高通量组学应用汇编-医学篇

Human Research in Next Generation Sequencing

版权所有2016华大基因

此书献给

广大前沿的科研工作者
希望此书能助您找到
新的科研思路……

目录

CONTENTS

人类研究

引言 人类遗传学研究 2

基于高通量测序的种群特征研究方案 4

疾病免疫组库研究方案 17

复杂疾病的MHC易感位点研究方案 26

疾病甲基化分子标记物研究方案 34

复杂疾病*De novo*突变研究方案 44

复杂疾病肠道宏基因组研究方案 57

单基因遗传病的外显子测序研究方案 67

肿瘤发生的甲基化调控机制研究方案 76

肿瘤诊断的蛋白标志物研究方案 82

肿瘤单细胞基因组研究方案 89

病毒性肿瘤发生机制研究方案 102

肿瘤融合基因研究方案 113

肿瘤研究之非编码RNA生物标记物研究方案 123

引言

Introduction

人类研究

从上古时代起，人类就一直在不断地探索和认识自身的由来及人体的奥秘，人的生老病死、思维意识无不与遗传息息相关。人类遗传学是在普通遗传学的基础上形成和发展起来的一门学科，其目的是研究不同人在形态、结构、生理、生化、免疫、行为等各种性状方面在遗传上的相似与差别以及其物质基础，它的研究与发展极大地丰富了普通遗传学的内涵。但是由于人类自身的特殊性，相对于动物、植物以及微生物等物种来说，人类遗传学在研究方法和条件等方面受到较多限制因素，因此初期的人类遗传学仅仅停留在分析研究血型等正常性以及患病后所显示的异常性等的遗传方式方面。

近代人类遗传学由高尔顿开拓，他注意到“先天与后天”的区别和关系，提出了“优生学”这一名词，并首倡双生儿法研究遗传与环境的关系。孟德尔和摩尔根则建立起理论的基石，伽罗德又把孟德尔规律应用于人类遗传学研究，并结合医学与人类遗传学探索疾病发生原因。

1928 年，Griffith 利用转化实验发现 DNA 是主要的遗传物质，在这之后，对于 DNA 的研究越来越多，而最具有里程碑意义的是 1953 年，Watson 和 Crick 发现了 DNA 的双螺旋结构，生命之谜进一步被解开，开启了分子生物学的新时代。人们能够清楚地了解遗传信息的构成和传递的途径，特别是 20 世纪 70 年代以来采用了分子遗传学的方法，例如工具酶的应用，有力地推动了基因定位和产前诊断相关研究工作的发展。随后还相继出现了人类细胞遗传学、人类生化遗传学和人类分子遗传学等多个子学科。

继 2001 年人类基因组计划完成之后，“国际单倍体型（HapMap）计划”和“千人基因组计划”相继开展并积累了海量的遗传变异数据，尤其是 2008 年启动的“千人基因组计划”，它绘制出迄今为止最详尽、最有应用价值的人类基因组遗传多态性图谱。这些遗传变异数据为更深层次上了解种族之间、个体之间的基因差异，以及为各种疾病的关联分析提供详细的基础数据，极大地推动了群体遗传学、人类疾病研究、比较基因组学及药物基因组学等研究。以人类基因组作图为重点的人类遗传学研究成果和各种先进的技术手段，正在有力地带动整个生命科学的飞速发展。

随着人类遗传学的深入研究，科学家发现，基因遗传并不能解释所有表型或疾病的发生，例如 II 型糖尿病、癌症以及心血管疾病等，这些疾病一般都是在遗传与环境的共同作用下发生；例如肺癌与吸烟有着密不可

分的关系，肥胖也可能是因为肠道微生物的影响。至此，人类遗传学的研究范畴被大大拓宽，广泛的涵盖了人类迁徙、进化、疾病、药物、发育、环境、肠道微生物等各个领域。

华大基因通过多平台的测序技术在 DNA 水平、RNA 水平、表观遗传学水平以及宏基因组水平对疾病展开全方位的研究，并结合质谱技术开展蛋白质组水平和代谢水平的研究，甚至是利用各组学平台进行贯穿研究，由此得出的海量数据，与表型相关的各种信息（所处环境、年龄、疾病治疗史、家族史、易感性等）相结合，为全面揭示人类遗传学机制奠定了重要基础。

基于高通量测序的种群特征研究方案

一、 研究背景

对于一个特定的种群，我们首先需要知道这个种群的基因组特点才能对这个种群有针对性研究，有利于后续该种群在疾病诊断和遗传病筛查等方面的应用。近代分子生物学的研究表明，不同地区不同民族对环境适应性基因有明显差异[1]，蒙古族人则有着异于平原地区居民的体征：虽然高血压、高血脂、高血糖，但是非常健康， 无任何不适。

人类由不同的人种组成，在世界各地生活着各种特异的人类群体，如傣族人、藏族人、犹太人、冰岛人等。就人类进化史来讲，造成这些差异的实际时间并不漫长，同时在考古学、数学、计算机以及迅速发展的生物信息学的帮助下，使得详细地研究人类迁徙和进化成为可能。格陵兰岛上的古爱斯基摩人的测序与分析结果表明，该个体可能来源于欧洲大陆，通过尚未被海水淹没的白令海峡迁徙到美洲，并经美洲最终到达格陵兰岛[2]。对 100 年前的澳洲土著人头发的测序结果表明，澳洲土著人在欧洲人、亚洲人分离之前就已经迁徙到了澳洲，且为现代澳洲土著人的祖先[3]。西伯利亚古人[4]也同样为我们提供了大量的先祖基因信息。同时，千人基因组[5]、Hapmap[6] 等计划的完成，使人们知道了各个人种（欧洲人、约鲁巴人、中国人、日本人等）之间的基因差异，现代人的共同祖先来自于非洲。通过对生活在高原上的藏族人基因组的研究[1]，人们发现，*EPAS1* 上的一个 SNP 突变可以帮助藏族人适应高原的缺氧气候。

由于地理、历史、风俗的原因，人类有很多种群保留了很明确的血统。比如孔家人（孔子后代）有着非常完善的族谱；吉普赛人、非洲原始部落一般为族内通婚；冰岛人生活在北极圈的岛国上，也极少与外界通婚；印度不同种族之间也往往有很明显的隔阂。这些人类群体基因交流极少，故往往都有着很独特的基因，为我们研究他们的迁徙进化史提供了便利。

随着 GWAS 研究方法的兴起，人类疾病的研究取得了很大的进展，找到大量与复杂疾病相关的基因变异位点。现有数据表明，疾病在不同血统的人群中发病率迥异。比如印度大陆人群在 *MYBPC3* 上有 4%的变异，而在其他地区没有或非常稀少[7]。不同的种群中，不同分子标记集的偏向性和 Tag-SNP 不同[8]，如不确定种群血统， 将会影响 imputation 的准确性。因此，种群进化迁徙历史和血统的鉴定，为后续种群疾病的研究奠定了基础。

研究分析种群的变异情况，包括 SNP、InDel、CNV、SV 等，寻找种群特异性基因或基因结构；进行血统分析，确定种群在进化树上的位置，推断种群进化迁徙历史；确定血统，为种群疾病研究如 GWAS 分析方法等提供依据，以便为不同血统选择合适的研究方法。

二、 方案设计

图 1 研究方案设计

2.1 样本建议

(1) 以家庭为单位进行取样，三口之家或者有双胞胎的家庭（有利于发现新生变异）；

(2) 有完整的表型数据，方便后续进一步进行关联分析。

2.2 实验技术

采用全基因组重测序（WGS）或外显子测序（WES）技术，通过对群体样品进行中低深度测序，找出群体 特异基因或基因结构。

2.3 测序参数

建议全基因组测序深度 15X-20X，外显子测序深度 50X 以上。

2.4 分析内容

群体遗传学分析（基于 SNP、InDel 等变异检测基础）

表 1 信息分析内容

1	群体 SNP 检测和基于连锁不平衡(LD)的 Genotype 分型检测
2	SNP 注释与统计（包含 OMIM 注释）
3	群体 SNP 质控（base quality，map quality，allele balance，strand bias，mappability，homopolymer，Hardy-Weinberg Equilibrium 测试，InDel 附近的 SNP 过滤）
4	基于遗传数据的样本质控：亲缘关系检测，基于近交系数的样本污染检测
5	基于参考单体型集合(HapMap/1000 Genome)的基因型推断和单体型定相
6	群体结构分析，主成分分析和系统发育树构建
7	选择分析，主要是近期选择事件分析，基于 iHS 和 XP-EHH，同时辅以 DDAF、Fst、Tajima’s D 来做验证
8	对受选择基因进行 GO 功能分类，以及 GO/KEGG/PANTHER 通路分析
9	单体型分析，主要包括单体型域，变异，模式，多样性，群体间的保守单体型和特异单体型等
10	染色体 chrY 和 chrM 单体型组分析
11	群体迁移历史推断

2.5 分析结果

(1) 群体变异信息统计

群体测序数据通过变异检测软件得到单碱基变异信息，然后对群体 SNP 信息进行注释和统计。如图 2 所示，

对各个群体的 SNP 进行归类统计，展示不同群体 SNP 特点、各个种群基因组变异数目以及各个群体平均单体型数目。图 3 展示的是每个群体的常见变异但是在全球人群中是罕见变异，以及每个人群中特异的基因。

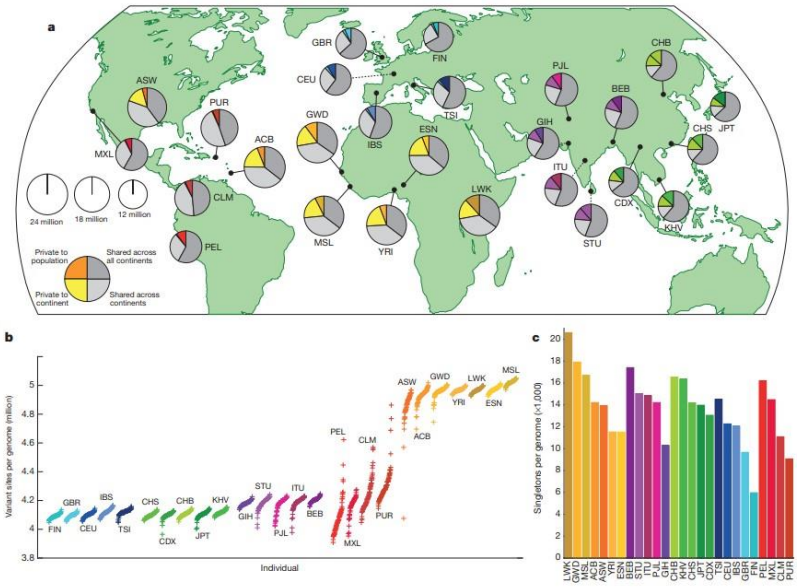


图 2 群体样品选择和群体 SNP 统计结果[9]

a) 各个地区群体 SNP 特点，橙色是群体特有的；亮黄色是单个大陆特有的，浅灰色是几个大陆群体共有，深灰色是所有大陆群体共有的;b) 各个基因组的变异数目;c) 各个基因组平均单体型数目。

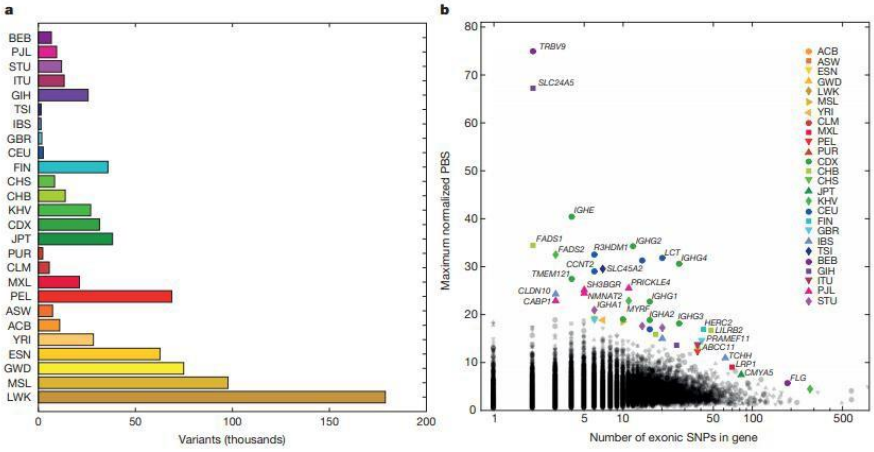


图 3 群体分化[9]

a) 特定群体的常见变异（MAF > 5%），但是在全球人群中是罕见变异（MAF < 0.5%）；b) 群体间分化严重的基因，基于 Fst 群体分支统计（PBS），纵坐标表示该基因的最大值，用彩色标注显示某群体中检测的基因。

(2) 群体结构分析、主成分分析和系统进化树构建

群体遗传学的研究，主要目的是区分不同群体的特点，根据群体特点进行分类，推断不同人种的遗传分离时间。图 4 展示的是用最大近似值的方法，用 8 个集群展示 27 个群体结构。PSMC 分析得到不同时间有效群体的大小。图 5a 也是展示不同群体结构，以及 51 个群体的系统进化树。图 6 是对荷兰人基因组计划数据进行 PCA 分析，选取 3 个主要因素，两两对比，从而区分了不同人群数据。

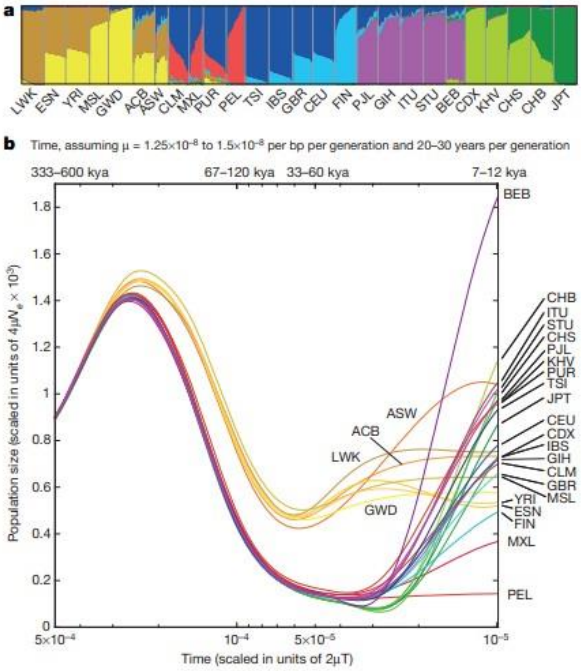


图 4 群体结构和人口统计[9]

a) 最大近似值方法，用 8 个集群展示群体结构；b) PSMC 分析得到不同时间有效群体的大小。

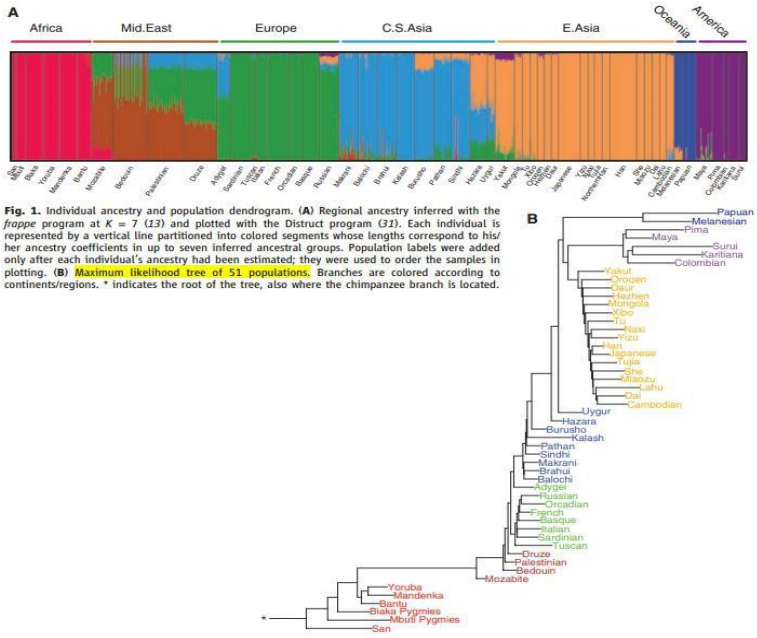


Fig. 1. Individual ancestry and population dendrogram. (**A**) Regional ancestry inferred with the *frappe* program at $K = 7$ (*13*) and plotted with the Distruct program (*31*). Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to his/her ancestry coefficients in up to seven inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting. (**B**) Maximum likelihood tree of 51 populations. Branches are colored according to continents/regions. * indicates the root of the tree, also where the chimpanzee branch is located.

图 5 个体血统分析和群体系统数构建[10]

a)利用 frappe 软件做出的区域血统图；b) 51 个群体的最大近似值数。

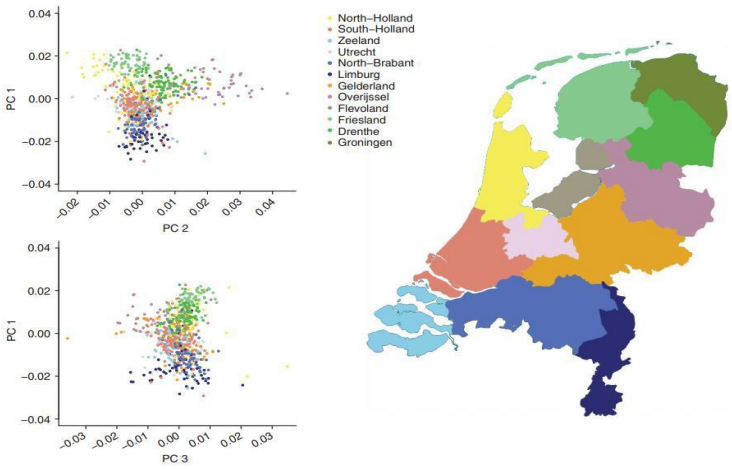


图 6 群体主成分分析[11]

荷兰人基因组计划PCA 分析，选取 3 个主要因素，两两对比。这样可以把不同地区的人分开。

(3) 群体迁移历史推断

除了考古学，基因组学是现在另一种推断群体迁移历史的工具。图 7 展示的是通过 MSMC 的分析方法得到世界各地人种分离的大致年代。

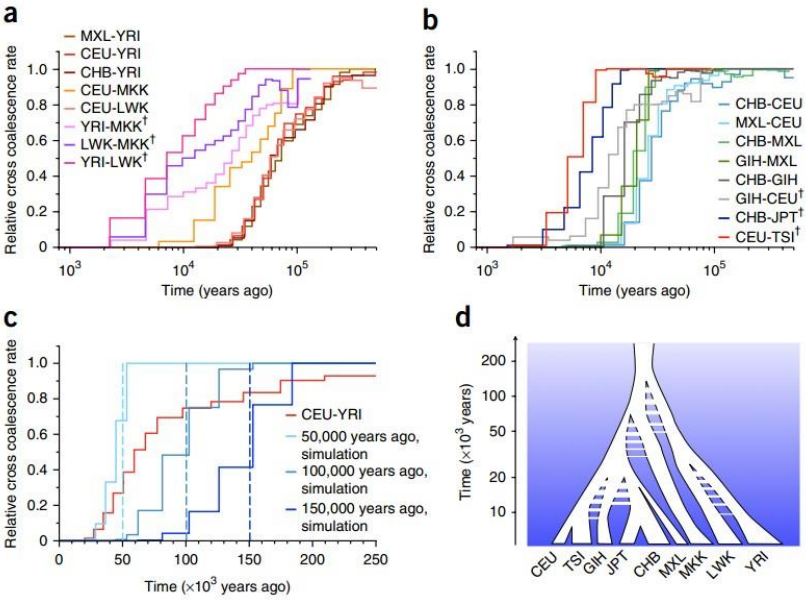


图 7 群体间的遗传分离[12]

a) 非洲人与其他人种的相对交叉聚合率,展示非洲人与其他人种分离的时间；b) 除了非洲人外，其他人种间的相对交叉聚合率,展示其他人种分离的时间；c) 非洲与欧洲人间的相对交叉聚合率，以及模拟三种情景的相对交叉聚合率（5 万年前分离；10 万年前分离；15 万年前分离）；d) 群体间分离的示意图。

2.6 项目周期

样品检测合格后，建库+测序+标准信息分析约需 40 个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.7 预期结果

研究分析种群的变异情况，包括 SNP、InDel、CNV、SV 等，寻找种群特异性基因或基因结构；进行血统分析，确定种群在进化树上的位置，推断种群进化迁徙历史；确定血统，为种群疾病研究如 GWAS 分析方法等提供依据，以便为不同血统选择合适的研究方法。

2.8 实验验证

分析得到的变异位点可以利用芯片分型、质谱分型进行更大群体范围验证，从而找到该群体确切的特异变异位点。

三、 应用案例

案例 华大参与——荷兰人基因组计划[11,13-14]

荷兰国家基因库是欧洲基因库的一部分，后者是为了获得欧洲人特有的基因数据，分析欧洲人的遗传资源并后续应用于欧洲人临床数据解读的研究计划。荷兰人基因组计划是一个全基因组重测序的项目，它选取了全荷兰具有代表性的 250 个家庭，目的就是构建一个荷兰人群体变异数据库，从而对这个群体进行深入的研究。

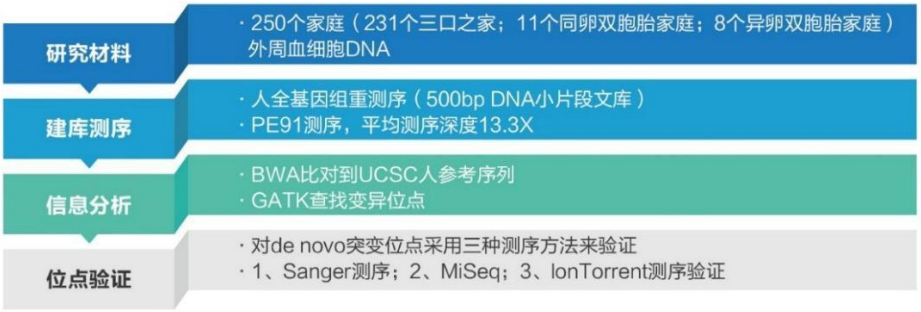


图 8 文章研究思路

重要的研究结果

表 2 荷兰人基因组计划参与者表型信息

	Birth year	*Average age at sampling (year)*	*Average height (cm)*	*Average BMI (kg/m²)*	*Average TC level (mmol/l)*	*Average HDL level (mmol/l)*	*Average LDL level (mmol/l)*	*Average TG level (mmol/l)*
Fathers (*N*=250)	1910–64	63.8 (46–87)	178 (160–198)	26.8 (18.1–39.6)	5.42 (2.98–8.23)	1.24 (0.60–2.30)	3.48 (1.29–5.70)	1.56 (0.10–4.80)
Mothers (*N*=250)	1910–64	61.7 (43–86)	166 (145–182)	27.0 (18.6–38.9)	5.67 (3.10–8.70)	1.51 (0.55–2.46)	3.54 (1.48–6.40)	1.35 (0.16–4.65)
Sons (*N*=105)	1945–89	36.1 (20–59)	183 (167–200)	25.1 (17.8–36.6)	4.86 (2.59–7.20)	1.17 (0.61–1.82)	3.07 (0.83–5.35)	1.44 (0.32–4.58)
Daughters (*N*=163)	1940–94	35.9 (19–58)	171 (156–185)	24.6 (18.1–41.6)	4.74 (2.20–7.61)	1.48 (0.61–2.48)	2.79 (0.70–5.40)	1.09 (0.24–4.50)

250 个家庭的父母年龄范围在 43-87 岁，其中父亲平均年龄 63.8 岁，母亲平均年龄 61.7 岁，子女的年龄范围是 19-59 岁。另外还有身高（子代身高平均比父辈高 5cm）、BMI、总胆固醇（TC）、高密度脂蛋白（HDL）、低密度脂蛋白（LDL）和甘油三酯等指标数据。

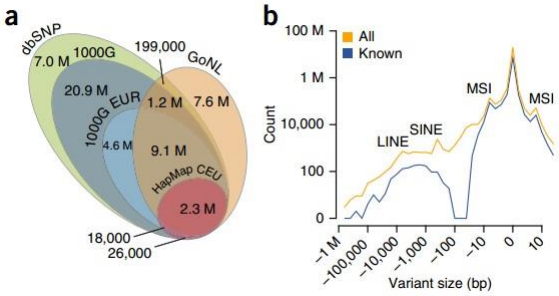


图 9 单碱基变异和结构变异的发现

a 图，荷兰人基因组数据与之前其他数据库发现的变异韦恩图，7.6M 荷兰人基因组特有变异中，大部分是罕见变异（MAF < 0.5%）；b 图，荷兰人基因组发现变异的片段大小分布图。横坐标负数代表的是片段缺失，突变类型包括长插入元件（LINE）、短插入元件（SINE）、微卫星不稳定性（MSI）。橘色是荷兰人基因组的变异数；蓝色的是千人基因组第一阶段的数据，图中可以看出荷兰人基因组计划发现了很多新的变异。

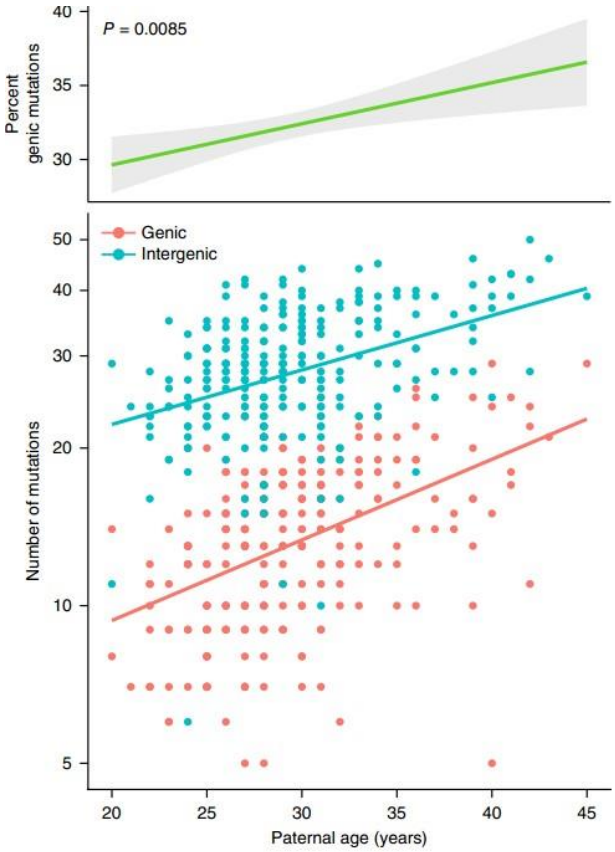


图 10 与父亲年龄相关的子代基因中 *De novo* 突变分布

上图可以直观看到，随着父亲年龄增长，不论是基因内还是基因间的突变都呈现上升趋势。

其他研究结果

（1）发现了 20.4M 的 SNVs，1.2M 的 InDel（小于 20bp），27500 个大的缺失（大于 20bp）。

（2）在发现的 SNVs 中，6.2M 是高频突变，4M 是低频变异，10.2M 是稀有变异。

（3）与参考序列基因型不同的 SNVs 中，有 99.5%与基因分型结果一致。

（4）展示了荷兰人不同地区的人口迁移历史。

（5）随着父亲生育年龄的增加，后代 *de novo* 变异的频率会随之增加。

荷兰人基因组计划的设计亮点

（1）样本选取具有代表性，首先人群是分布在荷兰全国各省，且是以家庭为单位，年龄从 19-87 岁，覆盖范围广。

（2）表型数据全面，这样对新的变异位点与疾病关系的研究提供了有参考价值的数据。

（3）另外样本是外周血细胞 DNA，没有经过体外培养，不会引入其他因素引起的突变。

四、 项目风险

在项目实施过程中，可能存在由于样本群体样本过小、数据覆盖度不够等因素的影响导致样本数据不具有统计学意义，在这种情况下，一般可以考虑通过增加样本数或增加测序数据等手段进行补充。

五、 华大发表文章

表 3 华大发表文章

研究领域	文献标题	发表期刊
炎黄一号[15]	The diploid genome sequence of an Asian individual	Nature (2008)
古人[2]	Ancient human genome sequence of an extinct Palaeo-Eskimo	Nature (2010)
千人[5]	A map of human genome variation from population-scale sequencing	Nature (2010)
澳洲土著人[3]	An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia	Science (2011)
HapMap[16]	Integrating common and rare genetic variation in diverse human populations	Nature (2010)
荷兰人[11-14,17]	The Genome of the Netherlands: design, and project goals	European Journal of Human Genetics(2013)
	Whole-genome sequence variation, population structure and demographic history of the Dutch population	Nature Genetics (2014)
	Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels	Nature Communications (2015)
	Genome-wide patterns and properties of *de novo* mutations in humans	Nature Genetics (2015)
千人[18-20]	A global reference for human genetic variation	Nature (2015)
	An integrated map of structural variation in 2,504 human genomes	Nature (2015)
	The 1000 Genomes Project: data management and community access	Nature Methods (2012)
UK10K[21-24]	Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel	Nature Communications (2015)
	The UK10K project identifies rare variants in health and disease	Nature (2015)
	Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture	Nature (2015)
	Height-reducing variants and selection for short stature in Sardinia	Nature Genetics (2015)

六、 华大优势

- **项目经验丰富。**目前华大基因已经完成 11 万多例人全基因组重测序，从建库实验、测序，到信息分析都有非常规范的流程和丰富的项目经验。华大基因实验室已经通过多项国际质量体系认证，包括质量管理体系 ISO9001:2008、医学实验室管理体系 CAP 认证。
- **计算实力雄厚。**BGI online 已经上线，可以在 1 天内完成千人基因组（外显子）分析，性能比传统分析模式提升数倍。且 BGI online 是一个开放的信息分析平台，集数据存储、数据分析、数据交付为一体，每个人都可以在这个平台玩转自己的项目数据，并与别人共享数据结果。

七、 参考文献

[1] Yi X, Liang Y, Huerta-Sanchez E, *et al*. Sequencing of 50 human exomes reveals adaptation to high altitude[J]. Science, 2010, 329(5987): 75-78.

[2] Rasmussen M, Li Y, Lindgreen S, *et al*. Ancient human genome sequence of an extinct Palaeo-Eskimo[J]. Nature, 2010, 463(7282): 757-762.

[3] Rasmussen M, Guo X, Wang Y, *et al*. An Aboriginal Australian genome reveals separate human dispersals into Asia[J]. Science, 2011, 334(6052): 94-98.

[4] Reich D, Green R E, Kircher M, *et al*. Genetic history of an archaic hominin group from Denisova Cave in Siberia[J]. Nature, 2010, 468(7327): 1053-1060.

[5] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing[J]. Nature, 2010, 467(7319): 1061-1073.

[6] International HapMap Consortium. A haplotype map of the human genome[J]. Nature, 2005, 437(7063): 1299-1320.

[7] Dhandapany P S, Sadayappan S, Xue Y, *et al*. A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia[J]. Nature genetics, 2009.

[8] Rosenberg N A, Huang L, Jewett E M, *et al*. Genome-wide association studies in diverse populations[J]. Nature Reviews Genetics, 2010, 11(5): 356-366.

[9] 1000 Genomes Project Consortium. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68-74.

[10] Li J Z, Absher D M, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation[J]. science, 2008, 319(5866): 1100-1104.

[11] Boomsma D I, Wijmenga C, Slagboom E P, *et al*. The Genome of the Netherlands: design, and project goals[J]. European Journal of Human Genetics, 2014, 22(2): 221-227.

[12] Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences[J]. Nature genetics, 2014, 46(8): 919-925.

[13] Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population[J]. Nature Genetics, 2014, 46(8): 818-825.

[14] Francioli L C, Polak P P, Koren A, *et al*. Genome-wide patterns and properties of de novo mutations in humans[J]. Nature genetics, 2015, 47(7): 822-826.

[15] Wang J, Wang W, Li R, *et al*. The diploid genome sequence of an Asian individual[J]. Nature, 2008, 456(7218): 60-65.

[16] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations[J]. Nature, 2010, 467(7311): 52-58.

[17] Van Leeuwen E M, Karssen L C, Deelen J, et al. Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels[J]. Nature communications, 2015, 6.

[18] 1000 Genomes Project Consortium. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68-74.

[19] Sudmant P H, Rausch T, Gardner E J, *et al*. An integrated map of structural variation in 2,504 human genomes[J]. Nature, 2015, 526(7571): 75-81.

[20] Clarke L, Zheng-Bradley X, Smith R, *et al*. The 1000 Genomes Project: data management and community access[J]. Nature methods, 2012, 9(5): 459-462.

[21] Huang J, Howie B, McCarthy S, *et al*. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel[J]. Nature communications, 2015, 6.

[22] UK10K Consortium. The UK10K project identifies rare variants in health and disease[J]. Nature, 2015, 526(7571): 82-90.

[23] Zheng H F, Forgetta V, Hsu Y H, *et al*. Whole [hyphen] genome sequencing identifies EN1 as a determinant of bone density and fracture[J]. Nature, 2015, 526(7571): 112-117.

[24] Zoledziewska M, Sidore C, Chiang C W K, *et al*. Height-reducing variants and selection for short stature in Sardinia[J]. Nature genetics, 2015.

疾病免疫组库研究方案

一、 研究背景

T、B 细胞是人体主要的淋巴细胞，分别负责细胞免疫和体液免疫，成熟过程中，这些细胞经历了可变区（V）、多样区（D）和接合区（J）基因片段的重排，以便形成独特的序列，编码 B 细胞免疫球蛋白和 T 细胞受体结构。T 细胞受体（TCR）和 B 细胞受体（BCR）由多条肽链组成，具有抗原结合特异性，每条肽链的互补决定区（CDR，又称超变区）氨基酸组成和排列顺序呈现高度多样性，构成容量巨大的 TCR 和 BCR 库。其中 CDR1 和 CDR2 都是由 V 基因编码，而 CDR3 则是由部分 V 基因片段、D 基因片段和 J 基因片段重组后编码形成，这也决定了 CDR3 的多样性要远大于 CDR1、CDR2。免疫组库研究重点主要集中在研究 CDR 基因的多样性上。

免疫组库应用方面很广，在病理研究上，涉及到和免疫相关的疾病几乎都可以从免疫组库找到研究思路，例如自身免疫疾病、感染类疾病、癌症、HIV 等；医学应用上，疫苗研发评价、药物研发、疾病诊断、器官和干细胞移植等。

1.1 免疫组库在肿瘤方面的应用[1-6]

肿瘤细胞是具有免疫原性的，会引起 T 淋巴细胞浸润到肿瘤组织中，即浸润淋巴细胞（Tumor-infiltrating lymphocytes，TIL），多项研究表明肿瘤组织中 TIL 的存在及数量与病人生存期相关，因此 TIL 也许可作为一种肿瘤预后的生物标记物。免疫组库技术已用于多种肿瘤类型的 TIL 研究，例如结直肠癌、卵巢癌、肝癌、黑色素瘤。

1.2 免疫组库在微小残留病（Minimal Residual Disease，MRD）检测的应用[7-9]

微小残留病指的是治疗后的病人体内仍残留的少量白血病细胞，它是导致白血病复发的主要因素。因此，微小残留病的检测对淋巴系统肿瘤的预后至关重要。免疫组库测序比流式分选技术表现出了更高的灵敏性和特异性，已被逐步用于 MRD 的临床监控上。

1.3 免疫组库在自身免疫性疾病方面的应用[10-12]

自身免疫性疾病是由于机体对自身抗原发生免疫反应而导致自身组织损害所引起的疾病。因此，利用免疫组库技术可分析异常表达的免疫细胞（T 细胞或抗体），作为临床诊断的标志物。例如 Klarenbeek 等人[10]分析了早期和长期类风湿性关节炎病人外周血和关节滑膜液的免疫组库，发现早期与长期病人的 T 细胞克隆的差异性。Robert Winchester 等[11]发现 EBV 引发的 TCR 克隆在多发性硬化症病人的脑脊液（CSF）中富集。

1.4 免疫组库在其他方面的应用

除了以上应用，免疫组库在干细胞移植[13-14]、用药[15]、罕见病方面[16]也有许多应用。比如干细胞移植前，患者接受化疗导致严重的免疫缺陷，需要几个月或几年的时间恢复，特别是对于 T 细胞，T 细胞长期缺乏的患者易发生感染和导致癌症复发。因此重建和恢复 TCR 的多态性，能够很大程度地减少移植相关的发病率和死亡率。Van Heijst JW 等评估了不同来源干细胞的 T 细胞库恢复情况[13]。

二、 方案设计

2.1 整体思路

本研究方案是针对疾病研究的大概研究思路，目的是为客户提供一个了解该技术在疾病领域研究应用的总体思路。具体疾病的方案需要结合实际情况（临床意义、研究目的等）进行调整（详见样品选择）

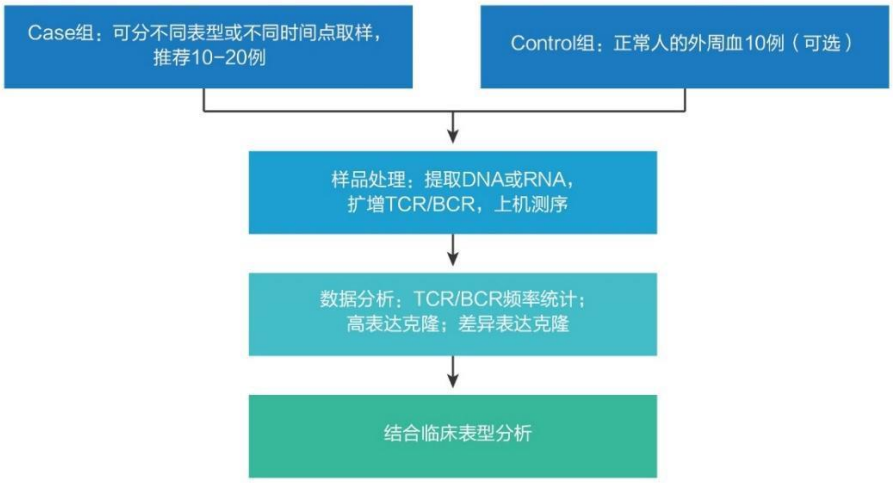


图 1 整体研究思路

注：

1、如需研究不同类型 T 细胞（CD4+、CD8+ T cell 等）在疾病发展不同时期、用药不同时间点、不同取样位置的差异，可在 case/control 组中加入不同细胞类型的分组。

2、case 组可设置相近疾病样品，例如肝癌亚型、肝炎，或者临床上易混淆疾病，寻找不同疾病诊断的 biomarker。

2.2 样本建议

可分不同疾病表型或不同时间点，例如移植前、后不同时间点，疾病发展不同时期，具有相近表型的不同疾病，不同癌症及癌症亚型等，每个点取样 10-20 例，样品类型外周血、T/B 细胞、炎症组织、癌组织等。

2.3 实验技术

采用多重 PCR 技术扩增 TCR/BCR CDR3 区域的编码基因，通过 PE101/151（数据量至少 2M raw reads）的测序策略得到完整的 CDR3 编码基因序列。

2.4 分析结果

（1）两种统计模型评估 V-J 基因多样性

针对测序数据结果序列，使用 ImMunoGeneTics （IMGT） 数据库进行比对，鉴定出 V、D、J 基因，通过配对找到 V-J 克隆的分布，并使用辛普森系数和香农-威纳指数对样品免疫多样性进行评估。

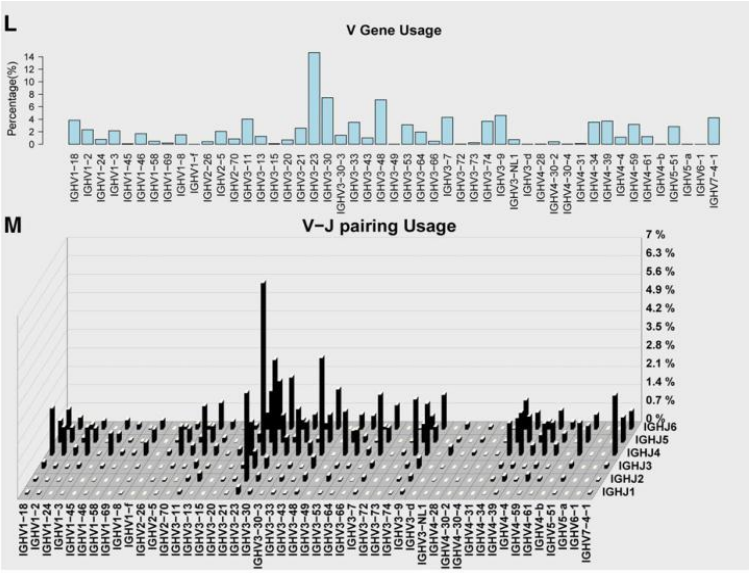


图 2 样品 BCR 库数据统计[17]

L：样品中V 基因用量分析，横坐标代表B 细胞重链V 基因的分类，纵坐标代表每种V 基因的占比。

M：样品中配对的 V-J 基因用量分析三维图，三个维度分别代表B 细胞重链V 基因子类、J 基因子类、配对的V-J 占比。

（2）针对不同表型的样品从 DNA、氨基酸、VJ 家族多个维度进行差异和特异克隆筛选。

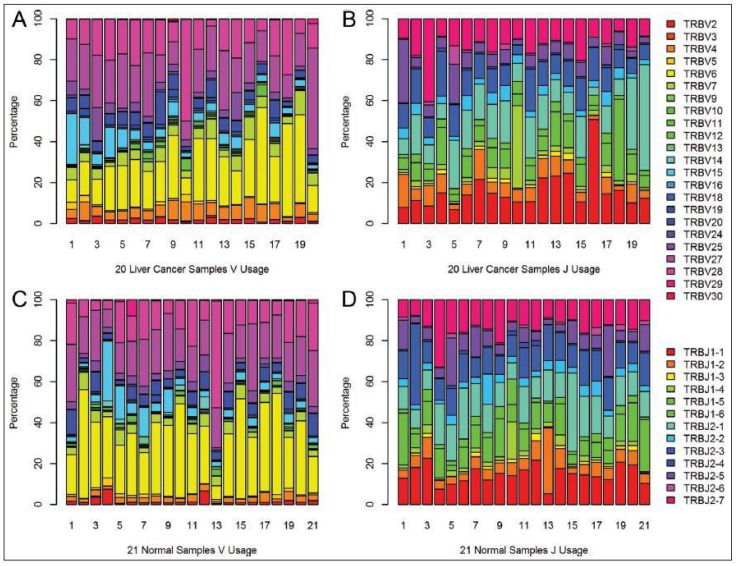


图 3 肝癌病人和正常人血液样品中V 基因和J 基因的半定量结果[18]

A. 20 个肝癌病人中V 基因片段的用量； B. 20 个肝癌病人中J 基因片段的用量；
C. 21 个健康人中V 基因片段的用量；D. 21 个健康人中J 基因片段的用量。

三、 应用案例

案例 1 抗 CTLA-4 单抗药物治疗患者的生存期与 T 细胞克隆稳定性有关[15]

本文选取了 25 例前列腺癌病人、21 例黑色素瘤病人和 9 例正常人，在用药前、用药后 1 个月分别取样，通过多重 PCR 技术扩增 TCRβ链的基因，分析抗 CTLA-4 单抗药物的抗肿瘤机制。研究发现，该单抗药物提高了 TCR 克隆的多样性，并且随着治疗进行，克隆也发生着改变。与临床结果对比发现，治疗前存在的高表达克隆持续稳定的病人，生存期较长，预示着治疗前就已经存在高效抗肿瘤的 T 细胞群体。

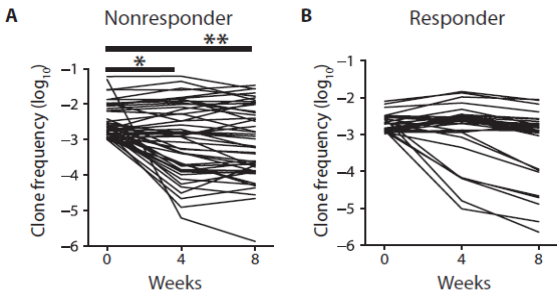


图 4 T 细胞克隆类型变化与生存期关系

选取 10^{-3} 以上 preexisting high-frequency clonoypes，对比发现，生存期短的病人高频克隆变化显著，而生存期长的病人高频克隆类型变化不显著。

案例 2　B 细胞可通过多发性硬化症病人的血脑屏障进行交换[12]

多发性硬化症（MS）是以中枢神经系统白质炎性脱髓鞘病变为主要特点的自身免疫病，其中枢神经系统和外周血中均存在抗原刺激的 B 细胞，两者之间关系如何一直未研究清楚。本文选取 6 个多发性硬化症病人和 7 个其他类神经系统疾病的病人的脑脊液和外周血，扩增 BCR 重链，进行进化分析，推测了多发性硬化症病人的中枢神经系统和外周血中 B 细胞的发展过程。

图 5 研究思路图

案例 3　定量评估干细胞移植后 T 细胞的恢复情况[13]

干细胞移植前，患者会接受化疗，这会导致严重的免疫缺陷，需要几个月或几年的时间来恢复，特别是对于 T 细胞。T 细胞长期缺乏的患者易发生感染和导致癌症复发，因此重建和恢复 TCR 的多态性，能够很大程度上减少移植相关的发病率和死亡率。

病例选择了骨髓干细胞、外周血干细胞、脐带血干细胞移植的患者共 28 例，在移植后 6 个月、12 个月分别取外周血 8ml 进行分析，对照选取供体 5 个人按相应时间点取样。文章采用 5′RACE 方法，回答了 TCR 恢复相关的两个基础问题：TCR 多态性的恢复时间是怎样的，不同来源的干细胞功能如何。

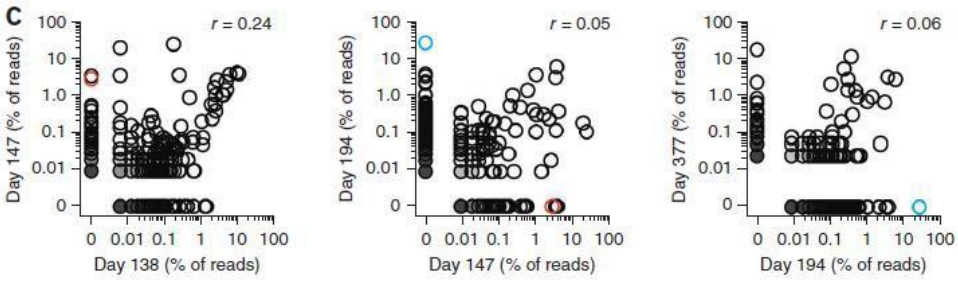


图 6 不同时间点的T 细胞克隆频率统计：138 天，147 天，194 天，377 天

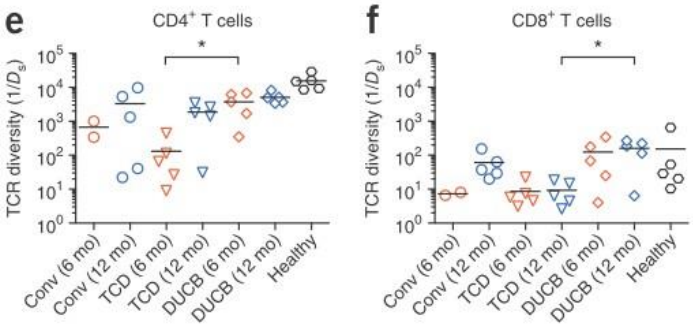


图 7 不同来源干细胞移植后的 T 细胞克隆恢复情况

四、　项目风险

项目设计不合理，可能会导致差异表达克隆不显著。因此，在项目执行前，需要充分了解疾病的背景，设计合理的疾病组。

五、 常见问题

问题 1 分选细胞需要提供多少细胞量？

答：至少需要 1×10^6 个细胞，细胞量越高，免疫细胞的多样性越高。

问题 2 免疫组库的样品类型是什么？

答：样品可提供全血、T/B 细胞、PBMC、提取好的 DNA/RNA。

六、 华大优势

- **丰富的项目经验。**已完成包括肿瘤、复杂疾病、移植等不同领域的项目，可提供从项目设计到个性化信息分析等全方位的服务，并已协助客户发表多篇免疫组库相关文章。
- **优化引物设计。**完成了多重 PCR 引物的优化，更精确的反映免疫组库克隆情况，其中部分引物设计已申请专利。
- **自主的信息分析软件。**分析流程中加入纠错环节，降低了扩增和测序错误，与其他分析软件比，具有更高的准确性。

七、 华大发表文章

表 1 华大发表免疫组库方面文章

研究领域	文献标题	发表期刊
肝癌	Identification of characteristic TRB V usage in HBV-associated HCC by using differential expression profiling analysis	Oncoimmunology (2015)
综述	Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma	Cancer Letters (2015)
信息分析方法学	IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis.	Genetics (2015)
实验方法学	Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire	PLoS One (2016)

八、 参考文献

[1] Gooden M J M, de Bock G H, Leffers N, *et al*. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis[J]. British journal of cancer, 2011, 105(1): 93-103.

[2] Han Y, Liu X, Wang Y, *et al*. Identification of characteristic TRB V usage in HBV-associated HCC by using differential expression profiling analysis[J]. Oncoimmunology, 2015, 4(8): e1021537.

[3] Sherwood A M, Emerson R O, Scherer D, *et al*. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue[J]. Cancer Immunology, Immunotherapy, 2013, 62(9): 1453-1461.

[4] Hwang W T, Adams S F, Tahirovic E, *et al*. Prognostic significance of tumor-infiltrating T cells in ovarian cancer: a meta-analysis[J]. Gynecologic oncology, 2012, 124(2): 192-198.

[5] Emerson R O, Sherwood A M, Rieder M J, *et al*. High- throughput sequencing of T- cell receptors reveals a homogeneous repertoire of tumour- infiltrating lymphocytes in ovarian cancer[J]. The Journal of pathology, 2013, 231(4): 433-440.

[6] Azimi F, Scolyer R A, Rumcheva P, *et al*. Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma[J]. Journal of Clinical Oncology, 2012, 30(21): 2678-2683.

[7] Wu D, Sherwood A, Fromm J R, *et al*. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia[J]. Science translational medicine, 2012, 4(134): 134ra63-134ra63.

[8] Faham M, Zheng J, Moorhead M, *et al*. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia[J]. Blood, 2012, 120(26): 5173-5180.

[9] Martinez-Lopez J, Lahuerta J J, Pepin F, *et al*. Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma[J]. Blood, 2014, 123(20): 3073-3079.

[10] Klarenbeek P L, De Hair M J H, Doorenspleet M E, *et al*. Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease[J]. Annals of the rheumatic diseases, 2012, 71(6): 1088-1093.

[11] Lossius A, Johansen J N, Vartdal F, *et al*. High- throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV- reactive CD8+ T cells[J]. European journal of immunology, 2014, 44(11): 3439-3452.

[12] von Büdingen H C, Kuo T C, Sirota M, *et al*. B cell exchange across the blood-brain barrier in multiple sclerosis[J]. The Journal of clinical investigation, 2012, 122(12): 4533-4543.

[13] van Heijst J W J, Ceberio I, Lipuma L B, *et al*. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation[J]. Nature medicine, 2013, 19(3): 372-377.

[14] Warnke C, Mausberg A K, Stettner M, *et al*. Natalizumab affects the T-cell receptor repertoire in patients with multiple sclerosis[J]. Neurology, 2013, 81(16): 1400-1408.

[15] Cha E, Klinger M, Hou Y, *et al*. Improved survival with T cell clonotype stability after anti–CTLA-4 treatment in cancer patients[J]. Science translational medicine, 2014, 6(238): 238ra70-238ra70.

[16] Wu J, Liu D, Tu W, *et al*. T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome[J]. Journal of Allergy and Clinical Immunology, 2015, 135(1): 209-216. e8.

[17] Zhang W, Du Y, Su Z, *et al*. IMonitor: a robust pipeline for TCR and BCR repertoire analysis[J]. Genetics, 2015, 201(2): 459-472.

[18] Han Y, Liu X, Wang Y, *et al*. Identification of characteristic TRB V usage in HBV-associated HCC by using differential expression profiling analysis[J]. Oncoimmunology, 2015, 4(8): e1021537.

复杂疾病的 MHC 易感位点研究方案

一、 研究背景

MHC，即主要组织相容性复合体，在人类中又被称作 HLA，也就是人类白细胞表面抗原，编码该抗原的基因位于 6 号染色体 6q21.3 区域，包括一系列紧密连锁的基因座，与人类免疫系统功能密切相关，涉及肿瘤、自身免疫性、免疫缺陷性、过敏性、感染类和代谢性等几乎所有复杂疾病的发生、发展，被称为人类的"主要抗原"。目前已发现这一区域至少与人类的 100 多种疾病相关，包括甲状腺癌、瘢痕瘤、开角型青光眼、自闭症、贫血症、风湿性关节炎、小儿哮喘、男性不育等各类复杂疾病及癌症。携带某一特定 HLA 型别的个体会对特定疾病表现为易感性或抗性，阐明该区与疾病的关系一直是生命医学研究领域的热点和难点。

MHC 区域与免疫疾病高度相关，该区域基因密度大，高多态性，高同源性。芯片上 MHC 区域位点有限，全基因组重测序费用太高。MHC capture 技术的出现，使得免疫类疾病可以更加有针对性的进行 MHC 区域上小范围的关联分析。本方案就是以 MHC capture 为技术基础，以复杂疾病 MHC 区域为关联对象，提供一种经济实惠的易感位点研究方案。

HLA-I 类基因区：远离着丝点一端，含 122 个基因座，其中，B、C、A 是经典的 I 类基因，又称 HLA-Ia 基因，编码相应的 HLAI 类分子的重链。

HLA-II 类基因区：近着丝点一端，含 34 个基因座，DR、DP、DQ 是经典的Ⅱ类基因，其编码产物均为双肽链（α、β）分子。

HLA-III 类基因区含 62 个基因座，包括编码某些补体（C2、C4、Bf）、CK（TNF-α、LT）、热休克蛋白 70（HSP70）等产物的基因，以及参与调节 NF-KB 活性的 NFKBIL1 基因等。

该图为 MHC 区域基因分布情况，主要分为 HLA-I,HLA-II,HLA-III 三类基因。

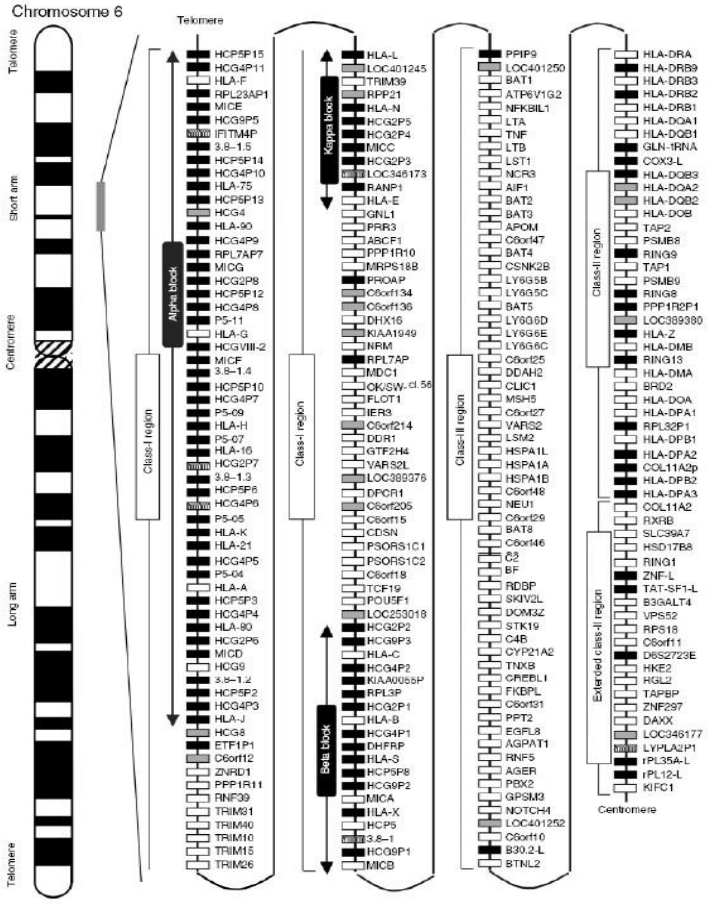


图 1 MHC 区域基因分布情况

二、 方案设计

2.1 研究目的

本方案拟解决，挖掘人不同疾病的 MHC 区域易感基因。

2.2 样本建议

（1）血液（case ，control 对） case 为疾病样本，control 为正常人样本。

（2）样品数。一期关联分析：常见变异（100-500 对）；罕见变异（2000~10000 对）；二期大样本量验证及关联分析：常见变异（500-1000 对）；罕见变异（10000 对以上）

2.3 测序参数

1G Raw Data （约 80×）,PE126 。

2.4 整体思路

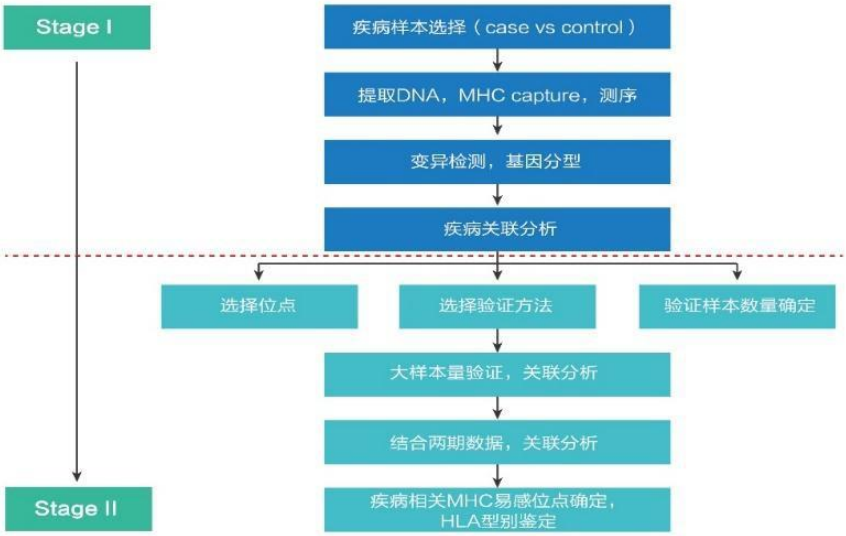


图 2 整体研究思路

Stage1 目的：通过 MHC 捕获测序找到 MHC 易感候选位点；

Stage2 目的：选择正确的验证方法，大样本量验证易感位点的可靠性。

2.5 实验技术

Stage1：MHC capture 芯片捕获技术。

首先血液中细胞基因组 DNA 被打断到 500bp 左右长度，回收建库;文库富集后，与 MHC 芯片进行杂交，收集目的片段，上机测序 PE126。

样品检测	? 基因组DNA较完整
样品打断	? 3ug基因组，打断
文库构建	? 末端修复，加“A”，加接头
片段选择	? 片段回收
文库富集	? LM-PCR进行捕获前文库扩增
杂交	? 用MHC芯片进行杂交
洗脱	? 将非目标片段洗掉并回收目的片段
捕获文库富集	? LM-PCR进行捕获后文库扩增
上机测序	? Illumina HiSeq 测序，测序类型PE126

图 3 MHC 捕获测序实验流程

Stage2：根据待验证位点数量，需要分型的基因种类，使用不同的技术（PCR+Sanger 测序、MHC 捕获测序、芯片分型等）进行验证。

2.6 项目周期

100 个样品，平均深度 80X，按标准流程（包含样品检测、建库测序和信息分析），约 50 个工作日（不包括芯片定制时间），实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.7 预期结果

利用 MHC capture 芯片捕获技术捕获 MHC 区域基因，高深度二代测序，关联分析，找到在 MHC 区域复杂疾病相关易感基因并验证。

三、 应用案例

案例 华大参与——深度测序揭示了 MHC 区域对复杂疾病研究的重要意义[1]

此项研究是首次对中国人群的 MHC 区域进行大规模样本深度测序。通过自主研发的 MHC 目标区域捕获芯片，对 20635 例人群样本的 MHC 目标区域进行高深度测序和分析，成功建立世界上最大样本量的中国汉族人群 MHC 全区域完整遗传变异数据库。通过对 MHC 区域准确的序列分析和基因分型，完整展示出中国人群 MHC 区域突变位点和 HLA 基因的多态性图谱，为开展中国人群复杂疾病与 MHC 区域的相关性研究奠定坚实基础。

表 1 Han-MHC 数据库的基本数据统计

Samples	10,689		
Mean coverage (×)	55.15 ± 12.15		
Mean SNPs	14,680 ± 1,670		
Mean indels	2,075 ± 326		
	Rare <0.5%	**Low frequency 0.5–5%**	**Common ≥5%**
SNPs			
Total sites	177,255	18,188	29,429
New sites (%)[a]	73.45	5.44	0.19
Synonymous	3,422	229	412
Nonsynonymous	6,728	368	668
Stop gain	192	2	11
Stop loss	10	3	1
Disease-associated variation			
OMIM	179	75	187
HGMD	18	9	42
Short indels[b]			
Total sites	29,109	2,789	4,402
New sites (%)[a]	81.93	28.25	13.38
Frameshift	949	17	15
Non-frameshift	88	0	1
Stop gain	8	0	0

[a]Not observed in dbSNP142 or the 1000 Genome Project Phase III data release. [b]Less than 100 bp in length.

该研究以银屑病为案例，首次采用测序的方法对 MHC 区域遗传变异在中国汉族银屑病人群发病中的相关性进行了全面、系统的大样本研究，发现了 MHC 区域 *HLA-C*，*HLA-B*，*HLA-DPB1* 和 *BTNL2* 等多个新的银屑病易感基因和位点，同时揭示了中国人群特异的银屑病易感位点。

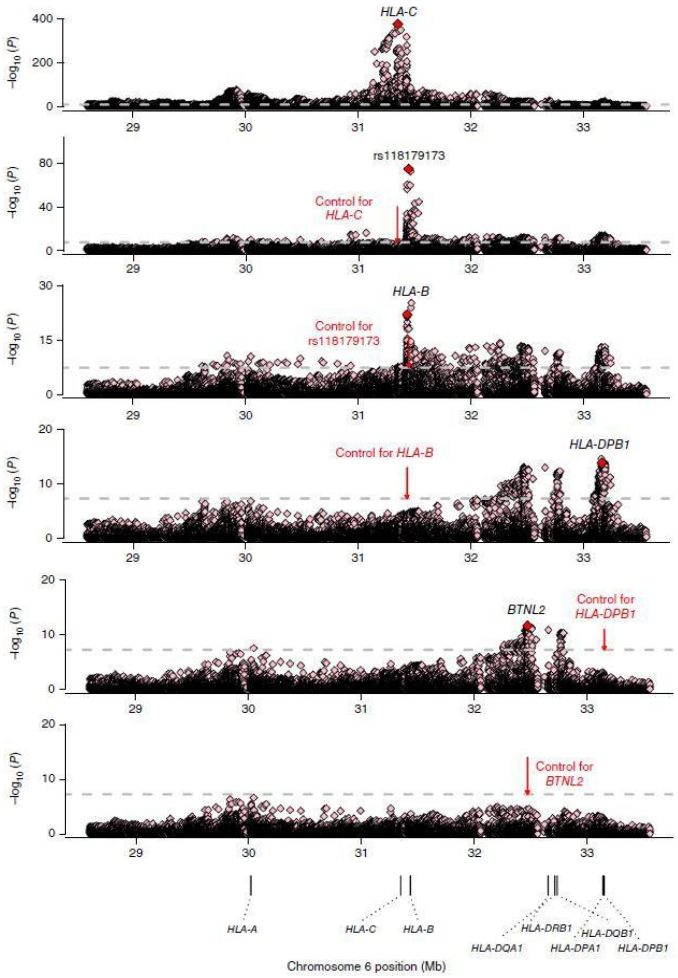


图 4 银屑病 MHC 区域的逐步条件关联分析（Stepwise conditional association）结果

进一步证明了基于完整测序构建的 Han-MHC 数据库在中国人群复杂疾病研究中的巨大应用价值。为探索 MHC 区域遗传多态性在自身免疫性疾病、感染性疾病、肿瘤、过敏性疾病等所有与 MHC 相关的复杂疾病中的作用机制，提供了一个完整的遗传变异比对图谱。该研究发现的 MHC 区域遗传变异靶标，将为免疫相关疾病的预测、早期诊断、药物疗效和副作用观察，以及靶向治疗的精准医学研究提供重要的参考数据。

四、 项目风险

因样本量不够或疾病涉及易感基因较多较复杂，芯片本身未覆盖所有易感基因，需加大研究样本量或增加芯片覆盖区域。

五、 常见问题

问题 1 MHC 捕获试剂盒的捕获区域有多大？

答：捕获区域共 4.97 Mb，包含了传统的 MHC 区域（约 3.37 Mb）及其侧翼区域（约 1.6 Mb），以及 8 个单体型信息。

问题 2 测序一般推荐多大的数据量？

答：一般推荐至少做 80X 以上。

问题 3 MHC 的主要应用领域有哪些？

答：与 HLA 变异相关的自身免疫病及多种复杂疾病、癌症等的研究；HLA 精准分型-器官移植配型；疫苗、药物等的靶向人群筛选；亲子鉴定和法医学；种族进化等。

问题 4 华大基因有相关 MHC 的发表文章吗？

答：由安徽医科大学、复旦大学与华大基因研究院共同开展的人类主要组织相容性复合体（MHC）区域全覆盖深度测序完成。该研究成功构建出迄今最完整的中国汉族人群 MHC 遗传变异数据库，并阐述了其在复杂疾病研究中的重要价值。2016 年 5 月 23 日，国际著名杂志《自然·遗传》以论著的形式在线发表了此项研究。详见参考文献[1]。

六、 华大优势

- **丰富的项目经验。**上万个 MHC 样品测序经验，大项目经验丰富，联合安徽医科大学、复旦大学构建出迄今最完整的中国汉族人群 MHC 遗传变异数据库，并阐述了其在复杂疾病研究中的重要价值。
- **行业领先的捕获技术。**华大基因的 MHC 芯片是和 NimbleGen 联合研发的，测试结果与基因分型 SNP 一致率高达 99.42%，覆盖度达到 97%。
- **全面覆盖 MHC 区域。**总共 4.97Mb，包含传统 MHC 区域 3.37 Mb 及 1.6 Mb 侧翼序列，并包含了 8 个单体型信息 PGF，APD，COX，DBB，MANN，MCF，QBL，SSTO。
- **个性化信息分析方案。**对于有特殊信息分析要求的客户，能根据客户需求制定个性化信息分析方案，信息分析团队会安排专人跟进，确保客户需求得到满足。
- **更清晰的结题报告。**结果图片更加直观，可用于文章撰写；结题报告结构优化，有效提升阅读体验。
- **性价比更高的科研方案。**华大基因自主研发的 MHC 芯片可实现多杂建库，降低成本，实现 MHC 全面区域覆盖，性价比超高。

七、 参考文献

[1] Zhou F, Cao H, Zuo X, *et al*. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease[J]. Nature genetics, 2016.

[2] Tang M, Lautenberger J A, Gao X, *et al*. The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove[J]. PLoS Genet, 2012, 8(11): e1003103.

[3] Bei J X, Li Y, Jia W H, *et al*. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci[J]. Nature genetics, 2010, 42(7): 599-603.

疾病甲基化分子标记物研究方案

一、 研究背景

近几十年，随着新的 DNA 甲基化检测技术建立和应用，DNA 甲基化（DNA methylation）在人类遗传表达及基因调控异常所致疾病中的作用，得到进一步证实。DNA 的甲基化对维持染色体的结构、X 染色体的失活、基因印记以及许多疾病的发生都起着重要的作用。许多基因的甲基化，被证明与各种临床疾病密切相关， 有些甚至可以被确定为疾病致病的独立因素。

DNA 甲基化，尤其是特定基因启动子部位的 DNA 甲基化，被认为是基因表达的开关。DNA 甲基化位点主要是 5 ̄-CpG-3’中胞嘧啶 C5，而且 CpG 在基因组中分布不均匀，主要存在于重复序列与 CpG 岛中，CpG 岛主要存在于管家基因（housekeeping gene）和组织特异性表达基因的启动子区。一般情况下，CpG 岛甲基化会造成基因沉默，某些预定程序以外的甲基化作用，会使某些功能基因的表达关闭或降低，产生包括致病等严重后果。

启动子区 CpG 岛的甲基化，常常可以抑制该基因的转录，对抑癌基因，凋亡相关基因，DNA 修复基因等的作用尤其明显，启动子区域发生甲基化会影响特定功能的发挥。因此，基因启动子区的异常甲基化可以作为一种高灵敏度的 Biomarker，该区域甲基化在疾病发展机制的研究中受到越来越多的重视，也是目前新研究热点之一。

二、 方案设计

2.1 研究目的

寻找特定疾病相关的基因启动子区 DNA 异常甲基化模式，作为一种高灵敏度的 Biomarker 对疾病进行检测，或研制使异常甲基化还原的药物等。

DNA 甲基化与众多疾病的发生和发展密切相关，比如Ⅱ型糖尿病、自身免疫疾病以及各种癌症。特别是癌症，其发生过程中癌细胞基因组整体的低甲基化和局部区域的超甲基化是其典型特征[1]。超甲基化体现在抑癌基因的启动子区被异常甲基化，整体的欠甲基化同重复序列(如转座子)以及癌基因的启动子区的甲基化程度减小、基因组遗传不稳定性的增加密切相关。

Bibikova 等[2]对结肠癌、乳腺癌、肺癌和前列腺癌细胞系以及正常细胞中 371 个基因中的 1536 个位点的甲基化状态进行聚类分析，发现了 55 个可以区分癌细胞和正常细胞以及各种癌细胞类型的位点。该方法表明，各种癌症都存在其特异的 DNA 甲基化生物标记。

2.2 研究思路

图 1 整体研究思路

2.3 样本建议

原则上分成两组选取，病例组和对照组。

如研究对象为癌症的，可取癌组织和癌旁组织，进而对癌组织中，甲基化程度升高的 marker 进行筛选，找到可以作为癌症诊断相关的 Biomarkers。

有研究者使用 Bisulfite sequencing 的方法，取结肠癌组织细胞以及正常结肠组织进行研究和对比，证实了在结肠癌细胞和组织中，有两个基因的启动子区发生了高甲基化，而正常结肠组织中未见相应甲基化，进而认为其可作为结肠癌早期预测的 Biomarker[3]。

2.4 实验技术

捕获区域的选择：CG 岛（CpG islands, CGI）的启动子区域。

此类区域的甲基化可能性最大，可直接调控基因的转录活性[4–6]。DNA 甲基化在维持基因结构和功能的完整性方面起着重要的作用，因为在一些内源性的逆转录因子中，含有一些高度甲基化的启动子，这些区域的去甲基化，会导致肿瘤细胞中某些基因的启动作用激活，促进异常基因的表达以及新陈代谢[7,8]。

2.5 测序参数

目标区域甲基化测序技术介绍：目标区域甲基化测序（Targeted Bisulfite Sequencing, TBS）技术，利用探针杂交捕获目标区域，鉴定相关区域 C 位点的甲基化水平，并通过成对样品间的比较以及筛选，寻找差异甲基化区域，并对差异甲基化区域所在基因的功能进行分析。

目标区域甲基化测序技术特点：可以对基因组的特定区域进行甲基化分析，例如启动子区，而全基因组甲基化测序是对全部基因组的甲基化水平进行分析。相对于全基因组甲基化测序，TBS 测序量低，可增加研究的样品数，并且得到的信息更加集中和有针对性。

2.6 预期结果

（1）通过甲基化水平分析对病例组样本和对照组样本进行比较分析，找出病例组的 DNA 甲基化特征。不同分布类型的 C 碱基(CG、CHG 和 CHH)，其甲基化水平在不同样本间，甚至同一样本的不同细胞类型、不同条件下都存在差异。下图显示的是每种类型 (CG、CHG 和 CHH)甲基化水平分布，反映了该物种 DNA 甲基化特征。

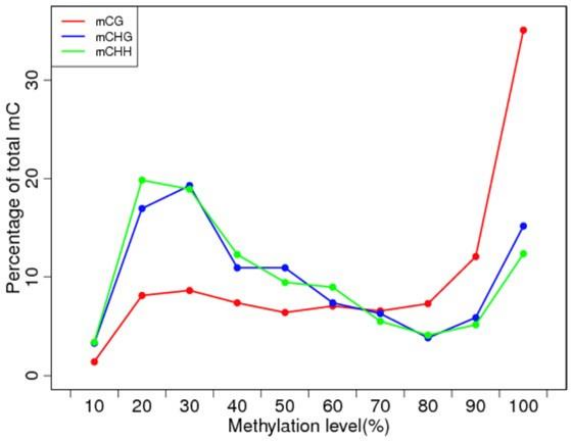


图 2 mC 的甲基化水平分布图

横轴表示甲基化水平，从左往右为 0%～100%，每 10%为一档，纵轴表示特定甲基化水平的 mC 在全部 mC 中所占比例。

（2）根据不同基因组区域的甲基化分布特征，可定位至病例样本组里，基因组区域内所有患病样本中，共有的甲基化分布特征，从而缩小疾病相关的甲基化区域范围。从分析结果的热图能够体现甲基化分布、CpG 密度分布以及甲基化和 CpG 密度之间关联的信息。下图中以热图的形式表示基因组不同特征区域的甲基化水平，有助于进一步了解这些区域的甲基化特征。

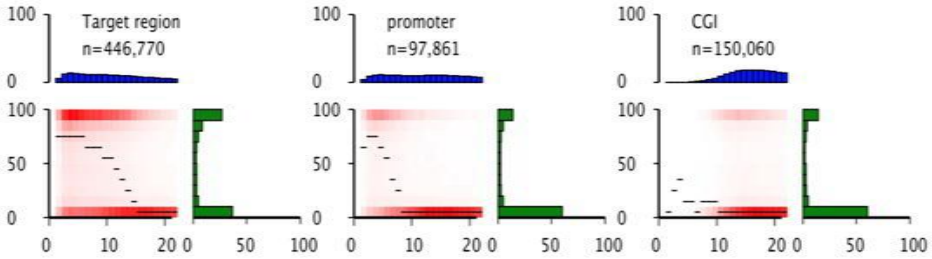


图 3 CGI、启动子以及目标区域的甲基化热图

n 为该原件中包括的 CpG 位点数（深度≥4），CpG 密度（横坐标）为 200bp 窗口中 CpG 位点的个数，甲基化水平（纵坐标）为 CpG 中 C 位点的平均甲基化水平。每个热图中的黑线为在特定 CpG 密度下的甲基化水平中位数。红色区域从浅到深表示特定甲基化水平和 CpG 密度下的CpG 丰度。上方的蓝色柱状图表示 CpG 密度的分布,映射到横轴上，右侧的绿色柱状图表示甲基化水平的分布，并且映射到纵轴上。

（3）差异甲基化区域(DMR)分析，初步对疾病相关的 DMR 进行筛选。通过差异甲基化分析，可以发现疾病组与对照组中的差异甲基化区域，后续可进行功能分析。例如，子代的甲基化特征与父本或母本的甲基化状态一致，亲本与子代甲基化模式的差异常常导致表观遗传缺陷[9]。

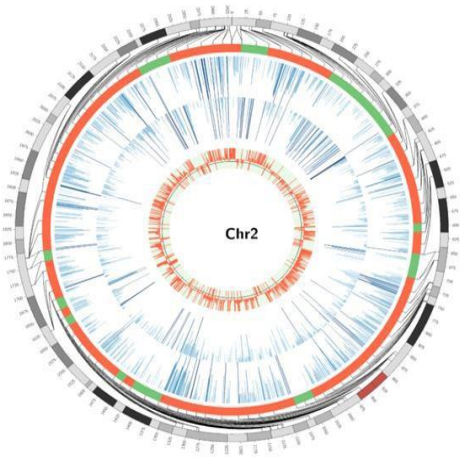


图 4 DMRs 的甲基化水平分析。

最外圈表示基因组染色体的位置；第二圈为 DMR 区域，红色区域代表样品 1 的甲基化水平高于样品 2，绿色区域表示样品 1 的甲基化水平低于样品 2；第三圈表示样品 1 各个位点的甲基化率；第四圈表示样品 2 各个位点的甲基化率；第五圈表示样品 1 和 2 甲基化率的差异程度。

（4）一般情况下，DNA 的甲基化，会导致基因沉默，从而导致部分基因功能失活，通过 DMR 相关的基因进行GO 和Pathway 分析，根据GO 基因功能分析结果与KEGG (Kyoto Encyclopedia of Genes and Genomes)

的 Pathway 分析结果，定位到疾病的致病基因功能失活紧密相关的 DMRs，进而找到疾病相关的 DNA 甲基化 Biomar。

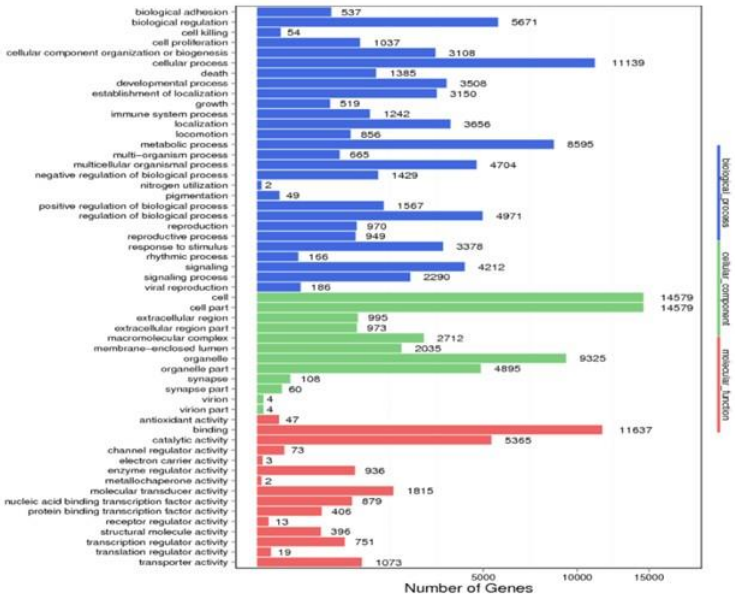


图 5 DMR 相关基因的GO 聚类分析

图形横轴表示DMR 相关基因的数量，纵轴表示各种 GO 条目，所有 GO 条目分为三类，蓝色为生物学过程，绿色为细胞组分，红色为分子功能。

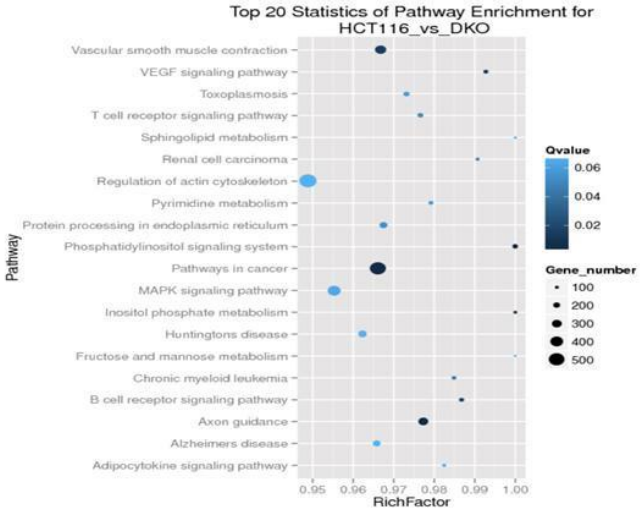


图 6 DMR 相关基因的Pathway 功能显著性富集分析

横轴（rich factor）是在通路中 DMR 相关基因占所有基因总数的比例，Rich Factor 值越高则在该 pathway 中越富集。Q 为修正的 p value，其值为 0-1，Q 值越小，富集度越高。在此图中仅显示前 20 个富集通路。

三、 应用案例

案例 华大参与——通过研究单绒毛膜双胎胎盘组织启动子区甲基化，揭示了胎儿宫内生长发育迟缓相关的甲基化图谱[10]

样本是来自 19 例单绒毛膜双胎的脐带组织样本，其中 13 例的胎盘组织被临床诊断为选择性胎儿宫内发育迟缓（Selective intrauterine growth restriction, sIUGR），作为实验组；6 例取自正常胎盘组织作为对照组。对其特定区域进行捕获甲基化测序，找到疾病相关的 Biomarker。

研究材料

单绒毛膜双胎的脐带组织部位共 19 例样本；13 例被临床诊断为 sIUGR；6 例取自正常胎盘组织。

甲基化处理与目标区域捕获

华大专利的 Promoter 探针设计，转录起始位点(transcriptional start sites)的上游 2200bp 到下游 500bp 的区域覆盖率达 91.8%。

文库构建测序

小片段文库，Hiseq2000，5.73Gb clean data/sample

结果分析

（1）胎盘组织中特定的基因启动子区的甲基化水平分析：得到病例组与对照组的差异甲基化图谱（图 7），再做分区细化，按照 CpG 的统计比率，分为高中低(HCP, ICP, LCP)三个区，进一步定位，sIUGR 的 DMRs 主要发生在中区(ICP)和低区(LCP)（图 8）

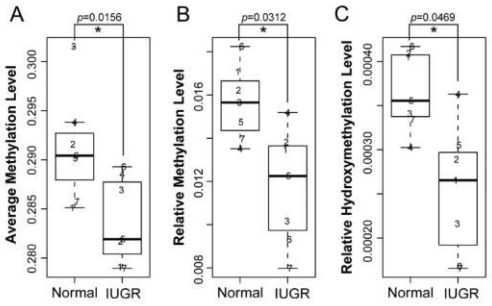


图 7 甲基化水平分析

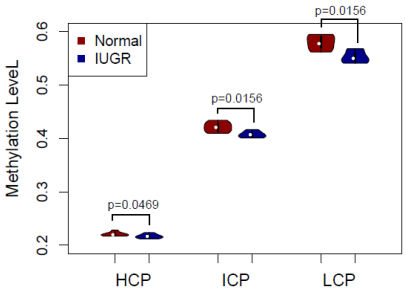


图 8 分区甲基化水平

（2）基于甲基化水平的 PCA 分析，结果发现 IUGR 胎盘明显与健康胎盘聚成两个类别（图 9A）由于主要

的差异是来自 ICP 和 LCP 两种类型的，所以文章针对了 CGI（图 9B）以及非 CGI（图 9C）的启动子的平均甲基化水平做了进一步的分层次聚类分析，结果也显示了与 PCA 结果的趋势是一致的。DMSs 结果的无监督聚类分析可清晰区分出形态病理学诊断为 IUGR 的胎盘组织以及形态病理学诊断为健康的组织（图 9D）。文章数据还显示了 IUGR 胎盘组织的 DNA 去甲基化大部分是发生在非 CGI 启动子区域和 CGI 岛。总而言之：研究都表明，IUGR 胎盘的启动子区具有不同于健康胎盘的甲基化水平，这一特征可作为新的 Biomarker，从而监控并改变胎儿在孕妇体内的成长状态。

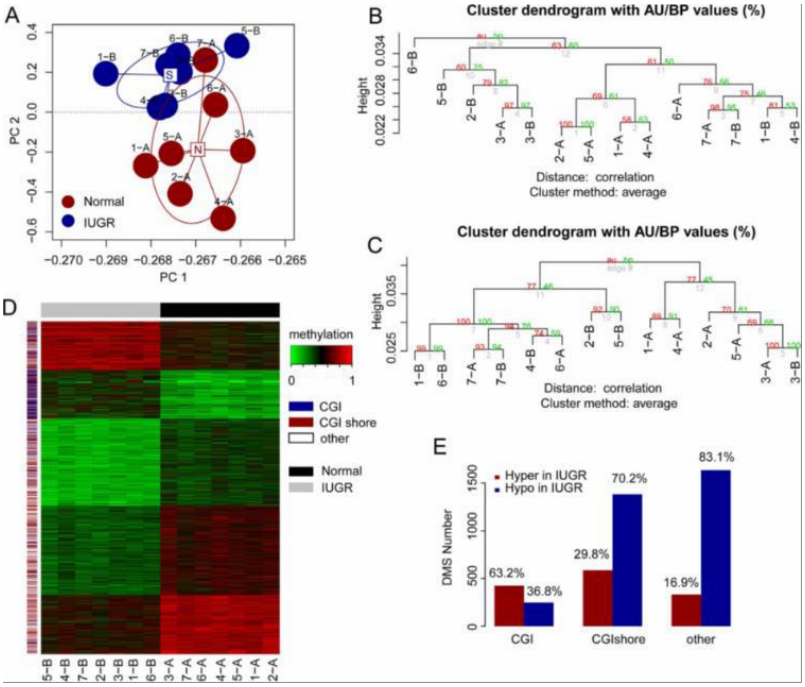


图 9 IUGR 胎盘特异性启动子区域差异甲基化结果的聚类分析

A：PCA 分析；B，C：层次聚类分析；D：聚类分析图；E：DMR 分布统计结果

（3）KEGG 和 GO 功能分析的结果显示，差异的甲基化水平主要是通过影响胎盘母体部分 DNA 结合的活力从而影响母体的生理代谢（图 10）

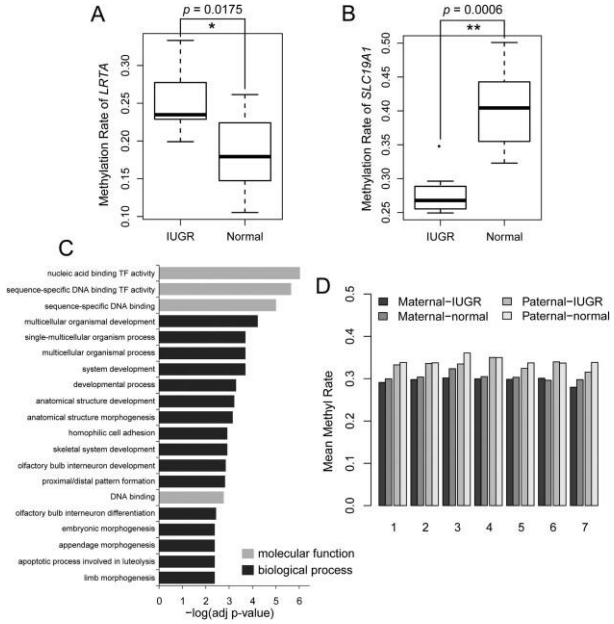


图 10 KEGG 以及 GO 功能分析

A B 特定基因的甲基化水平比较，C 是差异启动子区 DMRs 的 GO 分析，D 是父系和母系印迹基因的甲基化率对比分析。

四、 项目风险

需对疾病的分子机制有比较深入的了解，对探针有较高的依赖性，区域覆盖度低会影响marker 的筛选结果。

五、 常见问题

问题 1 目标区域甲基化测序项目开始之前需要考虑哪些因素?

答：是否为低甲基化率的物种；该物种的基因组完成情况如何（影响 BS-SEQ 的比对）；基因组是否存在复杂因素：GC 含量偏高、杂合度偏高、转座子、重复区域等。

问题 2 我们是否可以捕获任意物种的任意区域，并进行 TBS 研究?

答：在理论上，我们可以通过 NimbleGen SeqCap Epi Enrichment System 和 Agilent SureSelectXT Methyl-Seq Target Enrichment System 对任意物种的研究区域进行捕获 BS 测序研究。当然，实际上还需要考虑上个问题所提到的情况。

问题 3　如何评价 TBS 产品的准确性?

答：通过评估,我们可以确定 TBS 产品得到的甲基化结果和 WGBS 及 Illumina 450K 甲基化芯片结果有很高的相关性，其中与 WGBS 的数据比较，相关系数可达 0.925。

问题 4 为什么 NimbleGen SeqCap Epi Enrichment 探针对目标区域有极高的敏感性?

答：该探针使用创新的探针生产工艺，针对完全甲基化、部分甲基化、完全没有甲基化的情况分别进行独立设计，从而提高了它的敏感度。

问题 5 Bisulfite-Seq 中如何定义支持甲基化的 reads？

答：Reads 是否支持甲基化是针对某一个位点而言的，在这个位点上，若 read1（正链），C 与 C 比对上（即：甲基化的 C，没有发生转化），若 read2（负链），G 与 G 比对上，那么这个 read 在这个位点上就是支持甲基化的 read，反之就是支持非甲基化的 read。

问题 6 Bisulfite-Seq 中如何定义 C 是一个甲基化的C？

答：如果该 C 位点支持甲基化的 reads 数大于等于期望值，则判断为甲基化的 C。

问题 7 Bisulfite-Seq 是否可以统计甲基化频率？

答：可以，但是并不包含在标准信息分析流程中。甲基化频率，就是指相隔多少个碱基（ATCG），会出现一个甲基化的 C，拟南芥和人中都做过相关统计。

六、　华大优势

- **华大注册专利的捕获甲基化高通量检测技术。**专利号为 WO 2012159564 A1 的甲基化高通量检测方法，是一种结合捕获重硫酸测序方法，其它能同时有效地扩大样品的目标或甲基化态降低设计的复增操作过程的行性
- **高准确度。**华大 TBS 产品具有很好的重复性；华大 TBS 产品可以正确检测到已知的 DMR 区也能检测出未知的 DMR 区。
- **超高的性价比。**可对任意物种任意区域的甲基化状态进行正确检测，相对于全基因组甲基化测序，成本更低。
- **99%以上的 Bisulfite 转化率**，单碱基分辨率，与全基因组甲基化测序(WGBS)数据相关性高达 0.925。

七、 参考文献

[1] Jones P A, Baylin S B. The fundamental role of epigenetic events in cancer[J]. Nature reviews genetics, 2002, 3(6): 415-428.

[2] Bibikova M, Lin Z, Zhou L, *et al*. High-throughput DNA methylation profiling using universal bead arrays[J]. Genome research, 2006, 16(3): 383-393.

[3] Yi J M, Dhir M, Guzzetta A A, *et al*. DNA methylation biomarker candidates for early detection of colon cancer[J]. Tumor Biology, 2012, 33(2): 363-372.

[4] Baylln S B, Herman J G, Graff J R, *et al*. Alterations in DNA methylation: a fundamental aspect of neoplasia[J]. Advances in cancer research, 1997, 72: 141-196.

[5] Miranda T B, Jones P A. DNA methylation: the nuts and bolts of repression[J]. Journal of cellular physiology, 2007, 213(2): 384-390.

[6] Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome[J]. Human molecular genetics, 2007, 16(R1): R50-R59.

[7] Szpakowski S, Sun X, Lage J M, *et al*. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements[J]. Gene, 2009, 448(2): 151-167.

[8] Estécio M R H, Gallegos J, Vallot C, *et al*. Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer[J]. Genome research, 2010, 20(10): 1369-1382.

[9] Heyn H, Li N, Ferreira H J, *et al*. Distinct DNA methylomes of newborns and centenarians[J]. Proceedings of the National Academy of Sciences, 2012, 109(26): 10522-10527.

[10] He Z, Lu H, Luo H, *et al*. The promoter methylomes of monochorionic twin placentas reveal intrauterine growth restriction-specific variations in the methylation patterns[J]. Scientific reports, 2016, 6.

复杂疾病 *De novo* 突变研究方案

一、 研究背景

De novo 突变又称为新生突变或从头突变，即在一个家系里子代有但父母不具有的突变。*De novo* 突变分为*de novo*体细胞突变与 *de novo* 胚系突变。人类 *de novo* 胚系突变从染色体水平到单碱基水平都有发生，常见的 *de novo* 突变类型包括SNVs（Single-nucleotide variants）、InDels(Small insertions or deletions)，CNVs(Copy number variants)以及 SVs（Structural variants），不同 *de novo* 突变类型的突变率存在差异。文献报道显示，目前人类胚系 *de novo* SNVs 的平均突变率约为 $1.18X10^{-8}$，与父母相比，每个子代基因组产生约 74 个新的 SNVs。*De novo* InDels 以及 *de novo* CNVs 突变频率比 *de novo* SNVs 频率相对要低，因其长度更长，影响更多的碱基对。*De novo* InDels 突变率约为 4×10^{-10}，即每个子代基因组会产生约 3 个在父母中不具有的新的 inDels，其中小片段缺失(Small deletions)突变率约为小片段插入(Small insertions)的 3 倍，且片段越长，突变频率越低；大于 100 kb 的 *de novo* CNVs 突变频率约 1/50，小于 100 kb 的 *de novo* CNVs 突变频率未知。在整个外显子组水平，每个子代会产生约 1 个 *de novo* 突变。影响这些突变频率的因素包括父母性别、年龄、种族等[1]。

De novo 突变比遗传变异经历了更少的进化选择，特定条件下对疾病的贡献度会更大。有假说认为散发性疾病的遗传基础可能不同于很多家族性个体，因为前者更可能是由于 *de novo* 突变而非遗传变异引起[2]。大量的研究结果也证实，*de novo* 突变与散发性的罕见疾病以及精神类疾病的发生密切相关，如下表 1 和表 2 所示。

表 1 常见精神类疾病 *de novo* 突变研究进展

文章	时间	杂志	研究疾病	方法	主要结果	影响因子
Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population [3]	2016.3	Nature genetics	自闭症	对 >38000 个 ASD 及其他样本数据进行关联分析，研究自闭症和精神类疾病相关遗传风险因素分析	利用 LD score 分析和 *de novo* 突变分析，发现多个与自闭症和其他精神类疾病相关的遗传变异	29.352
CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders [4]	2016.5	Mol Psychiatry	自闭症	对 1534 患者和 8936 对照进行 SNP 芯片测序检测 CNV，并对 212 个患者和 217 个 controls 进行 Sanger 测序检测 SNV，最后对 200 个 trios 和 89 个 sib pairs 进行全外显子测序评估	发现两个具有 *de novo* 突变的目的基因	14.496
The contribution of *de novo* coding mutations to autism spectrum disorder [5]	2014.11	Nature	自闭症	2517 个 SSC 家系进行≥40X 外显子组测序	识别出 27 个高置信度的基因目标	41.456
De novo variants in sporadic cases of childhood onset schizophrenia [6]	2016.6	Eur J Hum Genet	精神分裂症	17 个 trios 进行全外显子分析	发现20 个*de novo* 突变	4.349

Biological insights from 108 schizophrenia-associated genetic loci [7]	2014.6	Nature	精神分裂症	研究人员对 36989 名患者和 113075 名正常人金银组进行了关联分析	发现与精神分裂相关的 *de novo* 突变的基因与自闭症，智障的相关基因有	41.456
Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes [8]	2014.11	PLoS One	精神分裂症	53 个 trios 进行全外显子测序	共发现 49 个 *de novo* 突变，18 个候选致病基因	3.234
De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects [9]	2016.4	PLoS Genet	先天性心脏病	对 59 个患病 trios 和 59 个健康 trios，100 个患病个体和 533 个健康个体进行全外显子测序	在 *NR1D2* 上发现一个 *de novo* 突变	7.528
De novo mutations in histone-modifying genes in congenital heart disease [10]	2013.5	nature	先天性心脏病	对362 个case 和 264 个 control 的 trios 家系进行全外显子测序	发现 *de novo* 突变导致了 10%的先天性心脏病案例	41.456
De novo loss- or gain-of-function mutations inKCNA2 cause epileptic encephalopathy [11]	2015.3	nature genetics	癫痫	首先用包含 265 个已知和 220 个候选癫痫基因的 panel 对 33 个患者进行目标区域测序，之后对 86 个trios 进行全外显子重测序	找到 4 个不同的 *de novo* 突变在 KCNA2 中	29.352

De novo mutations in epileptic encephalopathies [12]	2013.8	nature	癫痫	对 264 个癫痫 trios 家系进行全外显子测序	共发现 329 个 *de novo* 突变	41.456
Exome sequencing of case-unaffected-parents trios reveals recessive and *de novo* genetic variants in sporadic ALS [13]	2015.3	Sci Rep	ALS	对 44 个ALS trio 家系进行平均 52.5X 外显子测序	在 27%的患者中发现 81 个 *de novo* 突变，其中 54 个通过 Sanger 验证	5.578
Genome sequencing identifies major causes of severe intellectual disability [14]	2014.6	nature	智障	对50 个智障病人家系进行全外显子测序，CNV 芯片分析	找到84 个*de novo* SNVs，8 个 *de novo* CNVs	41.456

表 2 罕见疾病类 *de novo* 突变研究进展

文章	时间	杂志	研究疾病	方法	主要结果	影响因子
De Novo Mutations in PDE10A Cause Childhood-Onset Chorea with Bilateral Striatal Lesions [15]	2016.4	Am J Hum Genet	舞蹈症	对3 个独立患者和其中两个患者的健康父母进行全外显子测序	发现位于 *PDE10A* 上的一个杂合 *de novo* 突变	10.931
De novo PMP2 mutations in families with type 1 Charcot-Marie-Tooth disease [16]	2016.3	Brain	Type 1 Charcot-Marie-Tooth disease	对1 个病人进行全外显子测序，并对136 个患者进行筛查	发现一个 *de novo* 突变在 *PMP2* 上	9.196
De novo GABRA1 mutations in Ohtahara and West syndromes [17]	2016.2	Epilepsia	Qhtahara 和 West 综合症	526 位和 145 位患者分别进行全外显子测序和 GABRA1 的目标区域捕获测序	发现 5 个 *de novo* 突变在 *GABRA1* 上	4.571
Mutations in LTBP3 cause acromicric dysplasia and geleophysic dysplasia [18]	2016.4	J Med Genet	肢端发育不良	全外显子测序	2 个 *de novo* 杂合突变在 *LTPB3*	6.335
De novo mutations in PLXND1 and REV 3L cause Möbius syndrome [19]	2015.7	Nature Communications	Möbius 综合症	对 2 个 trios 家系和 6 个单独患者进行全外显子测序	找到 2 个含有 *de novo* 突变的基因 *PLXND1* 和 *REV3L*	11.47

二、 方案设计

2.1 整体思路

图 1 *De novo* 突变的设计方案

2.2 样本建议

本研究方案中疾病样本选择比较特殊，建议人重测序样品选择>30 组，外显子测序样品选择>50 组 trios 家系（正常父母＋1 个患病孩子）或 quads 家系（正常父母＋患病同卵双胞胎）。

对于自闭症等这些 de novo 已经广泛被研究的精神类疾病，需要更大的样本量，具体要根据文献调研结果而定；对于研究罕见疾病的 de novo 突变，需要 1-3 个家系即可。

2.3 实验技术

全基因组重测序或者外显子测序。De novo SNVs 检测采用了改进过滤法以及基于机器学习的方法 forestDNM 联合分析，可以获得传统过滤法不能找到的 de novo 突变且验证率更高。

2.4 测序参数

全基因组重测序>50X；外显子测序>100X。

2.5 预期结果

旨在对极端罕见疾病或精神类疾病的家系样本进行高通量测序，检测 de novo 突变，解析de novo 突变在疾病发生过程中的作用。

2.6 分析流程

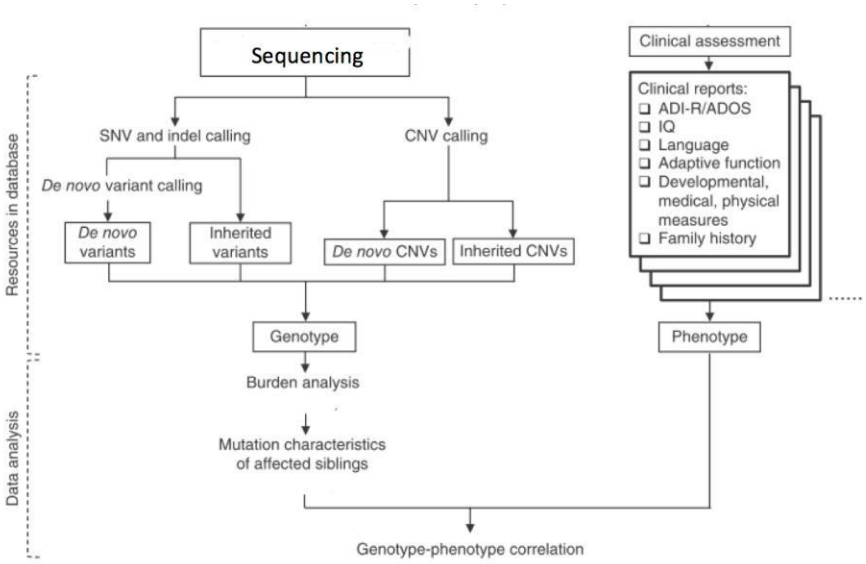


图 2 *De novo* 突变检测信息分析流程图

全基因组重测序可以分析 SNVs，InDels，CNVs，SVs；外显子重测序只可分析 SNVs，InDels。

2.7 项目周期

样品检测合格后，建库+测序+标准信息分析：约 40 个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.8 实验验证

1. 技术验证：采取 Sanger 与质谱验证。
2. 大样本量验证：建议>30 个 trios 或 quads 家系；>100 个对照样本。
3. 功能验证：生物学实验层面，采用动物模型或者基因产物表达水平研究等。

三、 应用案例

案例 1 华大参与——全基因组测序发现基因胚系突变与自闭症风险的关联[20]

文章对 10 对同卵双胞胎的自闭症患者及其正常父母进行全基因组 40X 测序。采用机器学习工具 forestDNM 和 Hard filtering 联合分析的新方法，共发现668 个潜在germline DNMs（*De novo* mutations），并采用Sanger sequencing 与 Sequenom genotyping 验证，父母子代验证了 652 个位点，565 个位点（87%）确定为 DNMs。对不确定为 DNMs 的 87 个 SNPs 进行了校验，其中 34 个（6%）是假阳性突变，其余 53 个（9%）是由于在父或母中的遗传 SNP 变异检测错误，结果确定了 forestDNM 方法的准确性。此外，研究人员巧妙设计父亲年龄偏大（>40 岁）和偏小（<30 岁）两组各五个家系进行研究，结果表明，父亲的年龄越大，其后代患自闭症的风险也越大。研究进一步表明胚系新生突变在基因组中表现出高度的非随机性，还对全基因组范围的变异全貌进行了研究，为深入探索疾病致病机制提供了新的研究思路。

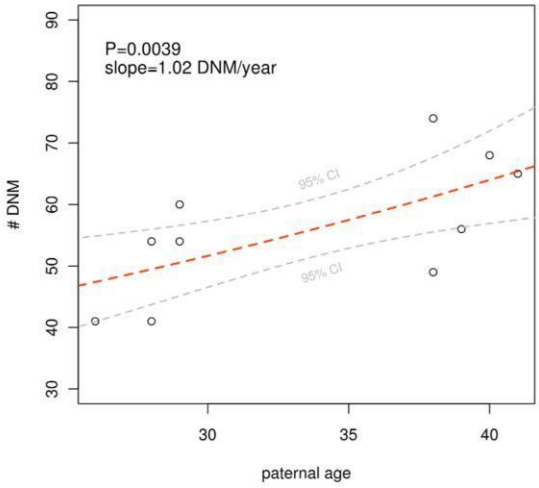


图 3 父亲年龄与 DNMs 数量关系图

案例 2 华大参与——全基因组重测序揭示自闭症罕见遗传变异 [21]

文章对 32 个自闭症谱系障碍（ASD）家系进行至少每样本 30X 的全基因组重测序。研究人员在 32 个家系中发现 6 个家系（19%）存在有害的 *de novo* 突变，并且找到 10 个家系（31%）有 X 染色体连锁或常染色体变异，比例均高于之前的相关报道。研究人员还发现了 4 个新的有害突变，9 个已知突变和 8 个可能的有害突变，包括 *CAPRIN1* 和 *AFF2*，VIP 及其它基因如 *SCN2A* 和 *KCNQ2*，*NRXN1* 和 *CHD7*，这些基因引起了 ASD

相关的 CHARGE 综合征。

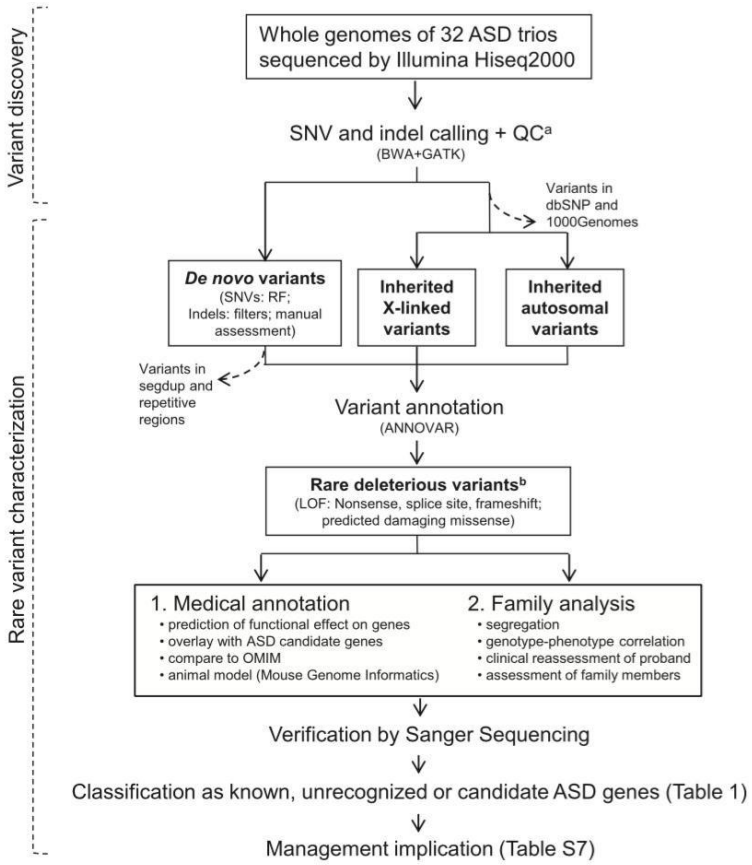


图 4 文章分析及验证流程图

案例 3 外显子组测序的自闭症相关遗传因素分析 [22]

研究人员对 2517 个 SSC 家庭(其中每个都有一个患 ASD 的孩子)，2508 个患病孩子，1911 个未患病的兄弟姐妹和父母进行了≥40X 外显子组测序。识别出 27 个高置信度的基因目标，并且估计 13%的新生错义突变和 43%的“可能会破坏基因的”(LGD)新生突变分别造成所诊断出病例中的 12%和 9%。无论患病女性的 LGD 的靶区是否与患病男性的重合，研究结果都显示，在 IQ 患者中，男性的 IQ 要高于女性患者。

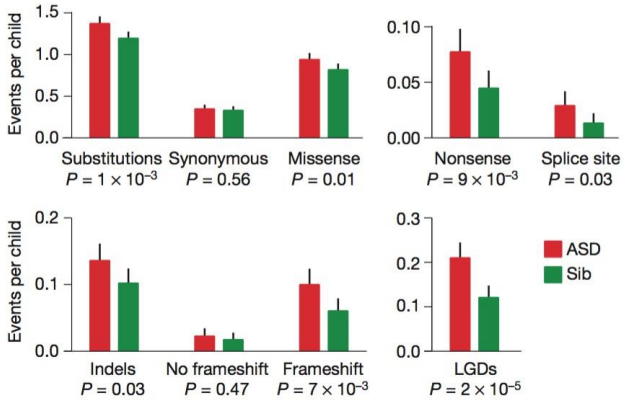


图 5 SSC 中 *de novo* 突变类型频率

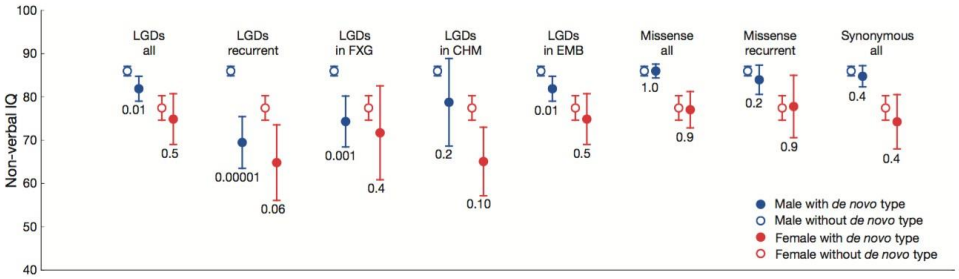


图 6 *De novo* 突变与非语言 IQ 关系

四、 项目风险

在项目实施过程中，由于家系样本量偏小，可能会导致 *de novo* 突变的验证率低，建议加大样本量测序。

五、 常见问题

问题 已知的与 *de novo* mutation 相关的疾病有哪些？

答：目前与 *de novo* mutation 相关的已知疾病有 ALS（肌萎缩性侧索硬化症）、自闭症、精神分裂症、癫痫、智障、先天性心脏病及与发育相关的罕见病等。

六、 华大优势

- *De novo* SNVs 检测采用了改进过滤法以及基于机器学习的方法 forestDNM 联合分析，*de novo* InDels 检测采用了改进过滤法。信息分析方法的升级，可以获得传统过滤法不能找到的 *de novo* 突变且验证率更高。

表 3 *De novo* 突变检测信息分析方法比较

传统过滤法	改进过滤法	forestDNM *
人为设立过滤阈值	人为设立过滤阈值（更严格)	数据驱动
偏向性大	偏向性较小	降低偏向性
假阳性高	假阳性较高	假阳性低
技术验证率低（~20%）	技术验证率较低（~80%）	技术验证率高（~90%）

- ForestDNM 基于数据驱动，能降低偏向性，促进 de novo 突变分析流程优化，其主要目的是将真正的 de novo 突变与由于测序、比对或突变 calling 错误引起的假阳性 de novo 突变区分开，能有效降低候选 de novo SNVs 的假阳性，提高技术验证率。

七、 参考文献

[1] Veltman J A, Brunner H G. De novo mutations in human genetic disease[J]. Nature Reviews Genetics, 2012, 13(8): 565-575.

[2] O'Roak B J, Deriziotis P, Lee C, *et al*. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations[J]. Nature genetics, 2011, 43(6): 585-589.

[3] Robinson E B, St Pourcain B, Anttila V, *et al*. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population[J]. Nature genetics, 2016, 48(5): 552-555.

[4] Mercati O, Huguet G, Danckaert A, *et al*. CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders[J]. Molecular psychiatry, 2016.

[5] Iossifov I, O'Roak B J, Sanders S J, *et al*. The contribution of de novo coding mutations to autism spectrum disorder[J]. Nature, 2014, 515(7526): 216-221.

[6] Ambalavanan A, Girard S L, Ahn K, *et al*. De novo variants in sporadic cases of childhood onset schizophrenia[J]. European Journal of Human Genetics, 2015.

[7] Ripke S, Neale B M, Corvin A, *et al*. Biological insights from 108 schizophrenia-associated genetic loci[J]. Nature, 2014, 511(7510): 421.

[8] Guipponi M, Santoni F A, Setola V, *et al*. Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes[J]. PloS one, 2014, 9(11): e112745.

[9] Priest J R, Osoegawa K, Mohammed N, *et al*. De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects[J]. PLoS Genet, 2016, 12(4): e1005963.

[10] Zaidi S, Choi M, Wakimoto H, *et al*. De novo mutations in histone-modifying genes in congenital heart disease[J]. Nature, 2013, 498(7453): 220-223.

[11] Syrbe S, Hedrich U B S, Riesch E, *et al*. De novo loss-or gain-of-function mutations in KCNA2 cause epileptic encephalopathy[J]. Nature genetics, 2015, 47(4): 393-399.

[12] Phenome E, Epi4K Consortium. De novo mutations in epileptic encephalopathies[J]. Nature, 2013, 501(7466): 217-221.

[13] Steinberg K M, Yu B, Koboldt D C, *et al*. Exome sequencing of case-unaffected-parents trios reveals recessive and de novo genetic variants in sporadic ALS[J]. Scientific reports, 2015, 5.

[14] Gilissen C, Hehir-Kwa J Y, Thung D T, *et al*. Genome sequencing identifies major causes of severe

intellectual disability[J]. Nature, 2014.

[15] Mencacci N E, Kamsteeg E J, Nakashima K, *et al*. De Novo Mutations in PDE10A Cause Childhood-Onset Chorea with Bilateral Striatal Lesions[J]. The American Journal of Human Genetics, 2016, 98(4): 763-771.

[16] Motley W W, Palaima P, Yum S W, *et al*. De novo PMP2 mutations in families with type 1 Charcot–Marie–Tooth disease[J]. Brain, 2016, 139(6): 1649-1656.

[17] Kodera H, Ohba C, Kato M, *et al*. De novo GABRA1 mutations in Ohtahara and West syndromes[J]. Epilepsia, 2016.

[18] McInerney-Leo A M, Le Goff C, Leo P J, *et al*. Mutations in LTBP3 cause acromicric dysplasia and geleophysic dysplasia[J]. Journal of Medical Genetics, 2016: jmedgenet-2015-103647.

[19] Tomas-Roca L, Tsaalbi-Shtylik A, Jansen J G, *et al*. De novo mutations in PLXND1 and REV3L cause Mobius syndrome[J]. Nature communications, 2015, 6.

[20] Michaelson J J, Shi Y, Gujral M, *et al*. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation[J]. Cell, 2012, 151(7): 1431-1442.

[21] Jiang Y, Yuen R K C, Jin X, *et al*. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing[J]. The American Journal of Human Genetics, 2013, 93(2): 249-263.

[22] Iossifov I, O'Roak B J, Sanders S J, *et al*. The contribution of de novo coding mutations to autism spectrum disorder[J]. Nature, 2014, 515(7526): 216-221.

复杂疾病肠道宏基因组研究方案

一、 研究背景

人类消化道中居住着大量的微生物，它们被统称为肠道微生物组。肠道微生物组在人类代谢食物、抵御感染和应答药物等过程中起到了重要的作用。许多人类疾病都与微生物组失衡有关。了解菌株水平上的多样性，对于理解肠道菌组成及其对健康和疾病的影响，是非常关键的。

不同样本中同种细菌的不同菌株之间，基因或基因拷贝数会出现很大的差异。这种基因拷贝数变异能显著改变菌株的多种能力，包括耐药性、毒力、从环境摄取生化分子、分解物质获得能量、移动方式等等。这些因素决定着菌株的生活方式，能对人类宿主的健康产生影响。研究人员还发现，一些菌株的基因拷贝数变异与肥胖症、炎症性肠病等疾病有关，具有一定的临床意义。

二、 方案设计

2.1 研究目的

针对肠道样本进行宏基因组测序，构建肠道参考基因集；分析不同疾病程度样本间基因、功能、物种间的差异；将疾病发生、进展或治疗过程与肠道微生物进行关联分析，结合实验验证，深入研究复杂疾病发生机理。为更好的诠释肠道微生物与人类健康提供科学依据，同时也推动肠道微生物研究在临床上的应用，有助于疾病的预防和治疗。

2.2 样本建议

推荐 50 对样本以上（case-control），具体依据研究目的及样品分组情况而定。

样本中可包括不同治疗时期、疾病不同阶段/不同型别、不同治疗方法等。

（1）粪便样品采集建议：样品用粪便专用收集管收集后立即放入干冰或液氮中冷却速冻，建议样本量为拇指大小的粪便，如果是短期保存可存于-20℃，长期保存必须放入-80℃，运输方式采用干冰运输。

（2）对采样人群进行饮食及生活习惯调查。

（3）临床指标采集：主要以血样中各种理化指标以及临床医生认为与疾病相关性较大的各种因素为主。

2.3 测序参数

（1）文库构建策略：350bp 插入片段文库。

（2） 测序策略：测序平台为 Illumina，PE151 读长测序，推荐测序数据量为 5~7G clean data/样本。

2.4 分析内容

（1）基因分析

将测序所得数据进行 De novo 组装，对组装结果进行基因预测、去冗余构建肠道宏基因组参考基因集。

（2）基因功能和物种注释

通过比对公共数据库（包括 nr、Swiss-Prot、COG、KEGG、GO、CAZy、eggNOG 以及 ARDB）对基因集进行注释，获取基因功能和物种注释信息。

（3）基因、功能、物种多样性分析

将 reads 比对回基因集，计算各样品的基因丰度以及物种丰度情况；基于丰度数据，可进行物种多样性分析、PCA 分析、聚类分析、差异分析、功能富集分析等多项分析内容。

（4）疾病关联分析

疾病相关表型信息筛选，鉴定疾病相关 markers，构建分类器。

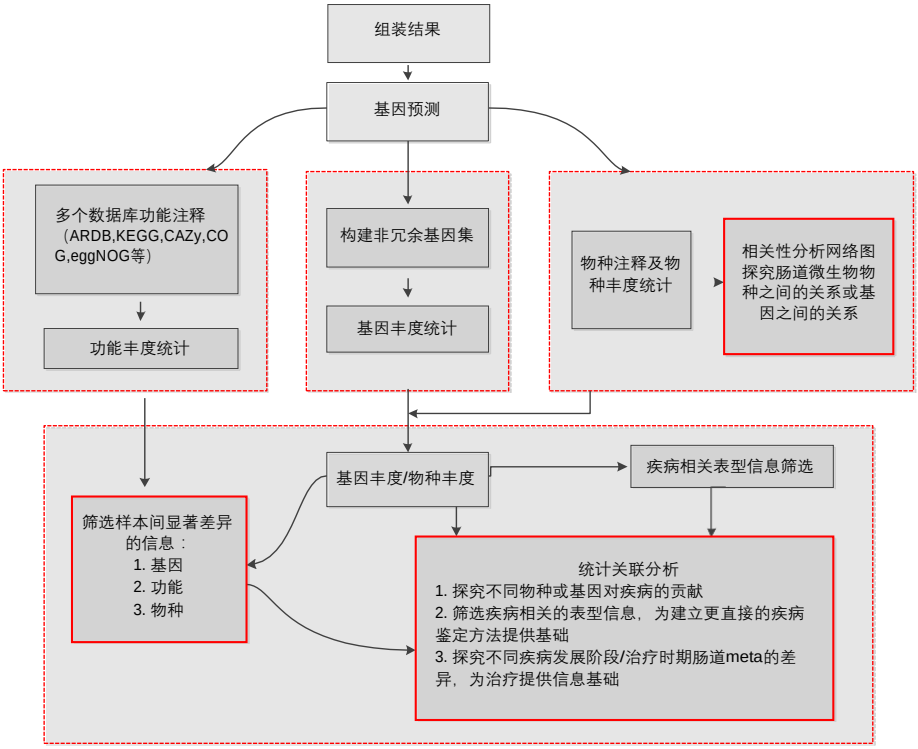


图 1 宏基因组调查流程图

2.5 分析结果

(1) 基因功能注释

所有预测得到的基因使用 blast 软件对公共数据库进行搜索（blast，e-value<0.00001），包括 eggNOG, CAZy, GO, COG, Swiss-Prot, KEGG, ARDB, nr。通过比对，选取相似度最高的蛋白序列并最终得到相应的蛋白功能注释。

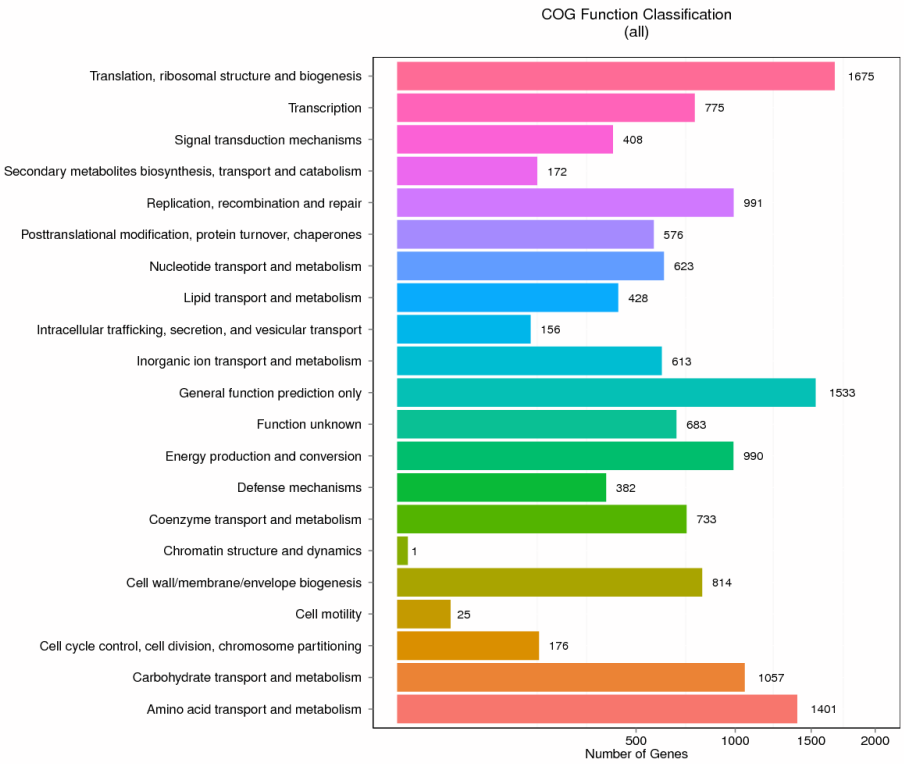


图 2 COG 功能分类图

纵轴表示不同的 COG 功能分类，横轴表示注释到该分类的基因个数

(2) 物种注释及物种丰度统计

使用 MEGAN（version 4.6）对 nr 比对结果进行处理。MEGAN 根据一种名为 LCA 的算法，对 BLAST 比对结果根据 NCBI 物种分类树进行重新梳理。从而得到每条序列的物种注释信息。之后，将同一物种注释的基因丰度累加起来，得到该物种在样品中的物种丰度。

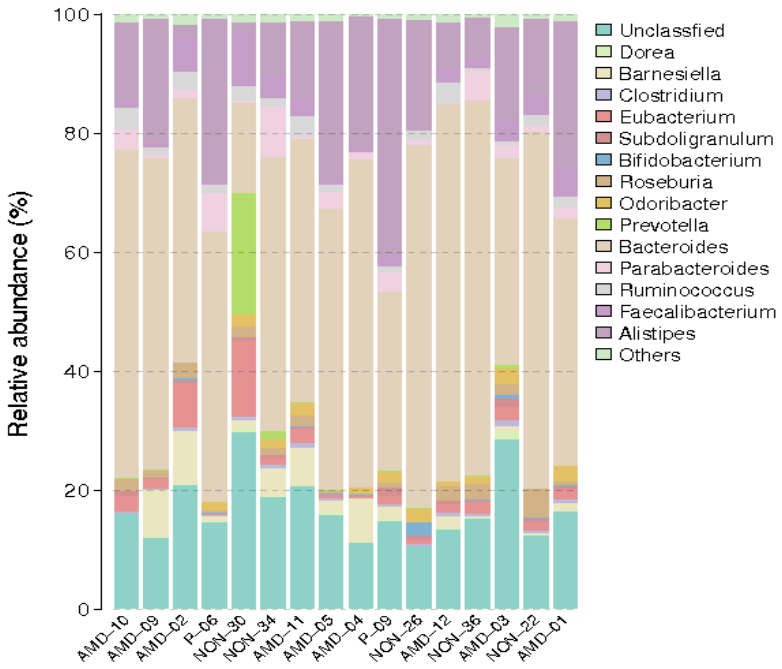


图 3 物种丰度柱状图

图中展示了每个样品的物种组成与分布。每个柱状表示一个样品，不同的颜色表示不同物种。y 轴为相对丰度。

(3) 差异基因分析

差异基因分析旨在找出疾病样本与正常样本间具有丰度差异显著的基因，以便后续更深入的功能挖掘。

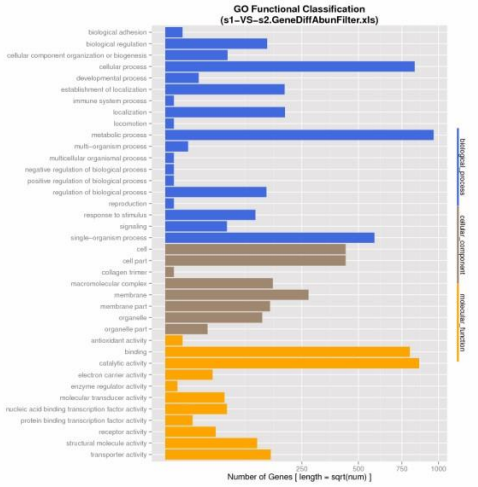


图 4 差异基因的 GO 功能注释分类统计

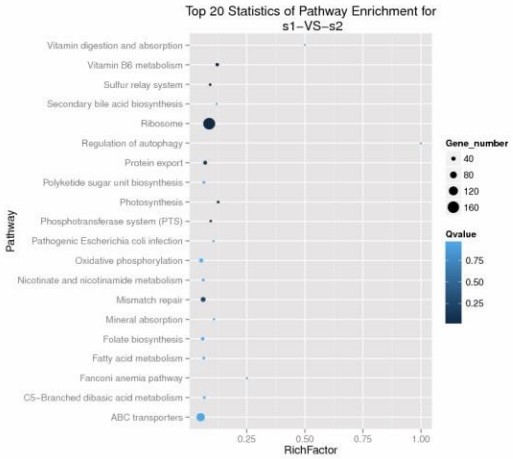


图 5 差异基因 pathway 富集统计散点图

(4) Biomarker 筛选

利用已有的丰度值信息筛选显著 biomarker，并对筛选得到的 biomarker 匹配effect size, 得到疾病样本与正常样本中差异显著的 biomarker。

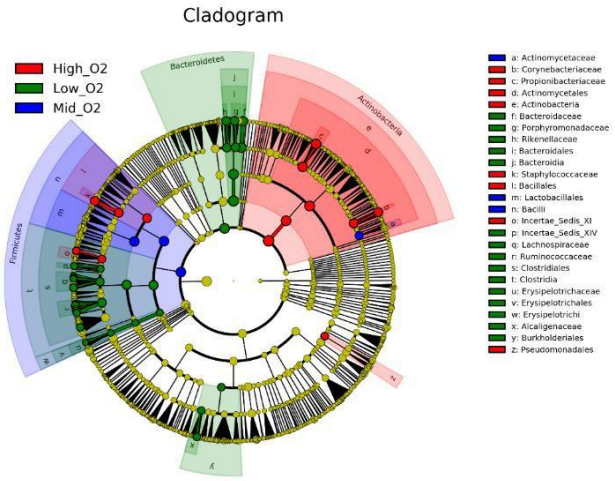


图 6 LEfSe 分析结果

图中为 LEfSe 聚类树，不同颜色表示不同分组，不同颜色的节点表示在该颜色所代表的分组中起到重要作用的微生物群，一个颜色圈点代表一个 biomarker，图例为 biomarker 名称。黄色节点表示的是在不同分组中没有起到重要作用的微生物类群。

2.6 项目周期

样品检测合格后，建库+测序+标准信息分析：约 55 个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.7 预期结果

（1）筛选疾病的物种或基因 marker，建立合适的分类器，来区分开正常人和病人样本。

（2）筛选不同治疗时期疾病样本的微生物差异，推测药物作用机制。

（3）研究疾病发生或治疗效果的个体差异；结合研究背景分析差异产生的因素。

（4）为探究疾病发病机制（微生物方向）及药物治疗前景和治疗方案做初步的探索。

2.8 实验验证

验证分类器对相关样本的效果，验证分类 marker 在样本分类中的效果；通过生物学手段验证疾病发生机制或药物作用机制。

三、 应用案例

案例 1 华大参与——类风湿性关节炎患者治疗前后肠道微生物研究[1]

通过宏基因组研究类风湿性关节炎（rheumatoid arthritis, RA）治疗前后对人肠道及口腔微生物的影响，并基于菌群研究结果构建分类器，有助于疾病的诊断和治疗。

材料与背景

样本来源：未经药物治疗的 RA 患者；健康人群（包括直系亲属和共同生活无血缘关系亲属）；经治疗的 RA 患者。

采样部位及样品数：粪便（212 例：77 例 RA 患者，80 健康人，17 对直系亲属（RA 患者&健康人)，21 例经类风湿性关节炎药物（disease-modifying antirheumatic drugs，DMARD）治疗 RA 患者；牙菌斑（105 例，54 例 RA 患者，51 健康人）；唾液（98 例，51 例 RA 患者，47 健康人）。

方法与策略

1. 对肠道、口腔样品构建基因集、物种分类等基础分析；
2. Case & control 比较分析健康人群与 RA 患者在牙菌斑、唾液、肠道三个部位的差异；
3. 与 RA 的临床医学指标的关联分析；
4. 口腔、肠道分析结果一致性评估；
5. 基于口腔与肠道微生物菌群的关联分析构建区分 RA 患者和健康对照人群的分类诊断模型；
6. DMARD 治疗结果对构建的分类模型进行验证。

结果与分析

7. 口腔与肠道菌群在组成及功能上具有一致性（overlap）；
8. RA 病人的口腔、肠道菌群失调，在经过治疗后，得到部分恢复；
9. 基于菌群构建分类器，该分类器与 RA 的医学检测指标相关；

10. 某种嗜血杆菌（Haemophilus sp.）在 RA 患者中呈现相对缺失的状态，并且其丰度与 RA 自身免疫抗体的滴度成反比。而唾液乳杆菌（Lactobacillus salivarius）在 RA 患者的牙菌斑、唾液和粪便中均显著富集，尤其是在病情高度活动患者中表现得尤为明显。

案例 2 华大参与——Ⅱ型糖尿病、二甲双胍和肠道菌群之间剪不断理还乱的关系[2]

通过宏基因组检测Ⅱ型糖尿病（T2D）及药物治疗对人肠道菌群结构的影响，揭示了肠道微生物与 T2D 以及相关药物治疗的关系，更进一步揭示肠道微生物影响二甲双胍作用效果的分子机制；为更深入研究复杂疾病发生或药物治疗提供依据。

材料与背景

784 份人肠道宏基因组样本。

MetaHIT（Metagenomics of the Human Intestinal Tract）项目：

丹麦人群：277 例非糖尿病个体，75 例 T2D 患者，31 例 T1D 患者 （MHD）；

瑞典人群：53 例女性 T2D 患者；92 例非糖尿病个体（43 例正常葡萄糖耐受水平，49 例葡萄糖耐受损伤）（SWE）；

中国人群：71 例 T2D 患者，185 例非糖尿病个体（CHN）。

方法与策略

研究 T2D 及药物治疗对人肠道菌群的影响；

分析肠道菌群对常用 T2D 药物二甲双胍作用的影响，以及可能的作用机制。

主要结果

1. T2D 患者的肠道菌群发生改变：

（1）菌群改变：

产丁酸盐相关菌群（*Roseburia* spp., *Subdoligranulum* spp.和产丁酸盐梭菌目）有减少；

厚菌壁门丰度也有变化，一般是降低，二甲双胍处理后丰度有回升。

（2）功能变化：

氧化氢酶和核糖，甘氨酸和色氨酸降解功能富集；而苏氨酸和精氨酸降解功能降低；这些功能改变与菌群变化有对应关系。

2. 二甲双胍和肠道菌群互相影响：

（1）T2D 患者和二甲双胍治疗的 T2D 患者间菌群差异比 T2D 和正常人群差异小；只在 family 水平有差异；

（2）二甲双胍作用下 *Escherichia* spp.增加（CHN 群体无该发现），*Intestinibacter* spp.降低（普遍存在

于所有群体）；

（3）T2D 病人肠道菌群中产丁酸菌含量下降；导致功能微生物的改变，二甲双胍可以部分缓解这一情况；

（4）微生物通过丁酸或丙酸等短链脂肪酸合成刺激肠道糖异生（IGN），进而影响二甲双胍的治疗效果；

（5）埃希氏杆菌菌属的丰度相对增加与二甲双胍治疗的副作用【如毒力因子、气体代谢基因（gas metabolism genes）功能富集】有关。

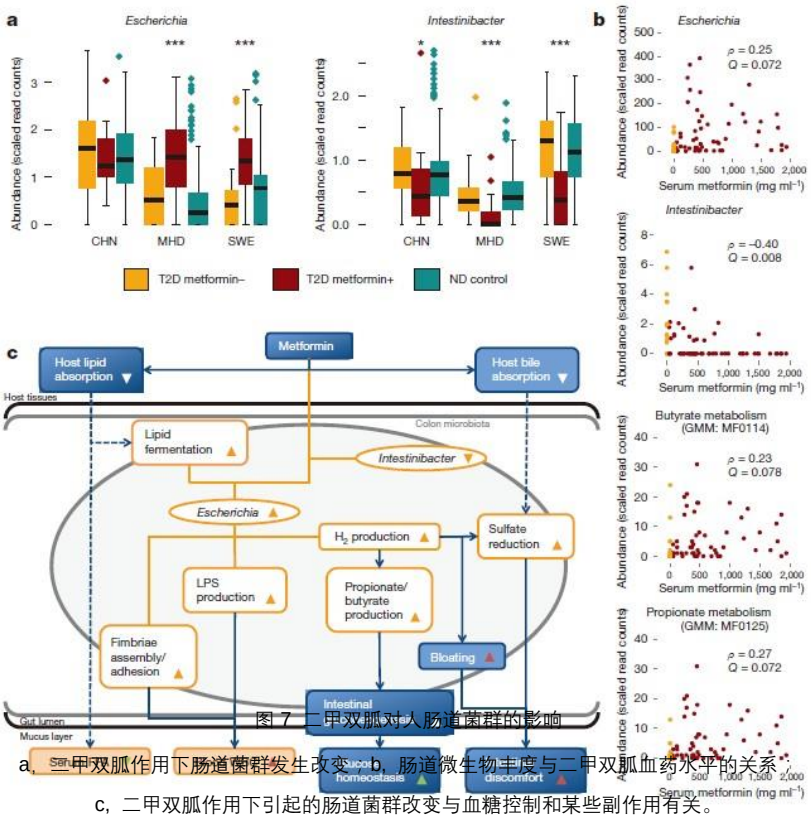


图 7 二甲双胍对人肠道菌群的影响

a，二甲双胍作用下肠道菌群发生改变；b，肠道微生物丰度与二甲双胍血药水平的关系；

c，二甲双胍作用下引起的肠道菌群改变与血糖控制和某些副作用有关。

四、 项目风险

样本量不足可能对统计结果产生较大的影响，影响结果的准确性。

五、 华大优势

- **丰富的项目经验。**承担MetaHIT等多个大型国际合作项目，已进行土壤、水体、动物及人体胃肠道、

极端环境、发酵罐等多种环境宏基因组调查，技术人员有丰富的宏基因组研究经验，可以为合作伙伴的后续工作提供极大的助力。

- **个性化的分析方案。**对于有特殊信息分析要求的客户，能根据客户需求制定个性化信息分析方案，信息分析团队会安排专人跟进，确保客户需求得到满足。
- **结题报告文章化。**深入分析相关领域研究思路，注重生物学意义的挖掘，提供完全文章化的项目结题报告，极大简化客户撰写文章的过程。

六、 参考文献

[1] Zhang X, Zhang D, Jia H, *et al*. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment[J]. Nature medicine, 2015.

[2] Forslund K, Hildebrand F, Nielsen T, *et al*. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota[J]. Nature, 2015, 528(7581): 262-266.

单基因遗传病的外显子测序研究方案

一、 研究背景

单基因遗传病,又称为孟德尔遗传病,是由 DNA 水平的单个碱基突变或碱基序列插入缺失引起的,这种单碱基突变、碱基序列的插入缺失往往带来非常严重的后果。至今,人类已知的单基因遗传病已有 5000 多种。

全外显子组测序通过对基因组1%序列测序，可以得到绝大多数编码区信息，可以通过一次检测同时分析同一疾病的多个基因或多种疾病的相关致病基因,可用于致病机制比较复杂的遗传病的检测以及单基因遗传病的鉴别诊断,能够有效地对临床使用常规方法难以诊断的遗传疾病提供分子水平的诊断依据。

华大外显子在孟德尔遗传病研究案例

- 研究脊髓小脑性共济失调，发现致病基因-*TGM6*
- 研究逆反性痤疮，发现致病基因-*NCSTN*
- 研究高度近视，发现致病基因—*ZNF644*
- 研究Olmsted 综合征，发现潜在靶标-*TRPV3*
- ……

二、 方案设计

图 1 已知致病基因排查

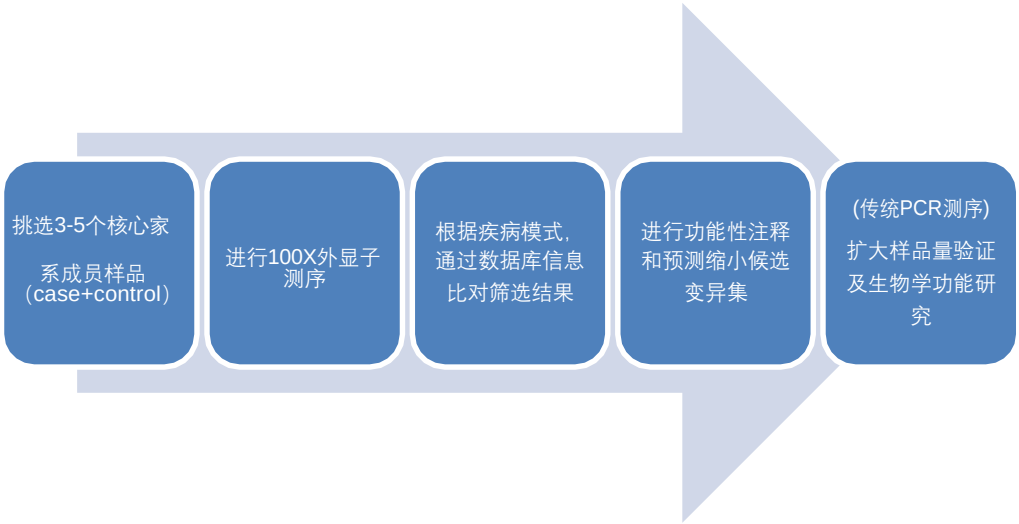


图 2 单基因病研究策略

2.1 样本建议

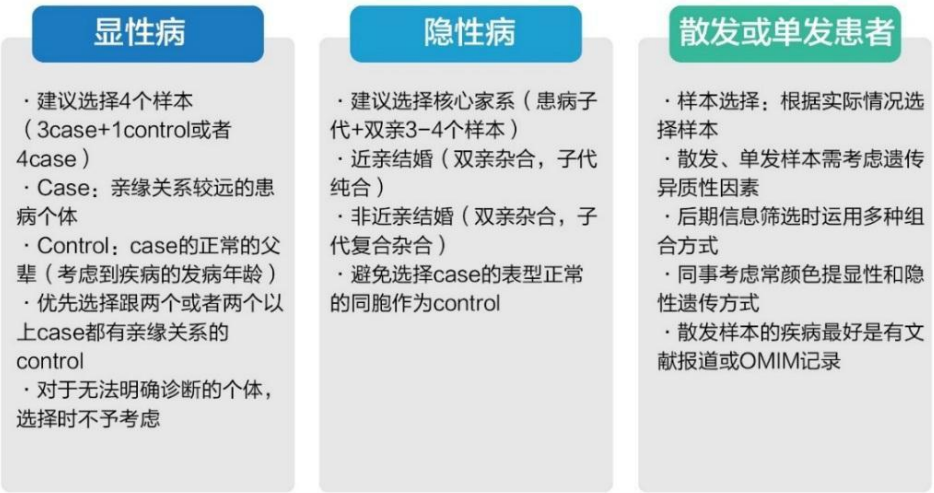


图 3 样本选择建议

样品选择总结：

? 显性病，case 越远越好，control 较老为好

? 隐性病，患病子代+正常双亲

? 散发，多个样本，多种组合分析

? 有研究基础，可减少测序样本数

? 特殊遗传方式，需慎重

2.2 测序参数

二代测序平台全外显子测序，PE101/PE151，测序深度建议100X以上。

2.3 实验技术

捕获平台主要指全外显子捕获平台，通常有Agilent SureSelect_Human_All_Exon、BGI Exome、NimbleGen SeqCapEZ_Exome三个平台作为选择。每个捕获平台都有多款捕获芯片，不同的捕获芯片具有不同的捕获区域和性能,研究人员可以根据具体的项目情况进行选择。

2.4 分析内容

(1) 标准信息分析

标准信息分析部分由比对处理，变异检测和注释三个主要分析内容组成。信息分析的结果将作为筛选解读的对象。

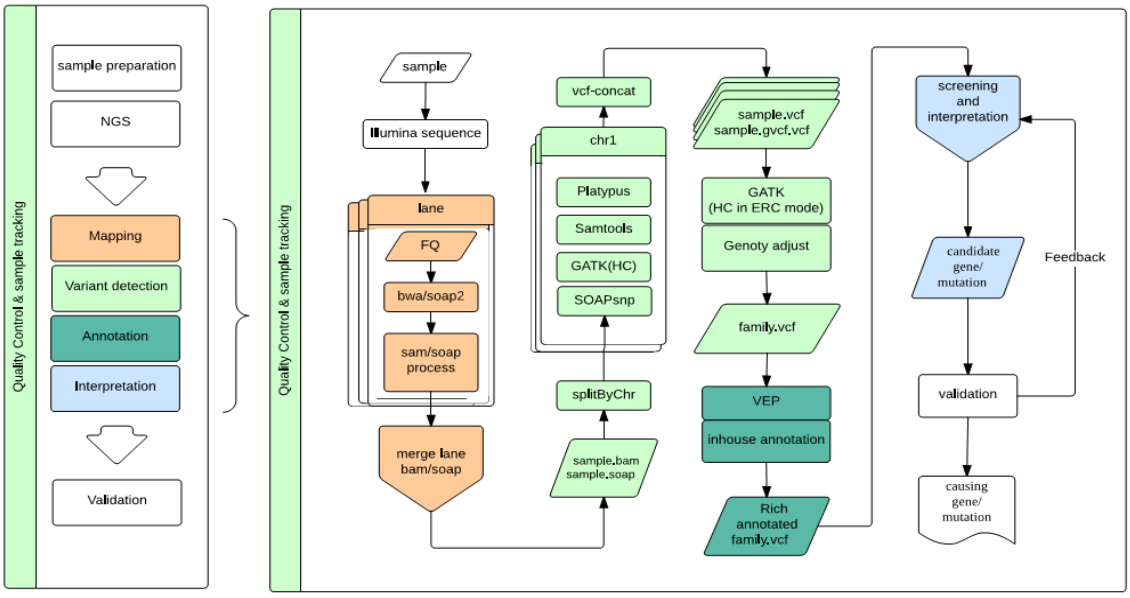


图 4 标准信息分析流程图

(2) 筛选解读

基于医学遗传学基本理论和测序数据的特点，对变异信息进行一系列过滤，最终得到可能与所研究疾病相关的变异集。对这部分变异集进行证据搜集，患者临床表征，按照一定规则对变异的致病性进行判定，评估其与所研究家系的疾病或临床表型之间的关系。筛选解读的结果将作为候选，通过进一步的验证进行确认。筛选解读时主要考虑方面如下：

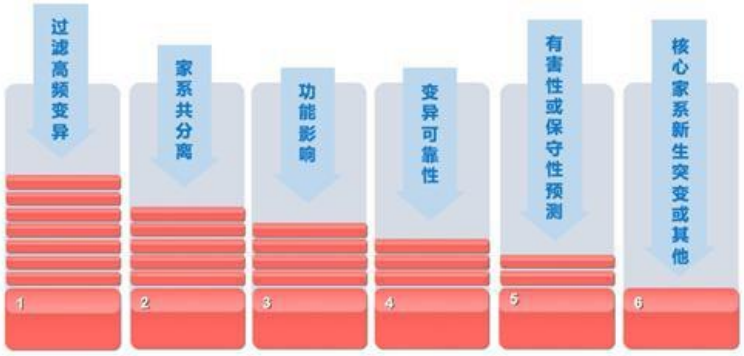


图5 筛选解读

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：约40个工作日。实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

利用人全外显子测序平台对单基因遗传病家系进行测序研究，前期准备时提供疾病调研，已知致病基因排查，测序样本选择等多方位研究策略，使得研究目的明确。信息分析部分采用多种数据库，有害性预测工具，最新版本KEGG通路注释等，筛选出单基因病家系的致病变异。可为您提供常显，常隐，X连锁和*de novo*突变分析。

2.7 实验验证

筛选解读的结果将作为候选，通过进一步的验证进行确认。验证需要对变异所在的基因和变异本身，从遗传学、生物信息学等角度加以考察，并结合相关实验，最终得出变异与功能(致病性)是否相关的结论。

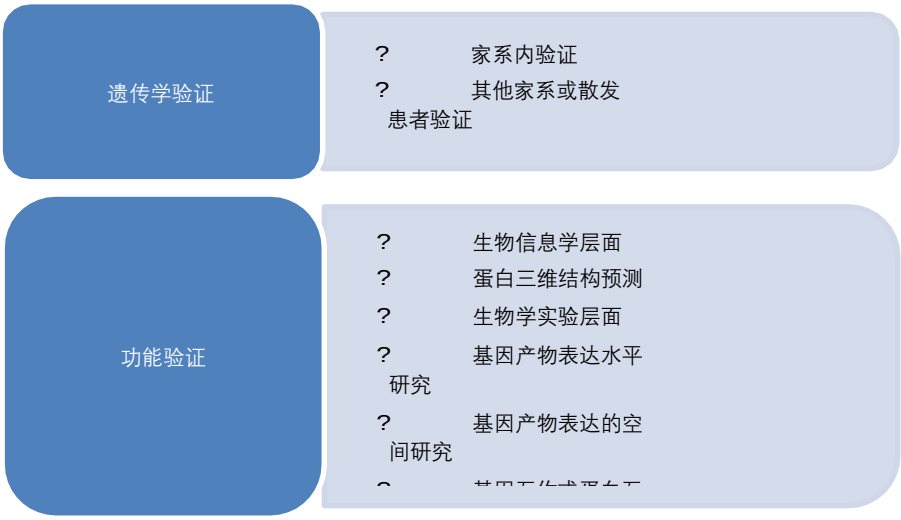


图6 后期验证方法

三、 应用案例

案例 1 华大参与——外显子研究小脑共济失调 [1]

这是国内科学家首次用全外显子组测序技术对单基因病进行研究。文章共选择了一个家系中的 4 个患者（III:6, III:7, III:17, IV:1）进行外显子测序，测序深度为 65X，覆盖度为 99.6%。研究人员发现一个新的小脑共济失调致病基因 *TGM6*，并且采用 Sanger 测序检测该家系里其他患者和正常人，观察该基因突变和表型的共分离进行验证。此外，用 Sanger 测序的方法在 84 个其他家系的先证者中检测这个基因，其中一个家系中发现这个基因有突变，而且是一个不同的突变。最后，用Sanger 测序方法，在 500 个正常对照中验证，并没有发现这两个突变。

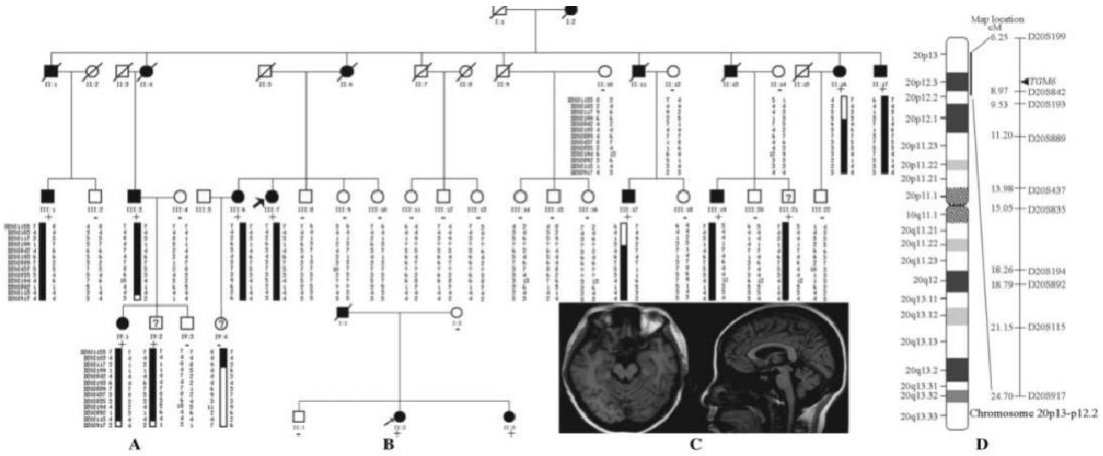


图 7 小脑共济失调家系图，先证者大脑核磁共振图以及 20 号染色体的比对区域

案例 2 外显子测序解析卵巢早衰的遗传因素 [2]

卵巢早衰通常指女性 40 岁之前闭经，1%的妇女患有此病，病因复杂，被认为与遗传因素有关。本研究利用外显子测序技术首次对中东家系（MO1DA）的卵巢早衰病人样本进行分析，DNA 由这个家族中的两个姊妹提供，姊妹俩其中一个是健康的，另一个是不育的。研究发现减数分裂基因中的 *STAG3* 基因突变可以导致隐性遗传卵巢早衰，也在小鼠动物模型和卵巢早衰病患中得到了证实。为探索卵巢早衰或卵巢功能不全的发生机理，以及阐明该病的临床高度异质性和遗传病因复杂性开辟了一个新的研究思路。

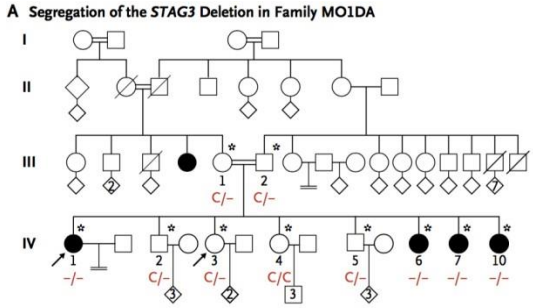


图 8　MO1DA 家系图谱

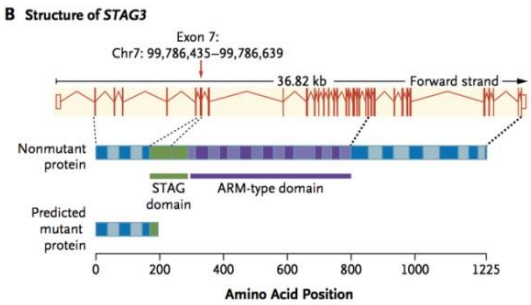


图 9 *STAG3* 基因结构图

案例 3　华大参与——外显子测序在视网膜色素变性的应用 [3]

视网膜色素变性（RP）是一种慢性、进行性、遗传性，营养不良性视网膜退行性疾病。研究人员对 RP 四代家系中，选取 4 个患者（II-2, II-3, II-4, II-7）做为 case，选取一个健康人（II-9）做为 control。通过全外显子测序，深度 100X，找到非同义突变（OR2W3 R142W）为 RP 家系新的致病基因。最后，用了三代独立家系的 3 个患者（II-1，II-2，III-1）和 1 个正常对照（I-1）进行验证，同样存在家系共分离现象。另外，研究人员通过 RT－PCR 发现 *OR2W3* 基因在 HESC－RPE 中（视网膜色素上皮细胞 RPE 诱导胚胎干细胞）中表达。

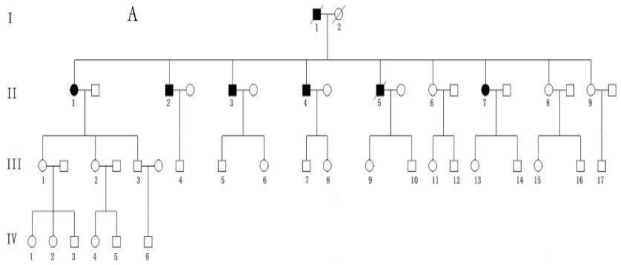


图 10 RP 四代家系

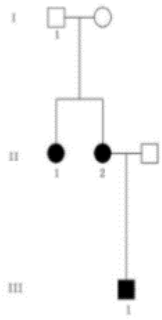


图 11 用来验证的独立 RP 三代家系

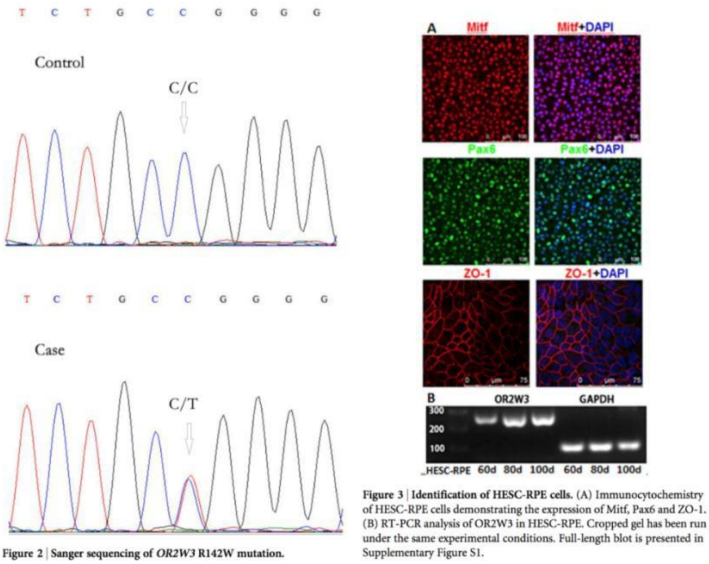


图 12 Sanger 验证图，RT－PCR 发现 *OR2W3* 基因在HESC－RPE 中表达

四、 项目风险

当前应用二代测序技术研究单基因病的成功率在50%左右，影响成功率的原因主要有两方面，一是疾病的复杂性，二是技术的局限性。首先，推荐的方法所基于的假设是家系遗传的疾病由DNA变异导致，且完全外显。但疾病通常受到遗传背景或环境的影响，存在不完全显性遗传，共显性遗传，遗传异质性，延迟显性，X染色体失活，生殖腺嵌合，拟表型等情况，可能会找不到致病基因。其次，在技术上全外显子捕获芯片对基因的覆盖度通常不都是100%，会漏掉覆盖不到的区域；且我们暂时只关注SNP/InDel，不能处理染色体异常、CNV、SV、STR等变异，也会导致某些情况下找不到致病基因。此外，非编码区变异解读的困难，验证过程中取样的困难等问题，同样会影响致病基因的发现。总之，疾病的复杂性和技术的局限性真实存在，准备工作越充分，研究策略越合理，则成功率越高。

五、 常见问题

问题 什么是家系共分离?

答：在单基因遗传病家系中，基因组序列的改变在后代中与遗传性状共同出现。即患者都有致病的基因型，正常对照都没有。在假设疾病为完全外显的情况下，通常会观察到家系共分离的现象。

六、 华大优势

- **超高的性价比。**可提供重复的疾病调研，已知基因排查，测序样本选择等多方位研究策略。通过筛选解读，极大缩小疾病相关的候选集，进一步降低后续验证的位点数。
- **高质量的信息分析结果。**由经验丰富的信息分析团队，利用华大自主开发的最新版单基因病流程进行分析。多软件变异检测获取最准确变异结果，并通过 12 种有害性或保守性预测工具进行打分，6 种公共群体等位基因频率数据库及华大基因内部群体等位基因频率数据库注释，此外还有 OMIM 疾病数据库，正常组织表达蛋白数据库，最新版 v76.0 KEGG 通路注释等数据库注释。对于 *de novo* 突变，采用过滤与基于机器学习（forestDNM）联合分析方法检测。
- **高效的筛选解读。**自主开发的筛选标签，多种遗传模式和分析内容在同一候选列表中进行筛选。
- **丰富的经验。**2010 年，华大发表了全国第一篇应用外显子测序技术研究单基因病的文章，研究论文共发表超过 600 篇，其中发表于顶级科学杂志 Nature 与 Science 100 余篇，发表单基因遗传病文章 90 余篇。

七、 参考文献

[1] Wang J L, Yang X, Xia K, *et al*. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing[J]. Brain, 2010, 133(12): 3510-3518.

[2] Caburet S, Arboleda V A, Llano E, *et al*. Mutant cohesin in premature ovarian failure[J]. New England Journal of Medicine, 2014, 370(10): 943-949.

[3] Ma X, Guan L, Wu W, *et al*. Whole-exome sequencing identifies OR2W3 mutation as a cause of autosomal dominant retinitis pigmentosa[J]. Scientific reports, 2015, 5: 9236.

肿瘤发生的甲基化调控机制研究方案

一、 研究背景

以往研究人员通常认为原癌基因的激活和抑癌基因的失活主要是因为基因点突变、插入缺失突变以及结构变异导致 DNA 序列发生改变，从而导致了癌症的发生。随着研究的深入，研究人员们发现这些基因表达的异常通常是由 DNA 甲基化调控实现的，原癌基因的低甲基化和抑癌基因的高甲基化改变是肿瘤细胞的一个重要特征。DNA 甲基化调控在肿瘤的发生与发展以及诊断治疗过程中起到了十分重要的作用。

二、 方案设计

2.1 整体思路

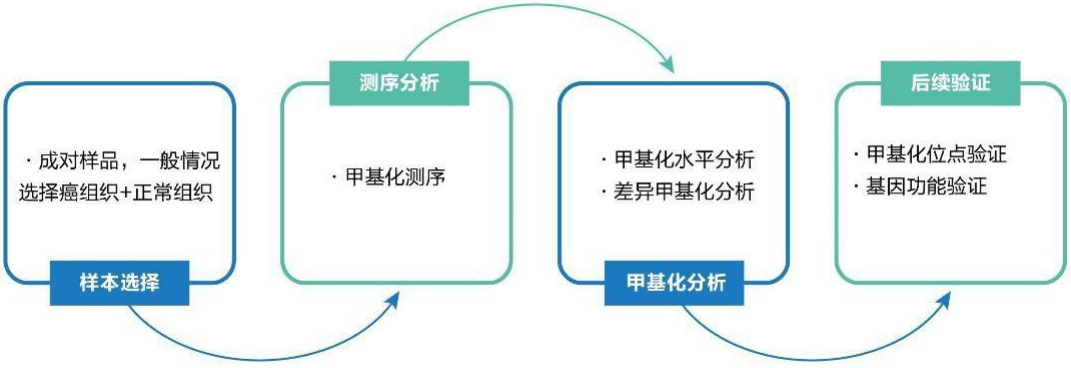


图 1 整体研究思路

2.2 样本建议

（1）对于实体瘤，一般可以选择成对的癌/癌旁组织样品，建议样本数不少于 5 对。

（2）同卵双胞胎样本，一个患病、一个正常，取相同部位的癌组织与正常组织，或外周血样品，建议样本数不少于 5 对。

2.3 实验技术

采用甲基化测序技术，例如全基因组甲基化测序（WGBS）或目标区域捕获甲基化测序（TBS）技术，鉴定 C 位点的甲基化水平，并通过成对样品间的比较以及筛选，寻找差异甲基化区域，并对差异甲基化区域所在基因的功能进行分析。

2.4 测序参数

测序深度建议：全基因组甲基化测序（WGBS）不低于 90G Clean Data，目标区域甲基化测序（TBS）不低于 100X。

2.5 分析结果

(1) 甲基化位点鉴定

C 碱基的甲基化水平等于该 C 碱基上覆盖到的支持 mC 的序列数除以有效覆盖的序列总数，(详细算法请参考方法部分)。通常 CG 甲基化存在于基因和重复序列中，在基因表达调控过程中起到非常重要的作用。非 CG 类型的序列（CHG 和 CHH）在基因中十分少见，主要存在于基因间区和富含重复序列的区域，在沉默转座子过程中起关键作用。

mCG，mCHG 和 mCHH 三种碱基类型的构成比例在不同物种中，甚至在同一物种不同样品中都存在很大差异。因此，不同时间、空间、生理条件下的样品会表现出不同的甲基化图谱，各类型 mC（mCG、mCHG 和 mCHH）的数目，及其在全部 mC 的位点中所占的比例，在一定程度上反映了特定物种的全基因组甲基化图谱的特征。

图 2 各类型 mC 统计图

(2) 差异甲基化分析

差异甲基化区域（DMRs）是指不同样品中基因组表现出不同的甲基化模式的某些 DNA 片段。 DMR 与遗传印记相关，在个体中表现为与父本或母本的甲基化状态一致。甲基化的等位基因经常表现为沉默状态。亲本与子代甲基化模式的差异常常导致表观遗传缺陷，而人工繁殖技术可能会导致异常甲基化的比例升高，并导致疾病

的发生。

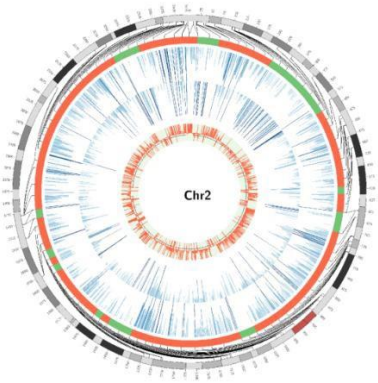


图 3 差异甲基化区域展示

(3) GO 与 Pathway 分析

针对差异甲基化区域包含的基因进行基因功能注释和代谢通量的分析。基因本体论（Gene ontology，GO）是所有物种中最主要的了解基因和基因产物属性的生物信息学分析手段，GO 分析能够用于鉴定基因产物的性能，它包含了三类基因功能信息：细胞组分（Cellular Component），分子功能（Molecular Function）和生物学过程 （Biological Process）

KEGG (Kyoto Encyclopedia of Genes and Genomes)是有关 Pathway 的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是测序得到的基因集与细胞、生物体以及生态环境的系统性功能相关联。所有样品中的DMR 相关基因均用 KEGG 数据库进行分析。

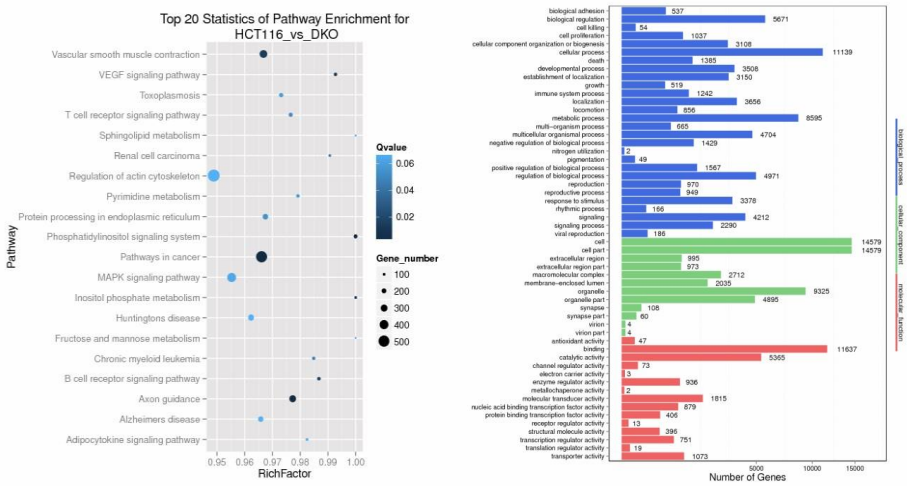


图 4 Go 与 Pathway 分析结果

2.6 项目周期

样品检测合格后，建库+测序+标准信息分析：约 60 个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.7 预期结果

利用表观遗传学手段，借助高通量测序平台，通过对不同处理的组织之间的比较，从基因组范围内的 DNA 甲基化水平对基因表达的影响的分子机理进行深入挖掘，并了解其生化通路以及调控网络，为药物反应以及研发提供基础平台，为疾病的治疗提供理论依据。

我们期望通过 case 与 control 的组合分析，找到有显著差异的基因区段，并通过对这些基因区段的验证从功能方面对其进行深入研究，从而对我们的疾病诊断、治疗提供有利帮助。

2.8 辅助研究策略

可以通过 RNA 测序研究全基因组表达量，并比较不同样品表达量的差异，通过 DNA 甲基化与 RNA 表达数据的联合分析研究甲基化与基因表达量的关系，从而辅助发掘与肿瘤发生、发展以及治疗相关的功能基因。

2.9 实验验证

分析得到的差异基因/差异甲基化区域可以利用 Q-PCR 检测表达量变化，从而将表观修饰与基因表达量进行关联分析，从而找到直接与疾病相关的基因或区域。

此外还可以利用动物模型对候选基因功能进行深入研究，例如转基因小鼠等。

三、 应用案例

案例 华大参与——捕获甲基化测序揭示肝细胞癌的发生与机理[2]

文章采用华大设计的 Promoter 捕获 kit，对 8 个肝细胞癌组织及其癌旁组织进行了目标区域捕获甲基化测序，并找到 2972 个 DMR 区域。后续通过联合 RNA-Seq 结果进行分析，鉴定出 12 个 DMR 相关基因，并用 BSP 方法验证了这 12 个基因的甲基化状态，其中 7 个基因在大样本量的群体（78 对 HCC 样品）中用 BSP 手段进行了验证。本文联合了 RNA-Seq 和 Promoter 甲基化测序的方法来发掘 HCC 的原癌基因和抑癌基因，并提供了一种性价比极高的表观遗传学研究方案。

研究策略：华大自主专利的 Promoter 捕获试剂盒＋RNA-seq，综合两种技术联合分析，缩小筛选范围获取癌症相关基因。

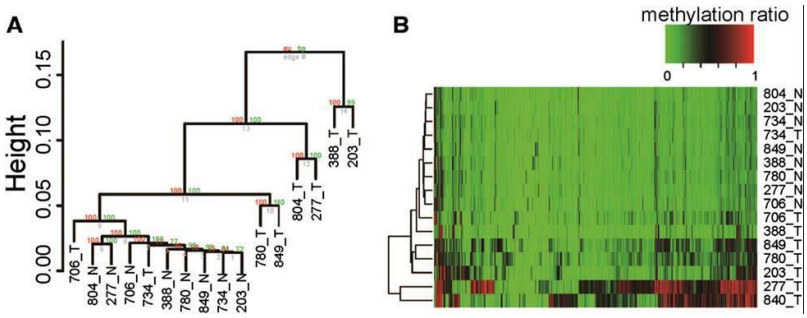


图 5 甲基化水平聚类分析

四、 项目风险

在项目实施过程中，可能存在由于样本群体样本过小、数据覆盖度不够或者由于肿瘤异质性高等因素的影响导致找不到甲基化水平呈现显著差异的区域或基因，在这种情况下，一般可以考虑通过增加样本数或增加测序数据等手段进行补充。

五、 常见问题

问题 1 WGBS 与 TBS 有什么区别？

答：WGBS 是针对全基因组范围内的甲基化测序，覆盖范围最大，几乎能够检测到全基因组的所有 C 位点的甲基化状态，而 TBS 是针对全基因组内某些特定范围进行甲基化检测，区域更加集中，在节省成本的同时减少资源消耗，检测甲基化水平也更加精确。

问题 2 WGBS 能够覆盖全基因组多少区域？

答：理论上 WGBS 可以检测到全基因组所有的 C 位点的甲基化状态，但是由于测序数据量的限制，可能会有少量的 C 位点无法覆盖，因此一般情况下 WGBS 可以覆盖全基因组范围内 99% 的 C 位点。

六、 华大优势

- 技术全面。可以提供 DNA 甲基化、羟甲基化、组蛋白修饰、转录因子结合位点研究等全面的科研技术服务。
- 经验丰富。华大已完成甲基化研究项目涉及超百种物种，广泛覆盖了动物、植物以及微生物等各种领域；上万份 ChIP 样本的测序研究；数十篇高影响因子文章发表。
- 优化的分析流程。绝对标准和成熟的分析流程，能够最大限度地缩短项目周期，得到最准确的分析结果。

- 强大的分析能力。华大基因拥有世界一流的生物信息学研究中心和世界最具竞争力的生物信息学团队，能够为不同的客户提供个性化的分析方案。

七、 参考文献

[1] Li Y, Zhu J, Tian G, *et al*. The DNA methylome of human peripheral blood mononuclear cells[J]. PLoS Biol, 2010, 8(11): e1000533.

[2] Gao F, Liang H, Lu H, *et al*. Global analysis of DNA methylation in hepatocellular carcinoma by a liquid hybridization capture-based bisulfite sequencing approach[J]. Clinical epigenetics, 2015, 7(1): 1.

[3] Wang J, Jiang H, Ji G, *et al*. High resolution profiling of human exon methylation by liquid hybridization capture-based bisulfite sequencing[J]. BMC genomics, 2011, 12(1): 1.

[4] Gao F, Wang J, Ji G, *et al*. Clustering of cancer cell lines using a promoter-targeted liquid hybridization capture-based bisulfite sequencing approach[J]. Technology in cancer research & treatment, 2015, 14(4): 383-394.

肿瘤诊断的蛋白标志物研究方案

一、 研究背景

肿瘤的早期诊断和确诊对于患者至关重要，直接影响到癌症患者的治疗和康复效果。现阶段蛋白质组学技术的飞速发展，为寻找肿瘤诊断标志物提供了有利条件。通过蛋白质组学技术，科研工作者可以实现肿瘤生物标志物的高通量筛选，做到对肿瘤的早期诊断和确诊，为医务工作者开展切实可行的肿瘤预防和治疗工作提供重要的指导。

二、 方案设计

2.1 整体思路

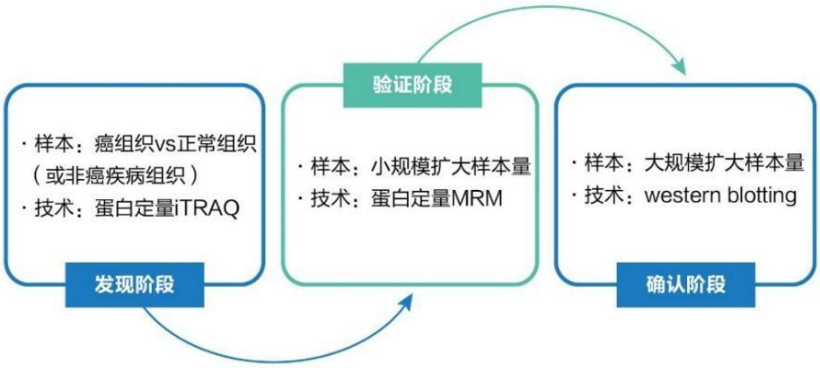


图 1 整体研究思路

2.2 样本建议及实验技术

(4) 发现阶段

本阶段旨在初步筛选与癌症相关的差异表达蛋白，得到候选的蛋白标志物。

样本选择：癌症组织 vs 正常组织（或非癌疾病组织）各 200mg，可使用混合样本。癌症组织可根据实验目的选取癌症发生不同时段的组织样本，与正常组织或非癌疾病组织进行比较分析。

技术介绍：蛋白定量 iTRAQ 技术，是一种体外同重同位素标记的蛋白相对定量技术。该技术利用多种同位素试剂标记蛋白多肽 N 末端或赖氨酸侧链基团，经高精度质谱仪串联分析，可同时比较 8 种样品之间的蛋白表

达量，是近年来定量蛋白质组学常用的高通量筛选技术。

(5) 验证阶段

本阶段旨在缩小发现阶段筛选出的候选蛋白标志物范围，得到经过验证的蛋白标志物。

样本选择：推荐使用在发现阶段选取的样本进行初步验证，要求进行单个样本验证而非混合样本。也可适当小规模扩大样本量进行验证，但样本类型建议不变（如发现阶段采集组织样本，验证阶段也同样采用组织样本）。

技术介绍：蛋白定量 MRM 技术，是指利用基于质谱的多反应监测技术（MRM）有目标地分析检测可能与特殊功能相关的关键蛋白在不同样本中的表达量，进而推测这些蛋白的生物学功能，每个样本可一次性检测上百种目标蛋白质的表达量，相比传统的 western blotting 复杂的制备抗体等试验环节，采用蛋白定量 MRM 技术可大大缩短项目周期，节省成本。

(6) 确证阶段

本阶段目的是将经过验证的蛋白标志物（一般情况下≤3 个），在大范围的样本中进行广泛的确证，最终得到可信的并且易于获取的癌症诊断蛋白标志物。

样本选择：推荐大规模扩大样本量进行深入验证，可考虑改变样本类型，采用易于获取的样本类型进行（如发现和验证阶段都采用癌症组织样本，确证环节可选用血液样本）；也可考虑从动物模型样本转移到临床样本进行验证。

技术介绍：Western Blotting 技术是通过特异性抗体对凝胶电泳处理过的细胞或生物组织样品进行着色，分析着色位置和着色深度获得特定蛋白质在所分析的细胞或组织中表达情况的信息。

2.3 分析结果

(1) 蛋白定量 iTRAQ 技术结果——样本间差异表达蛋白信息

统计样本间的上调蛋白、下调蛋白总数，作为候选蛋白标志物，并提供详细蛋白列表。

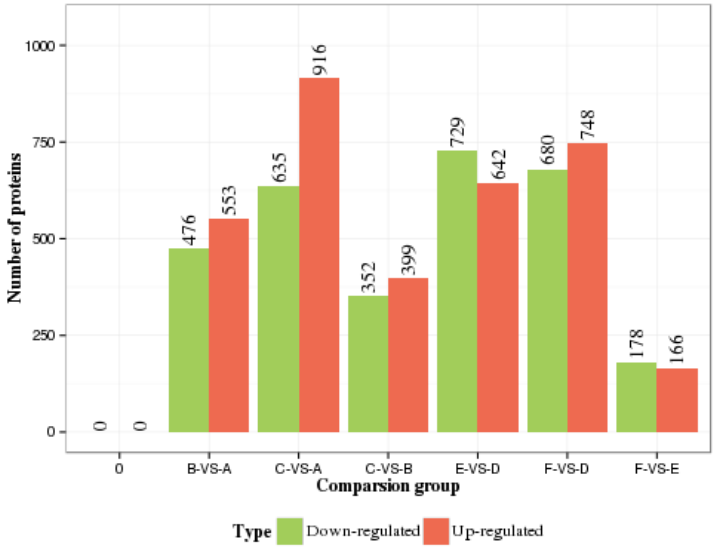


图 2 差异蛋白柱形图

X 轴: 比较组信息; Y 轴:差异蛋白数量。红色柱为显著上调蛋白; 绿色柱为显著下调蛋白。

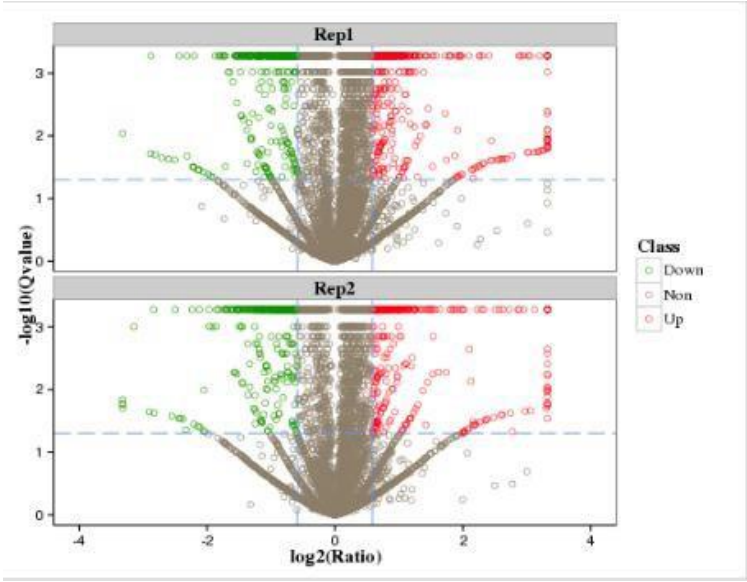


图 3 显著差异蛋白火山图

该图X 轴为蛋白差异倍数（取 log2），Y 轴为相应的-log10（Qvalue）。Qvalue<0.05 和 Foldchange>=1.5 是显著差异蛋白的筛选条件。图中红色点为显著上调蛋白，绿色点为显著下调蛋白，灰色点为无显著变化蛋白。

(2) 蛋白定量 MRM 技术结果——蛋白标志物的验证

对单个目标蛋白进行验证，得到各组样本间蛋白标志物表达量的相对比值，与 iTRAQ 结果进行比较，进一步缩小蛋白标志物的范围。

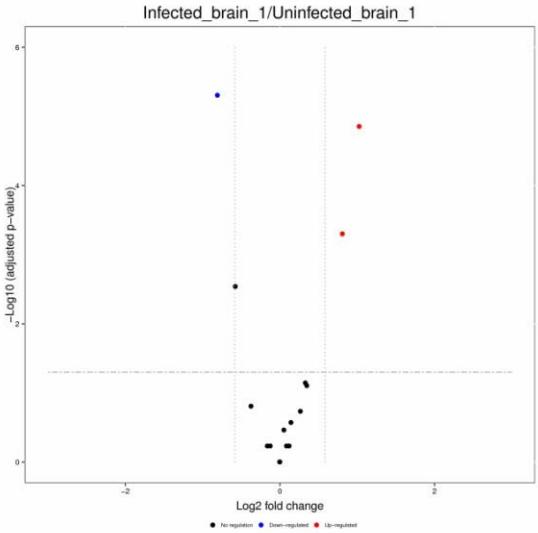


图 4 目标蛋白的相对定量分布

该火山图展示了目标蛋白在比较组的比值和校正 p 值，横坐标为比值取 log2，纵坐标为校正p 值取-log10，目标蛋白在 1.5 倍差异以上且校正 p 值<0.05 的条件下认为是差异蛋白。

2.4 项目周期

样品提取检测合格后，蛋白定量 iTRAQ 技术标准周期约 40 个工作日；蛋白定量 MRM 技术标准周期约 60 个工作日。实际项目完成时间由所选样本数以及信息分析条款决定。

2.5 预期结果

期望通过蛋白质组学技术手段，借助高通量质谱平台，通过对癌症组织和正常组织（非癌疾病组织）之间的比较，寻找特异性好、分辨度高、易于取样的癌症诊断蛋白生物标志物。

2.6 辅助研究策略

可以通过免疫组化技术验证癌症组织中特异的蛋白标志物，进一步确认筛选出的蛋白生物标志物具备特异性好、分辨度高的特点，辅助筛选癌症蛋白标志物。

三、 应用案例

案例 1 华大参与——寻找结直肠癌发生发展过程中的蛋白生物标志物 [1]

文章在发现阶段采用三个时期小鼠模型的癌症组织和对照组织进行蛋白定量 iTRAQ 分析，找到了 144 个差异蛋白，选取其中 62 个持续变化的蛋白进行 MRM 验证，得到 18 个趋势一致的蛋白标志物，再挑选其中的 12 个持续上调的蛋白标志物在小鼠模型的血清样本中进行 MRM 验证，得到 3 个显著上调表达的蛋白标志物，最后将这三个蛋白标志物在 16 组临床血清样本中进行验证，其中两个蛋白标志物（LRG1,TUBB5）得到了验证，可以作为候选结直肠癌的蛋白生物标志物。

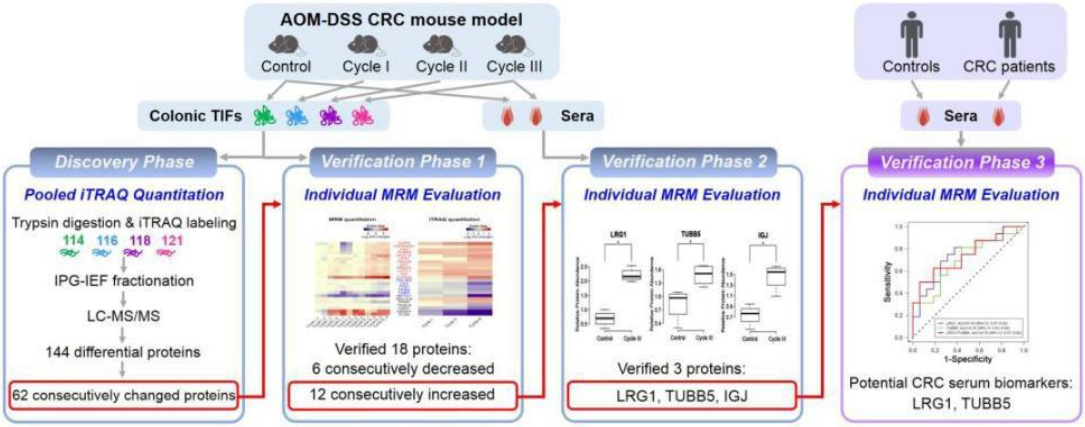


图 5 实验整体思路

案例 2 寻找胰腺癌早期诊断的蛋白生物标志物[2]

该研究采用 iTRAQ 技术对乙酰胆管腺癌（PDAC）病人的血清样本进行研究，并采用 MRM 技术、Western blotting 对筛选的蛋白进行验证，Thrombospondin-1 （TSP-1）蛋白在 PDAC 癌病前 24 个月就有明显的下降；PDAC 病人与良性胆道阻塞病人以及健康人样本相比，TSP-1 表达也是下调的。TSP-1 的低表达与 PDAC 病人存活期短有关。在被确诊为伴随糖尿病的 PDAC 病人样本中 TSP-1 下调经常被观察到，可能 PDAC 病人血清 TSP-1 含量低与糖尿病相关。TSP-1 可以和 CA19-9 一起作为早期诊断 PDAC 的标志物。

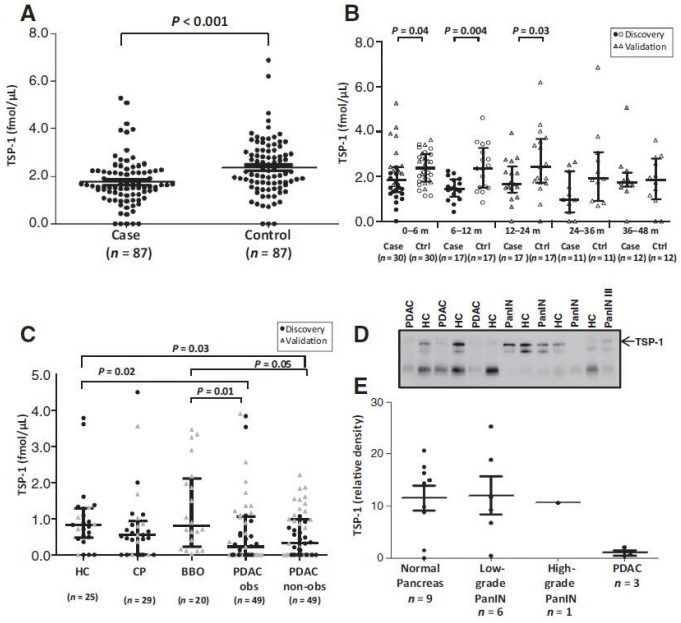


图 6 MRM 方法对 346 个样本中的 TSP-1 蛋白进行验证

四、 项目风险

在项目实施过程中，由于某些蛋白的丰度低，蛋白定量 iTRAQ 和 MRM 技术方法存在一定的差异，用 iTRAQ 初步筛选出的差异蛋白，用 MRM 方法可能存在验证不到的情况，此时对于某些有重要作用的关键蛋白，建议改用 western blotting 的方法进行验证，该方法的灵敏度更高，对于低丰度的蛋白具有更好的检测效果。

五、 常见问题

问题 1 蛋白质组学的主要研究内容有哪些？基于质谱的蛋白质组学的研究主要应用在哪些方面？

答：蛋白质组学是研究生物体中所有蛋白质全局变化规律的学科。谈到全局变化，就涉及到所有蛋白质的表达与否、表达量的相关变化规律、蛋白质的修饰规律、结构变化以及相互作用等很多方面。可以说，近几年质谱技术的飞速发展是蛋白质组研究的福音。现在的质谱技术不仅可以大规模鉴定蛋白质的表达情况，还可以对不同状态、处理的生物样品的蛋白质组进行全局的定量比较。不仅如此，结合一些富集、交联的实验技术，质谱还可用以研究大规模蛋白质组的修饰变化规律、相互作用的规律。

问题 2 只通过蛋白定量 iTRAQ 技术找到的差异蛋白，可否直接认定为 biomarkers？

答：从生物实验的角度来讲，通过一种实验得到的结果是需要再进行验证的。在发现阶段，用蛋白定量 iTRAQ 技术找到的差异蛋白，一般还需要通过其它技术进行验证后，再得出最终的结论，推荐采用目标蛋白定量 MRM 技

术，或者传统的 western blotting 技术进行验证。

六、 华大优势

- 成熟的蛋白定量 iTRAQ 和 MRM 技术，项目经验达上千个 [3,4]，涉及人类、动植物、微生物等多个领域。
- 完善的蛋白定量 iTRAQ 和 MRM 技术打包验证方案，在蛋白质组学知名杂志上发表多篇文章[1,5]。

七、 参考文献

[1] Wang Y, Shan Q, Hou G, *et al*. Discovery of potential colorectal cancer serum biomarkers through quantitative proteomics on the colonic tissue interstitial fluids from the AOM–DSS mouse model[J]. Journal of Proteomics, 2016, 132: 31-40.

[2] Jenkinson C, Elliott V L, Evans A, *et al*. Decreased serum thrombospondin-1 levels in pancreatic cancer patients up to 24 months prior to clinical diagnosis: association with diabetes mellitus[J]. Clinical Cancer Research, 2016, 22(7): 1734-1743.

[3] Luo J, Tang S, Peng X, *et al*. Elucidation of cross-talk and specificity of early response mechanisms to salt and PEG-simulated drought stresses in Brassica napus using comparative proteomic analysis[J]. PloS one, 2015, 10(10): e0138974.

[4] Liu L, Li G, Sun P, *et al*. Experimental verification and molecular basis of active immunization against fungal pathogens in termites[J]. Scientific reports, 2015, 5.

[5] Zi J, Zhang J, Wang Q, *et al*. Stress responsive proteins are actively regulated during rice (Oryza sativa) embryogenesis as indicated by quantitative proteomics analysis[J]. PLoS One, 2013, 8(9): e74229.

肿瘤单细胞基因组研究方案

一、 研究背景

单细胞技术从 2009 年开始出现单细胞 RNA 定量测序，至今已得到突飞猛进的发展，发表了大量高水平文章（图 1）。目前单细胞技术主要包括单细胞基因组测序（全基因组、外显子、目标区域）、单细胞转录组测序（定性、定量）、单细胞甲基化、单细胞平行测序（对一个细胞同时做单细胞基因组和转录组测序）。应用领域主要集中在肿瘤、发育、微生物、方法学方向，少部分集中在嵌合体、神经细胞、免疫细胞方面的研究（图 2）

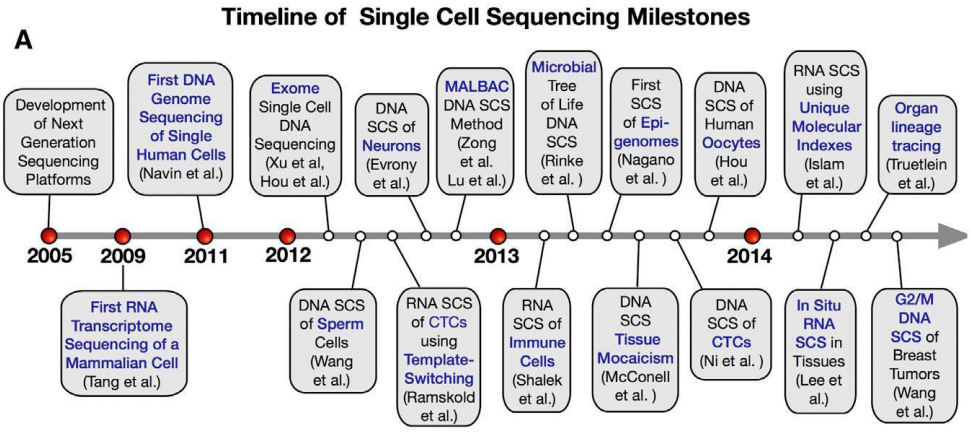


图 1 截止到 2015 年单细胞测序中里程碑事件[1]

2009 年发表第一篇单细胞 RNA 文章，

2011 年发表第一篇单细胞基因组文章，

2012 年华大在 *Cell* 同期发表两篇单细胞肿瘤文章。

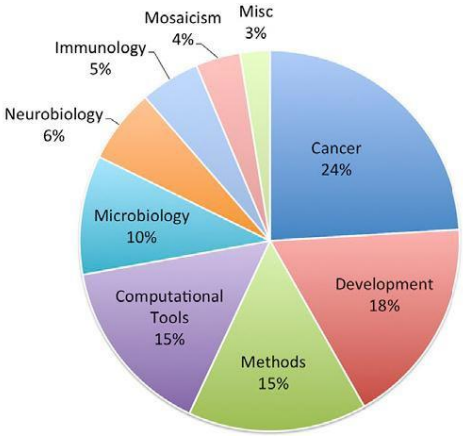


图 2 单细胞测序技术应用方向[1]

肿瘤的发生发展是由正常细胞开始，通过不断的变异积累，形成肿瘤，演化成不同的克隆群体。克隆多样性在肿瘤发展过程，例如肿瘤浸润癌旁、克隆演化、远处转移中起到关键作用。肿瘤克隆多样性使得有些克隆群体在选择压力（例如缺氧、化疗、免疫监控等）下存活下来[1]。因此，了解肿瘤克隆多样性，挖掘主要和次要克隆群体的关键突变基因，对于癌症发展机制研究、药物研究及治疗策略制定都是非常关键的。

第一篇研究癌症的单细胞基因组文章，是 2011 年冷泉港实验室发表的三阴性乳腺癌单细胞基因组研究。研究人员从 2 个乳腺癌病人不同病灶取了 100 个乳腺癌单细胞进行 DOP-PCR 扩增及测序，检测 CNV 变异，研究发现 CNV 在进化过程中出现断点式的爆发（punctuated bursts），进而形成稳定的克隆群体[2]。德克萨斯大学 MD 安德森癌症研究中心的研究人员，采用 NUC-seq 的方法（利用细胞周期将 DNA 起始量加倍，再进行单细胞基因组和外显子测序，提高覆盖度），发现非整倍体重排发生在肿瘤发生早期，并且持续稳定存在。相反，点突变是逐渐进化的，形成不同克隆群体[3]。单细胞测序研究脑胶质瘤时发现，不同的克隆群体出现 EGFR 突变趋同进化[4]。

单细胞测序也被用在血液系统肿瘤的研究，例如研究人员对 3 例急性髓系白血病患者进行单细胞测序，构建突变图谱[5]。而另一项研究对 6 例急性淋巴性白血病（ALL）患者，共 1479 个单细胞进行目标区域测序，发现 ALL 存在多个显性克隆，大部分结构变异发生在点突变之前[6]。

单细胞基因组测序也被用于移植瘤研究中，利用单细胞测序技术，研究免疫缺陷鼠肿瘤移植后，肿瘤细胞的传播和转移情况。研究发现，15 例移植瘤中，原发癌和转移癌的移植瘤均发现克隆选择，这种选择可能是小克隆群体，也可能是大的克隆群体[7]。

华大基因于 2012 年 3 月在 Cell 杂志上同期发表了两篇单细胞外显子测序技术研究癌症的文章，研究人员分别从原发性血小板增多症(ET)和肾癌中分离 90 个和 25 个单细胞，研究发现这两个病例均属于单克隆群体[8,9]。其采用的全基因组扩增方法为多重置换扩增（Multiple Displacement Amplification, MDA），DNA 起始量仅 需单个细胞基因组，最高可覆盖单个细胞 90%以上的全基因组区域，假阳性率仅 10-5，适合检测单碱基突变。2014 年华大基因又发表一篇结肠癌单细胞外显子文章，研究发现该例结肠癌病人中存在两个克隆群体，其中次要克隆群体中存在的低频突变在组织测序中会被掩盖掉[10]。

二、 方案设计

2.1 研究目的

（1）肿瘤异质性研究寻找关键变异，指导针对不同亚群的个性化治疗；

（2）细胞亚群分析，寻找与预后相关的变异信息；

（3）不同分期的癌症进化分析，研究癌症的发生发展机制，寻找与复发、转移相关的细胞亚群和变异；

（4）治疗前后的癌症进化分析，寻找与疗效相关的分子机制。

2.2 整体思路

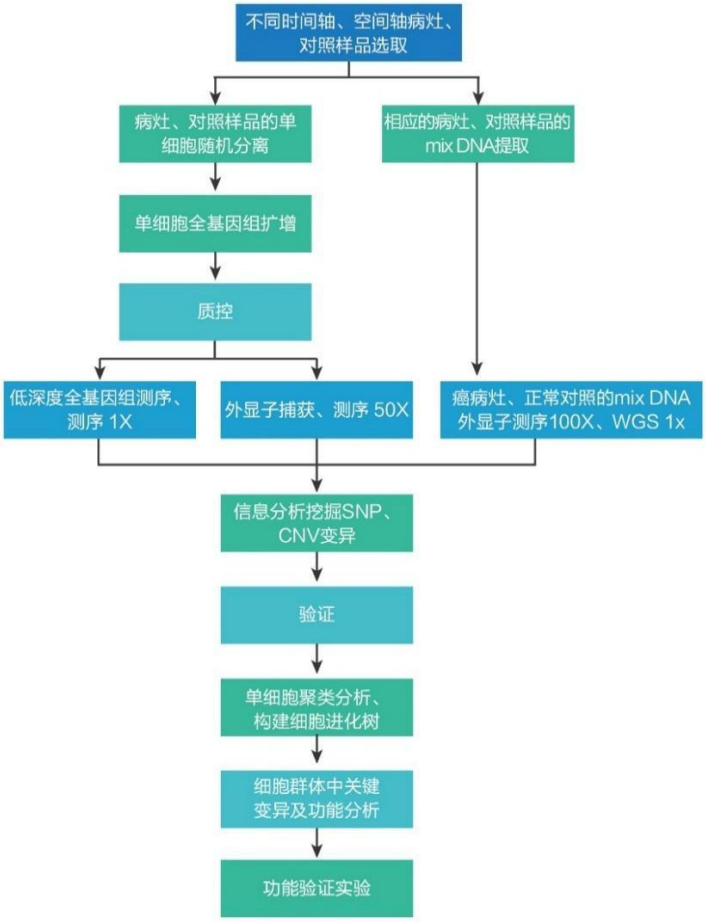


图 3 整体研究思路

2.3 样本建议

研究发现，不同时期（原发-转移、治疗前-后）的肿瘤会发生变化[11-13]，不同区域的肿瘤（例如胰腺癌、肝细胞癌、肾癌）也会存在差异[14-16]。目前，癌症异质性及进化分析主要有两种方法：（1）肿瘤个体间的比较研究；（2）肿瘤个体内部的比较研究。前者通过分析不同分期的多个病人的组织样品，分析出癌症早期和晚期

的高频突变，这种方法往往需要大量临床样本。后者从单个肿瘤个体中取多个异质性样本，构建肿瘤进化树，推断肿瘤进化过程和次要克隆群体中特有的低频突变。

选样建议：选择具有不同时间、空间的病灶样品，样品具有完整的临床信息和预后信息。建议同种类型病例至少 2 例。

时间轴

（1）治疗前活检样本

部分病人在诊断时发现原发肿瘤，在手术前会先进行一期化疗或放疗，如果选择这类病例，建议在治疗前取有相应的活检样本（biopsy），以便在后续分析时确定所测癌细胞（手术切除样本）的突变率和异质性是否受化疗影响。

（2）手术切除样本

完整的样本包括原发肿瘤、近源转移肿瘤、癌旁或血。

注：建议原发灶做组织病理学检查，确定肿瘤组织中癌细胞与间质细胞的比例。对原发灶建议进行分割处理（见“空间轴”部分）。

（3）复发灶或远端转移样本

手术切除肿瘤后进行治疗，一段时间后部分病人会发生原位复发或远端转移，选取再次手术切除或活检的样本。

空间轴

肿瘤不同病灶的异质性已在多种癌症中被证实，因此对于具有较大肿瘤组织的样本（原发/转移），需要进行分割处理，研究其异质性。

分割区域：建议 3 个以上，可以根据前期实验初步判定各分区的异质性，针对性的选取。

细胞数目选取建议

肿瘤组织具有高度异质性，需要选取大量细胞才能充分了解组织的克隆多样性，但同时也需要考虑投入经费的问题。建议策略有三种：（1）关注非编码区：单细胞全基因组测序，所需样品数量较少（2）关注编码区：单细胞外显子测序，所需样品量较多（3）关注特定区域，单细胞目标区域测序，可进行大量样本、深度测序[17]。

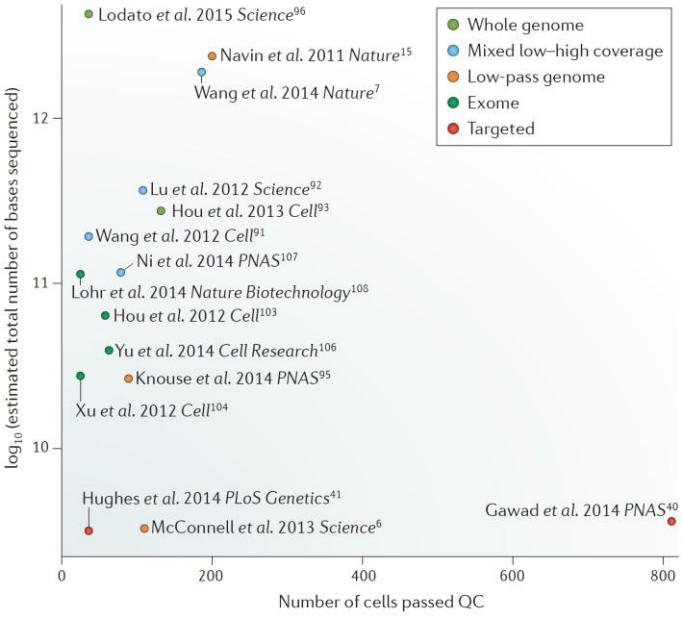


图 4 已发表单细胞文章中样品数选取情况[17]

单细胞全基因组测序和外显子一般在 50 个左右，单细胞目标区域测序，可以做到几百。理论上细胞数目越多，对克隆及进化分析的结果越准确。

2.4 测序参数

（1）组织样品测序

分离部分癌组织样品，进行外显子测序，数据一方面作为单细胞测序的对照，另一方面作为不同病灶区域样品异质性的初筛。

建议测序深度：100X。

注：可增加 1X 低深度的全基因组测序，分析 CNV。

（2）单细胞测序

对每个病灶癌组织样品，随机分离 30 个左右肿瘤单细胞，癌旁随机分离 10 个左右细胞作为对照。单细胞经过裂解、全基因组扩增、质控合格后，进行外显子测序 50X。另外，可分析癌症的 CNV 变异情况，可进行低深度全基因组重测序，建议深度 1X。

注：根据异质性不同，肿瘤细胞数量可灵活调整。

2.5 分析内容

信息分析挖掘 SNP 和 CNV 突变及突变频率，变异经过验证后，基于 SNP/CNV 变异进行癌细胞群体结构分析和聚类分析，分析肿瘤细胞的克隆群体和进化关系，并找到克隆群及其变异频谱。

分析内容展示：

（1）去除接头污染序列及低质量数据

（2）比对，产出数据统计

（3）SNP 检测、注释和统计

（4）InDel 检测、注释和统计

（5）CNV 检测、注释和统计（仅全基因组测序）

（6）SV 检测、注释和统计（仅全基因组测序）

（7）变异基因 GO 和 KEGG 通路富集分析

（8）单细胞高级分析内容：

评估细胞的 FP 和 ADO

绘制突变的 landscape（heatmap）

构建细胞的系统发生树

细胞群体的演化模式（单克隆/多克隆）

2.6 实验验证

采用细胞系或动物模型来验证关键变异的功能，可增加文章的亮点。

三、 应用案例

案例 华大参与——单细胞测序发现一例结肠癌为双克隆起源，不同克隆展现不同的突变特征[10]

文章概述

深圳华大基因联合香港中文大学、北京大学肿瘤医院/研究所的研究人员，对一例 T3N0M0 结肠癌病人的 63 个肿瘤单细胞、4 个正常单细胞样本进行外显子测序。单细胞群体遗传学分析发现肿瘤细胞群体存在两个克隆群体，不同克隆具有不同的突变特征。

研究背景

结肠癌的癌变过程有明显的分阶段特征，从息肉到癌变到转移等过程。结肠癌最常见的变异是 APC、KRAS、TP53 和 18q 丢失。肿瘤内部的异质性与表型异质性和肿瘤发展、转移、药物反应有关。

理论认为，肿瘤发展过程是符合达尔文逻辑的遗传进化过程，当细胞亚克隆获得关键的 driver 基因，就会变成快速增殖的细胞群体，进而成为肿瘤组织内的主要群体，甚至可能发生转移。依据这个模型，通过不同时期或不同病灶癌细胞的进化分析，可分析结肠癌的发生发展及其关键变异。

实验设计

图 5 文章实验设计

研究成果

1. 群体遗传学分析发现该例结肠癌存在两个细胞亚群

利用 106 个非同义突变，进行 PCA 分析，发现肿瘤细胞明显分为两个群体（图 6A）。为了进一步分析细胞的群体结构及各细胞的体细胞突变情况，进行基于 UPGMA 的系统树分析（图 6B）

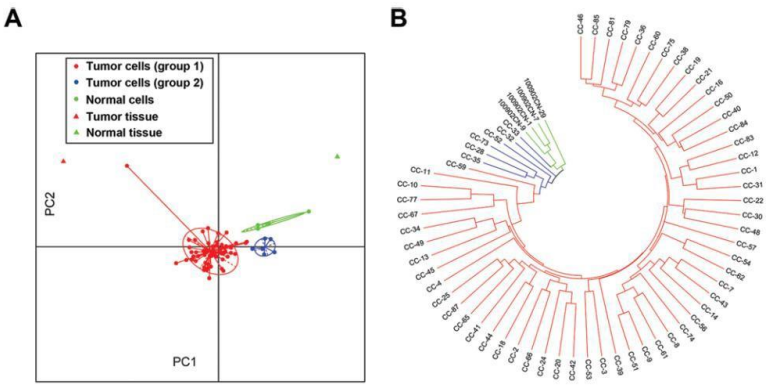


图 6 单细胞群体遗传学分析

A.主成分分析（PCA） B. 系统树分析(UPGMA)

2. 通过 driver events 聚类分析发现，两个亚群具有不同的细胞起源

群体遗传学分析发现该例肿瘤为双克隆起源。其主要克隆群体发生了 APC、TP53 突变，次要克隆群体 CDC27 和 PABPC1 为主要突变，但 APC 和 TP53 未发生突变（图 7）。单细胞水平发现 SLC12A5 为高频突变，但群体水平突变率极低。

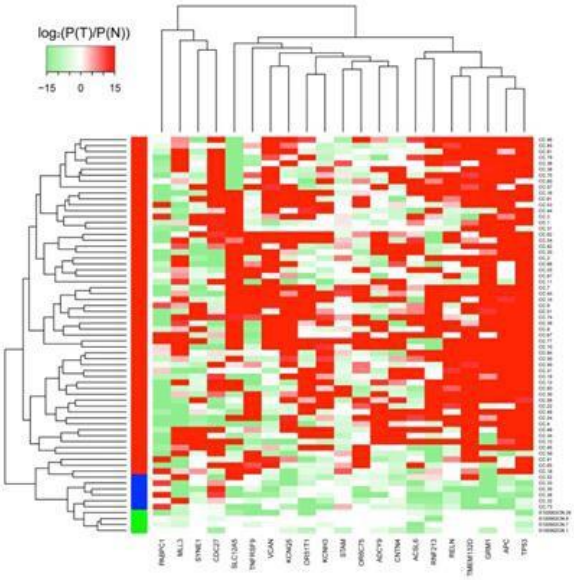


图 7 结肠癌单细胞不同克隆群体的 driver event 分析

3. 分析不同克隆群体的突变基因频率和变异频谱

对结肠癌单细胞及另外 21 例结肠癌的变异结果进行分层聚类，发现 major clone (red)主要包含 TP53 和

APC，minor clone (blue)主要包含 CDC27 和 PABPC1，normal cells (green)不含突变。

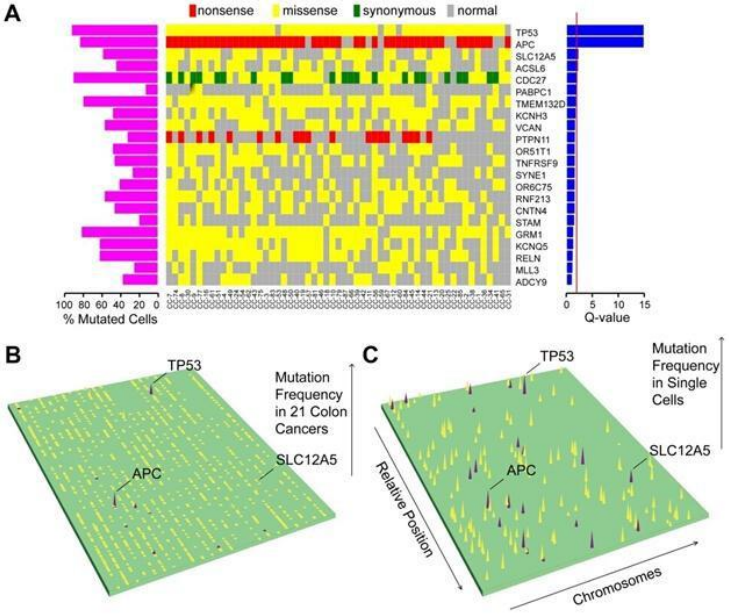


图 8 突变图谱及 dirver genes 预测

(A) major clone 中体细胞突变。左侧代表突变频率，右侧表示预测的 driver genes Q-score;

(B) 21 例结肠癌的体细胞突变图谱，尖峰高度代表突变频率; (C) 结肠癌单细胞的突变图谱。

4. 功能验证发现 *SLC12A5* 可能对结肠癌具有致癌效应

单细胞测序发现的稀有突变 SLC12A5，利用一系列的试验，发现 SLC12A5 可能通过扰乱细胞增殖和凋亡的平衡，促进结肠癌的发生。

（1）利用稳定转染后的细胞进行 RT-PCR 在正常组织和癌细胞系中检测，发现 SLC12A5 在癌细胞系中过表达（图 9B）

（2）colony formation assay、xCELLigence 实时细胞分析系统的检测发现，突变的 SLC12A5 增强了菌落形成能力和细胞增殖能力（图 9C-E）

（3）利用流式细胞仪检测，发现 SLC12A5 突变减少了细胞凋亡（图 9F）

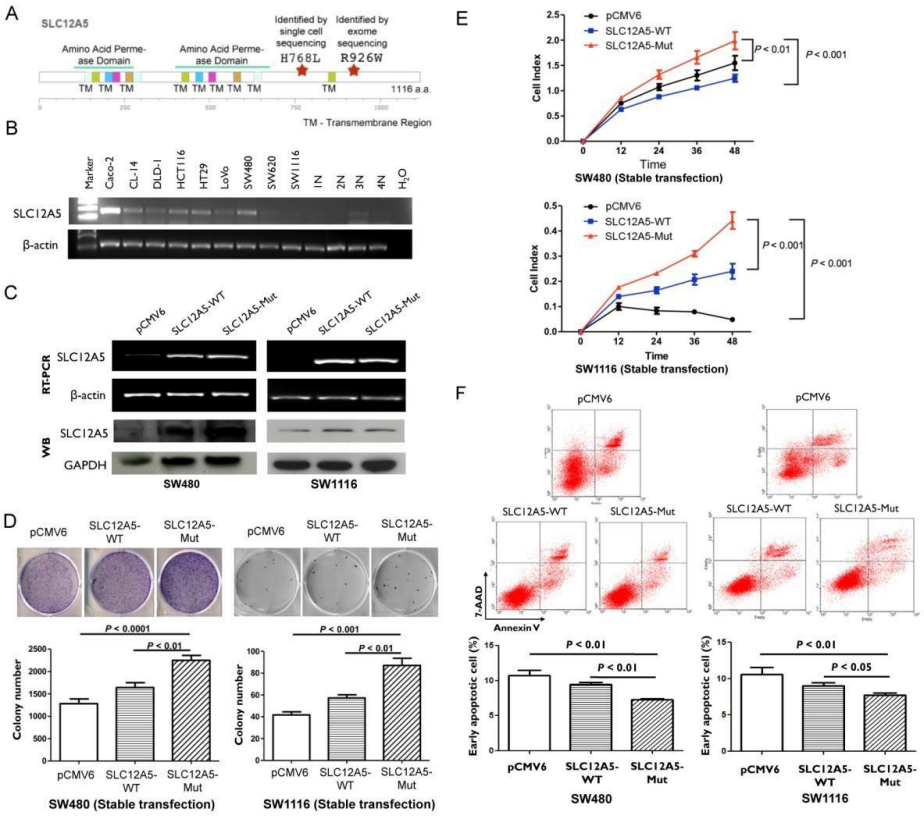


图 9 *SLC12A5* 功能验证试验

(A) *SLC12A5* 结构域及点突变位点，H768L 为单细胞测序检测到的，R926W 为 1/21 个结肠癌中通过组织测序发现的; (B) RT-PCR 检测 *SLC12A5* mRNA 在癌细胞系和正常肠组织中的表达; (C) *SLC12A5* 在稳定转染（pCMV6 载体、pCMV6-SCL12A5(WT)、pCMV6-SCL12A5(Mut)）的两个结肠癌细胞系 SW480 和 SW1116 中异常表达情况; (D) colony formation assay; (E)通过 xCELLigence System 检测细胞增殖情况; (F) 通过流式细胞术检测细胞凋亡情况。

四、 常见问题

问题 1 单细胞分离技术有什么建议？

答： 单细胞测序技术研究的前提是，获得正确的目标细胞。目前可以使用的技术有显微操作、流式细胞术（FACS）、激光捕获显微切割（LCM）、微流控芯片等。

显微操作最常见的方法是口吸管法，该方法简单、便宜、对细胞没有损伤。缺点是由于没有对细胞进行特殊处理，因此无法在显微镜下分辨细胞类型，通量低。

流式细胞术可利用特定的细胞标志物，从细胞悬液中富集目标细胞。与传统的荧光镜检相比，具有速度快、

精度高等优点，通量也比较大。但缺点是需要大量细胞制作细胞悬液，不适合稀有样品的分离。另外，富集到的细胞状态无法观察，对细胞有一定的损伤。

激光捕获显微切割可以从组织切片或细胞涂片上将所要研究的细胞直接从组织三维构造中分离出来，是较为理想的细胞提取工具。该技术缺点是需要专业的人工操作、通量低，由于分离单细胞时很容易切割到相邻细胞的胞质，因此该技术分离的细胞样品不太适合单细胞转录组测序。并且其处理的样本往往已经经过染色、冰冻或包埋处理，对细胞活性损伤较大，不太适合用于单细胞测序。

微流控芯片技术近年来发展很快，在这一分析平台上也建立了用于单细胞分析的方法。微流控芯片具有网络式通道结构，可结合多种物理、化学或电学手段，操纵细胞通过或驻留在微流控芯片通道内的任意位置，从而使单细胞技术、筛选等分析操作大大简化，实现对目标细胞的高效识别和捕获。

目前华大采用的单细胞分离技术，主要有口吸管法、基于微流控技术的单细胞分离技术。单细胞 DNA 可以送组织来华大做口吸管法，也可以客户自行分离。其他方法，如流式细胞术也可以用于单细胞分离，但扩增效果需要预实验来评估。

问题 2　单细胞全基因组扩增会出现 ADO（Allele Dropout），如何避免？

答：华大采用的全基因组扩增技术为 MDA（Multiple Displacement Amplification），具有一定的等位基因缺失的情况，由于这种情况是随机出现的，可通过增加细胞数目来降低 ADO 对结果的影响。

问题 3　华大单细胞技术可以做到哪些水平的？

答：华大基因从 2010 年开始就投入单细胞研发中，属于国内早期就开展单细胞测序研发的公司。目前已经能从 DNA、RNA（转录组）、表观三个水平对单细胞进行研究，并且可以对同一个细胞进行 DNA 和 RNA 的平行测序。

问题 4　对于已经发表过单细胞文章的癌症类型，有何建议？

答：科研方案设计时，首先需调研一下该类型肿瘤的研究进展，对于已做过单细胞研究的，可以考虑从以下思路进行：（1）不同癌症亚型，例如 AML、ALL（2）单细胞目标区域的角度进行更大细胞数量的研究（3）结合临床情况，选取特殊的案例。

五、 华大优势

- **高保真扩增**。Phi29 DNA 聚合酶具有 3’ → 5’核酸外切酶活性，扩增时的保真度比 Taq DNA 聚合酶高 1000 倍。
- **扩增产物长**。多在 10kb。
- **扩增覆盖度高**。在人细胞中进行测序，最高可达到 99%覆盖度。
- **扩增错误率低**。扩增错误率低至 10^{-6}。
- **扩增偏好性降低**。技术优化后，可极大地降低扩增偏向性，实现全基因组较均一的扩增，使该技术存在同时可检测 SNP 和 CNV、SV 的可能性。
- **适合多个物种的扩增**。随机引物扩增，可针对多个物种（植物除外）。
- **丰富的项目经验。**已完成上千个细胞的扩增及测序，并发表文章成果 6 篇。

六、 参考文献

[1] Wang Y, Navin N E. Advances and applications of single-cell sequencing technologies[J]. Molecular cell, 2015, 58(4): 598-609.

[2] Navin N, Kendall J, Troge J, *et al*. Tumour evolution inferred by single-cell sequencing[J]. Nature, 2011, 472(7341): 90-94.

[3] Wang Y, Waters J, Leung M L, *et al*. Clonal evolution in breast cancer revealed by single nucleus genome sequencing[J]. Nature, 2014, 512(7513): 155-160.

[4] Francis J M, Zhang C Z, Maire C L, *et al*. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing[J]. Cancer discovery, 2014, 4(8): 956-971.

[5] Hughes A E O, Magrini V, Demeter R, *et al*. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing[J]. PLoS Genet, 2014, 10(7): e1004462.

[6] Gawad C, Koh W, Quake S R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics[J]. Proceedings of the National Academy of Sciences, 2014, 111(50): 17947-17952.

[7] Eirew P, Steif A, Khattra J, *et al*. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution[J]. Nature, 2015, 518(7539): 422-426.

[8] Hou Y, Song L, Zhu P, *et al*. Single-cell exome sequencing and monoclonal evolution of a

JAK2-negative myeloproliferative neoplasm[J]. Cell, 2012, 148(5): 873-885.

[9] Xu X, Hou Y, Yin X, *et al*. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor[J]. Cell, 2012, 148(5): 886-895.

[10] Yu C, Yu J, Yao X, *et al*. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing[J]. Cell research, 2014, 24(6): 701-712.

[11] Welch J S, Ley T J, Link D C, *et al*. The origin and evolution of mutations in acute myeloid leukemia[J]. Cell, 2012, 150(2): 264-278.

[12] Ding L, Ley T J, Larson D E, *et al*. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing[J]. Nature, 2012, 481(7382): 506-510.

[13] Ding L I, Ellis M J, Li S, *et al*. Genome remodelling in a basal-like breast cancer metastasis and xenograft[J]. Nature, 2010, 464(7291): 999-1005.

[14] Yachida S, Jones S, Bozic I, *et al*. Distant metastasis occurs late during the genetic evolution of pancreatic cancer[J]. Nature, 2010, 467(7319): 1114-1117.

[15] Tao Y, Ruan J, Yeh S H, *et al*. Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data[J]. Proceedings of the National Academy of Sciences, 2011, 108(29): 12042-12047.

[16] Gerlinger M, Rowan A J, Horswell S, *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing[J]. New England journal of medicine, 2012, 366(10): 883-892.

[17] Gawad C, Koh W, Quake S R. Single-cell genome sequencing: current state of the science[J]. Nature Reviews Genetics, 2016, 17(3): 175-188.

病毒性肿瘤发生机制研究方案

一、 研究背景

1909 年 Rous 肉瘤病毒的发现使人们意识到肿瘤的发生与某些病毒的感染息息相关。肿瘤的发生是遗传因素和致癌因素相互作用的结果。致癌因素包括物理、化学和生物因素。生物因素中的病毒与肿瘤的发生关系最为密切。许多研究发现，某些病毒的感染能诱发肿瘤。流行病学调查和分子生物学的研究表明二者之间确实存在着密切的关系。1989 年在“DNA 病毒在人类肿瘤中的作用”国际研讨会上首次确定了至少有 3 种病毒与人类肿瘤的关系密切。这就是肝炎病毒与肝细胞癌；EB 病毒（Epstein-Barr virus，EBV）与 Burkitt 淋巴瘤、鼻咽癌；人乳头瘤病毒（human papilloma virus，HPV）与宫颈癌有直接关联。另外，腺病毒、人类免疫缺陷病毒（HIV）、多瘤病毒等亦与肿瘤发生有一定的关系[1]。

那么病毒侵入人体之后发生了什么，怎么影响和调控基因表达，从而引发癌症？目前为止还知之甚少。有些研究发现病毒侵入人体之后会整合到宿主的基因组内，使得宿主基因发生断裂突变，同时还会影响宿主基因组的甲基化和组蛋白的修饰，从而改变宿主基因表达，甚至引起癌变[2]。所以综合的多组学研究病毒性肿瘤的发生发展机制可以提高临床决策,促进个性化医学,加快肿瘤生物标志物和治疗靶标的临床应用[3]。

二、 方案设计

图 1 病毒性肿瘤发生机制研究方案

2.1 样本建议

样品量建议 100 对以上癌组织和癌旁组织或外周血。癌症具有高度特异性，即不同病人之间、不同癌组织间的突变特异性非常高。对于 100 个病人及其对照的深度测序研究，这个数量级能具备临床上的统计学意义，我们从目前的文章成果也可以看出，研究规模已经呈现出从最初的小样本向大样本群体发展的趋势。

肿瘤组织中肿瘤纯度在 70%以上。样本最好要有完整的临床信息，包括性别，年龄，种族，以及详细的临床病例特征等，这对后续的关联分析至关重要。

2.2 实验技术

采用全基因组重测序和转录组测序技术，通过不同样品组的比较及筛选，寻找病毒插入整合位点和突变频谱，在转录水平验证基因整合和突变带来的影响，并对断点基因进行 Pathway 的富集分析。

2.3 测序参数

建议全基因组测序每个样品测序量≥150Gb，转录组测序每个样本≥8G。

2.4 分析结果

(1) 病毒整合位点

对于有病毒感染史的肿瘤病例，首先是查看肿瘤基因组中病毒基因组整合的位置以及整合频率。

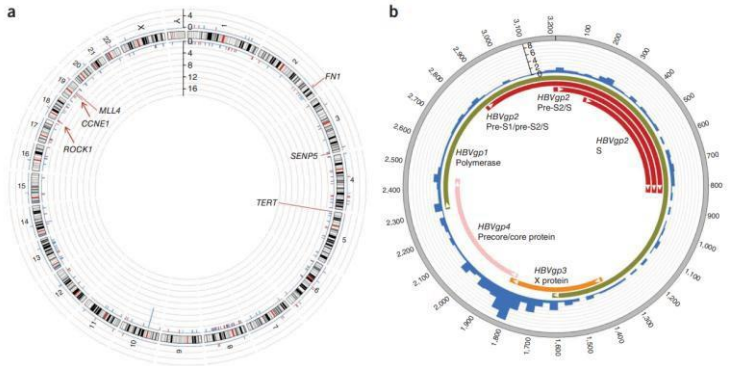


图 2 病毒整合位点展示[4]

a 图展示的是 88 个感染 HBV 患者的基因组病毒基因整合分布情况，内圈是肿瘤和癌旁组织基因组病毒整合情况，外圈是非肿瘤个体基因组病毒整合情况。b 图展示的是乙肝病毒基因组整合频率情况，蓝色柱状图展示的是基因整合频率，绿色、粉色、红色、橘色展示的是乙肝病毒基因。

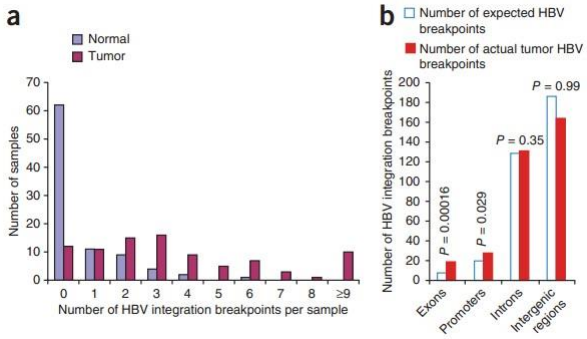


图 3 病毒整合位点频率和类型[4]

a 图展示的是 88 个患者中每个样品病毒基因整合数目统计，蓝色的是非癌症病人的乙肝病毒整合数量，基本在 4 次以内，大部分是没有病毒整合的；红色的是癌症病人的乙肝病毒整合数量，少量样品是没有 病毒整合的。b 图展示的是乙肝病毒在人基因组上的整合位置分布，白色柱状图是预期中的基因整合位置频率，红色展示的是实际上的基因整合位置频率，二者基本吻合，且主要集中在内含子和基因间区。

(2) 断点统计分析

分析肿瘤基因组中断点所在的基因信息（如基因名称、插入位置、病毒基因等），一般用表格展示，如下表所示。也可以用图展示（如图 4）

Table 1 Characterization of recurrent genes with HBV integration breakpoints in HCC

Human genes	Integration locations	Affected HBV proteins	Affected samples
TERT	Intron 2	Polymerase, pre-S2/S, S	13T, 268T
	Intron 6	Precore/core protein, core and e antigen	198T
	Promoter	X protein	14T, 22T, 38T, 60T, 64T, 81T
		X protein, precore/core protein	65T, 73T, 261T
		Polymerase	266T
		Precore/core protein	63T
		Precore/core protein, core and e antigen	22T, 266T
		Pre-S1/pre-S2/S	58T
		Pre-S2/S	34T, 261T, 92T
		Pre-S2/S, S	46T
MLL4	Exon 3	Precore/core protein, core and e antigen, X protein	70T, 95T
	Exon 6	Precore/core protein, X protein	186T
	Intron 3	Polymerase, X protein	116T, 159T
		X protein, precore/core protein	159T
		X protein	49T, 53T, 62T
		Precore/core protein, core and e antigen	62T
		Pre-S2/S, S, X protein	71T

表 1 肝癌中HBV 整合断点统计表[4]

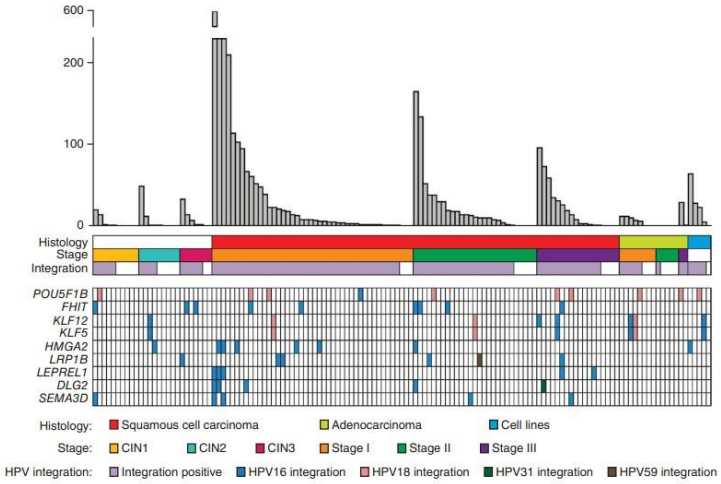


图 4 135 个宫颈癌样品信息和HPV 整合信息展示[5]

135 个样品的临床信息注释：组织学水平分为 26 个宫颈癌前病变患者（CINs，有三个阶段的样品 CIN1、CIN2、CIN3）；104 个宫颈肿瘤（鳞癌和腺癌）；5 个宫颈癌细胞系。

病毒基因组整合阳性样品（紫色标示），详细列出具体整合的基因。

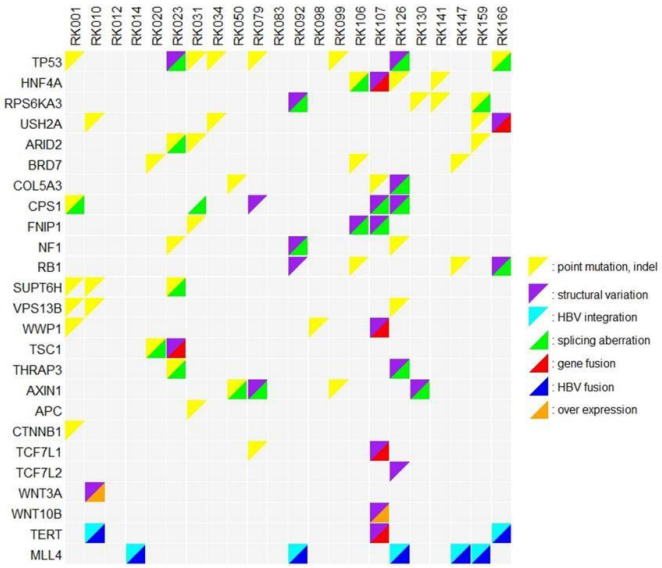


图 5 22 个肝癌样品中突变类型展示[6]

横坐标代表 22 个样品；纵坐标代表基因。不同颜色三角表示不同类型的变异（包括 SNP、InDel、结构变异、HBV 整合、可变剪切、基因融合、HBV 融合、超表达）。

(3) 断点基因表达分析

结合全基因组测序和转录组测序结果，查看样品中有断点的基因表达情况，研究病毒基因组插入整合会如何影响宿主的基因表达。研究方法如下图 5 所示：

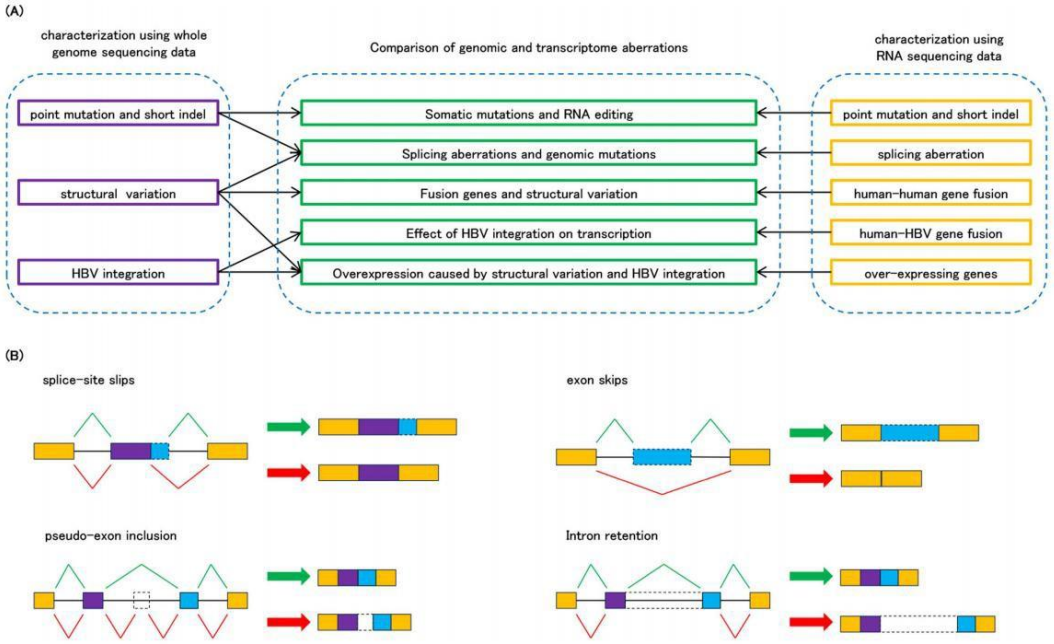


图 6 全基因组测序联合 RNA-seq 的研究思路[6]

a 图展示的是全基因组测序联合 RNA-seq 研究的大致思路；

b 图展示的是四种异常剪切类型。绿色的线和箭头代表的是正常剪切；红色的线箭头代表的是异常剪切。

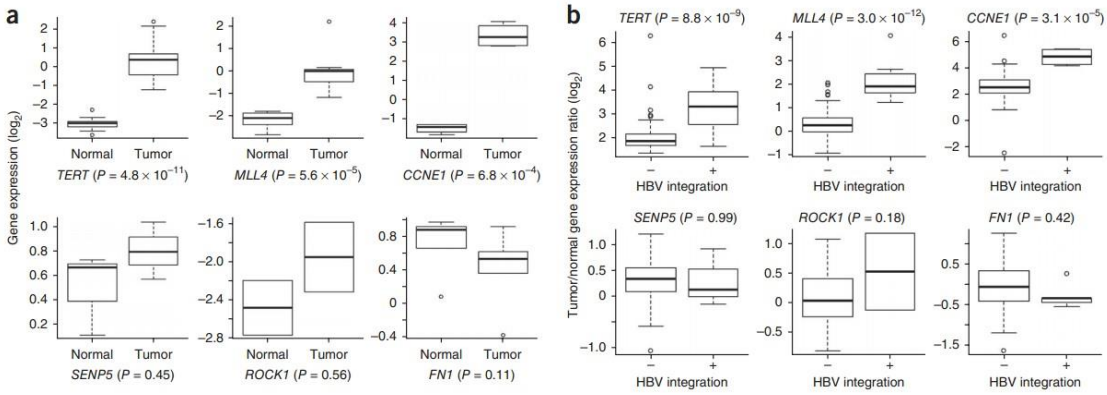


图 7 HBV 整合对肝癌基因表达的影响[4]

(4) 断点基因 Pathway 分析

发现断点基因后需要继续探究病毒插入宿主基因组是如何导致癌症发生发展的，那么需要对断点基因进行 pathway 分析，是哪些通路发生了变化，从而推断癌症发生发展的机制。

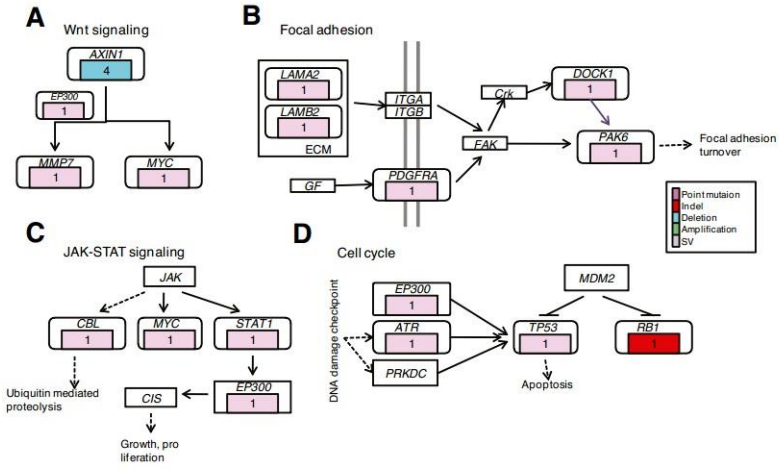


图 8 肝癌中突变频繁的癌症通路[7]

本图展示了高频突变基因影响的通路，包括(A) Wnt, (B)焦点粘连, (C) JAK-STAT 信号和 (D) 细胞周期通路。

2.5 项目周期

样品检测合格后，建库+测序+标准信息分析：约 40 个工作日，实际项目完成时间根据所选具体样本数以及信息分析条款决定。

2.6 预期结果

借助高通量测序平台，通过对感染组织和非感染组织（对照样品）进行 DNA 和 RNA 层面的分析，找出病毒整合位点和整合机制，确定病毒感染与肿瘤发生发展的机制。

2.7 实验验证

挑选目标或者候选基因，通过 PCR 和 Sanger 测序进行验证，检测测序结果是否和 Sanger 测序结果一致。还可以使用荧光原位杂交（FISH）来验证病毒基因组的整合位置。

三、 应用案例

案例 1 华大参与——全基因组测序研究肝癌组织的 HBV 整合位置[4]

图 9 研究流程图

摘要

用高通量测序的方法对 81 对 HBV 阳性的 HCC（肝细胞癌）和 7 对 HBV 阴性的 HCC 病例进行研究，发现 HBV 在肝癌中的整合频率是 86.4%，明显高于在癌旁组织中的整合频率 30.7%。CNV 在断点处（即 HBV 整合位点处）明显增加，同时染色体的不稳定性也增加。大约 40%的 HBV 插入片段是 HBV 基因组上坐标在 1800bp 左右的 X gene 和core gene。确定并验证了已知的与癌症相关基因 TERT, MLL4 和 CCNE1 中的 HBV 整合事件（在 4 个以上的样本中出现），这些基因在肿瘤中表达上调。同时证据表明 HBV 整合与病人生存期相关。

方法思路

样本： 88 对原发中国 HCC 患者其中 81 对 HBV 阳性+7 对 HBV 阴性，全基因组重测序，6 个 sample 100X，其余是 30-40X，PE reads 分别比对 HG19 和 HBV（NC-003977）序列，每个整合位点至少 2 条 PE reads 支持。

其他研究方法

所有的样本都做过芯片分析，来研究 HBV 整合对宿主基因组拷贝数变异（CNV）的影响。

其中 9 对样本做转录组测序，来验证部分断点对基因表达的影响。

断点检测准确率验证

在 6 个高频插入基因中随机选取 32 个断点，在 22 个样本中进行 PCR 验证，验证率为 82%。

结论

比对 HBV 在 HCC 上的整合位点：

（1）在肿瘤中的整合频率是 86.4%，明显高于在癌旁组织中的整合频率 30.7%。

（2）40%以上的 HBV 插入片段是在 HBV 基因组的 1800bp（HBx 和 core 基因 overlap 的地方）位置附近，可能因为 HBV 的增强子和 ORF 复制位点在这。

（3）确定了 3 个 gene（*TERT, MLL4, CCNE1*）为 HBV 整合热点区域，HBV 阳性肿瘤中有 42%整合位点是在这三个 gene 上。

（4）通过 Sanger 测序发现整合出现了 human-HBV 的阅读框内融合基因，可能 HBV 通过这种机制来控制癌基因和抑癌基因，RNA-Seq 和芯片的数据显示 HBV 的整合使这 3 个基因的表达上调。

（5）HBV 的整合增加了 HBV 整合位点周围片段的拷贝数。

（6）HBV 整合数与肿瘤大小、血清 HBsAg 和α-胎蛋白水平都正相关，有 HBV 整合的患者患癌年纪相对较小，HBV 整合数>=3 的患者存活期显著比整合数<=3 的患者短(P=0.037)。

案例 2　华大参与——宫颈癌全基因组测序鉴定了 HPV 整合热点和整合机制[5]

Genome- wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology - mediated integration mechanism

摘要

宫颈癌中 HPV 整合是一个关键事件。通过全基因组测序，在 26 个宫颈上皮瘤、104 个宫颈癌和 5 个细胞系中发现了 3667 个 HPV 整合断点。除了已经报道的 6 个频繁整合位点（POU5F1B (9.7%), F H I T (8.7%), K L F 1 2 (7.8%), K L F 5 (6.8%), L R P 1 B (5.8%)和 L E P R E L 1 (4.9%)），研究还发现了 3 个新的热点【H M GA 2(7.8%), D LG 2 (4.9%)和 S E M A 3 D(4.9%)】。当HPV 整合到 F H I T 和 L R P 1 B，会导致这两个蛋白表达下降；当 HPV 整合到 MYC 和 HMGA2 的内含子，会提高二者的蛋白表达。另外，还发现在整合断点附近会有人和 HPV 基因组之间微同源序列富集，说明病毒和人 DNA 融合可能是通过微同源介导的 DNA 修复途径来实现的。

研究方法

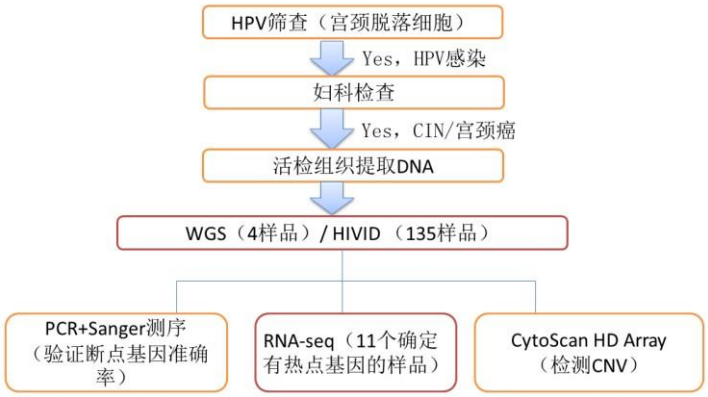


图 10 研究流程

结果

1、发现整合断点

对 135 个样品进行 HIVID（高通量病毒检测技术），在 103 个样品中共检测出 3666 个 HPV 整合位点。HIVID 技术具有高灵敏度，不仅检测了 10/11 个 WGS 鉴定的断点，还检测出了 135 个 WGS 没有鉴定出的断点。这些断点虽然遍及全基因组，但是呈现出整合热点。值得注意的是，3320 个断点（大于 90%）是分布在已注释基因的 500Kb 以内，其中 1546 个断点（大约 42%）是分布在基因内的，且很多是在肿瘤发生相关的通路内。说明病毒基因组最初是根据就近原则，随机分布在宿主基因组。但是随着肿瘤的发生发展，一些反复整合热点对肿瘤细胞有选择优势，所以导致了在不同样品中整合热点基因出现。

HPV 病毒基因整合是全方位的，可能这有利于病毒适应肿瘤发展过程中的环境变化。

把病毒整合与临床病理指标联合分析，期望找到病毒整合特异的生物标记，用于指导后续的临床诊断。发现不论是病毒整合基因数目还是频率，癌症早期都要高于晚期。

2、断点基因表达改变

全基因组和转录组的数据结果显示当 HPV 整合到 FHIT 和 LRP1B 的内含子，他们的蛋白表达会下调；当 HPV 整合到 MYC 和 HMGA2 的侧翼序列，他们的蛋白表达会提高。

3、病毒基因整合假说

整合位点处的人基因组和病毒基因组附近存在微同源序列（2-5bp），微同源序列越小、侧翼序列越长，发生整合的概率越大。因为 FoSTeS 和 MMBIR 通路通常是由自身基因组不稳定引发的。我们分析了整合位点侧翼序列，发现整合位点侧翼序列基本是基因组不稳定元件，包括短的串联重复、SINE-Alu、脆性 X 断点簇重复、长末端重复-内生逆转录病毒序列（LTR-ERV1）、微卫星序列。Array-CGH 的结果显示，拷贝数变异不仅

发生在所有的 HPV 整合位点，也在微同源序列附近富集。我们提出一个新的病毒整合模型：某种环境下（比如病毒入侵），自身基因组元件变得不稳定，倾向形成 DNA 断点，或者停止复制叉，导致 FoSTeS 和 MMBIR 通路激活。由于微同源侧翼序列出现断点，HPV 基因组趁火打劫，利用微同源介导的 DNA 修复机制整合到宿主基因组。

四、 项目风险

做完全基因组重测序/转录组测序以后，可能有部分确定有病毒感染的肿瘤却找不到病毒基因组插入位点，这个是正常的，因为癌变的机制是复杂的，已经报道的研究结果也显示病毒感染的肿瘤基因组没有 100%的病毒基因组插入。所以需要尽量加大样品量，增加统计意义和结果的可靠性。

五、 华大优势

- **丰富的项目经验。**专业的信息分析团队，已完成 50,000 多例肿瘤样品测序，发表多篇高水平文章。
- **全面的样品类型。**提供血液、FFPE、微量等特殊样品服务，微量样品建库低至 ng 级，完全解决样品准备的后顾之忧。
- **贯穿的研究思路**。从基因组、转录组、表观组、蛋白组等多组学角度提供解决方案，一站式服务。
- **清晰的结题报告。**结果图片更加直观，可用于文章撰写；结题报告结构优化，有效提升阅读体验。

六、 参考文献

[1] 王振城, 赵海燕. 病毒感染与人类肿瘤发生的关系[J]. 医学综述, 2010, 17(13): 1983-1985.

[2] Tian Y, Ou J J. Genetic and epigenetic alterations in hepatitis B virus-associated hepatocellular carcinoma[J]. Virologica Sinica, 2015, 30(2): 85-91.

[3] Miao R, Luo H, Zhou H, *et al*. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis[J]. Journal of hepatology, 2014, 61(4): 840-849.

[4] Sung W K, Zheng H, Li S, *et al*. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma[J]. Nature genetics, 2012, 44(7): 765-769.

[5] Hu Z, Zhu D, Wang W, *et al*. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism[J]. Nature genetics, 2015, 47(2): 158-163.

[6] Shiraishi Y, Fujimoto A, Furuta M, *et al*. Integrated analysis of whole genome and transcriptome

sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers[J]. PloS one, 2014, 9(12): e114263.

[7] Limei Ouyang, Jeeyun Lee, Cheol-Keun Park, Mao Mao, *et al*. Whole-genome sequencing of matched primary and metastatic hepatocellular carcinomas. BMC Medical Genomics, 2014, 7: 2.

肿瘤融合基因研究方案

一、 研究背景

基因融合是指染色体上两个异位的基因嵌合在一起，形成一个嵌合基因的现象。这种现象一般是由于染色体发生易位、缺失或者倒置造成的，它们在癌症的发生机制上扮演着重要的角色，并且可以作为诊断和治疗癌症的靶标。基因融合现象最早在血液系统恶性肿瘤中被发现，其中以慢性粒细胞白血病中 *BCR-ABL* 基因融合最为经典[1]。该融合是由于人基因组上第 9 号染色体上的 *ABL* 原癌基因和第 22 号染色体上的 *BCR* 基因相互易位导致的，融合导致蛋白激酶持续性激活，使得白细胞过分增殖而出现慢性粒细胞白血病的症状。随着对基因融合的深入研究，科研人员发现，无论是血液系统肿瘤还是实体瘤中，都存在着基因融合的现象。在实体瘤基因组学的研究中，研究人员发现了若干高发的基因融合，例如前列腺癌中的 *TMPRSS2-ERG* [2-3]、小细胞肺癌中的 *EML4-ALK* [4]、结直肠癌中的 *VTI1A-TCF7L2* [5]等等。这些都表明了基因融合在实体瘤发生发展过程中扮演着重要的角色。

传统基因融合研究方法主要基于 PCR 和荧光原位杂交（FISH）技术，这两种技术具有通量低、操作复杂、不便于大规模样品筛查的缺点。随着第二代测序技术的发展，尤其是高通量的 RNA 测序技术（RNA-Seq），大大加快了基因融合研究的进展。RNA-Seq 具有通量高、研究成本低、检测精度高和检测范围广的优点，这些优点使得 RNA-Seq 技术广泛应用于肿瘤的基因融合研究。此外与全基因组测序相比，RNA-Seq 除了能找到由于重排导致的融合外，还能找到更多转录水平上的融合，并且与全基因测序相比，RNA-Seq 的价格更加低廉。基于 RNA-Seq 二代高通量测序数据，市面上已经有很多检测融合基因的软件，比如 chimerascan、deFuse、FusionHunter、SnowShoes-FTD TopHat-Fusion，还有华大自主开发的 SOAPfuse 软件。

本方案利用 SOAPfuse[6]、FusionCatcher、JAFFA 等主流软件对样本中存在的融合基因进行全面的筛查，另外转录组测序的二代测序数据还可以结合转录组数据信息分析流程，进行基因表达、基因结构优化、可变剪接、新转录本预测及注释、SNP 检测等一系列后续分析，并从基因表达结果中，筛选出样品间差异表达的基因，基于差异表达基因，进行 GO 功能显著性富集分析和 pathway 显著性富集分析。后续研究还可以结合下游的蛋白功能验证，筛选合适的靶向药物小分子等。

二、 方案设计

2.1 整体思路

研究对象：常见肿瘤类型

研究目的：鉴定肿瘤组织中存在的融合基因

研究手段：转录组测序结合生物信息学分析

图 1 整体研究思路

2.2 样本建议

（1）所选取的**样本应具有有明确的临床信息**。包括：首次发病时间；性别、年龄；家族史；病理学报告，分级，分期等；

（2）组织样本要求保存良好，能够提取符合测序要求数量和质量的核酸；

（3）根据 ICGC 取样标准，**癌组织中癌细胞的纯度要达到 80%以上**，如果由于样本本身原因达不到此纯度，可以通过增加测序深度减少偏差；

（4）肿瘤研究**一般需要成对的样本**，对于融合基因检测来说，由于结构变异较大，如果无正常对照样本，可先对病例样本进行分析，然后再在正常大样本中进行验证；

（5）转录组测序至少需要 RNA 200 ng，样品浓度≥ 20 ng/μL，OD260/280= 1.8~2.2。

2.3 实验技术

构建小片段文库：200 bp 或 250~300 pb 插入片段。具体流程如下：

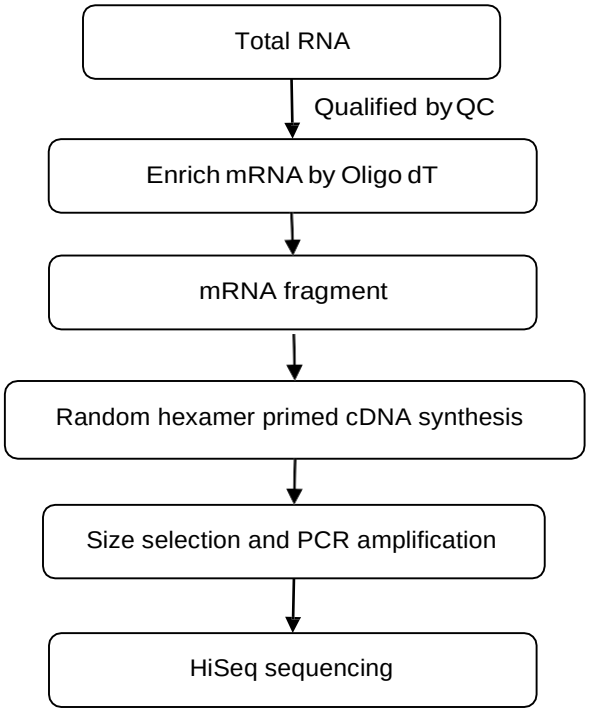


图 2 转录组建库流程图

2.4 测序参数

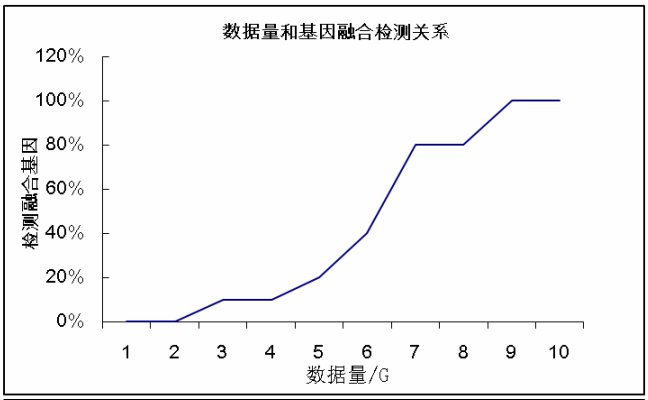
利用 Illumina HiSeq 平台进行转录组测序，应用华大自主研发的 SOAPfuse 对融合基因进行检测。

测序平台：Illumina HiSeq 平台。

测序策略：101PE，151PE。

测序数据量：建议 8G clean data。根据图 3 模拟数据可看出，数据量达到 9-10 G 基本能够检测出所有的基因融合。考虑后续的实验，一般推荐 8 G 数据量。

数据量（G）	gene fusion个数
1	1
2	1
3	3
4	3
5	4
6	6
7	6
8	7
9	8
10	8



数据量（G）	gene fusion个数
1	0
2	0
3	1
4	1
5	2
6	4
7	8
8	8
9	10
10	10

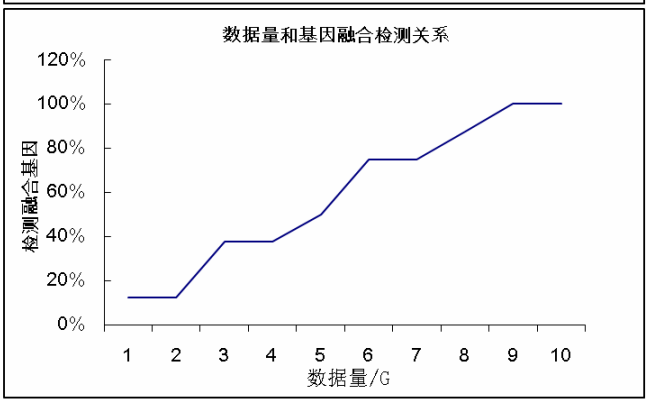


图 3 测序数据量与融合基因检测饱和度模拟数据分析

2.5 分析内容

首先对 Illumina HiSeq 平台测序所得的数据 raw data 进行质控（QC），以确定测序数据是否适用于后续分析。质控后，经过滤得到 clean reads，用比对软件[7]将 clean reads 比对到参考序列。比对后，统计 reads 在参考序列上的分布情况及覆盖度，判断比对结果是否通过第二次质控（QC of alignment）。若通过，则利用分析软件进行融合基因的检测。另外对符合质控的数据还可以进行常规的表达谱信息分析，包括基因表达、可变剪接、新转录本预测及注释、SNP 检测等分析。

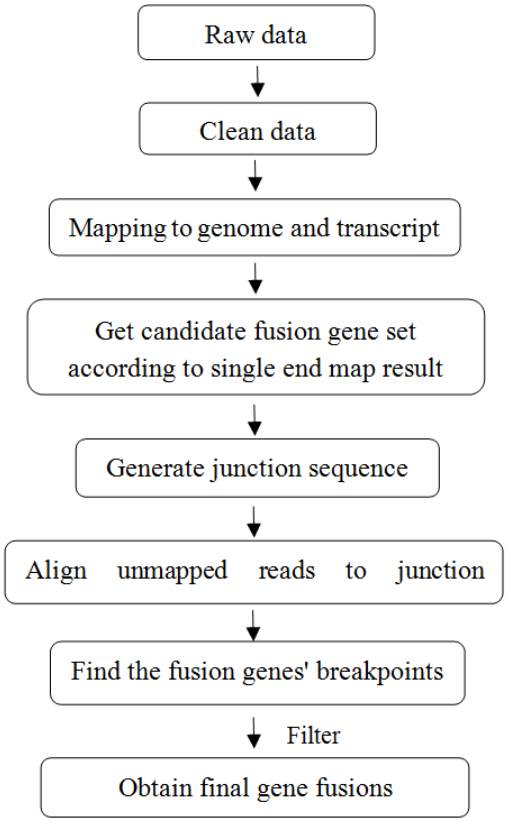


图 4 融合基因鉴定信息分析流程图

2.6 分析结果

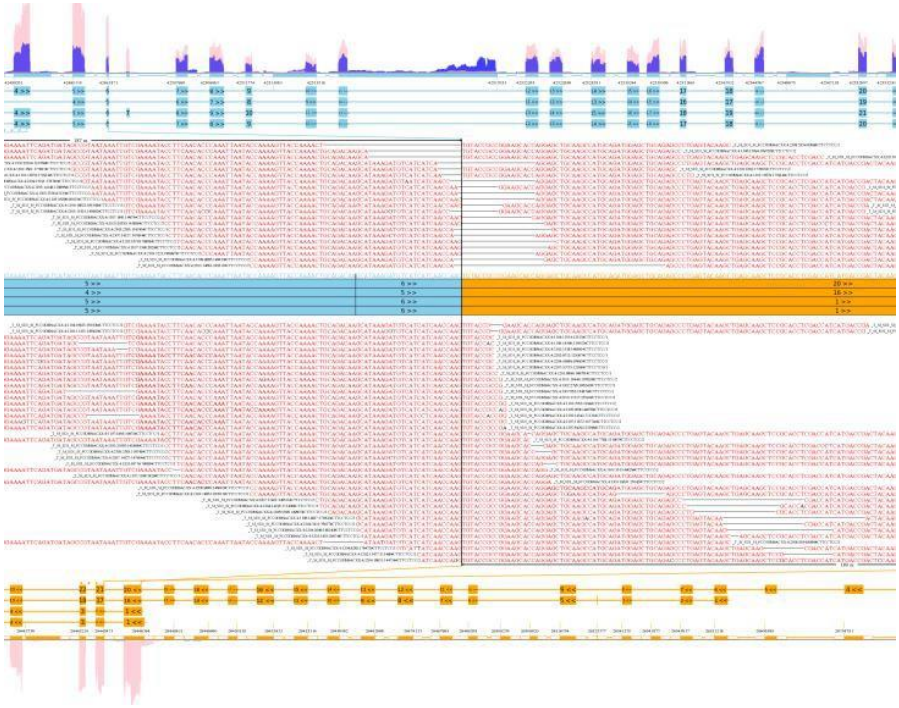
得到所提供样本中融合基因的详细信息，以及基因融合对后续表达的影响，了解基因融合在肿瘤发生、发展中的重要作用。

结果展示-融合基因列表

表 1 融合基因列表

No. of samples sharing this fusion	5' gene	chr	strand	break point	edge or internal	3' gene	chr	strand	break point	edge or internal	largest No. of read pairs (No.>=1)	largest No. of fusion-spanning reads(No.>=1)	Sample1	……	Sample x	……	Sample Y
same chromosome																	
1	Gene1	chr1	-	49021911	edge	Gene2	chr1	-	49123891	edge	1	2	√				
different chromosome																	
2	Gene3	chr2	-	47901781	edge	Gene4	chr12	+	51891342	edge	10	10			√		

➢ **结果展示-融合基因图示**



5 融合基因展示图

2.7 实验验证

采用 FISH、Sanger 测序或者 RT-PCR 的方法进行验证，用基因敲除、质谱、Western blot 等技术研究融合基因表达对细胞的影响，研究致癌机理。

2.8 项目周期

样品检测合格后，项目周期约 40 个工作日，可完成全部数据交付。

三、 应用案例

案例 1 基因融合导致骨上皮样血管瘤 *FOS* 基因被截短[8]

骨上皮样血管瘤是一种局部侵袭性的血管肿瘤。通过转录组检测到骨上皮样血管瘤 *FOS* 基因的三种融合现象，*FOS-MBNL1*，*FOS*-lincRNA(RP11-326N17.1)和 *FOS-VIM*。

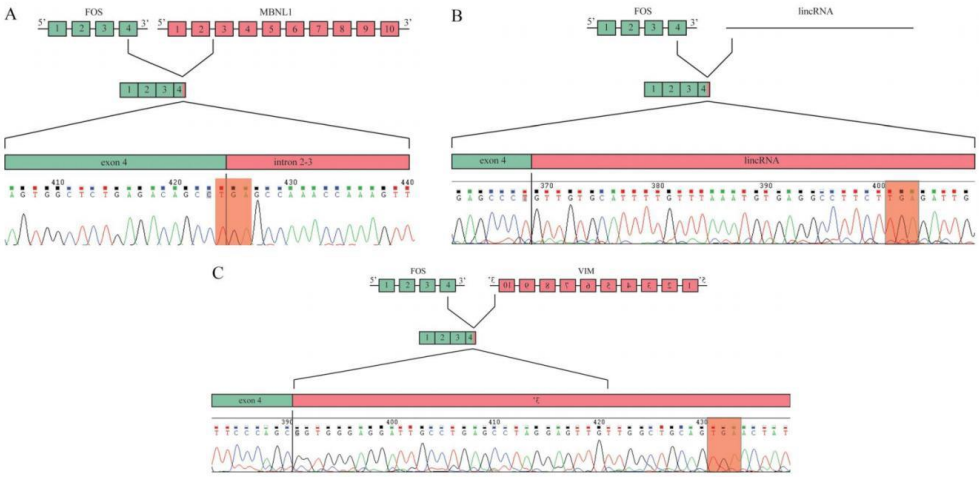


图 6 *FOS* 的三种融合基因

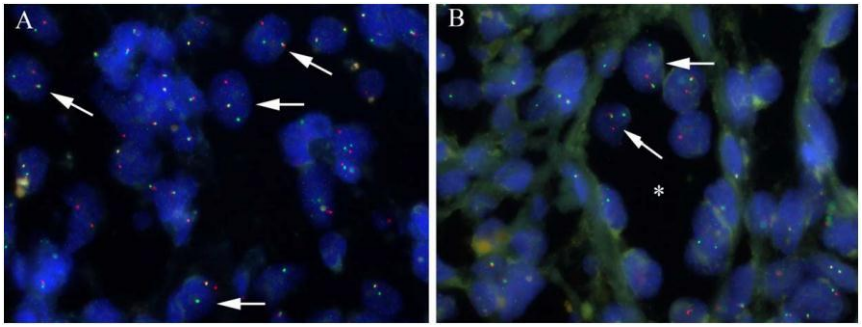


图 7 *FOS* 融合基因的验证

上述三种基团融合中，FOS 的融合断裂位点都在第 4 个外显子。在指示病例中，FOS 第 4 号外显子的 c.858 融合到 MBNL1 的第 2 内含子，出现一个终止密码，明显缩短了编码区，导致 FOS 的 C 端减少了 95 个氨基酸。FOS 第 4 号外显子的 c.828 融合到 VIM 下游的反向互补链，导致 FOS 蛋白 C 端多出了 13 个氨基酸。

案例 2　膀胱癌转录组 RNA-Seq 测序发现新的融合基因转录本[9]

泌尿上皮细胞癌（urothelial carcinoma, UC）是最常见的膀胱癌，也是检出率很高的泌尿生殖系统肿瘤， UC 的基因融合现象对于膀胱癌的诊断有重要价值。本研究利用 Illumina HiSeq 平台，对浅表性（superficial）、侵袭性（invasive）和侵袭转移性（invasive metastatic）膀胱癌三个样本进行了转录组分 析，预测的融合基因中，除去已知的，还发现了四个新的融合基因。经验证，SEPT9/CYHR, IGF1R/TTC23 和 CASZ1/DFFA 融合基因转录本在所选取的 48 个病例样本中未检测到，SYT8/TNNI2 在病例样本中检测率为 37.5% (18/48)，在正常组织中检测率为 22.7% (5/22)。

表 2 本研究中发现的融合转录本特征

Characteristic of identified fusion transcripts.

5' Gene	3' Gene	Number of spanning reads	5'gene junction	3'gene junction	Predicted_effect	Fusion description	Tumor type
PPP6R3	*LRP5*	25	chr11:68460827	chr11:68406524	5'UTR exon/CDS	known read-through	Lung squamous cell carcinoma
CLTC	*VMP1*	10	chr17:59679519	chr17:59817712	inframe	known intra-chromosomal	Breast cancer
HEPHL1	*PANX1*	5	chr11:94067743	chr11:94129328	inframe	known read-through	Head and neck squamous cell carcinoma
GOLT1A	*KISS1*	7	chr1:204213882	chr1:204192914	CDS/UTR	known read-through	-
GPHN	*MPP5*	4	chr14:67023675	chr14:67292511	inframe	known intra-chromosomal	Bladder cancer, breast cancer, lung adenocarcinoma
VPS45	*PLEKHO1*	6	chr1:150110627	chr1:150150912	out of frame	known read-through	Lung cancer, prostate cancer
IGF1R	*TTC23*	4	chr15:98708107	chr15:99161867	inframe	novel intra-chromosomal	revealed *de novo*
SYT8	*TNNI2*	6	chr11:1836861	chr11:1839675	out of frame	novel read-through	revealed *de novo*
CASZ1	*DFFA*	3	chr1:10796564	chr1:10469338	5'UTR exon/CDS	novel intra-chromosomal	revealed *de novo*
SEPT9	*CYHR*	2	chr17:75303278	chr8:145678842	5'UTR exon/CDS	novel inter-chromosomal	revealed *de novo*

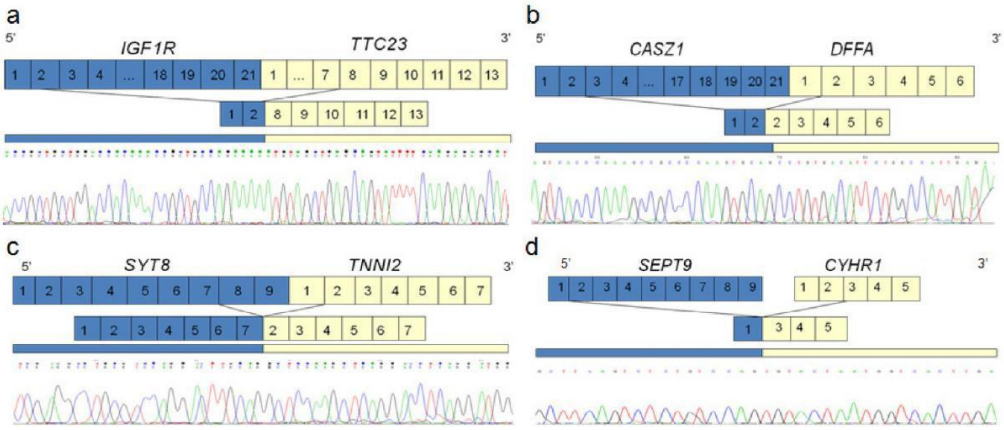


图 8 四个新融合基因的结构

案例 3 转录组测序发现新的前列腺癌融合基因[10]

对 12 例良性前列腺良性增生和 41 例前列腺癌的冷冻切片进行低深度全基因组测序和全转录组测序。经分析，在 41 例前列腺癌中检测出 32 例包括 ETS-转录因子或 SPINK1 参与的基因融合现象。在一例 CRPC 样本中发现一个新的 TMPRSS2-SKIL 融合基因。对另外 76 例前列腺癌和 22 例 LuCaP 移植瘤进行融合基因验证，发现有两例 SKIL 过表达，对这两例进行转录组分析，发现移植瘤 LuCaP-77 中的融合基因是 SLC45A3-SKIL，而临床样品 PC-11423 中的融合基因是 MIPEP-SKIL。

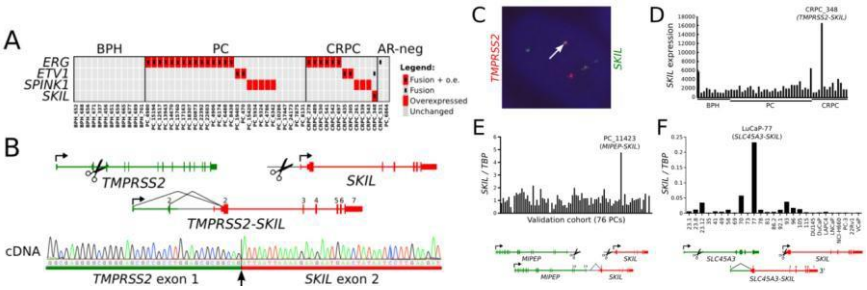


图 9 融合基因的检测及验证

A-D: *TMPRSS2-SKIL* 融合基因的结构及验证；E-F: *SLC45A3-SKIL* 和 *MIPEP-SKIL* 的结构。

从 TCGA 提供的 423 例转录组数据中还发现有额外的 4 例 *SKIL* 参与的基因融合，都表现出 *SKIL* 的过表达。

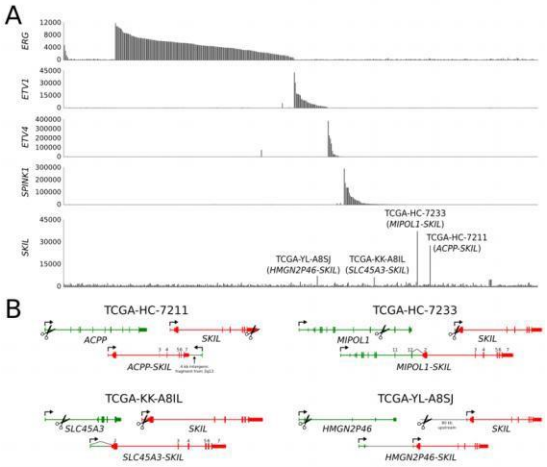


图 10 四例 SKIL 参与的基因融合

四、 常见问题

问题 1 融合基因检测一般推荐多大的数据量？

答：一般推荐 8 G 的数据量，数据量达到 9-10 G 时基本能够检测出所有的基因融合，但是考虑后续的实验，一般推荐 8 G 数据量。

问题 2 样本量应该用多少？

答：常规样本量一般 200ng 就可以，如果样本比较珍贵，可以做到 ng 级及 pg 级。这个时候采用的是其他的试剂盒，价格也会比常规转录组的价格更贵一些。

五、 华大优势

- **丰富的项目经验**。专业的信息分析团队，已完成 100,000 例转录组测序，发表多篇高水平转录组文章。

- **全面的样品类型**。提供单细胞、FFPE、微量样品等特殊样品服务，微量样品建库低至 pg 级，完全解决样品准备的后顾之忧。
- **更精准的分析结果**。独创的插入片段文库，转录本组装长度更长，可变剪接、基因融合鉴定更灵敏准确。
- **更丰富的信息分析内容**。独有信息分析内容（主成分分析、样本相关性分析、RNA 编辑分析），并且可以提供多组学关联分析，有效信息更加全面。
- **更清晰的结题报告。**结果图片更加直观，可用于文章撰写；结题报告结构优化，有效提升阅读体验。

六、 参考文献

[1] Tkachuk D C, Westbrook C A, Andreeff M, *et al*. Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization[J]. Science, 1990, 250(4980): 559-562.

[2] Tomlins S A, Rhodes D R, Perner S, *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer[J]. Science, 2005, 310(5748): 644-648.

[3] Tomlins S A, Laxman B, Dhanasekaran S M, *et al*. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer[J]. Nature, 2007, 448(7153): 595-599.

[4] Soda M, Choi Y L, Enomoto M, *et al*. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer[J]. Nature, 2007, 448(7153): 561-566.

[5] Bass A J, Lawrence M S, Brace L E, *et al*. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion[J]. Nature genetics, 2011, 43(10): 964-968.

[6] Jia W, Qiu K, He M, *et al*. SOAPfuse: an algorithm for identifying fusion transcripts from paired- end RNA-Seq data[J]. Genome biology, 2013, 14(2): 1.

[7] Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment[J]. Bioinformatics, 2009, 25(15): 1966-1967.

[8] van IJzendoorn D G P, de Jong D, Romagosa C, *et al*. Fusion events lead to truncation of FOS in epithelioid hemangioma of bone[J]. Genes, Chromosomes and Cancer, 2015, 54(9): 565-574.

[9] Kekeeva T, Tanas A, Kanygina A, *et al*. Novel fusion transcripts in bladder cancer identified by RNA-seq[J]. Cancer letters, 2016, 374(2): 224-228.

[10] Annala M, Kivinummi K, Tuominen J, *et al*. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer[J]. Oncotarget, 2015, 6(8): 6235.

肿瘤研究之非编码 RNA 生物标记物研究方案

一、 研究背景

迄今为止，对于晚期的癌症尚无有效的治疗措施，癌症死亡率仍然居高不下。因此，寻找新的生物标记物用于癌症的早期诊断和预后等更显得十分迫切。癌症生物标记物是癌症直接产生或由身体对癌症特定响应分泌的物质，这些物质或过程的存在表明癌症存在于人体。目前研究癌症生物靶标的方法有多种，包括生理学、影像学、特异性分子、遗传变异、基因和蛋白表达、以细胞为基础的标志物及其他的形式等。长期以来，非编码 RNA 一致被认为是不参与编码蛋白的“垃圾信息”，但是越来越多的研究表明，非编码 RNA 与肿瘤的形成、发展关系密切，并且在体内稳定存在，因此非编码 RNA 可作为肿瘤研究的一项重要诊断、预后的生物标志物。目前已在乳腺癌、肺癌、前列腺癌、结直肠癌、卵巢癌、黑色素瘤和白血病等癌症研究中发现用于诊断和预后的非编码 RNA 生物标记物，这将有助于在临床上尽早发现和预后干预治疗等。

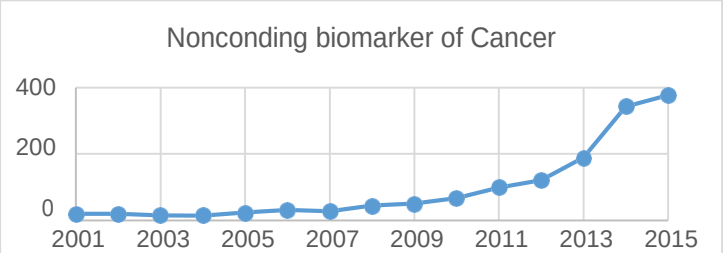


图 1 非编码 RNA biomarker 文献历年发表篇数统计（NCBI）

关于癌症 biomarker 研究发表文献也逐年上升，在 NCBI 上收录的关于非编码 RNA 生物标记关键词的文献数 2015 年高达 377 篇（图 1）

2016 年 2 月，研究发现一个用于喉癌诊断环状 RNA（hsa_circRNA_100855 l）biomarker[1]。2016 年 3 月，研究人员发现 MicroRNA-143 可作为胰腺癌治疗的潜在 biomarker[2]。2016 年 4 月，研究人员鉴定 lncRNA-HOTAIR 作为卵巢癌生物标记，用于早期诊断[3]。

高通量测序研究技术方法，可在相对短的时间内可以得到较多的 RNA 表达有效数据，所以利用非编码 RNA 高通量测序技术手段，将有助于加快癌症发生、肿瘤发育、以及治愈复发等过程中分子生物学机制的相关研究进度，可作为科研的高效研究方式选择。

二、 方案设计

2.1 整体思路

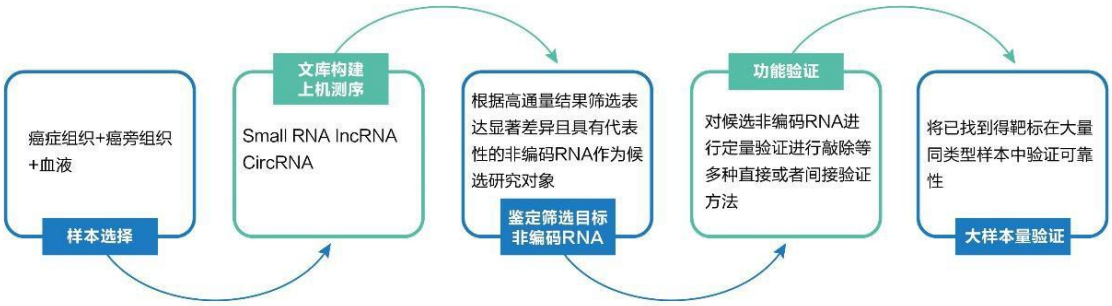


图 2 整体研究思路

2.2 样本建议

癌症组织+癌旁组织，血液样本(Case + Control)，可选择淋巴细胞、血清/血样样本。

注意：选择样品需要考虑不同时期以及癌组织的异质性；可选择治疗前活检样本，手术切除样本，复发或转移样本。根据文献经验一般推荐 10 对及以上。

2.3 实验技术

(1) Small RNA：切胶回收 18-30nt 片段范围，进行文库构建和 SE 测序，推荐 10M clean Reads。

(2) LncRNA：通过去核糖体 RNA 方法，构建文库和 PE 测序，推荐 10G clean data。

(3) CircRNA：通过去除核糖体 RNA 及线性 RNA 后，构建文库和 PE 测序，推荐 10G clean data。

2.4 分析内容

鉴定筛选目标非编码 RNA：

a) 已知非编码 RNA 鉴定

与已知数据库信息比对，鉴定已知表达的 lncRNA、miRNA 信息，确定其在基因组位置信息。

b) 未知非编码 RNA 预测

针对数据库中未知信息，根据结构、基因组位置、软件打分情况，对满足条件的 novel RNA 进行预测鉴定。

c) 非编码 RNA 表达及差异分析

与正常组织进行比对后，查找表达显著差异的非编码 RNA 作为候选 RNA，将表达差异比较结果排名前 10 或者前 15 的非编码 RNA 进行目的性分析。

d) 非编码 RNA 调控机制研究

将非编码 RNA 与基因进行靶向分析，确定其调控作用机制。

2.5 实验验证

(1) 功能验证

a) 定量验证：qPCR/颈环 qPCR 定量验证。

b) 体外验证：miRNA 敲除/miRNA 拮抗物/target protecter technology、荧光素酶标记等。

c) 体内验证：构建小鼠模型。

(2) 大样本量验证

a) 利用 qPCR 结果及表型结合验证。

b) 样本选择量在 100 及以上（实验应当包含足够数量的样品，才能充分代表每个肿瘤亚型，以建立统计学可信度。多名患者肿瘤中共有非编码 RNA 的存在足以创建此假设，即非编码 RNA 可能在疾病中发挥作用。通常需要大量患者来检验此假设，并建立统计学可信度得以验证）。

三、 应用案例

案例 1 华大参与——肝癌诊断的 miRNA biomarker[4]

利用高通量测序技术深度挖掘正常肝组织、肝炎患者肝组织和肝癌组织小 RNA 表达信息，发现在人肝脏中 9 个重要的小 RNA， 占总小 RNA 量的 88.2%， 其中表达量第三高的 miR-199a/b-3p，在肝癌组织中均出现表达下调，并且下调的程度与病人生存率下降显著相关。在体内、体外进行验证，表明 miR-199a/b-3p 靶向作用 PAK4（癌症促进因子）,通过抑制 PAK4/Raf/MEK/ERK 通路，从而抑制肝癌组织的生长。此项研究得到了正常肝脏和肝癌组织的小 RNA 谱，可用于肝癌和其他肝脏疾病的 biomarker。

主要结果展示

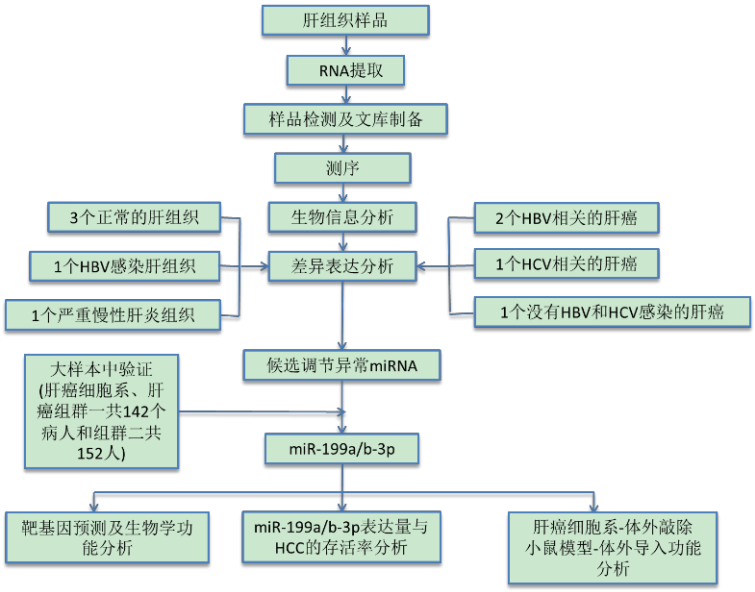


图 3 文章研究思路

1、 miRNA 表达统计分析

发现在正常肝组织中 85.9%的 miRNAs 是低表达的（<10TPM），13.2%的 miRNAs 适量表达（10~10,000TPM），只有 0.9%的 miRNAs 是高表达的（>10,000TPM）。这 0.9%的 miRNAs 的表达量占全部 miRNAs 的 88.2%。前 3 位表达量最高的是 miR-122，miR-192 和 miR-199a/b-3p，其表达量分别为 52.0%、16.9%和 4.9%（表 1）

表 1 正常组织和肝癌组织中 miRNA 高表达统计表

miRNA	Average in Normal Liver	HBV(+) Adjacent Sample 1	HBV(+) HCC Sample 1	HBV(+) Adjacent Sample 2	HBV(+) HCC Sample 2	HCV(+) Adjacent	HCV(+) HCC	HBV(-) HCV(-) Adjacent Tissue	HBV(-) HCV(-) HCC
miR-122[a]	520,204	497,436	581,733	478,691	516,360	449,921	511,457	597,746	17,280
miR-192[a]	169,186	184,230	182,858	153,346	205,819	111,731	166,201	111,530	164,532
199a/b-3p[b]	48,885	79,099	2588	58,199	2211	44,000	2831	39,779	2979
miR-101[a]	36,671	37,538	35,730	54,202	44,686	26,361	38,555	57,519	39,321
let-7a [b]	32,760	26,392	5515	39,339	40,747	36,908	28,419	26,423	110,611
miR-99a[b]	21,923	22,139	6019	27,594	10,326	20,701	6107	18,287	984
let-7c[b]	21,358	20,724	2930	19,340	15,231	18,996	17,059	19,840	1902
let-7b[b]	16,718	20,556	1531	11,819	2951	20,767	8864	12,944	19,074
let-7f [a]	14,757	12,778	10,047	14,195	12,139	14,230	14,552	19,600	133,023
let-7d[b]	8792	5326	1264	8605	6390	9110	6471	6256	56,226
miR-100[b]	3606	4166	81	5959	7863	4763	7708	2742	998
125b-5p[b]	2665	2043	234	2634	992	1542	1391	1368	76
miR-143[b]	2269	2254	750	2635	996	6167	763	1278	4573
miR-29c[b]	2009	2067	670	2189	1062	3012	2301	3059	2266
miR-21[c]	5465	9561	16,241	10,099	12,314	9133	9356	10,948	82,662
miR-146b[c]	1614	698	7817	602	171	2737	643	597	555
miR-92a[c]	955	4775	16,142	874	420	2009	1650	1370	4269

Expressions of the abundant miRNAs seem.
[a] unaltered.
[b] decreased.
[c] increased in HCC. See also Tables S2 and S3.

2、体外功能验证

肝癌细胞系 miR-199a /b-3p 下调表达（图 4A）；miR-199a/b-3p 恢复表达后，抑制细胞的生长，诱导细胞凋亡（图 4C）；阻断细胞周期，使其停留在 G1 期（图 4E）。

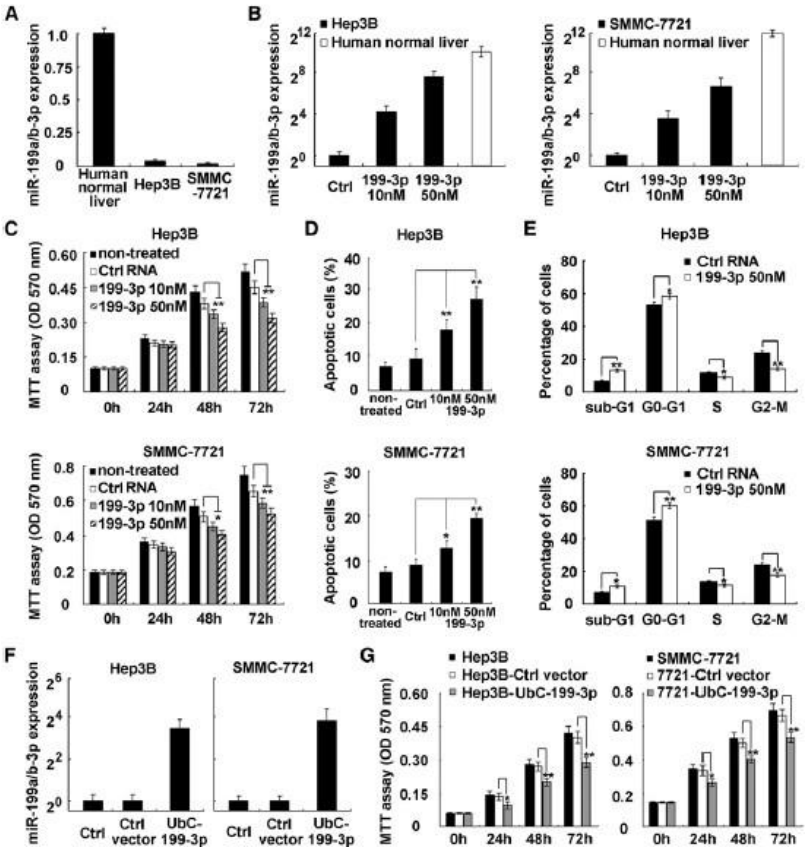


图 4 miR-199a/b-3p 在体外实验中表现可阻止肝癌细胞生长

3、体内验证

小鼠移植瘤实验——移植瘤 miRNA 表达与肝癌中 miRNA 表达相似（图 5A/B/C）；miR-199a/b-3p 模拟物注射导致 miR-199a/b-3p 表达量升高，同时能够抑制肿瘤生长，降低血清中的 AFP 含量，并引起肿瘤坏死。构建了 AAV-8 载体系统验证，结果一致（图 5D/E/G/H）

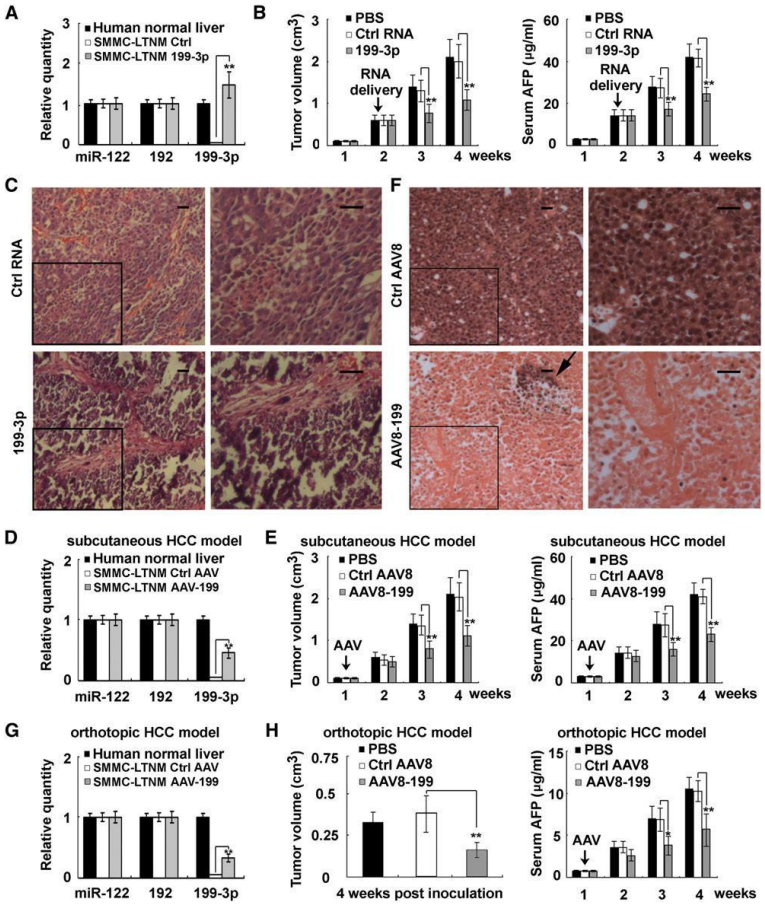


图 5 miR-199a/b-3p 在体内验证表现可阻止肝癌细胞生长

案例 2 前列腺癌 miRNA 靶标筛查及功能验证[5-6]

利用高通量测序方法，对 546 个经过手术切除的复发的前列腺癌患者跟踪研究，对其中 31 个患者进行 miRNA 高通量测序，鉴定了与前列腺癌复发相关的 5 个 miRNA，其中 miR-310a 与前列腺癌复发显著相关。进一步研究：利用 qPCR 方法验证这 5 个小 RNA 在 515 个前列腺癌患者的术后样品中是否表达，以便作为预测术后转移的 biomarker；并对 miR-301a 进行体内和体外过表达验证，发现其在小鼠模型中影响肿瘤生长和转移。

进一步深入研究：在 585 个前列腺癌根治术病人中进一步筛查 miR-301a 的表达情况，跟踪前列腺癌复发和转移。主要采用 miRNA 高通量测序、qPCR 定量和 Western 印迹方法，发现：miRNA-301a 可通过靶向 p53 基因影响癌细胞的间充质细胞转化生物学过程，荧光素酶标记也进一步证明了 p53 作为 miR-301a 的直接靶向基因，p53 缺失导致 miR-205 表达下调，并释放 Zeb1 和 Zeb2 抑制钙黏蛋白；已通过在前列腺癌细胞系

和病人肿瘤组织中证明，miR-301a 表达上调可调控此代谢通路。

结论：miRA-301a 可作为前列腺癌预后的新生物标记。

方案设计

主要结果展示

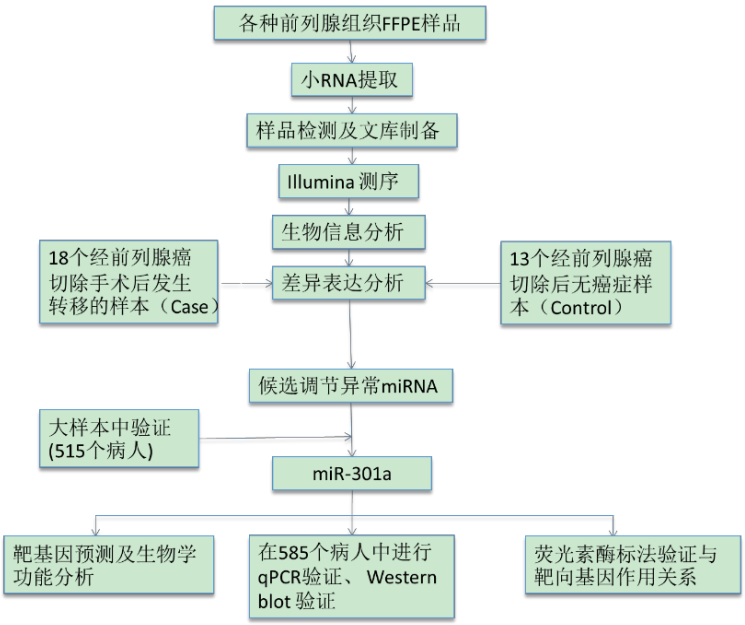


图 6 文章研究思路

1. 前列腺根除手术后复发病人和未复发病人 miRNA 表达比较，其中 miR-301a 和 miR -652 呈现 2 倍以上显著差异表达。

表 2 前列腺根除手术后复发病人(PR)和未复发病人(PNMNR)miRNA 表达差异表

Table 2: List of 33 microRNAs identified by whole miRNome analysis to be associated with prostate cancer recurrence.

Gene Identifier	Mean Expression Level*		Ratio	Direction	p-value	adj. p-value†
	PNMNR	PR				
hsa-mir-671	6.94	7.75	1.76	Up	3.75E-06	0.00072
hsa-mir-301a	7.55	8.78	2.34	Up	2.64E-05	0.00124
hsa-mir-652	6.73	8.30	2.97	Up	3.69E-05	0.00124
hsa-mir-223	11.90	10.66	2.36	Down	7.46E-05	0.00179
hsa-mir-545	4.04	5.10	2.08	Up	0.00049	0.00524
hsa-mir-501	4.91	5.74	1.77	Up	0.00052	0.00525
hsa-mir-1249	4.47	5.37	1.86	Up	0.00079	0.00622
hsa-mir-454	7.26	8.22	1.94	Up	0.00148	0.00981
hsa-mir-154	4.26	5.28	2.03	Up	0.00159	0.01016
hsa-mir-497	10.28	11.19	1.89	Up	0.00179	0.01111
hsa-mir-708	5.45	6.67	2.32	Up	0.00214	0.01206
hsa-mir-452	5.15	6.02	1.84	Up	0.00303	0.01385
hsa-mir-1302-3	2.75	4.42	3.18	Up	0.00354	0.01529
hsa-mir-1302-1	2.75	4.53	3.45	Up	0.00348	0.01529

2. 体内验证：在大样本量分析中，对 515 个前列腺癌术后病人进行回归系数分析发现：miR-301a 与前列腺癌复发显著相关。
3. 体外验证：在前列腺癌细胞系中发现，miRNA-301a 与癌细胞生长显著相关；在构建小鼠前列腺癌模型中发现，miRNA-301A 与癌细胞转移显著相关。

案例 3 长链非编码 RNA-LUNAR1 成为血癌治疗靶标和鉴定 biomarker[7]

利用高通量测序技术鉴别出了 6,023 种长链非编码 RNA，这些 RNAs 只在 15 名罹患急性T 淋巴细胞白血病的患者的免疫系统 T 细胞中表达，而在 3 名健康人的 T 细胞中则没有表达。进一步分析发现，采用化学方法阻断其中一种非编码 RNA：LUNAR1 的作用，可以阻止白血病的进展。研究证实，LUNAR1 高度特异地存在于急性 T 淋巴细胞白血病中，在这类癌症的形成中起至关重要的作用。测试的几乎所有（90%）的白血病患者均过度表达 LUNAR1。

结论：LUNAR1 可以作为诊断这种血癌的新的 biomarker。

主要结果

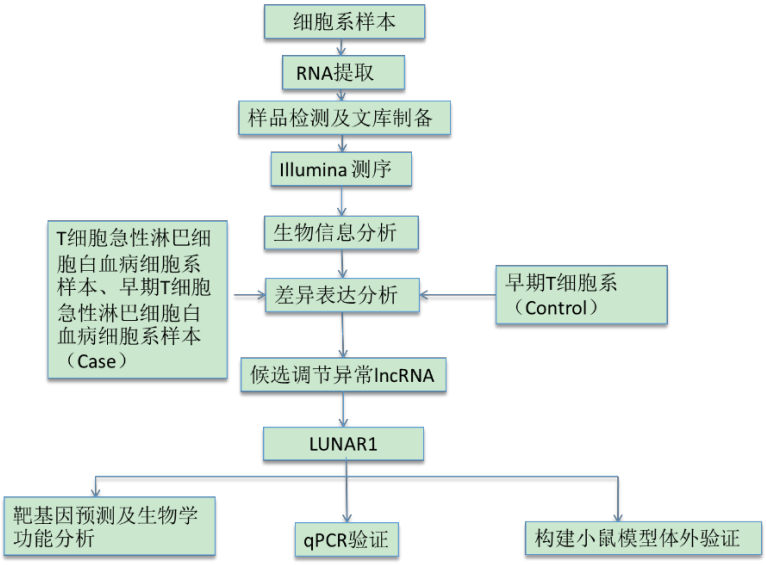


图 7 文章研究思路

1 通过高通量测序，得到 6023 个 lncRNA 转录本，发现已知 lncRNA 占 46%，未知 lncRNA 占 54%（图 8B），且 lncRNA 表达量低于mRNA 表达量（图 8D）； T-ALL 细胞系与其他组织相比，其 lncRNA 表达具有组织特异性（图 8E）。

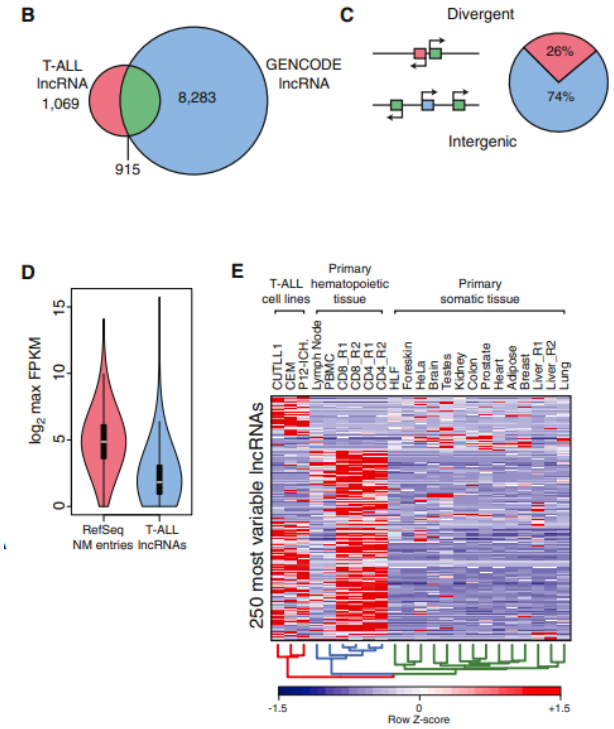


图 8 T-ALL 细胞系与其他组织中 lncRNA 表达谱对比

2. 通过筛选 lncRNA 与邻近编码基因 cis 调控关系，对相关性高于 0.75 的 lncRNA 进行后续分析发现，*LUNAR1* 与其编码 RNA 邻近基因（*IGF1R*）呈现很高的相关性，且在 Notch 代谢途径中呈现负调控（图 9C），在血癌细胞系中呈现高表达（图 9D），另外发现该 lncRNA 在细胞核中富集（图 9E）。

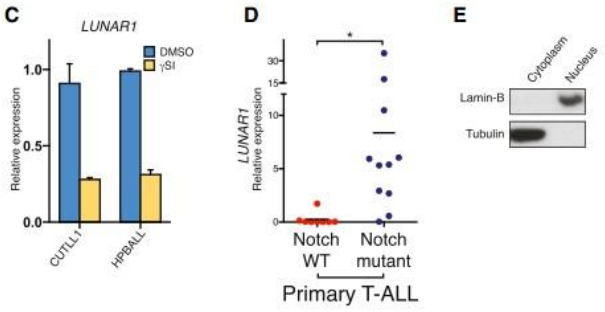


图 9 新的长链非编码 RNA-LNUAR1 功能验证图

3. 体外验证：通过 RNAi 干扰技术研究 lncRNA-LUNAR1 和 mRNA-IGF1R 调控关系，发现敲除 lncRNA

后，导致 IGF1R 低表达。

4. 体内验证：将白血病患者的 T 细胞移植到小鼠体内，并在动物中采用化学方法阻断了 LUNAR1 表达，结果显示：只有 LUNAR1 失活的小鼠体内的肿瘤停止生长（图 10C/D）

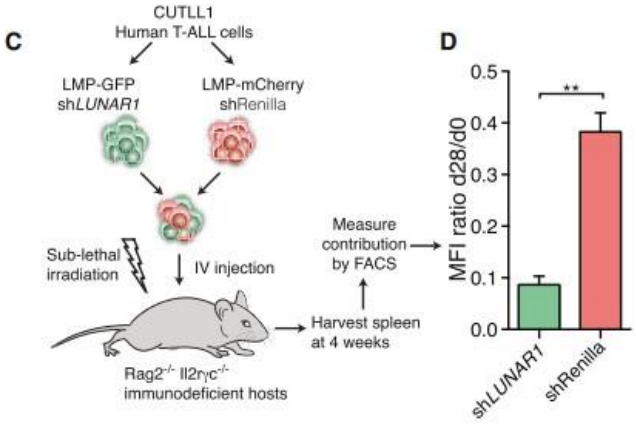


图 10 小鼠移植瘤验证长链非编码 RNA LUNAR1 的功能

四、 项目风险

- 在样品选择的过程中，如若未注意癌症样品不同时期的特征，有可能会找不到表达差异显著的非编码 RNA 用于 biomarker 研究。
- 在样本准备过程中，因样品切割时间和温度等细节问题影响到 RNA 质量问题，在分析结果中也会因 RNA 降解导致假阳性结果。
- 样本数建议选取 10 对及以上，分析结果与表型相关分析部分才具有统计学意义。
- 验证过程需要选择合适的验证方法，如对低表达的非编码 RNA 进行验证，可能会出现验证率低的风险。

五、 常见问题

问题 1 样品选择一定要 10 对以上吗？

答：尽量准备 10 对以上，如没有，建议 5 对以上，后续验证样本建议 100 例以上。

实验应当包含足够数量的样品，以建立统计学可信度。另外，需要注意异质性是癌症样品共有的问题，不同的病因和机理往往共有一个特定的癌症表型。因此，严格的分析需要有足够的样品量，才能充分代表每个肿瘤亚型。

问题 2　血液样本如何取样？

答：哺乳动物白细胞的分离方法：

a)在已加入抗凝剂的新鲜全血中加入等体积 PBS（1×），充分混匀；

b)缓慢转入另一已加入淋巴细胞分离液的离心管中，并使上述混合液处于淋巴细胞分离液液面之上（即两种液体不要混合，保留清晰的界面），3000 g 离心 30 min；

c)用移液器小心分离出白细胞层；用 PBS（1×）清洗白细胞，离心回收白细胞，弃去上清；

d)加入白细胞 20 倍体积的 TRIzol 试剂，用一次性注射器进行反复抽打细胞直至看不见成团的细胞团块，整个溶液呈清亮而不粘稠的状态；

e)放入干冰或于-80℃冰箱中保存。

问题 3　验证方法要多种吗？

答：建议尽可能选取 1 种及以上进行验证，文章发表相对容易。对于目标非编码 RNA 的筛选，可选择表达量排名 TOP10 进行验证。

六、　华大优势

- **丰富的项目经验。**已完成万份以上的样品测序分析工作，承接国内外中小型项目。
- **精湛的实验工艺。**可接受多种类型样本，包括 FFPE 样本。样本起始量低至 100ng/单次，可满足珍稀样本的研究需求。
- **严格的质量管理。**实验全程采用全方位质控体系，应用质量管理体系金标准。
- **贴心的售后支持。**7*24 小时全天候技术支持，第一时间响应客户需求。

七、　参考文献

[1] Xuan L, Qu L, Zhou H, *et al*. Circular RNA: a novel biomarker for progressive laryngeal cancer[J]. American journal of translational research, 2016, 8(2): 932.

[2] Li W, Wu H, Li Y, *et al*. MicroRNA-143 promotes apoptosis of osteosarcoma cells by caspase-3 activation via targeting Bcl-2[J]. Biomedicine & Pharmacotherapy, 2016, 80: 8-15.

[3] Wu H, Shang X, Shi Y, *et al*. Genetic variants of lncRNA HOTAIR and risk of epithelial ovarian cancer among Chinese women[J]. Oncotarget, 2016.

[4] Hou J, Lin L, Zhou W, *et al*. Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma[J]. Cancer cell, 2011,

19(2): 232-243.

[5] Nam R K, Amemiya Y, Benatar T, *et al*. Identification and validation of a five microRNA signature predictive of prostate cancer recurrence and metastasis: a cohort study[J]. Journal of Cancer, 2015, 6(11): 1160.

[6] Nam RK, Benatar T, Wallis CJ, *et al*. MiR-301a regulates E-cadherin expression and is predictive of Nam R K, Benatar T, Wallis C J D, *et al*. MiR- 301a regulates E- cadherin expression and is predictive of prostate cancer recurrence[J]. The Prostate, 2016, 76(10): 869-884.

[7] Trimarchi T, Bilal E, Ntziachristos P, *et al*. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia[J]. Cell, 2014, 158(3): 593-606.

关注华大科技，尽享精彩科研！

华大基因（深圳总部）

地址：深圳市盐田区北山工业区综合楼（518083）
电话：400-706-6615
邮箱：info@bgi.com
网站：www.bgi.com

美洲华大基因（波士顿）

地址：One Broadway,14th Floor,Cambridge,MA 02142,USA
电话：+1-617-500-2741
邮箱：info@bgi.com
网站：www.bgi.com

欧洲华大基因（哥本哈根）

地址：Ole Maaløes Vej 3,DK-2200 Copenhagen N,Denmark
电话：+45-7026-0806
邮箱：info@bgi.com
网站：www.bgi.com

日本华大基因（神户）

地址：Kobe KIMEC Center BLDG.8F,1-5-2 Minatojima-min-amimachi,Chuo-ku,Kobe City,Hyogo-pref.650-0047,JAPAN
电话：+81-785-996-108
邮箱：info@bgi.com
网站：www.bgi.com

亚太、港澳台华大基因（中国香港）

地址：香港新界大埔工业村大富街16号
电话：+852-3610-3510
邮箱：info@bgi.com
网站：www.bgi.com

出品：深圳华大基因股份有限公司科技服务产品部、市场部。

本手册仅供客户学习、交流和研究使用，请勿用于商业用途，违者必究。
任何其他个人或组织均不得以任何形式将本手册中的各项内容进行复制、 编辑或翻译为其他语言。