



中国科学院大学

University of Chinese Academy of Sciences

## 研究生学位论文开题报告

报告题目 基于自闭症谱系障碍 GWAS 分析  
的相关生物学作用探讨

学生姓名 刘可钦 学号 20202801923006

指导教师 孔祥银 职称 研究员

学位类别 理学硕士

学科专业 遗传学

研究方向 疾病易感基因识别和功能研究

培养单位 中科院上海营养与健康研究所

填表日期 2021 年 10 月 27 日

中国科学院大学制

## 基于自闭症谱系障碍 GWAS 分析的相关生物学作用探讨

**摘要** 自闭症谱系障碍 (ASD) 是一种复杂的异质性疾病, 其发病由遗传、神经、免疫和环境因素相互作用而共同导致, 美国国家卫生统计中心 (NCHS) 在 2014 年进行的全国健康访谈调查显示, 每 45 名 3 至 17 岁的儿童中就有 1 名被诊断为 ASD。但是, 尽管最近的科学进展揭示了 ASD 的分子因子和生物机制, 有助于发现和验证其致病基因, 但确切的致病因素和基因互作网络仍然难以捉摸, 也没有出现关于 ASD 分子病理学的统一假说。在这篇研究中, 我们基于 GBA 工具的筛选结果确定了 8、17 和 20 号染色体上与 ASD 相关的风险位点; 同时还确定 ASD 基因水平上的部分关联; 另外, 当前研究成果展示了 ASD 相关基因在不同脑区和不同身体组织中的表达情况, 同时定位了部分风险基因座, 此外, 这篇报告还提出了 ASD 基因潜在的互作网络与其参与的生物学过程的理解, 特别是与神经元功能有关的见解。

**Abstract** Autism spectrum disorder (ASD) is a complex and heterogeneous disorder whose pathogenesis is caused by a combination of genetic, neurological, immune, and environmental factors, and the National Health Interview Survey conducted by the National Center for Health Statistics (NCHS) in 2014 revealed that 1 in 45 children aged 3 to 17 years were diagnosed with ASD. However, despite recent scientific advances that have revealed ASD molecular factors and biological mechanisms that contribute to the discovery and validation of its causative genes, the exact causative factors and gene interaction networks remain elusive, and no unifying hypothesis on the molecular pathology of ASD has emerged. In this study, we identify risk loci associated with ASD on chromosomes 8, 17 and 20 based on the screening results of the GBA tool; we also identify partial associations at the ASD gene level; in addition, the current findings demonstrate the expression of ASD-associated genes in different brain regions and different body tissues, while localizing some risk motifs, and in addition, this article proposes insights into the potential reciprocal networks of ASD genes and the biological processes they are involved in, particularly in relation to neuronal function.

## 1 研究背景

自人类历史有记载以来, 人类社群就倾向于排斥或忽视那些在身体或精神层面偏离正常标准的人。这一倾向多数来自社群内部并无科学依据的神话和信仰, 以及未知本身。自闭症谱系障碍患者因其独特的神经发育与古怪的言行举止, 曾经被视为这类边缘人群中的一部分。而自闭症本身也被视为是一种具有极端多变表型的神经发育综合征, 从而被认为是人类发展的障碍。事实上, 自闭症谱系障碍 (Autism Spectrum Disturbances, ASD) 是一种早发性的神经发育障碍, 其特点是大脑功能紊乱和行为缺陷, 核心症状是互惠社会化受损, 沟通技能受损, 以及重复或限制性兴趣和行为。其症状一般开始于儿童早期, 并持续到成年, 严重影响患者的正常生活 (Mattila et al. 2011)。这种终生疾病是全世界发展最快的发育障碍之一, 其中男性患者比女性多出 4 倍, 其发病率仍在以惊人的速度增长 (Zablotsky et al. 2015)。2005 年前的大规模调查估计发病率约为 1-2%。美国疾病控制和预防中心 (CDC) 在 2013 年指出, 每 68 个美国儿童中就有一个患有 ASD (Blumberg et al. 2013)。而美国国家卫生统计中心 (NCHS) 在 2014 年进行的全国健康访谈调查发现, 每 45 名 3 至 17 岁的儿童中就有 1 名被诊断为自闭症 (Christensen et al. 2018)。现在, 随着 ASD 发病率在人群中日渐上升, 而确诊患者仍然缺少有效的治疗药物, 同时对患者进行教育和护理的高昂成本正在成为患者家庭的严重经济负担, 这种神经发育谱系障碍正在成为公共卫生领域的一个焦点问题 (Jobski et al. 2017)。

显然, ASD 是一种复杂的异质性疾病, 其发病由遗传、神经、免疫和环境因素相互作用而共同导致 (Matelski and Van de Water 2016), 同时也具有高度复杂的遗传性, 因为罕见和常见的遗传变异都会导致 ASD 风险 (Lai, Lombardo, and Baron-Cohen 2013), 这为 ASD 的基因预测和基因诊断提供了理论基础。但是, 由于 ASD 的遗传异质性和复杂性、以及行为特征和认知特征的巨大差异使得定义特定的遗传风险成分具有挑战性 (Yoo 2015)。因此, 尽管最近的科学进展揭示了 ASD 的分子因子和生物机制, 有助于发现和验证其致病基因 (Geschwind and State 2015), 但确切的致病因素和基因互作网络仍然难以捉摸, 也没有出现关于 ASD 分子病理学的统一假说, 这也导致对尚未表现出明显症状的 ASD 患者进行基因预测和诊断较为困难 (Diaz-Beltran et al. 2017)。但是, 一些大规模的临床研究已经证实了与 ASD 相关的高并发症率。这些合并症代表了额外的疾病负担 (Leyfer et al. 2006)。事实上, 约有 70% 以上被诊断为自闭症的人同时患有其他医学疾病, 其频率明显高于神经发育正常谱系人群 (Lai, Lombardo, and Baron-Cohen 2013)。其中一些疾病, 如癫痫或抑郁症, 可以在 ASD 患者的青春期甚至生命后期首次出现, 加重了对患者终身的损害。据报道, 近 45% 的自闭症患者同时受到智力障碍的影响, 28-44% 被诊断为注意力缺陷多动症, 12-70% 有临床抑郁症, 8-30% 的 ASD 患者有癫痫, 42-56% 有焦虑症, 9-70% 表现为胃肠道问题 (Lai, Lombardo, and Baron-Cohen 2013)。考虑到 ASD 症状与许多其他人类疾病的高共病率——无论是神经性的还是非神经性的——这就提出了一个可检验的假设: 与自闭症共同发病率较高的疾病可能与 ASD 有许多共同的基因, 这导致它们涉及的生物学过程也有重叠。因此, 检测存在于 ASD 和几种并发疾病之间的关键基因可能有助于破译共同的分子机制以及潜在的共同的病理生理学, 并最终在理解自闭症病因方面提供巨大帮助。

此外, 在过去的二十年里, 医疗数据的产生、获取和存储都有了显著的增长。尽管针对 ASD 的研究还没有像其他领域那样充分利用“大数据”; 然而, 在简单、经济的数据收集和分析方面的进步可能很快就会使这成为现实。近年来, 人工智能 (artificial intelligence, AI) 作为一种有前途的替代技术得到了发展。人工智能是建立在人脑的生物网络基础上的, 它涵盖了通过模仿人类智能来执行认知功能的广泛技术。人工智能已经在其他领域 (例如, 工程、商业和日常应用) 已经显示出了可喜的结果, 如今正在被越来越多地被纳入医疗保健领域 (Jiang et al. 2017; Yu, Beam, and Kohane 2018)。人工智能研究中最常用的子领域之一是机器学习 (machine learning, ML)。现在, ASD 的高患病率和异质性使得一些研究人员转向机器学习而不是传统的统计学方法进行数据分析 (Hyde et al. 2019)。考虑到当前针对 ASD 的主流诊断方法仍然是基于观察和访谈的患者行为评估量表, 这使得 ASD 的可靠诊断几乎不可能在患者三岁之前做出, 事实上, ASD 患者接受诊断的平均年龄在 38 至 120 个月之间 (A. M. Daniels and Mandell 2014); 更为重要的是, ASD 是一个广泛的神经发育谱系障碍, 这使得患者症状具有极强的个体差异性, 根据报道, 许多没有明显智能障碍的女性患者往往要等到成年后才能够被确诊, 这使得她们错过了儿童行为干预疗法的最佳时期 (Camm-Crosbie et al. 2019), 而基于人工智能的预测诊断方法具有提前预测 ASD 诊断的潜力, 同时也可作为主流诊断方法的参考, 使其做出更为准确的诊断。因此, 基于患者遗传变异识别的 ASD 预测算法是值得深入研究的一个课题, 也显得十分必要和紧迫。

## 2 研究目标与内容

ASD 是一种多因素的疾病, 这意味着多种因素 (环境和遗传) 参与了这一疾病的发展。因此, 罕见的遗传变异, 包括小的缺失和复制 (indels)、CNVs (拷贝数变异) 和 SNVs (单核苷酸变异) 对 ASD 发病有很大贡献。此外, 遗传模式 (常见或罕见) 和这些突变的来源 (遗传或从头开始) 也会对遗传风险有不同的贡献。事实上, 罕见的遗传变异 (MAF<1%) 只能解释 3% 左右的 ASD 遗传风险 (Sanders et al. 2015) 而常见变异占 ASD 遗传性的很大一部分 (约 50%) (Anttila et al. 2018; Glessner, Connolly,

and Hakonarson 2014)。

不过近年来 GWAS 的研究结果揭示了罕见突变或许在 ASD 发病中贡献了大部分效应 (Quick, B. Wang, and State 2021)。人类基因组中的大部分遗传突变是常见的, 这些突变通过孟德尔经典遗传模式从父母传递给后代。下图显示了在普通人群中呈正态分布的常见多基因风险的理想化分布图。图 b 中的红色虚线代表 ASD 的当前诊断阈值。图 a 与图 b 模拟了 ASD 在人群中的分布: 绝大多数的共同等位基因人群风险存在于没有获得临床诊断的个体中, 他们主要分布在 ASD 诊断阈值以下 (红色虚线左侧); 而被诊断为 ASD 的个体则位于 ASD 诊断阈值以上 (红色虚线右侧), 他们可能同时具备以下并发症: 癫痫 (epi epilepsy)、注意缺陷多动障碍 (attention deficit hyperactivity disorder, ADHD)、精神分裂症 (schizophrenia, SCZ) 和特殊语言障碍 (specific language impairment, SLI) 等。图 b 底部的紫色肩头表示具有大效应的罕见新发突变 (de novo mutations), 一个具有较大效应的新发突变可能使得一个具有中等风险和无症状可能性的个体超过诊断阈值。这些新发突变可能并不符合孟德尔遗传学, 这些罕见的大效应突变也许并不直接显示表型, 而是在单个效应非常低的多基因风险背景下发挥作用。虽然这种新发突变在人群中风险比例非常小, 但它们在超过临床阈值的个体中却占了相当大一部分 (Quick, B. Wang, and State 2021)。此外, 绝大多数研究者认为, ASD 研究领域的一个关键问题是疾病机制与 ASD 风险基因所编码的高度多态性生物学相区别。证明某一风险突变的一系列生物学后果 (如不同的基因表达、电生理特性的变化、细胞增殖、分化或迁移的改变) 比确认某一特定风险突变是人类疾病表型的促成因素更为直接。现在, 用于研究 ASD 病理学的系统生物学方法涉及使用大型数据集来确定从基因到表型的某处多个风险基因之间的交叉点或汇合点, 而这一关键交叉点可能与分子途径、细胞类型、解剖区域和/或发育阶段有关。越来越多的研究者们正在采用组学方法, 因为此后将会有越来越多的大规模基础数据库记录不同物种间特定区域和细胞的转录活动、基因表达和调节的发育模式以及蛋白质组学途径 (A. J. Willsey et al. 2018)。

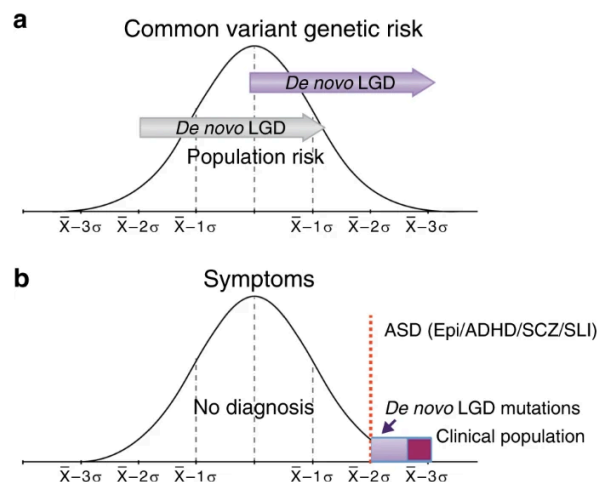


图 1: 罕见的大效应从头突变与常见风险等位基因共同作用的模型 (A. J. Willsey et al. 2018)

近十年来 GWAS 数据的大规模增长使得可供研究的 ASD 数据大幅增多。精神病学基因组联盟 (Psychiatric Genomic Consortium, PGC) 进行的最新 ASD GWAS 荟萃分析就是一个证明。PGC 测序队列中的样本量达到了 10000 个病例和对照, 而且还制定了明确的质量控制和归因流程。在这里, 我们使用 PGC 数据库 (<https://www.med.unc.edu/pgc/download-results/>) 中的 ASD GWAS 统计摘要数据集 (<https://pubmed.ncbi.nlm.nih.gov/30804558>), 采用按照基于基因的分析 (gene-based analysis, GBA) 方法进行分析。基因内每个 SNP 的 p 值将被使用路径评分算法 PASCAL 结合起来, 从而获得

整个基因组内每个单一基因的单一统计数据。GBA 方法使用 GWAS 的汇总统计数据作为输入数据，而不需要基因分型后的数据。本研究希望将基因作为基本分析单位，并与每个基因参与的实际生物学作用相结合，从而帮助确定参与疾病发病机制的分子和细胞机制的特征，以及 ASD 病理过程中多个风险基因之间的关键交叉点，并为基于患者遗传变异识别的 ASD 预测算法开发做出帮助。

## 3 已有研究进展和初步结论

### 3.1 背景与方法

来自 GWAS 的总结统计数据可以用于执行基于基因的分析 (gene-based analysis, GBA)。GBA 方法 (将基因作为一个单元进行测试) 为准确理解 ASD 复杂的生物学机制提供了一个优势，它允许整合基因上的所有遗传信息，以创建一个单一的统计数据 (每个基因的 p 值)。一种新型 GBA 工具 (PASCAL, 路径评分算法) 此前被开发出来，并被推荐用于最新的 ASD 荟萃分析中 (Alonso-Gonzalez et al. 2019)。因此，可以利用 GBA 方法筛选出来的基因进行基因网络分析和 KEGG 和/或 GO 的富集分析，并使用 GENE2FUNC 创建基因表达注释和表达热图，使用 Locus Zoom 绘制 Regional Plots。

#### 3.1.1 GWAS 荟萃分析

从 PGC 网站 (<http://www.med.unc.edu/pgc/results-and-downloads>) 的公共资料库中获得了最新的 ASD GWAS 荟萃分析的汇总统计数据。本篇研究中采用了以下数据集:iPSYCH\_PGC\_ASD\_Nov2017.gz (Grove et al. 2019)。该数据集共包括 18381 个病例和 27969 个对照，数据采用的基因分型和质量控制方法可在 PGC 网站上找到参考。

#### 3.1.2 GBA 分析

使用 PASCAL 评分算法 (<https://www2.unil.ch/cbg/index.php?title=Pascal>) 作为 ASD 荟萃分析的汇总统计方法。首先将 GWAS 结果中的单个 SNPs 映射到基因上，在 5 和 3 UTRs 周围采用默认的  $\pm 50\text{kb}$  窗口。PASCAL 允许的每个基因的 SNPs 的默认最大数量为 3000。LD 信息来自 1000 Genomes European 统计结果。Bonferroni 校正将显著性分界线设定为  $2.26 \times 10^{-6}$  ( $0.05/22135$  个基因)。

#### 3.1.3 功能注释

使用 FUMA (全基因组关联研究的功能图谱和注释) (<http://fuma.ctglab.nl/>) 的 GENE2FUNC 功能，以 PASCAL 运行结果作为输入数据，对 ASD 基因及其相互作用基因进行功能注释。

#### 3.1.4 区域关联图

使用 LocusZoom 工具 (<http://locuszoom.org/>) 构建包含 PASCAL 相关基因的区域关联图 (regional association plots)。使用 iPSYCH PGC ASD Nov2017 数据中的 GWAS 统计汇总作为输入数据，指定 marker 名称、P 值、OR、染色体位置 (起始端) 和索引 SNP。用于构建这些区域图中 SNP 之间的  $r^2$  相关矩阵的 LD 信息来自于 g19/1000 Genomes Nov 2014 EUR。

## 3.2 结果与讨论

### 3.2.1 GENE2FUNC 的注释结果

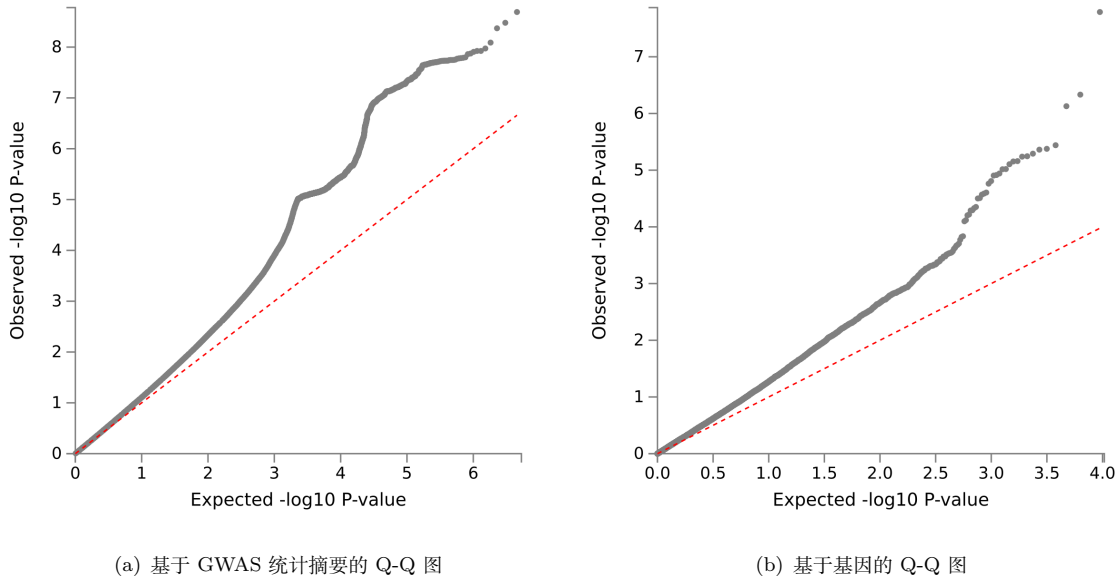


图 2: 左下角是显著性低的位点, 即确定与性状不关联的位点, 这些位点的 P value 观测值应该与期望值一致 (小于  $10^{-1}$  和  $10^{-0.5}$ )。右上角为显著性较高的位点, 是潜在与性状相关的候选位点。这些点位于对角线的上方, 即位点的 P value 观测值超过了期望值, 说明这些位点的效应超过了随机效应, 进而说明这些位点是与性状显著相关的。图 2 中 a 和 b 中左下角说明了模型的合理性, 右上角说明找到关联位点, 证明使用该 GWAS 数据集的分析结果符合期望 ( $P$  值  $\geq 1e^{-5}$  的 SNP 被过滤掉)。

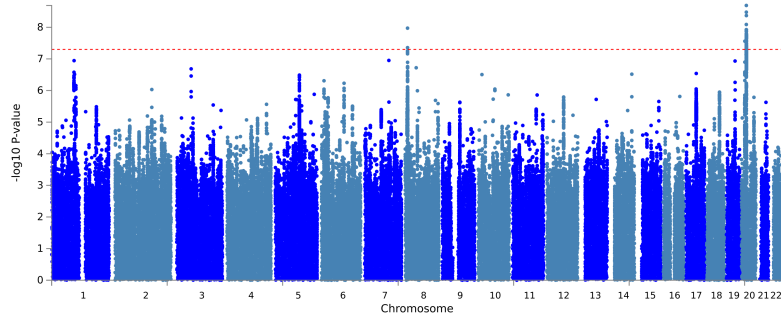


图 3: FUMA 基于 GWAS 统计摘要绘制的曼哈顿图。X 轴为染色体编号, 且每个基因组 SNP 位点沿染色体序列排列; Y 轴为该位点相关的统计显著性 Pvalue 值的  $\log_{10}$  变换; 图中水平线为设定的显著性水平阈值 ( $P = 0.05/18937 = 2.64e^{-6}$ )。输入的 SNP 被映射到 18937 个蛋白编码基因 ( $P$  值  $\geq 1e^{-5}$  的 SNPs 被过滤)。图 3 显示在 8 号和 20 号染色体上, SNP 的信号峰最强, 暗示其与 ASD 表型关联程度较高。

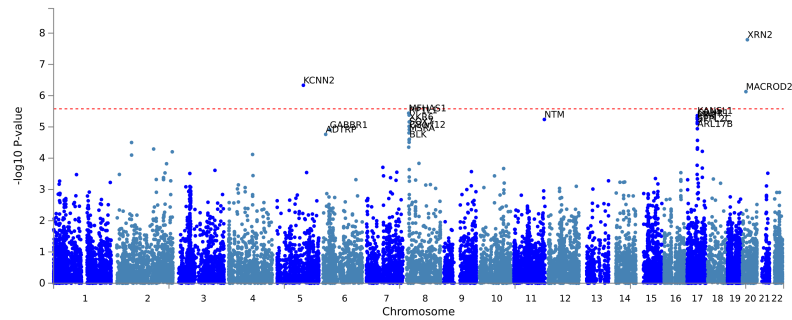


图 4: FUMA 基于基因绘制的曼哈顿图。图 4 显示在 5、8、17 和 20 号染色体上有与 ASD 表型关联程度较高的基因。

表 1: MAGMA 基因组分析的结果展示。MAGMA 基因集合分析 (MAGMA gene-set analysis) 是针对从 MsigDB 获得的策划的基因组和 GO 术语进行的。表 1 中展示的是  $P_{bon} < 0.05$  的显著基因集合。需要注意的是, MAGMA 基因组分析使用 SNP p 值的完整分布, 这种方法与 GENE2FUNC 中实现的路径富集测试不同, 后者只测试优先基因的富集。

Gene Set	N	Beta	Beta STD	SE	PB	Pbon
	genes					
GO bp:go macrophage colony stimulating factor production	5	1.394	0.022649	0.35169	3.7034e-05	0.573434456
GO cc:go extrinsic component of postsynaptic membrane	8	1.057	0.021721	0.28832	0.00012356	1
GO bp:go granulocyte colony stimulating factor production	51.3071	0.021236	0.36585	0.00017718	1	
GO mf:go neurotransmitter receptor activity	108	0.3108	0.023405	0.088422	0.00022052	1
GO cc:go extrinsic component of synaptic membrane	12	0.87453	0.022008	0.252630	.00026897	1
GO bp:go extraembryonic membrane development	9	1.1923	0.025987	0.35019	0.00033198	1
Curated gene sets:pidncadherin pathway	33	0.49317	0.02057	0.1451	0.00033919	1
GO bp:go histone h3 k27 demethylation	2	1.65	0.016957	0.49	0.0003802	1
GO mf:go histone demethylase activity h3 k27 specific	2	1.65	0.016957	0.49	0.0003802	1
GO cc:go t tubule	52	0.43768	0.022904	0.13185	0.00045174	1



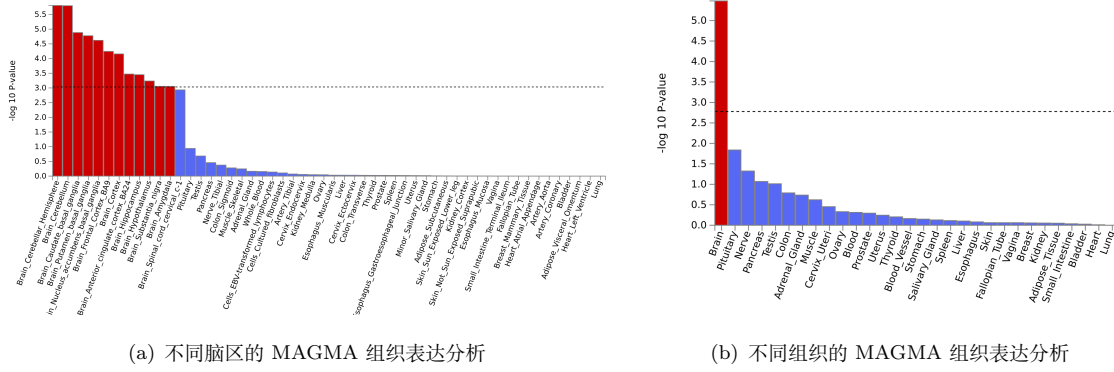


图 5: MAGMA 组织表达分析。如图 5 所示, MAGMA 基因特性分析是针对我们选择的数据集的基因表达进行的。需要注意的是, MAGMA 基因特性分析使用 SNP p 值的完整分布 (这与 GENE2FUNC 中实现的 DEG (差异表达基因) 的富集测试不同, 后者只测试优先基因的富集程度)。图 5 是用 GTEX v7 (53 个组织) 构建的 ASD DEG 图。显著富集的 DEG 集合 ( $P_{bon} < 0.05$ ) 以红色突出显示。图 5(a) 为不同脑区的 MAGMA 组织表达分析, 图中可见高于设定阈值 (图中虚线,  $Pvalue \leq 10^{-3}$ ) 的主要脑区由 P 值从高到低依次为: 小脑半球、小脑、尾状神经节、基底神经节、伏隔核基底节、大脑额叶皮质、大脑前扣带皮层 BA24、大脑海马区、脑黑质和杏仁核。图 5(b) 为不同组织的 MAGMA 组织表达分析, 图中可见高于设定阈值 (图中虚线,  $Pvalue \leq 10^{-3}$ ) 的组织为脑。

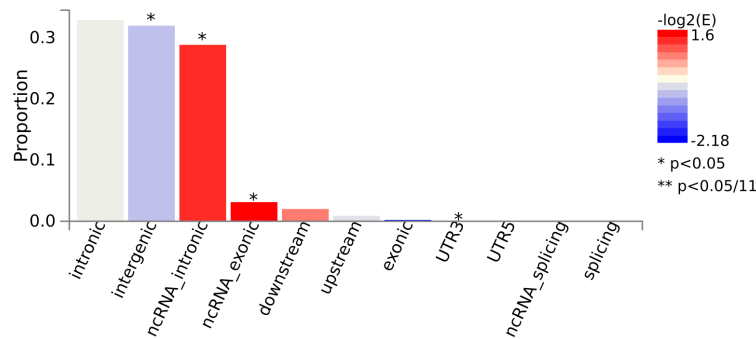


图 6: SNPs 对基因的功能影响。根据 FUMA 计算的结果, 全基因组的显著风险基因座有 3 个, 主要 SNP 有 3 个 (rs10099100, rs71190156 和 rs910805), 显著 SNP 有 7 个, 候选 SNP 有 447 个, 候选 GWAS 标记 SNP 有 422 个, 对应基因有 8 个。如图 6 所示, FUMA 计算出的功能性 SNP 主要位于内含子区域、基因间、非编码 RNA 的内含子区域和非编码 RNA 的外显子区域 ( $pvalue \leq 0.05$ )。

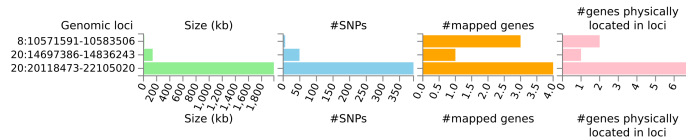


图 7: 每个基因风险位点的摘要图。ASD 关联的风险基因座主要位于 8 号 (10571591bp-10583506bp) 和 20 号染色体 (14697386bp-14836243bp 和 20118473bp-22105020bp) 区域内。

### 3.2.2 Locus Zoom 绘制区域关联图

根据 FUMA 的计算结果，这里重点关注第 8、17 和 20 号染色体。图 8、9 和 10 是 ASD GWAS 荟萃分析的区域关联图（8、17 和 20 号染色体）。关联图在周围标记、基因注释、估计重组率和周围标记与假定相关变体之间的成对关联结果的背景下突出关联的统计强度。图中菱形标记为被正确分型的 SNP，圆形标记为被估算的 SNP，估计的重组率被绘制在图上，以反映相关 SNP 周围的局部 LD 结构（红色表示高度相关，深蓝色表示弱相关）。图表下方 SNP 在我们选定的染色体区域内对应的基因。路径评分算法 PASCAL 强调了 NKX2-2 和 NKX2-4（20 号染色体）、KANSL1-AS1（17 号染色体）和 C8orf74（8 号染色体）等基因在基因水平上的关联。

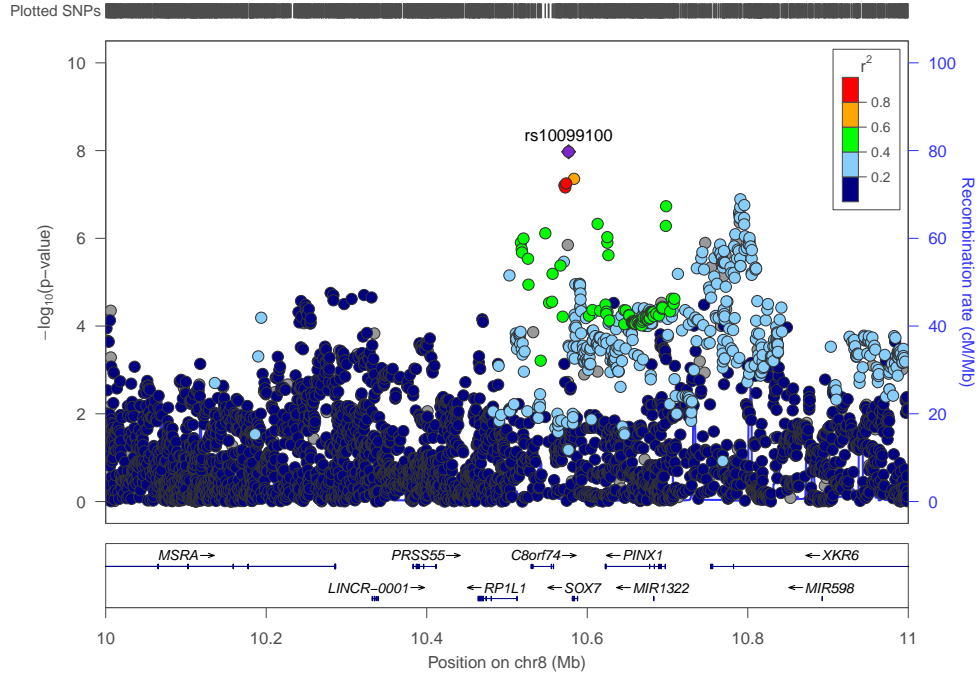


图 8: EUR 8 号染色体 (10mb-11mb 区域)

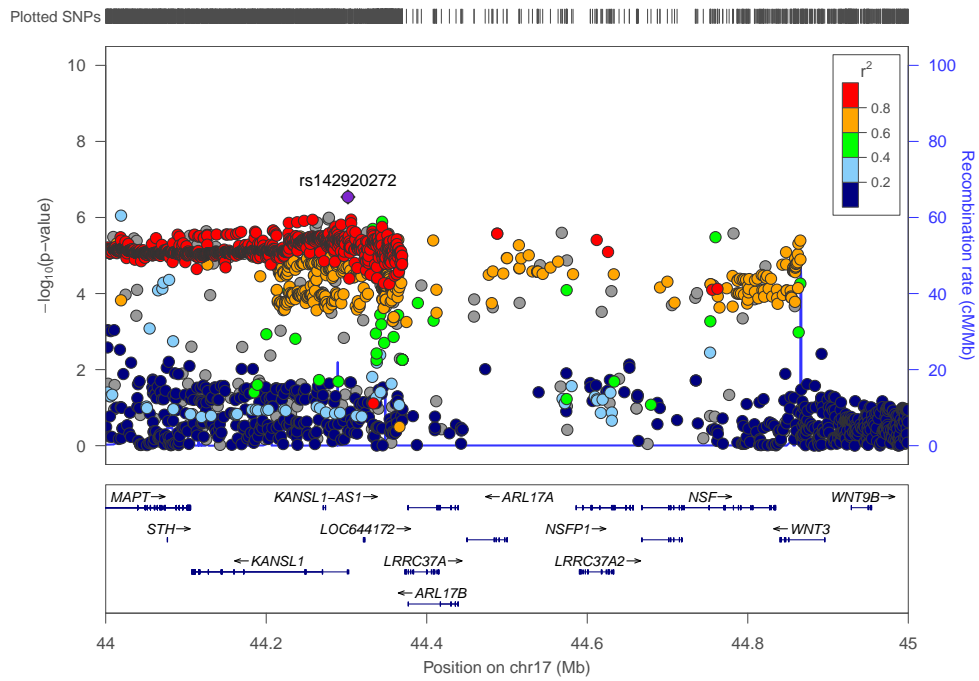


图 9: EUR 17 号染色体 (44mb-45mb 区域)

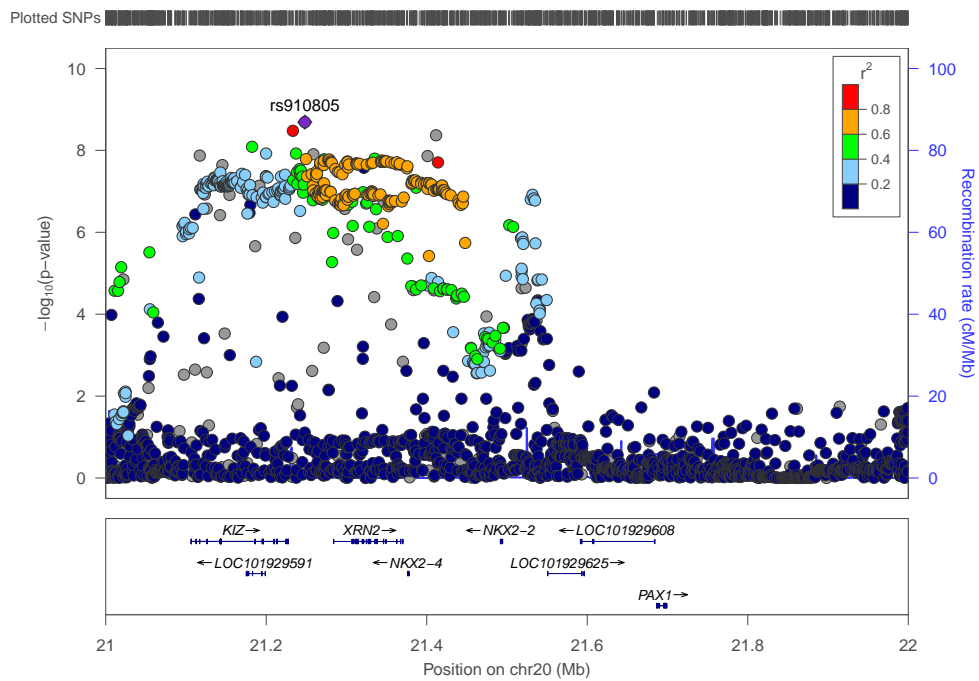


图 10: EUR 20 号染色体 (21mb-22mb 区域)

### 3.3 结论

当前研究基于 GBA 工具的筛选结果确定了 8、17 和 20 号染色体上与 ASD 相关的风险位点；同时还确定 ASD 基因水平上的部分关联；另外，当前研究成果展示了 ASD 相关基因在不同脑区和不同身体组织中的表达情况，同时定位了部分风险基因座，此外，描述了 ASD 基因潜在的互作网络与其参与的生物学过程。

## 4 未来研究计划

当前，我们已经取得了一系列初步成果。在本篇研究中，ASD GWAS 荟萃分析的结果报告了 8 号染色体、17 号染色体和 20 号染色体等个别 SNPs 的关联，这些 SNPs 附近的不同基因——如此前报道过 (Grove et al. 2019) 的 *XRN2*、*KCNN2*、*KIZ* (或 *PLK1S1*) 和 *PASCAL* 分析新发现的额外相关基因 *NKX2-4*、*NKX2-2*、*CRHR1-IT1*、*C8orf74* 和 *LOC644172*——被关联起来。而且，值得注意的是，根据此前的报道，当 MTAG (multi-trait analysis of GWAS, GWAS 的多性状分析) 与遗传相关的表型 (精神分裂症和教育程度) 进行分析时，位于 20 号和 8 号染色体上的线索 SNP 与 ASD 相关，但类似 *PASCAL* 的路径评分算法 *MAGMA* 本身的计算结果并不能显示这些 SNPs 与 ASD 相关 (Turley et al. 2018; Grove et al. 2019)。这可能暗示，在此后的研究中，我们需要将 ASD 视为一类范围宽广且可能与其他疾病性状高度相关的表型，并且可能需要将 ASD 与并发症/表型进行关联研究。

此外，关于对 ASD 相关基因的生物学作用的深入研究，以下基因座可能值得重点关注：*NKX2-2*、*NKX2-4* 和 *c8orf74*。*NKX2-2* 和 *NKX2-4* 是 homeobox-转录因子家族的成员。*NKX2-2* 编码的转录因子参与中枢神经系统的形态发生，它在早期发育过程中位于后脑和脊髓的神经群的分化过程中至关重要 (Briscoe and Ericson 1999)。*NKX2-4* (Homeobox Protein *Nkx-2.4*) 在大脑发育中具有关键作用，它在发育中的前脑中下调会促进神经祖细胞的增殖和抑制分化，从而导致神经发生受损 (Shen et al. 2018)。此前有报道称 *NKX2-2* 最初在分化的少突胶质细胞前体细胞 (oligodendrocyte precursor cells, OPCs) 中表达，然后被下调。然而，它被证明在后来的神经发育阶段也可以再次上调其表达，以维持髓鞘结构 (parencitecai2010co)。这一点非常有趣，因为此前有文献提到可以将整个少突胶质细胞系的基因异常表达与 ASD 病理联系起来 (parenciteli2014integrated, zeidan2016up)。*NKX2-2* 和 *OLIG-2* 的功能富集关键词是蛋白质、染色质和 TF 结合，这些都是参与 ASD 发病机制的生物过程。这些结果表明，需要进行后续的功能研究，以进一步确定其生物学作用的特点。另外，*C8orf74* 是一个具有翻译能力的阅读框的一部分，但它还没有被定性。因此，这个基因座似乎由于其生物学特征不明显而没有纳入功能和遗传数据库。然而，此前有关基因表达谱荟萃分析的文献显示，*C8orf74* 在 ASD 脑组织和对照组之间有不同的表达 (Alonso-Gonzalez et al. 2019)，这表明 *C8orf74* 需要进一步的功能鉴定。

另外，本项研究希望通过公共数据库中的 GWAS 数据构建潜在的 ASD 风险基因评分预测方法。考虑到一些大规模的临床研究已经证实了与 ASD 相关的高并发症率——如 ADHD (注意力缺陷障碍, attention deficit hyperactivity disorder)、ANX (焦虑障碍, Anx disorders factors) 和 SCZ (精神分裂症, Schizophrenia)——这些合并症代表了 ASD 患者额外的疾病负担 (Leyfer et al. 2006)，未来研究中将会考虑将几类高发的并发症与 ASD 合并表型，从而简化计算复杂度。此外，由于绝大多数研究者认为，ASD 研究领域的一个关键问题是将疾病机制与 ASD 风险基因所编码的高度多态性生物学相区别。证明某一风险突变的一系列生物学后果 (如不同的基因表达、电生理特性的变化、细胞增殖、分化或迁移的改变) 比确认某一特定风险突变是人类疾病表型的促成因素更为直接；而且用于研究 ASD 病理学的系统生物学方法涉及使用大型数据集来确定从基因到表型的某处多个风险基因之间的交叉点或汇合点，而这一关键交叉点可能与分子途径、细胞类型、解剖区域和/或发育阶段有关。(A. J. Willsey

et al. 2018)。本研究计划进一步分析 GWAS 中出现的高风险基因位点，并采用组学方法，尝试将它们纳入基因互作网络研究的范畴，试图寻找潜在的关键交叉点，为 ASD 风险基因评分预测算法提供线索。

因此，未来研究计划申请 PGC 数据库中有关 ADHD 和 ANX 的 GWAS 统计汇总数据使用权，希望作为后续风险基因评分预测分析的训练集；同时申请 iPSYCH（伦贝克基金会综合精神病学研究计划，Lundbeck Foundation Initiative for Integrative Psychiatric Research）从丹麦收集的 14584 个病例和 22492 个对照人群的 GWAS 数据与精神诊断量表评分，希望作为后续风险基因评分预测分析的测试集。

## References

- Alonso-Gonzalez, Aitana et al. (2019). “Novel gene-based analysis of ASD GWAS: Insight into the biological role of associated genes”. In: *Frontiers in genetics* 10, p. 733.
- Anttila, Verner et al. (2018). “Analysis of shared heritability in common disorders of the brain”. In: *Science* 360.6395.
- Blumberg, Stephen J et al. (2013). *Changes in prevalence of parent-reported autism spectrum disorder in school-aged US children: 2007 to 2011-2012*. 65. US Department of Health and Human Services, Centers for Disease Control and ...
- Briscoe, James and Johan Ericson (1999). “The specification of neuronal identity by graded Sonic Hedgehog signalling”. In: *Seminars in cell & developmental biology*. Vol. 10. 3. Elsevier, pp. 353–362.
- Camm-Crosbie, Louise et al. (2019). “ ‘People like me don’ t get support’ : Autistic adults’ experiences of support and treatment for mental health difficulties, self-injury and suicidality”. In: *Autism* 23.6, pp. 1431–1441.
- Christensen, Deborah L et al. (2018). “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012”. In: *MMWR Surveillance Summaries* 65.13, p. 1.
- Daniels, Amy M and David S Mandell (2014). “Explaining differences in age at autism spectrum disorder diagnosis: A critical review”. In: *Autism* 18.5, pp. 583–597.
- Diaz-Beltran, Leticia et al. (2017). “Cross-disorder comparative analysis of comorbid conditions reveals novel autism candidate genes”. In: *BMC genomics* 18.1, pp. 1–14.
- Geschwind, Daniel H and Matthew W State (2015). “Gene hunting in autism spectrum disorder: on the path to precision medicine”. In: *The Lancet Neurology* 14.11, pp. 1109–1120.
- Glessner, Joseph T, John J Connolly, and Hakon Hakonarson (2014). “Genome-wide association studies of autism”. In: *Current Behavioral Neuroscience Reports* 1.4, pp. 234–241.
- Hyde, Kayleigh K et al. (2019). “Applications of supervised machine learning in autism spectrum disorder research: a review”. In: *Review Journal of Autism and Developmental Disorders* 6.2, pp. 128–146.
- Jiang, Fei et al. (2017). “Artificial intelligence in healthcare: past, present and future”. In: *Stroke and vascular neurology* 2.4.
- Jobski, K et al. (2017). “Use of psychotropic drugs in patients with autism spectrum disorders: a systematic review”. In: *Acta Psychiatrica Scandinavica* 135.1, pp. 8–28.

- Lai, M-Ch, MV Lombardo, and S Baron-Cohen (2013). *Autism (seminar)*.
- Leyfer, Ovsanna T et al. (2006). “Comorbid psychiatric disorders in children with autism: interview development and rates of disorders”. In: *Journal of autism and developmental disorders* 36.7, pp. 849–861.
- Matelski, Lauren and Judy Van de Water (2016). “Risk factors in autism: Thinking outside the brain”. In: *Journal of autoimmunity* 67, pp. 1–7.
- Mattila, Marja-Leena et al. (2011). “Autism spectrum disorders according to DSM-IV-TR and comparison with DSM-5 draft criteria: an epidemiological study”. In: *Journal of the American academy of child & adolescent psychiatry* 50.6, pp. 583–592.
- Quick, Veronica B Searles, Belinda Wang, and Matthew W State (2021). “Leveraging large genomic datasets to illuminate the pathobiology of autism spectrum disorders”. In: *Neuropsychopharmacology* 46.1, pp. 55–69.
- Sanders, Stephan J et al. (2015). “Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci”. In: *Neuron* 87.6, pp. 1215–1233.
- Shen, Tianjin et al. (2018). “Brain-specific deletion of histone variant H2A. z results in cortical neurogenesis defects and neurodevelopmental disorder”. In: *Nucleic acids research* 46.5, pp. 2290–2307.
- Willsey, A Jeremy et al. (2018). “The psychiatric cell map initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders”. In: *Cell* 174.3, pp. 505–520.
- Yoo, Heejeong (2015). “Genetics of autism spectrum disorder: current status and possible clinical applications”. In: *Experimental neurobiology* 24.4, p. 257.
- Yu, Kun-Hsing, Andrew L Beam, and Isaac S Kohane (2018). “Artificial intelligence in healthcare”. In: *Nature biomedical engineering* 2.10, pp. 719–731.
- Zablotsky, Benjamin et al. (2015). “Estimated prevalence of autism and other developmental disabilities following questionnaire changes in the 2014 National Health Interview Survey”. In: