Analysis and Application of SNP and Haplotype in the Human Genome

LI Jing[1], PAN Yu-Chun[2], LI Yi-Xue[3,①], SHI Tie-Liu[3,4,①]

(1. *College of Life Science & Biotechnology, Shanghai Jiao Tong University, Shanghai* 201101, *China*;
2. *College of Agriculture & Biology, Shanghai Jiao Tong University, Shanghai* 201101, *China*;
3. *Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai* 200031, *China*;
4. *Bioinformatics Center, Shanghai University, Shanghai* 200436, *China*)

Abstract: Single nucleotide polymorphism (SNP) is the most common type of genetic variant in human genome. Haplotype, defined as a specific set of alleles observed on a single chromosome, or a part of a chromosome, has been an integral part of human genetics for decades. The goal of the international HapMap project is to determine the common patterns of DNA sequence variation and find the Tag SNPs representing all SNPs in the human genome. Some studies demonstrated that the analyses of haplotype defined by the grouping and interaction of several variants rather than any individual SNP correlated with complex phenotypes. Here, we describe the definitions of SNPs, genotype, haplotype and some information of the HapMap project. In this review, we summarize the current three haplotype-inference methods, including Clark 's method, EM algorithm and Byes approach, and the different defining methods for haplotype block, as well as the methods for choosing tag SNPs and association studies of complex diseases using haplotype. The major public SNP databases and applications of SNPs and haplotype in common complex diseases and drug response are also introduced in the paper.

Key words: SNPs; haplotype; haplotype block; tag SNPs; complex disease

人类基因组单核苷酸多态性和单体型的分析及应用

李 婧[1], 潘玉春[2], 李亦学[3,①], 石铁流[3,4,①]

(1. 上海交通大学生命科学技术学院, 上海 200030;
2. 上海交通大学生物与农业学院, 上海 201101;
3. 中国科学院上海生命科学研究院生物信息中心, 上海 200031;
4. 上海大学生物信息学中心, 上海 200436)

摘 要: 单核苷酸多态性是人类基因组中最丰富的遗传变异。单体型是指位于一条染色体上或某一区域的一组相关联的 SNP 等位位点, 单体型已经成为近年来人类遗传研究的组成部分。人类基因组单体型图(HapMap)计

收稿日期: 2004 - 10 - 13; 修回日期: 2004 - 12 - 06

基金项目: 国家高技术研究发展计划项目(国家"863 计划")(编号: 2002AA231051, 2001AA231011, 2001AA231111)[Supported by Chinese National Programs for High Technology Research and Development (No. 2002AA231051, 2001AA231011, 2001AA231111)]

作者简介: 李婧(1978 -), 女, 在读博士, 研究方向: 生物信息学

① 通讯作者。石铁流: 副研究员, 博士生导师, 研究方向: 生物信息学。E-mail: tlshi@sibs.ac.cn;
李亦学: 教授, 博士生导师, 研究方向: 生物信息学。E-mail: yxli@sibs.ac.cn

划的目标就是构建人类 DNA 序列中多态位点的常见模式,找出代表整个人类基因图谱之中的 SNP 集合的标签 SNP。在复杂性疾病研究中,由多个变异位点组合构成的单体型分析优于单个 SNP 的分析。文章论述了 SNPs、基因型、表现型的定义与 HapMap 计划的一些情况,综述了单体型的 3 种推断算法和单体域的不同定义与构建方法,同时介绍了标签 SNP 的选择及单体型与复杂疾病关联分析的方法,可利用公共 SNP 数据库的情况以及 SNPs 与单体型在复杂疾病与药物反应方面的应用。

关键词:SNPs;单体型;单体域;标签 SNPs;复杂性疾病
中图分类号:R394 文献标识码:A 文章编号:0379-4172(2005)08-0879-11

近年来,由 SNP 研究委员会(The SNP Consortium,TSC)与美国国立人类基因组研究院发起的对单核苷酸多态性(Single nucleotial polymorphisms ,SNPs)的大规模研究表明,单核苷酸变异能够为研究多基因复杂疾病及个体间患病风险与药物反应的不同提供新方法。人类所有群体中存在大约 1 500 万个 SNP 位点(稀有 SNP 位点的频率至少为 1%),平均约每 300 ~600 bp 存在一个碱基突变[1]。这些 SNPs 位于基因的编码区或非编码区,可能是人类基因组中疾病易感基因的遗传标记,甚至是直接影响癌症、心脏病、糖尿病与其他常见病的易感性基因位点。随着人类基因组计划的完成,如何利用人类基因组 SNP 多态信息探究遗传性状,特别是复杂疾病与药物反应的遗传机制已经成为当前的研究热点。

1 SNP 与 HapMap

尽管任意两个不相关的人的 DNA 序列有 99.9%是一致的,正是剩下的 0.1%差异造成了人们罹患疾病的不同风险和对药物的不同反应[2]。发现这些与常见疾病相关的 DNA 多态位点,是揭示人类疾病复杂致病原因的最重要途径之一。在基因组中,最普遍的 DNA 变异就是单个碱基的差异,可以分为转换与颠换、单碱基的插入与缺失等不同类型。对于其中发生率大于 1%的变异被称作单核苷酸多态性(SNPs)。例如:某些人染色体上某个位置的碱基是 A,而另一些人染色体的相同位置上的碱基则是 G。同一位置上的每个碱基类型叫做一个等位位点。除性染色体外,每个人体内的染色体都有两份。一个人所拥有的一对等位位点的类型被称作基因型(genotype)。对上述 SNP 位点而言,一个人的基因型有 3 种可能性,分别是 *AA*、*AG* 或 *GG*。基因型这一名称既可以指个体的某个 SNP 的等位位点,也可以指基因组中很多 SNPs 的等位位点[3]。检定个体的基因型,被称作基因分型(genotyping)。目前许多高通量 SNP 检测方法应用到 SNP 的基因分型中,大大加快了人类 SNPs 数据的增长。由不同基因型与环境共同作用所产生的生物体(人类)可观测的物理或生理性状称为表现型(phenotype)。寻找基因型与表现型(疾病或性状)的关系一直就是遗传学的基础目标。

人类基因组计划使 SNPs 信息已经成为人类基因组最丰富的遗传变异[4]。人类基因组中,相邻近的 SNPs 等位位点倾向于以一个整体遗传给后代。位于一条染色体上或某一区域的一组相关联的 SNP 等位位点被称作单体型(haplotype)。如果一个单体型有 N 个变异位点,理论上就可能有 2^n 种可能的单体型。实际上,大多数染色体区域只有少数几个常见的单体型(每个常见单体型具有至少 5%的频率),它们代表了一个群体中人与人之间的大部分多态性。一个染色体区域可以有很多 SNP 位点,但是只用少数几个标签 SNPs,就能够提供该区域内大多数的遗传多态模式,这样将大大减少用于基因型与疾病关联分析中的 SNPs(图 1)。

人类基因组单体型图(Haplotype map,HapMap)计划是人类基因组计划的自然延伸(http://www.hapmap.org/abouthapmap.html.zh)。HapMap 计划将由日本、英国、加拿大、中国、尼日利亚和美国的科学家们合作完成。项目正式开始于 2002 年 10 月的 HapMap 计划第一次会议,预计进行 3 年。中国将在 HapMap 项目中做出 10%的贡献,HapMap“中国卷”的具体内容是构建人 3 号、21 号染色体和 8 号染色体短臂的 HapMap 以及提供一半的亚洲样品。HapMap 的目标是构建人类 DNA 序列中多态位点的常见模式,运用单体型分型的方法找出约 50 万个标签 SNP(tag SNPs)来代表整个人类基因图谱之中的 SNP 集合,这些标签 SNP 与表现型间的关联也更容易显现出来。HapMap 将成为研究人员确定对人类健康和疾病以及

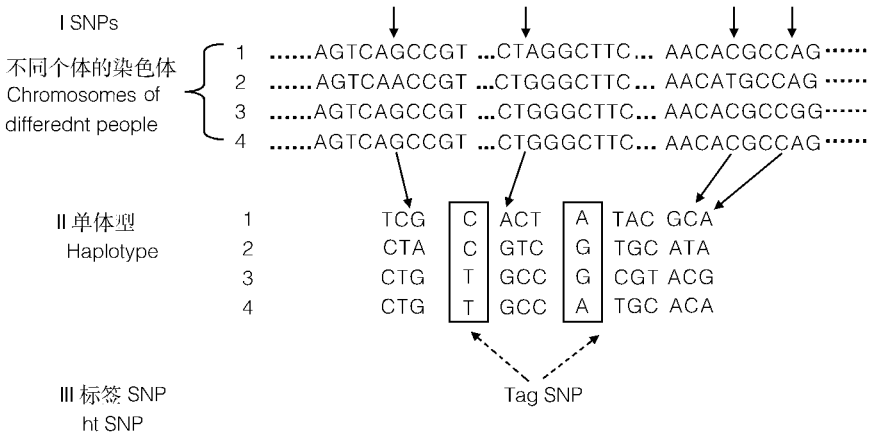

图 1　SNPs、单体型与标签 SNPs

Fig.1　SNPs, haplotypes and tag SNPs

对药物和环境的反应有影响的相关基因的关键信息。

单体型图将描述人类常见的遗传多态模式,它包括染色体上具有成组紧密关联 SNPs 的区域(haplotype block),这些区域中的单体型,以及这些单体型的标签 SNPs[2]。同时,单体型图还将标示出那些 SNP 位点关联不紧密的区域。研究者一般通过比较患者和非患者来发现影响某种疾病,例如糖尿病的基因。在两组单体型频率不同的染色体区域,就有可能包含疾病相关基因。

为了构建单体型图,要对样本至少 100 万个 SNPs 进行全基因组规模的基因分型检测。截至到 2004 年 7 月 20 日,此项目第九次数据公布共发布 639 110 个 SNP 位点的基因分型、基因型频率和实验数据(其中共有 57 519 900 个基因分型数据),现在这项工作仍在继续进行。

2　单体型的推断方法

如前所述,SNPs 位点并不是独立遗传的,而是在染色体上倾向于以一个整体遗传给后代。成组遗传的 SNPs 位点在一代又一代的遗传中绝少发生重组。于是,这样的一组 SNPs 位点类型也就是单体型。由于单体型包含着多个 SNP 的遗传信息,许多研究表明,在与复杂性状的相关分析中,采用单体型比单个 SNP 具有更好地统计分析效果[5~8]。单体型的推断方法主要有 3 类:实验法、系谱推断和统计算法。

单体型推断的实验方法目前有单分子稀释法(single-specific dilution)、AP-PCR(allele-specific PCR)、长插入克隆法(long-insert cloning)与双倍型-单体型转化(diploid-to-haploid conversion)等[9~12]。尽管有报道表明这些实验方法可以得到比某些统计方法更多的信息[13,14],但由于这类方法费用昂贵,耗时长等特点,因此不适合大规模应用[15]。

系谱推断是通过家系中相关个体的基因型,即追溯染色体片段的传递来推断单体型状态。尽管家系内推断单体型并不比在连锁作图中构建单体型的众多方法更简单[15,16],但这样的推断可以为紧密连锁的 SNPs 提供真实的连锁相信息[14]。然而,当某些家系成员资料的无法获得、数据缺失以及某些基因型方式无法提供信息都可能使 SNPs 间的关系状态模糊不清。这种不确定性可能会导致完全假的单体型与疾病的相关[17]。此外,这个方法在个体无相关的群体或很小的家系中将失去作用。

广泛应用于大规模人类基因组单体型推断是统计算法,目前主要存在 3 类演算方法。

2.1　Clark 算法

Clark 第一个提出了在无相关个体间利用基因分型数据推断单体型的算法[18]。假设有二倍体生物的一组序列,并具有多个突变位点,该算法首先找出样本中所有纯合子与仅有单突变位点的杂合子,将这些个体的单体型作为已识别(已分型)的单体型。然后确定每一个已识别的单体型是否为那些尚未确定单体型并有变异位点的序列的等位基因,如果是,就将这种 SNP 的组合确定为新单体型,(例

如,我们观察到一段序列 *ATGGTAC*,在所有位点上均为纯合子,就把它作为一个确定的真实的单体型。如果又检测有两个变异位点的序列 $AT^{G}_{C}G^{C}_{T}AC$,这样就有 4 种可能的单体型。我们假设该基因型包含已确定的单体型,那么可以推断出另一个新的单体型就是 ATCGCAC,然后以新确定单体型为基础继续对其他未分型个体的单体型进行分型,这样的推断链一直到所有的单体型确定或没有新的单体型被发现。这一过程也使得被推断个体的单体型依赖于样本中个体的检测顺序。图 2 就是运用 Clark 算法进行推断的示意图。假设 H_1H_1 对单体型 A_1 而言为纯合子,这里的 H_1 代表任意长的序列。杂合子 H_1H_2 可能有多个变异位点,但如果所有可能的单体型中包括 H_1 与 H_2 的组合,那么我们就得到了 H_1 与 H_2 两个单体型。然后以新的单体型 H_2 来继续推断其他单体型,这个过程直至所有单体型被发现或剩余个体的单体型状态无法由已知的单体型来推断(如基因型 H_8H_9)。

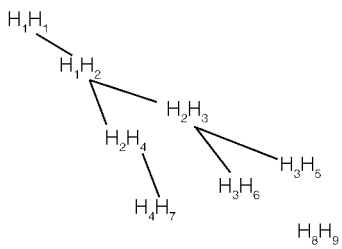

图 2 Clark 算法单体型推断示意图

Fig. 2 Diagram showing the cascade of inferences in the haplotype-inferring algorithm

当遇到多种可能的单体型推断链时,按照简约规则,从中选择产生最少孤儿(如果一个基因型 H_iH_j 的两个单体型 H_i 与 H_j 均没有出现在已识别的单体型中,那么这个单体型将无法被确定,则被称为“孤儿”,如 H_8 与 H_9)的结果。值得注意的是,Clark 的算法没有给出哈代-温伯格平衡状态的前提假设,这与下面所介绍的其他方法是不同的。

Clark 方法存在的问题是,当样本含量较小时,可能会出现没有纯合子或单个突变的杂合子,从而导致单体型的推断无法开始。结合 AS-PCR,Clark 等检测了来自 3 个群体的 72 个个体的脂蛋白脂肪酶基因(LPL)9.7 kb 区域 DNA 序列的变异,通过 Clark 算法确定了 88 个单体型[19]。

2.2 最大似然算法

如上所述,在运用 Clark 算法时,如果样本中没有纯合子或单突变的杂合子就无法开始单体型的推断。此外,当推断过程中出现多种可能的单体型时,依据已有的单体型来确定其中一种为新单体型的结果将受到抽样中个体排序的影响。为了克服上述问题,Excoffier 等提出了最大似然算法[20]。以群体处于哈代-温伯格平衡状态为假设前提,该方法采用 EM 算法进行样本单体型频率的最大似然估计。该方法首先建立关于各单体型频率的似然函数。然后对单体型的频率或其对数求偏导,得出其最大似然估计。但当单体型很多时,运算将十分繁琐,若采用 EM 算法进行迭代求解大大提高运算效率。EM 算法是一种获得参数最大似然估计的迭代方法,它首先假设待估参数(单体型频率)的一组初始值,将初始值看作真实频率,从而求出基因型(两个单体型的特定组合)频率(E 步);然后将此期望值代入似然函数,求出新的一组单体型频率的估计值(M 步)。如此迭代下去,直至两次迭代所得到的参数估计值的差异小于某一个给定的常数,迭代停止(也称为迭代收敛),此时得到的单体型频率的估计值就是它的最大似然估计值。

最大似然算法与 Clark 算法罗列所有可能的单体型不同,它可以直接求得单体型频率。此外,EM 算法对样本中个体检测顺序并不敏感。

最大似然法中初始值的选择很重要,当频率分布存在局部最大值,迭代可能产生错误的最大似然估计。好的初始值可以大大减小风险。保险的做法是广泛选择一组不同的初始值,这样可以增加获得最大似然估计的可能性。2004 年提出的随机 EM 算法(stochastic-EM algorithm)可以有效地解决不收敛和局部收敛的问题[21]。

2.3 贝叶斯算法

尽管最大似然法比较简单,但除了不收敛或收敛至局部最大值的问题以外,另一个问题是随着公共数据库中 SNP 位点的增多,需要考虑的单体型将呈指数增加,这会使 EM 算法的统计效力降低[22,23]。为了克服 EM 算法的上述缺点,贝叶斯理论被 Stephens 等[23]引入单体型的推断。他们采用蒙特卡罗-马尔卡夫链(Markov Chain Monte Carlo,MCMC)方法进行推断,他们的算法也被称为

SSD(Stephens-Smith-Donnelly)算法。根据单体型频率先验分布的不同,SSD 算法又包括了两个算式,第一个是伪 Gibbs 抽样法(PGS 算法),采用的是 Dirichlet 先验分布;另一个算法结合群体遗传的溯祖理论,采用了近似溯祖的先验分布。比较而言,采用近似溯祖的先验分布优于 Dirichlet 先验分布,采用近似溯祖的先验分布的 SSD 算法被应用到程序 PHASE v1.0 中。SSD 算法与 EM 和 Clark 算法的模拟比较表明:SSD 算法的错误率比以前的推断方法减少近 50%,较 EM 算法有两大优点:处理数据规模很大;可以给出单体型构建的不可靠性的估计。

随后,Niu 等[24]与 Lin 等[25]分别采用 Dirichlet 先验分布提出另两种贝叶斯算法。Niu 提出的 PL 算法(partition-ligation algorithm)引入了分割连接(partition ligation,PL)与预先退火(prior annealing)两个新的计算技术,节约了运算时间。PL 算法与 PGS 算法、EM 算法和 Clark 算法在模拟数据与实际数据分析的比较显示,无论样本是偏离了 HW 状态,出现缺失数据,还是出现了重组热点,PL 算法都可获得稳健估计。

Lin 等[25]对 SSD 算法提出修正,在原来的伪 Gibbs 抽样法基础上考虑了缺失数据问题,同时考虑了所有可能的单体型,取消了对不确定性的估计。Stephens 和 Donnelly[26]又提出了新算法(在程序包 PHASE v2.0 中),仍采用近似溯祖先验分布,吸收了 PL 算法中新技术的思路,提高了运算效率与规模。新算法考虑了重组和连锁不平衡随距离的变化,而且可以从群体基因型数据中估计重组率,确定重组热点。Lin[27]针对核心家系中个体数据缺失的情况,在 2002 年的算法基础上重新构建了算法。新算法选择了无穷等位基因模型(infinite-alleles model),添加了对高度连锁不平衡区域间状态的分析,其模拟结果表明,面对缺失数据,该方法无论是对单体型的推断还是对缺失数据的等位基因状态的估计都具有很高的准确度。

3 单体域的定义与构建

2001 年,Daly 等的研究表明,人类染色体 5q31 上 500 kb 的片段上,其单体型结构可以被分为一系列分离的单体域,其大小为 3 ~92 kb。每个单体域有 2 ~4 个常见的单体型(这些单体型包含了所有染色体 90%以上的 SNP 信息),并且域的内部几乎不发生重组[28]。

几乎与此同时,Jeffreys 等的单精分型(single-sperm typing)实验数据表明,Ⅱ型主要组织相容性复合体(MHC)基因序列的大部分重组都限定在狭窄的重组热点处,这就暗示了一个有趣的假设:基因组可以被重组热点分割为一些高度连锁不平衡的区域[29]。

如图 3 所示,染色体上存在着连续、稳定的几乎没有被重组所打断的单体型范围,称之为单体域(Haplotype Block),单体域很可能是遗传的最小单位,在极端情况下,它可以是一个单独的 SNP,或者是一整条染色体。事实上,往往很少的一部分单体型就可以包括大部分的基因状态,在分型时可省节大量工作量,而对延续范围更广的单体域则可以节省更多工作量。

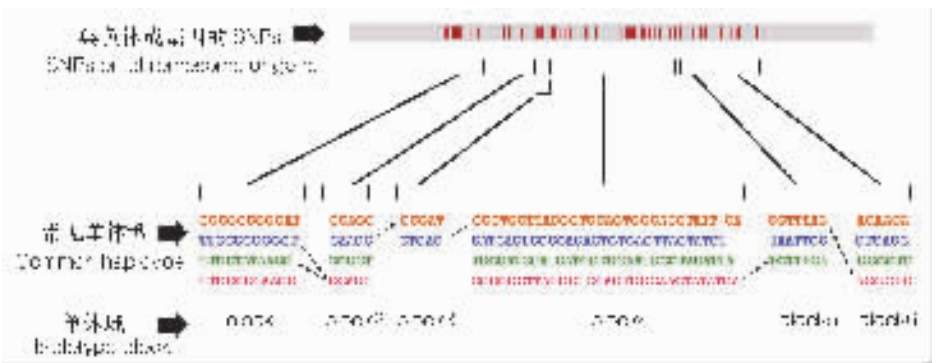

图 3 人类染色体或基因单体型的域状结构示意图(修改自 Daly 等[28])

Fig.3 Block-like haplotype diversity at human chromosome or genes(revised from Daly *et al.* [28])

自从产生了单体型的概念,就存在一些不同的定义法,主要是基于单体型多样性或连锁不平衡的两类方法。基于单体型多样性的方法将单体域定义为具有有限的单体型的片段[28,30,31]。基于连锁不平衡的方法使用配对不平衡来寻找重组多发的区段,将其作为单体域的边界[32~34]。

3.1 基于单体型多样性的定义

基于单体型的多样性,Patil 等[30]首先将单体域定义为包括样本中所有单体型中 80% 以上多态(即出现一次以上的单体型)的区域。根据这个定义,Patil 等提出了获得单体域近似分割的贪婪算法。简而言之,就是首先考虑由连续 SNPs 形成的所有可能的单体域,然后从中选出一个单体域,使得该域中的 SNP 数目与所需最少的标签 SNPs(用来区分的出现一次以上单体型)数目之比值达到最大,也就是用最少的标签 SNP 区分出最多的 SNP。这一过程不断重复直至一套覆盖整个染色体的相互连接且无重叠的单体域被选择出来,这样每个 SNP 都被安排一个单体域中。他们对人类 21 号染色体的单体域结构进行了分析,用 4 563 个标签 SNP 确定了总计 4 138 个单体域。结果表明,大部分单体域包含的 SNPs 少于 10 个,平均每个域中有 2.7 个 SNPs。所有单体域的大小与其在染色体上的顺序无关,平均为 7.8 kb。由 21 号染色体的部分单体域结构[30]可以看出,用不到 10% 的域中 SNP 就可以确定整个样本中 80% 的单体型结构。Patil 等的单体域分割算法以样本的遗传信息含量为基础,因此单体域没有绝对的边界。

Zhang 等[35,36]提出了单体域分割的动态程序算法,算法的原理是使每个单体域中能代表域中大部分性质的标签 SNPs 达到最少。该算法对人类 21 号染色体数据的分析,与 Patil 等的研究相比,使得单体域与标签 SNPs 的数目减少 20% ~40%,分别为 2575 和 3582。他们的算法已经被开发为程序 HAPBLOCK (http://hto-b.usc.edu/~msms/HapBlock/)。

3.2 基于连锁不平衡的定义

基于连锁不平衡的定义法首先由 Gabriel 等[32]给出,他将单体域定义为只包含一些常见单体型,且几乎没有发生过重组的一套连续的位点,于是在生物学方面确定单体域就是检测每个区间内的重组交换方式。估计成对 SNPs 重组历史的常用方法就是检测两两位点间连锁不平衡的方法——D'[37,38],这里所谓连锁不平衡是指相邻的两个位点的等位基因同时出现在一个单体型中的次数多于自由分离重组的期望值[39]。估计染色体区段内两两位点间 D',依据 D' 与预先所确定的域值范围(判断位点间是强连锁不平衡还是曾经发生重组)的关系而进行染色体上单体域的分割。Gabriel 等[32]在不同人群(Nigeria/Yoruba, Asia, African Americans, Europeans)中,分析了 51 个常染色体片段(共计 13 Mb),检测了群体内与群体间单体域结构的相似性,不同人群中的单体域的差异,不仅可以解释其自身独特的历史迁徙情况,也可以解释人群间某些疾病易感性的差异。依据 Gabriel 对单体域的定义,Philips 等[33]使用公共遗传标记数据构建了 19 号染色体的单体型,结果发现染色体的三分之一被单体域所覆盖。

单体型的不同定义与算法,甚至是相同定义,个人主观决定的域值不同,都会导致单体域结构的变化,这样使得比较不同研究的结果与研究单体域结构的机制变得非常困难。基于位点间重组交换的分布,Wang 等[34]利用四配子检验法(four-gamete test,FGT)提出了另一种单体域的算法。算法首先对成对的 SNPs 进行四配子检验(检测到 4 个配子就表示曾经发生重组),将两两位点的四配子状态用矩阵表示,有 4 个配子出现计为 1,否则为 0;单体域被定义为没有重组现象发生的一组有序 SNP 标记,也就是根据 FGT 的结果,只要配子数不超过 3 个,就不断累加 SNP 到一个域中,直到第 k 个位点出现 4 个配子而结束。位点 k 可作为另一个新域的一个突变起始点。FGT 算法的优点之一就是无需预先设定域值。

尽管上述两类方法各具优点,但 Wall 等[34]指出更倾向于第二类方法的原因:其一,使用 D' 直接检测历史性重组的发生看起来更符合单体域的定义;其二,对于二倍体的遗传数据,两两配对的方法更容易应用;最后,两两配对连锁不平衡的系数更易于可视化[40]。

4 标签 SNP 的选择

研究者一般通过比较患者和非患者的基因组的不同来发现影响某种疾病例如糖尿病的基因。在两

组单体型频率不同的染色体区域,就可能包含着疾病相关基因。理论上,研究者通过对全部 1 500 万个 SNP 位点都进行基因分型,也能够寻找到这样的区域。但是,目前用这种方法进行鉴定的成本是过于昂贵。通过单体图计划将鉴定出约 50 万个标签 SNP 位点,从而提供与一千万个 SNP 位点大致相同的图谱信息。这样将大幅度地减少成本使研究易于进行。

有研究表明,用少部分的遗传标记仍可保留单体型的大部分信息[30]。这就引出了如何选择单体型标签 SNP(haplotype tag SNP,htSNP)的问题。一些文献将 htSNP 定义为构建样本单体型或进行与相关分析所必须的一组遗传标记(SNP)[30,41]。htSNP 的选择问题与计算中的最小化问题相似。目前主要方法如下:

Patil 等[30]将 htSNP 解释为能够分辨样本中至少 α% 的单体型,且位点数目最少的一组 SNPs。Zhang 等[35]所构建单体域的动态程序中,采用枚举法来选择 htSNP。这些方法就是先将染色体分割成连续的单体域,然后通过肉眼观察或程序运算从每个域中选择出可以代表域中多样性的标签 SNPs,并要求标签 SNPs 的数目达到最小。

Johnson 等[41]提出的算法是以连锁不平衡为基础的,原理是首先计算两两 SNPs 间连锁不平衡程度,如果高度连锁,那么就可以用一个来预测另一个,就是说只选择其中一个作为 htSNP。算法依据连锁不平衡的参数,剔除冗余的 SNPs,列出所有可能的 htSNPs 子集,然后根据各组 htSNP 能够说明样本中单体型变异的多少(多样解释比例,proportion of diversity explained,PDE)来确定最佳的一组为 htSNPs。

Clayton 算法的原理是让进行基因分型的标签 SNPs 能够对剩余的不分型 SNPs 进行很好的预测[42],依据其原理建立算法选出可能的 htSNPs 子集,最后同样依据 PDE 来选出 htSNPs。与 Johnson 的算法不同的是该方法可以按照使用者的要求在 PDE 分析之前剔除覆盖率达不到要求或有缺失的数据的 htSNPs 子集。

Stram 等[43]选择标签 SNPs 算法就是让标签 SNPs 能够对总体 SNPs 的分布进行较好的预测。算法考虑每个个体真实的单体型拷贝数与通过标签 SNPs 所预测的单体型拷贝数的相关程度,并将相关系数的平方 R^2 作为选择 htSNPs 的参数。

5 单体型的关联分析方法

经典的连锁作图与克隆定位无论现在还是将来仍是确认稀有、高风险、与疾病相关突变的有效方法之一,但对于相对低风险、复杂的性状,关联分析将成为更有效的方法[44]。单体型频率的估计,单体域的确定以及标签 SNPs 的选择可以用作基因的精细定位或候选基因分析,但更重要的作用是进行复杂遗传疾病或药物反应的关联分析。估计单体型(标签 SNPs)与感兴趣的性状间的相关,最简单的方法就是将单体型多态作为一个因素,同时考虑年龄、性别等其他影响因素进行关于性状的回归分析。Logistic 回归可以进行疾病研究中常碰到的二级计分或二类评定性状(例如"生"与"死"或疾病与对照)的回归分析[45]。简单回归分析面临的一个问题是单体型不确定性存在,对一个个体的单体型的推断往往得出的是以不同的概率出现的多个单体型。采用混合模型[46]或比分检验(score test)[47]进行疾病性状与单体型的关联分析可以将概率作为权重直接整合到统计模型中去,解决了单体型不确定的问题。此外,无需明确的模型,在病例对照(case-control)研究中将病例与对照间单体型频率进行简单的卡方检验也可作为性状与单体型间相关分析的一个选择[48]。若以家系为基础,进行病例双亲对照(case-parental)研究(即将双亲作为患病孩子的对照),可以采用传递不平衡检验(the Transmission/Disequilibrium Test,TDT)进行关联分析[49,50]。

6 SNP 数据库的利用

无论 SNPs 的基因型的确定、单体型的推断、单体域的构建,还是与目标性状的关联分析,都要涉及到 SNP 数据库的使用。目前主要的公共数据库:

(1)美国国立生物技术信息中心(NCBI)的遗传变异数据库(dbSNP)[51,52] http://www.ncbi.nlm.nih.gov/SNP/index.html ;

(2)SNP 研究委员会(The SNP Consortium,TSC)提供的数据库[53] http://snp.cshl.org/db/snp/mHp ;

(3)人类基因组突变数据库(the Human Genome Variation,HGVbase)[54] http://hgvbase.cgb.ki.se/ 。

此外,可利用的公开 SNP 网上资源还包括:

美国国立卫生院(National Institutes of Health,NIH)提供的主要是与癌症和肿瘤相关的候选 SNP 数据库:http://lpg.nci.nih.gov/;

美国怀特和特研究所(Whitehead Institute for Biomedical Research Genome Institue)建立的人类 SNP 数据库:http://www.genome.wi.nit.edu/SNP/humHn/index.html;

华盛顿大学的按染色体位置组织的 SNP 数据库:http://www.ibc.wustl.edu/SNP;

瑞典卡尔林斯卡研究院(Karolinska Institute of Sweden)建立的 HGBase(Human Genic Bi-Allelic SEquences)数据库:http://hgbHse.cgr.ki.se。

dbSNP 数据库包括从近 300 个来源得出的约 985 万无冗余的人类 SNPs(http://www.ncbi.nlm.nih.gov/SNP/snp_summHry.cgi),我们还可以通过 dbSNP 数据库连接的 OMIM(Online Mendelian Inheritance in Man)等数据库查询某一遗传变异所引发的疾病或某一疾病的致病遗传变异。HGVbase 包含从 800 个不同来源得到的近 100 万个无冗余的 SNPs。目前公用数据库中的 SNPs 距离预测的 1 500 万这个总数还有较大差距,而且只有少部分 SNPs 有总群频率和特定种族频率的说明。

随着 HapMap 项目的完成,将向公众公布所有的实验数据至 dbSNP 数据库,包括 SNP 位点、SNP 基因分型实验设计、SNP 检定结果和频率,以及构建的单体型。当对染色体区域进行了足够的 SNP 分型来确定紧密连锁的区域时,这些区域的单体型、个体的基因型和标签 SNPs 将无条件地公开发布。

中华民族的单核苷酸多态性(SNP)数据库日前在国家人类基因组南方研究中心(南方中心)建成。该数据库收集了约 800 个疾病相关基因或具有重要功能以及药物代谢和效应相关基因,以及 21 号染色体上的 127 个已知基因的 SNP。中国人群在这些基因上的多态以及这些多态的性质,如 SNP 在基因的位置,是否引起氨基酸的改变及其在某些人群中的频率等等,其中约 3/4 为首次发布的数据,与现有国际公共数据库数据形成了良好互补。

7 SNP 与单体型的应用

心血管疾病、癌症、肥胖、糖尿病、精神病等常见病都是由多个基因与环境共同作用的结果。运用人类基因组的 SNPs 与单体型信息来挖掘这些常见病的遗传因素将使人们对人类疾病的发病机理,诊断和治疗产生全新的认识。连锁分析已经成功定位了许多单基因疾病的遗传变异,然而在定位影响常见复杂疾病的遗传变异时连锁分析往往是失败的,这些遗传变异影响着个体的疾病风险。寻找遗传疾病风险因子的一个补充方法就是通过比较患病组与未患病的对照组来寻找遗传变异与疾病之间的相关。SNP 是人体 DNA 序列变异最普遍的形式,单体型与单体域的构建使我们可以用很少的标签 SNP 来代表全部的 SNP 或整个基因或染色体的单体型来进行相关的研究应用。通过基因组的标签 SNPs 与复杂性疾病或药物反应的相关分析,可以揭示复杂性疾病的致病机理与疾病的不同临床表型,也可作为实行个体化治疗的根据[44,55,56]。

通过对候选基因或相关区域的单个或多个 SNPs 或单体型与复杂性状的关联分析寻找复杂疾病风险的遗传因素已经有很多报道。已有的研究表明:编码蛋白激酶 AKT2/PKB β 中一个单核苷酸的变化导致酶的活性变化,从而引起糖尿病[57]。对于更多的复杂疾病,往往是由多个突变位点共同影响和决定的。对维生素 D 受体(VDR)基因与前列腺癌的早发风险的相关研究表明:*VDR* 基因 4 个 SNPs 位点(*Fok* Ⅰ,*Bsm* Ⅰ,*Apa* Ⅰ,*Taq* Ⅰ 酶切位点)的单体型频率在美国黑人群体的患病组与对照组中差异达到极显著[58]。胰岛素样生长因子 3(*IGFBP*3)基因 5 个遗传变异位点(A-202C,G227C,C3804G,5606InsA,and C5827T)多态与中国妇女乳腺癌患病风险的关联分析显示,在 A-202C,G227C,5606InsA 和 C5827T 位点纯合的个体乳腺癌的患病风险增加 30% 至 60%,而且 5 个位点的单体型也与乳腺癌的患病风险显著相关[59]。通过比较患病组与对照组 *AOPE* 基因单体型频率证明了阿滋海默症与 *AOPE* 基因的相关[48];此外,还有研究表明:*HLA-DPB*1 基因的 SNP 多态与 1 型糖尿病的患病风险相关[60];墨西哥裔美国人中脂蛋白脂酶基因 6 个多态位点的单体型 1 对冠状动脉疾病有保护作用[61]。近期,我国军事医学科学院贺福初院士等日前发现人类雌激素受体基因上的单核苷酸多态性能够影响慢性乙型肝炎的发病风险[62]。这众多的研究结果将为复杂疾病预防、诊断和治疗打下坚实的基础。

除了用于常见复杂疾病及患病风险的关联分析，SNPs 与单体型还可以应用到药物基因组的研究中。药物基因学（Pharmacogenomics）就是研究遗传变异如何影响个人对药物反应的不同的科学。β_2 肾上腺素受体基因上 13 个 SNPs 所形成的不同单体型间，哮喘病治疗药物的药物反应显著相关，这个研究就是 SNPs 在药物基因组学中应用的典型例子[63]。通过 SNPs 与药物反应的相关分析能够显示出在不同个体的药物作用目标或药物代谢途径中的某个酶的差异，揭示个体的基因组多态与疾病治疗药物反应之间的关系。这就让我们可以预测出哪种药或疫苗对那些携带特殊基因型的个人最有效，让医生做到对人下药，增加临床试验的成功率。此外，对基因组多态与药物反应研究将促进个人化药物的开发。

随着人类基因组单型图（HapMap）的逐步完成，SNPs 与单体型的研究必将在探究复杂性遗传疾病的遗传机理、患病风险与药物反应不同中扮演重要角色。尽管在过去的几年中，关于人类基因组单体型结构及应用的研究已经取得了很大的进步，但仍存在很多问题有待解决与完善。例如，如何结合单体型数据寻找形成人类单体域结构的因素；如何利用单体型来进行可疑致病位点的定位。伴随单体型数据的剧增，更为迫切需要的是发展更为有效的设计与统计分析手段，从而在研究中考虑到更多的多态位点、更复杂的性状（如纵向性状）、单体域间的连锁、系谱信息、上位效应以及与环境的互作等问题，使得我们对单体型、单体域的研究更为有效与完善。

参考文献（References）：

[1] Jiang R, Duan J, Windermuth A, Stephens J C, Judson R, Xu C B. Genome-wide evaluation of the public SNP databases. *Pharmacogenomics*, 2003, 4(6): 779 ~ 789.

[2] The International HapMap Consortium. The international HapMap project. *Nature*, 2003, 426: 789 ~ 796.

[3] ZENG Chang-Qing. The international HapMap project. *Bulletin of Biology*, 2004, 39(2): 1 ~ 3.
曾长青. 国际基因组单体型图计划. 生物学通报, 2004, 39(2): 1 ~ 3.

[4] Nelson M R, Marnellos G, Kammerer S, Hoyal C R, Shi M M, Cantor C R, Braun A. Large-scale validation of single nucleotide polymorphisms in pene regions. *Genome Research*, 2004, 14: 1664 ~ 1668.

[5] Hoehe M R. Haplotypes and the systematic analysis of genetic variation in genes and genomes. *Pharmacogenomics*, 2003, 4(5): 547 ~ 570.

[6] Morris R W, Kaplan N L. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol*, 2002, 23: 21 ~ 233.

[7] Hoehe M R, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd K K, Berrettini W H, Church G M. Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. *Hum Mol Genet*, 2000, 9(19): 2895 ~ 2908.

[8] Davidson S. Research suggests importance of haplotypes over SNPs. *Nat Biotechnol*, 2000, 18(11): 1134 ~ 1135.

[9] Ruano G, Kidd K K, Stephens J C. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc Natl Acad Sci USA*, 1990, 87: 6296 ~ 6300.

[10] Michalatos-Beloin S, Tishkoff S A, Bentley K L, Kidd K K, Ruano G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res*, 1996, 24(23): 4841 ~ 4843.

[11] Bradshaw M S, Bollekens J A, Ruddle F H. A new vector for recombination-based cloning of large DNA fragments from yeast artificial chromosomes. *Nucleic Acids Res*, 1995, 23: 4850 ~ 4856.

[12] Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland C R, Lynch H T, Chadwick RB, de la Chapelle A, Berg K, Eshleman J R, Yuan W, Markowitz S, Laken S J, Lengauer C, Kinzler K W, Vogelstein B. Conversion of diploidy to haploidy. *Nature*, 2000, 403: 723 ~ 724.

[13] Douglas J A, Boehnke M, Gillanders E, Trent J M, Gruber S B. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet*, 2001, 28: 361 ~ 364.

[14] Schaid D J. Relative efficiency of ambiguous vs directly measured haplotype frequencies. *Genet Epidemiol*, 2002, 23: 426 ~ 433.

[15] Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet*, 1996, 58(6): 1323 ~ 1337.

[16] Kruglyak L, Daly M J, Reeve-Daly M P, Lander E S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, 1996, 58(6): 1347 ~ 1363.

[17] Schaid D J, McDonnell S K, Wang L, Cunningham J M, Thibodeau S N. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet*, 2002, 71(4): 992 ~ 995.

[18] Clark A G. Inference of haplotypes from PCR-amplified samples o diploid populations. *Mol Biol Evol*, 1990, 7: 111 ~

122.

[19] Clark A G, Weiss K M, Nickerson D A, Taylor S L, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing C F. Haplotype structure and population genetic inferences form nucleotide sequence variation in human lipoprotein lipase. *Am J Hum Gene*, 1998, 63: 595 ~ 612.

[20] Excoffier L, Slatkin M. Maximization likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 1995, 12: 921 ~ 927.

[21] Tregouet D A, Escolano S, Tirel L, Mallet A, Golmard J L. A new algorithm for haplotype based assocaition analysis: the stochastic EM. *Ann Human Gent*, 2004, 68: 165 ~ 177.

[22] Zhao H Y, Pfeiffer R, Gail M H. Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, 2003, 4(2): 171 ~ 178.

[23] Stephens M, Smith N J, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Gent*, 2001, 68: 978 ~ 989.

[24] Niu T, Qin Z S, Xu X, Xu X, Liu J S. Bayesian haplotype inference for multiple kinked single nucleotide polymorphisms. *Am J Hum Gent*, 2002, 70: 157 ~ 169.

[25] Lin S, Cutler D J, Zwick M E, Chakravarti A. Haplotype inference in random population samples. *Am J Hum Genet*, 2002, 71(5): 1129 ~ 1137.

[26] Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Gent*, 2003, 73: 1162 ~ 1169.

[27] Lin S, Chakravarti A, Cutler D J. Haplotype and missing data inference in nuclear families. *Genome Res*, 2004, 14(8): 1624 ~ 1632.

[28] Daly M J, Rioux J D, Schaffner S F, Hudson T J, Lander E S. High-resolution haplotype structure in the human genome. *Nature Genet*, 2001, 29: 229 ~ 232.

[29] Jeffreys A J, Kauppi L, Neumann R. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet*, 2001, 29: 17 ~ 222.

[30] Patil N, Berno A J, Hinds D A, Barrett W A, Doshi J M, Hacker C R, Kautzer C R, Lee D H, Marjoribanks C, McDonough D P, Nguyen B T N, Norris M C, Sheehan J B, Shen N, Stern D, Stokowski R P, Thomas D J, Trulson M O, Vyas K R, Frazer K A, Fodor S P A, Cox D R. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 2001, 294: 1719 ~ 1723.

[31] Dawson E, Abecasis G R, Bumpstead S, Chen Y, Hunt S, Beare D M, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley D R, Cardon L R, Dunham I. A first generation linkage disequilibrium map of human chromosome 22. *Nature*, 2002, 418: 544 ~ 548.

[32] Gabriel S B, Schaffner S F, Nguyen H, Moore J M, Roy J, Blumenstie B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero S N, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander E S, Daly M J, Altshuler1 D. The structure of haplotype blocks in the human genome. *Science*, 2002, 296: 2225 ~ 2229.

[33] Phillips M S, Lawrence R, Sachidanandam R, Morris A P, Balding D J, Donaldson M A, Studebaker J F, Ankener W M, Alfisi S V, Kuo F S, Camisa A L, Pazorov V, Scott K E, Carey B J, Faith J, Katari G, Bhatti H A, Cyr J M, Derohannessian V, Elosua C, Forman A M, Grecco N M, Hock C R, Kuebler J M, Lathrop J A, Mockler M A, Nachtman E P, Restine S L, Varde S A, Hozza M J, Gelfand C A, Broxholme J, Abecasis G R, Boyce-Jacino M T, Cardon L R. Chromosome-wise distribution of haplotype blocks and the role of recombination hotspots. *Nature Genetics*, 2003, 33: 382 ~ 387.

[34] Wang N, Akey J M, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am J Hum Genet*, 2002, 71: 1227 ~ 1234.

[35] Zhang K, Deng M, Chen T, Waterman M S, Sun F. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA*, 2002, 99: 7335 ~ 7339.

[36] Zhang K, Sun F, Waterman M S, Chen T. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am J Hum Genet*, 2003, 73: 63 ~ 73.

[37] Pritchard J K, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 2001, 69: 1 ~ 14.

[38] Lewontin R C. The interaction of selection and linkage I General considerations; heterotic models. *Genetics*, 1964, 49: 49 ~ 67.

[39] Jeffrey D W, Jonathan K P. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews*, 2003, 4: 587 ~ 597.

[40] Wall J D, Pritchard J K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews*, 2003, 4: 587 ~ 597

[41] Johnson G C, Esposito L, Barratt B J, Smith A N, Heward J, Di Genova G, Ueda H, Cordell H J, Eaves I A, Dudbridge F, Twells R C, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough S C, Clayton D G, Todd J A. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 2001, 29: 233 ~ 237.

[42] Ke X, Cardon L R. Efficient selective screening of haplotype

tag SNPs. *Bioinformatic*,2003,19 :287 ~288.

[43] Stram D O,Haiman C A,Hirschhorn J N,Altshuler D,Kolonel L N,Henderson B E,Pike M C. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered*, 2003, 55(1) :27 ~36.

[44] Botstein D,Risch N. Discovering genotypes underlying human phenotypes :past success for mendelian disease,future approaches for complex disease. *Nature Genet*,2003, 33 :228 ~237.

[45] Wallenstein S,Hodge S E,Weston A. Logistic regression model for analyzing extended haplotype data. *Genet Epidemiol*,1998,15 :173 ~181.

[46] Keavney B,McKenzie C A,Connell J M,Julier C,Ratcliffe P J,Sobel E,Lathrop M,Farrall M. Measured haplotype analysis of the angiotensin-1 converting enzyme (ACE) gene. *Hum Mol Genet*,1998,7 :1745 ~1751.

[47] Schaid D J,Rowland C M,Tines D E,Jacobson R M,Poland G A. Score tests for association between traits and haplotypes when lingkage phase is ambiguous. *Am J Hum-Genet*,2002,70 :425 ~434.

[48] Fallin D,Cohen A,Essioux L,Chumakov I,Blumenfeld M,Cohen D,Schork N J. Genetic analysis of case/control data using estimated haplotype frequencies :application to APOE locus variation and Alzheimer's disease. *Genome Res*,2001,11 :143 ~151.

[49] Spielman R S,McGinnis R E,Ewens W J. Transmission test for linkage disequilibrium :the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*,1993,52 :506 ~516.

[50] Clayton D,Jones H. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet*,1999,65 : 1161 ~1169.

[51] Sherry S T,Ward M,Sirotkin K. dbSNP database for single nucleotide polymorphisms and other calasses of minor genetic variation. *Genome Res*,1999,9 :677 ~670.

[52] Sherry S T,Ward M H,Kholodov M,Baker J,Phan L,Smigielski E M,Sirotkin K. dbSNP :the NCBI database of genetic variation. *Nucleic Acids Res*,2001,29 :308 ~311.

[53] Masood E. A consortium plans free SNP map of human genome. *Nature*,1999,398(6728) :545 ~546.

[54] Fredman D,Siegfried M,Yuan Y P,Bork P,Lehvaslaiho H,Brookes A J. HGVbase :a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res*,2002,30(1) :387 ~391.

[55] Lander E S. The new genomics :Global views of biology. *Science*,1996,274 :536 ~539.

[56] Reich D E,Cargill M,Bolk S,Ireland J,Sabeti P C,Richter D J,Lavery T,Kouyoumjian R,Farhadian S F,Ward R,Lander E S. Linkage disequilibrium in the human genome. *Nature*,2001,411(6834) :199 ~204.

[57] George S,Rochford J J,Wolfrum C,Gray S L,Schinner S,Wilson J C,Soos M A,Murgatroyd P R,Williams R M,Acerini C L,Dunger D B,Barford D,Umpleby A M,Wareham N J,Davies H A,Schafer A J,Stoffel M,O'Rahilly S,Barroso I. A family with severe Insulin resistance and Diabetes due to a mutation in AKT2. *Science*,304 :1235 ~ 1238.

[58] Oakley-Girvan I,Feldman D,Eccleshall T R,Gallagher R P,Wu A H,Kolonel L N,Halpern J,Balise R R,West D W,Paffenbarger R S Jr,Whittemore A S. Risk of early-onset prostate cancer in relation to germ line polymorphisms of the Vitamin D receptor. *Cancer Epidemiol Biomarkers Prev*, 2004,13(8) :1325 ~1330.

[59] Ren Z,Cai Q,Shu X O,Cai H,Li C,Yu H,Gao Y T,Zheng W. Genetic polymorphisms in the IGFBP3 gene :association with breast cancer risk and blood IGFBP-3 protein levels among Chinese women. *Cancer Epidemiology Biomarkers & Prevention*,2004,13 :1290 ~1295.

[60] Cruz T D,Valdes A M,Santiago A,Frazer de Llado T,Raffel L J,Zeidler A,Rotter J I,Erlich H A,Rewers M,Bugawan T,Noble J A. DPB1 alleles are associated with type 1 Diabetes susceptibility in multiple ethnic groups. *Diabetes*,2004,53 :2158 ~2163.

[61] Goodarzi M O,Guo X,Taylor K D,Quinones M J,Samayoa C,Yang H,Saad M F,Palotie A,Krauss R M,Hsueh W A,Rotter J I. Determination and use of haplotypes :ethnic comparison and association of the lipoprotein lipase gene and coronary artery disease in Mexican-Americans. *Genet Med*,2003,5(4) :322 ~327.

[62] Deng G,Zhou G,Zhai Y,Li S,Li X,Li Y,Zhang R,Yao Z,Shen Y,Qiang B,Wang Y,He F. Association of estrogen receptor polymorphisms with susceptibility to chronic hepatitis B virus infection. *Hepatology*,2004,40(2) :318 ~326.

[63] Drysdale C M,McGraw D W,Stack C B,Stephens J C,Judson R S,Nandabalan K,Arnold K,Ruano G,Liggett S B. Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci USA*, 2000, 97(19) :10483 ~10488.