三、全基因组关联分析法在人类基因组研究中遭遇的挑战与取得的成绩

自从人类基因组计划开展以来，十几年时间已经过去。从最初大肆炒作给予人们无限希望到现在逐渐归于理性，科研界也慢慢对许多人类常见疾病背后的遗传基础机制有了进一步的了解。全基因组关联分析法（genome-wide association study）在其中起到了决定性的作用，它帮助人们发现了很多与疾病相关的基因突变位点。现在，在研究疾病发病学的过程中使用遗传学方法和数据已经是一条常规技术路线。同时，遗传学在疾病的防治过程中也起到了很大的作用。

全基因组关联分析法（GWA）是一种用来寻找某种基因变异与表型之间关系的方法。人们对种群中个体的表型，例如是否患有某种疾病进行研究，然后从成千上万种SNP中找出与该表型相关的基因型。如果能够从统计学上确认这些SNP与该表型之间具有联系，那么就可以初步认为它们是与该表型相关的基因型。随后，还需要在另一个种群中进行试验，验证上述发现是否正确。

到目前为止，研究人员已经对70多种疾病（表型）进行了超过300次重复的GWA研究（http://www.genome.gov/gwastudies）。结果，他们发现了30多个与自身免疫性疾病——克隆氏病（crohn's disease）相关的基因位点、20多个与II型糖尿病相关的基因位点以及40多个与人体身高相关的基因位点。而就在两年前，上述这些基因位点的数目还都分别只是2、3和0。

在GWA研究中，研究人员发现某种相关SNP的几率会随着该SNP位点中罕见突变的数量增多而降低。因此，到目前为止，在他们发现的基因型中，大部分都是常见的SNP位点，只有少数是罕见的SNP位点。在GWA研究中，另一个常见的现象就是这些SNP位点对表型（疾病）的影响力都不大。一般来说，每增加一个突变SNP拷贝只会增加10%~30%的患病几率。

除了少数几起例外，研究人员对大部分基因型与表型之间的具体相互关系及其背后的分子信号机制的了解还有待进一步加强。尽管如此，他们还是获得了一些进展，尤其是发现了与不同疾病相关的基因位点之间具有重合现象（知识框1）。不过出乎意料的是，研究人员发现的这些SNP位点并不能帮助他们确定究竟是哪些基因与该疾病相关。

研究人员现在还只是发现了冰山的一角，还需要在目前的工作基础之上进行更深入、更广泛的研究以及荟萃分析（小词典1），籍此发现更多的与该表型相关的基因型信息。同时，还要进行更多新的GWA研究。目前，英国资助医学研究最多的威康信托基金会（Wellcome Trust）准备投资3,000万英镑（约合4,500万美元）用于资助在12万人中对27种表型进行研究的项目。同时，他们还承诺会继续资助该项目。研究人员认为，基于目前在这方面取得的成绩以及如何使用这么巨额的资金，有必要好好审视GWA方法将会面临哪些挑战，这样才能在今后的工作中取得更好的成绩。

通过GWA研究得到了基因型－表型因果关系之后，可以提示人们有哪些信号通路可能参与致病过程，但还需要进行大量的研究以阐明这些基因型－表型之间相互作用的机制。而且，人们还发现了一个有趣的现象，就是完全不同的疾病却可能具有相同的易感基因位点。

下面就来介绍一下最近发现的几种疾病易感基因（disease-susceptibility）。在乳腺癌中，基因BRCA1和基因BRCA2是大家熟知的乳腺癌相关基因，它们参与了DNA修复过程。如果它们发生了突变，就极有可能会患上乳腺癌。不过，最近通过GWA研究发现的几种与乳腺癌相关的基因都有可能会激活生长促进基因（growth-promoting gene），但该结果还有待进一步验证。

有很多与II型糖尿病相关的基因都位于某些特殊的基因座中，这些基因座中还有能限制胰岛β细胞的功能和大小的基因。这一发现有助于解决长久以来一直存在的有关II型糖尿病发病机理方面的一个问题。在克隆氏病的研究中，研究人员发现该病与自体吞噬作用有关，天然免疫机制（innate immune mechanism）也可能在致病过程中起到了重要的作用。

有趣的是，人们发现在不同的疾病之间，它们的致病基因会有所重叠。比如，在人类染色体的8q24部分有一个区域，该区域包含好几个独立的与前列腺癌相关的基因座，但这些基因座中有一个同时也和结肠癌相关，与该位点临近的另一个基因座又与乳腺癌相关。目前，人们还不知道这些位点是否都是通过同一作用机制来引发肿瘤，也不知道具体是哪个基因在发挥作用，只知道这里面最有可能的致病基因就是众所周知的癌基因——MYC。

研究人员不断发现了新的致病位点，这些位点与不止一种的自身免疫病相关，但它们又都与主要组织相容性复合物（major histocompatibility complex）无关。其中有一个重要的位点与好几种自身免疫病都相关。不过，这种致病基因位点相互重叠的机制并不像想象的那样简单。比如，在一个基因座中，有一个等位基因既与克隆氏病相关也与I型糖尿病相关，而另一个等位基因虽然与I型糖尿病相关，但对克隆氏病却具有保护作用。

在GWA研究中，最让人感兴趣的一个研究结果，就是在9号染色体上发现了一段长约120kb的区域，该区域与冠心病（coronary artery disease）相关，不过其中的具体机制还有待研究。有两个参与调节细胞周期的基因——CDKN2A（编码p16，也被称作INK4A）和CDKN28（编码p15，也被称作INK4B）有可能起到了重要作用。人们以前并没有发现这些基因与冠心病有关，相反一直都认为它们与某些癌症相关。此外，人们还在该区域发现有些变异与II型糖尿病有关，这更是增加了该区域的研究热度。

还有一个出乎意料之外的重叠现象，即与II型糖尿病和前列腺癌相关的位点之间的重叠现象。一个位点包含TCF2基因（即转录因子2基因，也被称作HNF1B），该基因突变会增加患前列腺癌的风险，但会对II型糖尿病起到保护作用；与此相反，另一个基因位点（包含基因JAZF1，该基因编码锌指蛋白）中与上述两种疾病相关的等位基因的作用似乎又截然相反。

1.GWA方法将会面临的挑战

随着高通量基因分型（high-throughput genotyping）产品越来越多地进入临床科研工作，GWA研究法也会被更多的人采用。于是，问题就出现了。要进行一项GWA研究，首先需要获取大量患者的DNA样本材料，同时还需要获得足够的经费支持。尽管并不需要基因分型试剂盒（geno-typing kit）或者多么专业的基因组学和数据分析知识，但是有一点非常重要，那就是要有一个好的质控。这是因为，在GWA研究中到处都充满了假象，非常容易得出错误的结果。因此，要想在GWA研究中获得正确的结果，一定要非常注意质控环节，尤其是在GWA被广泛应用的今天，这一点更加需要注意。

另一个挑战就是，GWA研究结果的“快速”获取。在GWA研究中，往往很容易就能发现几个“成果”。不过，研究人员逐渐明白，真正的成果其实并不会这么容易就能获得。比如，在好几种疾病的研究中，他们是通过荟萃分析以及广泛的追踪研究才发现了真正的相关基因型变异。不过，现在也许正好是改变的时机了。与以往的研究不同的是，研究人员现在更注重基因分型的规模效应，因为这样更经济、更有效率，在一次试验中就可以对大量样品的位点（一般是1,536个检测位点，不过也有多达5,000~10,000个检测位点）进行分析，比如最近报道的腹部疾病（coeliac disease）GWA研究就是如此。不过，初始研究（initial study）的规模同样很重要，因为只有借助初始研究发现的结果才能更好地指导后续实验。在某些多阶段研究中，大规模初始研究的效果会更好，因为它可以获取更多、更有效的数据供荟萃分析使用。

通过GWA分析及后续试验验证的基因位点都局限在基因组中很小的几个区域内，通常都在50~250kb之间，少数区域的长度会超过1mb。因此，另一个需要考虑的问题就是，人们发现的与疾病相关的SNP位点似乎都不太可能是“功能性的变异（functional variant）”，而只是为真正功能性变异服务，或与之相关的“代理人”而已。因为在人类基因组中，同一条染色体上相邻的SNP位点等位基因之间通常都具有类似的进化过程，并且彼此相关（连锁），于是人们往往会发现一个SNP位点携带“A”等位基因，而与其相邻的SNP位点则携带“C”等位基因，这种现象被称为连锁不平衡（linkage disequilibrium）现象。这意味着，在关联研究中，实际上并不需要弄清楚实际的（真正的）因果关系。在SNP分析中发现的基因型往往都与真正的基因变异有关，这就说明，只要SNP分析的样本量足够大，这些SNP位点中就会有真正的相关SNP位点。因此，人们现在面临的问题是如何对相关区域进行深入研究，发现真正的致病基因，以及弄清它们的致病机制。

基于这个思路，如果对患病人群和健康人群的基因组目标区域进行更精细的作图，是不是会得到比GWA更好的结果呢？那些致病SNP的信号应该会比其它SNP的信号要强一些。最理想的状态是，能对相关热点区域里尽可能多的SNP位点进行作图。因此，最好在精细作图（fine mapping）之前，对待测样本的相关区域再次测序，以尽可能多地发现SNP位点。不过，现在的千人基因组计划（1000 Genomes Project，http://www.1000genomes.org）至少可以为人们提供健康人群对照组的信息，这在一定程度上减少了重复测序的必要性。现在，学界已经发表了几篇对相关热点区域进行精细作图的研究文章，不过对于如何更好地设计这类试验，还有很多重要的问题尚待解决。例如，还没有一种基因分型试验（genotyping experiment）可以很好地区分致病SNP位点和其它与该SNP紧密相关的SNP位点。因此，即使精细作图试验取得了成果，也不太容易确切地指出致病SNP，但至少它有助于缩小研究的范围，指导后续的功能试验。

这一阶段的工作将非常困难。很多已知遗传机制的人类遗传疾病都是因为基因发生截短或改变而使基因产物（蛋白质）功能改变而导致的。这在其它人类常见疾病中也是非常常见的现象。不过，越来越多的证据显示，有相当一部分甚至是大部分的功能变异（致病变异）实际上是发生在基因调节水平上，而不是在基因产物水平上。比如，在GWA研究中发现的好几个基因组相关区域实际上并不包含开放阅读框。在评价GWA研究发现的位点功能试验中，使用基因敲除等技术可以获得更精细的信息。这也对基因调控研究领域以及试验技术领域提出了更高的要求。

还有一种方法可以对GWA的结果进行进一步验证，那就是在基因组的相关热点区域筛查基因中的罕见突变。如果在患病人群中出现该突变的几率在统计学上要高于健康对照组人群，那么就说明该基因与该疾病相关。不过和上述精细作图法一样，这个方法在如何设计实验等方面也还存在一定问题。

2.预测患病风险

图片说明：最近的科研进展帮助人们绘制出了与疾病相关的DNA序列图，但就像早期的世界地图一样，还有很多细节有待进一步研究。

使用GWA研究不仅能帮助人们了解疾病的病理生理学（pathophysiology）和病因学（aetiology），同时还能帮助人们预测未来罹患该病的风险。不过，虽然看起来在医疗工作中不可避免地会用到遗传信息，但人们似乎还是不太清楚该如何使用它们。

有一个问题就是，人们现在还不能根据已知的致病变异来判断患病风险，而且很有可能会低估患病风险，因为有大部分的致病变异还没有被发现。目前，人们主要通过检测个人基因组中是否含有由GWA发现的“代理（proxy）SNP”位点来判断自身的患病风险。由于这不是一个非常精确的测试，所以，个体的患病风险通常都被低估了。基于同样的原因，要使用统计学方法来发现正确的疾病模型，比如判断某个表型是隐性的还是显性的也非常困难，除非清楚知道哪些变异是真正的致病变异，或者知道哪些变异是与致病变异高度相关的基因变异。因此，随着人们发现的致病位点越来越多，人群中相应的发病几率也会越来越高，结合多个位点进行预测的准确性也将越来越高。

至于如何将这些基因风险因子（genetic risk factor）与环境风险因子（environmental risk factor）结合起来还有待研究。比如，有一个55岁的男性，他有20年的吸烟史，同时具有心血管疾病相关遗传背景，那么他患上心血管疾病的风险有多高呢？不同的疗法针对不同遗传背景的患者的疗效有多高呢？要回答这些问题需要进行大量的、昂贵的研究，需要进行前瞻性群组研究（prospective cohorts study）来分析环境危险因素与遗传风险因素之间的相互作用关系，需要进行临床试验来评价遗传信息对于疗法选择有什么帮助等等。越早进行这些研究，就会获得越好的结果。

随着评价患病风险的准确性越来越高，已经有好几家公司开始提供“个性化基因组学（personal genomics）”服务项目。他们主要是针对人体基因组中500,000~1,000,000个SNP位点进行检测，并据此得出相应的患病风险结果。这类检测是否可靠、是否

有科学依据非常重要。此外，还需要保证客户能够真正理解测试结果的意义。

众所周知，遗传分子标志物只能在一定程度上帮助科研工作者预测个体是否会患上某种疾病。个体基因组学则可以帮助人们从另一个角度进行预测。对于某些疾病来说，大部分的人体都可能会遗传到一些致病变异，同时也可能遗传到一些保护性变异，因此，会有一个平均的患病风险。不过，有一小部分人可能主要获得的都是致病变异遗传，所以他们的患病风险就要高一些。以克隆氏病为例，使用最简单的风险倍增疾病模型（simplest multiplicative-risk disease model），按照保守的、较低的患病风险估计，在英国携带有克隆氏病相关基因位点的人群中有5%的人的患病风险比平均风险要高出5~8倍，有1%的人的患病风险比平均风险要高出9~15倍。在II型糖尿病高危人群中进行类似的研究发现，有5%的人的患病风险比平均风险要高出2~3倍，有1%的人的患病风险比平均风险要高出4倍。

在50种疾病中，按照简化的假设，即每一种疾病的易感性之间互相独立，毫无关系，那么几乎每个人至少都会属于易患某种疾病的高危人群中的前5%，几乎有一半的人会属于易患某种疾病的高危人群中的前1%。因此可以预测，例如某人很容易会患上某些疾病，因为他也遗传了一些常见的基因变异。虽然现在某人还不清楚会得哪种病，但随着个体化基因组学时代的来临，终究会知道的。如果不清楚情况，就不知道他是否该做出某些改变，比如改变生活方式、进行身体检查或接受某些“干涉（intervention）”等以避免患病。尽早了解患病风险，就可以尽早进行相应的处理以避免发病，由此可见遗传风险因子的重要性。需要指出的是，尽管还不清楚家族史在疾病预测中的作用有多大，但了解家族史也同样重要，因为它很有可能帮助人们发现在GWA研究中无法发现的、非常重要的以及作用很强的致病变异。

3.展望

尽管研究人员已经揭开了人类基因组中的部分秘密，但还有很多问题等着他们去研究。在发现与疾病相关的遗传变异方面，GWA取得了不错的成绩，不过今后的研究历程会更加艰难。

研究人员使用GWA获得了一些成果，同时也对GWA有了更深入的认识。但是，使用GWA可能不再那么容易获得结果了。更先进的新一代测序技术也削弱了GWA技术的优势地位。尽管借助GWA分析获得了很多结果，但人们逐渐发现，这些结果越来越难以解释遗传性疾病的特点了。

不过，如果要就此快速做出结论，全盘否定GWA的试验结果也是不对的。随着人们对表型－基因型因果关系研究的深入，毫无疑问会对遗传现象了解得越来越多，越来越清楚。如果不考虑这些遗传位点在预测患病风险上的问题的话，就帮助人们更好地认识疾病的病因学、预测药物靶点等方面来说，还是有很大用处的。同样，如果认为有了基因组完整的测序数据，就不需要对GWA研究进行改进了，这种想法也是非常天真的。比如，在GWA研究后要确定一个基因型－表型因果关系就很困难，因为由于连锁不平衡的原因，相邻的SNP之间会有连锁现象发生。同样，在测序时同样存在连锁不平衡现象。而且即使测序的费用降到非常低的水平，要想如GWA研究一样，获得大量样本的基因组数据，至少在几年之内是不太可能实现的。

旨在探索“遗漏的遗传现象（missing heritability）”的并行研究（Parallel study）也非常重要。在改善人类健康状况的征途上还有很长、很艰巨的路要走，但这是非常值得走下去的。

虽然人们可能无法取得前几年那么大的成就，但当最终弄懂了人类常见病背后的发病机制时，就会发现，今天的工作还是非常有意义的。

原文检索：Peter Donnelly. (2008) Progress and challenges in genome-wide association studies in humans, *Nature* 456 (7223): 728-731.

筱玥/编译

小词典

1. 荟萃分析（meta-analyse）

荟萃分析，又称“Meta分析”。Meta意指较晚出现的更为综合的事物，而且通常用于命名一个新的相关的并对原始学科进行评论的学问，不但包括数据结合，而且包括结果的流行病学探索和评价，并以原始研究的发现取代个体作为分析实体。荟萃分析产生的主要的理由是：对于多个单独进行的研究而言，许多观察组样本过小，难以产生任何明确意见。在过去的十年中，医学研究领域中有关荟萃分析的已发表的论文数量急剧地增加。那么，究竟什么是荟萃分析？Huque给出了一个有用的定义：“一种对被分析家认为可组合的多个各自独立的临床实验结果进行组合或整合性统计学分析。”荟萃分析的过程必须有统计学的参与，目前有相当多的统计学软件中已经包含了相应的分析模块，专门的Meta软件也已经问世。

四、全球联手，消灭疟疾
——疟疾基因流行病学的研究

对疟疾基因组内的遗传变异现象进行大规模研究，将有望攻克疟疾这一顽症。但是，在疟疾肆虐的地区——发展中国家，进行这样的研究会受到科技发展水平和伦理道德的制约，同时还存在可操作性方面的问题。疟疾基因流行病学计划（Malaria Genomic Epidemiology Network，Malaria GEN）就是在这样的大背景下出现的。该计划集合了全球21个国家的众多科学家。人类希望集全球之力一起攻克疟疾。

疟疾是由单细胞寄生性的疟原虫（*Plasmodium*）造成的一种疾病。疟原虫是一种原生动物，生活史非常复杂（图1）。

自然界中有100多种疟原虫，其中5种能感染人类。而这5种疟原虫中，恶性疟原虫（*Plasmodium falciparum*）、间日疟原虫（*Plasmodium vivax*）、卵形疟原虫（*Plasmodium ovale*）以及部分诺尔斯（氏）疟原虫（*Plasmodium knowlesi. P. falciparum*）能引起严重的疟疾，它们是大部分致死性疟疾的元凶。疟疾通过按蚊（*Anopheles*）的叮咬在人群之间传播。按