基于SNP的连锁不平衡分析

刘智广

第四军医大学药学系
School of Pharmacy The Fourth Military Medical University

主讲内容

一、单核苷酸多态性

二、连锁不平衡分析

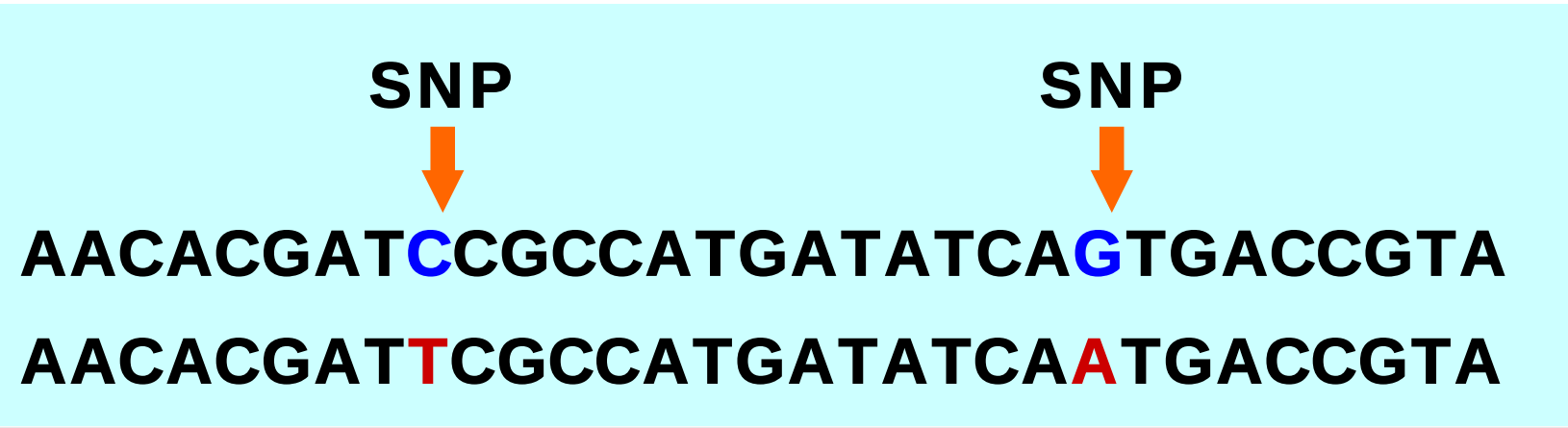
三、单体型分析

四、应用举例

一、单核苷酸多态性
(single nucleotide polymorphisms,SNPs)

1. SNPs概念

⑤ SNPs指染色体DNA 序列中的某个位点由于单个核苷酸的变化而引起的多态性，在群体中的频率>1%。



SNPs的基本类型

⑤ SNPs属于二等位基因,有两种基本类型:

转换：嘧啶置换嘧啶 C-T

嘌呤置换嘌呤 G-A

颠换：嘧啶与嘌呤互换

C-A(G-T)

C-G(G-C)

T-A(A-T)

⑤ 转换:颠换＝2:1

GpC岛SNPs发生率较高，约占总SNPs 25%，主要是C-T。可能胞嘧啶是最易发生突变位点；且大多数是甲基化的，自发脱氨基形成胸腺嘧啶。

2. SNPs的特点

⑤ **数量多、分布广**：一个个体至少携带300万SNPs，平均300-1000pb有一个SNPs。有学者推测基因组约有1000万个SNPs。

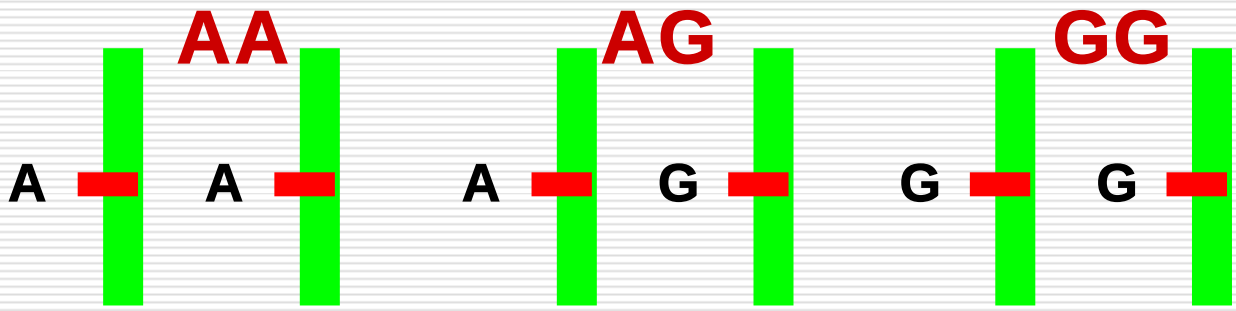
⑤ **相对稳定**：每一代中每个核苷酸变异频率极低(10^{-8})，且这种变化的随机性。

⑤ **易于快速筛查和基因分型**：SNPs的二态性标记，非此即彼。有利于实现高通量、自动化的筛查和分析。

3. SNPs的基因型

⑤ 人体除性染色体外，每个染色体都有两份，个体所拥有的一对等位基因的类型称作基因型。

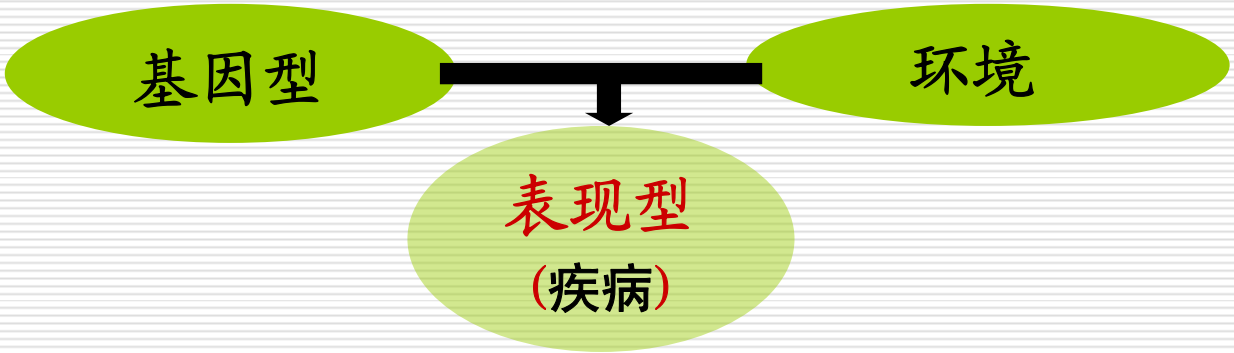
⑤ 例如，一SNPs(A/G)，则个体在该位点的基因型则：



⑤ 检定个体的基因型，被称作基因分型。

基因型与表现型

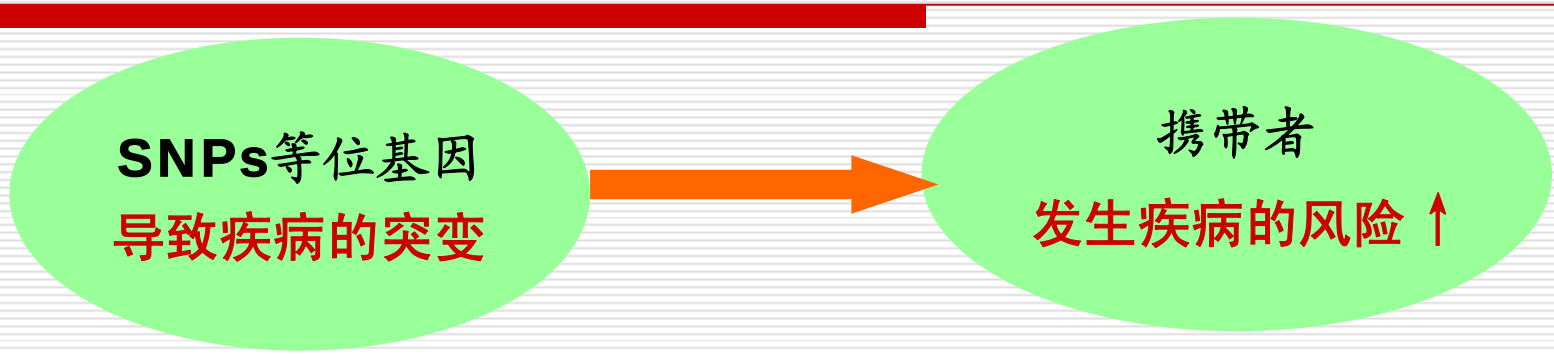
⑤ **表现型(表型)：**指由不同基因型与环境共同作用，而生物体可观测到的物理或生理性状（如疾病）。



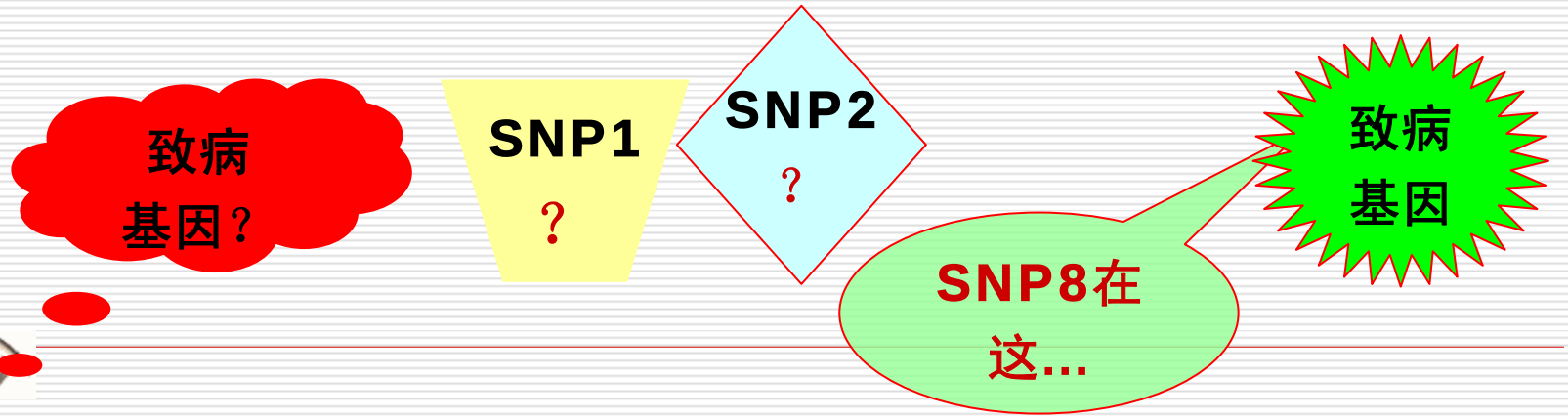
⑤ 寻找**基因型与表现型**的关系是遗传学的基本目标。

基因型 ⟷ 药物基因组学 ⟷ 药　物
(耐药、不良反应)

4. SNPs可用于发现致病基因



⑥ 大部分SNPs都不具有这种功能性的变异，但是可以作为寻找致病基因的标志（路标）。



⑥ 为了寻找致病基因所在的区域，可以将病人和正常人的SNPs等位基因的频率进行比较。

某些等位基因频率比正常人高 → 这些SNPs可能与该疾病相关

SNPs-疾病相关性提示：

致病基因可能存在于SNPs所在的染色体区域

5. SNPs分析: 基于实验的分析方法

未知SNPs

温度梯度凝胶电泳(TGGE)
变性梯度凝胶电泳(DGGE)
单链构象多态性(SSCP)
变性高效液相色谱检测(DHPLC)
限制性片段长度多态性(RFLP)
随机扩增多态性DNA(RAPD)
发现含有SNP的DNA链: 测序

已知SNPs

突变错配扩增检验
实时定量PCR技术
焦磷酸微测序技术
荧光偏振光技术
基因芯片技术

SNPs的实验分析方法
可用于基因型的分析

SNPs分析：基于公共数据库的方法

⑤ 利用数据库中的大量序列信息，采用生物信息学软件，

用计算机自动识别，是发现SNPs的新策略和重要方法。

⑤ 与癌症和肿瘤相关的候选SNP数据库：

http://cgap. ncbi. nih. gov/ GAI

⑤ 适于生物医学研究的SNP数据库：

http:// www. ncbi. nlm. nih. gov/ SNP

⑤ 人类SNP数据库：

http://hgbas. cgr. ki.sei 或http//hgbase.interactiva.de /

二、连锁不平衡分析

（一）连锁不平衡概念

连锁不平衡（ linkage disequilibrium, LD），又称等位基因关联，是指同一条染色体上，两个等位基因间的**非随机相关**。即，当位于同一条染色体的两个等位基因（**A，B**）同时存在的概率，大于人群中因随机分布而同时出现的概率时，就称这两个位点处于**LD**状态。

SNP1(A，a)　　**SNP2(B，b)**

假设：位于同一条染色体相邻两个SNP：

SNP1(A，a)　　SNP2(B，b)

③组合方式（单体型）：AB，Ab，aB：ab。

③如果A与B无LD：两个SNP的等位基因相互独立，随

机组合，概率为AB:Ab:aB:ab=0.25: 0.25:0.25:0.25，

AB组合的频率：fAB= fA×fB（等位基因频率）

③如果A与B存在LD：A与B连锁，当完全连锁时概率为AB:ab=0.5:0.5， AB组合的频率fAB= fA×fB +D，（D表示两位点间LD程度)

③LD定义式: D= fAB - fA×fB

LD的产生原因

⑥ LD 是由突变或重组形成的。在染色体某一SNP附近有新的突变产生时，则LD出现。

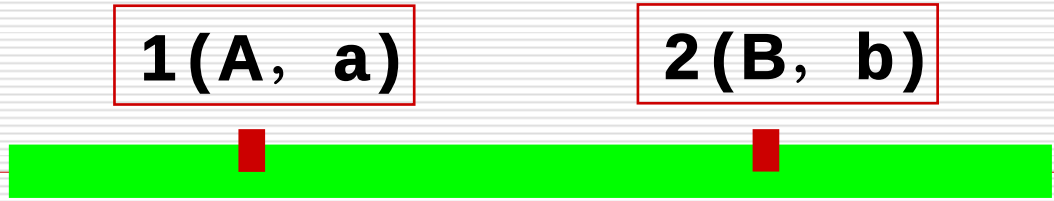
重组的发生：两位点间LD程度↓ 。

⑥ 理论上, LD强度与2个SNP间的距离有关：

距离越小：发生重组机会越小→ LD强；

距离越大：发生重组机会越大→ LD弱。

⑦ 实际上,也有距离很近不存在LD，而距离相当远（超过100 kb）存在LD。



LD的度量

⑥ LD的度量一般不直接使用LD定义式，而对D进行归一化后，用LD系数D'和r^2进行检验。

$$|D'| = D^2 / \min(fA\ fb,\ fa\ fB) \quad (D < 0)$$

$$|D'| = D^2 / \min(fA\ fB,\ fa\ fb) \quad (D > 0)$$

$$r^2 = D^2 / fA\ fa\ fB\ fb$$

⑥ 取值范围：0（无LD）——1（完全LD）。

D′ 的意义

⑥ D'是与频率无关的量，两位点间无重组时，D'＝1

⑥ D'＝1 称为完全LD，说明两个位点间没有发生重组；
两位点组成的单体型最多出现3种。

⑥ D'＝0 称为无LD或连锁平衡，即4种单倍型频率相等。

⑥ D'＜1 说明两位点间发生过重组或突变；
4种单倍型均可出现； D'相对值意义模糊 。

D ' 接近1： 提示：两位点间发生重组的可能性很小；

D '中间值：无法比较两位点LD 的差别。
D'值的95%可信区间(confidence inteeval,CI）进行比较。

D′值的95%可信区间（95%CI）

⑥ **95%CI**：对每对SNP，采用重复采样算法(一般大于1000次), 建立一个95%可信区间。

⑥ **95% CI的定义**：

区间上限值C_U＞0.98
区间下限值C_L＞0.70 } “强LD”

区间上限值C_U＜0.90： “重组证明明显”；

其 余： “无提示意义”。

r^2的意义：

⑥ r^2是与频率有关的量，在两位点间无重组时，r^2也不一定达到最大值1。

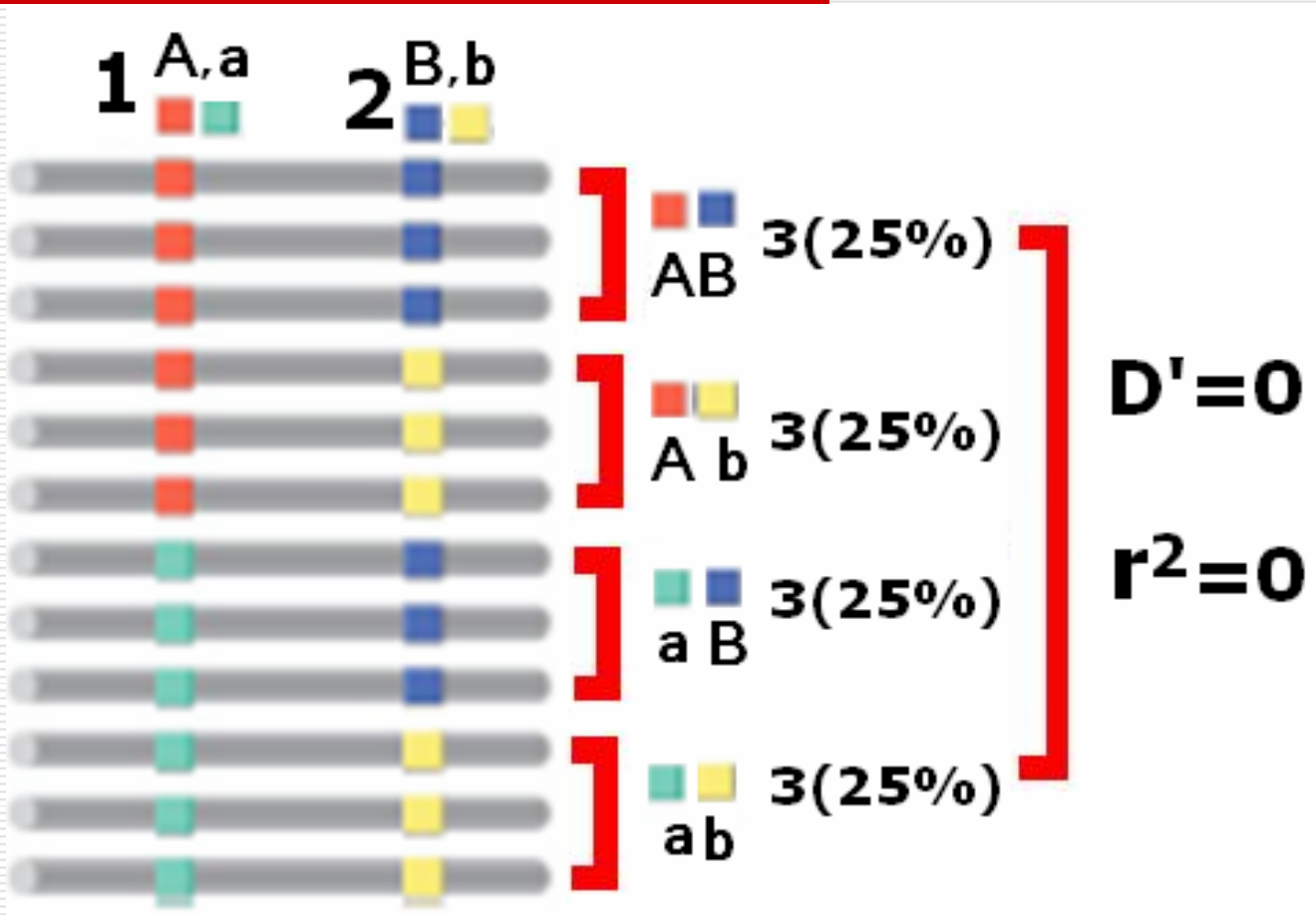
⑥ $r^2=1$ 说明两位无重组； 4种单倍型最多只能出现2种(AB, ab)，且等位基因频率相同。

称为完美LD：观察一个标记即可得到另一标记的全部信息。

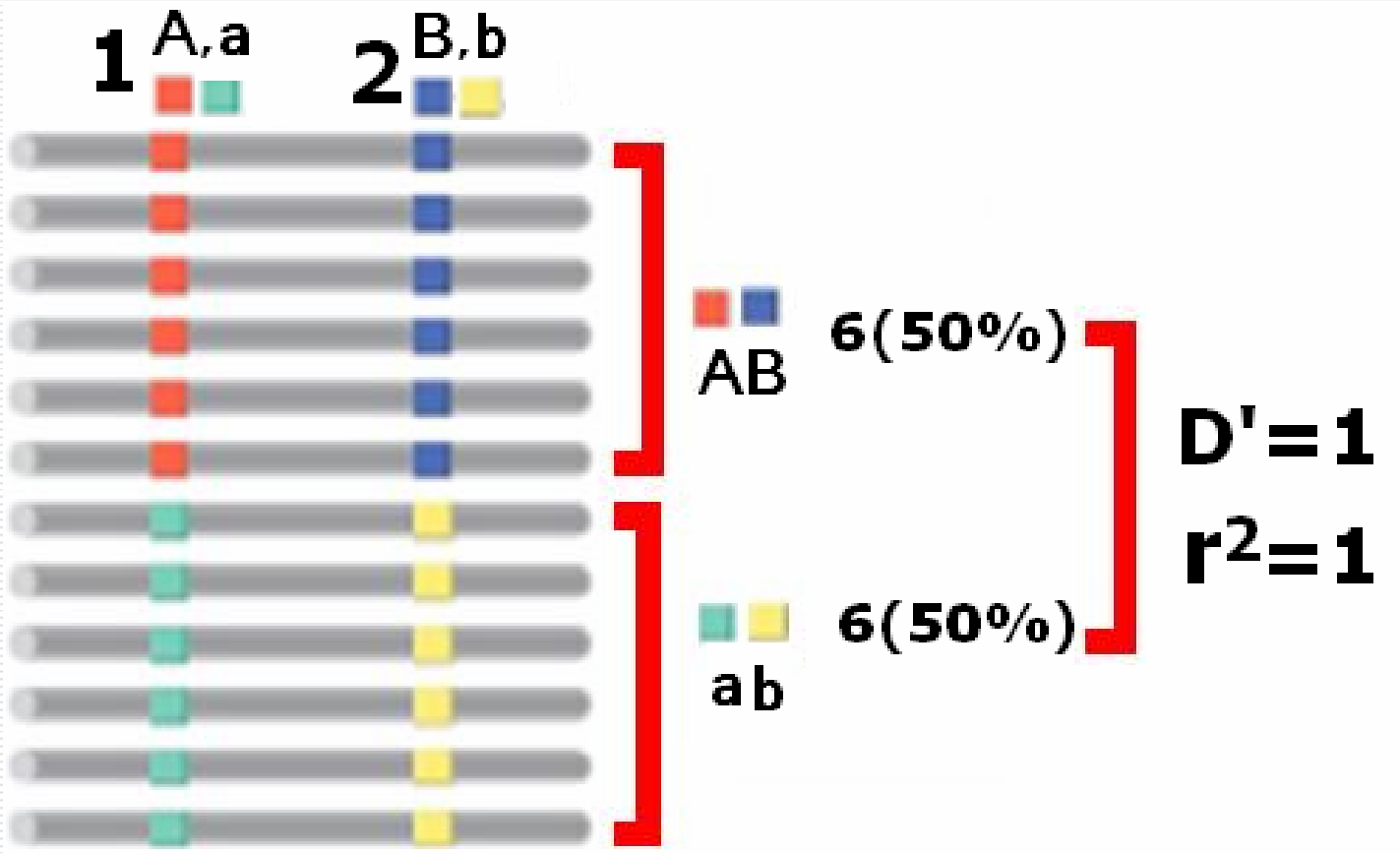
⑥ $r^2=0$ 与D'＝0意义相同.

⑥ $r^2>0.33$：提示“强LD”.

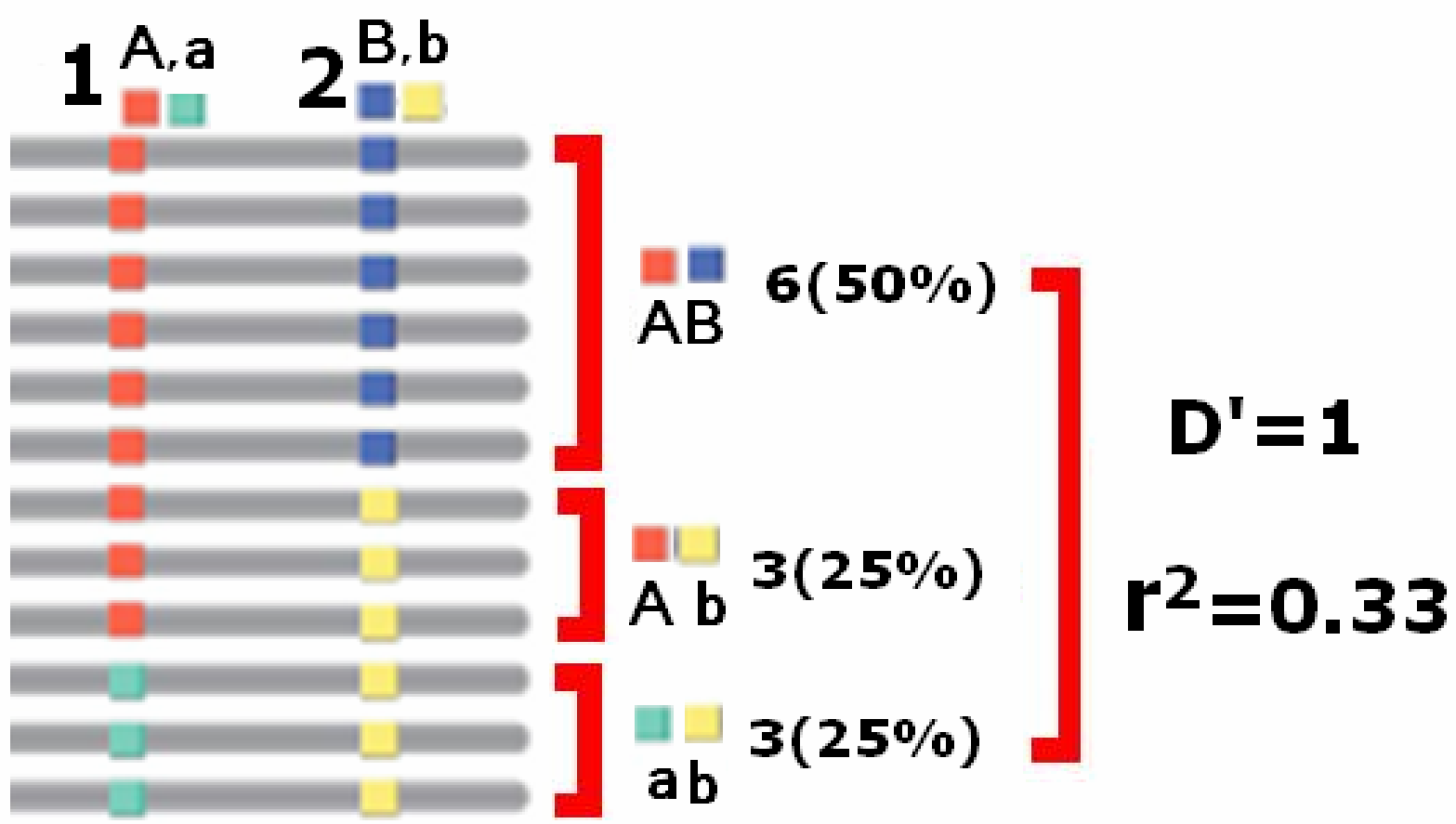
D’=0, r^2=0



D’=1, r^2=1



D’=0, r^2=0.33



（二）影响LD的因素

⑥ **遗传漂变：**群体较小，导致群体中基因频率随机波动的现象称为遗传漂变。

一般认为：群体越小，漂变效应越大→ LD程度↑。

⑥ **“奠基者效应”：**是一种剧烈的漂变；指一个小群体从一个大群体中分离出来，并逐渐发展壮大的现象。

“奠基者效应”→ LD程度↑

⑥ **人口增长：**人口增长会降低遗传漂变，LD强度减弱。

群体的增长→LD程度↓；

群体的再分→LD程度↑（“奠基者效应”）。

⑥ **重组率的变化：** LD程度与重组率呈反比。

重组率↑ → LD↓

重组区域 → LD↓

非重组区 → LD↑

⑥ **突变率的变化：** 与重组类似，突变率↑ → LD↓

突变率高的SNPs间几乎无LD。

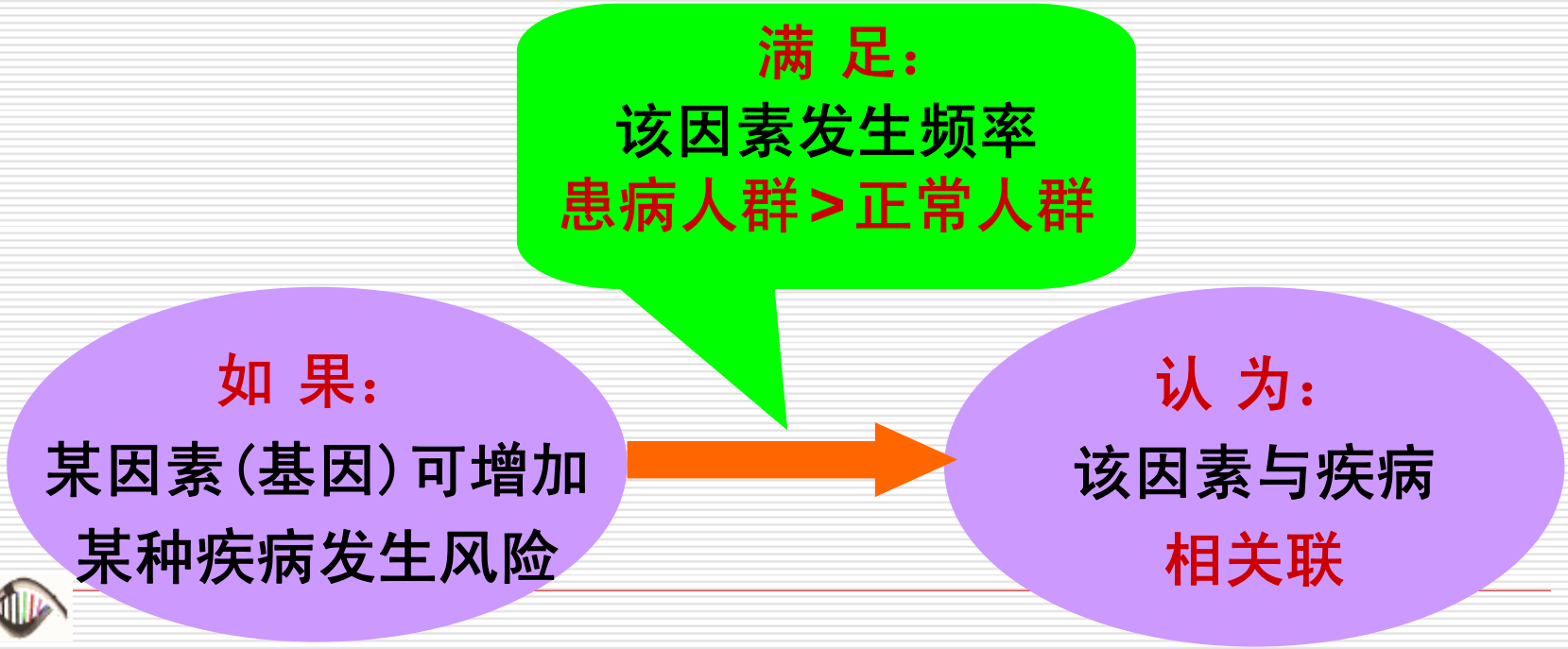
⑥ **基因转换：** 指染色体的部分片段在减数分裂过程中转移到另一片段的过程。基因转换在人类的发生率较高。类似重组或突变，基因转换→ LD↓。。

基因转换对紧密相邻SNPs间的LD影响最大。

（三）基于SNP的LD关联分析

⑥ 在关联分析中，主要采用基于LD 的关联分析。

⑥ 将LD应用于关联研究，可定位复杂的疾病基因。



1. 基于LD 的关联分析原理

比较遗传标记差异：

患者-正常人

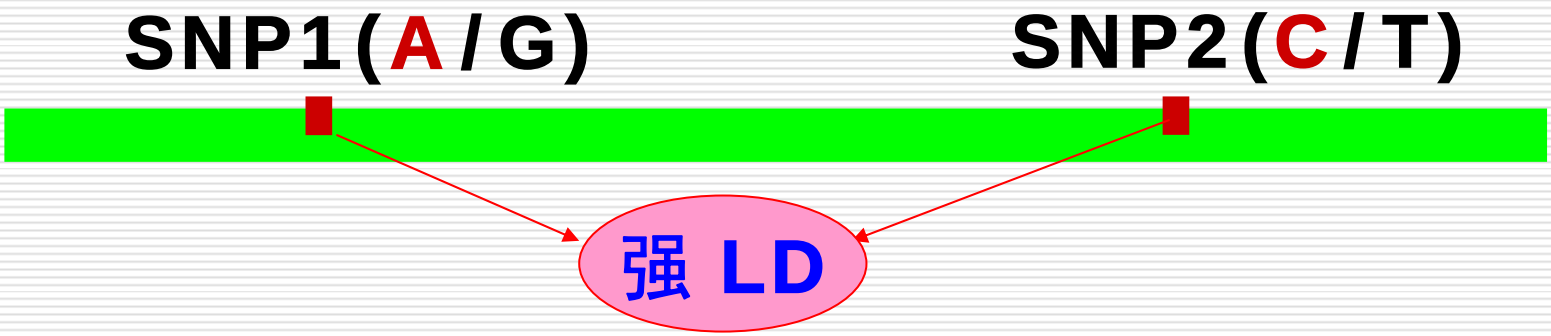
致病基因-遗传标记

强 LD

致病基因在疾病

发生中相对危险度

基于SNP的LD分析原理



当SNP1A与
疾病易患性有关

观察到

SNP2C频率
患病群体高于对照群体

等位基因A：与该疾病相关
单体型AC：确定了与疾病相关的风险因子

2. LD作图

⑥ LD作图是将一段基因的所有SNPs的LD关系标记在基因序列中。用来观察重组热点。

⑥ 作图方法有：

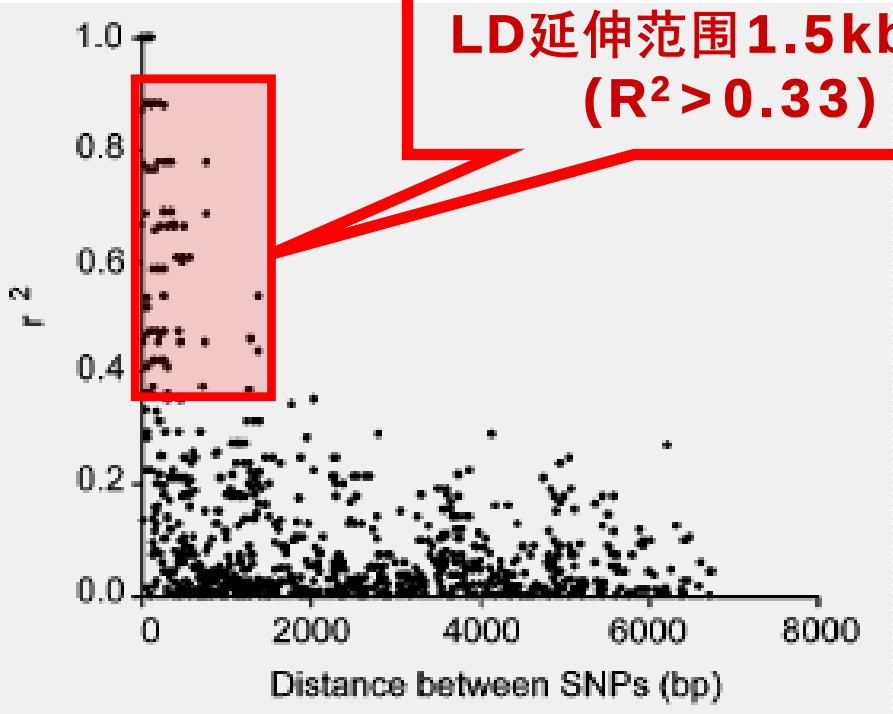
LD散点图（dot plot）

LD矩阵图（ LD matrix ）

邻近LD窗口分析(adjacent LD window analysis)

（1） LD散点图（dot plot）

以两个SNPs间的LD值与其两点间的物理距离（bp）绘图。用于观察LD与物理距离之间的关系，即SNPs间的**LD延伸范围**（extent of LD）

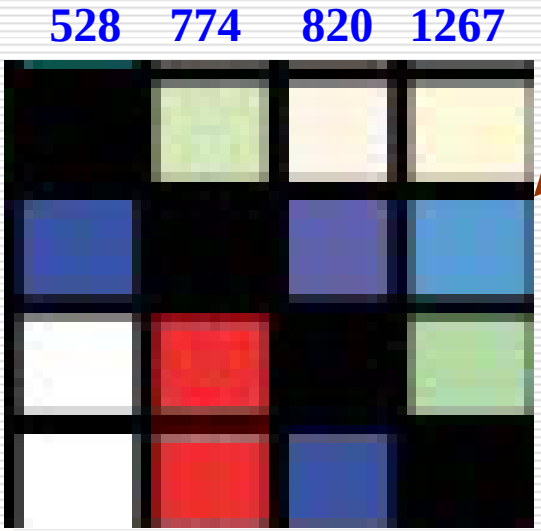


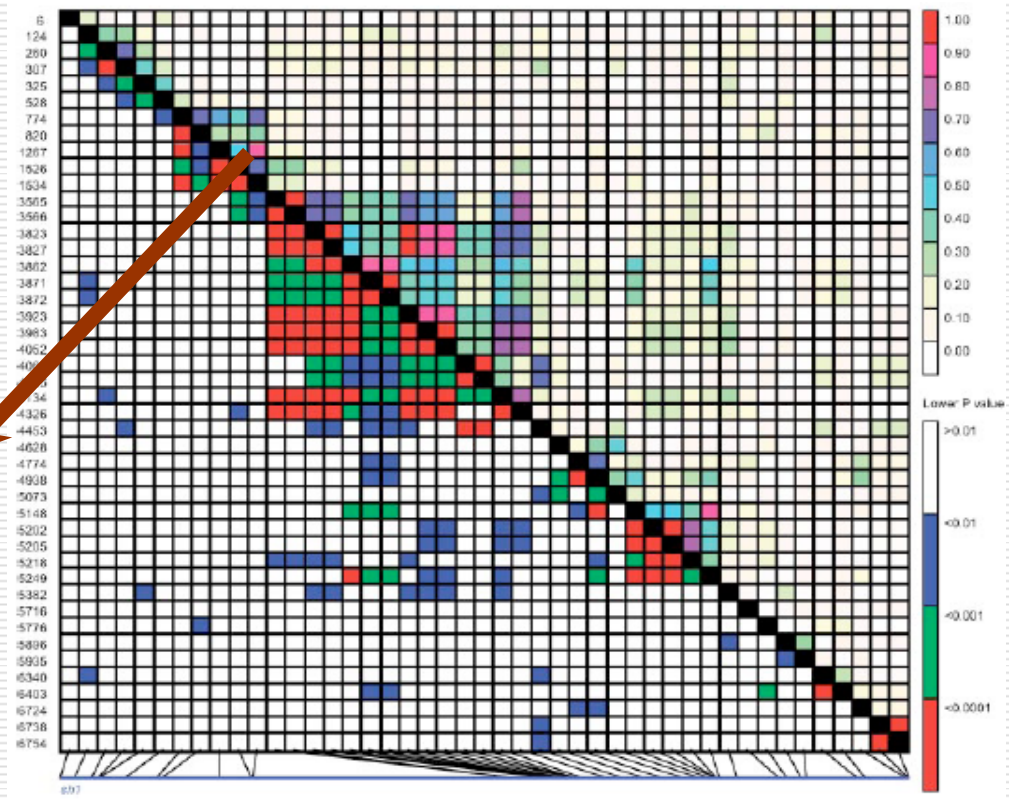
LD decay plot of *shrunken 1* (*sh1*) in maize.

（2） LD矩阵图（ LD matrix ）

以SNPs在基因序列中的位点组成阵列，将SNPs间的LD或P值填到相应的阵列中。

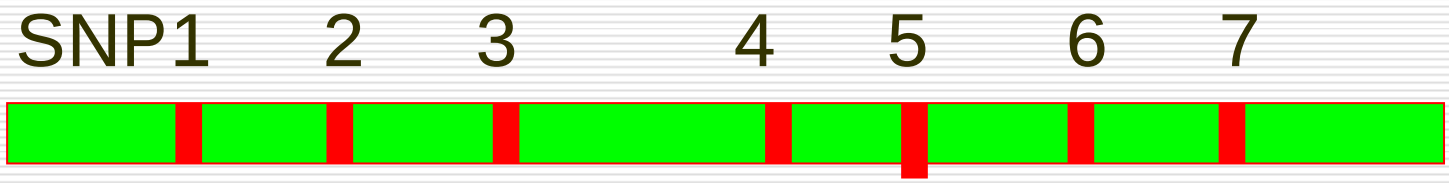
可直接观察LD与物理距离bp之间的关系。





LD matrix for polymorphic sites within *sh1*.

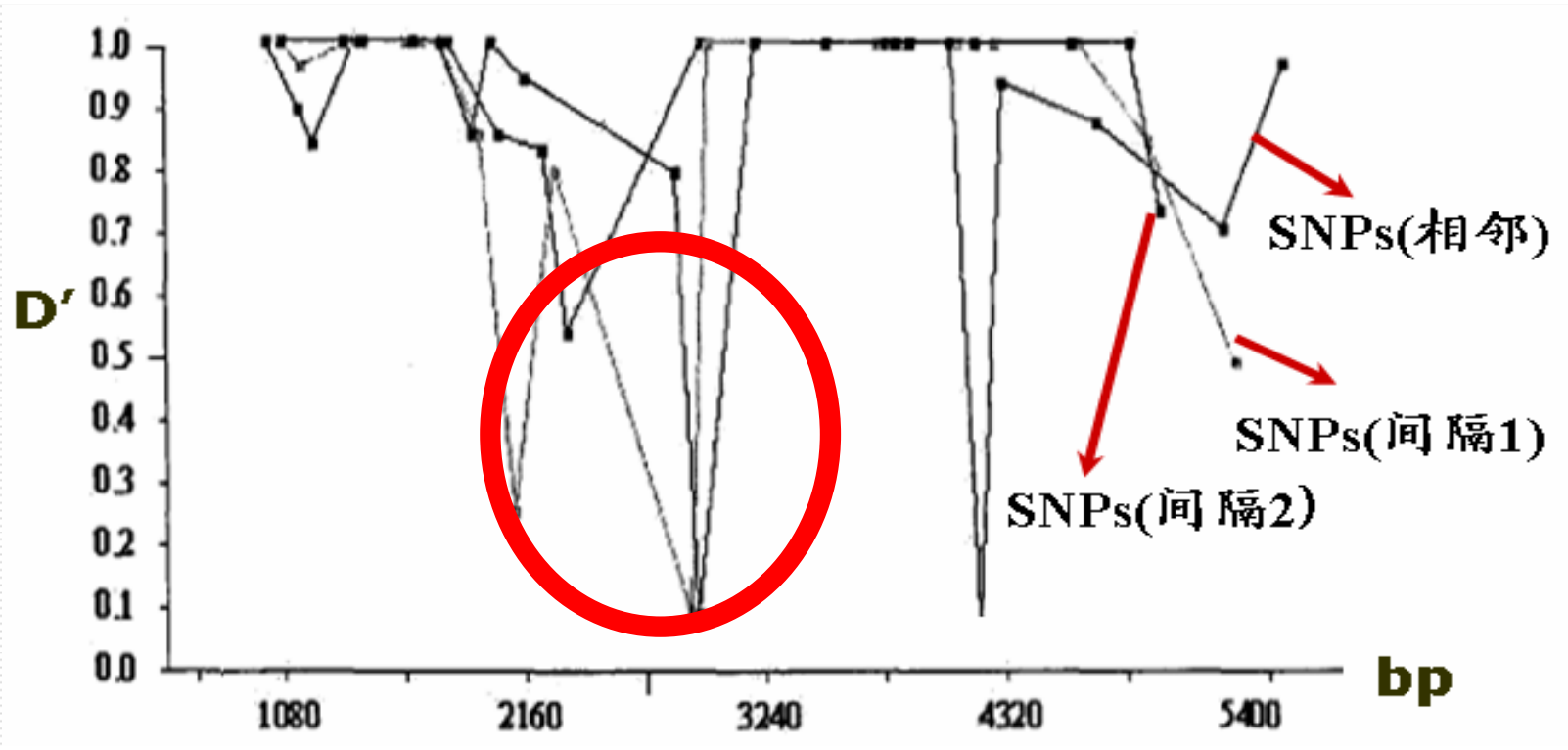
3. 邻近LD窗口分析（adjacent LD window analysis)



⑦ **方法：**是将相邻SNPs（1-2，2-3...）、间隔1个SNPs（1-3，2-4，3-6...）、间隔2个 SNPs（1-3，2-5，3-6...），与其对应的LD值绘制散点图再连线即可。

⑦ **作用：**观察强LD区域，分析推断在扫描的基因组区域潜在的重组热点（波谷或较低的LD区域）。

CDKN1A 基因调控区21个SNPs邻近LD窗口分析



发现：在~2800bp有较低的LD值及波谷；

提示：在该位置可能有较高的重组率。

三、单体型分析

⑥ **单体型：**一条染色体区域中所有SNPs等位基因的集合称为单体型或单倍型（haplotype)。

⑥ **单体型理论数量：**有n个SNP → 2^n个单体型。

如：

SNP1(A,G)　　SNP2(C,T)

AC、AT、GC、GT

LD存在，实际上只存在少数几个常见的单体型：

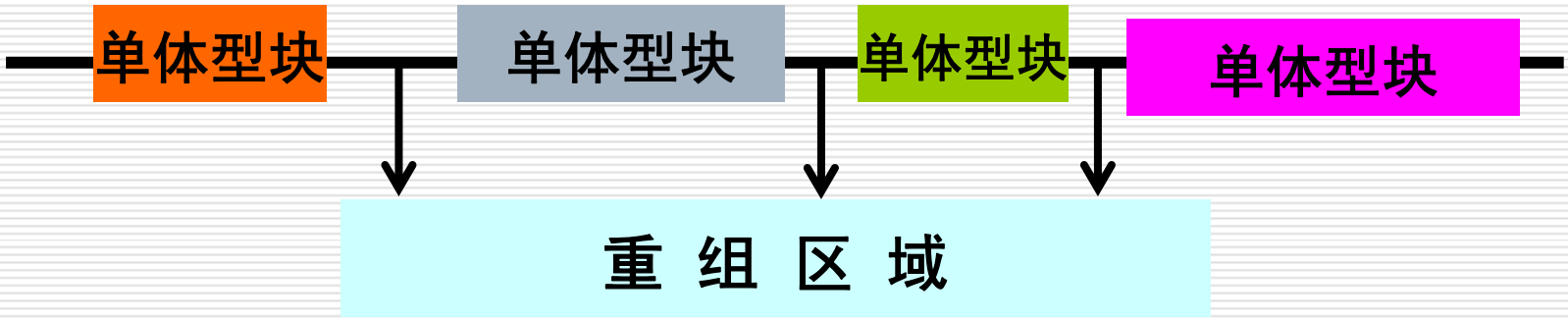
⑥ 例如，在一段含有6个SNPs区域中，理论上应有2^6=64种单体型，实际上只有3种常见的单体型(频率90%)。

⑥ 对1和2：4种单体型中实际只有AC和GT是常见的。

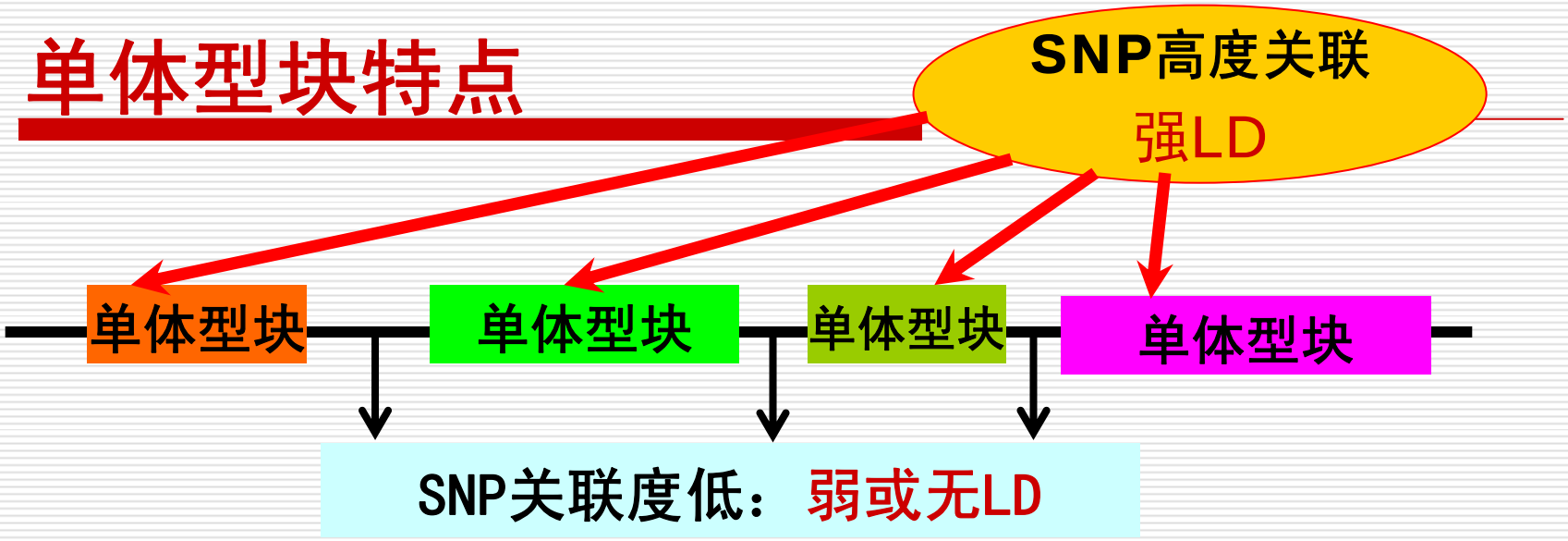
1	2	3	4	5	6	频率
… A …	C …	A …	T …	G …	T	40%
… A …	C …	C …	G …	C …	T	30%
… G …	T …	C …	G …	G …	A	20%
…	其		他		…	10%

单体型块(haplotype block)

⑥ **单体型块概念:** 染色体在传递中同源片段发生重组，多代之后祖先染色体片段的原有排布已被打乱，染色体形成没有发生重组的区域被重组区域相互隔开，这些没有发生重组的区域称为单体型块或单体型区域、单体型域。重组区域称为重组热点。



单体型块特点



⑥ 单体型块的形成： 由重组区域所致。

⑥ 单体型块的大小： 从lkb~数百kb；

⑥ 人体之间单体型块的大小及单体型种类非常相似；

⑥ 一个单体型块一般只有几个常见单体型，用几个SNP位点，就可以确定单体型块的类型。

例如，Daly等用103个常见SNPs (频率>5％)，研究250个欧洲人5号染色体上500 kb范围内的单体型结构。发现：

⑤500 kb区段被分为11个单体型块；

⑤单体型块大小：3kb～92 kb；

⑤每个单体型块中，有2～4个单体型，频率95％；

⑤单体型块：LD较高；

重组区域：LD较低。

Daly MJ, et al. High-resolution haplotype structure in the human genome. Nat Genet, 2001, 29 (2) : 229

标签SNPs
(haplotype tag SNPs, htSNPs)

- ⑥ **htSNPs：**指确定染色体某一区段的单体型结构所必须的、少量的、关键的SNPs。
- ⑥ 用htSNPs可以确定一个单体型或一个基因，从而使基因型的检测工作量大大降低。
- ⑥ 例如，有学者在研究疾病基因单体型时发现：
- ❶ 2-5个htSNPs就可以确定单体型结构；
- ❶ 基因型的检测工作量：从122个SNPs减少到34个。

Johneon GC,et al. Haplotype tagging for the identification of common discase genens. Nat Genet, 2001, 29:233-237

单体型的确定方法：
（实验法、系谱推断法、统计算法）

实验法

③ 单分子稀释法（single-specific dilution）

等位基因特异性PCR （AS-PCR法）

长插入克隆法（long-insert cloning）

双倍型-单体型转化(diploid-to-haploid conversion)

③ 实验法可以得到更多的信息，但由于费用昂贵，耗时长，因此不适合大规模应用。

系谱推断法

③ 系谱推断法是依据家系中相关个体的基因型来确定单体型。

③ 该法可以为紧密连锁的SNPs(强LD)提供真实的信息，但当家系中某些成员的资料无法获得或数据缺失时，会使SNPs间的关系模糊不清，可能导致完全错误的单体型与疾病的相关结论。

③ 该方法仅适用于家系的单体型确定。

统计算法

- 目前最经济、最实用、应用广泛的单体型推断的方法。
- 克拉克算法(Clark's)：是试图使观察样本中单体型数目最小化的一种算法。计算软件是Hapinferx程序。
- 最大似然算法(Expectation-Maximization)：采用EM算 法进行样本单体型频率的最大似然估计。计算软件：Haploview和EH (estimation of haplotype)。

  http://linkage.rockefeller. edu/ software/eh
- 贝叶斯算法(Bayesian):按照在自然人群中的理论值预测单体型的类型。计算软件：Phase

  http：//www. stat. washington. edu/ stephens/

单体型确定的影响因素

③ SNPs密度：如果SNPs密度太低，导致单体型种类及重组事件的检测灵敏度降低。

建议：最合适的密度为2kb选择1-4个SNPs。

通常：只选择常见SNPs位点中的一部分进行检测。

（常见SNPs：等位基因频率>5%或10%）

③ 样本量：样本量少会导致：单体型被漏检；单体型块的大小比真实值高。

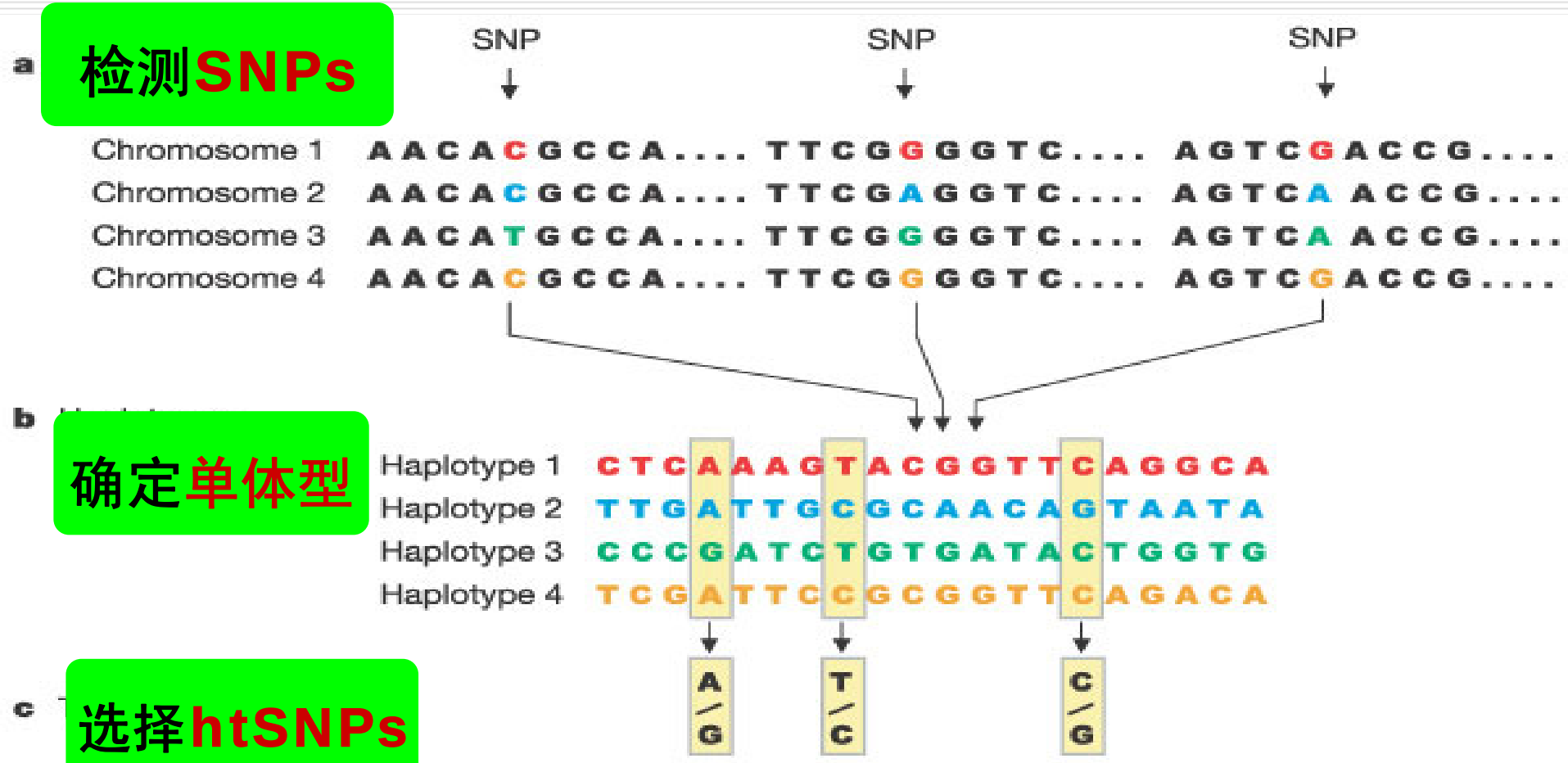
最小样本量：100个样品（染色体），以确保频率>5%单体型的检出概率>95%。

SNPs-单体型-标签SNPs

检测SNPs

确定单体型

选择htSNPs



用htSNPs关联分析

单体型的关联分析

- ⑥ 如果我们在全基因组300-1000万个SNP中，逐一进行基因型与疾病的相关分析，工作量将是非常庞大。
- ⑥ 如果我们以单体型块为单位，只要检测几个标签SNP，就可以识别出相应的单体型结构，进而确定是否与疾病相关。
- ⑥ 以单体型块为基础，进行疾病与基因关联分析的方法，称为单体型方法(haplotype method)，该法是目前最经济、信息量最丰富的LD分析的方法。

四、应用举例

⑥ 例一、全反式视黄醇脱氢酶基因(RDH8)

SNPs连锁不平衡图谱的建立

⑥ **RDH8：** 是视循环代谢中最早发现的一种酶，属于乙醇脱氢/还原酶家族，催化视黄醛到视黄醇的还原反应，该酶代谢可能与屈光不正及眼球生长有关。该基因位于19号染色体，有6个外显子，长度约9000bp，是高度近视的候选基因。

韩伟等．中华眼科杂志，2006,42(7):642-648

方法（基于实验的分析方法 ）

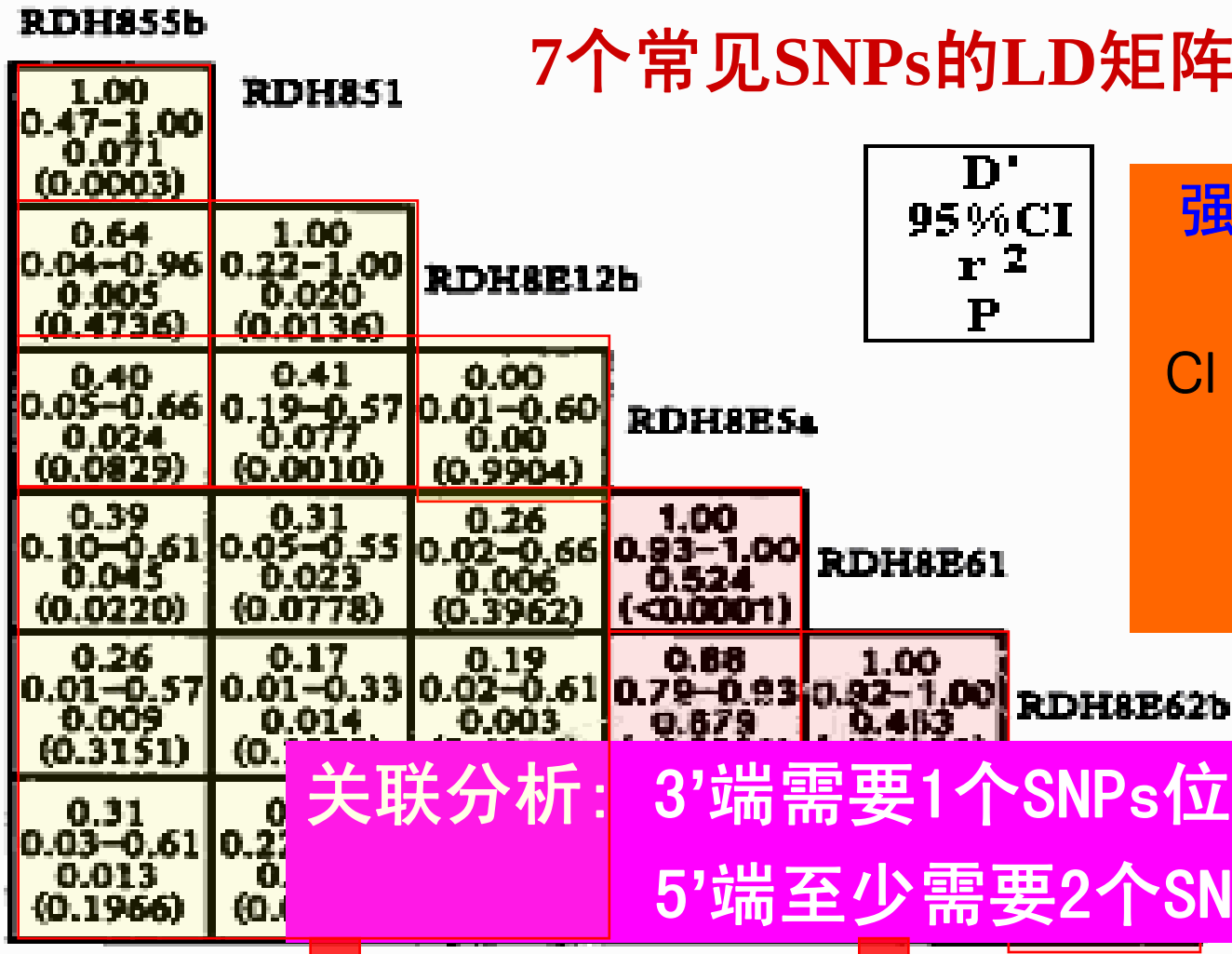
- 用DHPLC和测序技术进行基因型测定；
- 在20个汉族人样本样品池中筛查SNPs；
- 在150个汉族人样本中测定基因频率；
- 用Haploview和EH软件对SNPs进行LD和单体型分析。

结果 筛查到15个SNPs，新发现10个，常见7个

SNP 名称	命名	位置	NCBI 的 SNP 序列号	等位基因型	基因频率
-2881G>C	RDH8S8a	5′ 端上		G/C	0.0033
-2720C>G	RDH8S8b				0.0033
-2710T>C	RDH8S8c				0.0033
-2076G>A	RDH8S6	5′ 端上		G/A	0.0033
-1799A>G	RDH8S5a	5′ 端上游序列	未报道	A/G	0.0067
-1715G>A	RDH8S5b	5′ 端上游序列	未报道	G/A	0.1633
-472C>T	RDH8S1	5′ 端上游序列	rs2233789	C/T	0.2667
-130C>A	RDH8E1a	外显子 1 的 5′UTR	未报道	C/A	0.0033
-126A>G	RDH8E1b	外显子 1 的 5′UTR	未报道	A/G	0.0533
7826T>C	RDH8E5a	外显子 5	rs1644731	C/T	0.4367
7827G>A	RDH8E5b	外显子 5	未报道	G/A	0.0133
8117C>T	RDH8E61	外显子 6	rs747574	C/T	0.4033
8344C>T	RDH8E62a	外显子 6 的 3′UTR	未报道	C/T	0.0067
8566-8567delGA	RDH8E62b	外显子 6 的 3′UTR	rs3217240	-/GA	0.4067
10254G>A	RDH8S36	3′ 端下游序列	rs1644727	G/A	0.4000

7个常见SNPs
（等位基因频率>5%）

7个常见SNPs的LD矩阵图



强LD判断标准：

D’>0.75；

CI：CU＞0.98 且 CL＞0.70；

r^2>0.33；

P<0.031

关联分析：3’端需要1个SNPs位点；5’端至少需要2个SNPs。

5’端3个SNPs间LD弱

3’端4个SNPs间存在强LD，组成一个单体型块

7个常见SNPs组成的单体型分析结果

⑥ 理论2^7=128种单体型；实际有16种单体型。

RDH855b	RDH851	RDH8E12b	RDH8E5a	RDH8E61	RDH8E62b	RDH836	单倍型频率	累计频率
1	1	1	2	2	2	2	0.2065	0.2065
1	1	1	1	1	1	1	0.1569	0.3634
1	2	1	1	1	1	1	0.1322	0.4956
2	1	1	2	2	2	2	0.0995	0.5951
1	1	1	2	1	2	2	0.0923	0.6874
1	2	1	2	2	2	2	0.0664	0.7538
1	2	1	1	1	2	1	0.0308	0.7846
2	1	1	1	1	1	1	0.0273	0.8119
1	1	1	1	1	1	2	0.0243	0.8362
1	2	1	2	1	2	2	0.0240	0.8602
1	1	2	1	1	1	2	0.0169	0.8771
1	1	1	1	1	2	1	0.0164	0.8935
2	1	1	2	1	2	2	0.0162	0.9097
1	1	1	2	1	1	1	0.0125	0.9222
1	1	1	1	1	2	2	0.0109	0.9331
1	1	2	2	2	2	2	0.0102	0.9433
G	C	A	C	C	–	G		
A	T	G	T	T	GA	A		

75%

例二、蛋白激酶Cζ 亚型(PRKCZ)基因 SNPs的LD分析

⑤**PRKCZ基因**：是2型糖尿病易感基因，有18个外显子，位于1号染色体。已知SNP（rs436045）与2型糖尿病相关。

⑤**目的**：寻找该基因在中国北方汉族人群2型糖尿病的相关单体型。

孙红霞等. 中国医学科学院学报，2002，24(5)：474-480

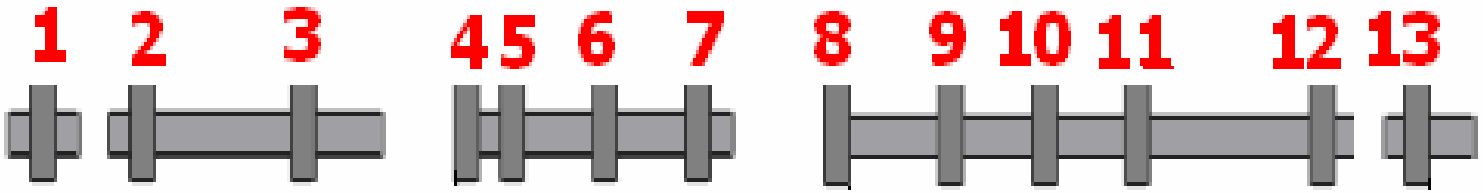
方法（基于公共数据库的方法 ）

- ⑥ 用生物信息学方法，在公共SNP数据库中查找该基因中的SNP位点, 在阳性位点rs436045上下游选择了13个SNP。
- ⑥ 样本：2型糖尿病患者173例，对照组152例
- ⑥ 用单碱基延伸反应法(SBE) 进行基因型分析；
- ⑥ SNP间的LD分析用DnaSP3.5程序；

  单体型分析用phase程序。

结果： 该基因13个SNP中有9个与疾病（P<0.05）

No.	SNP name	Position in the gene	Distance with rs436045	χ^2	
				Statistic	*P*-values
1	rs1878745	exon4	−79138	1.549	0.213
2	rs1467217	intron4	−24924	3.887	0.049
3	rs1401136	intron4	−13943	1.066	0.302
4	rs411021	intron5	−1235	7.998	0.005
5	rs436045	intron5	+1	8.612	0.003
6	rs427811	intron5	+3960	6.812	0.009
7	rs385039	intron6	+5809	8.631	0.003
8	rs809912	intron7	?	8.631	0.003
9	rs262669	intron9	+4235(from rs809912)	1.649	0.199
10	rs262662	intron9	+6779(from rs809912)	4.797	0.029
11	rs381664	intron10	+9950(from rs809912)	7.998	0.005
12	rs262650	intron10	+15488(from rs809912)	5.252	0.022
13	rs262642	intron13	?	3.371	0.066



对照 | 1-4弱LD(重组) | 4-13 强LD：单体型块

病例 | 1-4弱LD(重组) | 4-8单体型块 | 8-13弱LD(重组)

4-13 SNP组成的单体型频率(%)

	Haplotypes	Control	Case	Haplotypes	Control	Case
1	CGTAGCTTGC	63.5	64.7	CGTAGTCTAC	0	1.0
2	TAGGATCCAC	22.5	13.1	TAGGATTCGT	0	1.0
3	CGTAG[illegible]	7.8	6.2	TAGAGTCCAC	0	0.3
4	[illegible]GATCCAT	4.5	2.9	CAGGATCCAT	0	0.3
5	CGTAGCCTGC	0.8	1.3	CGGAGTTTGC	0	0.3
6	CGGAGTCTGC	0.4	0.7	TAGGATTCGC	0	0.3
7	CAGAGTCCAT	0.4	0	CGTAGTTCGC	0	0.3
8	CGTAGCCTAC	0	2.9	TAGGATTTGT	0	0.3
9	CGTAGCTTGT	0	2.3	TAGGACCCAT	0	0.3
10	CGTAGTTTGC	0	2.3			

对照:95%

病例:84%

重组?

5个SNP4-8组成的单体型（%）

	Haplotype	case	control
1	CGTAG	81.0	72.1
2	TAGGA	17.3	27.0
3	CGGAG	1.0	0.4
4	CAGAG	0	0
5	TAGAT	0.3	0
6	CAGGA	0.3	0

OR=1.652, P<0.007

提示：

单体型1,2可能

与疾病发生相关

小　结

基于SNP的LD关联研究策略

⑤ 选取10-20样本，对某一区域内所有SNPs进行基因型检测，或用生物信息学方法在数据库中查找SNPs；

⑤ 用LD确定单体型块结构，用统计算法确定单体型，选择或确定标签SNPs；

⑤ 对所有样本中的标签SNPs进行基因型鉴定；

⑤ 用X^2检验进行病例-对照关联分析。

谢谢！

等位基因频率计算

⑥ 等位基因频率=（2倍纯合子+杂合子）/2倍例数

$$A=2AA+Aa/2n$$

$$a=2aa+Aa/2n \quad 或 \quad a=1-A$$

⑥ 如一SNP（C/T）的等位基因频率为：

N	Proportions of genotypes(%)			Allelic frequencies(%)	
	CC	CT	TT	C	T
137	29(21. 17)	72(52. 55)	36(26. 28)	47. 45	52. 55

$$C=2CC+CT/2n=47.45\%$$

$$T=2TT+CT/2n=52.55\%$$

⑥ 连锁（linkage）：一条染色体上的两个或多个标记（如多态性）离的越近，它们在DNA修复或复制过程中分开的概率就低。因此，它们的关联度越高，共同遗传的概论就越大。

⑥ 连锁图（linkage map）:关于一条染色体上遗传位点的相对位置的图谱，根据共同遗传的位点数据绘制。距离用厘摩（cM）。1cM＝100万bp

影响关联分析效力的因素

⑦ 所研究疾病位点的危险度;

⑥ 疾病位点等位基因的频率;

⑥ 标记位点等位基因的频率;

⑥ 两者之间的LD强度 ;

⑥ 群体数据（群体遗传平衡：基因型的频率不变；基因的频率不变）等。

例二、炎症性肠病IL-1β基因多态性及LD研究

⑥**IL-1:** 是一种由单核巨噬细胞产生的炎症反应细胞因子,包括α和β单体,位于第2号染色体长臂，其活性主要由 IL-1β表达。 IL-1β的基因长达 7 kb,含有7个外显子和 6个内含子。 在启动子区域存在 C-31T和T-511C位点基因多态性。

⑥**目的:** 探讨 IL-1β基因多态性与炎症性肠病(BD)的发病关系。

⑥**方法:** 采用 PCR-RFLP法,对BD组(N=48)和正常对照组(N=137) IL-1β启动子区–31和-511位点进行基因分型,计算等位基因频率分布,并进行连锁不平衡分析及其单体型与疾病的关联分析。

薛惠平等. 上海交通大学学报(医学版), 2006, 26(8): 912-915

结果1. 两组基因型频率和等位基因频率无差异(P>0.05)

-31 C-T

Group	N	Proportions of genotypes(%)			Allelic frequencies(%)	
		CC	CT	TT	C	T
Normal	137	31(22.63)	78(56.93)	28(20.44)	51.09	48.91
BD	48	16(33.33)	19(39.58)	13(27.08)	53.13	46.88

-511 T-C

Group	N	Proportions of genotypes(%)			Allelic frequencies(%)	
		CC	CT	TT	C	T
Normal	137	29(21.17)	72(52.55)	36(26.28)	47.45	52.55
BD	48	14(29.17)	21(43.75)	13(27.08)	48.95	51.04

*BD-Normal:基因型频率无差异(P>0.05)

BD-Normal:等位基因频率无差异(P>0.05)

结果2. -31 C/T与-511 T/C间存在强LD:

(D'=0.915, r^2=0.735)

结果3. -31C/-511T单体型发生BD的风险增加

Item	BD group (frequency)	Normal group (frequency)	OR	95%CI
C-C	3. 03 (0. 032)	16. 43 (0. 060)	0. 510	0. 146 - 1. 778
C-T	47. 97 (0. 500)	123. 57 (0. 451)	1. 216	0. 763 - 1. 937
T-C	43. 97 (0. 458)	127. 57 (0. 466)	0. 970	0. 609 - 1. 547