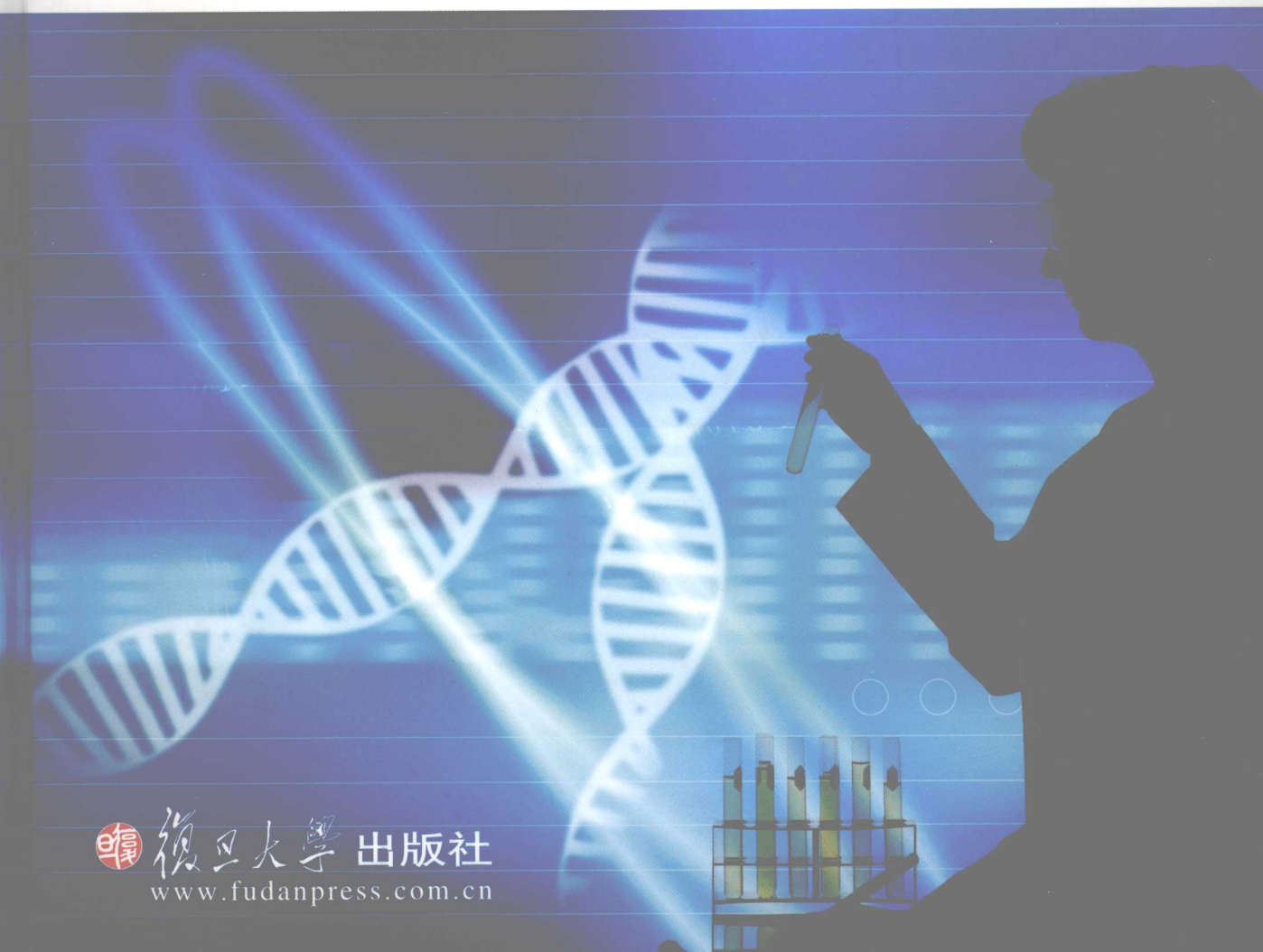
国家科学技术学术著作出版基金
NFAPST

生物学前沿技术在医学研究中的应用

主编 马 端



复旦大学出版社
www.fudanpress.com.cn

国家科学技术学术著作出版基金资助

生物学前沿技术在医学研究中的应用

主　编　马　端
副主编　杜　新　孔德升

复旦大学出版社

图书在版编目(CIP)数据

生物学前沿技术在医学研究中的应用/马端主编. —上海：
复旦大学出版社,2007.9
ISBN 978-7-309-05639-6

Ⅰ. 生… Ⅱ. 马… Ⅲ. 生物医学工程-研究 Ⅳ. R318

中国版本图书馆 CIP 数据核字(2007)第 109626 号

生物学前沿技术在医学研究中的应用
主编 马 端

出版发行 复旦大学出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 宫建平
总 编 辑 高若海
出 品 人 贺圣遂

印 刷 上海第二教育学院印刷厂
开 本 787×1092 1/16
印 张 26.25
字 数 575 千
版 次 2007 年 9 月第一版第一次印刷
印 数 1—3 100

书 号 ISBN 978-7-309-05639-6/R·987
定 价 58.00 元

内容提要

伴随着功能基因组时代的到来，生物技术日新月异，推动着生命科学的高速发展。掌握和运用这些新技术，是一个生物研究工作者在科研和产品开发中居于领先地位的关键。《国家中长期科学和技术发展规划纲要（2006-2020）》中指出：生物技术和生命科学将成为21世纪引发新科技革命的重要推动力量，基因组学和蛋白质组学研究正在引领生物学技术向系统化研究方向发展。据此，我们联合了复旦大学、中国科学院上海生命科学研究院、上海交通大学、美国Rockefeller大学、美国Van Andel研究所等机构的学者和研究人员，共同编著了本书。

本书共有15章，分别从基因克隆、遗传调控、信号转导、细胞凋亡、基因定向敲除、芯片技术、基因治疗、干细胞与组织工程、肽库构建、蛋白质组学、疫苗构建、纳米材料与药物制剂、蛋白质工程、生物信息学等15个层面对每个领域中的关键技术进行了深入浅出的介绍，并附有实例和操作方法。本书的对象是有志于应用最新生物学技术开展医学研究的医学工作者和研究生，对本科阶段学习的医学生也有指导作用。

编委会名单

（按姓氏笔画排序）

前言

自20世纪90年代以来，生物学新技术层出不穷，极大地推动了医学研究的进步。在如此之多的新技术中，有些与医学关系密切，有些则不然。从医学角度来看，寻找疾病的发病原因，阐明疾病的发病机制，建立疾病早期诊断方法，发明高效低毒的药物，始终是受到密切关注的关键问题。尽管医学模式已由生物医学模式发展到生物-心理-社会医学模式，但疾病的发生、发展和防治仍然是建立在基因-RNA-蛋白质-细胞水平变化之上的。寻找致病基因、药物靶基因和疾病相关蛋白质，探索它们的调控机制，在疾病发生的早期进行诊断和干预治疗，是每一个医学研究者感兴趣的课题。我们在多年医学研究和教学工作中体会到，让每一个初级和中级科研人员，特别是硕士和博士研究生，在浩如烟海的大部头专著中去寻找他们感兴趣的研究方法，不仅工作量大，而且事倍功半。因此，我们组织了一批国内外在本专业领域中有造诣的青年学者编写了本书，希望能够在有限的篇幅里，为广大的医学研究生和科研人员提供当今前沿和实用的生物学技术的理论和方法，让他们在较短的时间里为自己的研究方向设计出研究内容和实验方案。如果这种目的能够达到，我们将感到非常欣慰。

本书涉及致病基因的寻找与克隆、遗传和表观遗传调控、信号转导、细胞凋亡、基因芯片和蛋白质组学、基因的组织特异性敲除、肽库的建立及应用、遗传性疾病的基因治疗、干细胞与组织工程、纳米材料与新型药物制剂、蛋白质工程与新药开发、生物信息学等内容。特点是既有理论知识，又有具体实验方法及其应用，是一本理论联系实际的实用型参考书。

马　端

2007年5月

目录

第一章

疾病相关基因的筛选与克隆

流行病学和遗传学研究表明，疾病是环境因素与遗传因素相互作用而形成的一种特殊生命过程。除了获得性疾病外，糖尿病、精神分裂症等常见疾病和血色素沉着病等罕见疾病的发病均与遗传因素密切相关，存在着遗传易感性(susceptibility)。即便是感染性疾病，感染与否亦与个体本身的遗传因素息息相关。根据涉及基因的多少，可初步把疾病分为单基因疾病(monogenic disease)和复杂疾病(complex disease)。单基因疾病是疾病发病的遗传因素中以单基因缺陷占主导地位，表现为单个基因的缺失或突变，且在家系成员中疾病的传递符合孟德尔规律者，也称为孟德尔遗传病。符合孟德尔遗传的单基因疾病约 6 000 多种，这类疾病多数症状明显，但发病率低，对人群的危害性不大。复杂疾病则是由多个基因及环境因素相互作用所致，在家系成员中疾病的传递不符合孟德尔规律，又称为多基因疾病(polygenic disease)。肿瘤、冠心病、糖尿病、精神分裂症和高血压等疾病均属于多基因疾病，这些疾病发病率高，严重危害着人类健康。因此，定位克隆这些疾病的相关基因其社会意义重大，同时具有很大的现实和潜在的经济意义。

人类基因组含 2 万多个基因，如何从众多的基因中搜寻到与疾病相关的基因，不仅是生命科学研究的难点，亦是研究的热点。由于单基因疾病遗传模式简单，涉及基因数少，多采用定位克隆(positional cloning)方法，目前约有 5 000 多个单基因疾病的相关基因已被定位，并克隆出约 1 300 个致病基因。经典定位克隆的成功例子有血色素沉着病(hemochromatosis)、指甲髌骨综合征(nail－patella syndrome)和乳糖耐受不良症(lactose intolerance)等。

定位克隆策略在单基因疾病上获得的巨大成功促使人们将这一研究策略应用于常见的复杂疾病。但迄今为止，对于糖尿病等大多数常见的多基因疾病，鲜有用定位克隆技术获得相关基因的成功报道。仅有的少数成功报道提示，严格选择研究样本有着重要意义，如采用发病年龄早的家系(接近孟德尔方式遗传)，成功的例子有发现 *BRAC*1 和 *BRAC*2 在乳腺癌和卵巢癌中的作用，以及 *hMSH2* 基因突变与遗传性非息肉性结肠癌(hereditary non－polyposis colon cancer，HNPCC)的相关性；或采用隔离群体(减少遗传异质性)，内皮

素受体B(EDNRB)第4外显子的一个无义突变是荷兰Mennonites地区先天性巨结肠(Hirschsprung's disease)的一个致病因素，以及过氧化氢酶基因启动子区－844 C/T多态与原发性高血压关联就是这样被发现的。

多基因疾病相关基因研究进展缓慢，主要是因为与疾病危险性相关的突变往往在机体遗传因素和环境因素双重作用下，才导致机体发病，并且多基因疾病确切发病机制目前尚不十分清楚，以及存在着遗传异质性(heterozygosity)、种族差异等因素，这在客观上限制了多基因疾病易感基因的研究进展。

一、疾病相关基因的搜寻与克隆的研究历程

1. 从功能克隆到遗传分析

20世纪40～60年代，许多研究生物化学和生物大分子的手段如电子显微镜、生物分子分离、提纯、蛋白质电泳相继出现，使人们在分子水平上研究疾病病因成为可能。许多酶缺陷的疾病就是利用这些方法将其基因确定的，如苯丙酮尿症等。功能克隆是基于疾病与正常之间明显可见的或直接与生化功能相关的线索确定与疾病相关的基因，主要是利用基因的产物蛋白质或RNA。但对于许多无法得到与疾病相关产物的情况，仅仅从疾病的表型特征几乎是不可能得到致病基因的，即使按前述方法得到了许多基因，人们也不可能得到每一基因的具体功能。因此，随着许多疾病研究的进行，人们逐渐认识到了遗传分析的优越性，至少人们可以获得从表型到基因型的直接联系。

2. 从定位克隆到候选基因关联研究

随着人类基因组遗传图谱和物理图谱的制作，出现了"定位克隆"的全新思路，发现了包括囊性纤维化(cystic fibrosis，CF)、Huntington舞蹈病、遗传性结肠癌、乳腺癌等一大批疾病基因。但在多基因疾病研究中，由于关联研究相对于连锁分析的有效性，因此受到更多的重视。人类基因组序列图已公布，所有人类基因均被精确地定位于染色体的各个区域，使得候选基因关联研究成为可行。候选基因关联研究不仅提高了复杂疾病关联研究的统计学效率，同时也促进了对诸如表型、组织、基因和蛋白质等可能与疾病有关因素的了解。

3. 从单基因疾病到多基因疾病

单基因疾病由于因素单一，基因型-表型关系清晰，已经形成了较为成熟的致病基因定位、分离技术体系，迄今已鉴定克隆了很多单基因疾病的致病基因。目前研究的趋势是更加注重对基因功能以及相关生物化学通路的阐明。由于基因组研究成果的不断改观和遗传统计分析方法的进展，目前国际上疾病基因组学的研究热点是多基因疾病。已知多基因疾病是由多个基因的累加作用和某些环境因素作用所致，这些基因的单核苷酸多态性(single nucleotide polymorphisms，SNP)及其特定组合可能是造成疾病易感性最重要的原因。因此，对疾病相关调节通路的候选基因进行SNP的关联研究，可能是多基因疾病研究取得突破的希望所在。

二、疾病相关基因的搜寻与克隆的研究策略

当前疾病相关基因的搜寻与克隆的研究策略依然主要采用关联研究和定位克隆，其他如全基因组扫描、全基因组关联研究、候选基因关联研究等策略和方法都是从上述两种策略衍生出来的。另外，尚有基于染色体异常和动物模型等一些方法和策略。但值得注意的是所采用的策略方法并非孤立，而是相互借鉴，综合应用。随着对基因组特性和各种策略方法认识的加深、新的研究方法的出现以及分型技术的发展和成本的降低，有助于阐明患病个体差异的根本原因，有助于探讨许多疾病的发病机制，因而能从根本上提供疾病的预防及个性化治疗。

第一节 变异与遗传分析

基因组 DNA 是人类遗传信息的载体，人类所有的生命活动均是通过基因经由其编码产物蛋白质来执行的。任何引起基因功能变化的突变（变异或体细胞突变）或表观遗传学（epigenetics）的改变必然对机体产生影响，甚至形成疾病。疾病相关基因的搜寻与克隆，其实就是寻找这些引起基因功能变化的突变或表观遗传学改变。找到了这些导致基因功能改变的突变或表观遗传学改变，也就找到了致病基因。这其中尤以变异最为常见，绝大多数疾病均与变异密切相关。目前疾病相关基因的研究策略，不管是定位克隆和关联研究，还是动物模型等其他一些方法，最终目的是找到疾病相关的突变位点。这些方法和策略主要是基于连锁（linkage）或连锁不平衡（linkage disequilibrium）。表观遗传学的相关内容请参阅第三章。

一、DNA 变异

人类基因组计划于 1990 年 10 月启动，目标是破译人类全部遗传信息，发现人类所有基因并确定其染色体位置，以及绘制遗传图谱和物理图谱。该计划对揭示生命奥秘、加深了解疾病本质、提高医疗健康水平有极其重要的意义。2000 年 6 月人类基因组草图绘成，并于 2001 年 2 月发表了初步分析结果，完成图于 2004 年 10 月公布。这些成果主要反映了基因组基本序列，未能全面反映其变异或多态的一面。只有阐明 DNA 序列的差异以及基因组的多态性，才能真正了解与疾病特别是多基因疾病有关的遗传机制，才有可能深入准确地了解人类起源、进化和迁徙过程中的 DNA 序列变化。因此，后基因组计划的主要研究内容之一即是揭示人类基因组序列变异和多态性。

人类是一个具有多态性的群体，除了同卵双生子，没有两个人的 DNA 完全一致。不同群体和个体在对疾病的易感性、抵抗性以及其他生物学性状（如对药物的反应性等）等方面存在的差别，其遗传学基础是人类基因组 DNA 序列的变异性。迄今为止，人们通过研究所发现的疾病相关基因大部分属于不常见的单基因遗传疾病，如血色素沉着病、镫骨硬化症

等。对于这类疾病，根据其家系资料及利用当前的技术找到其致病基因相对比较容易。但是，大多数常见疾病的遗传因素则要复杂得多，并且其发病与否尚受到众多环境因素的影响。被广泛接受的“常见疾病-常见变异”(common disease－common variants)假说认为，大多数常见疾病如糖尿病、心血管疾病、精神病和肿瘤等疾病，与多个基因序列上较为常见的变异有直接关系。这些变异单独存在时也许对于患病的作用很微弱，但多个基因的变异共同起作用则大大增加了个体的疾病易感性。因此，对于人类基因组中的遗传多态的研究是阐明疾病遗传基础的前提之一。

基因组多态主要包括散在重复序列(如 Alu Ⅰ序列等)、可变数目的串联重复多态序列(variable number of tandem repeat，VNTR)、SNP 和大片段 DNA 拷贝数多态性(copy number polymorphisms，CNP)。VNTR 是第二代遗传标记，包括卫星 DNA、小卫星 DNA(minisatellite DNA)和微卫星 DNA(microsatellite DNA)。其中，微卫星 DNA 又称为短串联重复序列(simple tandem repeats，STR)，主要表现为 2～6 个核苷酸的串联重复，因此杂合度和信息度高。由于其位点在人类基因组中分布较广，且符合孟德尔遗传规律，是很好的遗传标记，在基因定位研究中运用较多(有关微卫星的详细内容本章不赘述，请参阅《基因组科学与人类疾病》和《解码生命》等专著)。第三代遗传标记是 SNP，是基因组中含量最为丰富的 DNA 变异。而第一代遗传标记是限制性片段长度多态性(restriction fragment length polymorphism，RFLP)，系 SNP 或 CNP 引起基因组序列中限制性内切酶特异酶切位点的改变，造成能切与不能切的两种情况，从而产生不同长度的片段(多态性)。CNP 是随着人类基因组序列被公布后，才被人们所熟知的一种新的多态性。

(一) 单核苷酸多态性

SNP 是指在基因组水平上由单个核苷酸的变异所引起的一种 DNA 序列多态性。这种变异可由单个碱基的转换(transition)或颠换(transversion)所引起，也可由碱基的插入或缺失所致。SNP 是人类可遗传的变异中最常见的一种，也是基因组中最为稳定的变异，占所有已知多态性的 90%以上。任何一个群体中频率不低于 1%的 SNP 位点大约有 1 000 万个，占总的 SNP 位点的 90%。SNP 最大限度地代表了不同个体之间的遗传差异，因而成为研究多基因疾病、药物遗传学及人类进化的重要遗传标记。至今在人类基因组中已有 700 多万个 SNP 位点被确认。许多科研机构和制药公司斥巨额资金用于 SNP 研究，并在发现 SNP 位点方面取得了巨大成就，同时在一些疾病的多态性研究方面亦取得了重要进步。但研究全基因组 SNP 的最大局限性仍是其高额的经费投入，且工作量极大。因此，目前需要快速而低成本的 SNP 鉴定技术或策略。

实际上，SNP 位点并不是独立遗传的，而是在染色体上成组遗传，即一组在 DNA 上位置比较接近的 SNP 位点，在一代又一代的遗传中绝少发生重组。这样的每组 SNP 位点所代表的序列形成一个单倍域(haplotype block)，单倍域内全部 SNP 的类型称之为单倍型(单倍型也就是指一条染色体上紧密连锁的多个等位基因的线性排列)。如果在某一段 DNA 片段上发现 10 个 SNP，理论上可能存在 1 024(2^{10})种单倍型，但由于连锁不平衡的存

在，实际发现的单倍型数目一般远远低于理论值。

Drysdale 等人通过构建克隆再测序方法，在 β_2 肾上腺素的启动子区 1.6 kb 内共发现了 13 个 SNP。虽然理论上单倍型有 2^{13}（即8 192）种，但在 77 人中只有 12 种单倍型。Stephens 等人分析了不同祖先的 82 位个体的 313 个基因位点，共发现了 3 899 个 SNP，每个基因平均有 12.5 个 SNP。理论上每个基因应该有 2^{12} 种单倍型，但实际上在 313 个基因中共发现 4 304 种不同的单倍型，平均每个基因只有 14 种单倍型。而且不同的单倍域，其 SNP 数、单倍型种类及单倍域的跨度亦不尽相同。Patil 等将人类 21 号染色体的 24 047 个常见 SNP（频率为 0.01 以上）划为 4 135 个单倍域，其中 14%单倍域中的 SNP 个数超过 10 个；52%单倍域中的 SNP 为 3～10 个；其余 34%单倍域中的 SNP 则少于 3 个。在4 135个单倍域中，平均每个单倍域含 2.7 种常见单倍型。最大的单倍域跨度为 115 kb，含 114 个 SNP。基因组中重组率变化非常大，单倍域内部的重组率很低或几乎不被重组。同一单倍域中的 SNP 间连锁不平衡，所有 SNP 有一同遗传的趋势。而单倍域间的重组率较大，形成重组热点区（hot spot）。Goldstein 的研究进一步证实了重组热点区域连锁不平衡相对较低，单倍型数目相对丰富；在非重组热点区域，连锁不平衡相对较高，单倍型数目相对稀少。

单倍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。如果单倍域一旦被确认，可以精确地检查特定的 SNP，以特异性区分常见单倍型。这些特定的 SNP 即为标签 SNP（图 1-1）。通过其中极少数的几个标签 SNP，就可

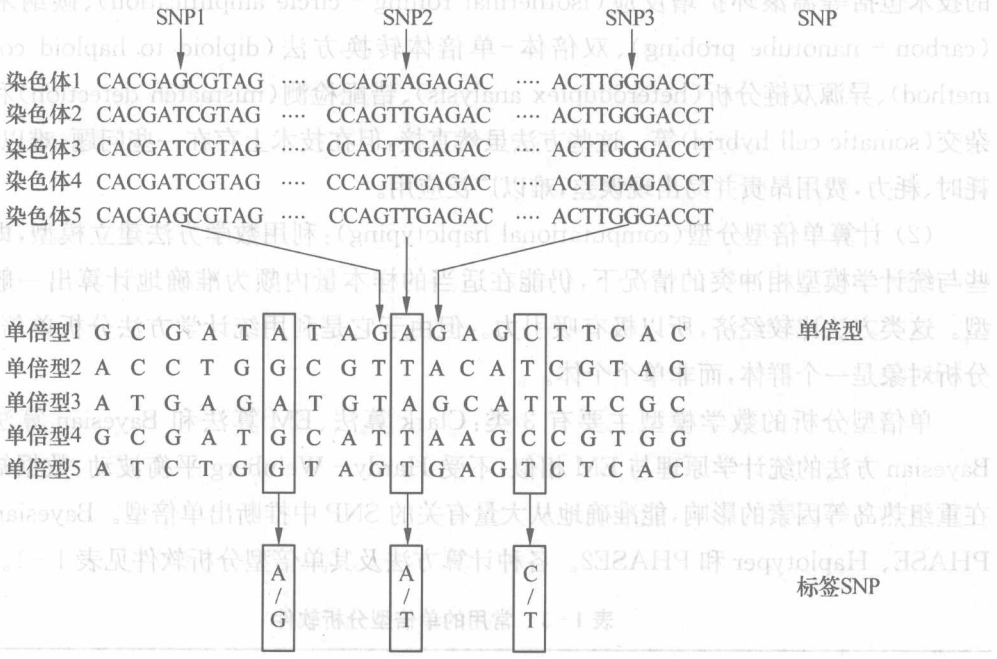


图 1-1　SNP、单倍型和标签 SNP

注：① SNP：来自不同个体的 5 条同一染色体的这一段 DNA 序列大部分完全相同，但有 3 个位点存在差异；

② 单倍型：一个单倍型由邻近的一系列 SNP 位点组成；

③ 标签 SNP：图中的 3 个 SNP 是标签 SNP（包括 SNP2），只需对这 3 个 SNP 进行分型，研究者就可以确定每个个体拥有图示的 5 个单倍型中的哪一个，而无需对这 20 个 SNP 都进行分型

以识别出不同的单倍域。因而在检测上只要对几个标签 SNP 而无须对所有的位点进行检测，便可确定一个单倍域。如果单倍域一旦被确认，还可把每个单倍域看作一个等位基因来进行连锁不平衡分析，这样用单倍域比用单个 SNP 更能精确反映基因组的多样性。这是因为用单个 SNP 分析研究连锁不平衡，产生的是杂乱的非单调的图，而用单倍型分析研究连锁不平衡，产生的则是清晰、单调、有梯度的图。

单倍域的发现，提示人类基因组中有着高度的复杂结构，绝非单纯碱基变异的任意组合。确定单倍域的存在，就可以找到相当部分的统一的单倍域结构。不同人群中单倍域的差异，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。国际单倍型图谱计划的目标，正是为了绘制人类基因组的单倍域及不同单倍域的标签 SNP。该计划包含了中国人样品(45 人)、日本人样品(45 人)、欧裔样品(90 人)和非洲裔样品(90 人)，从这几个种族所组成的 270 个样品中，推断出常见的单倍型/单倍域，为不同群体的遗传多态性研究、致病相关基因和致病因子的确定、药物基因组学、个性化治疗、人类起源进化迁徙历史的研究等提供完整的人类基因组信息和有效的研究工具。可以说，单倍体图谱(HapMap)的完成，简化了人类常见疾病相关基因分型，使基因组分型更全面、准确和精细，为致病相关基因的确定等提供了更多的信息。

单倍型的检测并非易事，主要有两种方法，它们各有其优缺点。

(1) 分子单倍型分型(molecular haplotyping)：利用实验的方法直接获得单倍型。已有的技术包括等温滚环扩增反应(isothermal rolling - circle amplification)、碳纳米管检测(carbon - nanotube probing)、双倍体-单倍体转换方法(diploid to haploid conversion method)、异源双链分析(heteroduplex analysis)、错配检测(mismatch detection)和体细胞杂交(somatic cell hybrid)等。这些方法虽然直接，但在技术上存在一些问题，难以自动化，耗时、耗力，费用昂贵并易出现误差，难以广泛应用。

(2) 计算单倍型分型(computational haplotyping)：利用数学方法建立模型，即使在一些与统计学模型相冲突的情况下，仍能在适当的样本量内颇为准确地计算出一般的单倍型。这类方法比较经济，所以极有吸引力。但由于它是利用统计学方法分析单倍型，所以分析对象是一个群体，而非单个个体。

单倍型分析的数学模型主要有 3 类：Clark 算法、EM 算法和 Bayesian 算法。其中 Bayesian 方法的统计学原理与 EM 相似，不受 Hardy - Weinberg 平衡波动、数据缺失和存在重组热岛等因素的影响，能准确地从大量有关的 SNP 中推断出单倍型。Bayesian 方法有 PHASE、Haplotyper 和 PHASE2。各种计算方法及其单倍型分析软件见表 1 - 1。

表 1 - 1　常用的单倍型分析软件

原　理	软　件	作者(年份)
Parsimony	HAPINF	Clark(1990)
Likelihood	MLHAPFRE	Excoffier and Slatkin(1995)
	PL - EM	Qin, et al(2002)
	NEMO	Gretarsdottir, et al(2003)

续 表

原 理	软 件	作者(年份)
Bayesian	HAPMCMC	Morris, et al(2003)
	HAPLOTYPER	Niu, et al(2002)
	PHASE	Stephens, et al(2001)
	PHASE2	Stephens and Donnelly(2003)

目前单倍域没有一个公认的标准,仅仅处于推断阶段。比较常用的有3种方法:整体优化法、连锁不平衡法和重组判定法(4-gamete)。整体优化法判断标准非常含糊,具体实施方法是由计算机反复搜索寻找,用少的单倍域中的单倍型代表多的样本数据的最佳解决方案。连锁不平衡法实质上是将单倍域和连锁不平衡区段等同起来,较大片段的连锁区域即为一个单倍域。连锁不平衡法计算简单,通过遗传标记两两配对相关分析,肉眼就可判断连锁不平衡的情况,网络上也有在线计算的,因此应用得较多。较精确的方法是重组判定法,通过假定重组点来界定单倍域。

(二) 大片段 DNA 拷贝数多态性

由于SNP在基因组丰富及与疾病和药物疗效等的相关性,备受关注。但在基因组中除了SNP和短片段插入(缺失)外,尚存在大片段DNA拷贝数多态性(长度约几十kb或以上)。受到净化选择(purifying selection)的作用,这些大片段DNA中的基因较少。至今已发现的受大片段DNA多态性影响的基因约有300个,其中有些基因与一些复杂性状(complex traits)或疾病相关,如内源性和外源性物质代谢相关基因(如*CYP2D6*、*UGT2B28*、*UGT2B17*、*CYP2A6*、*GSTT1*和*GSTM1*等)、常染色体隐性遗传病基因(如*STRC*、*FSHB*、*GCNT2*和*NEB*等)、嗅觉受体基因(*OR51A2*和*OR4F5*)以及肿瘤相关基因(如*DLEU7*、*TUSC3*、*BCAS1*、*HIC2HE*、*LOH12CR1*等)等。通过功能分类分析和基因组基因平均分布情况比较发现,与免疫防御、感官知觉、细胞黏附和信号传导相关的基因显得特别易于缺失,而编码核酸结合蛋白和核酸代谢相关的基因缺失却很少见。大片段DNA拷贝数多态性与SNP相似,不同群体其频率和拷贝数存在差异。大片段DNA拷贝数多态性在基因组中的分布特征及其在不同群体中的差异,目前仍知之甚少,有待于进一步研究发现。

大片段DNA拷贝数的多态性主要影响这些片段中的基因功能,引起这些基因的重复(duplication)、部分或全基因缺失,从而导致基因表达或功能改变。已知多基因疾病是由多个微效基因共同作用的结果,因等位基因拷贝数不同而引起基因表达量的改变与调控区多态性导致易感基因表达量的改变相似,并且已发现的多态性大片段DNA中的基因本身与一些复杂性状或疾病相关,即大片段DNA拷贝数多态性可能参与了疾病的发病过程。这有可能是当前一些多基因疾病虽然定位了不少易感位点,但往往难以找到易感基因的原因之一。

自人类基因组序列草图公布以来,虽然对基因组DNA序列进行了重复研究,但大多停留在对几个数据库中的人类基因组序列进行分析,还未见有对大片段DNA拷贝数的多态

性与疾病的相关性进行研究。这一方面固然是由于人类基因组序列公布不久，对大片段DNA拷贝数多态性所知有限，然而更为主要的是大片段DNA拷贝数多态性的检测目前尚缺乏十分有效的手段，主要依赖于测序和DNA芯片。尽管这两种技术目前已非常完善，但费用仍不低。利用大片段DNA拷贝数多态性搜寻疾病相关基因或群体研究，恐非一般的实验室所能承受，新的检测技术和生物信息学分析方法仍有待于发展。

最新研究显示，大片段DNA的缺失与SNP有共同的进化历史，检测与其连锁的SNP，将能有效地预测出大片段DNA的缺失。因此通过大规模的群体研究，寻找常见的大片段DNA拷贝数多态性，研究这些大片段DNA多态性在基因组的分布特征，并推断出检测这些常见大片段DNA拷贝数多态性的SNP集合，将有可能降低大片段DNA拷贝数多态性的检测费用。大片段DNA拷贝数多态性有可能是继SNP以后，基因组多态性的一个新的研究热点。

（三）体细胞突变

除了可遗传的变异影响疾病的发病外，有时组织细胞在各种因素的作用下，其DNA也会发生突变，即体细胞突变（somatic mutation），使得该突变细胞的基因组不同于人体其他正常细胞的基因组。体细胞突变最常见于肿瘤细胞。在肿瘤细胞中，除了单个碱基的突变外，通常还涉及DNA大片段的缺失或扩增。如原发性肝细胞癌，*p*53等抑癌基因常因突变或缺失失活，而β-连接素（*β-catenin*）等原癌基因常突变激活，并且在1p、1q、4q、5q、8p、8q、9p、10q、11p、11q、16q、17q、19p和22q区段均可见到杂合子缺失（loss of heterozygosity，LOH），而在1q、5q、6p、7、8q、10q、11q和20q等区段常可见到染色体区段扩增。

事实上，体细胞突变除了与肿瘤形成有关外，尚与一些非肿瘤疾病有关。如神经纤维瘤Ⅰ型和麦-奥综合征（McCune-Albright Syndrome）等一些罕见疾病（表1-2）。最近的研究发现，在动脉粥样硬化斑块处的平滑肌细胞可见到LOH、MSI（microsatellite instability）等染色体异常现象，应用微核试验（micronucleus test）和彗星检验（comet assay）在冠心病患者外周血淋巴细胞也可观测到DNA损伤情况，并与动脉粥样硬化疾病严重程度及与一些动脉粥样硬化的危险因素有关，表明动脉粥样硬化可能同时亦与DNA不稳定密切相关。因此，在肿瘤形成过程中具有重要作用的体细胞突变，可能也参与了其他一些诸如动脉粥样硬化等慢性退行性疾病的发病。

表1-2　体细胞突变与疾病

疾　病	基　因	染色体	突　变
神经纤维瘤Ⅰ型	*Neurofibromin*	17	微小和大段缺失
神经纤维瘤病Ⅱ型	*NF2*	22	无义和移码突变
麦-奥综合征	*GNAS*1	20	错义突变，Arg201X
阵发性睡眠性血红蛋白尿	*PIGA*	X	移码、无义和剪接点突变，插入/缺失