

## 复杂疾病的遗传易感基因区域的精细定位

宋庆峰<sup>1,2,3,4</sup>, 张红星<sup>2,3,4</sup>, 马亦龙<sup>1</sup>, 周钢桥<sup>2,3,4</sup>

1. 广西医科大学附属肿瘤医院介入治疗科, 南宁 530021;

2. 军事医学科学院放射与辐射医学研究所, 北京蛋白质组学研究中心, 蛋白质组学国家重点实验室, 北京 102206;

3. 蛋白质药物国家工程研究中心, 北京 102206;

4. 国家蛋白质科学中心(北京), 北京 102206

**摘要:** 以单核苷酸多态性 (Single-nucleotide polymorphism, SNP) 为遗传标记, 采用全基因组关联研究 (Genome-wide association studies, GWAS) 的策略, 已经在 660 多种疾病 (或性状) 中发现了 3 800 多个遗传易感基因区域。但是, 其中最显著关联的遗传变异或致病性的遗传变异位点及其生物学功能并不完全清楚。这些位点的鉴定有助于阐明复杂疾病的生物学机制, 以及发现新的疾病标记物。后 GWAS 时代的主要任务之一就是精细定位研究找到复杂疾病易感基因区域内最显著关联的易感位点或致病性的易感位点并阐明其生物学功能。针对常见变异, 可通过推断或重测序增加 SNP 密度, 寻找最显著关联的 SNP 位点, 并通过功能元件分析、表达数量性状位点 (Expression quantitative trait locus, eQTL) 分析和单体型分析等方法寻找功能性的 SNP 位点和易感基因。针对罕见变异, 则可采用重测序、罕见单体型分析、家系分析和负荷检验等方法进行精细定位。文章对这些策略和所面临的问题进行了综述。

**关键词:** 全基因组关联研究; 精细定位; 遗传易感性; 单核苷酸多态性

## Fine mapping of complex disease susceptibility loci

SONG Qing-Feng<sup>1,2,3,4</sup>, ZHANG Hong-Xing<sup>2,3,4</sup>, MA Yi-Long<sup>1</sup>, ZHOU Gang-Qiao<sup>2,3,4</sup>

1 *Interventional Radiology Department of Affiliated Tumor Hospital of Guangxi Medical University, Nanning 530021, China;*

2 *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China;*

3 *National Engineering Research Center for Protein Drugs, Beijing 102206, China;*

4 *National Center for Protein Sciences in Beijing, Beijing 102206, China*

**Abstract:** Genome-wide association studies (GWAS) using single nucleotide polymorphism (SNP) markers have identified more than 3 800 susceptibility loci for more than 660 diseases or traits. However, the most significantly associated variants or causative variants in these loci and their biological functions have remained to be clarified.

收稿日期: 2013-05-08; 修回日期: 2013-09-24

基金项目: 国家自然科学基金项目 (编号: 81222027; 30901707; 30901231) 和北京市科技新星计划项目 (编号: 2010B006) 资助

作者简介: 宋庆峰, 硕士研究生, 专业方向: 复杂疾病的遗传易感性。Tel: 0771-5332738; E-mail: qfsong2010@hotmail.com

**通讯作者:** 周钢桥, 研究员, 教授, 博士生导师, 研究方向: 医学遗传学和基因组学。E-mail: zhougq114@gmail.com

These causative variants can help to elucidate the pathogenesis and discover new biomarkers of complex diseases. One of the main goals in the post-GWAS era is to identify the causative variants and susceptibility genes, and clarify their functional aspects by fine mapping. For common variants, imputation or re-sequencing based strategies were implemented to increase the number of analyzed variants and help to identify the most significantly associated variants. In addition, functional element, expression quantitative trait locus (eQTL) and haplotype analyses were performed to identify functional common variants and susceptibility genes. For rare variants, fine mapping was carried out by re-sequencing, rare haplotype analysis, family-based analysis, burden test, etc. This review summarizes the strategies and problems for fine mapping.

**Keywords:** genome-wide association study; fine mapping; genetic susceptibility; single nucleotide polymorphism

截止到 2012 年 4 月, 以单核苷酸多态性 (Single-nucleotide polymorphism, SNP) 为遗传标记, 采用全基因组关联研究 (Genome wide association studies, GWAS) 的策略已在 666 种疾病 (或性状) 中发现了 3 869 个显著关联 ( $P < 5.0 \times 10^{-8}$ ) 的遗传易感基因区域<sup>[1]</sup>。但是, 在这些区域内, 与复杂疾病最显著关联的遗传变异或致病性遗传变异都有待进一步确认, 其生物学功能也尚待深入阐明。当遗传易感基因区域内的 SNP 位点之间存在较强的连锁不平衡 (Linkage disequilibrium, LD) 以及存在遗传因素和环境因素交互作用时, 上述工作变得更加具有挑战性。后 GWAS 时代的主要任务之一是对复杂疾病易感区域内的致病性遗传变异进行精细定位 (Fine mapping), 即在通过 GWAS 鉴定到的疾病易感区域内获取高密度的遗传变异目录及其基因型, 从中鉴定出易感区域内最显著关联或致病性的遗传变异, 并阐明其生物学功能<sup>[2]</sup>。目前, 已出现一些系统性的策略用于复杂疾病的精细定位研究 (表 1)。

## 1 针对常见遗传变异 (Common variant) 的精细定位研究

### 1.1 确定最显著关联的 SNP 位点

SNP 是可遗传的变异中最常见的一种变异形式, 在群体中其次要等位频率 (Minor allele frequency, MAF) 大于 1%。目前, GWAS 采用的商业化 SNP 分型芯片已经可以同时检测 100 万个甚至更多的 SNP 位点。但是, 这些芯片仍远未能覆盖人类基因组中的全部 SNP 位点, 一些与复杂性状最显著关联的 SNP 位点可能会被遗漏。因此, 获得易感基因区域内高密度的 SNP 目录是进行精细定位的前提之一。可以通过以下两种方法增加易感基因区域内的 SNP 密度, 然后再进行遗传关联分析, 以确定最显著关联的 SNP 位点。

#### 1.1.1 根据参考数据集进行 SNP 的推断 (Imputation)

由许多国家共同参与的“人类基因组单体型图计划”（HapMap 计划）和“千人基因组计划”（1000 genome project）为人们提供了比较全面的人类基因组 DNA 序列变异数据。以这些研究计划产生的 SNP 数据为参考集，可以通过计算推断出与已分型 SNP 位点相邻的未

分型 SNP 位点的基因型,从而大大降低遗漏的可能性<sup>[3]</sup>。用于推断的代表性软件有 MACH<sup>[4]</sup>和 IMPUTE<sup>[4]</sup>等,现已得到广泛应用。例如, Raychaudhuri 等<sup>[5]</sup>以 2 767 个个体的全基因组数据为参考集,对 6 个独立的类风湿性关节炎的 GWAS 数据集进行推断,并对推断结果进行了遗传关联分析,结果在 MHC 区域发现了 5 个新的与疾病显著关联的 SNP 位点 ( $P < 1.0 \times 10^{-550}$ )。Peters 等<sup>[6]</sup>在 20 488 个非洲裔美国人中,对体重指数 (Body mass index, BMI) 的候选易感基因区域——*FTO* 基因所在的基因组区域 (646 kb) 进行推断和关联分析,结果鉴定出一个新的与体重指数显著关联的 SNP 位点 rs56137030 ( $P = 8.3 \times 10^{-6}$ )。Liu 等<sup>[7]</sup>在来自英国的 2 861 例原发性胆汁性肝硬化患者 (Primary biliary cirrhosis, PBC) 和 8 514 例对照中,对约 20 万个多态性位点进行分型、推断和关联分析,结果新发现了 3 个与 PBC 显著关联的 SNP 位点 ( $P < 5.0 \times 10^{-8}$ )。以往在日本人群中开展的 GWAS 研究发现, *MICA* 基因上的 rs2596542 与丙型肝炎 (Hepatitis C virus, HCV) 相关的肝细胞癌 (Hepatocellular carcinoma, HCC) 显著关联。Lange 等<sup>[8]</sup>进一步在瑞士人群中对此区域进行了分型和推断,关联分析结果表明, *MICA* 基因上游 *HCP5* 基因上的 rs2244546 是一个新的疾病易感性标志物。

SNP 的推断基于单体型结构和单体型内各 SNP 位点之间的 LD 程度,参考集的样本量越大,推测的成功率和准确率就越高。但是在现阶段,各项研究所能提供的参考集的样本量均较小。另外,有些 SNP 位点在人群中位频率较低 ( $MAF < 2\%$ ),且与邻近的已被芯片检测到的 SNP 位点并不存在较强的 LD ( $r^2 < 0.8$ ),从而无法被成功推断出来<sup>[9]</sup>。因此,用推断这种方法无法彻底解决最显著关联的 SNP 被遗漏的情况。

### 1.1.2 对目标区域进行重测序 (Target region resequencing)

对 GWAS 鉴定的易感基因区域进行重测序,将有助于全面了解一个区域内所有的遗传变异信息,并发现与复杂性状最显著关联的 SNP 位点。此外,还可以发现一些新的功能性罕见变异 (Rare variation)<sup>[10]</sup>或结构变异,如拷贝数变异 (Copy number variation, CNV)<sup>[11]</sup>和小的插入-缺失变异 (Indel) 等<sup>[12]</sup>,从而鉴定到易感基因区域内与复杂性状最显著关联的或致病性的其他各类遗传变异形式。近年来,高通量的二代测序技术得以迅速发展,大大降低了测序的费用及缩短了测序的时间,使得在大量样本中进行目标区域的重测序成为现实。例如, Xiang 等<sup>[13]</sup>对已知的高原低氧适应基因 *EGLN1* 所在的基因组区域 (长约 59.4 kb) 进行重测序,新发现一个非同义突变 (rs186996510, D4E) 的频率在高原藏族人群和低海拔汉族人群中存在显著差异 ( $F_{st} = 0.709$ ),提示 rs186996510 可能是高原低氧适应的成因性 SNP。在对镰刀状红细胞贫血的易感基因重测序研究中,除了发现新的显著关联的常见变异位点外,还在 *MYB* 基因上发现了 3 个新的与疾病显著关联 ( $P = 5.0 \times 10^{-3}$ ) 的罕见遗传变异<sup>[14]</sup>。

**表 1 精细定位的研究策略**

变异类型	研究策略	优势	不足	参考文献
常见变异	推断	成本较低，方便快捷	低频变异 (<2%) 的推断准确率较低； 不能完全解决遗漏问题	[3~9]
	重测序	能够覆盖所有遗传变异，没有遗漏	测序成本较高	[10~14]
	基因调控元件分析	能够迅速确定功能性遗传变异和易感基因	一般仅能针对已知的功能元件进行分析	[16~23]
	eQTL 分析	能够在全基因组范围内确定候选易感基因	研究人群可能存在抽样偏倚； 基因表达谱的成本较高	[24~28]
	单体型分析	弥补了单个 SNP 的信息量不足； 跨种族分析有助于迅速确定候选易感基因	功能验证比较困难； 不同种族人群的招募比较困难	[2, 14, 29~32]
罕见变异	重测序	能够迅速确定功能性遗传变异和易感基因	测序成本较高； 需要较大的研究人群	[35~37]
	罕见单体型分析	能够迅速确定功能性遗传变异和易感基因	前期分析需要有大样本量的 GWAS 数据	[38, 39]

家系分析	提高了罕见变异的检测效能; 有助于理解疾病的遗传模式	家系的招募比较困难; 需要较大的研究人群	[40]
罕见变异的累积效应	弥补了单个罕见变异的信息量不足	检测成本较高; 功能验证比较困难	[41]

---

## 1.2 通过 SNP 位点筛选候选易感基因

确定了与复杂疾病等性状最显著关联的遗传变异后，还要阐明其生物学功能。通过 GWAS 鉴定到的最显著关联的 SNP 位点大多位于内含子、基因间区等非编码区域内，这对于直接阐明易感 SNP 位点的生物学功能造成了一定的困难。研究显示，这些 SNP 位点主要通过改变相关基因的转录表达水平或者转录本的剪接等方式影响疾病的发生风险<sup>[15]</sup>。

### 1.2.1 通过调控元件分析寻找功能性的 SNP 位点和易感基因

启动子、沉默子和绝缘子等基因表达调控元件，能够通过正调控<sup>[16]</sup>或负调控<sup>[17]</sup>的机制来调节相关基因的表达量。而组蛋白、非编码 RNA 等表观调控元件，能够通过末端修饰<sup>[18]</sup>、自身活性的改变<sup>[19]</sup>或量的改变<sup>[20]</sup>等机制来调节相关基因的表达量。因此，位于各类调控元件序列中的与疾病易感性显著关联的遗传变异，有可能通过自身等位型的变化影响这些调控元件的调控机制而影响相关基因的表达水平，从而影响疾病发生的风险<sup>[21]</sup>。例如，Pomerantz 等<sup>[22]</sup>研究发现，结直肠癌的易感基因区域 8q24 中最显著关联的 SNP 位点是 rs6983267，该 SNP 位于 *MYC* 基因的增强子序列内，其风险等位型能够增加 *MYC* 基因的增强子与转录因子 TCF7L2 的结合，导致 *MYC* 基因的表达量增加，从而增加结直肠癌的发生风险。Peters 等<sup>[6]</sup>在 *FTO* 基因所在基因组区域发现了一个新的与体重指数显著关联的 SNP 位点 rs56137030 ( $P = 8.3 \times 10^{-6}$ )，并进一步发现有多与之呈强 LD 的 SNP 位点处于调控元件之中，其中 rs1421085 在转录因子 CUX1 的结合上具有等位特异性。也有研究发现，某些遗传变异可以影响基因网络内不同中枢节点之间的联络，进而影响下游基因的表达量，最终影响疾病的发病风险<sup>[23]</sup>。

### 1.2.2 通过 eQTL 分析寻找功能性的 SNP 位点和易感基因

在人类基因组上，能够影响基因的 mRNA 或蛋白表达水平的遗传变异位点称为表达数量性状位点 (Expression quantitative trait locus, eQTL)，检测遗传变异与 mRNA 或蛋白质表达量之间是否有关联的统计分析，称为 eQTL 分析<sup>[24]</sup>。对 GWAS 鉴定到的易感基因区域内的 SNP 位点进行 eQTL 分析，有可能发现直接影响相关基因表达、进而改变疾病发生风险的功能性 SNP 位点。按照功能性 SNP 位点与受其调控的基因之间的距离，可以分为顺式 (cis-eQTL, 1 Mb 范围内) 和反式 (trans-eQTL, 1 Mb 范围之外) 调控两种类型<sup>[25]</sup>。在以往的研究中，eQTL 分析大多是在淋巴细胞系中进行，例如 Morley 等<sup>[26]</sup>在 14 个大家系的永生 B 细胞中发现，近 1 000 个 SNPs 能够作为顺式或者反式 eQTL 位点调节 3 554 个基因的表达量。但是，在复杂疾病的研究中，eQTL 分析更应该在与疾病对应的特定组织中进行<sup>[27]</sup>。例如，在肝脏组织中，对高脂血症的易感基因区域 1p13 开展的 eQTL 分析显示，最显著关联的 SNP 位点 (rs12740374,  $P < 1.0 \times 10^{-40}$ ) 的风险等位型能增强 *SORT1* 基因启动子与转录因子 C/EBP 的结合，从而增加该基因在肝脏中的表达，使肝脏中极低密度脂蛋白 (Very low-density lipoprotein, VLDL) 的分泌增加，进而增加血清中低密度脂蛋白 (Low-density lipoprotein cholesterol, LDL) 和 VLDL 的浓度，最终增加高脂血症的发生风险<sup>[28]</sup>。

### 1.3 通过确定显著关联的单体型和跨种族的单体型对比来确定候选易感基因

单体型是指在后代个体中没有发生重组的祖先染色体片段。由于构成单体型的 SNP 位点都位于同一条染色体上的某一区域内，且各位点之间具有一定程度的 LD，所以有时难以通过独立性检验判定哪个 SNP 位点与疾病更具有关联性。在这种情况下，这些 SNP 位点可能以单体型的形式与疾病相关联<sup>[2]</sup>。例如，Galameau 等<sup>[14]</sup>在对镰刀状红细胞贫血的研究中，发现 rs7599488 和 rs10189857 ( $r^2=0.96$ ) 与血红蛋白浓度之间存在显著关联；进一步的单体型分析发现，与既往研究中所发现的关联位点 rs4671393 ( $P=3.7\times 10^{-37}$ ) 相比，rs7599488、rs10189857 和 rs4671393 构成的单体型具有更加显著的关联程度 ( $P=4.0\times 10^{-45}$ )，从而提示这 3 个 SNP 位点通过单体型的形式共同在镰刀状红细胞贫血的易感性上发挥作用。

在确定了单体型与疾病的相关性后，如何确定候选易感基因是极其关键的一步。对于一些长度较短、只包含有一个基因的单体型而言，可直接将该基因确定为候选易感基因。对于长度较长、包含有多个基因的单体型而言，可以尝试在不同种族人群之间比较 SNP 位点之间 LD 程度的差异来定位候选易感基因。

HapMap 计划和千人基因组计划的单体型数据显示，同一基因组区域内的单体型结构在不同种族之间具有差异性。非洲人的单体型比其他种族的单体型更短，原因在于非洲人有更长的历史，从而有更多的重组来打破原有单体型的结构，从而形成新的单体型。同时，大量 GWAS 显示，同种疾病在不同种族之间具有共同的易感区域<sup>[29]</sup>。例如，针对发作性睡病 (Narcolepsy) 的 GWAS 显示，19q13.2 区域是高加索人、亚洲人和非洲裔美国人 3 个不同种族共同的易感基因区域<sup>[30]</sup>。6q23 区域是高加索人和亚洲人共有的系统性红斑狼疮的易感基因区域，进一步对比两个种族中该区域的风险单体型结构，发现其中有连续 6 个 SNP 位点的基因型完全一致。因此，推测这 6 个位点所构成的单体型可能为两种人群共有的风险单体型，据此可将易感基因区域缩小至 48.5 kb，并确定了 *TNFAIP3* 是系统性红斑狼疮的候选易感基因<sup>[31]</sup>。最近，Wu 等<sup>[32]</sup>通过在不同种族人群中对血脂水平的候选易感基因进行精细定位分析，成功地将 *GCKR*、*PPP1R3B*、*ABO*、*LCAT* 和 *ABCA1* 等易感基因的致病性位点进一步缩小了范围，例如发现 *GCKR* 基因中的功能性变异 rs1260326 (P446L) 可能是甘油三酯水平的成因性 SNP。因此，在不同种族中比较同一易感基因区域的单体型结构，有利于缩小易感基因的范围并最终确定候选易感基因。

## 2 针对罕见遗传变异 (Rare variant) 的精细定位研究

GWAS 常常是基于“常见疾病-常见变异”的假说而开展的。但是基于“常见疾病-罕见变异”假说开展的研究显示，基因编码区内新近发生的罕见遗传变异 ( $MAF<1\%$ ) 中富集了较多的有害变异，因此也能影响复杂疾病发生的风险<sup>[33]</sup>。例如，Dickson 等<sup>[34]</sup>对耳聋 GWAS 鉴定到的易感区域进行重测序研究，发现该易感区域内的罕见变异同样可影响疾病的患病风



险。Azzopardi 等<sup>[35]</sup>对结直肠肿瘤的易感基因 *APC* 进行重测序后发现,在未携带已知常见风险基因型的个体中,基因上多个罕见变异与患病风险具有显著的相关性 ( $P = 1.7 \times 10^{-2}$ )。

由于罕见变异产生的时间较短,在人群中的频率很低,想要有效的发现这些变异需要比发现常见变异更大的样本量和更多的经费<sup>[36]</sup>,这都极大地限制了对罕见变异的研究。例如,在一项对炎症性肠病的 70 个已知易感基因的重测序研究中,在第一阶段对 112 个病例和 112 个对照进行重测序后,未发现与疾病发生风险显著关联的罕见变异,于是在第二阶段扩大样本量,对 896 个病例和 1 216 个对照进行重测序,最终在 *IL23R* 基因上发现了 p.Arg86Gln、p.Gly149Arg 和 p.Val362Ile 与疾病发生风险显著关联的罕见变异<sup>[37]</sup>。

最近的研究结果显示了罕见变异的一些新规律,包括下文中提及的罕见单体型携带罕见变异、罕见变异在疾病家系中具有较高的频率和多个罕见变异具有累积效应等。针对这些规律制定的研究策略不但有效地提高了检测罕见变异的检验效能,而且减少了需要研究的样本量。

## 2.1 罕见单体型携带罕见变异

DNA 序列中新的罕见变异可以与常见变异一起构成新的单体型,这些单体型由于产生的时间较短,没有足够的时间传播,所以在人群中的频率小于 1%。对与疾病相关联的由常见变异构成的罕见单体型进行重测序,有可能在这些罕见单体型上发现致病的功能性罕见变异<sup>[38]</sup>。例如, Raychaudhuri 等<sup>[39]</sup>对老年性黄斑变性 GWAS 鉴定的易感区域进行深入研究,发现在对照组中频率为 0.048% 的 H5 单体型 (*CFH* 基因)能增加疾病的患病风险。对该易感区域内携带不同结构单体型 的 84 个个体进行重测序后发现,携带 H5 单体型 的 6 个个体在 *CFH* 基因的 22 号外显子上均有一个能够直接改变氨基酸的罕见变异 R1210C (在人群中的频率小于 0.1%),而在其他携带非 H5 单体型的个体中均未发现该变异。通过上述研究策略, Raychaudhuri 等成功地从罕见单体型上发现了一个与疾病易感性显著关联的罕见变异 ( $P = 9.4 \times 10^{-3}$ )。

## 2.2 罕见变异在疾病家系中具有较高的频率

在家系中,各成员来源于共同的祖先,其中某一患者携带的致病变异可能会传递给下一代子女,并导致他们患病。因此,在普通人群中频率较低的致病性罕见变异,在家系患病个体中会具有较高的频率,从而更容易被发现并了解其遗传模式。例如, Ewing 等<sup>[40]</sup>在对前列腺癌易感区域 17q21-22 的研究中,发现与前列腺癌发生风险相关的 rs138213197 在 5 083 个患者中的等位频率为 1.4%,在 1 401 个对照个体中的频率为 0.1%。但是,在 4 个前列腺癌的大家系中该变异的频率为 34%,在携带该变异的成员中有 82% 的个体患病,大大高于在随机人群中的比例。上述研究结果显示,基于家系的研究策略在精细定位研究中可能更具有优势。

### 2.3 多个罕见变异具有累积效应

因为罕见遗传变异的发生具有随机性,所以易感区域内与疾病发生风险相关联的罕见变异在不同的患病个体中可能发生在同一易感基因的不同外显子上,对于这些罕见变异,主要通过负荷检验(Burden test)来鉴定其所在的易感基因,同时鉴定它们与疾病发生风险的关联程度<sup>[2]</sup>。例如,对高脂血症的易感基因 *ABCA1*、*APOA1* 和 *LCAT* 进行重测序,对发现的多个位于不同外显子的新的罕见变异进行了负荷检验,显示与疾病发生风险显著关联 ( $P < 1.0 \times 10^{-4}$ ) 的罕见变异主要富集在 *ABCA1* 基因的不同外显子上,这些罕见变异能够直接改变 *ABCA1* 基因编码的蛋白质,进而降低血清中高密度脂蛋白的浓度,最终增加高脂血症的发病风险<sup>[41]</sup>。

## 3 结语

以 SNP 为遗传标记,采用 GWAS 的研究策略成功地发现了许多复杂疾病及其他性状的易感基因区域。但是,目前也面临着一些巨大挑战,包括缺乏快速、准确和可重复使用的方法用于从这些易感基因区域中精确定位疾病的致病位点或致病基因,以及缺乏简单、流程化的功能研究方案用于阐明致病位点的生物学功能,这将是今后研究工作的瓶颈<sup>[2]</sup>。因此,应用新的基因组序列检测技术(如高通量测序)和采用更为有效的分析方法,在 GWAS 鉴定出的易感区域中精确定位与疾病发生风险最显著关联的或致病性的遗传变异,同时采用快速和流程化的功能验证实验来阐明其生物学功能,是后 GWAS 时代精细定位研究的主要内容之一。

目前,针对常见变异的精细定位研究比较多。这类研究主要通过推断或重测序增加 SNP 密度,寻找最显著关联的 SNP 位点,并通过功能元件分析、eQTL 分析和单体型分析等方法寻找功能性的 SNP 位点和易感基因。随着高通量测序成本的迅速降低,以及基因组功能元件的全面阐释(例如 ENCODE 计划)<sup>[42]</sup>,预计今后针对常见变异的精细定位研究将会更多的发现。另一方面,由于未受遗传净化选择的制约和具有潜在的致病性功能<sup>[43]</sup>,罕见变异在遗传易感性中的作用在近年来受到越来越多的关注。今后研究的重要方向之一,将是在通过常见变异鉴定的候选易感基因组区域内寻找致病性的罕见变异和易感基因。进一步对所定位的常见变异或罕见变异进行后续的功能验证,将是精细定位研究的关键所在。只有充分理解了这些变异的生物学意义,才能推动对人类复杂疾病或性状的遗传机制的全面认识。

此外,采用 DNA 序列的保守性<sup>[44]</sup>、基因的调控网络<sup>[45]</sup>和染色质的结构状态<sup>[46]</sup>等策略对易感区域内的遗传变异进行研究,可以作为上述研究策略的有益补充。另有研究显示,同义突变虽然不改变编码的氨基酸,但是有可能通过三联体核苷酸影响蛋白质的合成速率,从而影响疾病的发生风险<sup>[47]</sup>。这提示,同义突变也可能是今后精细定位疾病易感基因研究领域的一个全新方向。

## 参考文献(References):

- [1] Asking for more. *Nat Genet*, 2012, 44(7): 733.
- [2] Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. *Cell*, 2011, 147(1): 57-69.
- [3] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 2010, 11(7): 499-511.
- [4] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 2009, 5(6): e1000529.
- [5] Raychaudhuri S, Sandor C, Stahl E A, Freudenberg J, Lee H S, Jia X, Alfredsson L, Padyukov L, Klareskog L, Worthington J, Siminovitch KA, Bae SC, Plenge RM, Gregersen PK, De Bakker PI. Five amino acids in three hla proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*, 2012, 44(3): 291-296.
- [6] Peters U, North KE, Sethupathy P, Buyske S, Haessler J, Jiao S, Fesinmeyer MD, Jackson RD, Kuller LH, Rajkovic A, Lim U, Cheng I, Schumacher F, Wilkens L, Li R, Monda K, Ehret G, Nguyen KD, Cooper R, Lewis CE, Leppert M, Irvin MR, Gu CC, Houston D, Buzkova P, Ritchie M, Matise TC, Le Marchand L, Hindorff L A, Crawford DC, Haiman CA, Kooperberg C. A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet*, 2013, 9(1): e1003171.
- [7] Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, Ducker SJ, Day DB, Heneghan MA, Neuberger JM, Donaldson PT, Bathgate AJ, Burroughs A, Davies MH, Jones DE, Alexander GJ, Barrett JC, Sandford RN, Anderson CA. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet*, 2012, 44(10): 1137-1141.
- [8] Lange CM, Bibert S, Dufour JF, Cellerai C, Cerny A, Heim MH, Kaiser L, Malinverni R, Mullhaupt B, Negro F, Semela D, Moradpour D, Kutalik Z, Bochud PY. Comparative genetic analyses point to HCP5 as susceptibility locus for HCV-associated hepatocellular carcinoma. *J Hepatol*, 2013 , 59(3):504-509.
- [9] Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 2012, 44(8): 955-999.
- [10] Ji W, Foo JN, O'roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*, 2008, 40(5): 592-599.
- [11] Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo

- JS. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet*, 2010, 42(5): 400-405.
- [12] Yu TW, Mochida GH, Tischfield DJ, Sgaier SK, Flores-Sarnat L, Sergi CM, Topcu M, McDonald MT, Barry BJ, Felie JM, Sunu C, Dobyns WB, Folkerth RD, Barkovich AJ, Walsh CA. Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. *Nat Genet*, 2010, 42(11): 1015-1020.
- [13] Xiang K, Ouzhuluobu, Peng Y, Yang Z, Zhang X, Cui C, Zhang H, Li M, Zhang Y, Bianba, Gonggalanzi, Basang, Ciwangsangbu, Wu T, Chen H, Shi H, Qi X, Su B. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol*, 2013, 30(8): 1889-1898.
- [14] Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet*, 2010, 42(12): 1049-1051.
- [15] Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet*, 2011, 43(6): 513-518.
- [16] Levine M, Tjian R. Transcription regulation and animal diversity. *Nature*, 2003, 424(6945): 147-151.
- [17] Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, Miclaus K, Jadallah SJ, Goldstein DB, Wolfinger RD, Gibson G. Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet*, 2010, 42(1): 62-67.
- [18] Kouzarides T. Chromatin modifications and their function. *Cell*, 2007, 128(4): 693-705.
- [19] Raval A, Tanner SM, Byrd JC, Angerman EB, Perko JD, Chen SS, Hackanson B, Grever MR, Lucas DM, Matkovic JJ, Lin TS, Kipps TJ, Murray F, Weisenburger D, Sanger W, Lynch J, Watson P, Jansen M, Yoshinaga Y, Rosenquist R, De Jong PJ, Coggill P, Beck S, Lynch H, De La Chapelle A, Plass C. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell*, 2007, 129(5): 879-890.
- [20] Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, Van De Vijver MJ, Sukumar S, Chang HY. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 2010, 464(7291): 1071-1076.
- [21] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 2009, 457(7231): 854-858.
- [22] Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R,

- Coetzee GA, Freedman ML. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet*, 2009, 41(8): 882-884.
- [23] Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu K S, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 2006, 38(11): 1341-1347.
- [24] Damerval C, Maurice A, Josse JM, de Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics*, 1994, 137(1): 289-301.
- [25] Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet*, 2006, 7(11): 862-872.
- [26] Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 2004, 430(7001): 743-747.
- [27] Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, Guhathakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, Van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 2008, 6(5): e107.
- [28] Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B, Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Lund-Katz S, Phillips MC, Wong J, Cantley W, Racie T, Ejebe KG, Orho-Melander M, Melander O, Koteliansky V, Fitzgerald K, Krauss RM, Cowan CA, Kathiresan S, Rader DJ. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 2010, 466(7307): 714-719.
- [29] Manku H, Langefeld CD, Guerra SG, Malik TH, Alarcon-Riquelme M, Anaya JM, Bae SC, Boackle SA, Brown EE, Criswell LA, Freedman BI, Gaffney PM, Gregersen PA, Guthridge JM, Han SH, Harley JB, Jacob CO, James JA, Kamen DL, Kaufman KM, Kelly JA, Martin J, Merrill JT, Moser KL, Niewold TB, Park SY, Pons-Estel BA, Sawalha AH, Scofield RH, Shen N, Stevens AM, Sun C, Gilkeson GS, Edberg JC, Kimberly RP, Nath SK, Tsao BP, Vyse TJ. Trans-ancestral studies fine map the SLE-susceptibility locus TNFSF4. *PLoS Genet*, 2013, 9(7): e1003554.
- [30] Kornum BR, Kawashima M, Faraco J, Lin L, Rico TJ, Hesselton S, Axtell RC, Kuipers H, Weiner K, Hamacher A, Kassack MU, Han F, Knudsen S, Li J, Dong X, Winkelmann J, Plazzi G, Nevsimalova S, Hong SC, Honda Y, Honda M, Hogg B, Ton TG, Montplaisir J, Bourgin P, Kemlink D, Huang YS, Warby S, Einen M, Eshragh JL, Miyagawa T, Desautels A, Ruppert E, Hesla PE, Poli F, Pizza F, Frauscher B, Jeong JH, Lee SP, Strohl KP, Longstreth

- WT, Jr., Kvale M, Dobrovolna M, Ohayon MM, Nepom GT, Wichmann HE, Rouleau GA, Gieger C, Levinson DF, Gejman PV, Meitinger T, Peppard P, Young T, Jennum P, Steinman L, Tokunaga K, Kwok PY, Risch N, Hallmayer J, Mignot E. Common variants in P2RY11 are associated with narcolepsy. *Nat Genet*, 2011, 43(1): 66-71.
- [31] Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, Bates JS, Hu Y, Kelly JA, Kaufman KM, Guthridge JM, Alarcon-Riquelme ME, Anaya JM, Bae SC, Bang SY, Boackle SA, Brown EE, Petri MA, Gallant C, Ramsey-Goldman R, Reveille JD, Vila LM, Criswell LA, Edberg JC, Freedman BI, Gregersen PK, Gilkeson GS, Jacob CO, James JA, Kamen DL, Kimberly RP, Martin J, Merrill JT, Niewold TB, Park SY, Pons-Estel BA, Scofield RH, Stevens AM, Tsao BP, Vyse TJ, Langefeld CD, Harley JB, Moser KL, Webb CF, Humphrey MB, Montgomery CG, Gaffney PM. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. *Nat Genet*, 2011, 43(3): 253-258.
- [32] Wu Y, Waite LL, Jackson AU, Sheu WH, Buyske S, Absher D, Arnett DK, Boerwinkle E, Bonnycastle LL, Carty CL, Cheng I, Cochran B, Croteau-Chonka DC, Dumitrescu L, Eaton CB, Franceschini N, Guo X, Henderson BE, Hindorff LA, Kim E, Kinnunen L, Komulainen P, Lee WJ, Le Marchand L, Lin Y, Lindstrom J, Lingaas-Holmen O, Mitchell SL, Narisu N, Robinson JG, Schumacher F, Stancakova A, Sundvall J, Sung YJ, Swift AJ, Wang WC, Wilkens L, Wilsgaard T, Young AM, Adair LS, Ballantyne CM, Buzkova P, Chakravarti A, Collins FS, Duggan D, Feranil AB, Ho LT, Hung YJ, Hunt SC, Hveem K, Juang JM, Kesaniemi AY, Kuusisto J, Laakso M, Lakka T A, Lee IT, Leppert MF, Matisse T C, Moilanen L, Njolstad I, Peters U, Quertermous T, Rauramaa R, Rotter JI, Saramies J, Tuomilehto J, Uusitupa M, Wang TD, Boehnke M, Haiman CA, Chen YD, Kooperberg C, Assimes TL, Crawford DC, Hsiung CA, North KE, Mohlke KL. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet*, 2013, 9(3): e1003379.
- [33] Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 2008, 40(6): 695-701.
- [34] Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Bio*, 2010, 8(1): e1000294.
- [35] Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, Colley J, Moskvina V, Frye C, Sampson JR, Wenstrup R, Scholl T, Cheadle JP. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res*, 2008, 68(2): 358-363.
- [36] Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 2010, 11(11): 773-785.
- [37] Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, Cleynen I, Colombel JF, De Rijk P, Dewit O, Finkel Y, Gassull MA, Goossens D, Laukens D, Lemann M, Libioulle C, O'morain C, Reenaers C, Rutgeerts P, Tysk C, Zelenika D, Lathrop M,

- Del-Favero J, Hugot JP, De Vos M, Franchimont D, Vermeire S, Louis E, Georges M. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat Genet*, 2011, 43(1): 43-47.
- [38] Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am J Hum Genet*, 2010, 86(5): 730-742.
- [39] Raychaudhuri S, Iartchouk O, Chin K, Tan PL, Tai AK, Ripke S, Gowrisankar S, Vemuri S, Montgomery K, Yu Y, Reynolds R, Zack DJ, Campochiaro B, Campochiaro P, Katsanis N, Daly MJ, Seddon JM. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet*, 2011, 43(12): 1232-1236.
- [40] Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, Bizon C, Yan G, Gielzak M, Partin AW, Shanmugam V, Izatt T, Sinari S, Craig DW, Zheng SL, Walsh PC, Montie JE, Xu J, Carpten JD, Isaacs WB, Cooney KA. Germline mutations in HOXB13 and prostate-cancer risk. *N Engl J Med*, 2012, 366(2): 141-149.
- [41] Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, Mcpherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 2004, 305(5685): 869-872.
- [42] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489(7414): 57-74.
- [43] Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. *Cell*, 2011, 147(1): 32-43.
- [44] Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 2000, 288(5463): 136-140.
- [45] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Fietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 2012, 489(7414): 91-100.
- [46] Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*, 2012, 489(7414): 109-113.
- [47] Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 2007, 315(5811): 525-528.