

人类基因组作图测序计划的进展

Progress in the Human Genome Project

王钦南

(国家自然科学基金委员会生命科学部, 研究员 北京 100083)

一、人类基因组作图

1. 遗传图

遗传图亦称遗传连锁图,是以“遗传标志”为位标,以经典的“遗传距离”厘摩(centimorgan, cM, 即两标志间发生 1/100 重组的可能性)为图距单位,“定点测距”而绘制的基因组图谱。

遗传图的绘制首先是寻找“遗传标志”。经典的遗传标志包括各种性状标志,如血型标志(红细胞 ABO 血型)、蛋白质和同工酶标志、HLA 免疫蛋白标志等。DNA 测试技术的进展产生了第一代 DNA 标志——限制性片段长度多态性(Restriction fragment length polymorphism, RFLP)。第一张以 RFLPs 为位标的遗传图对亨廷顿舞蹈病(Huntington's chorea)的基因定位提供了依据,它标志着经典遗传学的新生。

第二代 DNA 标志“简单串联重复序列(Simple tandem repeat, STR)即微卫星 DNA”是分布在人类基因组中以几个核苷酸为单位的重复序列(主要是 CA),它所提供的遗传多态性任何其他标志所不能比拟的,而且便于机器人操作。最近,法国 CEPH 中心的 Jean Weissenbach 小组报告,CEPH 与美国 NIH 等几个中心合作所公布的第二代“遗传连锁图”,由 6 000 多个 STR 组成,其平均分辨率(两个标志间的平均图距)达 0.7cM,相当于 700kb,已超前达到了“人类基因组”计划的原定目标。这不仅可满足单基因性状定位的需要,也为多基因性状定位奠定了基础。与此同时,模式动物小鼠的遗传图更为精细,有 7 377 个标志,定位 4 897 个基因,其中 4 200 个已被克隆,分辨率达 0.2cM。人类肥胖基因的分离,曾得益于小鼠连锁图所作出的贡献。

绘制遗传图第二方面的工作是利用“基因组扫描(genome scanning)”进行遗传连锁分析,即选择多世代、多个体的大家系,用 PCR 系统地检测分布在整个基因组上的遗传标志(微卫星 DNA),找出其

中与靶基因连锁的遗传标志,确定它们在基因组上排列关系。目前由机器人通过“基因组扫描”克隆大家系与疾病相关基因,如乳腺癌易感性基因的分离就是其中一例。

最近兰德尔(E. Lander, MIT 人类基因组中心)提出并开展了第三代遗传标志“DNA 序列多态性(Sequence polymorphism)”研究,将把现有的遗传图(STR)、物理图(STS)与正在进行的序列图整合集成,以指导并加速基因组作图、测序及基因定位的进程,使人类基因组计划进入新阶段。

2. 物理图

选用序列标志位点(Sequence-tagged site, STS)为位标,以物理长度为图距的基因组图。人类基因组计划的初步目标是建立分辨率为 100kb,即两个 STS 之间的平均距离为 100kb 的物理图。兰德尔牵头实施的 STS 图,目前已贮存 2 万多个人类基因和 1 000 个遗传多态性的 STS 标志,平均分辨率达 150kb。用于绘制此图的 PCR 流水线一次能完成 15 万份反应,使分子遗传学跨入了技术现代化的新时代。

物理图需通过构建 DNA 片段克隆的重叠群(contig)来实现。克隆载体已由质粒、 λ 噬菌体,发展到 cosmid、酵母人工染色体(Yeast artificial chromosome, YAC),插入片段长度在 200~2 000 kb 之间;细菌人工染色体(Bacterial Artificial Chromosome, BAC)插入片段长度在 100~250kb 之间;插入片段长度在 80~120kb 的 λ 噬菌体来源的新型载体 PI,以及正在发展的新载体体系 PAC(噬菌体人工染色体),还有 Fosmid 和哺乳类人工染色体 MAC 等,它们对绘制染色体的物理图将发挥重大的作用。人类 Y 染色体和 21 号染色体已完成了全部 YAC contig 构建,其它染色体载体片段的 contig 构建也接近完成。

3. 转录图

生物性状都是由具有转录功能的结构基因决定的。“转录图”将绘制出 cDNA 在基因组上的排列位置。cDNA 是在反转录酶的作用下,合成与模板

RNA 互补的 DNA 片段,相当于基因的编码序列即外显子。可以此为探针确定基因在染色体上的排列位置,有助于弄清基因组中编码序列与非编码序列的分布,成为人类基因组计划和基因组计划后研究(Post-genome Project)的接合部。

1990 年布伦纳(Brenner)等提出大规模 cDNA 测序战略,艾特尔(C. Venter)及其同事为实现这一战略建立了表达序列标志 EST 技术。至 1996 年 8 月,公共数据库中已收集了 60 万个 EST 序列(其中 45 万是人类的),占 GenBank 记录数的 65%,从而提示在现有的 EST 数据库中约有一半是人类基因。一个基因常包含多个 EST,例如血清白蛋白有 1 300 个 EST,它们各自代表不同的转录方式或不同的转录部位。分析 mRNAs 的 3' 非翻译区(3'UTRs,基因专一代表序列),可将 EST 序列转换成相应的 STS 图。Merck Index 虽然每天以数千条的速率将 EST 输入 GenBank,但已在染色体上定位的还不够多,主要是费用贵(平均定位一个 EST 为 1 000~2 000 美元)。因此,构建转录图急待克服的技术关键就是如何建立基因位点与已知 EST 之间的联系。新近发展的放射杂种细胞作图技术(Radiation Hybrid Cell Mapping),适用于人的任何单拷贝 STS,将加速基因定位的进程。

二、人类基因组测序

作为人类基因组计划中最富挑战性的目标,对全基因组 30 亿对核苷酸测序也取得了飞速进展。基因组计划实施之初,获得的最长 DNA 序列属细胞肥大病毒,约 250kb,该工作经过了数年才得以实现。而现在,对模式生物基因组的测序多个实验室都超过 1mb,甚至在 10mb 以上;对人基因组已测出长度达 500kb 以上的连续序列,仅 T 细胞受体 β 位点,长度就达 685kb。随着自动化测序水平的提高和测序技术的改进,测得数据将会迅速扩大。

现在,伦敦 Sanger 中心与 Washington 大学序列分析中心已完成了全基因组测序的前期实验,做好了测全序列的技术准备。美国 NIH 最近出资 6 000 万美元用于新的序列分析。在物理图的基础上,利用新的自动化设备,如期(2005 年 9 月 30 日)完成全序列分析是乐观的。

三、单基因遗传病基因的克隆

1. 定位克隆

人类基因组计划的进展,还反映在“定位克隆(Positional cloning)”若干遗传病基因的成功。定位克隆是在不知道疾病的发病机理,也不知道其基因结构与功能的情况下,通过遗传连锁分析所定位的

疾病基因位点的分子结构,并鉴定、分离疾病相关基因的技术体系。其策略分 4 步:(1)对确诊的家系成员进行基因型分型,(2)根据疾病基因型与遗传标志之间的连锁关系,预测基因在基因组中的可能位置,(3)对候选区段进行基因组分析,构建重叠群,建立详尽的区段物理图与转录图,(4)将得到的候选基因,进行突变检测,根据疾病相关性改变,辨认其疾病基因。定位克隆的关键是基因定位与突变筛选,其基础决定于收集遗传病家系的规模和诊断的可靠性。目前用此法已克隆 40 多种疾病基因,如:1987 年杜氏肌症(Duchenne Muscular Dystrophy,DMD)基因克隆的成功是因为发现了 X 染色体的一个小缺失;1989 年囊性纤维化(Cystic fibrosis,CF)基因被克隆是几个核心家系起了关键作用;大肠癌易感基因得以克隆,就是因几个家系最后都追踪到共同的“始祖(founder)”;1992 年克隆亨廷顿病(Huntington)基因并发现其发病机制是由于在三核苷酸序列重复动态突变(Dynamic mutation)下,通过对大家系的遗传连锁分析而成功的。此外,1994 年用连锁分析法还克隆了乳腺癌基因与小鼠肥胖基因,这预示定位克隆不仅用于克隆单基因疾病基因,还可用于鉴定分离多基因病基因。

通常致病基因是因其突变(或多态)才被发现、被分离。例如,对着色性干皮病(Xeroderma pigmentosum,XP)基因的分离,使我们了解了 DNA 修复的重要机制;对视网膜母细胞瘤(Retinoblastoma,Rb)基因的克隆不仅为该病的治疗打下基础,更重要的是发现视网膜母细胞瘤基因对细胞周期调控起关键作用。遗传病患者提供了认识、克隆人体基因功能的机会,应当受到社会的关心。

2. 功能克隆

根据基因表达的蛋白质功能分离基因的技术称之为功能克隆(functional cloning)。这种技术是克隆基因的经典、可靠而有效的方法,是克隆基因的首选策略。在已分离的 1 400 个人类基因中,大多数是通过功能克隆获得的。是先有了“定位克隆”后,才把根据蛋白质结构和功能分离基因的技术称为“功能克隆”。如珠蛋白基因的克隆,是在发现镰形细胞贫血的珠蛋白电泳迁移异常,提示其病因可能与蛋白质结构改变有关后,从中找出氨基酸组分差异,用合成的寡核苷酸片段为探针筛查血细胞的 cDNA 文库,才最终克隆到的。

3. “候选克隆”

决定表型的基因定位于某一区段后,再将这一区段的已知基因、cDNA、EST 都作为候选者,直接进行与疾病相关的突变筛选,从而鉴定出致病基因的技术体系称候选克隆(candidate cloning),即致病基因鉴定。如过氧化物歧化酶(superoxide dismutase,SOD)基因与肌萎缩侧索硬化症之间的关

系,就是这样发现的。近年来用这种方法克隆到的基因还有巨结肠症(Hirschsprung)基因等。随着人类基因组图谱的完善,基因数据库的充实以及越来越多的EST和新基因的定位,候选克隆的应用更加广泛,发挥的作用更大。由于候选克隆并不能省去基因定位,因而基因定位仍是今后需要重点发展和发挥优势的重要技术。

四、多基因遗传病基因的克隆

人类遗传病的研究,特别是致病基因的克隆,已取得很大进展。但大多数疾病并不缘自单基因遗传方式,它们涉及一个以上的基因,即“多基因病”。目前对多基因病的研究已成为生物医学的前沿。

多基因病是指在临床或流行病学方面具有一定程度的家族倾向,但又不表现为单基因遗传方式,称为“复杂性状(complex traits)”疾病。它可能是“多因微效”的数量性状,也可能是孟德尔显性不全(incomplete penetrance)或遗传异质性(genetic heterogeneity)所致。

多基因疾病对人类健康危害极大,例如,我国人口前三位的死因是脑血管病,心血管病、恶性肿瘤。每年,在北京约3万人死于心血管病,在上海发生3万例癌症。常见的多基因遗传病还有支气管哮喘、糖尿病、高血压、肥胖、骨质疏松和精神分裂症等。多基因病殃及的患者甚多。例如,在我国40岁以上的人群中,高血压的患病率达10%左右。

1. 克隆多基因病基因的可行性

对人类基因组研究已将其分割成6 000多个区段,每个区段都有“标志”。凡基因决定的疾病,均可通过“连锁分析”,找到疾病相关基因和DNA标志的连锁关系。所绘制的人类基因组遗传图和物理图,为确定复杂性状疾病基因在基因组中的位点提供了可能性。如前所述,肥胖基因的克隆就得益这些图谱的指导。此外,各种多基因遗传病大家系的收集,将是克隆多基因病基因的源泉。

2. 几种多基因病基因的研究进展

(1) 肿瘤易感性基因

肿瘤发生涉及遗传与环境,多数具有明显的家庭聚集性,很少为典型的孟德尔遗传。最早定位的肿瘤易感基因是家族性非息肉型结肠癌(HNPCC)基因,通过对几十个有关大家系的连锁分析,把HNPCC的易感基因定位于2号染色体短臂15~16区,进一步研究表明,这一基因牵扯DNA碱基错配修复。另一个相似的HNPCC基因则被定位于3号染色体短臂21~23区。乳腺癌易感基因(brca1),也是通过大家系连锁分析,定位于17号染色体长臂21区后,以定位克隆分离得到的。

(2) 支气管哮喘相关基因

目前认为,哮喘病的成因是基因缺陷和环境因素影响各半。据统计,英国有1/7的儿童患哮喘。作为一种多基因遗传病,哮喘的发生可能有一个或几个主基因参与。利用哮喘大家系和隔离群基因连锁分析,找到主基因的位点所在。免疫球蛋白E受体基因被认为与哮喘有关,大量的连锁分析资料表明哮喘的主基因可能位于11号染色体长臂13区和6号染色体短臂。由于多家公司投资,克隆哮喘主基因的研究有较大进展。美国Sequona公司利用南大西洋中Loneliest岛上的一个哮喘大家系(301口,全是近亲婚配的表兄弟姐妹的隔离群,其中1/3患哮喘)和中国东海某岛上的哮喘大家系,把一个哮喘相关基因定位在300kb的YAC插入片段上,预期将有突破性进展。

(3) 糖尿病相关基因

胰岛素依赖型糖尿病(IDDM)易感基因的定位是对人类基因组进行整体分析的尝试。首先用290个遗传标志,对96个英国人家系进行连锁分析,将所发现的18个可能相关位点,进一步分析,确定5个首选位点。虽然分离糖尿病相关基因尚在进行中,但已显示出多基因病研究策略已目臻成熟。

(4) 高血压相关基因

从个体间的症状差异得知,高血压发生有遗传因子参与。到目前为止,已发现3个突变基因参与人类高血压病发生:①醛固酮合成酶位点的突变,导致产物的异位表达,引起GRA的高血压病(glucocorticoid remediable aldosterone),②氨氯吡嗪(Amloride)敏感性上皮钠通道的 β -亚基突变导致该通道的组成型激活,产生liddle综合症,③高加索人群原发性高血压,与血管紧张素原基因位点的变异有关。高血压作为一种多因素遗传病,对其致病的遗传因子尚在研究中。

(5) 肥胖相关基因

肥胖(obesity, ob)是遗传、生理和心理等因素综合作用的结果。通过小鼠肥胖模型的连锁分析,发现至少有5个位点与肥胖有关。使用新的连锁图及连锁标志,将其中1个位点的范围缩小到650kb,再以定位克隆分离到这一基因,然后以小鼠的ob为探针从人类基因组中钓出人的ob基因。最近有报道认为该基因与人的肥胖关系不大。

五、几个值得注意的问题

1. 大家系

疾病基因克隆是本领域的焦点。迄今报道的重大疾病基因的克隆,都与掌握大家系资料有关。多基因疾病的克隆仍沿用单基因病基因的克隆策略,故而隔离群与大家系就显得更加重要。例如芬兰多个大家系均可追踪到同一个祖先,目前正用于对12

种疾病基因的克隆。

我国人口众多,遗传资源丰富,但由于投入少,对大家系疾病基因尚未很好地开展研究;国外某些大公司已进入我国收集遗传病家系的资料和 DNA 样品,致使我国的人类遗传资源(隔离群和大家系)尚未开发就已外流。我国有关部门应像保护自然资源和珍稀濒危生物一样采取必要的措施。

2. 基因专利问题

基因专利是基因组研究的知识产权问题所在。尽管人类基因组组织(HUGO)创立之初便率领全球科学家抵制基因专利化提案,引发了持续至今的基因专利,特别是 EST 专利之争。美国 NIH 率先资助序列分析,要求被资助者将得到的新序列当天送上网络,使其公开化。而 Incyte 及 SB 等公司却呈送了 50 万条 EST 的专利申请。

1992 年初,NIH 提出了 2 750 条 EST 的专利申请,当即遭到全球科学家的反对。当时,中国科学家也支持 HUGO 反对 cDNA 专利的声明。美国专利局驳回了 NIH 的申请。随后 SB 公司聘用 NIH 的该项目主要负责人 C. Venter,成立了 HGS 与 TIGR(The Institute Genome Research),投资 1.2 亿美元,用先进技术分离 EST,并建立专用数据库,使用数据库的条件是“接受监督、洽谈转让”。如接受这一条件,所分离的人类基因都将被打上 SB 的标签。

Merck 公司以“非盈利”方式赞助建立新的 cDNA 数据库,并聘请 HUGO 前任主席卡斯基(T. Caskey)负责以更大的规模生产 EST,这一势态打破了垄断的格局。为此,SB 公司波斯特(G. Poste)撰文(《Nature》,1995. 12)指责 Merck 公司公开序列损害了其它公司的利益,还对 HUGO 的反对专利

声明进行全面驳斥。

根据人类基因组计划发展的趋势,《Nature》发表一项调查说,从 1981 年至 1995 年,美国和欧洲专利机构已批准人类基因专利 1 175 项。在拥有人类基因专利最多的 10 家公司中,日本武田公司位居榜首,为 63 项;美国基因技术公司居第二,为 41 项。虽然很多科学家不主张过多批准人类基因专利,但从目前形势看,人类功能基因的专利化基本上已成定局。因此我国必须加大克隆人类功能基因的投入力度,得到更多的人类新基因,争取对基因专利化问题拥有更多的主动权,并联合各国科学家遏制 EST 专利化的企图。

3. 伦理、法律和社会问题

人类基因组研究的另一个中心议题是伦理学问题。个人基因组信息属“个人隐私”,是否应归保险公司“健康档案”掌管?对基因信息的隐私权以及在生育过程中能否以社会、专业和法律影响双亲的选择等,目前难于以法律形式固定下来。联合国教科文组织和世界卫生组织正在做这方面的准备,HUGO 的 ELSI(Ethical Legal Social Implications)委员会也有类似的报告,拟专门讨论这些问题。

4. 人类基因组多样性问题

HUGO 公布了“人类基因组多样性计划”。我国是亚太地区实施这一计划的重要国家。“多样性”是我国资源的优势所在,是与国际接轨的桥梁,也是我国人类基因组计划的重要内容。现急待统筹安排,组织合作,统一群体和家系标准,制定相应的调查程序,规范取材方式,确立研究方法。

(责任编辑 蔡德诚)

(上接第 23 页)

了评价,而不必再进行鉴定;发明专利的实质性审查,事实上也是一种应用技术成果的评价,取得发明专利的成果,也没有必要另行鉴定。社会科技评价体系的这种多样性,为科技成果完成人、完成单位和有关当事人选择科技成果鉴定所采用的适当方式,创造了条件。**选择何种鉴定与评价方式,完全是科技成果完成人、完成单位和有关当事人自己的权利,因此,新的科技成果鉴定制度必须奉行当事人自愿的原则。**

科技成果鉴定的当事人自愿原则,包括以下三层含义和规则:一是是否提交科技成果鉴定机构鉴定与评议由当事人自主决定,没有经过鉴定的科技成果照样可申请科技奖励。因为各级各类科技奖励的评审过程,本身就是对科技成果的评价与鉴定过程。二是何时提交鉴定与评议由当事人自主决定,履行技术合同完成的成果,合同有约定的,按约定的时间提交鉴定;合同没有约定,当事人双方愿意

鉴定的,按双方商定的时间提交鉴定。执行研究开发计划完成的成果,计划任务书或计划管理办法要求鉴定的,按其要求的时间提交鉴定。三是委托哪个鉴定机构进行鉴定,也由当事人自主决定。委托人不论是成果完成单位、成果完成人、技术合同的当事人双方,还是受理科技成果争议案件的法院或仲裁机构,都是作为技术咨询合同的委托方,鉴定机构则作为具有特定资格的顾问方,双方在科技成果鉴定过程中,各自承担相应的义务,享有相应的权利,受技术合同法的保护,同时又要受科技成果鉴定规范的制约。鉴定机构制定和承诺执行的鉴定规范,应视为技术咨询合同的组成部分,当事人不得违反。

按照上述办法,并实行政府主管机关和社会公众的监督,有利于提高鉴定的公正性;而实行当事人自愿选择鉴定机构的办法,则有利于鉴定机构提高工作效率和鉴定质量。 (责任编辑 孙立明)