·创刊30周年纪念·

全基因组关联研究与复杂疾病风险预测的现状与展望

沈洪兵 靳光付

【关键词】 遗传标志物；遗传风险；风险预测；全基因组关联研究；基因组流行病学

Genome-wide association study (GWAS) and risk prediction of complex disease: advances and prospects *SHEN Hong-bing, JIN Guang-fu. Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing* 210029, China

Corresponding author: SHEN Hong-bing, Email: hbshen@njmu.edu.cn

【Key words】 Genetic marker; Genetic risk; Risk prediction; Genome-wide association study; Genome epidemiology

人类基因组计划(human genome project, HGP)的完成预示着生命科学研究进入了基因组时代，在利用基因组学方法进行流行病学研究的过程中产生了基因组流行病学。在人群中研究与疾病发生发展或健康相关的遗传变异，即遗传标志物(genetic marker)用于疾病的预防和治疗、促进健康，是基因组流行病学的主要研究内容。在21世纪初的几年里，基因组流行病学研究者将精力集中在与疾病密切相关的功能已知基因上，认为这些基因关键区域的多态性(polymorphism)可能影响基因和(或)蛋白产物的生物学功能，以外显子区导致氨基酸改变(nonsynonymouse)的多态性或者位于基因启动子区域的多态性为主，如TP53 Arg72Pro与肿瘤等；后期也出现了基于连锁不平衡(linkage disequilibrium, LD)原理，挑选代表整个基因的标签(tagging)多态性位点进行的关联研究，这些研究无一例外都是基于所研究基因具有重要生物功能这一假设，此类关联研究被称之为候选基因(candidate gene)策略。然而，此类研究经常出现不一致甚至相反的结果，目前认为主要是样本量不足导致的机遇性假阳性。随着国际人类单体型计划(HapMap)的推进，研究者对于人类基因组遗传变异，特别是单核苷酸多态性(single nucleotide polymorphism, SNP)有了更全面的认识。在基因组中有上千万的常见SNP，即较小等位基因频率(minor allele frequency, MAF)不低于5%的SNP位点，位于同一染色体区域多个SNP位点倾向于整体遗传，多个位点之间具有较强的相关性(即高度LD)，其中一个或多个位点(标签位点)即可以代表该区域的所有位点[1]，因此检测时仅需检测这些标签位点即可。与此同时，基因分型技术不断推陈出新，特别是高通量芯片的应用使得同时进行几十万甚至上百万SNP基因分型成为可能，这些芯片可以代表基因组80%左右的常见SNP位点，随着基因分型成本的大幅下降，结合在候选基因研究策略过程中积累的大样本资源，共同催生了全基因组关联研究(genome-wide association study, GWAS)[2,3]。该研究策略在较大样本量(一般大于1000对研究对象)的研究人群中，筛选与疾病或性状显著相关的位点，并利用额外的一个或多个人群进行验证，最终确定表型相关SNP位点[2]。多种类型的研究设计均可采用GWAS，包括病例对照研究、队列研究和临床试验等。GWAS分析过程中需要对每个SNP位点进行至少一次统计检验，应用Bonferroni等方法进行多重比较的校正是GWAS通常采用的方法，因此根据GWAS分析SNP的数量其显著性水平将低至10^{-7}或10^{-8}。两阶段或多阶段的研究设计是GWAS的另一个特点，即在全基因组扫描分析后，对其中显著的位点(显著标准根据各个研究的情况具体选择)在额外人群中验证。这种多阶段的筛选和验证策略，也是近年来分子流行病学研究中一个新趋势，既节约研究成本，又能有效降低假阳性，确保研究结果的真实性；独立的验证人群和增加的样本量则可以降低由于单个研究人群或研究设计缺陷所导致的虚假结果或虚假效应，增加研究结果的可靠性[2]。

DOI: 10.3760/cma.j.issn.0254-6450.2011.07.002

作者单位：210029 南京医科大学公共卫生学院流行病与卫生统计学系

通信作者：沈洪兵，Email：hbshen@njmu.edu.cn

由于具备上述特征，GWAS的结果往往具有较高的真实性和可靠性，可重复性强。目前，600余项GWAS覆盖了近200种疾病或性状，已经发现P值在5×10^{-8}以下的表型相关位点上千个[3]。有关GWAS结果的特点已有阐述[4]，本文不再赘述。目前GWAS的研究成果在应用转化方面，主要集中在疾病遗传风险预测（genetic risk prediction），为解读该领域的最新进展，本文将对GWAS研究较为深入的几种复杂疾病，如前列腺癌、乳腺癌、2型糖尿病和冠心病进行回顾，简述利用GWAS成果进行风险预测的相关研究结果，在分析和解读这些研究结果的基础上，指出当前遗传风险预测研究面临的挑战，展望未来的研究趋势，以期为国内的研究提供借鉴。

一、复杂疾病遗传风险预测研究现状

1. 前列腺癌：双生子研究结果表明，前列腺癌是受遗传因素影响最大的恶性肿瘤，其病因可归咎于遗传因素的部分为42%[5]。前列腺癌是最早开展GWAS的恶性肿瘤，也是迄今为止发现遗传易感位点最多的一类肿瘤，在欧美人群已经发现30余个前列腺癌易感位点[6]，日本人群也报道了5个新的前列腺癌易感位点[7]。遗传位点已广泛用于前列腺癌风险预测研究（表1）。Zheng等[8]在瑞典人群的研究发现，年龄和肿瘤家族史对前列腺癌风险预测能力[以受试者工作特征曲线（ROC）下面积（AUC）作为指标]为0.607，利用11个前列腺癌易感位点信息后可将这一数值提高至0.648。Johansson等[9]在另一个瑞典人群的研究中发现，利用33个SNP位点信息可将前列腺特异抗原（PSA）预测前列腺癌风险的能力从0.862增加至0.872，这两项研究都显示了遗传标志物对传统因素预测前列腺癌发病风险有显著的改善（至少在统计学意义上），但提高的绝对值却有限，AUC升高仅为0.04和0.01。

2. 乳腺癌：在欧美国家，Gail模型已经广泛用于女性散发性乳腺癌风险评估，该模型包括年龄、家族史（一级亲属乳腺癌家族史）、生殖因素（初潮年龄和首胎活产年龄）和疾病史（乳腺活检次数和乳腺不典型增生）等因素[10]。Gail[11]应用美国国家癌症研究所（NCI）乳腺癌风险评价工具（BCRAT）对GWAS早期发现的7个乳腺癌易感位点进行评价，分析发现这7个易感位点联合后可以将Gail模型的预测能力从0.607提高至0.632。随后，Mealiffe等[12]应用大型临床试验的乳腺癌病例和对照进行实例分析，发现7个SNP的单独预测能力为0.587，优于传统的Gail模型（0.557），两者结合后AUC为0.594。Wacholder等[13]应用4项美国队列研究和1项波兰的病例对照研究（累计超过1万例的大样本人群），系统评价10个乳腺癌易感位点在乳腺癌风险预测中的价值，研究结果与Mealiffe等较为一致，遗传模型的AUC（0.597）优于Gail模型（0.580），两者结合后AUC可达到0.618。此外，Zheng等[14]利用上海地区乳腺癌研究人群评价易感位点在中国人群的预测价值，该研究包括3039例乳腺癌病例和3082例对照，初潮年龄、首胎活产年龄、腰臀比、乳腺癌肿瘤家族史和乳腺良性疾病史等传统因素对乳腺癌发病风险的预测能力为0.6178，应用8个易感位点的信息后，预测能力提高至0.6295。

3. 2型糖尿病：2型糖尿病的GWAS研究较为广泛，至今已经发现40余个易感位点，预计可以解释10%左右的糖尿病遗传性[15]。目前，已有10余项研究评价了易感位点对2型糖尿病发病的预测能力[16]，

表1　常见复杂疾病遗传标志物风险预测的研究结果

疾病	研究类型	人群	样本量（病例/对照）	非遗传标志物	遗传标志物（SNP）	AUC[a] 非遗传	AUC[a] 遗传	AUC[a] 合并	P值[b]
前列腺癌[8]	病例对照研究	瑞典	2893/1781	年龄和家族史	11	0.607	-	0.648	<0.001
前列腺癌[9]	巢式病例对照研究	瑞典	520/988	血清总PSA和游离/总PSA比值	33	0.862	0.643	0.872	0.002
乳腺癌[12]	基于临床试验的病例对照研究	非西班牙裔白人	1664/1636	年龄、种族、家族史、初潮年龄、首胎活产年龄和乳腺活检次数	7	0.557	0.587	0.594	<0.001
乳腺癌[13]	4项队列研究和1项病例对照研究	美国和波兰	5590/5998	家族史、初潮年龄、首胎活产年龄和乳腺活检次数	10	0.580	0.597	0.618	-
乳腺癌[14]	病例对照研究	中国	3039/3082	初潮年龄、首胎活产年龄、腰臀比、家族史和良性乳腺病史	8	0.6178	-	0.6295	<0.001
2型糖尿病[17]	队列研究	美国	446/3025	年龄、性别、家族史、BMI、FPG、SBP、HDL-C和TG	40	0.903	-	0.906	0.04
冠心病[20]	前瞻性队列研究	芬兰	715/12 065	年龄、性别、LDL-C、HDL-C、吸烟、BMI、高血压、抗血压治疗和糖尿病	13	0.871	-	0.872	0.19

注：[a]采用受试者工作特征（ROC）曲线下面积（AUC）评价；[b]包括遗传标志物（合并）的预测模型与不包括遗传标志物（非遗传）的预测模型AUC之间统计学检验P值

其中大部分研究应用15～20个易感位点，遗传模型的AUC约为0.60，而年龄、性别和体重指数（BMI）等传统因素所构成预测模型的AUC为0.7～0.8，易感位点对传统因素预测能力的改善有限（AUC差值≤0.02）。

最近，de Miguel-Yanes等[17]在弗明汉子代研究（Framingham Offspring Study）中评价了40个SNP对该人群2型糖尿病的风险预测能力，年龄、性别和家族史等传统因素的AUC为0.903，应用40个SNP后AUC增至0.906，尽管两者间的差异有统计学意义，但绝对数量仍然很小。研究者将研究人群分为低龄组（<50岁）和高龄组（≥50岁）后，易感位点在低龄组AUC从0.908增至0.911，在高龄组从0.883增至0.884，两个年龄组的AUC增加均不显著[17]。但是，研究者发现低龄组的再分类优化差值（net reclassification improvement，NRI）将提高10.2%，而高龄组仅提高0.4%，显示这些标志物对低龄人群2型糖尿病的风险分类有较大的改善作用[17]。

4. 冠心病：遗传因素预计可以解释30%～60%冠心病发病风险，结合最新发现的一系列易感位点，目前已发现的冠心病易感位点接近30个，可以解释约10%的遗传因素[18,19]。目前，评价易感位点预测冠心病风险的研究有限，Ripatti等[20]利用芬兰的前瞻性队列研究评价早期发现的13个SNP对冠心病风险的预测能力（表1），在该队列中传统因素对冠心病的预测能力为0.871，应用13个SNP后提高至0.872，差异无统计学意义。研究者进一步分析了13个遗传标志物对该人群风险分类的影响，分析显示，遗传标志物提高了2.2%的再分类优化差值，差异仍无统计学意义（P=0.18），但是总体识别优值（integrated discrimination index，IDI）则显著提高了0.004（P=0.0006）[20]。由此可见，这些遗传标志物对冠心病的预测能力相对较弱。可以预计，如果应用最新发现的易感位点，遗传标志物的预测能力有望得到改善，但其具体效果还需将来的研究进行评价。

二、遗传风险预测研究结果解读与展望

上述4种疾病是目前GWAS研究较为成功的，通过风险预测研究结果可以发现：遗传标志物确实可以提高疾病风险预测能力，在某些疾病中（如乳腺癌）与传统预测因素效果相当，但是对传统因素风险预测能力提高的幅度有限。鉴于此，有研究者对GWAS提出异议[21,22]，认为GWAS投入较大，且在转化应用上难有作为。尽管如此，更多的研究者认为，GWAS是近年来科学研究最为成功的范例之一。首先，GWAS的研究结果已引起基础机制研究方向的转变，特别是在功能基因组相关研究领域，GWAS发现的大量疾病易感位点位于以往认为的基因组“垃圾序列”（非编码区或非基因区），最新的研究认为这些区域虽不编码蛋白质，但可能调控所有蛋白的功能，已经逐渐成为目前生命科学研究最热门的领域，如非编码RNA、增强子等研究。这些研究可能开拓生命科学研究的新领域，揭示疾病发生发展深层的机制，有望在药物开发和疾病治疗中形成突破性进展[23]。其次，以往通过宏观研究如双生子研究等认识到遗传因素在复杂性疾病中的重要作用，但真正的致病位点或可靠的遗传标志物则鲜见报道，而GWAS在短短几年间发现了上千个疾病相关的遗传位点，这对认识和阐明疾病的发生机制是一个伟大的进步[3]。再则，GWAS不是“一次性”的研究，第一波GWAS过后产生了大量的基因组数据，这些数据可以与各种表型建立联系，这些数据仍将是一个“宝藏”，是以后应用研究的基础数据，研究者可以从不同角度利用和挖掘，揭示遗传的和非遗传的机制。因此，GWAS对于疾病的研究来说是一个极其重要的“原始积累”过程，这也是欧美国家不惜投入，源源不断在各种疾病上开展GWAS研究的原因所在[24]。

GWAS除上述贡献外，对于其研究结果的转化应用予以正确解读，盯住其某个指标或基于特定的出发点，易造成全盘否认或者盲目扩大，不利于形成科学的认识。为此，本文将从以下几个方面分析和讨论GWAS及后GWAS时代应用遗传标志物进行复杂疾病风险预测的前景。

1. 认识遗传因素在复杂疾病中的效应。探讨遗传标志物在复杂疾病中的应用前景，首先应分析遗传因素在复杂疾病中作用。复杂疾病是遗传因素和环境因素共同作用的结果，其中归因于“纯遗传因素”的仅占一小部分。例如，欧洲人群双生子研究表明，在恶性肿瘤中前列腺癌归因于遗传因素的比例最高，可达到42%，其他恶性肿瘤一般约为30%[5]。基因-环境交互作用则是构成复杂疾病病因的主要部分。因此可以预见，即使能够解释全部遗传因素的遗传位点完全被发现，在不考虑部分传统因素事实上已经代表了部分遗传因素的情况下（如肿瘤家族史，前列腺癌的PSA水平等），单纯应用这些位点预测疾病风险的效果可能有限，除非能够更好的认识环境因素，特别是阐明与环境存在交互作用的那部分遗传因素。此外，遗传因素在各种疾病中的作

用则不同，因此对遗传标志物预测疾病风险的期望值也应视疾病而异，遗传度高的复杂疾病可能更适合应用遗传标志物进行疾病预测，而环境因素占主导作用的疾病可能应该策略性的从其他角度探讨遗传标志物的应用价值。

其次分析目前GWAS发现的遗传位点对遗传因素可解释的程度。尽管GWAS发现了上千个疾病易感位点，但具体到某一特定疾病，大部分尚不足10个；即使对于个别已经发现了几十个易感位点的疾病，这些位点也仅仅能够解释疾病小部分的遗传机制。例如，GWAS已发现Cohn's病的71个位点，这些位点共同可以解释该疾病23.2%的遗传性[25]。目前GWAS发现的这些位点频率一般较高（MAF>0.10），效应相对较强（OR=1.10～1.50），尚有大部分频率低或效应低的遗传位点有待进一步阐明。增加研究样本量是一个非常有效的策略，最近基于多中心GWAS的Meta分析，使得GWAS分析样本量大大增加，研究结果也展现了巨大的成效[18,25,26]。例如，冠心病GWAS以往共发现了12个易感位点，Schunkert等[18]对14项GWAS（共2万余例病例和6万余例对照以及5万余例的验证样本）进行Meta分析，一次性新发现13个冠心病易感位点。数据模拟也显示（表2）[27]，随着GWAS样本量的增加，确实能够发现更多的遗传位点，Crohn's病约有142个常见易感位点，乳腺癌、前列腺癌和结直肠癌等遗传度较高的肿瘤常见易感位点个数预计平均可以达到67个。但是随着易感位点数量的不断增加，对疾病遗传因素可解释程度的增加趋势将减缓，预计最终GWAS发现的易感位点仅能解释疾病15%～20%或略高一些的遗传性[27]。

表2　肿瘤和Crohn's病GWAS预期发现的易感位点及其可解释的遗传易感性比例[27]

样本量[a]（×10⁴）	肿瘤[b]		Crohn's病	
	预期发现的易感位点	可解释遗传性(%)	预期发现的易感位点	可解释遗传性(%)
1	2.8	2.8	26.0	11.1
2	10.1	5.8	64.4	14.6
3	21.2	8.7	108.2	17.7
4	33.6	11.4	132.7	19.3
5	44.5	13.5	140.1	19.8

注：[a]病例和对照的比例为1∶1；[b]乳腺癌、前列腺癌和结直肠癌平均估计值

2. 发现更多的遗传位点以解释疾病的遗传性。目前明确的是GWAS增加样本量后可以发现更多的疾病易感位点，但显而易见GWAS样本量的增加有限，且GWAS也不可能检出疾病全部的遗传位点，如何发现那些被GWAS遗失的遗传性（missing heritability）已引发研究者广泛探讨[28,29]。目前被关注的是罕见（MAF<0.01）单碱基变异（或称之为遗传突变）[21,28]。这类遗传位点很难被现有的GWAS检出，一方面是罕见SNP数量众多，相互之间LD程度较低，GWAS芯片难以较好覆盖这些位点；另一方面由于SNP频率较低，若非效应很强的位点，现有的GWAS研究样本量还不足以检出这些位点[28]。针对特定区域的重测序（resequencing）和精细作图（fine-mapping）分析以及全基因组测序（whole-genome sequencing）研究，可能为疾病风险相关罕见SNP的发现提供帮助[30]。研究者最近提出了下一代关联研究（next-generation association study）的概念[31]，这种研究策略将基于已有的GWAS数据，整合全基因组测序结果，将GWAS研究领域从高频（>0.05）向低频（0.01～0.05）甚至罕见遗传变异推进[32]，也许可为将来（全基因组测序成本降至足以开展大样本研究前）一段时期内的罕见遗传变异研究提供可借鉴的思路。

另一个被广泛关注的是基因组拷贝数变异（CNV）[28]。CNV是指大于1 kb的基因组结构变异，由于其对基因组改变程度较大，因此被认为可能对生理病理过程具有较大的影响，在复杂疾病遗传机制中具有重要作用[33]。人群研究确实发现了一些与复杂疾病有关的罕见CNV[34-37]，但由于其数量有限，且频率较低（<1%），对复杂疾病遗传性的解释较小。目前大部分已知的常见（common）CNV与SNP存在高度连锁不平衡[38,39]，Wellcome Trust协作组开展的CNV全基因组关联研究发现了3个疾病相关CNV，但这些区域均已被以往GWAS研究所发现，该结果进一步显示，已知（或可检测到）的常见CNV也不可能很好解释GWAS未能阐明的遗传机制。随着千人基因组计划（1000 Genome Project）的不断推进，人类基因组CNV的图谱有望更加清晰[40]，检测和分析技术的发展有望为阐明CNV在复杂疾病遗传机制中的作用提供更好的平台[41]。

此外，千人基因组计划首批数据除了发现1500万个SNP和2万余个CNV，还发现了多达100万个短序列插入和缺失（indel）[42]，虽在数量上小于SNP但远远超过CNV，可能蕴藏着疾病的部分遗传机制，该领域限于目前高通量分型技术的不成熟，尚未进行大规模的人群研究，可能是将来研究的一个增长点。

当然，阐明多个遗传位点之间以及遗传与环境

之间的交互作用，这将是未来研究的重点和难点。目前GWAS均是单纯基于遗传因素，忽略或淡化了环境因素，正如前文所述，遗传因素和环境因素在复杂疾病致病过程中的单独作用有限，很大程度是二者共同作用的结果，因此如何识别并检出二者之间交互作用是提高疾病风险预测能力的关键所在[43]。但此类研究还面临很大的不确定性，因为目前在研究设计和分析方法上还不成熟，将来的发展在很大程度上将依赖未来相关支撑条件的进展而定。

3. 优化遗传预测模型构建和评价的策略。在应用GWAS发现的遗传位点进行疾病风险预测中，模型的构建和评价方法也不断改进。在遗传位点效应累积构建遗传模型方法上，目前的研究已经基本上摒弃了单纯累计危险位点（等位基因）个数的分析方法，大多采用基于位点效应和频率信息加权累积的方法计算相对风险[44]，或对其进行对数转换使其呈正态分布，并应用人群登记数据的发病率和死亡率信息估计个体发生某种疾病的绝对风险。

模型评价常用的方法是ROC曲线，通过在所有可能界值处用真阳性率（灵敏度）对假阳性率（1-特异度）作图，比较不同模型AUC，综合评价模型的灵敏度和特异度等真实性。但是目前该方法受到质疑[45]。在实际研究中，当AUC达到一定程度后，额外的标志物将很难明显提高AUC。在遗传标志物的模型评价中，遗传标志物对传统模型AUC的提高有限，同理，若以遗传标志物作为基础模型，评价个别传统的因素其对模型的额外贡献也非常有限。鉴于此，研究者提出NRI和IDI[17,20,45]。前者指通过分类表方式评价新的预测因子对原有分类表正确分类（如高、中和低危等）能力提高的程度。其基本方法为计算阳性事件（如发病或死亡）在各风险分类（组）评价中提高比例与降低比例的差值，并计算阴性事件在各风险分类（组）降低比例与提高比例的差值，二者之和即为NRI；该指标可以通过simple asymptotic test评价其显著性[45]。NRI的弊端容易受到分类（组）标准的限制和影响。为此研究者将分类（组）的概念进一步延伸，即将每个个体归为一组，并根据其变动情况进行赋值，阳性事件和阴性事件获益的比例即为灵敏度和特异度提高值，二者之和即为IDI[45]。该指标反映了特异度不变的情况下，应用新预测因子灵敏度的平均提高状况，同样可以通过simple asymptotic test评价其显著性。

模型的构建和评价不仅与待评价遗传标志物有关，也受到多种其他因素影响，如人群的选择、传统因素的定义和采集、缺失值的处理等。鉴于目前遗传风险评价研究良莠不齐，最近一些国际期刊同时刊登了遗传风险评价领域研究者共同起草的遗传风险预测研究（Genetic Risk Prediction Studies）声明[46]及其操作细则[47]，以规范该类研究。

4. 调整遗传标志物的应用期望。在人类基因组测序完成之际，科学家曾提出美好的蓝图：在婴儿出生时，只需要对其进行DNA检测，就能够获知其将来可能发生的疾病。GWAS无疑是迄今为止接近这一目标最为有效的方式。但如前分析，仅仅依靠基因组（遗传）可能无法达到这一目标，因为遗传在疾病中的作用决定了研究所及的位置。但是，遗传标志物对复杂疾病风险的预测能力确定无疑，现在的问题是如何应用。随着更多的遗传标志物被发现，将来在以下几方面有望开展应用：

首先，尽管标志物整体的预测能力受到普遍关注，但在现实应用中，该标志物对于高危人群的检出能力更受关注，高危人群往往是预防和早期诊断关注的重点，筛选的高危人群可以早期进行生活方式干预或者药物预防，可作为进一步开展其他检查的目标人群。因此，遗传标志物对于高危人群的筛选可能是将来其应用的重点。

其次，遗传标志物对亚组人群的风险预测能力将是其应用的另一个重点。例如，吸烟可能是肺癌最大的风险预测因子，若将人群分为吸烟和非吸烟两个亚人群，分别应用遗传模型预测吸烟者和非吸烟者的肺癌风险，检出两个不同暴露亚人群的高危个体，将比粗略评价遗传标志物对整个人群的预测能力更有价值，如de Miguel-Yanes等[17]已在弗明汉子代研究评价2型糖尿病遗传模型上做了类似分析。在特殊人群中检出高危人群进行预防和干预，将是未来遗传标志物应用的一个重要策略。

再则，辅助诊断或辅助检查可能是遗传标志物将来应用的一个途径。例如，Aly等[48]通过一项5241人前列腺穿刺活检队列评价遗传标志物作为穿刺评判标准的价值，与年龄、前列腺癌家族史、总PSA和游离/总PSA比值等传统非遗传因素相比，利用35个前列腺癌易感标志物信息将使得22.7%的研究对象免于穿刺，但将造成2%～3%的进展型前列腺癌患者漏诊。这一结果显示了遗传标志物也许在前列腺活检穿刺上可以作为一项重要的判断指标。

此外，与传统因素（特别是环境暴露和生活因素）相比，遗传标志物拥有许多优势。首先这一指标极其稳定，可以在婴儿出生之时进行评价，也可以在

人生的任何时期甚至出生前的妊娠期进行评价，结果，不会发生改变。但是传统预测因子一般很难保持稳定，如环境暴露因素、体内指标(如PSA水平)的改变，甚至家族史也会有变化。其次是这一指标的早期预测性能较好，遗传指标可以在出生前或出生时检测，在接触环境暴露之前进行风险评价，真正做到"早期"预防，这是其他因素难以实现的。再有，尽管一些研究者认为，许多传统因素不需要检测，没有成本的考虑，但目前基因分型的成本已经极其低廉，其成本仍将会继续下降，而重要的是，在一次检测中可以同时检测多种表型的遗传位点，用于多种疾病风险的预测。遗传标志物这种稳定、早期、快速和综合应用的优势将增加其应用前景。

三、结语

近年来，随着我国科研投入的不断加大及国内外交流日益深入，在基因组流行病学领域的研究正在逐渐缩小与发达国家的差距，但在GWAS及其成果转化等方面与欧美发达国家的差距还很大。GWAS与一般医学研究不同，在技术成熟的情况下，更依赖于样本的积累和经费的投入。我国人口资源丰富，经过多年的基因组流行病学研究，已具备了较好的样本积累。与欧美甚至日本等国家相比，我国在研究投入的时间上滞后，投入的力度也较小，尽管在乳腺癌、鼻咽癌、胃(食管)癌和肝癌等恶性肿瘤以及银屑病、白癜风、系统性红斑狼疮等皮肤病上取得了一些成绩，但这些研究结果对于阐明我国人群复杂疾病的遗传机制，尚显微不足道。因此，基于我国人群特有的易感位点进行遗传风险预测的研究几乎未见报道。鉴于不同种族之间疾病的易感性存在明显差异，欧美人群的研究结果不可能照搬到我国人群。因此，在全基因组测序等研究成本较高、短期内尚难以开展大样本人群测序研究的情况下，广泛开展GWAS仍是阐明复杂疾病遗传机制最为有效的途径和进行遗传机制研究必要的"原始积累"，也是应用基因组进行个体化预防的重要基础。

参 考 文 献

[1] Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science, 2002, 296: 2225-2229.

[2] McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet, 2008, 9: 356-369.

[3] Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med, 2010, 363: 166-176.

[4] Shen HB, Jin GF. Genome-wide association studies in cancer: current status and challenge. Chin J Prev, 2009, 43: 632-639. (in Chinese)
沈洪兵，靳光付. 肿瘤全基因组关联研究的现状与挑战. 中华预防医学杂志，2009，43：632-639.

[5] Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med, 2000, 343: 78-85.

[6] Kim ST, Cheng Y, Hsu FC, et al. Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic Variation. Prostate, 2010, 70: 1729-1738.

[7] Takata R, Akamatsu S, Kubo M, et al. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat Genet, 2010, 42: 751-754.

[8] Zheng SL, Sun J, Wiklund F, et al. Genetic variants and family history predict prostate cancer similar to prostate-specific antigen. Clin Cancer Res, 2009, 15: 1105-1111.

[9] Johansson M, Holmström B, Hinchliffe SR, et al. Combining 33 genetic variants with prostate-specific antigen for prediction of prostate cancer: Longitudinal study. Int J Cancer, 2011 Feb 15. doi: 10.1002/ijc.25986. [Epub ahead of print]

[10] Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst, 1999, 91: 1541-1548.

[11] Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst, 2008, 100: 1037-1041.

[12] Mealiffe ME, Stokowski RP, Rhees BK, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst, 2010, 102: 1618-1627.

[13] Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med, 2010, 362: 986-993.

[14] Zheng W, Wen W, Gao YT, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. J Natl Cancer Inst, 2010, 102: 972-981.

[15] McCarthy MI. Genomics, type 2 diabetes, and obesity. N Engl J Med, 2010, 363: 2339-2350.

[16] Herder C, Roden M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. Eur J Clin Invest, 2011, 41: 679-692.

[17] de Miguel-Yanes JM, Shrader P, Pencina MJ, et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care, 2011, 34: 121-125.

[18] Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet, 2011, 43: 333-338.

[19] Coronary Artery Disease(C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies

five new loci for coronary artery disease. Nat Genet, 2011, 43: 339-344.

[20] Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet, 2010, 376: 1393-1400.

[21] Goldstein DB. Common genetic variation and human traits. N Engl J Med, 2009, 360: 1696-1698.

[22] Kraft P, Hunter DJ. Genetic risk prediction—are we there yet? N Engl J Med, 2009, 360: 1701-1703.

[23] Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. N Engl J Med, 2009, 360: 1699-1701.

[24] Stefánsson K. Icelandic genetic database not at risk from bankruptcy. Nature, 2010, 463: 25.

[25] Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet, 2010, 42: 1118-1125.

[26] Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet, 2011, 43: 246-252.

[27] Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet, 2010, 42: 570-575.

[28] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature, 2009, 461: 747-753.

[29] Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet, 2010, 11: 446-450.

[30] Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet, 2010, 11: 415-425.

[31] Zeggini E. Next-generation association studies for complex traits. Nat Genet, 2011, 43: 287-288.

[32] Holm H, Gudbjartsson DF, Sulem P, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet, 2011, 43: 316-320.

[33] Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature, 2010, 464: 704-712.

[34] Stefansson H, Rujescu D, Cichon S, et al. Large recurrent microdeletions associated with schizophrenia. Nature, 2008, 455: 232-236.

[35] Glessner JT, Wang K, Cai G, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature, 2009, 459: 569-573.

[36] Greenway SC, Pereira AC, Lin JC, et al. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. Nat Genet, 2009, 41: 931-935.

[37] Pinto D, Pagnamenta AT, Klei L, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature, 2010, 466: 368-372.

[38] McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet, 2010, 40: 1166-1174.

[39] Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, et al. Genome-wide association study of CNVs in 16 000 cases of eight common diseases and 3000 shared controls. Nature, 2010, 464: 713-720.

[40] Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. Nature, 2011, 470: 59-65.

[41] Handsaker RE, Korn JM, Nemesh J, et al. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet, 2011, 43: 269-276.

[42] 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. Nature, 2010, 467(7319): 1061-1073.

[43] Thomas D. Gene — environment-wide association studies: emerging approaches. Nat Rev Genet, 2010, 11: 259-272.

[44] Hsu FC, Sun J, Zhu Y, et al. Comparison of two methods for estimating absolute risk of prostate cancer based on single nucleotide polymorphisms and family history. Cancer Epidemiol Biomarkers Prev, 2010, 19: 1083-1088.

[45] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med, 2008, 27: 157-172.

[46] Janssens AC, Ioannidis JP, van Duijn CM, et al. Strengthening the Reporting of Genetic Risk Prediction Studies: The GRIPS Statement. PLoS Med, 2011, 8: e1000420.

[47] Janssens AC, Ioannidis JP, Bedrosian S, et al. Strengthening the reporting of Genetic Risk Prediction Studies (GRIPS): explanation and elaboration. J Clin Epidemiol, 2011. [Epub ahead of print]

[48] Aly M, Wiklund F, Xu J, et al. Polygenic risk score improves prostate cancer risk prediction: results from the stockholm-1 cohort study. Eur Urol, 2011, 60: 21-28.

(收稿日期:2011-04-15)

(本文编辑:张林东)