复杂疾病遗传学研究方法

严卫丽　主编

科学出版社
www.sciencep.com

复杂疾病遗传学研究方法

[redacted]

复杂疾病遗传学研究方法

主　编　严卫丽（博士、教授、硕士生导师，新疆医科大学）
参　编　陈　艳（博士、教授、硕士生导师，新疆医科大学）
葛东亮（博士、留美博士后，美国 Duke 大学）

科学出版社
北京

内 容 简 介

本书是基于21世纪世界生物医学领域发生的巨大进步和编者的研究积累而编写的。人类基因组计划的完成、单体型计划的完成和全基因组关联不断发现多种人类复杂疾病的致病基因,人类对于疾病的认识进入了一个崭新的阶段。本书第一至第六章,主要介绍了复杂疾病关联研究的设计、资料收集、生物信息数据库查询、基因分型技术和资料入机等基本知识;第七章至第十二章,介绍了包括全基因组关联研究在内的复杂疾病关联研究方法学进展,内容包括复杂疾病关联研究存在的问题、人类单体型计划与其对复杂疾病研究的意义,系统介绍了全基因组关联研究的最新进展,如设计原理、遗传标记选择、统计分析原理、多重比较以及重复问题。

图书在版编目(CIP)数据

复杂疾病遗传学研究方法 / 严卫丽主编. —北京:科学出版社,2009
ISBN 978-7-03-026361-2

Ⅰ.①复… Ⅱ.①严… Ⅲ.①医学遗传学-研究方法-医学院校-教材 Ⅳ.①R394-3

中国版本图书馆CIP数据核字(2010)第005472号

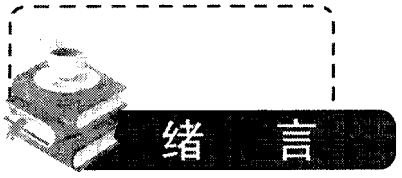
策划编辑:李国红 / 责任编辑:许贵强　李国红 / 责任校对:赵桂芬
责任印制:刘士平 / 封面设计:黄　超

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行　各地新华书店经销
*
2009年12月第 一 版　开本:787×1092　1/16
2009年12月第一次印刷　印张:5　1/4
印数:1—1 500　字数:131 000
定价:24.00元
(如有印装质量问题,我社负责调换)

目　录

绪言 …… (1)
一、复杂疾病关联研究概述 …… (1)
二、GWA 研究设计相关问题 …… (1)
三、GWA 的遗传标记 …… (3)
四、基因组拷贝数变异 …… (5)
第一章　复杂疾病遗传资源收集 …… (9)
一、概述 …… (9)
二、遗传资源收集 …… (10)
第二章　生物信息数据库的查询 …… (13)
第三章　应用 Oligo 软件进行 PCR 引物设计 …… (21)
一、引物设计的原则 …… (21)
二、常用的参数 …… (21)
三、引物设计的具体原则 …… (21)
四、应用 Oligo 6.0 进行引物设计的步骤 …… (22)
第四章　SNP 检测方法概述 …… (25)
一、聚合酶链式反应-限制性片段长度多态性(PCR-RFLP) …… (26)
二、聚合酶链式反应-单链 DNA 构象多态性分析法(PCR-SSCP) …… (28)
三、Taqman 法 …… (29)
四、变性高压液相色谱法(DHPLC) …… (30)
五、测序 …… (30)
六、DNA 芯片法 …… (32)
第五章　聚合酶链式反应(PCR)技术总结 …… (33)
一、PCR 技术的基本原理 …… (33)
二、PCR 反应体系与反应条件 …… (33)
三、PCR 反应特点 …… (35)
四、PCR 扩增产物的分析 …… (36)
五、PCR 反应过程中污染的控制及可能出现的问题 …… (36)
第六章　调查资料的入机 …… (38)
一、利用 Visual FoxPro 6.0 录入数据 …… (38)
二、利用 Excel 录入数据 …… (39)
三、利用 Access 窗体功能录入数据 …… (40)
第七章　复杂疾病关联研究中的若干问题 …… (41)
一、关联研究中的混杂 …… (41)
二、候选基因的确定 …… (42)
三、中间表型的应用 …… (42)

四、SNP 的选择 …… (43)
五、单体型分析 …… (44)
六、关联研究结果的判定 …… (44)
第八章 传统关联研究遗传统计分析 …… (47)
一、Hardy-Weinberg 平衡及检验 …… (47)
二、位点间连锁不平衡强度分析 …… (49)
三、单体型分析 …… (49)
第九章 单体型分析:复杂疾病基因定位的新希望 …… (50)
一、EM 算法 …… (50)
二、Clark's 算法 …… (51)
三、Phase …… (51)
四、混合 DNA 样本(DNA Pool)的单体型分析 …… (52)
第十章 连锁分析的方法学概述 …… (55)
一、连锁分析的基本原理 …… (55)
二、家系收集 …… (60)
三、实验室技术 …… (60)
第十一章 全基因组关联研究遗传统计分析 …… (65)
一、GWA 研究统计分析原理 …… (65)
二、GWA 研究多重假设检验调整 …… (67)
三、人群混杂 …… (68)
四、GWA 研究的重复 …… (70)
第十二章 SNP 检测技术进展与全基因组关联研究 …… (74)



一、复杂疾病关联研究概述

人类对疾病或者疾病性状的认识最早可以追溯到1956年对ABO血型遗传基础的发现(Clarke C. ABO blood groups and secretor character in duodenal ulcer. *BMJ*, 1956)。在过去半个世纪,人类对复杂疾病遗传基础的认识有了很大提高。20世纪80年代,运用第一代遗传标记RFLP研究主要限于基于候选基因策略的单个位点研究,其局限性在于无法发现未知基因与疾病的关联。90年代第二代遗传标记微卫星(microsatellite)的出现,以家系为基础的单基因遗传病致病基因连锁定位研究取得了显著的成果,大量单基因病的具有主基因效应的致病基因得到确认。这些基因在人群中频率非常低,对于复杂性状而言,这些罕见基因突变所能解释的变异非常少,人群归因危险度分数(population attributable fraction, PAF)常低于10%,对于复杂性状疾病的致病基因的发现效果并不理想。随着人类全基因组计划的完成,第三代遗传标记单核苷酸多态SNP被发现是人类基因组内分布最为广泛的基因序列变异,并被迅速、广泛应用到复杂疾病和性状的研究中,发现了许多与复杂疾病和性状关联的SNPs。这一阶段的研究也存在一些重要问题(见:复杂疾病关联研究若干问题. 遗传学报,2004.),比如假阳性较高、研究结果很难重复等。随着国际人类基因组计划测序完成和基于人类基因组分布最为广泛的序列变异——单核苷酸多态性(single nucleotide polymorphism,SNP)的单体型图谱构建完成,人类遗传学研究最近正在进入一个新的时期。与此同时,经济高效的高通量基因分型技术得到了迅猛发展,一个反应可以同时检测成百上千个SNPs。所有的这些进步,是一种系统的,甚至是"未知的"。在全基因组范围筛检与疾病关联的SNP方法成为可能,这就是全基因组关联研究(genome-wide association study,或者whole-genome association study,GWA)。与以往的候选基因策略明显不同在于,我们不再需要在研究之前构建任何假设。全基因组关联研究的实现将把我们对复杂疾病的病因的认识推动到一个全新的阶段。

二、GWA研究设计相关问题

(一)表型的选择

确定研究的表型是研究设计中的首要问题。研究表型的选择应当尽量基于以下三个原则[9]:

原则一 选择遗传度较高的疾病或表型。

疾病的遗传度(heritibility, h^2)表示疾病(或表型)在多大程度上遗传因素的影响。低

遗传度的疾病会降低遗传学关联研究的把握度(power)[10]。

原则二 表型(trait)优于疾病的原则。

疾病的状态有时很难测量,或者模糊不清,有时则多种疾病混在一起而难以判断。例如,2型糖尿病是一种诊断相对比较明确的疾病,但是有很多表面上健康的人患了2型糖尿病却不知晓[11];有人认为心理疾病的诊断不够精确,比如精神分裂症(schizophrenia)、双相情感性精神障碍(biopolar disorder)、孤独症等,但是这些疾病的诊断可以发现致病的遗传因素[12]。又如,脑卒中有不同的发病机制(比如,心脏或者主动脉栓子脱落,或者脑出血),但是临床上却常常很难区分。基于以上原因,研究疾病相关数量表型有时要优于研究疾病状态。

原则三 选择测量简单、准确和遗传度高的数量表型。

尽可能选择那些反映疾病危险的数量表型(比如BMI是糖尿病和其他许多疾病的危险因素)、有助于区分疾病临床亚型的表型(如胰岛素释放和胰岛素敏感性),或者那些用来诊断疾病的表型(如空腹血糖)。数量表型测量的难易程度直接和该表型的遗传度相关,因为降低了测量误差(比如通过重复测量),降低了噪声和总体变异,理论上就增加了该数量表型可以由遗传因素解释的变异的比例大小。例如,单次测量收缩压的遗传度为0.42,然而,多次测量的连续观察的(longitudinal)收缩压的遗传度可以达到0.57[13]。

(二) GWA研究设计类型

GWA设计的基本原理同经典的病例对照研究,即假设某个SNP与疾病发生关联,则理论上病例(患有该疾病)中该SNP的等位基因频率应当高于对照组(未患有该疾病)。然后通过假设检验来检验该假设。

考虑到研究成本、基因分型成本以及研究把握度(power)等方面的因素,GWA的研究设计目前分为单个阶段研究(one-stage design)和两阶段研究(two-stage design)或多阶段研究(multiple-stage design)设计。

One-stage design:即选择了足够的病例和对照样本后,一次性在所有研究对象中对所有选中的SNP进行基因分型。然后分析每个SNP与疾病的关联,分别计算关联强度和OR值。显而易见,该设计的最大缺陷在于基因分型耗资巨大。为节约基因分型的数量和成本,两阶段研究正在被更多研究者所采用。

Two-/multiple-stage design:第一阶段,在小样本中对全基因组范围选择的所有SNP进行基因分型,统计分析后筛选出较少数量的阳性SNPs;第二阶段,在更大的样本中对那些在第一阶段得到阳性结果的SNP进行基因分型,然后结合两个阶段的结果进行分析。第一阶段的基因分型可以是以个体为单位(individual genotyping),也可以采用DNA pool的办法,后者可大大降低基因分型的工作量。有多项研究证明,在GWA研究中,采用DNA pool结合Affymatrix微阵列试剂盒可以低成本、高效益进行SNP筛选[14,15]。理论上,如果对DNA pool的等位基因频率估计误差为零的话,我们可以认为这种方法与对所有个体进行基因分型的方法没有差别。然而事实上差异还是存在的,研究表明,DNA pool估计的等位基因频率其标准差在1%~4%的范围[16~19]。如果单独DNA pool的方法来估计等位基因频率,那么其1%~4%的标准差对全基因组的病例对照研究的把握度影响是不可忽视的[20,21]。但是如果第一阶段采用DNA pool的方法,而第二阶段采用个体基因分型的方法,这种结合则不失为一种理想的既经济、又有较高把握度的研究策略[22~24]。

事实上,两阶段的研究策略实际应用的时候仍然有一系列问题值得探讨,比如如何保证

第一阶段研究以最大的可能筛选出与疾病或者表型关联的 SNPs，然后在第二阶段里对它们进行基因分型；在统计分析中控制了假阳性率后如何同时保证研究的把握度；如果我们希望在两阶段研究完成后将与疾病或性状相关的遗传标记（SNP）筛选出来，那么一项研究至少筛选出一个疾病或性状相关位点的概率有多大？Zuo[25]等人通过理论数据推算研究发现，即使第一阶段 DNA pool 对等位基因频率的估计存在一定的误差，当病例和对照组间等位基因频率的差异不小于 0.05，以及样本量足够大时，至少一个与疾病相关的位点被发现的概率足够高（>0.90）。尽管研究表明，设计良好的两阶段的研究可以大大减少基因分型的工作量[26~28]，但是这样做是否降低了研究的把握度，研究者们针对这个问题进行了深入的探讨。关于两阶段研究（独立样本）后数据分析，Skol[29]等人比较了两种方案的把握度，即联合分析（joint analysis，将第一阶段和第二阶段的统计量合并分析）和基于重复原理的分析（replication-based analysis，基于第一阶段的结果，对进入第二阶段的 SNP 和样本单独进行分析来看是否能重复第一阶段的结果），发现在不同等位基因频率、不同第一、二阶段样本的分配比例情况下，联合分析都获得了较高的把握度。见图 1[29]。

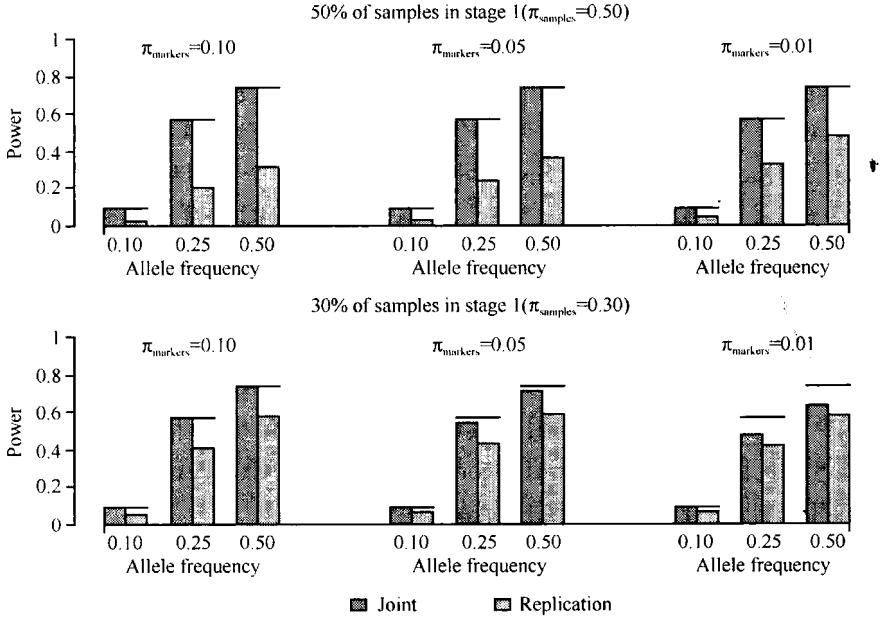


图 1 联合分析和两阶段独立分析在不同等位基因频率、显著性水平下的把握[29]

Wang[30]等人通过数学方法探讨了一个最优阶段设计（optimal two-stage design）所涉及的一系列具体问题。如第一阶段阳性 SNP 筛选的 α 水准的确定、达到不同把握度所需的第一阶段和第二阶段所需的最小样本量和分配比例，以及预计的实验花费。对于两阶段研究设计中的具体问题还在继续热烈讨论当中[31]，比如，第二阶段研究的样本是与第一阶段完全不同的样本好，还是包括第一阶段的样本好？把握度如何？假阳性率如何？这些问题的阐明还有赖于更多的研究和数据。

三、GWA 的遗传标记

GWA 研究的特点是不再选择候选基因或者候选染色体区域，而是针对基因组所有的

SNPs。人类基因组计划的完成为我们提供了人类基因组内广泛存在的浩瀚的单核苷酸多态(SNP)的信息(大约 700 万)。2003 年由国际单体型图谱计划协作组(the International HapMap Consortium)发起的国际人类单体型计划(HapMap)计划第一阶段的研究提供了人类 4 个种族共 269 个个体基因组超过 100 万个 SNP(大约每 3kb 一个)以及相关之间 LD 关系的图谱,第二阶段增加了 SNP 的密度(大约每 1kb 一个),为 GWA 提供了强有力的工具[32]。基于 HapMap 数据库平台,研究者可以因此筛选多达 250 000～500 000 个常见 SNPs 用于全基因组的关联研究,也可以选择功能 SNP 进行研究,也可以随机选择 SNPs 进行研究,所选 SNPs 可以覆盖全基因组 70%～80%的常见 SNPs,而通过 Tag SNP 筛选工具 Hapview 可以筛选覆盖、代表更大范围基因组序列变异的 TagSNPs。Bakker[33]和 Zeggini[34]率先对 HapMap 作为工具在 GWA 中的重要作用进行了评估。虽然 HapMap 主要包含常见 SNP(common SNP)的信息而较少少见 SNP(rare SNP) 的信息,早期的 GWA 研究确实成功地发现了许多常见 SNPs 在疾病发生中的作用。理论上讲,研究者可以根据研究的经费、所要达到的把握度(power)和已经掌握的某疾病的遗传背景知识,根据具体的研究需要从 HapMap 中选择一套 SNPs 在病例和对照组中分别进行基因分型。而实际情况是,只有少数的研究组织和机构能够做到根据研究实际需要来选择覆盖全基因组的 SNPs。大多数研究者只能被动选择有限的几个大公司提供的事先设计好的商品化检测试剂盒,这些试剂盒包含的 SNPs 是事先根据特殊的数学算法、能够达到预期把握度而选择的,如 Bakker[33]等人所采用的试剂盒包含的 SNPs 可以保证在 2000 个病例和 2000 个对照中,在全基因组范围以 $p<0.05$、单个位点 $p<10^{-7}$、达到 95%的把握度发现导致疾病发生的 SNP (causal SNP)。如果希望 HapMap 能够代表更多范围的低频率 SNPs,就需要继续增加更多基因组内低频率 SNPs 的基因型信息,而只有大量的测序工作才能实现这个目标,这无疑意味着更大的测序工作量。

与 HapMap 数据库主要涵盖常见 SNPs 不同,SeattleSNP 网站数据库则采用 DNA 测序方法,提供了与炎症反应相关基因及其侧翼序列的所有 SNP 信息,可供研究者们选择 tagger SNP 使用(包括两个人种:Europeon Amercian 和 Africa American 白人和黑人的信息)。网站同时提供了 Tagger SNP 的选择工具 perl 软件。

目前常用的市售的用于 GWA 的试剂盒有以下几种:HumanHap 300、HumanHap 550 Array Sets (Illumina Infinium 系列)、Mapping 100 K 和 Mapping 500 K Array Sets (Affymetrix GeneChip 系列),这些产品系列分别按不同目的设计的,比如不包含 nonsynonymous SNP (nsSNP)的试剂盒、基于 LD 关系选择的 Tag SNPs 以及全基因组随机选择的 SNPs 等。研究者们比较了它们对不同人种全基因 SNP 的覆盖程度(coverage)[35]。具体方法是:采用 HapMap phase Ⅱ 提供的 SNPs 信息(release 2-常见 SNPs 和 release 16c.1-少见 SNPs),采用 Haploview 软件[36],以 paired r^2 大于某一个界值选择 Tag SNPs,比较了 Illumina HumanHap300, Affymetrix 500k, Affymetrix 111k, Affymetrix 500k,+175k tag 和 Illumina Human-1 产品对 HapMap 所采用的四个人种 CEU, JPT+CHB,和 YRI 全基因组内常见 SNPs 和少见 SNPs 的覆盖程度。结合其他几项类似的研究结果后发现,第一代的商品化试剂盒对除非洲裔人种以外的其他人种的全基因组覆盖程度相近,对基因组内常见 SNP 的覆盖程度达到令人满意的效果[35,37,38]。最近的一项研究还比较了 HapMap 550 与其他产品,除了肯定了前几项研究的结论外,发现按照 HapMap CEPT 设计的试剂盒对高加索人以及其他欧洲人种的基因组 SNP 覆盖程度都相对较好[39]。

以下是相关的网站地址：

HapMap：http://www.HapMap.org

SeattleSNP：http://pge.mbt.washington.edu

Tagger：http://www.broad.mit.edu/mpg/tagger/

Wellcome trust Case Control Consortium：http://www.wtccc.org.uk/

Illumina：http://www.illumina.com

Affymetrix：http://www.affymetrix.com

四、基因组拷贝数变异

在研究遗传变异和环境因素交互作用在导致单基因疾病以及多基因疾病发生中的作用过程中，拷贝数变异（copy number variation，CNV）在基因组的广泛存在以及其作用在最近才开始被人们所认识[40]。

基于 HapMap 计划的研究样本，人类第一代基因组 CNV 图谱构建完成，基因组内存在大约 1500 个 CN 区域，覆盖大约 12%（大约 360Mb）的人类基因组范围，比任何一个最大的染色体包含的遗传信息还要多。这个发现，打开了人类基因组研究的新篇章。

CNVs 包括基因组内从 1kb 到几个 Mb 不等的序列缺失、插入和重复，不包括那些可遗传的片段里的插入和缺失。CNV 不仅存在人类基因组，也存在与其他物种的基因组，如小鼠和黑猩猩。CNVs 的大小和普遍性提示了它在决定人类复杂疾病、多基因疾病如心血管疾病的遗传易感性中的重要作用。原因十分简单：某个 CNV 所在的位置和范围可以涵盖许多基因。然而要想验证这个研究假设，需要大样本的研究和良好的表型资料，以及综合的、多种方法来根据 CNV 状态将研究对象正确分组，这在目前并不是十分容易的事。除了基于 SNP 方法外，Genome TilePath（WGTP）也发现了数量可观的 CNVs，大大补充了前一种方法的发现[41]。

CNVs 导致疾病的机制可能分为通过数量作用和质量作用两种机制[41]。CNV 在导致疾病发生中所起的作用可能与突变的作用相似。其生物学意义在于，当讨论疾病发生的遗传基础时，不能无视 CNV 的作用：它的存在与否、是否存在 CNVs 与其他序列变异的交互作用。目前关于 CNV 的认识仍然有限，包括：①OMIM 数据库里 14.5% 的基因与 CNVs 有重合；②CNVs 参与决定人类多样性；③一些 CNVs 参与决定某些疾病的易感性，如低拷贝数 CCL3L1、FCGR3B 和 DEFB4 基因与高 AIDs、免疫介导的肾小球肾炎和克罗恩病（节段性肠炎）易感性增高有关[42~44]；④CNVs 可能影响所在基因的基因表达水平。有关基因组内 CNVs 的信息可在以下链接查询：Database of Genomic Variants：http://projects.tcag.ca/variation/。

总之，全基因组关联研究时代已经到来。在全基因组关联的第一次浪潮中，为我们提示了很多未知基因的作用，用候选基因策略这些基因的作用几乎没有可能被发现。两阶段的研究设计，商品化全基因组 SNP 试剂盒价格不断下降逐渐为多数研究组所接受，在遗传统计分析方面在降低研究假阳性率、保证研究的把握度等方面的深入探讨和发展，为全基因组研究的第二次浪潮到来打下了良好的基础。

关于全基因组研究的第二次浪潮，我们需要回答研究什么（what to do）、如何研究（how to do）和现实状况（current fact）的三个问题：

第一，GWA 研究将致力于发现更多与疾病关联的基因变异，阐明变异-基因-环境因素之间的交互作用关系；

第二，数据共享是加快研究步伐的最佳途径。生命科学领域的顶尖科学家们在这个问题上已经达成了共识。研究结果需要通过在不同种族、人群中加以重复才能避免虚假的关联[45]。要实现数据共享，具体实施起来还存在一些问题，比如知情同意的问题，以及不同来源的表型数据在临床测量和数据收集的方法方面的差异可能导致的假阳性问题等。

第三，尽管有很大的难度，数据共享只能是 GWA 取得发现基因与疾病之间真正关联的必由之路。如果没有大型研究之间的合作，许多导致复杂疾病的遗传变异就不可能被发现。比如，在三个研究团队的合作下与 2 型糖尿病关联的遗传标记才得以被发现[4]。

我们正处在人类遗传学发展的关键时刻，让我们每个相关科研人员都做好准备，迎接这个时代的到来，并为这个时代做出自己的贡献。

（严卫丽）

参 考 文 献

[1] Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF. A common genetic variant is associated with adult and childhood obesity. *Science*, 2006,312:279～283

[2] Rosskopf D, Bornhorst A, Rimmbach C, Schwahn C, Kayser A, Kruger A, Tessmann G, Geissler I, Kroemer HK, Volzke H. Comment on "a common genetic variant is associated with adult and childhood obesity". *Science*, 2007, 315:187; author reply 187

[3] Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 2007,316:889～894

[4] Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 2007,316:1331～1336

[5] Ubeda M, Rukstalis JM, Habener JF. Inhibition of cyclin-dependent kinase 5 activity protects pancreatic beta cells from glucotoxicity. *J Biol Chem*, 2006,281:28858～28864

[6] Foley AC, Mercola M. Heart induction by wnt antagonists depends on the homeodomain transcription factor hex. *Genes Dev*, 2005,19:387～396

[7] Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF,

Bergman RN, Tuomilehto J, Collins FS, Boehnke M. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 2007,316:1341～1345

[8] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, Konig IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. Genomewide association analysis of coronary artery disease. *N Engl J Med*, 2007,357:443～453

[9] Newton-Cheh C, Hirschhorn JN. Genetic association studies of complex traits: Design and analysis issues. *Mutat Res*, 2005,573:54～69

[10] Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet*, 2000,66:1616～1630

[11] Harris MI, Klein R, Welborn TA, Knuiman MW. Onset of niddm occurs at least 4-7 yr before clinical diagnosis. *Diabetes Care*, 1992,15:815～819

[12] Risch N. Linkage strategies for genetically complex traits. Ii. The power of affected relative pairs. *Am J Hum Genet*,1990,46:229～241

[13] Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension*, 2000,36:477～483

[14] Docherty SJ, Butcher LM, Schalkwyk LC, Plomin R. Applicability of DNA pools on 500 k snp microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics*, 2007,8:214

[15] Meaburn E, Butcher LM, Schalkwyk LC, Plomin R. Genotyping pooled DNA using 100k snp microarrays: A step towards genomewide association scans. *Nucleic Acids Res*, 2006,34:e27

[16] Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci U S A*, 2001,98:581～584

[17] Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. In silico mapping of complex disease-related traits in mice. *Science*, 2001,292:1915～1918

[18] Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, Porteous DJ, Evans KL. Snp genotyping on pooled dnas: Comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res*, 2002,30:e74

[19] Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: A tool for large-scale association studies. *Nat Rev Genet*, 2002,3:862～871

[20] Zou G, Zhao H. The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet Epidemiol*, 2004,26:1～10

[21] Zou G, Zhao H. Family-based association tests for different family structures using pooled DNA. *Ann Hum Genet*, 2005,69:429～442

[22] Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet*, 1997,61:734～747

[23] Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. Association testing by DNA pooling: An effective initial screen. *Proc Natl Acad Sci U S A*, 2002,99:16871～16874

[24] Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet*, 2002,66:393～405

[25] Zuo Y, Zou G, Zhao H. Two-stage designs in case-control association analysis. *Genetics*, 2006;173:1747～1760

[26] Satagopan JM, Venkatraman ES, Begg CB. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, 2004,60:589～597

[27] Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association

studies. *Biometrics*, 2002,58:163～170

[28] Thomas D, Xie R, Gebregziabher M. Two-stage sampling designs for gene association studies. *Genet Epidemiol*, 2004,27:401～414

[29] Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, 2006,38:209～213

[30] Wang H, Thomas DC, Pe'er I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol*, 2006,30:356～368

[31] Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol*, 2007

[32] Kruglyak L. Power tools for human genetics. *Nat Genet*, 2005,37:1299～1300

[33] de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet*, 2005,37:1217～1223

[34] Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI. An evaluation of hapmap sample size and tagging snp performance in large-scale empirical and simulated data sets. *Nat Genet*, 2005,37:1320～1322

[35] Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet*, 2006,38:659～662

[36] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of ld and haplotype maps. *Bioinformatics*, 2005,21:263～265

[37] Nicolae DL, Wen X, Voight BF, Cox NJ. Coverage and characteristics of the affymetrix genechip human mapping 100k snp set. *PLoS Genet*, 2006,2:e67

[38] Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet*, 2006,38:663～667

[39] Magi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M. Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics*, 2007,8:159

[40] Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L, Aburatani H, Jones K, Redon R, Hurles M, Armengol L, Estivill X, Mural RJ, Lee C, Scherer SW, Feuk L. Genome assembly comparison identifies structural variants in the human genome. *Nat Genet*, 2006,38:1413～1418

[41] Pollex RL, Hegele RA. Genomic copy number variation and its potential role in lipoprotein and metabolic phenotypes. *Curr Opin Lipidol*, 2007,18:174～180

[42] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science*, 2005,307:1434～1440

[43] Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT. Copy number polymorphism in fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 2006,439:851～855

[44] Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon. *Am J Hum Genet*, 2006,79:439～448

[45] Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Jr., Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. Replicating genotype-phenotype associations. *Nature*, 2007,447:655～660

第一章 复杂疾病遗传资源收集

一、概　述

复杂疾病(complex disease)，又称为复杂性状疾病、多基因病(polygenic disease)。比如原发性高血压、2 型糖尿病、冠心病、精神分裂症和老年性痴呆(Alzehamers' disease)等，其发病的遗传基础和与环境因素间的相互作用已经成为现代遗传学的主要研究内容之一。单基因病(monogenic disease 或 single gene disorder)是指受一对主基因影响而发生的疾病，它的遗传符合孟德尔定律，所以也称孟德尔式遗传的疾病。依 McKusik 所著 *Mendelian Inheritance in Man* 第 11 版(1994)记载，人类单基因遗传的性状有 6678 种。单基因病在人群中的发病率通常较低，属于罕见疾病。人类的大部分遗传性状不是决定于一对主基因，其遗传基础是多对基因，每一对基因对该遗传性状或遗传病形成的作用是微小的，称为微效基因(minor effect gene)，多对微效基因累加起来可以形成明显的表型效应，成为加性效应(additive effect)，这些基因因此也称为加性基因。多基因性状或遗传病的发病除受微效基因影响外，也受环境因素的影响。近年来的研究表明，多基因病的遗传基础中，除微效基因外，也有一些主基因(major gene) 的参与，而且存在复杂的基因-基因间、基因-环境间相互作用，这类疾病因此被称为复杂疾病。

遗传与环境在疾病发病中的相互作用和相对关系呈一种连续的光谱。但在不同的疾病中，遗传因素和环境因素在发病上所起的作用是各不相同的，归纳起来，可以分为以下四类：

(1) 由遗传因素决定发病，基本看不到什么特定环境因素的作用，如成骨发育不全、血友病、先天性肌弛缓和一些染色体病。

(2) 基本上由遗传因素决定的，但需要一定的环境因素作诱因才发病，例如半乳糖血症纯合隐性基因的婴儿，只有在吃了含半乳糖的乳食后，才诱发生病。

(3) 遗传因素和环境因素对发病都有作用，但在不同疾病发病中，遗传因素的作用大小是不同的，即遗传因素和环境因素共同决定个体易患性的程度。其中遗传基础是多个基因的作用，遗传因素作用的大小称为遗传率(heritability)或遗传度。比如唇裂和腭裂、精神分裂症、先天性幽门狭窄等，遗传度为 75%～80%。遗传因素对发病有重要的作用，而消化性溃疡、各型先天性心脏病等遗传度不到 40%，表明在其发病中环境因素的作用较大，而遗传因素的作用较小。

(4) 基本上由环境因素决定发病，而无明显的遗传基础，比如外伤、中毒及某些急性传染病等。

近 20 年来，人类疾病的概念已经发生了根本性的改变，认为绝大多数疾病都是由环境与遗传相互作用的结果，即所有疾病的发生都与遗传因素有关，差别仅在于程度的不同。研

究疾病病因学中遗传因素和环境因素的相对重要性，成为现代医学的重大课题之一。

复杂疾病的研究策略有候选基因策略和全基因组扫描策略。候选基因策略是指选择那些编码的蛋白质参与疾病发病机制的已知基因作为候选基因。首先将候选基因位点的遗传标记与疾病进行连锁分析，进而通过关联研究比较遗传标记基因型在病例组人群和对照组人群之间的频率分布。候选基因策略方法简单易行，有较强的针对性。全基因组扫描策略是指利用广泛存在于人类基因组中的微卫星标记，采用覆盖整个基因组的 300 多对微卫星 DNA 引物进行 PCR 扩增，对遗传标记的基因型进行分型，经过连锁分析，可以将致病基因定位在染色体的某一区域，分辨率可达 10cM。进一步用区域内多态性标记进行精细定位，将疾病基因定位在 1cM 范围内，可以通过大规模的 DNA 测序，分离并克隆疾病相关基因。

遗传流行病学，人类遗传学和流行病学紧密结合而成的一门新兴的边缘学科，近年来在复杂疾病的研究中，尤其在研究遗传因素和环境因素的相互作用方面表现出很大的优势。对于复杂疾病的研究，目前常用的两种研究方法有连锁研究(linkage study)和关联研究(association study)。目前大样本、多中心合作、不同人群、种族之间比较正在成为国际上复杂疾病研究的趋势，有计划地建立不同种族、不同疾病的遗传资源库可以大大节约资源，是我国 863 重大科技攻关计划和“十五”攻关计划的内容。

二、遗传资源收集

在确定了研究设计之后，遗传资源的收集是研究过程的第一步。遗传资源的收集不仅仅是遗传物质样品的收集，而是指研究对象的入选、相关资料尤其是遗传分析所需的生物样品的收集的过程。本节将就遗传资源应当包含的内容、收集的方法以及质量控制等问题进行讨论。

(一) 遗传资源收集的内容

复杂疾病，以原发性高血压为例，研究设计通常涉及较大的群体，遗传资源收集的内容应当包括人口学资料、个人疾病史、疾病家族史、相关环境因素调查、体格检查和血液生化检查和重要中间表型资料等。

1. 人口学资料(demographic data) 包括姓名、性别、年龄、族别、婚姻状况、受教育情况、职业等，以及联系方式、第二联系人联系方式。

2. 个人疾病史(individual history of disease) 指研究对象与研究相关的疾病的既往史。如以高血压为例，包括高血压初次诊断年龄、是否服药、药物种类、接受药物治疗时间、服药频率，和调查之前两周内服药情况。此外，还应收集与高血压发病有关的疾病信息，如高脂血症、糖尿病、脑卒中、冠心病、心肌梗死、慢性消化道疾病、恶性肿瘤等。

3. 疾病家族史(family history of diseases) 收集被调查者父母、兄弟姐妹和子女的所研究疾病及与其密切相关的疾病比如肥胖、2 型糖尿病、冠心病、脑卒中等的既往史。

4. 相关环境因素调查 与高血压发病相关的环境因素包括吸烟、饮酒、饮食、身体活动等。收集资料应该尽可能详细并且尽可能量化。例如，对于吸烟，应调查开始吸烟年龄、每日吸烟量、持续时间、有否戒烟、现在是否吸烟、吸烟量，以及对于非吸烟者是否存在被动吸烟、被动吸烟的量、频率以及持续时间。

5. 体格检查 例如，身高、体重、腰围、臀围、皮脂厚度等。

6. 血液生化检查　包括与所研究疾病密切相关的血液生化指标，如总胆固醇水平、高密度脂蛋白、低密度脂蛋白、血肌酐、空腹血糖、胰岛素水平等，此外还可以增加载脂蛋白、炎性反应因子、肿瘤坏死因子、C-反应蛋白、黏附因子等非常规项目，也可以根据分子遗传学研究内容作为备测项目。

7. 生物样品的收集和保存　对于遗传学研究，高质量的 DNA 样品、蛋白样品、细胞株的采集和保存是样本收集的核心内容。生物样品的编号和研究对象的其他信息的编号应当保持高度的一致。生物样品的采集、分离和保存应当严格遵守操作程序，以保证样品的质量。

应当高度重视的是，研究应当严格遵循知情同意的原则。即在所有调查项目开始之前，应当先征得被调查者的知情同意，最好签署书面的知情同意书以备日后发生问题时作凭证。所谓知情同意，是指调查者应向被调查者介绍研究的目的、主要的研究方法、被调查者在研究中所承担的义务和责任、所能获得的权益、可能的风险，以及承诺对被调查者的个人信息予以保密，还须说明被调查者有权利随时因任何原因退出研究并不会受到歧视和报复。在了解了以上内容之后，被调查者自愿参加本研究，并签名。

（二）资料收集过程中的质量控制

实际上，质量控制应当贯穿于研究的整个过程，在研究方案的制定、资料收集的过程中以及资料处理中。质量控制的好坏直接影响表型的质量，进而影响关联分析的效力（power）。

1. 事先设计调查表（questionnaire）　遗传流行病学研究通常要求的样本量较大，比如传统的候选基因策略的病例对照研究设计的原发性高血压关联研究的样本量通常要求达到500，连锁研究核心家系样本量也通常超过 100，目前的全基因组关联研究样本量更是达到数千人，因此，在资料收集之前做好设计是至关重要的。所收集的资料内容应当尽量全面、详细，等到资料分析阶段发现信息的缺陷，补救几乎是不可能的。一份好的调查表应该满足如下几项要求：调查项目要精选；项目的定义要明确；项目的答案应简单，便于调查者填写；调查结果应当便于输入计算机；调查表设计好之后在大规模现场正式使用之前，应该在小范围内先进行预调查，及时发现问题，及时修改，保证调查表的可行性。

2. 疾病的诊断应当尽量采用金标准　对所研究的疾病应有明确的定义，有详细的诊断标准和获得有关诊断信息的一致而又可靠的途径。所有疾病表型资料的收集应当尽量采用国际统一的标准化的方法。如高血压的诊断可以采用国际高血压联盟推荐的诊断标准，血压的测量应当采用国际通用的水银柱血压计，而非电子血压计等。数量表型的测量要尽量采用标准、统一的方法，这对于保证研究的把握度和不同研究组之间的数据共享具有十分重要的意义。

3. 所有调查人员均应当事先经过培训和考核，合格后方可参与调查　培训的内容包括研究目的、意义，尤其是调查方法的标准化，比如问卷的填写、表型的测量、人体测量等方法的标准化，血液样品采集、运输、保存、分离方法的标准化，血液生化检查方法的标准化等。

4. 做好质量控制　质量控制的概念应当贯穿于整个研究过程。除其他环节外，值得重视的是，现场调查结束后，应当随机抽取 5%的样本采用新的调查表，由其他调查者操作，按照标准操作程序，对原有调查内容进行二次调查（采血等带有创伤性的项目除外，采用简单的统计方法，观察两次调查间的误差。如果误差较大，则考虑是否采用原始调查结果。