录，它描述了这些变异的形式、在DNA上存在的位置、在同一群体内部和不同人群间的分布状况。HapMap计划并不是利用HapMap中的信息来建立特定的遗传变异与某一疾病之间的联系，而是为其他研究者提供相关信息使之能够将遗传多态位点和特定疾病风险联系起来，从而为预防、诊断和治疗疾病提供新的方法。

两个阶段

1. HapMap项目的第 I 阶段： 5000bp密度

2. HapMap项目的第 II 阶段：密度增至2000bp

02

研究内容

1. HapMap第一阶段

- HapMap项目的第一阶段设定了一个目标，即在269个DNA样本中的每个基因组中至少对一个5千碱基（kb）的常见SNP进行基因分型。

- 出于实用性的考虑，并受人类基因组中变异体等位基因频率分布的影响，研究的目标是0.05或更高的次要等位基因频率（MAF）。

1. HapMap第一阶段

项目组织

Table 1 | Genotyping centres

Centre	Chromosomes	Technology
RIKEN	5, 11, 14, 15, 16, 17, 19	Third Wave Invader
Wellcome Trust Sanger Institute	1, 6, 10, 13, 20	Illumina BeadArray
McGill University and Génome Québec Innovation Centre	2, 4p	Illumina BeadArray
Chinese HapMap Consortium*	3, 8p, 21	Sequenom MassExtend, Illumina BeadArray
Illumina	8q, 9, 18q, 22, X	Illumina BeadArray
Broad Institute of Harvard and MIT	4q, 7q, 18p, Y, mtDNA	Sequenom MassExtend, Illumina BeadArray
Baylor College of Medicine with ParAllele BioScience	12	ParAllele MIP
University of California, San Francisco, with Washington University in St Louis	7p	PerkinElmer AcycloPrime-FP
Perlegen Sciences	5 Mb (ENCODE) on 2, 4, 7, 8, 9, 12, 18 in CEU	High-density oligonucleotide array

*The Chinese HapMap Consortium consists of the Beijing Genomics Institute, the Chinese National Human Genome Center at Beijing, the University of Hong Kong, the Hong Kong University of Science and Technology, the Chinese University of Hong Kong, and the Chinese National Human Genome Center at Shanghai.

1. HapMap第一阶段

样本采集

269个样本来自
（1）尼日利亚伊巴丹的约鲁巴人（简称YRI）的90个个体（30个父母-子女三人组）；
（2）美国犹他州90个人（30个三人组）（简称CEU）；
（3）中国北京45名汉族人（简称CHB）；
（4）44日本东京的日本人（缩写为JPT）。

1. HapMap第一阶段

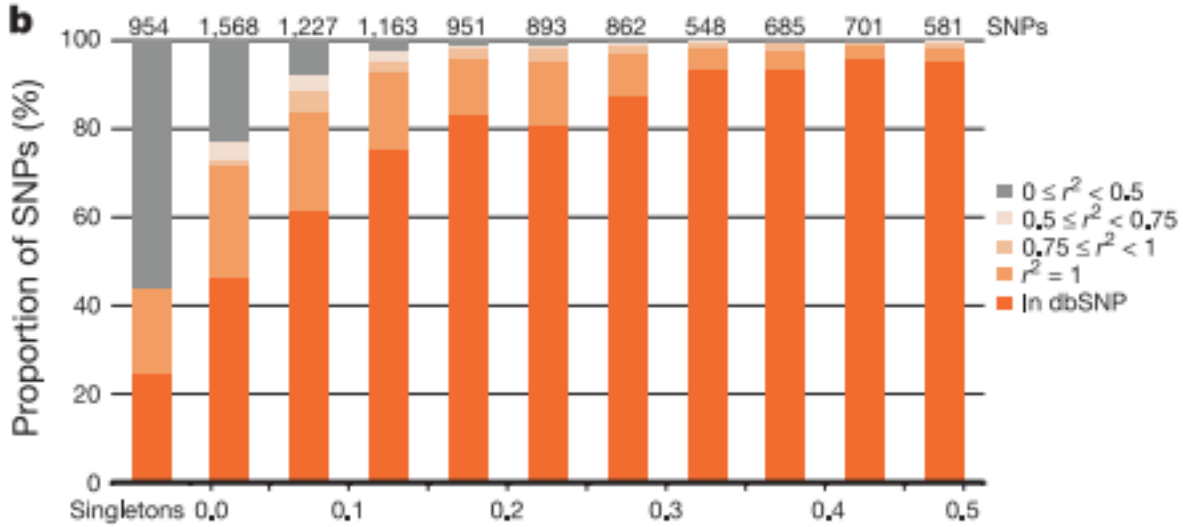
在ENCODE中选择具有代表性的10个区域，并进行测序

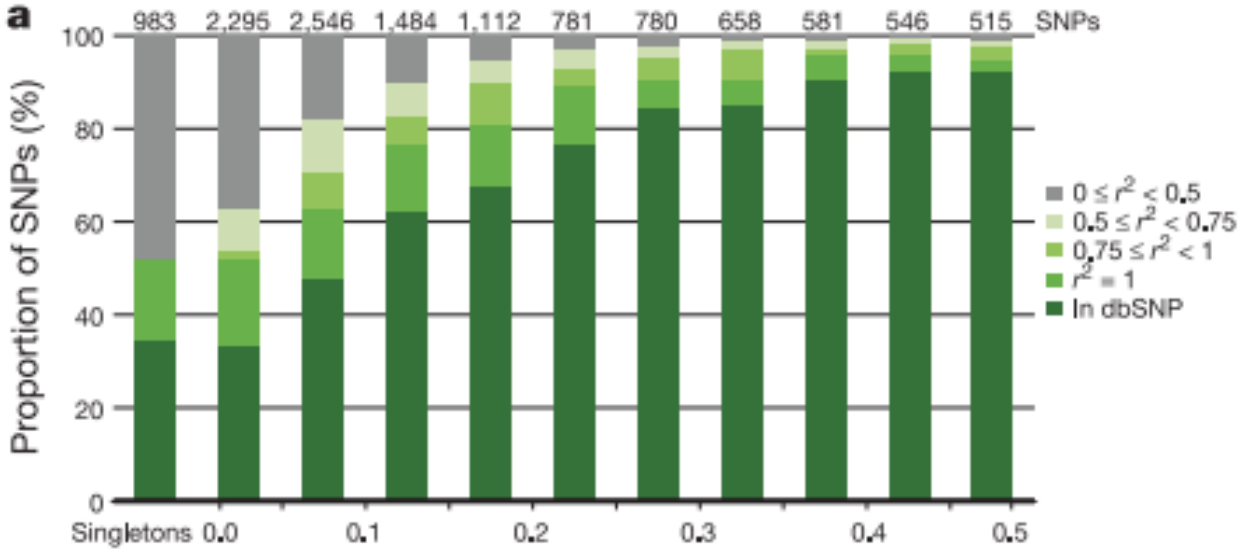
Table 2 | ENCODE project regions and genotyping

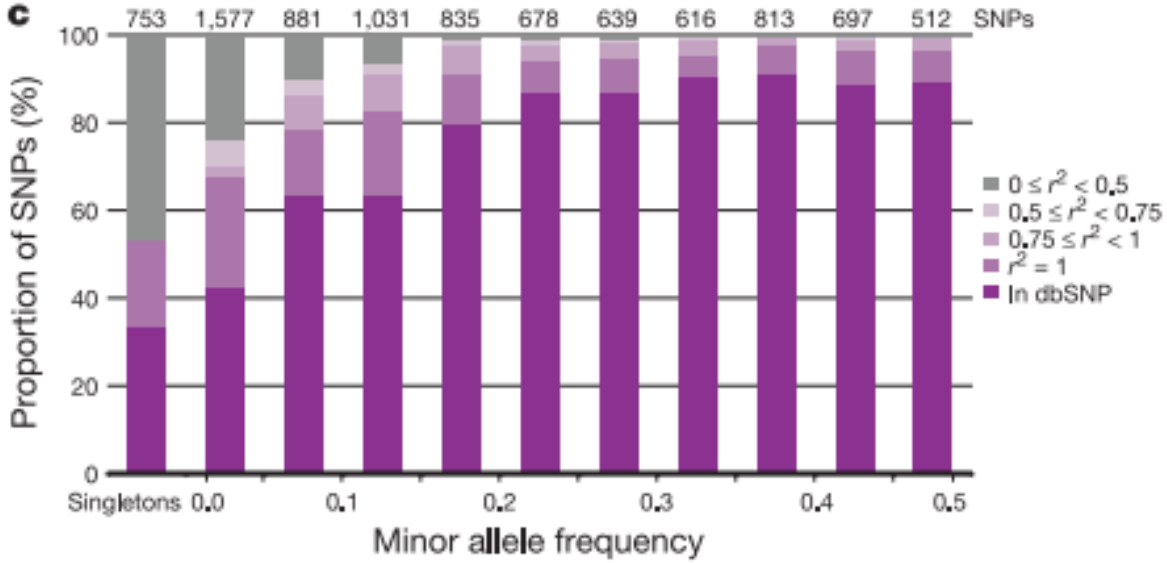
Region name	Chromosome band	Genomic interval (NCBI) (base numbers)†	Gene density (%)‡	Conservation score (%)§	Pedigree-based recombination rate (cM Mb^{-1})‖	Population-based recombination rate (cM Mb^{-1})¶	G+C content#	Available SNPs			Successfully genotyped SNPs††	Sequencing centre/ genotyping centre(s)‡‡
								dbSNP☆	Sequence**	Total		
ENr112	2p16.3	51,633,239–52,133,238	0	3.8	0.8	0.9	0.35	1,570	1,762	3,332	2,275	Broad/ McGill-GQIC
ENr131	2q37.1	234,778,639–235,278,638	4.6	1.3	2.2	2.5	0.43	1,736	1,259	2,995	1,910	Broad/ McGill-GQIC
ENr113	4q26	118,705,475–119,205,474	0	3.9	0.6	0.9	0.35	1,444	2,053	3,497	2,201	Broad/ Broad
ENm010	7p15.2	26,699,793–27,199,792	5.0	22.0	0.9	0.9	0.44	1,220	1,795	3,015	1,271	Baylor/ UCSF-WU, Broad
ENm013*	7q21.13	89,395,718–89,895,717	5.5	4.4	0.4	0.5	0.38	1,394	1,917	3,311	1,807	Broad/ Broad
ENm014*	7q31.33	126,135,436–126,632,577	2.9	11.2	0.4	0.9	0.39	1,320	1,664	2,984	1,966	Broad/ Broad
ENr321	8q24.11	118,769,628–119,269,627	3.2	11.4	0.6	1.1	0.41	1,430	1,508	2,938	1,758	Baylor/ Illumina
ENr232	9q34.11	127,061,347–127,561,346	5.9	8.3	2.7	2.6	0.52	1,444	1,523	2,967	1,324	Baylor/ Illumina
ENr123	12q12	38,626,477–39,126,476	3.1	1.7	0.3	0.8	0.36	1,877	1,379	3,256	1,792	Baylor / Baylor
ENr213	18q12.1	23,717,221–24,217,220	0.9	7.4	1.2	0.9	0.37	1,330	1,459	2,789	1,640	Baylor/ Illumina
Total	–	–	–	–	–	–	–	14,765	16,319	31,084	17,944	–

1. HapMap第一阶段

ENCODE所得SNP与dbSNP之间的比较







1. HapMap第一阶段

对269样本进行基因分型并筛选

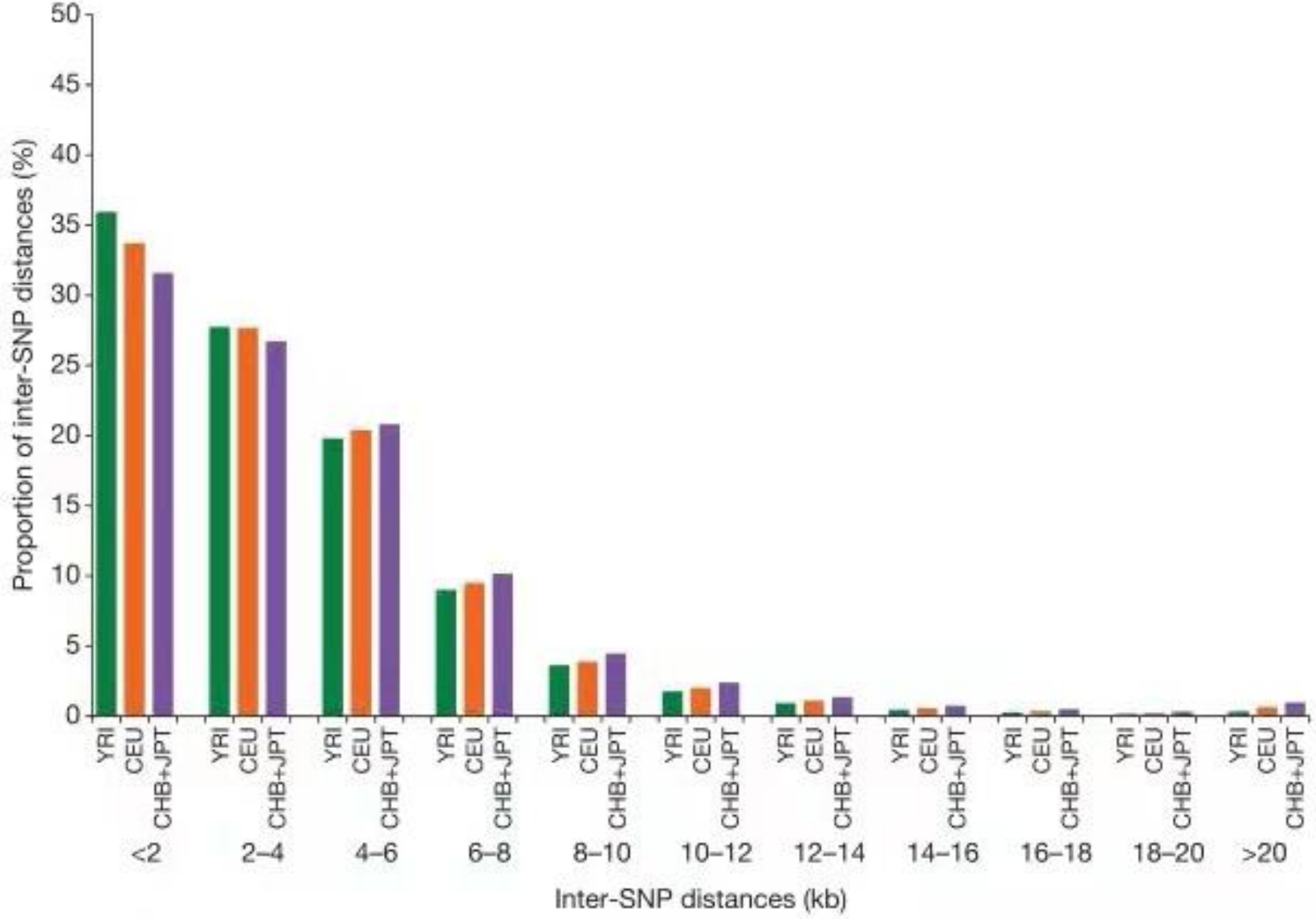
Table 3 | HapMap Phase I genotyping success measures

	Analysis panel		
SNP categories	YRI	CEU	CHB + JPT
Assays submitted	1,273,716	1,302,849	1,273,703
Passed QC filters	1,123,296 (88%)	1,157,650 (89%)	1,134,726 (89%)
Did not pass QC filters*	150,420 (12%)	145,199 (11%)	138,977 (11%)
> 20% missing data	98,116 (65%)	107,626 (74%)	93,710 (67%)
> 1 duplicate inconsistent	7,575 (5%)	6,254 (4%)	10,725 (8%)
> 1 mendelian error	22,815 (15%)	13,600 (9%)	0 (0%)
< 0.001 Hardy–Weinberg *P*-value	12,052 (8%)	9,721 (7%)	16,176 (12%)
Other failures†	23,478 (16%)	17,692 (12%)	23,722 (17%)
Non-redundant (unique) SNPs	1,076,392	1,104,980	1,087,305
Monomorphic	156,290 (15%)	234,482 (21%)	268,325 (25%)
Polymorphic	920,102 (85%)	870,498 (79%)	818,980 (75%)
	All analysis panels		
Unique QC-passed SNPs	1,156,772		
Passed in one analysis panel	52,204 (5%)		
Passed in two analysis panels	97,231 (8%)		
Passed in three analysis panels	1,007,337 (87%)		
Monomorphic across three analysis panels	75,997		
Polymorphic in all three analysis panels	682,397		
MAF ≥ 0.05 in at least one of three analysis panels	877,351		

*Out of 95 samples in CEU, YRI; 94 samples in CHB + JPT.
†'Other failures' includes SNPs with discrepancies during the data transmission process. Some SNPs failed in more than one way, so these percentages add up to more than 100%.

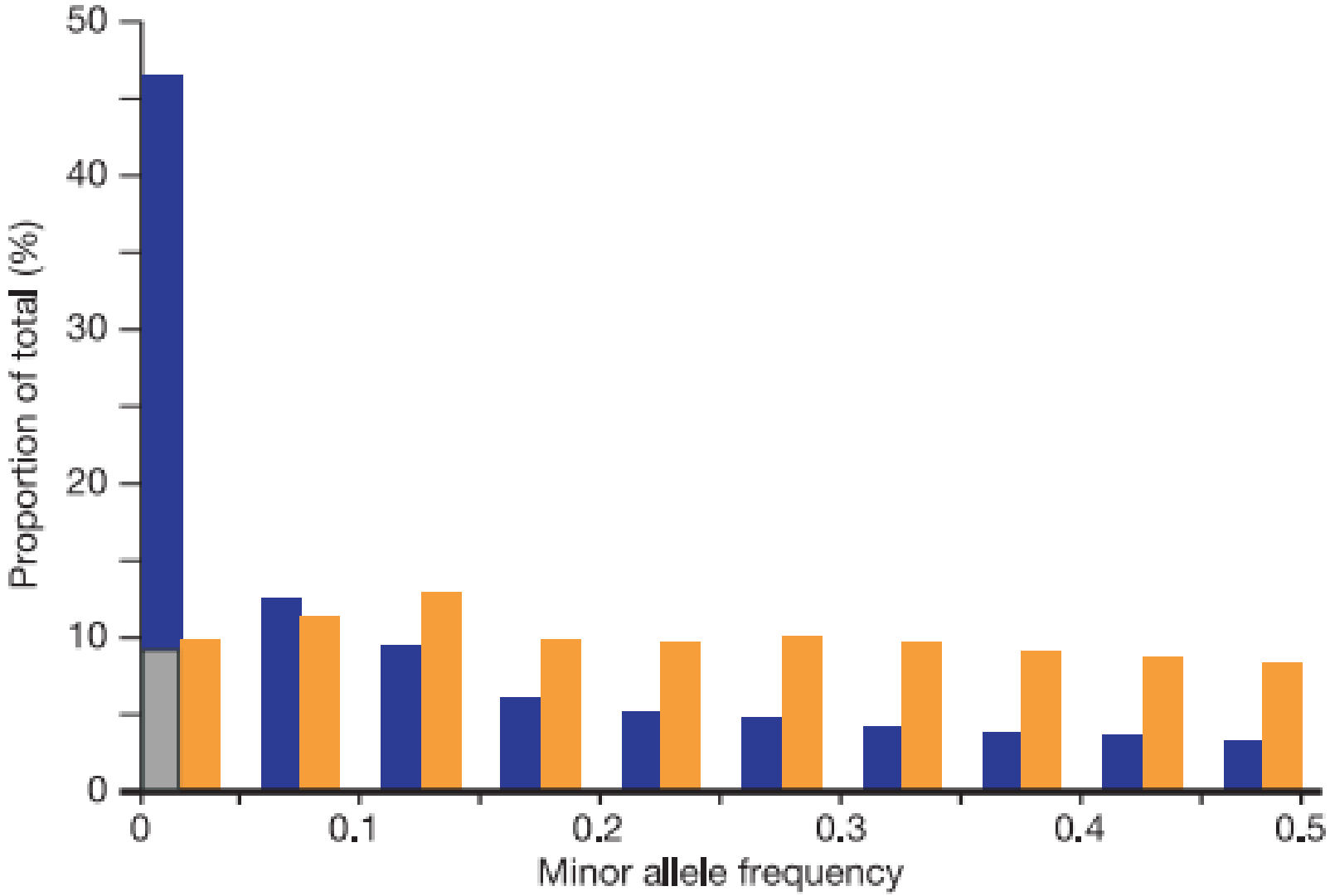
1. HapMap第一阶段

Hapmap中SNP之间的距离



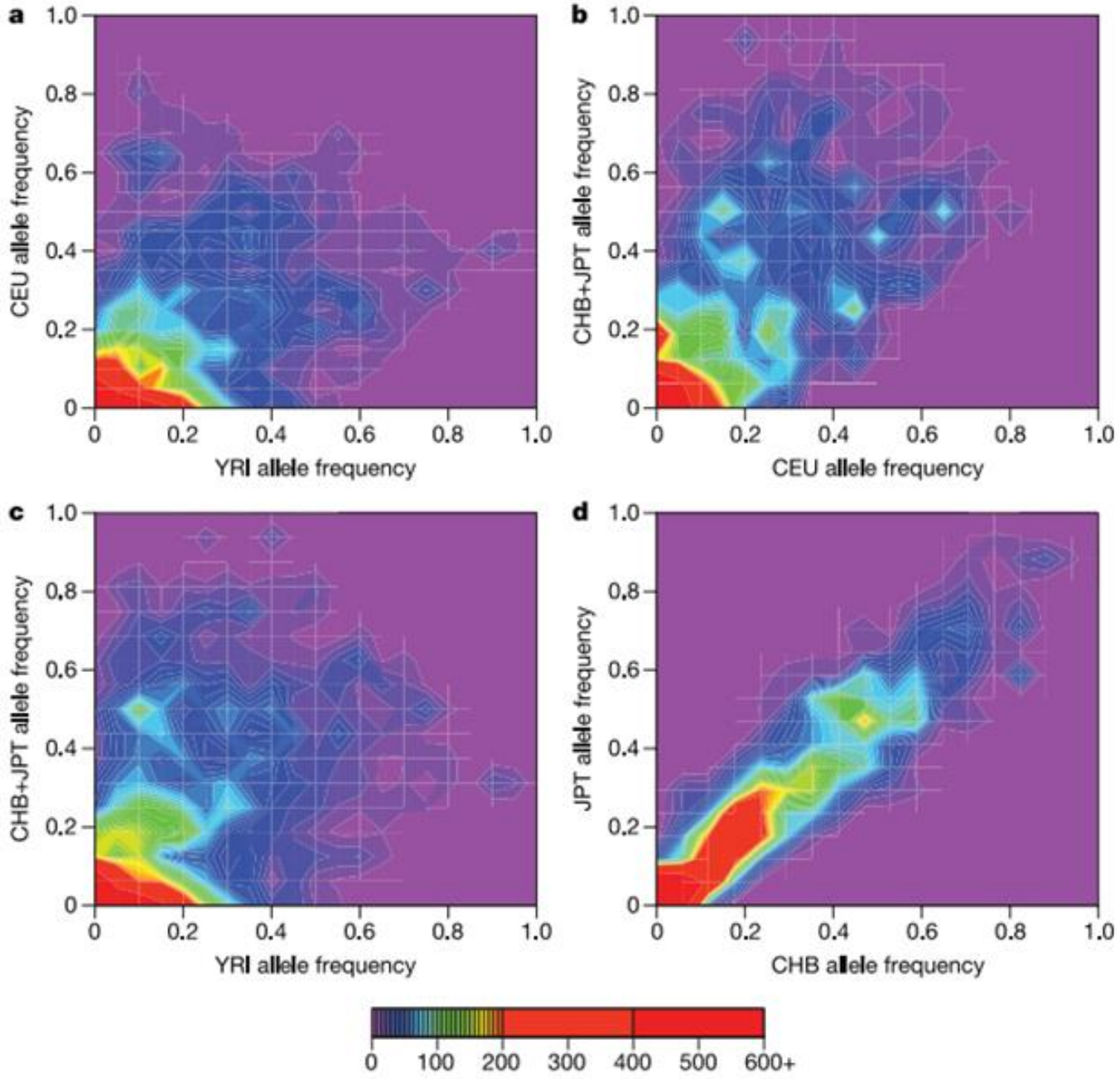
1. HapMap第一阶段

SNP种类及整体贡献率



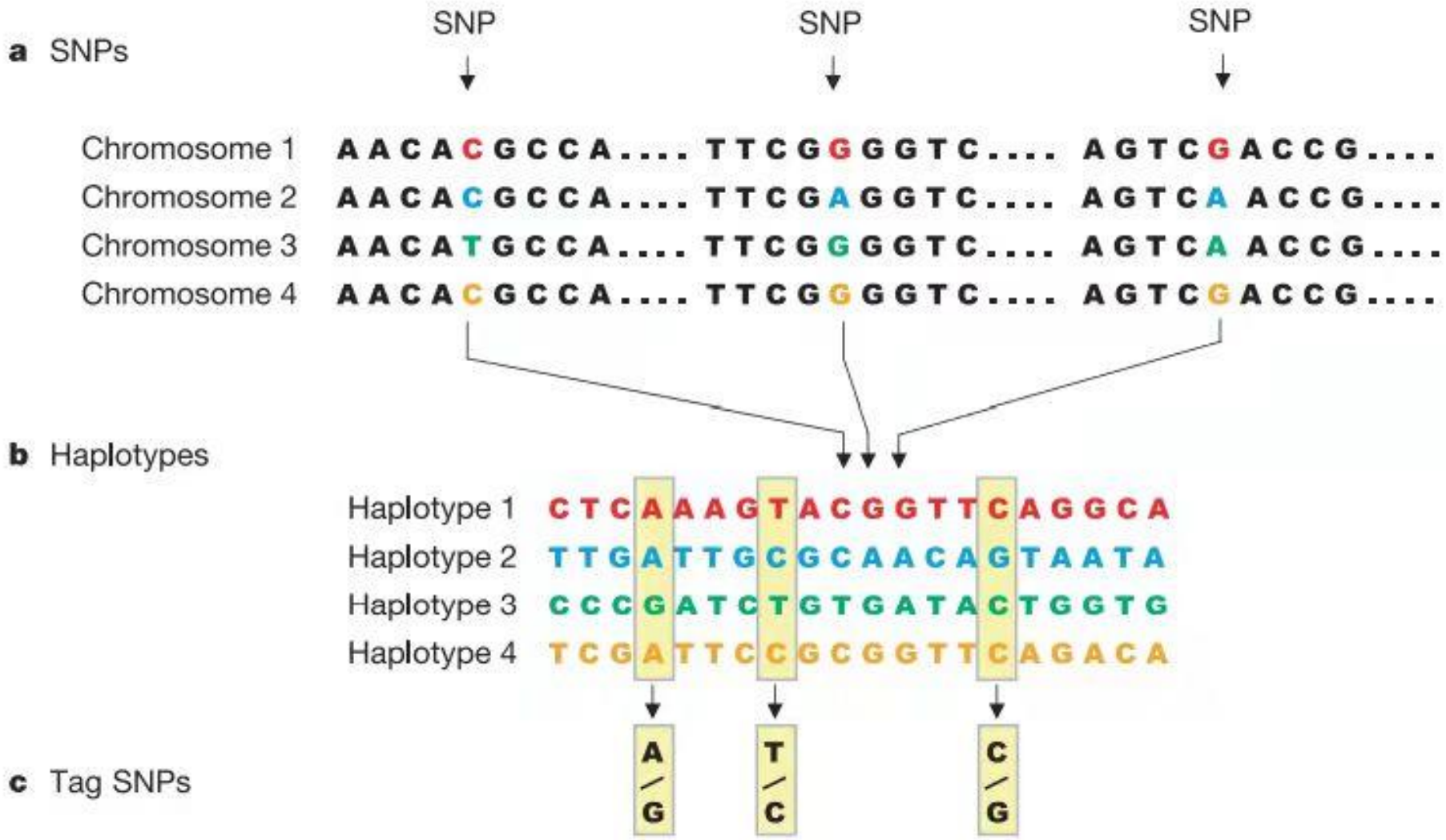
1. HapMap第一阶段

对所有分析小组之间的
等位基因频率的比较



1. HapMap第一阶段

SNP、Haplotypes以及 tag SNP



2. LD在人类基因组中的特性

- 何为LD?

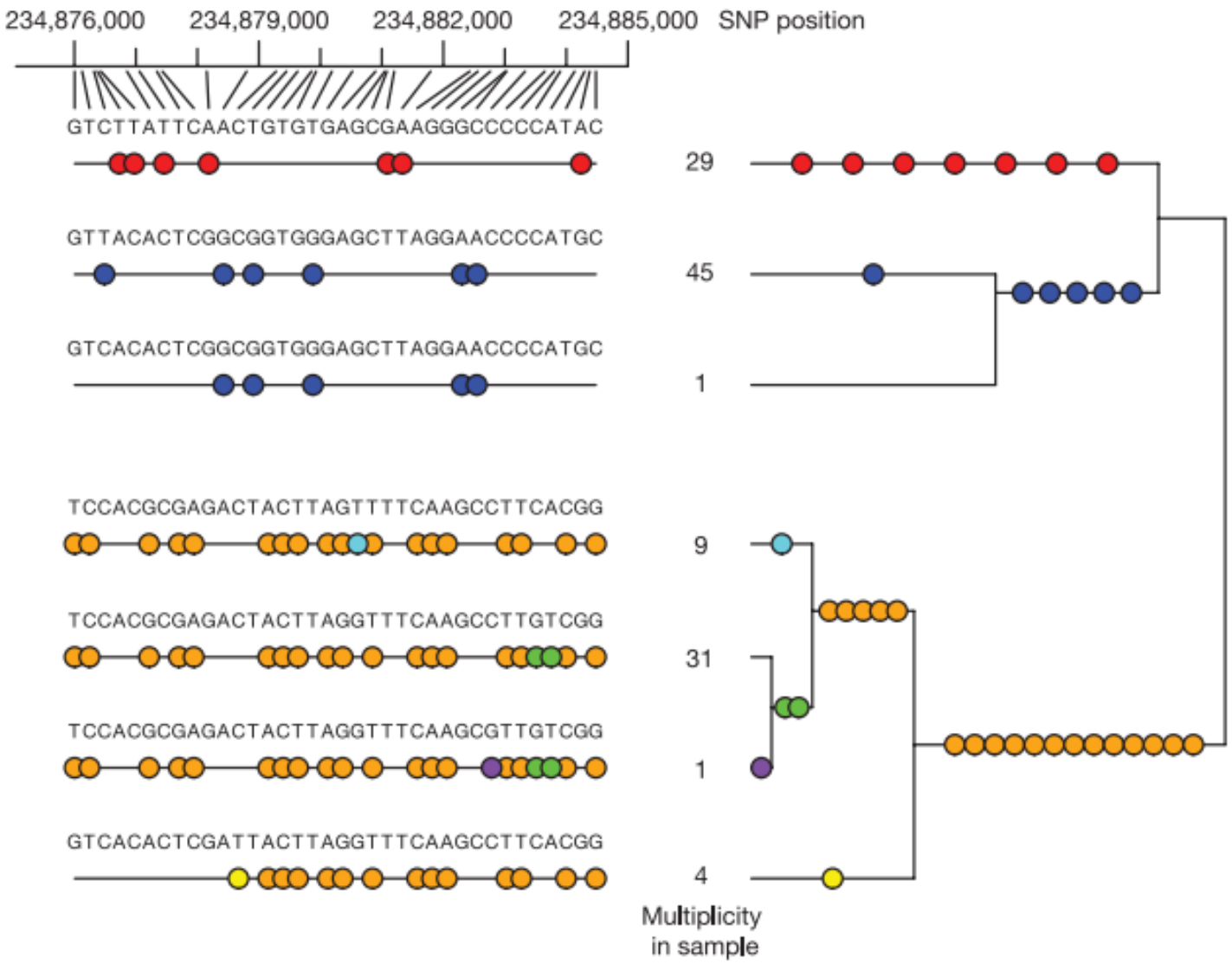
不同座位上某两个等位基因出现在同一条单元型上的频率与预期的随机频率之间存在明显差异的现象，称连锁不平衡 (Linkage disequilibrium, LD) 。

- 几个常用于度量LD的符号：D’，r^2，LOD

D’和r^2值为零时，连锁完全平衡；
D’和r^2值为1时，连锁完全不平衡。
LOD值为0，意味着连锁假设与不连锁假设的可能性相等；
LOD>1时，表示存在连锁；
LOD>3时，表示肯定连锁；
LOD<-2时，表示否定连锁。

2. LD在人类基因组中的特性

在没有强制性重组的区域中，单倍型和r^2值之间的遗传关系



2. LD在人类基因组中的特性

ENr131.2q37.1 和 ENm014.7q31.33 的重组比较

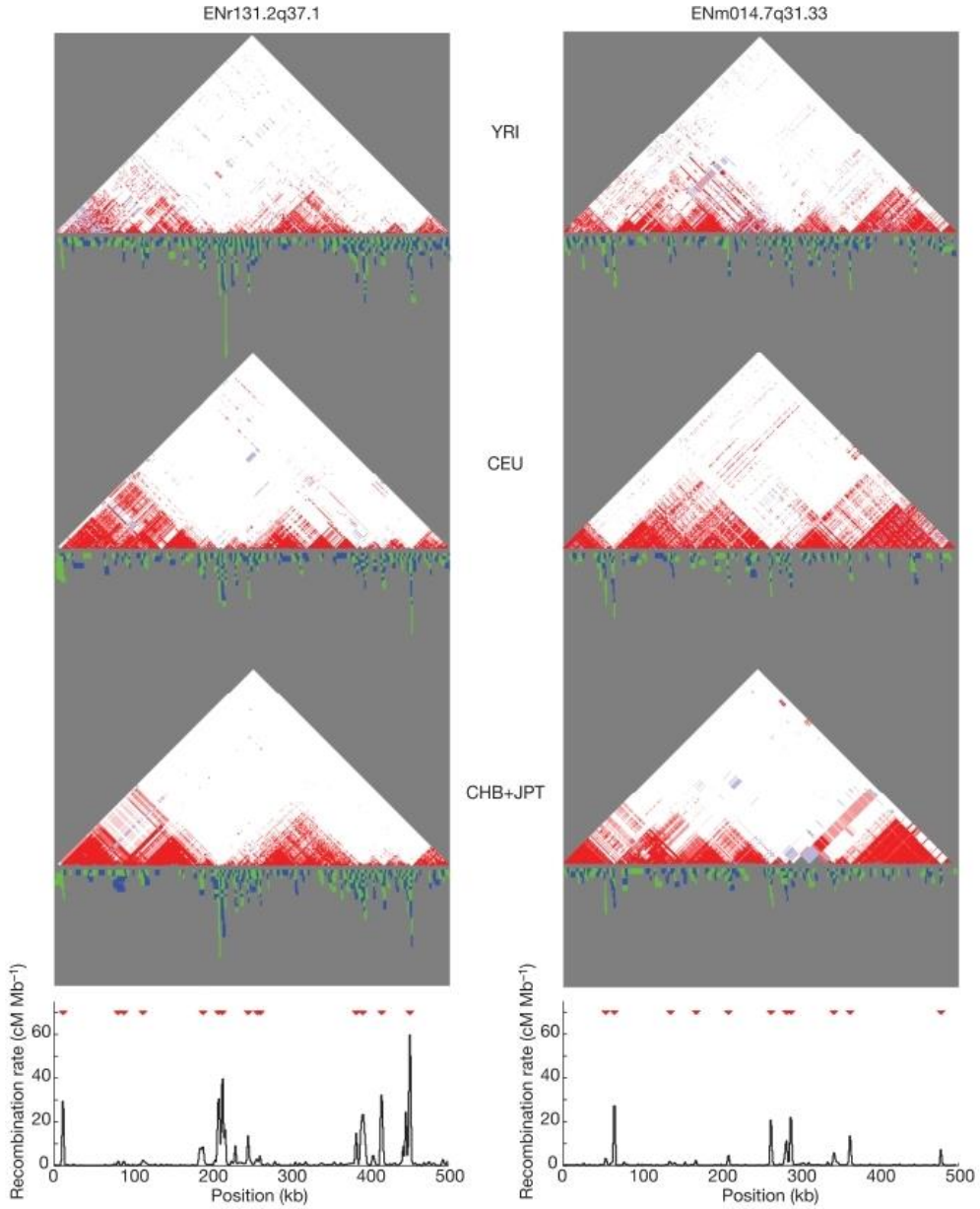
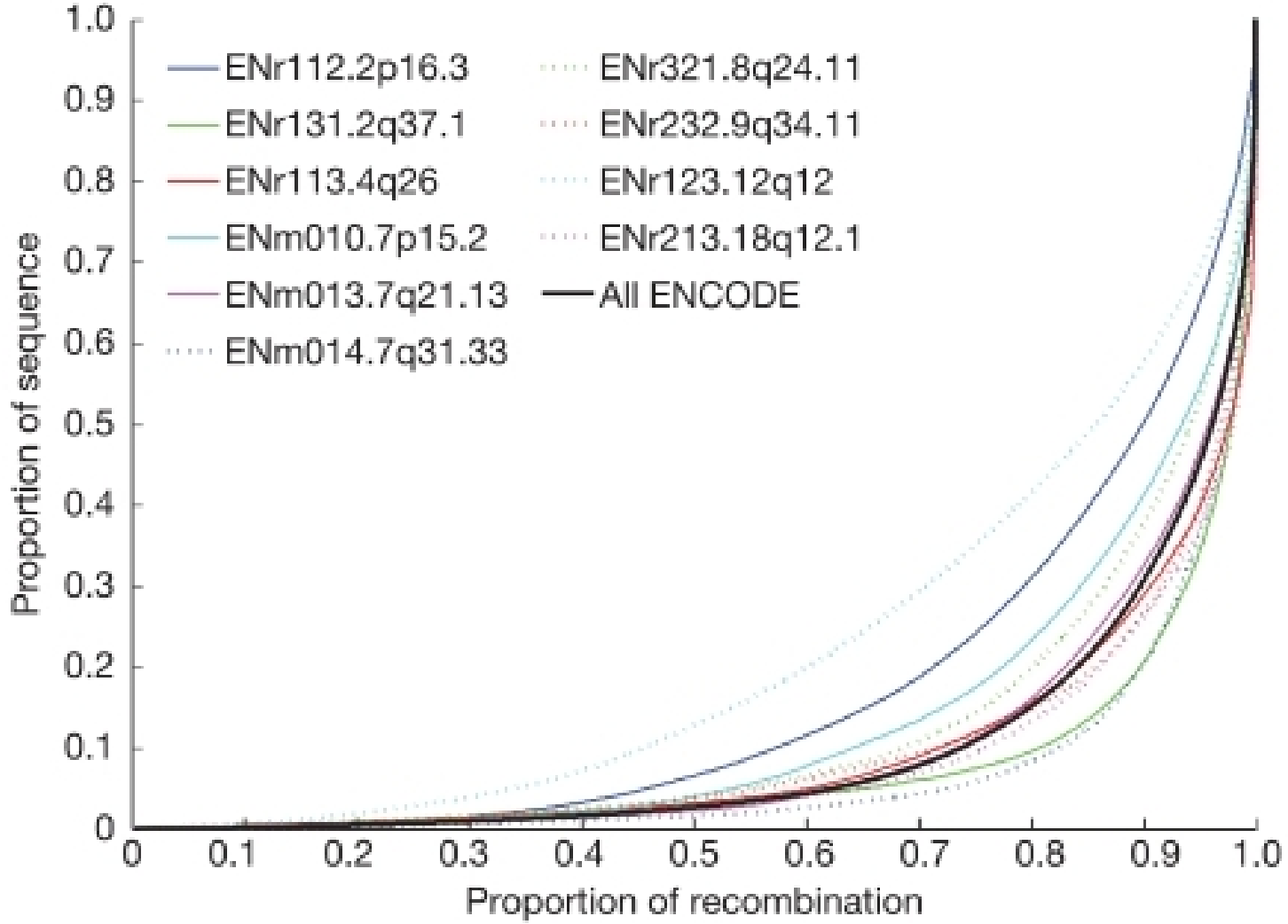


Figure 8 | Comparison of linkage disequilibrium and recombination for two ENCODE regions. For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB+JPT analysis panels are shown: white, $D' < 1$ and LOD < 2; blue, $D' = 1$ and LOD < 2; pink, $D' < 1$ and LOD ≥ 2; red, $D' = 1$ and LOD ≥ 2. Below each of these plots is shown the intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles[46].

2. LD在人类基因组中的特性

重组在ENCODE的分布



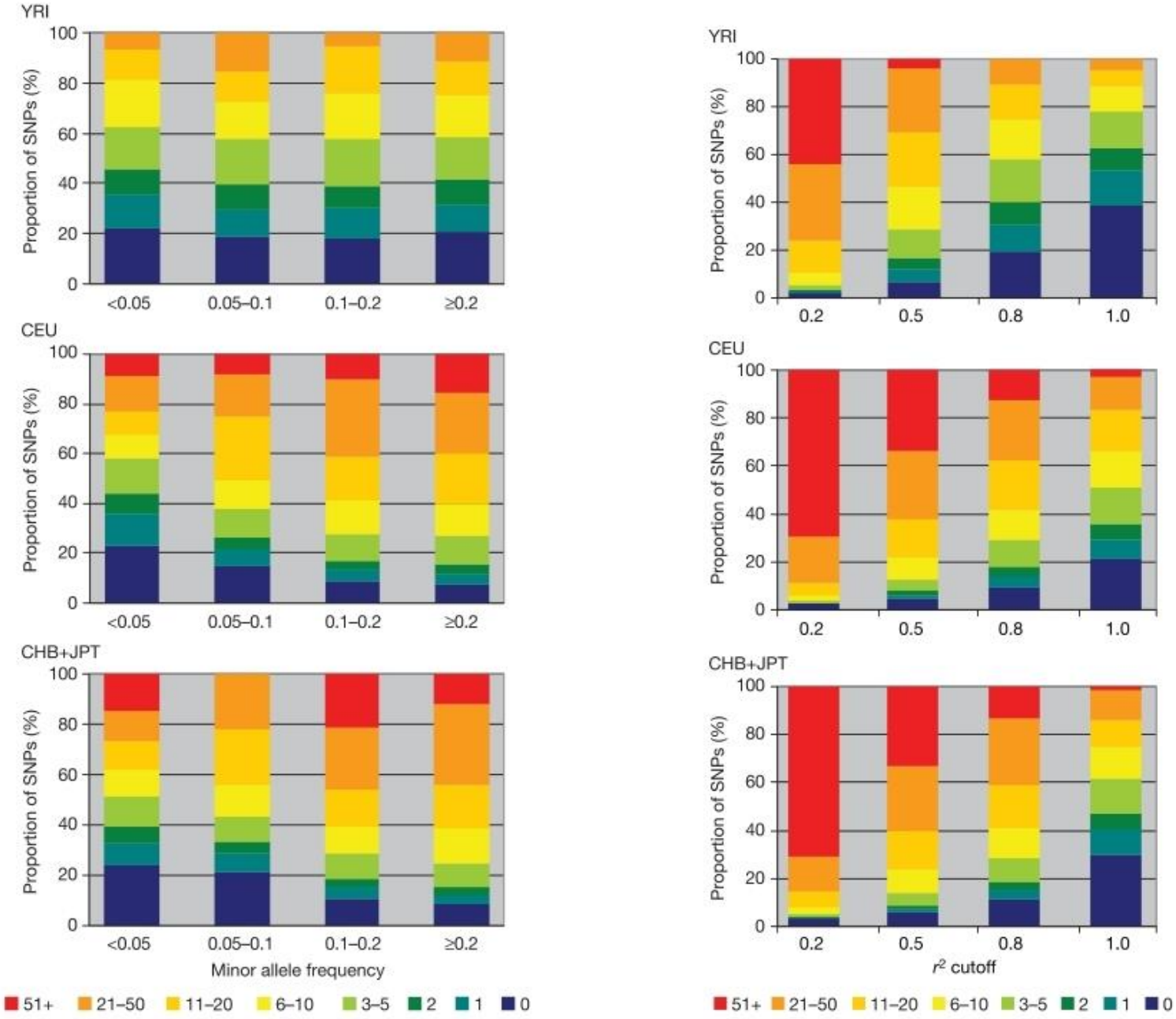
3.关于LD 中SNP的相关性

- ENCODE 数据显示，SNP 通常与附近的几个 SNP 完全相关，并且部分关联到许多其他 SNP。

- 我们使用“代理”来表示与一个或多个其他SNP关联性强。当两个SNP完全相关时，测试一个完全等同于测试另一个SNP：我们称此类 SNP 集合为“完美的代理集"。

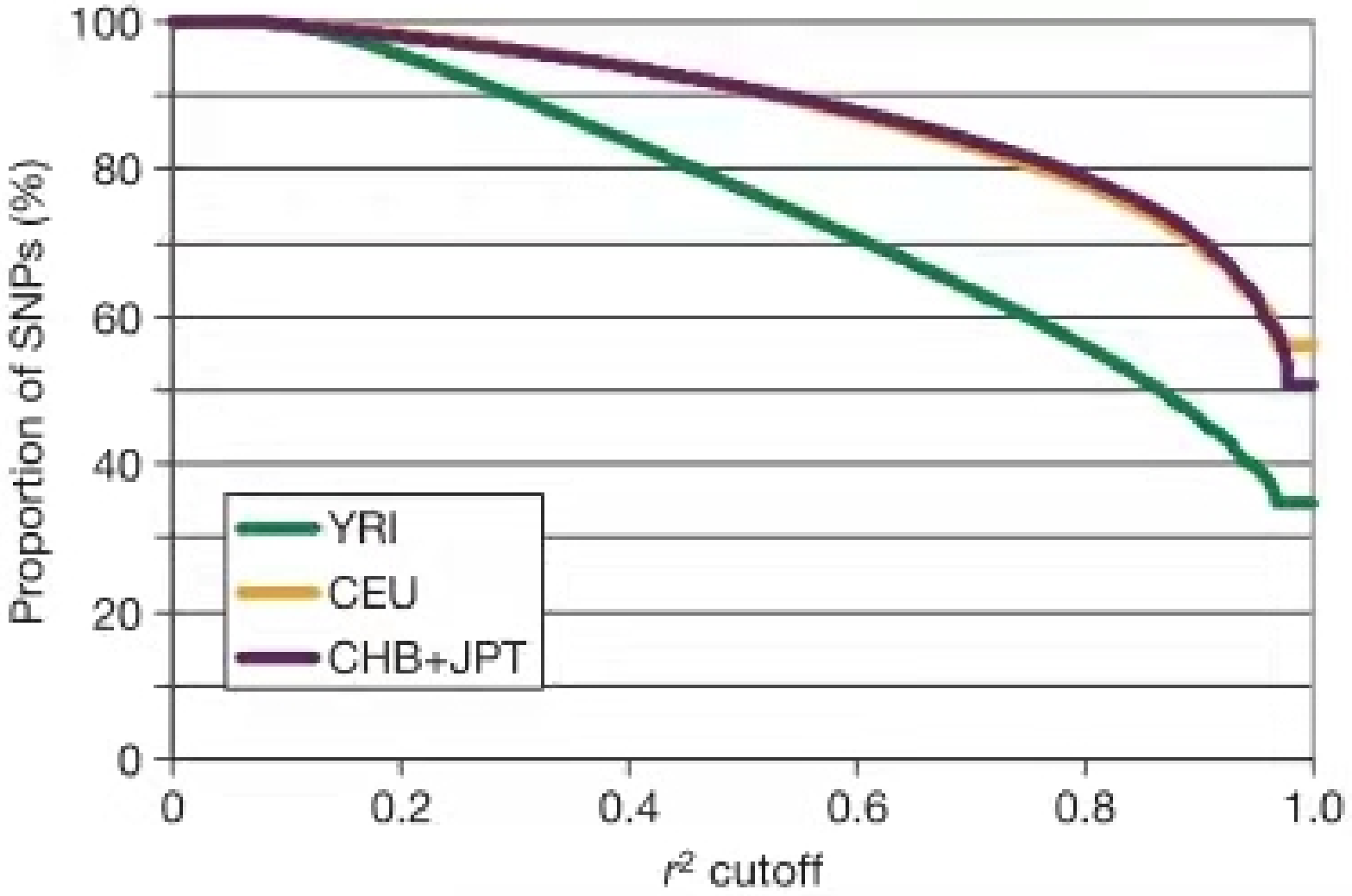
3.关于LD 中SNP的相关性

各样本集中具有代理作用的SNP的数量



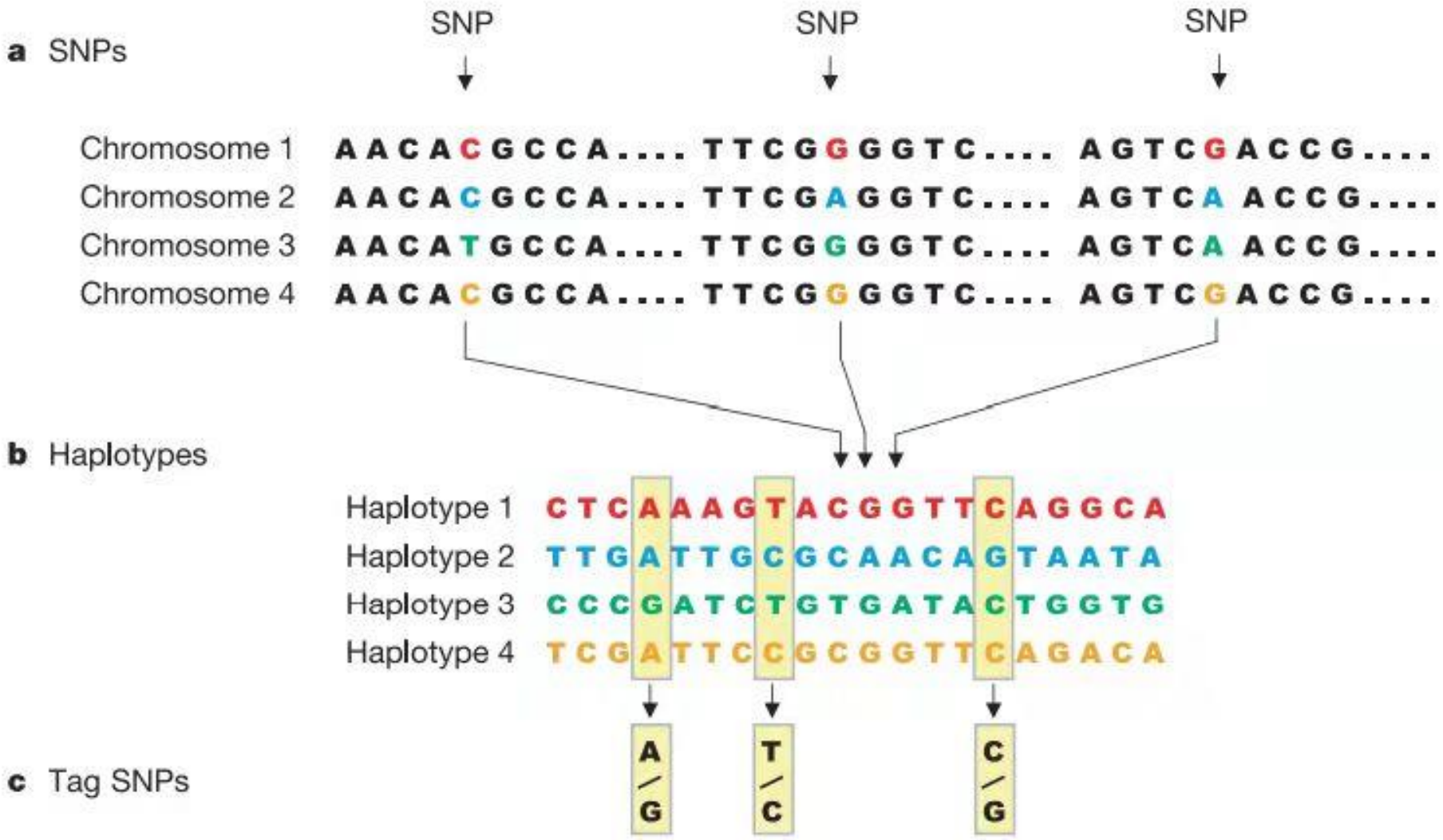
3.关于LD 中SNP的相关性

在Hapmap第一阶段中代理SNP与所有SNP的相关性的衡量



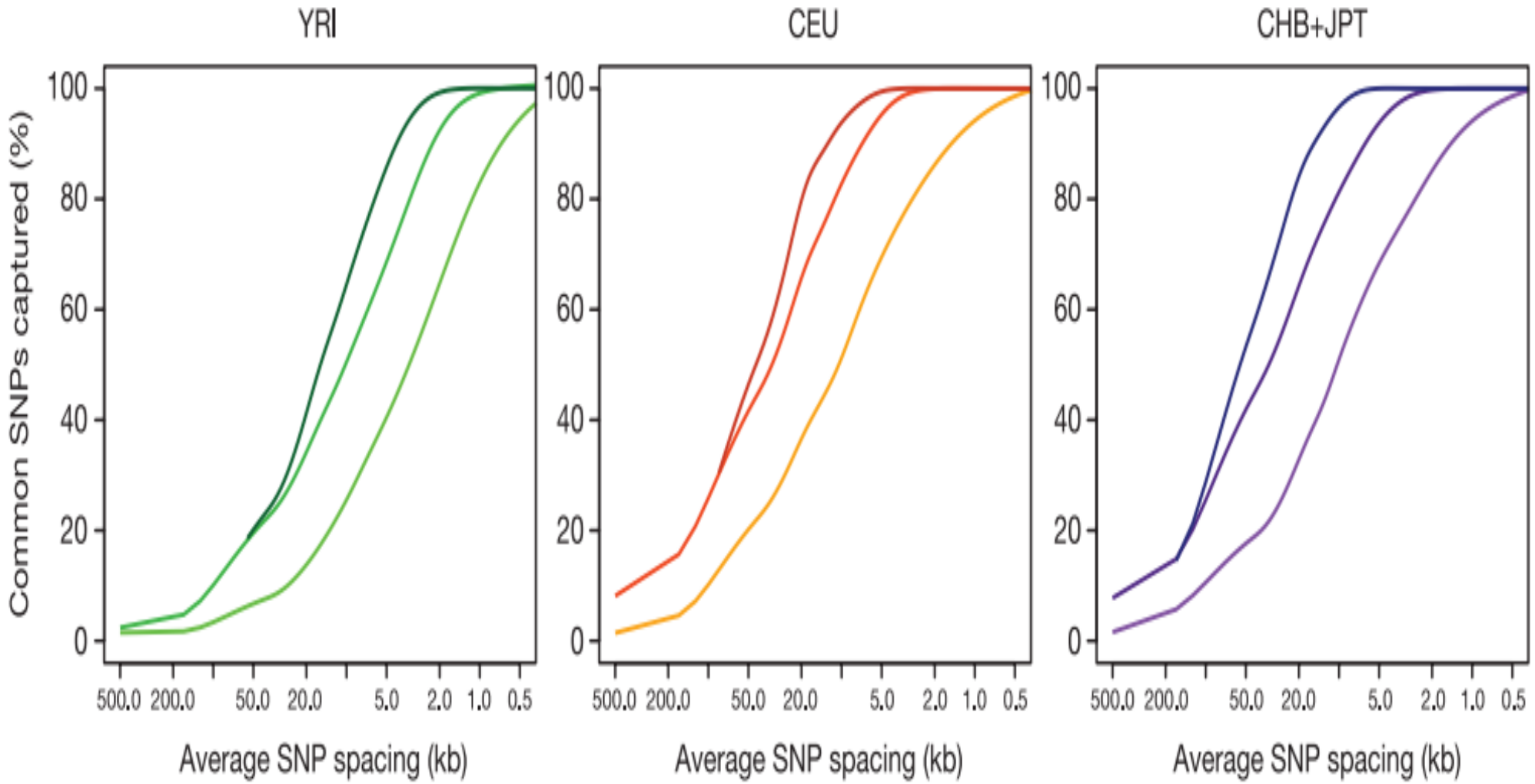
4.用于关联研究的tag SNP 选择

tag SNP 具有一定关联性的SNPs，居于代表性的SNP为tag SNP



4.用于关联研究的tag SNP 选择

不同样本集标签SNP的评估



4.用于关联研究的tag SNP 选择

提高效率的三种方法

1、放宽用于标记SNP选择的r^2阈值大大减少了被选择的标记SNP的数量

Table 7 | Number of selected tag SNPs to capture all observed common SNPs in the Phase I HapMap

r^2 threshold*	YRI	CEU	CHB + JPT
$r^2 \geq 0.5$	324,865	178,501	159,029
$r^2 \geq 0.8$	474,409	293,835	259,779
$r^2 = 1.0$	604,886	447,579	434,476

2、利用多标记单倍型

3、根据捕获的其他SNP的数量对标签进行优先排序

03

研究应用

研究应用

一、可以为疾病相关性研究的后续分析和解释提供信息

二、HapMap数据还提供了关于重组和自然选择历史的线索

04 研究总结及意义

研究总结

- 国际单体型基因图项目开始创建一种资源，以加速识别影响医学特征的遗传因素。
- 本文报道的分析证实了重组热点、长片段强连锁不平衡和有限的单倍型多样性的普遍性。以及通过选择标签SNPs和优化关联分析提高效率。
- 同时，为疾病分析以及自然选择的历史研究奠定了基础。

感谢您

Listen attentiv
ely
宝贵时间