华大基因
BGI

Cases Study

科技服务方案手册

医口篇

Human Research

基因科技造福人类

Omics for all

华大科技

电话：400-706-6615

邮箱：info@genomics.cn

网址：www.bgitechsolutions.com

地址：深圳市盐田区洪安三街21号（518083）

深圳华大基因股份有限公司

此书献给

广大前沿的科研工作者

希望您能从本书中寻找到

新的科研思路

CONTENTS
目 录

引言 人类遗传学研究

001 肿瘤免疫治疗多组学研究方案

011 肿瘤异质性和进化研究

020 肿瘤免疫组库研究方案

035 ctDNA肿瘤液体活检研究方案

045 外泌体非编码RNA疾病生物标记物研究方案

02

054 肿瘤发生的关键转录因子及其靶基因多组学研究方案

060 自身免疫性疾病免疫组库研究方案

074 感染类疾病免疫组库研究方案

086 复杂疾病*De novo*突变研究方案

097 疾病相关的肠道菌群多组学研究方案

108 疾病诊断及药效评估的蛋白标志物研究方案

03

117 药物作用机制的基因表达研究方案

132 基于高通量测序的种群特征研究方案

140 孟德尔遗传病基因组测序研究方案

147 肿瘤融合基因研究方案

155 病毒感染相关的转录组与蛋白质组关联分析研究方案

166 高通量单细胞疾病类研究方案

引言

人类遗传学研究

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的限制因素，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，人们得以清楚地了解遗传信息的构成和传递的途径，特别是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的、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。

肿瘤免疫治疗多组学研究方案

001

研究背景

免疫治疗是近年来肿瘤研究的热点之一，肿瘤免疫治疗在《科学》杂志2013年十大科学突破中位居首位。肿瘤免疫治疗主要分为两种：细胞免疫治疗和免疫检查点抑制剂治疗。细胞免疫疗法是指：提取患者的免疫细胞在体外进行改造或者诱导，使这些细胞具备对癌细胞更有效、更精准的免疫能力，改造和诱导后的免疫细胞回输到患者体内后，它们会定向消灭癌细胞，例如CAR-T治疗。免疫检查点抑制剂治疗是指：免疫细胞会产生抑制免疫的蛋白小分子，防止人体产生过多的免疫作用。肿瘤细胞就是利用这种机制，抑制免疫细胞，从人体免疫系统中逃脱存活下来。使用免疫检查点抑制剂类药物，可解除这种抑制作用，让免疫细胞重新激活工作，消灭癌细胞，例如PD-1和CTLA-4。

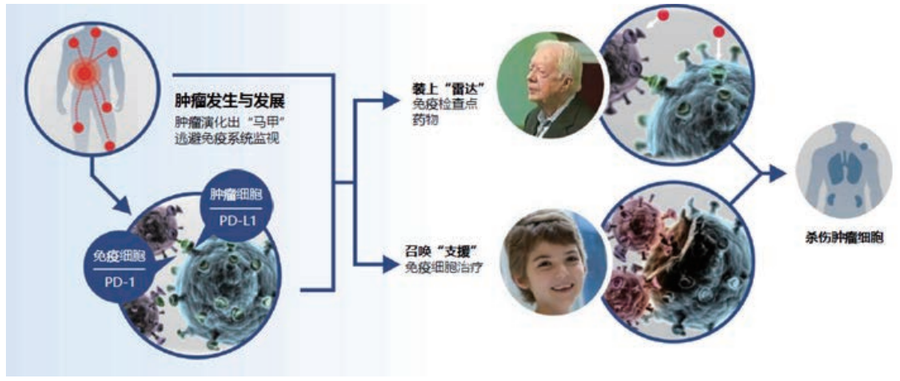


图1 免疫治疗分类

FDA在2017年5月批准了帕姆单抗用于治疗成人及儿童不可切除或转移的、MSI-H或错配修复缺失的实体肿瘤，且适用于在前期治疗后疾病进展、更换治疗方案后不满意的患者。和之前的药物有所区别，批准的这个药物是基于肿瘤基因组上的特征，而非我们之前知道的不同肿瘤用不同药。这个遗传特征就是MSI-H/dMMR，也就是微卫星不稳定程度高和错配修复基因缺陷。肿瘤中只要有MSI-H和dMMR特征，就可以用这个药。帕姆单抗的获批是基于5项多中心、单臂的临床试验数据的分析，这些研究纳入的149名实体瘤患者均为MSI-H或dMMR状态。结果显示，帕母单抗的客观缓解率为39.6%，其中11例完全缓解，48例部分缓解。结直肠癌患者OR为36%，其他肿瘤患者OR为46%。

同年，FDA批准了一项PD-1抑制剂帕姆单抗联合化疗用于晚期非鳞非小细胞肺癌的一线治疗，此方案适用于无EGFR和

ALK基因突变的患者，使用无需考虑PD-L1表达情况。123名受试者总体缓解率达到55%，效果也非常显著。这些免疫相关的治疗方式在肿瘤方面，可以说是有着重大突破。

虽然免疫治疗效果显著，但是仍有很大一部分病人不能受益，而且后期伴随耐药复发。目前已经有一些类似TMB, MSI-H/dMMR, PD1/PD-L1表达（详见常见问题解释）等标记物可以预示哪些病人可以受益，但并不能解释所有的案例。

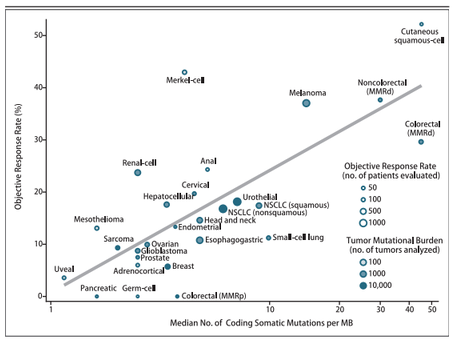


图2 27种癌症的抗PD-1/PD-L1治疗的客观缓解率与肿瘤突变负荷（TMB）的相关性[1]

研究表明，55%不同类型肿瘤的客观缓解率差异可以用TMB来解释。

另外，免疫组库作为一种新兴的技术，研究T淋巴细胞的克隆多样性，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免疫疗效相关，因此免疫组库也许可预测肿瘤免疫治疗的疗效。研究人员采用多组学分析（WES、转录组和TCR测序），对68例晚期黑色素瘤病人进行分析，发现抗PD-1治疗前后T细胞克隆发生了变化，并且发现TCR克隆特征与药物的相关性[8]。

本方案旨在通过高通量测序技术和生物信息分析手段，从肿瘤基因组突变特征、T细胞克隆特征两个角度，为肿瘤精准免疫治疗提供一整套临床科研方案。

图3 展示的是肿瘤免疫循环的过程，其中肿瘤突变负荷（TMB）及其衍生的新抗原负荷（TNB）、MSI-H（microsatellite instability-high, 微卫星不稳定）、dMMR（mismatch repair deficient, 错配修复基因缺陷）、浸润T细胞的多样性等都可以作为研究的切入点。

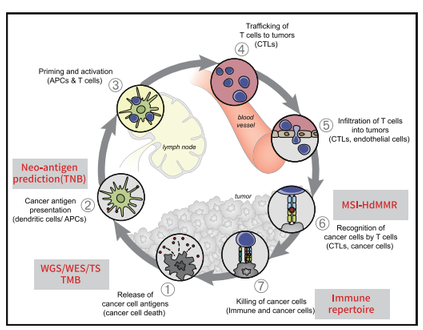


图3 肿瘤免疫循环过程[2]

理想情况下，肿瘤细胞死亡后会释放肿瘤细胞特有的抗原，称之为肿瘤抗原，这些抗原被树突细胞或抗原呈递细胞识别提呈，接下来在淋巴结中，抗原呈递细胞激活T细胞，使其活化增殖。活化的T细胞通过血液循环到达肿瘤部位，并且浸润到肿瘤组织中，识别并杀伤肿瘤细胞，死亡的肿瘤细胞再形成下一个循环。

方案设计

A. 研究目标

通过对免疫治疗不同预后的两组病人进行研究，鉴定其免疫治疗的标记物以及耐药复发的分子机制。

B. 样本类型

原发癌组织、不同阶段的外周血（包括ctDNA）、复发癌组织（视情况而定），每组样品10-20例。

C. 技术方法

全基因组测序（30x）或外显子测序（200x）、ctDNA目标区域测序（>3000x）、免疫组库测序（1G raw data）。

D. 研究内容

1. 通过对治疗效果好和效果差的两组肿瘤样品进行测序分析，鉴定包括TMB, MSI-H/dMMR等在内的分子标记物以及其他基因组范围内突变信息。

2. 通过对治疗效果好的人群不同时间段的ctDNA进行研究，鉴定肿瘤游离片段丰度、变异以及TMB信息。另外，通过免疫组库测序，监控T细胞、B细胞的克隆种类及频率。

3. 对于复发耐药人群，比较原发癌和复发癌的基因组变异信息，鉴定耐药复发的潜在分子机制。

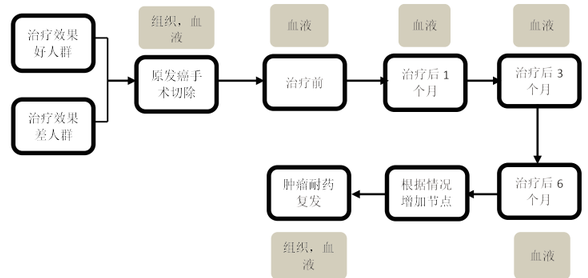


图4 方案研究思路

E. 数据分析方案

1. 肿瘤组织突变分析流程

使用CSAP(Cancer Sequencing Analysis Pipeline)流程，进行BWA比对，samtools，GATK，Varscan，Crest等软件分析得到somatic SNV，indel，SV和CNV等位点，再通过Annovar注释得到突变的详细信息。该结果将与ctDNA分析结果进行综合比较，分析ctDNA检测突变。

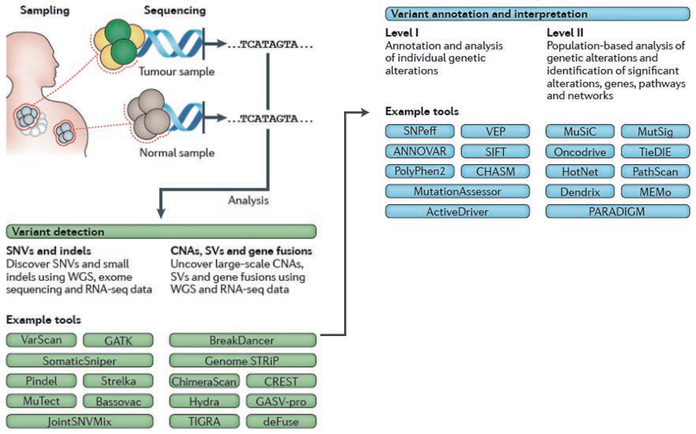


图5 肿瘤突变分析流程及软件

2. ctDNA突变检测流程

（1）UID分析去重复序列及随机错误 —— UID（unique identitor）是一段独特的序列，同一个UID家族95%以上序列含有相同的突变，则认为该突变是真实存在的。通过UID的识别能降低PCR和测序错误对突变鉴定的影响，提高检测结果的可信度。

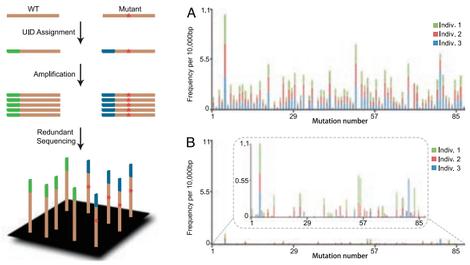


图6 ctDNA加入UID检测原理及检测灵敏度[3]

（2）ctDNA突变信号检测 —— 使用华大ctDNA变异检测流程BGI-in house pipeline，通过对不同大小的DNA片段进行分层比对，最大限度地检测到ctDNA上发生的突变信号，并通过后期的数据过滤和较正，保证检测结果的高度准确性。

- 突变检测的灵敏度高：能够检出VAF (突变等位基因频率)低至0.2%的点突变，其中VAF为1%的点突变检出率为100%；
- 突变检测的准确率高：标准品技术重复显示检出突变一致性至少70%以上。

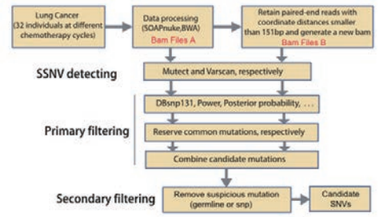


图7 ctDNA变异检测流程

通过对不同大小的DNA片段进行分层比对，最大限度地检测到ctDNA上发生的突变信号，并通过后期的数据过滤和较正，提高结果的准确性。

F. 项目执行周期

样品检测合格后，建库+测序+标准信息分析：约40个工作日，高级信息分析约1-1.5个月，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。

应用案例

案例一：《Science》新文章揭示影响肾透明细胞癌免疫检查点抑制疗法的关键因素[4]

发表期刊：《Science》

影响因子：37.205

发表日期：2018年1月

研究目的：PD-1免疫检查点抑制剂提高了部分肾透明细胞癌（ccRCC）患者的生存率，本文研究目的是找到与抗PD-1治疗疗效相关的基因组变异特征。

样品及方法：选取35个转移的ccRCC病人进行研究，根据疗效将病人分为免疫应答（CB, clinical benefit）、中间应答（intermediate benefit）和不应答（NCB, no clinical benefit）三个群体，进行外显子测序。另外选取63个转移的ccRCC病人做独立验证集。

研究结果：1、发现阶段：35个ccRCC病人的变异数据筛选后得到PBRM1基因，该基因突变为截短突变或基因功能丢失的突变，是唯一一个富集在免疫应答组（CB）的显著突变基因。结合临床数据，确定该基因的纯合丢失将会显著提升患者的总体生存率和无恶化生存率。

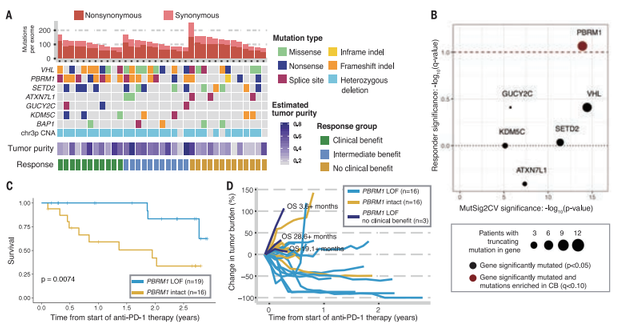


图8 PBRM1基因的纯合丢失将会显著提升患者的总体生存率和无恶化生存率

A. 不同预后的群体的肿瘤突变负荷（TMB）、显著突变基因；B. CB和NCB病人中显著富集的截短突变；C. PBRM1功能丢失与不丢失病人的生存曲线；D. PBRM1功能丢失与不丢失病人的肿瘤负荷。

2、验证阶段：选取63个同样接受免疫治疗的ccRCC患者作为独立验证集。

a)验证PBRM1的纯合丢失的确会提升患者的免疫应答率。

b)唯一一个发生PBRM1的纯合丢失患者却并没有免疫应答，发现其同时存在了另外一个在抗原呈递中起关键作用的B2M突变，这提示我们肿瘤的免疫应答涉及多方面，任何一个环节存在问题都可能会导致免疫应答失败。

3、PBRM1功能验证：PBRM1编码的蛋白质，是PBAF复合物的组成物质，该复合物主要起染色质重塑作用。深入比较PBRM1未突变和突变的肾癌细胞系的转录组表达差异，研究PBRM1是否会引起特定通路的基因表达水平变化。

案例二：抗PD-1免疫治疗过程中肿瘤微环境的变化[5]

期刊：《Cell》

影响因子：30.41

发表日期：2017年10月

研究目的：免疫检查点抑制剂对肿瘤进化的调整机制，以及基因组特征和TCR克隆特征与临床药物反应的相关性。

研究样本：68例晚期黑色素瘤病人，治疗前肿瘤组织，进行WES 150X。35个病人之前经过ipilimumab治疗（Ipi-P），33个病人没有ipilimumab治疗（Ipi-N）。Nivo（抗PD-1）药物反应在Ipi-P组为21%，Ipi-N组为22%。

研究方法：WES、转录组、TCR-seq

研究亮点：

- 抗PD-1治疗导致肿瘤突变负荷改变；
- 基因表达量变化与临床药物反应有关；
- 免疫检查点抑制治疗后TCR库发生改变。

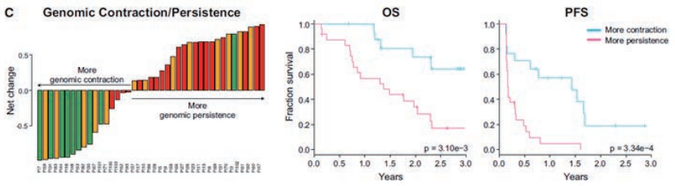


图9 基因组变化与药物反应和OS总体生存期强相关

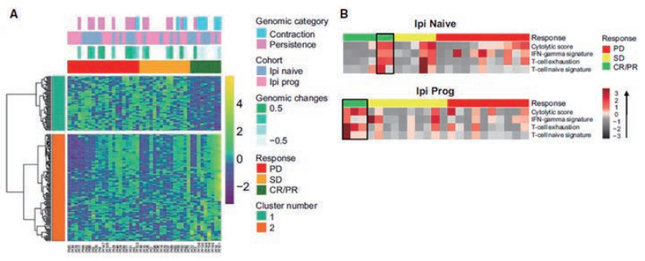


图10 CR/PR和PD对比治疗前组织RNA-seq，找到189个差异表达基因

高表达基因是免疫相关，GO注释发现是与T细胞激活、淋巴细胞聚集、调控免疫微环境的。

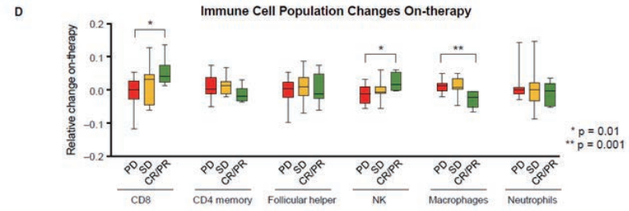


图11 TCR diversity组间差异分析

治疗前样品无差异，治疗中的样品有差异。治疗中样品，unique CDR3 sequences数量（richness）与Ipi-P组显著相关，与Ipi-N无关。
T cell evenness与Ipi-N药物有效有关，与Ipi-P无关。

可能存在的风险

ctDNA在病人体内含量，与肿瘤分期、肿瘤类型有关，因此可能存在检测不到肿瘤相关突变的情况。另外存在背景噪音干扰，需要通过加大测序深度，来增加检测极低频突变的灵敏度。技术上可以通过建库时加UMI（分子标签）的方法，提高检测的灵敏度。

常见问题

1、什么是TMB?

TMB：Tumor mutational burden，肿瘤突变负荷，指的是肿瘤基因组上非同义突变数目的个数，这张图描述了不同肿瘤TMB的状态，总体来说肺癌和黑色素瘤是最高的，主要是受吸烟和紫外线的影响。

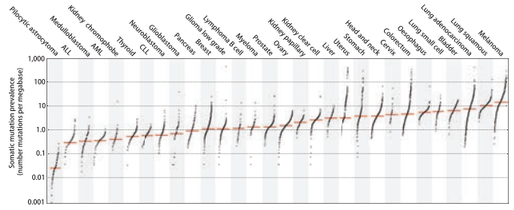


图12 突变负荷高的肿瘤倾向于对免疫治疗反应，而对传统治疗有抗药性[6]

2、什么是TNB?

TNB：Tumor neoantigen burden，肿瘤新抗原负荷，这个主要是通过检测可表达变异，并结合MHC的亲和力，预测能够别免疫系统识别的抗原。肿瘤基因组变异之后，经过转录翻译会形成新的可以被免疫细胞识别的肽段。可以想象，这些肽段非常多，并不是所有的肽段可以被自身免疫系统识别。我们现在可以运用技术来鉴定哪些肽段最容易被识别，后续可以用于疫苗的开发，来进行免疫治疗。这些可能引起免疫反应的肽段，被称之为肿瘤新抗原。TNB数目和TMB有直接的对应关系。

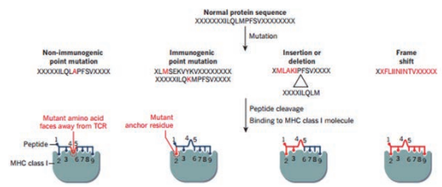


图13 肿瘤新抗原预测

3、什么是dMMR和MSI-H?

dMMR: mismatch repair deficient，错配修复基因缺陷。DNA在复制过程中会产生一些错误，正常情况下这些错误可以被错配修复基因识别并修复，如果这些基因的功能缺陷就没法达到修复的目的。dMMR相关基因主要包括MLH1、MSH2、MSH6、PMS2、MLH3、MSH3、PMS1、POLE、POLD1、FAN1。

MSI-H: microsatellite instability-high，微卫星不稳定，基因组上一些重复区域的数目发生了改变，包括单碱基或者多碱基。现在很多研究认为，dMMR可能是导致这种微卫星不稳定的原因之一。所以很多文章拿这两个指标一起来说。

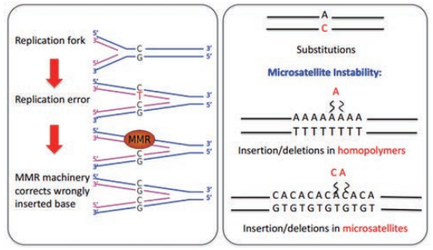


图14 dMMR和MSI-H

4、什么是免疫组库?

T/B细胞是适应性免疫系统的两大细胞群，细胞表面受体TCR/BCR存在一块区域叫互补决定区（Complementary Determining Region, CDR），包含CDR1、CDR2、CDR3，其中CDR3最高变，在抗原识别中起关键作用。华大的免疫组库测序是通过多重PCR和高通量测序技术，分析编码CDR3区的DNA/RNA序列，获得机体的免疫特征。

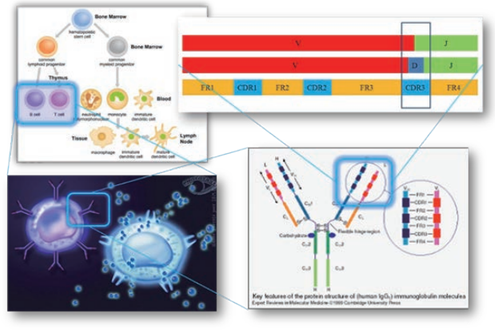


图15 免疫组库研究内容

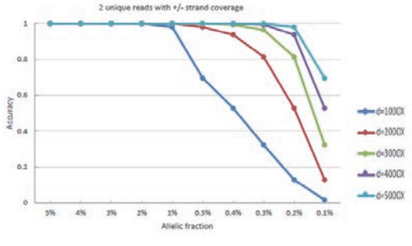
以B细胞为例，B细胞表面有BCR（B细胞受体），即Y字形的抗体。BCR顶端的区域是CDR区域（抗原互补决定区），分别由V、D、J基因编码，其中CDR1和CDR2是由V基因编码，CDR3是由V(D)J基因编码。免疫组库是通过对编码CDR3/CDR的V(D)J基因进行测序，通过基因频率反映B细胞克隆多样性。

5、华大可提供哪些肿瘤高级分析?

序号	分析条目	分析内容介绍
0	过滤、比对	标准分析
1	SNV &INDEL检测和注释	标准分析
2	超突变分类	根据 TMB, MSI 状态和 MMR 基因突变状态判断肿瘤是否超突变
3	生殖系突变	筛选肿瘤相关的生殖系突变
4	显著突变基因	1. 识别相对于背景突变率的高频突变基因 2. 识别具有功能偏向性的显著突变基因 3. 识别热点突变基因及突变分布图 4. mutation landscape 绘制
5	药靶注释	使用 CIViC 专业数据库鉴定特定体细胞突变对应的靶向药物，预测对靶向治疗的反应
6	突变特征	1. 提取突变特征并比较不同样品里各种突变特征的贡献度 2. 与机制已知的突变特征进行比较 3. 寻找突变特征活性相关的关键基因突变
7	突变链非对称性（仅 WGS）	1. 计算每个肿瘤转录链、复制链偏向性及对应碱基型 2. 比较不同组别肿瘤的转录复制偏向性；
8	新抗原预测	异常蛋白/肽序列检测，并预测与HLA的结合亲和力
9	拷贝数变异	1. 绝对拷贝数变异 2. 拷贝数中性杂合性缺失 3. 得到 allele-specific 拷贝数，分析等位基因不平衡
10	显著拷贝数变异	得到显著扩增或缺失的区域及对应基因
11	克隆进化	1. 计算每个肿瘤体细胞突变的主克隆和亚克隆比例 2. 估计肿瘤样品的纯度和倍性 3. 推测肿瘤关键基因如驱动基因和抑癌基因在肿瘤发生发展过程中扮演的角色 4. 对多位点取样病人构建肿瘤进化模型
12	突变网络	1. 突变互斥网络 2. 突变共生网络
13	分子分型	1. 根据点突变、拷贝数变异等分子特征对肿瘤进行分型 2. 结合临床预后数据比较不同组别生存差异
14	TCGA 注释	提供 TCGA 数据库里感兴趣基因的突变及拷贝数变异频率
15	结构变异（仅 WGS）	1. 识别每个样品里的结构变异及相应机制 2. 绘制每个样品的 sv, cnv, snv circos 图
16	SV 特征（仅 WGS）	提取结构变异特征，并解析其代表着不同的潜在机制
17	双微体（仅 WGS）	寻找携带关键癌基因的双微体

6、ctDNA推荐测序深度？

去重后5000x测序深度检测0.1%频率的灵敏度约为70%，去重后4000x测序深度检测0.2%频率的灵敏度约为95%，而cfDNA中SNV突变频率中位值在0.5%，因此推荐去重后的测序深度在2000x以上。



7、华大是否可提供ctDNA提取？

是，提取量根据ctDNA含量不同有变化，大约是在20ng以上。

华大优势

自主平台BGI-SEQ500，重复片段比率低于5%，准确度高：采用DNA纳米球技术，始终以同一个模板进行滚环复制，相较于PCR指数扩增，可以避免错误累积，有效提高测序准确度。

更全面的肿瘤分析流程：分析内容不仅限于体细胞突变，还包括CNV、SV、突变特征、突变网络、肿瘤新抗原识别、分子分型等分析。

全新开发的ctDNA分析流程，灵敏度高、准确度高，更适合极低频率突变的检测：

- 突变检测的灵敏度高：能够检出VAF (突变等位基因频率)低至0.2%的点突变，其中VAF为1%的点突变检出率为100%；
- 突变检测的准确率高：标准品技术重复显示检出突变一致性至少70%以上。

丰富的免疫组库研究经验：国内很早投入免疫组库研发的团队，累积发表文章13篇，可提供TCR/BCR的扩增、建库、测序、信息分析以及数据挖掘一系列全方位的服务；

专业的肿瘤研究团队：团队具有丰富的肿瘤基因组学分析经验，在国际顶级期刊发表文章30余篇，依托于华大基因在肿瘤领域长期积累的经验，为全行业提供方案设计、测序服务、数据分析、平台建设和技术优化等全面的解决方案。

参考文献

[1] Yarchoan M, Hopkins A, Jaffee E M. Tumor Mutational Burden and Response Rate to PD-1 Inhibition[J]. New England Journal of Medicine, 2017, 377(25): 2500-2501.

[2] Chen D S, Mellman I. Oncology meets immunology: the cancer-immunity cycle[J]. Immunity, 2013, 39(1): 1-10.

[3] Kinde I, Wu J, Papadopoulos N, et al. Detection and quantification of rare mutations with massively parallel sequencing[J]. Proceedings of the National Academy of Sciences, 2011, 108(23): 9530-9535.

[4] Miao D, Margolis C A, Gao W, et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma[J]. Science, 2018: eaan5951.

[5] Riaz N, Havel J J, Makarov V, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab [J]. Cell, 2017, 171(4): 934-949. e15.

[6] Alexandrov L B, Nik-Zainal S, Wedge D C, et al. Signatures of mutational processes in human cancer[J]. Nature, 2013, 500(7463): 415.

肿瘤异质性和进化研究

研究背景

仅2012年，全球有1400万新发病例和820万人患癌死亡，预计在未来的20年内将会增长70%[1]。癌症如此高的致死率和低治愈率及其普遍存在的抗药性，主要问题在于其内在的异质性。以往多区域测序的癌症研究项目发现，发生在癌症转移或复发时的亚克隆可编码突变数目基本在0～8000范围内高度变化[2]。在一些特定的癌症类型中，如黑色素瘤、肺癌，主克隆突变数目将会显著高于其他癌症类型，可能是此类癌症的发生发展容易受到外界环境因素如紫外线照射和尼古丁类致癌物的影响（图1）。同时，高丰度的主克隆突变并不意味着高丰度的亚克隆突变，如初级神经脑胶质瘤拥有最高丰度亚克隆突变，但其主克隆突变数却较低，表明该类癌症进化过程中存在多样的突变机制，这类肿瘤的高丰度亚克隆突变可能是跟错配修复缺陷有关[3-4]，而在非小细胞肺癌和膀胱癌中却可能是由于APOBEC家族导致[5-7]。

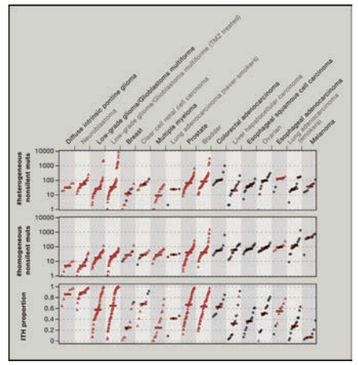


图1 不同癌症类型非沉默突变的异质性程度分布

癌症进化过程中，关键突变发生的顺序不同，会导致截然不同的癌症发展方向。在骨髓增殖类肿瘤中，先发生TET2突变将会导致造血干细胞等祖细胞的扩张，而红系祖细胞的增殖将会受到抑制，直到发生JAK2突变；相反，若先发生JAK2突变，红系祖细胞将会出现增值扩张而造血干细胞则会受到抑制。在临床层面，先发生JAK2突变的病人往往相比于先发生TET2突变的病人更可能出现红细胞增多症也更可能造成血栓。所以，两者不同进化模式的治疗方法将会有不同的侧重点，因为针对早期的突变进行靶向治疗，由于几乎所有癌细胞将会包含早期突变，所以在理论上受到影响的是所有癌细胞，Ruxolitinib作为JAK2抑制剂在JAK2早期突变病人身上则会相对取得最佳治疗效果[8]。

目前，约有16种癌症通过多区域取样测序构建了突变顺序进化图谱(表1)。癌症进化过程中除了可编码突变还会存在甲基化修饰、拷贝数变化、结构变异等一系列突变，这些突变发生在癌症进化的不同阶段从而分别扮演不同的角色，同时这些突变之间也会存在协同或是相互影响，比如在非小细胞肺癌中，早期染色体不稳定性会在一定程度上影响后期突变的异质性程度，也会显著提高癌症复发死亡的风险[9]。

表1 不同癌症类型非沉默突变的异质性程度分布

Table 1. Summary of Truncal and Branched Driver Events across Cancer Types

Tumor Type	Trunk Drivers[a]	Branch Drivers	References
AML	*DNMT3A, TET2,* t(15;17), t(8;21), t(16;16), inv(16)	*WT1, KRAS, NRAS, KIT*	Welch, 2014
Breast	*TP53, PIK3CA*	*BRCA2*	Martins et al., 2012; Nik-Zainal et al., 2012a; Shah et al., 2012
CLL	*MYD88*	*SF3B1, TP53*	Landau et al., 2013
Colorectal[b]	*KRAS, NRAS, BRAF*	*TP53, PIK3CA*	Brannon et al., 2014; Vakiani et al., 2012
Ewing Sarcoma	*EWSR1-ETS* fusion	*STAG2*	Tirode et al., 2014
Follicular lymphoma	BCl2-IGH (14;18), *MLL2, CREBBP, EZH2*	*MYD88, TNFAIP3, MYC, TP53*	Okosun et al., 2014
Glioma	*IDH1*	*SMARCA4, BRAF, TP53, ATRX*	Johnson et al., 2014
MDS	*SF3B1, SRSF2, U2AF1, DNMT3A*	*NRAS*	Papaemmanuil et al., 2013
Melanoma[c]	*BRAF*	*NRAS, MEK1*	Van Allen et al., 2014
Myeloma	*IgH* rearrangements	*KRAS, NRAS, BRAF, FAM46C*	Bolli et al., 2014; Lohr et al., 2014; Melchor et al., 2014
NSCLC	*BRAF, NF1, TP53, EGFR*	*HGF, MLL3*	Chen et al., 2012; de Bruin et al., 2014; Govindan et al., 2012
Esophageal adenocarcinoma	*TP53, SMAD4*	*MYO18B, TRIM58, CNTNAP5, ABCB1, PCDH9, UNC13C, SEMA5A, CCDC102B*	Weaver et al., 2014
Ovarian	*TP53*	*PIK3CA, CTNNB1, NF1*	Bashashati et al., 2013
Prostate	*ERG* rearrangements, 21q22 deletion, *NKX3-1* deletion *FOXP1, SPOP*	*PTEN, CDKN1B, AR* amplification	Baca et al., 2013; Haffner et al., 2013
Pancreatic	*KRAS, CDKN2A, TP53, SMAD4*	*OVCH1*	Yachida and Iacobuzio-Donahue, 2013
Renal	*VHL, PBRM1**, 3p loss of heterozygosity	*SETD2, BAP1, KDM5C, MTOR, TSC1, TSC2, TP53*	Gerlinger et al., 2012, 2014

[a]Genes with an asterisk have also been found to be subclonal in multiregion samples.
[b]Comparative sequencing analysis was used between matched primary and metastatic colorectal lesions to define potential branched status.
[c]Branched drivers defined in BRAF mutant melanoma.

方案设计

A. 研究目标

癌症进化过程中复杂多变的遗传和表观修饰突变给其进化发展方向带来多种可能性，通过对不同区域（空间）和时间的肿瘤取样进行二代测序分析，获取相关突变信息，深入研究癌症的进化机制，推断这些基因组变异在进化上的发生顺序，挖掘癌症进化过程中的关键事件，从而给癌症的治疗提供方向。

B. 研究样本

推荐20个病例，每个病例取3-5处病灶。

C. 技术方法

采用高深度全外显子测序（WES），视前期结果和研究目的而定是否进行RNA测序，ctDNA，TCR 测序；后期验证：CRISPR-Cas9细胞系实验验证。

D. 研究内容

选取肿瘤组织(实验组)和正常组织或血液（对照组），肿瘤组织根据癌变程度或空间位置进行切分。空间上，对于多个转移区域进行取样，一般按照3-5处，越多越能准确反映瘤内异质性的复杂程度。时间维度上，主要是干预治疗前后，或者复发前后，用药前后，设置多点取样。建议总样品份数在100以上，极端特殊病例可以减少样本数。采用高深度（500X）全外显子测序或者转录组测序对样品进行克隆结构分析，绘制进化树，区分早期克隆突变、晚期克隆突变，推测肿瘤的进化历史，研究不同治疗手段下抗药性的根源，指导下一步的治疗靶标方向。

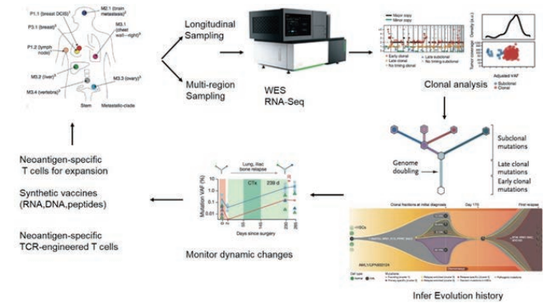


图2 肿瘤异质性和进化方案全景图

多区域取样测序同样也可研究肿瘤的异质性和进化历史，定位主克隆和亚克隆突变，在临床治疗和药物开发方面更有针对性。

E. 分析流程

图3 肿瘤进化分析流程图

F. 部分分析结果

1. 纯度和倍性分析

肿瘤组织样品在取样的时候，个体间的污染最难控制，尤其是在比较肿瘤样品和正常对照样品的研究中，即便很小比率的污染也会导致结果的假阳性。个体内污染（如癌症研究中肿瘤DNA的正常DNA污染）通常会导致敏感度下降。肿瘤纯度过低时无法保证分析结果的准确性，特别是拷贝数变异分析，我们建议剔除纯度小于30%的样品（备注：肿瘤样本中污染正常细胞时，会降低read count和read depth值，使BAFs值脱离理论值，影响分段步骤中CNV数量估计）。

当DNA从癌细胞和正常细胞的混合组织中提取和测序时，癌细胞的比例可通过ABSOLUTE软件[10]根据肿瘤样品基因组的拷贝数和体细胞突变点的等位基因频率，可以计算出肿瘤样品的纯度（Purity）和倍性（Ploidy）。

表2 纯度和倍性分析结果

SampleID	Purity	Ploidy	crossContamination
Sample_1	0.64	2	0.00463
Sample_2	0.82	2.56	0.00258
Sample_3	0.46	2	0.00617

2.拷贝数变异分析

拷贝数变异在肿瘤中非常常见，且它在致癌基因激活和抑癌基因失活上起着重要作用。体细胞拷贝数变异会同时影响上千个基因，但是只有极少量基因具有选择优势。通过评估拷贝数变异的频率和变化幅度来检测具有驱动特性的高频体细胞拷贝数变异。一般检测CNV的方法：1) read count；2) paired-end；3) assembly。随着测序成本的降低以及测序深度的增加，read count 成为最主要的方法。Read count 方法原理是利用一个非重复滑动的窗口去统计覆盖到与该窗口重叠的基因组区域内 reads 数量，从而推断发生 CNV 的位置。

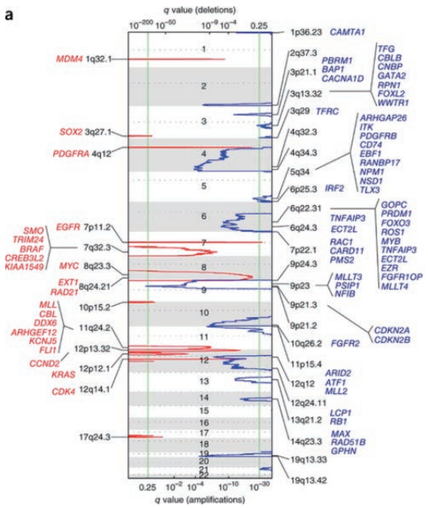


图4 通过体细胞CNV的方法识别驱动肿瘤生长的基因

3.异质性、克隆结构分析

进行高深度全外显子测序时，可用突变频率结合拷贝数、肿瘤纯度和LOH等信息，对肿瘤细胞进行聚类。聚类基于的假设：后期产生的肿瘤细胞在继承前代的突变同时，会产生自己所特有的新的突变。等位基因频率反映携带该突变的肿瘤细胞所占比例，频率越接近1表示越有可能是主克隆性突变；频率值越小，表明亚克隆突变可能性越高。以前的研究表明不同克隆状态下的相同突变对肿瘤发生的影响也不同，并且与不同的临床结果有关。以下是取自发表在*Nature Genetics*上的13名食管鳞状细胞癌患者的41个多区域样本的全外显子（WES）数据作为测试数据[11]得到的进化模型。

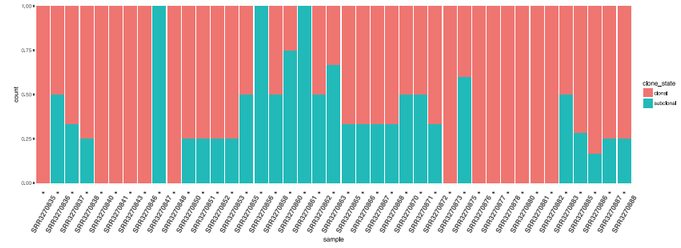


图5 41个样品的总克隆突变状况

展示同一个病人的不同区域样品的克隆比例不同，提示肿瘤内的异质性和亚克隆进化变异。

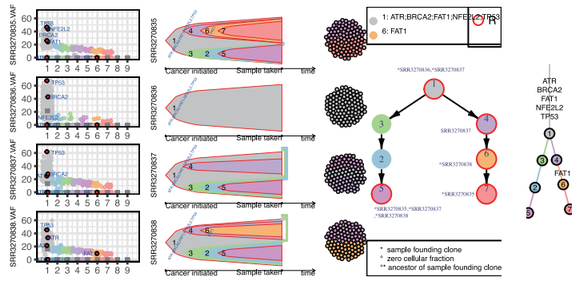


图6 通过Clonevol 推断和展示的ESCC1 病人的进化模型

利用同一患者多个区域样本测序的优势，我们可以整合多区域不同和常见突变的癌细胞信息，减少样本偏差，系统推断和可视化克隆进化史（图6）

可能存在的风险

1. 选用本方案前先文献调研是否已有研究构建过研究对象的进化历史。
2. 若有，要针对性地根据研究对象的特殊情况和优点进行个性化方案修改。
3. 理论上，测序深度越高，克隆进化结构的构建就越准确，但也存在对于部分低频的亚克隆突变未能检出的可能性，所以尽量在条件允许下提高测序深度。
4. TCR（免疫组库）测序技术较新，存在肿瘤浸润免疫细胞分离和TCR捕获的技术难点，该部分风险需要与技术人员确认。

常见问题

1. 什么是克隆?

假设认为肿瘤是起源于一个单个细胞，那么这个肿瘤被认为是一个克隆，基于这个假设，起始突变存在于每一个肿瘤细胞中，被称为CCF（cancer cell fraction肿瘤细胞分数）为1。CCF<1的细胞组成的肿瘤称为亚克隆。事实情况是即便一个特定的突变出现在一次活检中，CCF为1，但是在接下来的肿瘤采样中可能部分或者完全检测不到这个突变。

2. 后期验证主要是什么方法?

后期验证主要是通过Sanger测序(根据验证位点设计扩增引物,PCR扩增及扩增产物纯化,上机测序及序列分析,比较位点NGS结果的碱基型和Sanger测序结果碱基型)、目标区域测序(针对挑选出来用于验证的位点多,并且样品量大的情况,可以设计目标区域捕获芯片,通过目标区域捕获测序)等方法对二代测序检测到的突变位点进行准确性验证。

项目周期

测序数据下机后,从比对到体细胞检测、注释,然后计算每个肿瘤体细胞突变的主克隆和亚克隆比例、估计肿瘤样品的纯度和倍性、推测肿瘤关键基因如驱动基因和抑癌基因在肿瘤发生发展过程中扮演的角色、对多位点取样病人构建肿瘤进化模型,总周期约3-4周。后参数调整再分析的周期另算。

华大优势

自主平台测序平台,准确度高

采用DNA纳米球技术,始终以同一个模板进行滚环复制,相较于PCR指数扩增,可以避免错误累积,有效提高测序准确度。

更全面的肿瘤分析流程

分析内容不仅限于体细胞突变,还包括CNV、突变特征、突变网络、肿瘤新抗原识别、分子分型等分析。

全新开发的ctDNA分析流程,更适合极低频率突变的检测

- 突变检测的灵敏度高:能够检出VAF (突变等位基因频率)低至0.2%的点突变,其中VAF为1%的点突变检出率为100%;
- 突变检测的准确率高:标准品技术重复显示检出突变一致性至少70%以上。

专业的肿瘤研究团队

团队具有丰富的肿瘤基因组学分析经验,在国际顶级期刊发表文章30余篇,依托于华大基因在肿瘤领域长期积累的经验,为全行业提供方案设计、测序服务、数据分析、平台建设和技术优化等全面的解决方案。

案例解析

案例一:多区域测序揭示非小细胞肺癌进化历史[9]

发表期刊:《The New England Journal of Medicine》

影响因子: 72.406

发表时间: 2017年6月

研究背景:

迄今,非小细胞肺癌(NSCLC)患者瘤内异质性和癌症基因组进化数据的获取仅限于小规模的回顾性队列,研究者的目的是前瞻性探索瘤内异质性与临床之间的关联,以及确定早期NSCLC驱动事件的克隆性和进化过程。

研究方法:

对全身性治疗开始前切除的100个早期NSCLC肿瘤进行多区域的全外显子测序(WES,平均426X),测序并分析了327个肿瘤区域(323个原发肿瘤区域,4个淋巴结转移区域),以及100个从全血分离的匹配种系样品(中位数是每个肿瘤3个区域;范围为2~8),以定义肿瘤进化史,获得克隆和亚克隆事件的统计数字,并评估瘤内异质性和无再发生存之间的关联。

研究结果:

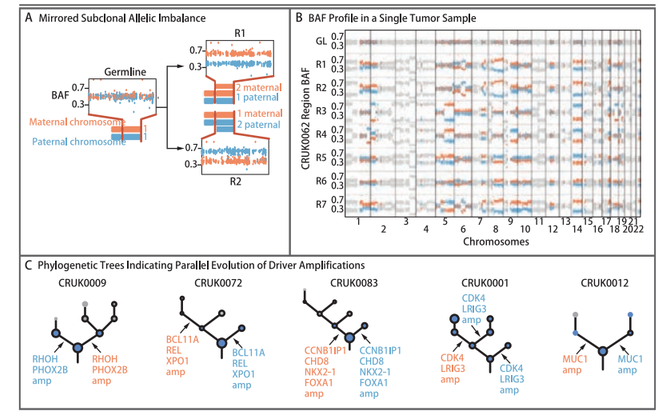


图7 瘤内异质性驱动力

观察到广泛存在的瘤内异质性,体现为体细胞基因拷贝数变异和突变。*EGFR*、*MET*、*BRAF*和*TP53*的驱动突变几乎总是克隆性的,但是在肿瘤进化的后期,超过75%的瘤内存在异质的驱动变异,且这些突变普遍发生在*PIK3CA*、*NF1*以及参与染色质修饰、DNA损伤应答和修复的基因内。基因组加倍和染色体的持续动态不稳定与瘤内异质性相关,导致驱动性体细胞基因拷贝数变异发生平行进化,其中包括*CDK4*、*FOXA1*和*BCL11A*的扩增。拷贝数异质性的加剧与再发或死亡风险的增加相关(风险比,4.9;P=4.4×10^{-4}),这种相关性在多变量统计分析中仍然显著。

本研究揭示,多区域取样测序同样也可以研究肿瘤的异质性和进化历史,定位主克隆和亚克隆突变,在临床治疗和药物开发方面更有针对性。

案例二:ctDNA监控非小细胞肺癌克隆结构动态变化[12]

发表期刊:《Nature》

发表时间:2017年5月

影响因子:40.137

研究背景:

肺癌的致死率居所有癌症的首位。目前转移的非小细胞肺癌(NSCLC)不能通过系统的化疗治愈,临床研究表明,只有5%的患者能够在术后的辅助化疗中获益。获益的患者可能是由于手术残留比较小体积的肿瘤组织,使得肿瘤内部的异质性降到了最低。基于ctDNA监测的液体活检技术与早期肺癌复发转移之间的关系还没有建立。类似的研究在乳腺癌和结直肠癌中已经开展,研究表明术后ctDNA跟踪能够早于临床发现复发,从而在早期为病人提供新的治疗方法。

研究方法:

100 例NSCLC患者(使用了96名,4名未通过质控),使用肿瘤组织和术前血液样品进行多区域外显子测序构建进化树。基于肿瘤组织的主克隆和亚克隆SNVs建立每个病人的multiplex-PCR assay panel,然后合并在一起,术前和术后多时间点进行ctDNA取样,用于液体活检中跟踪肿瘤的进化过程。

研究结果：

发现在NSCLC中存在的基因缺陷，可以用从肿瘤中游离出来的血液DNA片段（ctDNA）来监测。然后，他们分析了24例非小细胞肺癌患者手术后的血液，并在临床影像证实疾病复发前的一年，准确地鉴定出超过90%的人注定要复发。在复发前或复发后有13个（93%）NSCLC样品检测出至少两个SNVs。但是在临床上显示没有转移的样品中只有1个（10%）检测出至少两个SNVs。ctDNA突变检出比临床上CT确诊复发早，平均时间间隔在70天（range，10-346days）。除此之外，ctDNA profiling还能够显示辅助化疗的耐药性现象（图3a-c）。通过比较术后ctDNA上检测到的亚克隆SNVs和多位点取样的原发癌SNVs，可以预测亚克隆突变是否参与肿瘤复发。在四个NSCLC复发病人的术后ctDNA发现其某个特定进化树分支上的亚克隆突变VAF和主克隆VAF相似（图3b,f-g），这意味着该亚克隆突变在肿瘤复发的过程中占据主导地位。

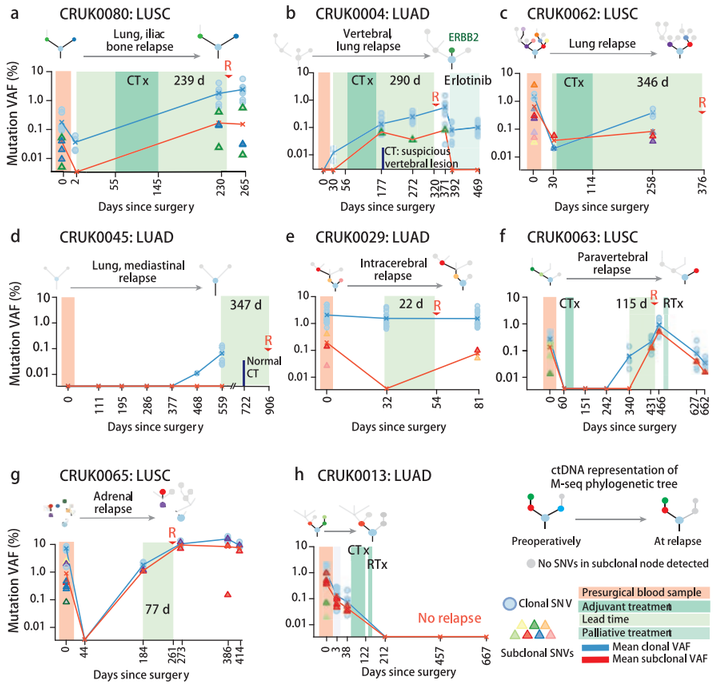


图8 术后ctDNA监测情况

参考文献

[1] Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013[J]. CA: a cancer journal for clinicians, 2013, 63(1): 11-30.

[2] McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future[J]. Cell, 2017, 168(4): 613-628.

[3] Johnson B E, Mazor T, Hong C, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma[J]. Science, 2014, 343(6167): 189-193.

[4] Kim H, Zheng S, Amini S S, et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution[J]. Genome research, 2015, 25(3): 316-327.

[5] de Bruin E C, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution[J]. Science, 2014, 346(6206): 251-256.

[6] McGranahan N, Favero F, de Bruin E C, et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution[J]. Science translational medicine, 2015, 7(283): 283ra54-283ra54.

[7] Zhang J, Fujimoto J, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing[J]. Science, 2014, 346(6206): 256-259.

[8] Ortmann C A, Kent D G, Nangalia J, et al. Effect of mutation order on myeloproliferative neoplasms[J]. New England Journal of Medicine, 2015, 372(7): 601-612.

[9] Jamal-Hanjani M, Wilson G A, McGranahan N, et al. Tracking the evolution of non–small-cell lung cancer[J]. New England Journal of Medicine, 2017, 376(22): 2109-2121.

[10] Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer[J]. Nature biotechnology, 2012, 30(5): 413.

[11] Hao J J, Lin D C, Dinh H Q, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma[J]. Nature genetics, 2016, 48(12): 1500.

[12] Abbosh C, Birkbak N J, Wilson G A, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution[J]. Nature, 2017, 545(7655): 446.

肿瘤免疫组库研究方案

研究背景

肿瘤的发生是由于部分细胞基因组发生突变，细胞调控发生异常，导致不断增殖。这种细胞的变化和异常增殖与微环境的支持是分不开的，微环境包括细胞和蛋白提供的结构、血管、免疫微环境等。正常情况下，免疫系统对异常细胞会进行及时清除，但当免疫微环境发生变化，例如肿瘤细胞启动免疫抑制程序（PD-1/PD-L1），会使得免疫系统对肿瘤细胞失去监控，导致肿瘤细胞未被及时清除，导致肿瘤的发生。

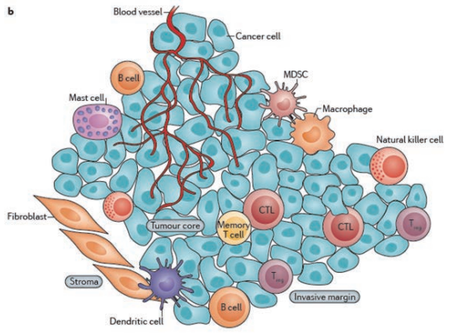


图1 肿瘤组织中的免疫微环境

免疫细胞会侵入肿瘤组织，包括适应性免疫系统中的B细胞、杀伤T细胞（CTL）、记忆T细胞、辅助T细胞、调控T细胞（Treg）；
也包括固有免疫系统中的巨噬细胞、树突状细胞、肥大细胞、自然杀伤细胞、髓源性抑制细胞（MDSCs）[1]

肿瘤细胞是具有免疫原性的，肿瘤细胞死亡后产生的肿瘤抗原会引起T淋巴细胞浸润到肿瘤组织中，启动杀伤肿瘤细胞作用。多项研究表明肿瘤组织中TIL（浸润淋巴T细胞）的存在及数量与病人生存期相关，因此TIL可作为一种肿瘤预后的生物标记物。如何对T细胞进行定量和定性?临床上常采用的技术是免疫组化，可以对特定抗原表达进行定性，例如PD-L1免疫组化检测的主要用于预测PD-1抑制剂的疗效。除此之外，一种全新的技术——免疫组库（TCR-seq），可以对TCR进行定性和定量研究，近年来被用于各种肿瘤预后和疗效研究。

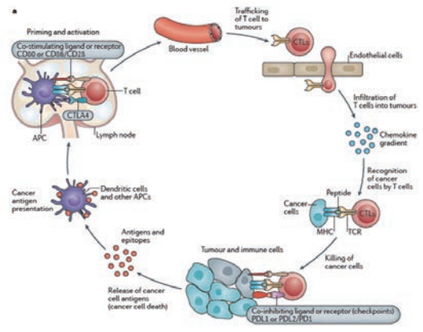


图2 肿瘤免疫循环

首先，肿瘤细胞死亡后会释放肿瘤抗原，肿瘤抗原被树突状细胞捕捉并通过MHC提呈给T细胞识别，从而启动和激活了效应T细胞。
激活后的T细胞通过血液循环浸润到肿瘤组织，通过TCR特异性识别肿瘤抗原，启动杀伤肿瘤细胞的作用[1]

免疫组库技术是利用高通量测序技术，分析编码TCR/BCR的V(D)J基因多样性，进而评估T/B细胞克隆多样性，其在肿瘤领域有着广泛的应用前景。

（1）肿瘤早诊：癌症是具有克隆性的，淋巴系统肿瘤是T细胞或B细胞发生了恶性增殖。大部分的肿瘤类型有特异性的TCR或BCR，外周血、骨髓、淋巴结中含有的一些高丰度、致癌性的TCR或BCR序列特征，可以辅助诊断白血病或淋巴瘤。

暴露于同一种病原的不同病人，他们的记忆淋巴细胞是相同的或非常接近的，即公共克隆（public clones）。通过大数据收集建立公共克隆数据库，未来可通过外周血免疫组库特征，对肿瘤进行早期辅助诊断。

（2）MRD检测：肿瘤治疗后，外周血中如果残留有特异性的克隆序列，可以用于预后，即微小残留病检测（MRD，minimal residual disease），灵敏度为1/10^6 有核细胞。

（3）检测TCR动态变化，评估免疫疗效：通过评估肿瘤TIL及外周血中的TCR分布和频率，可以分析肿瘤的免疫原性以及评估对免疫药物的反应和病人预后。不同病灶对比分析，展示的是肿瘤内部和微环境的异质性。

什么是免疫组库?

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。

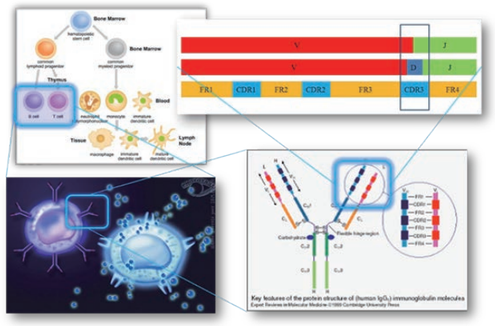


图3 免疫组库研究内容

以B细胞为例，B细胞表面有BCR（B细胞受体），即Y字形的抗体。BCR顶端的区域是CDR区域（抗原互补决定区），分别由V、D、J基因编码，其中CDR1和CDR2是由V基因编码，CDR3是由V(D)J基因编码。免疫组库是通过对编码CDR3/CDR的V(D)J基因进行测序，通过基因频率反映B细胞克隆多样性。

TCR/BCR CDR3多样性是如何实现的?

1、V(D)J recombination重排；

2、V-D和D-J间随机插入碱基；

3、抗体常发生体细胞突变。

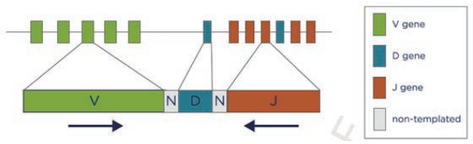


图4 CDR3编码基因的多样性是由V(D)J基因重排加上随机插入碱基产生的

免疫组库在肿瘤领域的研究进展

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等等；医学应用上，疫苗研发评价、药物研发、疾病诊断、器官和肝细胞移植等，发表文章呈逐年上升趋势。

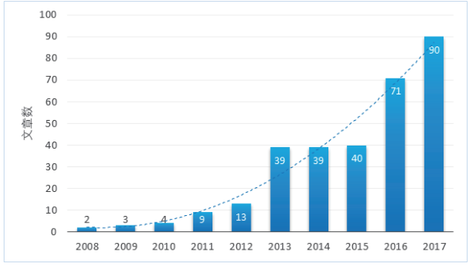


图5 免疫组库已发表文章统计（IF>5文章不完全统计）

免疫组库科研方面可用于肿瘤预后biomarker、疗效评估等

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[2-7]。近年来免疫组库常与肿瘤新抗原结合起来分析，肿瘤新抗原（neoantigen）会刺激T细胞进行抗肿瘤免疫反应，TCR克隆多样性与肿瘤复发和生存率低有关[8]。免疫组库也常用于免疫检查点抑制剂的药效评估，采用多组学（WES、转录组和TCR测序），对68例晚期黑色素瘤病人进行分析，发现药物治疗前后T细胞克隆发生了变化，并且发现TCR克隆与药物的相关性[9]。

免疫组库临床上可用于微小残留病检测

微小残留病（Minimal Residual Disease, MRD）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[10-12]。

表1 免疫组库已研究肿瘤类型

实体瘤	血液系统肿瘤
乳腺癌	滤泡淋巴瘤
结直肠癌	弥漫性大 B 细胞淋巴瘤
成神经细胞瘤	T 细胞淋巴瘤
脑胶质瘤	白血病
头颈鳞癌	
鼻咽癌	免疫治疗
甲状腺癌新抗原和 TIL	肿瘤新抗原
非小细胞肺癌	免疫检查点 PD-1、CTLA-4
肺腺癌	肿瘤疫苗
肝癌 肝炎	MHC & TCR
食管癌	
胃癌	MRD 微小残留病检测
胃肠间质瘤	ALL
黑色素瘤 微环境	B-lineage ALL
前列腺癌 免疫治疗	B 细胞淋巴瘤
卵巢癌 TIL	CLL
乳腺癌 TIL	ctDNA 淋巴瘤
肾癌	MRD
胰管腺癌	multiple myeloma
肿瘤浸润 B 细胞	T-ALL 急性淋巴细胞白血病

方案设计

A. 研究目的（临床切入点）

1、对比**病灶和外周血**的克隆特征，寻找外周血中能够预测疾病发展或者预后相关的克隆特征（例如V基因、V-J基因的突变特征），用于非侵入性早期诊断。

2、从不**同药效**入手，例如治疗前后取样，对比克隆频率变化，找到与疗效相关的biomarker。

3、对比**不同部位**的克隆特征，分析肿瘤组织内部的异质性，及其与复发、生存期的关系。

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，例如不同疾病亚型，或者临床上易混淆疾病，寻找不同疾病诊断的biomarker。

B. 样品选择

每个group 10-20例，分多点取样（时间点、外周血和病灶、不同亚型），对照样品外周血样品10-20例，总计50-100个样品。设置的组越多，每个组的样品越少；分析的组学越多，样品数量也可以变少。

选取样品类型包括肿瘤组织（浸润淋巴T细胞）、外周血等。

C. 实验技术

多重PCR在BCR或TCR的位于CDR3区两端的V、J基因保守区域设计PCR引物，通过多重PCR扩增得到互补决定区CDR3区域，扩增产物用于后续高通量测序PE151（数据量推荐1Gb raw data）。如果模板为RNA，则需要先进行反转录得到cDNA，再进行多重PCR。

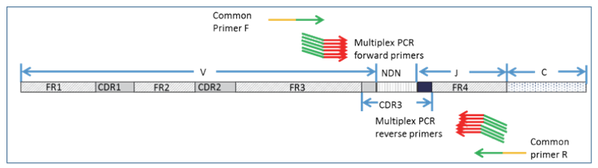


图6 多重PCR示意图

D. 信息分析流程

1）测序所得的数据称为raw reads或raw data，随后要对raw reads进行质控(QC)，以确定测序数据是否适用于后续分析；

2）经过滤得到的clean reads比对到参考序列，对于比对上的reads，做下一步的组装，得到具体的功能区域，例如CDR3区（clones）；

3）碱基质量符合要求的克隆序列会作为核心克隆（core clonotype），存在一个以上质量值较差碱基的克隆会以核心克隆作参考二次比对和校正；

4）然后，对相差一个碱基的克隆，进行层次聚类，每个分支间仅有一个碱基差别（mismatch），依次聚类下去，克隆频率低的克隆会合并到上一分支，最终保留最顶端的head序列；

5）将上述得到的克隆序列再次比对到V，D，J和C参考序列，最终得到的统计文件包含了克隆序列、氨基酸残基序列、克隆数量、克隆频率，V/J基因组合等信息。后续可以根据这些信息做克隆分布、基因重组、多样性分析等深入挖掘。

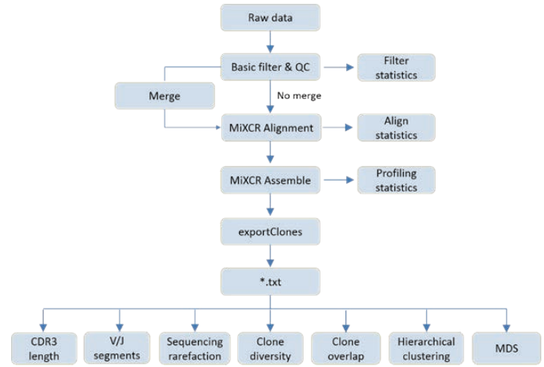


图7 信息分析流程

E. 信息分析内容

1、基本数据统计

数据过滤，对原始数据进行去除接头污染及低质量reads的处理

数据搭建，数据拼接，消除测序背景及有效数据构建

数据统计，数据产出统计及测序数据的成分和质量评估

2、数据比对分析

比对分析，与数据库V/D/J基因片段比对

比对结果统计

3、克隆序列特征注释

CDR3区核酸序列和氨基酸序列

鉴定无效序列（包含终止密码子，超出结构范围）

鉴定单碱基突变（替换、删除、插入）（for BCR）

4、单样品克隆群体特征分析

CDR3序列长度分布

V/J基因频率分布

V-J基因组合频率分布（3D,Circos）

克隆群体结构分析（频率分布，D50曲线，甜甜圈图）

5、样品间比较分析

测序饱和度分析

克隆多样性分析(辛普森系数、香农威纳系数等)

样品间共有克隆分析

聚类分析（层次聚类，MDS聚类）

组间差异分析

F. 部分分析结果展示

1. V-J基因频率

针对测序数据结果序列，使用IMGT数据库进行比对，鉴定出V、D、J基因，并对样本中所有克隆的V基因、J基因、V-J基因组合形式进行了统计，以每种克隆reads数计算权重，V-J组合结果以3D和Circos图分别展示。

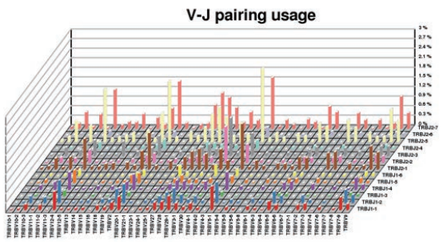


图8 V-J基因组合频率3D柱状图

平面上分别为V基因、J基因。柱子的高度代表一种V-J组合的频率。

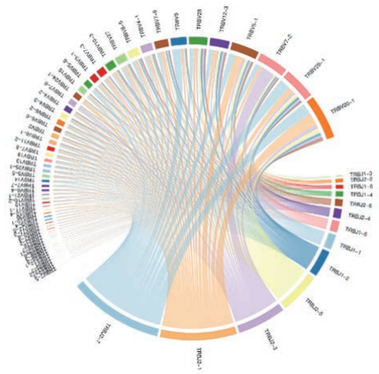


图9 V-J基因组合频率Circos图

每个颜色块代表一种基因，颜色块越宽，频率越高。色块间的连线代表一种V-J基因组合方式。

2. 克隆多样性分析

克隆多样性统计，是不同于V-J基因频率的统计。V-J基因会存在SNP（BCR存在超突变）、随机碱基插入等，增加了克隆的多样性。

样品克隆频率分布图直观反映每个样本中所有克隆类型频率分布情况，D50是近年来引入反映样本克隆群体结构的一个指标，值越低，反映克隆多样性越低，值越大，克隆多样性越高。

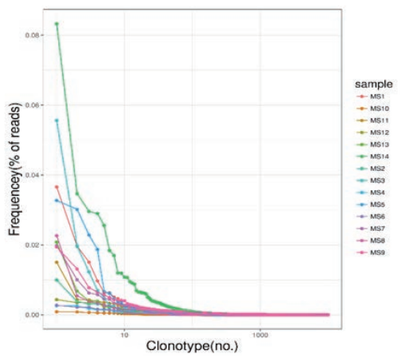


图10 克隆频率分布图

横纵坐标分别为克隆数和克隆频率。

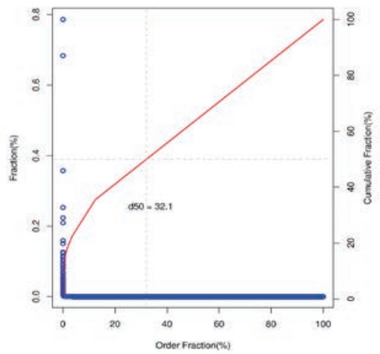


图11 D50曲线

X轴表示样本克隆组成累积百分比，左侧Y轴表示单个类型克隆频率，右侧Y轴表示克隆频率累计百分比。每个点表示单个克隆具体的频率，曲线为所有克隆的累计分布。其中的D50为累计频率达到50%时的克隆所在位置。

3. 组间差异分析

分组比较克隆多样性及top20高频表达V基因频率。

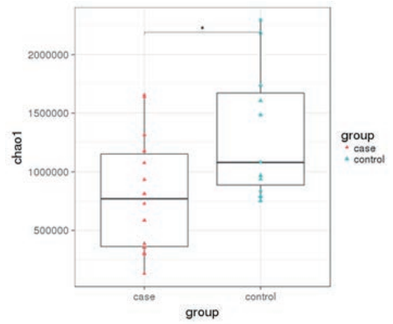


图12 克隆多样性箱线图

每个箱线代表一个group，每个箱线图对应五个统计量（自上而下分别为最大值，上四分位数，中值，下四分位数和最小值）。使用Student's t-Test进行差异显著性检验，其中ns表示差异不显著（P>0.05）；*表示有统计学差异（P<0.05）；**表示有显著统计学差异（P<0.01）；***表示有极其显著的统计学差异（P<0.001）。

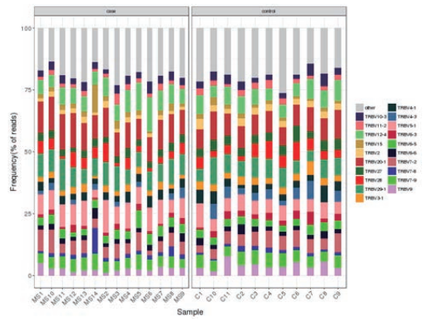


图13 Top20 V基因频率组间分布柱状图

X轴表示样品编号，Y轴表示重组结果中各基因的使用频率。

应用案例

案例一：TCR克隆异质性与新抗原异质性及肺癌复发的关系[7]

发表期刊：《Cancer Discovery》

影响因子：20.01

发表时间：2017年10月

研究目的：肿瘤突变产生新抗原，新抗原会刺激T细胞进行抗肿瘤免疫反应。因此，通过TCR测序可以反映克隆多样性及其与预后的相关性。

研究样本：11例未转移肺腺癌，共45 tumor regions (2 to 5 regions per tumor)

研究技术：WES、TCR sequencing

研究结果：病灶特有突变导致新抗原（neoantigen）的内部表达差异，因此，产生不同的免疫原性，形成了TIL克隆内部差异（TCR ITH）。TCR ITH（intratumor heterogeneity）高与术后疾病复发和低生存率有关。

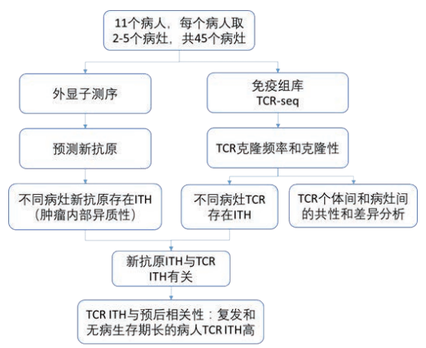


图14 文章研究思路

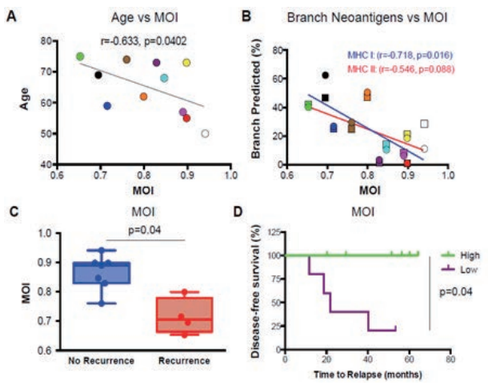


图15 TCR克隆异质性（ITH）与肺腺癌基因组和临床的相关性

（A）年龄与MOI相关性；（B）新抗原与MOI负相关；（C）复发的病人具有更高的TCR ITH（lower MOI）；（D）MOI越高（TCR克隆多样性越低），病人的无病生存期（disease-free survival）越长。MOI：克隆重叠指数，范围0-1，0代表克隆完全不同，1代表克隆完全相同。

案例二：抗PD-1免疫治疗过程中肿瘤微环境的变化[8]

期刊：《Cell》

影响因子：30.41

发表日期：2017年10月

研究目的：免疫检查点抑制剂对肿瘤进化的调整机制。

研究样本：68例晚期黑色素瘤病人，治疗前肿瘤组织，进行WES 150X。35个病人之前经过ipilimumab治疗（Ipi-P），33个病人没有ipilimumab治疗（Ipi-N）。Nivo（抗PD-1）药物反应在Ipi-P组为21%，Ipi-N组为22%。

研究方法：WES、转录组、TCR-seq

研究亮点：

- 抗PD-1治疗导致肿瘤突变负荷改变；
- 基因表达量变化与临床药物反应有关；
- 免疫检查点抑制治疗后TCR库发生改变。

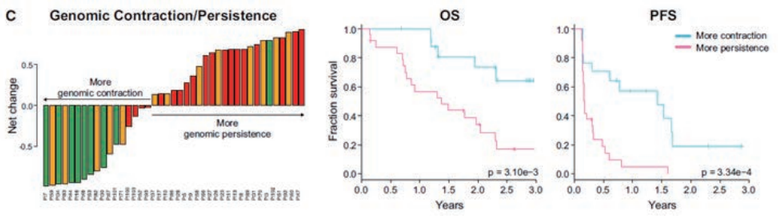


图16 基因组变化与药物反应和OS总体生存期强相关

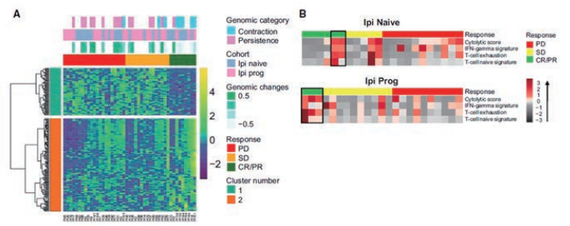


图17 CR/PR和PD对比治疗前组织RNA-seq，找到189个差异表达基因

高表达基因是免疫相关，GO注释发现是与T细胞激活、淋巴细胞聚集、调控免疫微环境的。

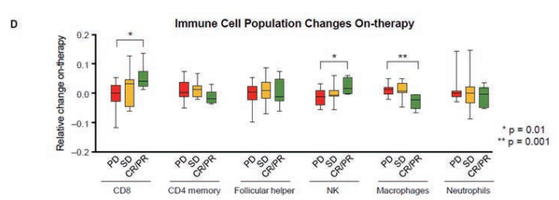


图18 TCR diversity组间差异分析

治疗前样品无差异，治疗中的样品有差异。治疗中样品，unique CDR3 sequences数量（richness）与Ipi-P组显著相关，与Ipi-N无关。T cell evenness与Ipi-N药物有效有关，与Ipi-P无关。

案例三：TCR V基因频率作为biomarker，可区分HBV相关的肝癌亚型[24]

杂志：《OncoImmunology》

影响因子：7.72

发表日期：2015年4月

研究目的：找到鉴定肝癌亚型的biomarker，用于非侵入性早期诊断

研究样品：160个RNA样品

样品来源	组织	外周血	癌组织	癌旁组织
健康人	9	21	–	–
HBV病人	17	17	–	–
HCC病人	–	20	21	21
MHC病人	–	–	3	3
ICC病人	–	–	5	5
结肠癌病人	–	10	4	4

注：HBV，乙肝；HCC，肝细胞癌；ICC，肝内胆管癌；MHC，混合型的肝细胞癌和肝内胆管癌

研究结果：

1. 不同群体间的TCRβ CDR3差异表达分析

肝癌、肝炎、健康人的外周血和组织TCR差异显著，可以用外周血作为非侵入性早期诊断的biomarker。

不同亚型癌症的癌组织和癌旁组织进行相关性分析，发现HCC癌组织和癌旁组织相关性低，可能说明HCC恶性程度低；而ICC癌组织和癌旁组织相关性高，可能说明癌细胞已经发生了转移。

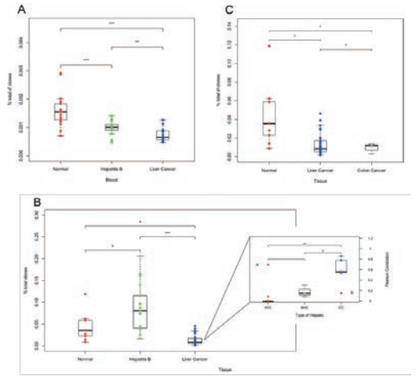


图19 不同样品的外周血和组织样品TCR克隆差异

2. HCC病人和健康人的V、J基因主成分分析

为了进一步寻找能够分辨HCC和健康人的生物标志物，用Vβ序列、Vβ亚家族（Vβ融合成Vβ亚家族）、VJ（V-J配对）来分析两个人群的差异。结果显示，外周血样品可以用这三种因素区分开。进一步用ROC分析发现，单独用Vβ序列分辨的效果更好。预示着，V基因也许可以作为HCC分类的参考。

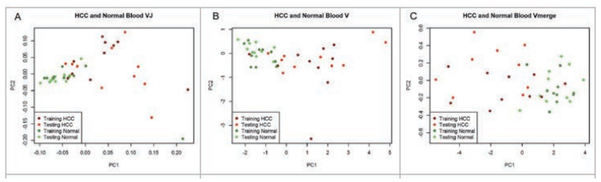


图20 肝细胞癌和正常人外周血中V基因、V-J基因（VJ）、V基因家族（Vmerge）的PCA分析

可能存在的风险

项目设计不合理或样品数太少，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，设计合理的疾病和对照组，并确定疾病与T细胞还是B细胞有关。

常见问题

Q1：TCR和BCR各条链编码基因的区别?推荐哪条链?

TCR Beta链和BCR重链是由V、D、J基因编码的，而TCR alpha链和BCR轻链是由V、J基因编码的。从发表文章来看，研究TCR beta链和BCR重链的比较多。

Q2：免疫组库测序可以区分IgG、IgM、IgD、IgE吗?

免疫球蛋白的亚型，通过C区序列可以区分，华大有相应的引物，但必须是RNA样品。利用多重PCR的方法，利用V区-C区的引物，用约30bp的序列区分。产物长度大约是在200-300bp。

Q3：免疫组库的测序深度?能得到多少序列?

分析数据显示，数据量的增加，主要影响低频克隆，并且这些克隆的排序在一千多到九千不等，而研究往往只关注top100的克隆和疾病的关系，所以推荐起始数据量1G raw data。但如果客户想关注更多低频克隆，可以加大测序数据量。

备注：表格中的样品（H1-H4，P1-P8）原始数据量为2-3G，截掉一半数据量为1-1.5G，表格中highest uniq_rank这一项表示unique 的克隆中频率最高的那个克隆在原始数据克隆中的排名。

Sample	origion	cut	overlap	uniq_by_origin	highest uniq rank
H1	26654	15662	15213	11441	2108
H2	17295	9975	9733	7562	1293
H3	27516	16301	15940	11576	2305
H4	25866	15153	14781	11085	2137
P1	28436	16748	16305	12131	3343
P2	27116	14921	14583	12533	2047
P3	25498	14453	14180	11318	2355
P4	26675	15407	15045	11630	2339
P5	55631	32664	31787	23844	5973
P6	46524	30249	28859	17665	8306
P7	63479	40190	38575	24904	8950
P8	49508	30062	29037	20471	5366

Q4：做免疫组库测序，用基因组DNA做模板好，还是RNA好?

A6：DNA水平侧重于研究基因重组信息，RNA水平侧重于研究基因的表达状态。使用DNA和RNA做模板各有优缺点：

gDNA的优点是：

1）因为每个基因只有两个拷贝，因此可准确地反映免疫细胞受体的克隆数；

2）DNA更稳定，易储存。

缺点是：

1）由于copy数不高，模板含量低，因此可能需要更多的样品；

2）由于J区和C区之间有很大的intron区，受测序长度的局限，缺少特异性扩增引物来扩增CDR全长。

RNA的优点是：

1）J区和C区之间无intron，可用C区进行引物设计，扩增全长CDR；

2）由于表达丰富，模板含量高，样品消耗量少。

缺点是：

1）免疫细胞受体的克隆性受到mRNA表达高低的影响，不能客观地反应本身的克隆数；

2）RNA不如DNA稳定，样品保存和操作要求较高。

华大优势

1. 丰富的项目经验：已完成包括肿瘤、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。

2. 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。

3. 扩增偏好性低：采用两步法建库，并优化引物配比，将扩增偏好性降低约70%。

4. 可重复性高：同一样本建库两次克隆一致性高。

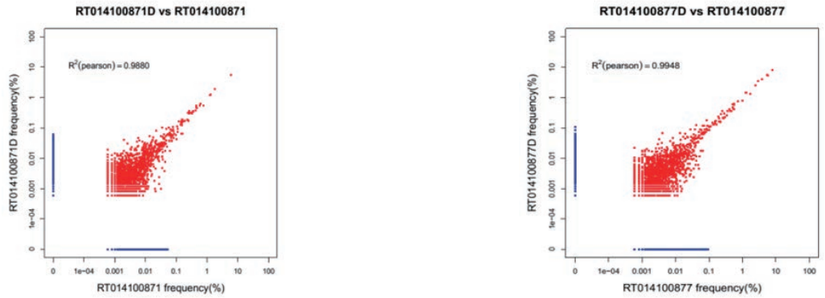


图21 同一样本实验重复性评估(HiSeq)

5. 更丰富的分析结果：新增多种结果统计图表；新增V/J基因频率分布统计、多样品聚类分析、共性分析、差异分析；新增四种多样性评估指标。

6. 更友好的xbio结题报告展现形式：全新升级的结题报告界面友好，对分析方法及结果解释详尽，图表按照发表文章要求展示，让您一目了然。

7. 更真实的克隆定量信息：将低质量reads与高质量克隆进行比对，挽回重要数据，让克隆定量信息不丢失。

8. 更强的纠错能力和错配处理能力：利用多层聚类方法，纠正PCR和测序引入的错误；错配处理能力提升，更适合分析BCR的高突变区域。

华大已发表文章

研究内容	发表时间	发表期刊	影响因子	文献标题
肝癌亚型差异分析	2015.04	*Oncoimmunology*	7.72	Identification of characteristic TRB V usage in HBVassociated HCC by using differential expression profiling analysis
分析软件	2015.08	*Genetics*	4.56	IMonitor: a robust pipeline for TCR and BCR repertoire analysis
骆驼	2016.09	*PLoS One*	2.81	Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies
实验方法学	2016.03	*PLoS One*	2.81	Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire.
肝癌	2015.07	*Cancer Letters*	6.38	Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma
原发性胆汁性胆管炎	2016.09	Journal of immunology	4.86	Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis
微小残留病	2016.10	Frontiers in Immunology	6.43	Minimal Residual Disease Detection and Evolved IGH Clones Analysis in Acute B Lymphoblastic Leukemia Using IGH Deep Sequencing
预测 V/J 基因软件	2016.11	Frontiers in Immunology	6.43	IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data
乳腺癌、癌旁和淋巴结的 TCR 分析	2017.02	Cancer Immunology Research	8.28	The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues
结肠腺瘤和结肠癌浸润淋巴细胞	2017.05	Journal of immunology	4.86	Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma
免疫缺陷	2015.12	Human Molecular Genetics	5.99	DCLRE1C (ARTEMIS) mutations causing phenotypes ranging from atypical severe combined immunodeficiency to mere antibody deficiency
肾移植	2016.11	Transplant Immunology	1.32	T cell repertoire following kidney transplantation revealed by high-throughput sequencing

参考文献

[1] Hackl H, Charoentong P, Finotello F, et al. Computational genomics tools for dissecting tumour–immune cell interactions[J]. Nature Reviews Genetics, 2016, 17(8): 441.

[2] M.J. Gooden, G.H. de Bock, et al., The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis, Br. J. Cancer 105 (2011) 93–103.

[3] Y.X. Han, X. Liu, et al., Identification ofcharacteristic TRB V usage in HBV-associated HCC by using differentialexpression profiling analysis, Oncoimmunology 4 (2015) e1021537.

[4] A.M. Sherwood, et al.,Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of Tcell receptor sequences that differ from the T cells in adjacent mucosal tissue, Cancer Immunol Immunother. 62 (2013) 1453–1461.

[5] W.T. Hwang, et al., Prognostic significance of tumor-infiltrating T cells in ovarian cancer: a meta-analysis, Gynecol. Oncol. 124 (2012) 192–198.

[6] R.O. Emerson, et al., High-throughput sequencing of T-cell receptors reveals ahomogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer, J. Pathol. 231 (2013) 433–440.

[7] F. Azimi, R.A. Scolyer, et al., Tumor-infiltrating lymphocyte grade is an independent predictor ofsentinel lymph node status and survival in patients with cutaneous melanoma, J. Clin. Oncol. 30 (2012) 2678–2683.

[8] Reuben A, Gittelman R, Gao J, et al. TCR repertoire intratumor heterogeneity in localized lung adenocarcinomas: an association with predicted neoantigen heterogeneity and postsurgical recurrence[J]. Cancer discovery, 2017, 7(10): 1088-1097.

[9] Riaz N, Havel J J, Makarov V, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab [J]. Cell, 2017, 171(4): 934-949. e15.

[10] D. Wu, A. Sherwood, J.R. Fromm, S.S. Winter, et al.,High-throughput sequencing detects minimal residual disease in acute Tlymphoblastic leukemia, Sci. Transl. Med. 4 (2012) 134ra163.

[11] M. Faham, J. Zheng, M. Moorhead, et al.,Deep-sequencing approach for minimal residual disease detection in acutelymphoblastic leukemia, Blood 120 (2012) 5173–5180.

[12] J. Martinez-Lopez, J.J. Lahuerta, F. Pepin, et al.,Prognostic value of deep sequencing method for minimal residual diseasedetection in multiple myeloma, Blood 123 (2014) 3073–3079.

[13] 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.

ctDNA肿瘤液体活检研究方案

研究背景

2000年6月26日，人类基因组草图绘制的完成，为基因组学的转化医学应用奠定了基础；2006年国际癌症基因组计划的开展，完成了上万例的癌症样本的测序及上千万癌症突变位点的发现，则为癌症的精准医学发展打下了基础。目前基于高通量测序的肿瘤精准医学研究正处于一个飞速发展的阶段，由于肿瘤异质性的影响，仅仅取某个部位的肿瘤组织检测并不能反映患者的整体情况，因此肿瘤的精准检测是肿瘤精准医学研究首要解决的课题。

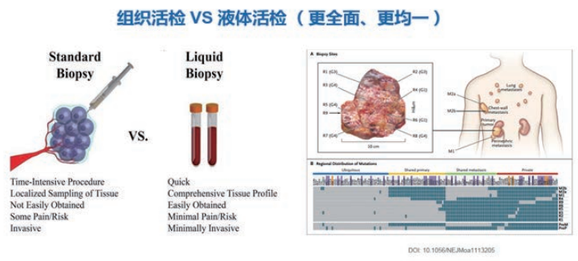


图1 液体活检可有效克服局部活检的局限性[1]

肿瘤液体活检技术，有效地克服了局部活检的局限性。该技术通过快速简捷的无创性诊断，不仅避免了穿刺活检带来的肿瘤转移的风险，还能实时的评估患者的治疗效果和整体情况，从而更加全面地指导临床药物的选择。由于液体活检技术的众多优势，2015年被《麻省理工科技评论》评价为年度10大科技突破。

1. 液体活检的概念发展

1948年，Mandel和Metais在血浆中发现了游离DNA（cfDNA）的存在[1]；1977年，Leon使用放射免疫化学证明，对于至少一半的癌症患者，血液中的cfDNA水平显着高于正常对照组，转移性癌症患者血液中的cfDNA水平显着升高[2]；1989年，Stroun发现cfDNA和癌细胞的DNA具有相同的生物物理特性，表明部分cfDNA是由肿瘤DNA游离出来的，液体活检（liquid biospy）的概念第一次被提出来[3]。1994年，癌症患者血液中突变RAS基因片段的突破性发现使循环肿瘤DNA(ctDNA)成为了研究的焦点[4, 5]。

外泌体发现于80年代初期，刚开始被认为是移除细胞内所不需要的蛋白的结构，类似于垃圾袋的功能，但在2007年发现了外泌体中有miRNA和mRNA的存在，揭示了外泌体在细胞间传递信号的功能[6]，在此之后关于外泌体的研究进入了一个指数增长期[7, 8]。

2. 什么是液体活检（Liquid Biopsy）

广义：一种直接从血液、唾液、尿液、腹水、胸水等体液中非侵入性检测生物标记物的技术

狭义：主要指检测血液内的循环肿瘤细胞（CTC）、游离肿瘤DNA（ctDNA）和外泌体(Exosomes)

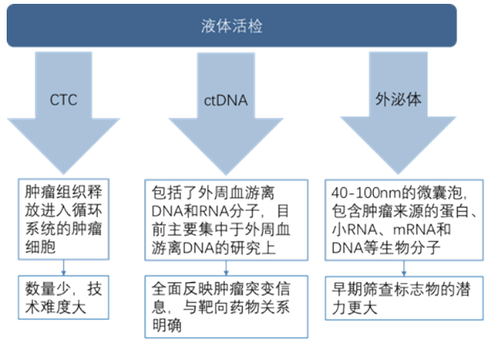


图2 液体活检的分类

作为体外诊断的一个分支，液体活检从广义上来讲是指所有的直接从血液、唾液、尿液、胸腹水等体液中非侵入性检测生物标记物的技术。但是从狭义的层面上来说，现阶段的液体活检技术主要是指通过循环肿瘤细胞(Circulating tumor cell, CTC)、游离核酸 (Circulating Tumor nucleic acids, ctNA)和外泌体等生物标志分子，进行肿瘤检测、诊断的技术。

CTC是来自于肿瘤组织而进入循环系统的肿瘤细胞；ctNA包括了外周血游离DNA和RNA分子，目前主要集中于外周血游离DNA (Circulating Tumor DNA, ctDNA) 的研究上；外泌体是由40-100nm的微囊泡，可以检测肿瘤来源的蛋白、小RNA、mRNA和DNA等生物分子。

3. 液体活检优劣势

CTC具有的主要优点是肿瘤来源特异性，但是受到数量及识别技术的制约，研究和应用的技术难度比较大。

ctDNA的优势是能够全面地反映肿瘤的突变信息，肿瘤相关突变特异性高，与靶向药物关系明确，但是由于在血液中游离DNA中的突变频率过低，容易受到背景噪音的影响。随着高通量测序的发展的测序成本价格的下降，通过超高深度测序克服背景噪音的影响成为可能，也预示了ctDNA更广阔的应用前景。

外泌体一方面数量较大，另一方面包含了蛋白质、DNA、RNA等多种信息，其在肿瘤检测中适用性更广，作为早期筛查标志物的潜力更大。

4. ctDNA特点

ctDNA主要是来源于死亡的肿瘤细胞，包括坏死、凋亡的肿瘤细胞释放的ctDNA以及进入血液的CTC及肿瘤外泌体释放的ctDNA[9, 10]。

ctDNA具有以下4个特点[11,12]：

(1) 半衰期短，平均的半衰期介于15分钟到2个小时之间。

(2) 浓度低，占总cfDNA的含量<1%。

(3) 片段短，主要片段集中在140bp左右。

(4) 大部分突变频率为低频突变，半数突变频率<0.4%。

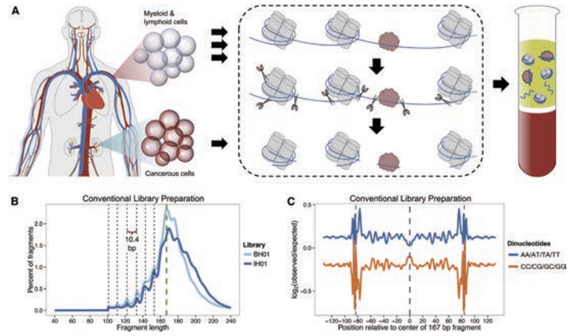


图3 血浆中cfDNA的来源及长度[11]

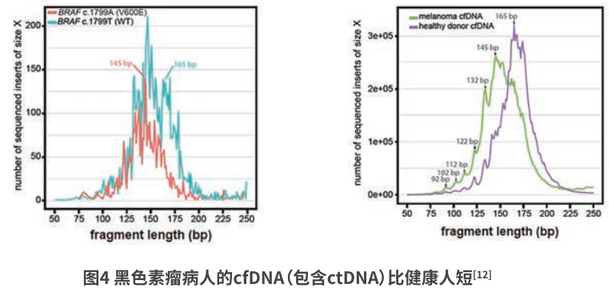


图4 黑色素瘤病人的cfDNA（包含ctDNA）比健康人短[12]

A. 黑色素瘤病人和健康人cfDNA对比，发现黑色素瘤病人cfDNA片段更短（132-145 bp vs. 165 bp）；B. 黑色素瘤病人cfDNA样品中，含有BRAF V600E突变的片段比不含突变的片段更短。

方案设计

A. 研究目标

通过ctDNA高深度测序技术，研究开发应用于肿瘤无创诊断的生物标志物，建立肿瘤高通量无创检测的方法与平台，为肿瘤的精准医疗打下基础。

B. 研究内容

采用BGI TumorCare Panel，或者综合可能的致病基因，设计辅助诊断或者分子分型的panel，以ctDNA作为检测材料，通过超高深度目标区域测序检测低频SNV，并结合低深度全基因组重测序（WGS）检测大片段的结构变异鉴定肿瘤类型，进行早期诊断，或者肿瘤分子分型，指导治疗。

C. 研究方案

收集癌症病人血液样本，分离cfDNA及淋巴细胞DNA。cfDNA加UID建库，然后使用BGI Tumor Care Panel或者重新设计panel，捕获并进行BGISEQ平台测序，深度达到3000x；同时用相同的cfDNA样品构建全基因组测序文库，进行低深度全基因组测序（5X）。淋巴细胞DNA直接用tumor care芯片捕获建Target region capture sequencing文库，BGISEQ平台测序，深度达到3000X。通过数据分析，寻找ctDNA上的已知突变，并与肿瘤组织检测到的突变进行比较，分析在外周血ctDNA中检测突变的敏感性和特异性。

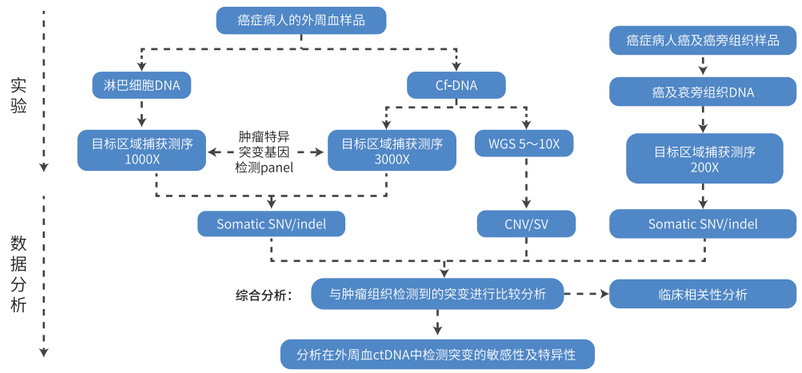


图5 ctDNA液体活检研究思路

D. 样本建议

≥5例 原发癌组织、癌旁组织、外周血、ctDNA。

E. 研究策略

低深度WGS（5~10X）、目标区域测序（组织200X、白细胞1000X、ctDNA 3000X）。

注：此测序深度未考虑duplicaton，实际比对回去，只有1500X左右，因此检测的突变频率>1%，如果要检测更低频突变，需要加大测序深度。

F. 样本要求

1. 病理信息

所有来自病人的样品，必须具备《知情同意书》或相关授权书，必须具备完整的病理报告及相关信息。

2. 样品保存和运输

用标记好Streck Cell-Free DNA BCT采血管，按照外周血标准采集操作采取不少于10 mL外周血，采血后请立即轻柔颠倒10次使血液与管内成分混匀，拖延颠倒混匀时间可能会造成检测失败。混匀后请将采血管直立置于试管架上室温放置（6-35°C）。采集的全血样本存储于6-35°C温度范围，满足从采血起72小时（3日）内抵达华大样本中心，否则不能够保证检测结果的可靠性。

【注意事项】

（1）未采血的Streck Cell-Free DNA BCT储存温度为18-25°C，采血后的储存温度为6-35°C；

（2）禁止将已采血的Streck Cell-Free DNA BCT存放于4°C或以下。

G. Panel选择方案

1. 使用BGI TumorCare Panel：TumorCare Cancer Panel是由华大基因研究院开发，基于NimbleGene探针杂交捕获测序技术，包含1053个肿瘤相关基因，其中176个基因是有明确用药解读的基因，255个基因是在COSMIC数据库中反复出现的基因，622个基因是在癌症重要通路中出现的基因。

2. 根据已发表基因组研究设计panel

设计标准：

- 高频突变基因和机理研究中比较重要的pathway基因。
- 根据TCGA、ICGC等数据集及已经发表的特定肿瘤基因组研究数据集挑选基因，同时满足基因在样品的履盖度较高的要求。

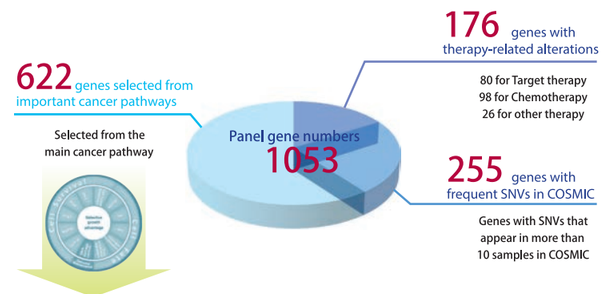


图6 BGI TumorCare Cancer Panel包含的基因

H. 数据分析方案

1. 肿瘤组织突变分析流程

使用CSAP(Cancer Sequencing Analysis Pipeline)流程，进行BWA比对，*samtools*，*GATK*，*Varscan*，*Crest*等软件分析得到*somatic SNV*，*indel*，*SV*和*CNV*等位点，再通过Annovar注释得到突变的详细信息。该结果将与ctDNA分析结果进行综合比较，分析ctDNA检测突变。

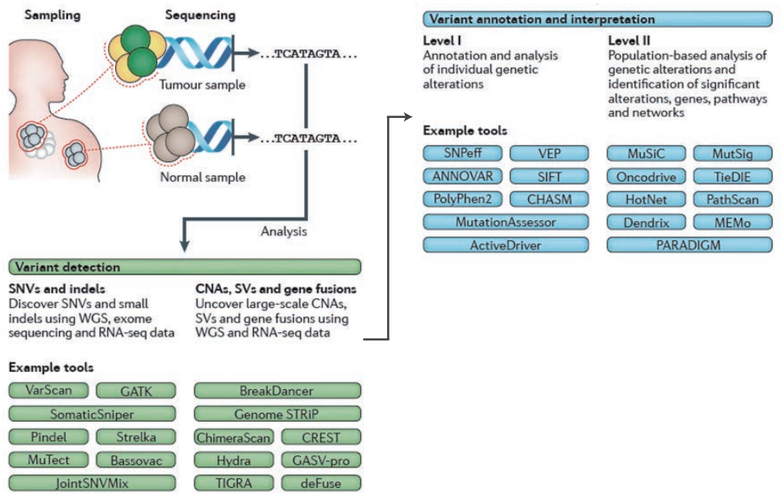


图7 肿瘤组织突变分析流程可选软件

2. ctDNA突变检测流程

（1）UID分析去重复序列及随机错误 —— UID（unique identitor）是一段独特的序列，同一个UID家族95%以上序列含有相同的突变，则认为该突变是真实存在的。通过UID的识别能降低PCR和测序错误对突变鉴定的影响，提高检测结果的可信度。

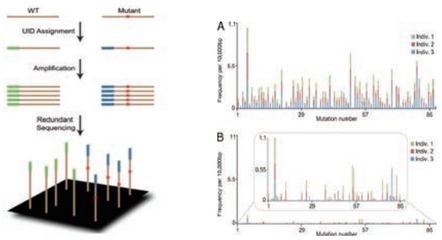


图8 ctDNA加入UID检测原理及检测灵敏度[3]

（2）ctDNA突变信号检测 —— 使用华大ctDNA变异检测流程BGI-in house pipeline，通过对不同大小的DNA片段进行分层比对，最大限度地检测到ctDNA上发生的突变信号，并通过后期的数据过滤和较正，保证检测结果的高度准确性。

- 突变检测的灵敏度高：能够检出VAF (突变等位基因频率)低至0.2%的点突变，其中VAF为1%的点突变检出率为100%；
- 突变检测的准确率高：标准品技术重复显示检出突变一致性至少70%以上。

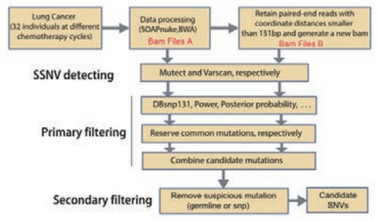


图9 ctDNA变异检测流程

通过对不同大小的DNA片段进行分层比对，最大限度地检测到ctDNA上发生的突变信号，并通过后期的数据过滤和较正，提高结果的准确性。

I. 项目执行周期

样品检测合格后，建库+测序+标准信息分析：约40个工作日，高级信息分析约1-1.5个月，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。

华大优势

案例一：ctDNA的目标区域高深度测序案例

2014年，Newman基于超高深度肺癌测序数据开发了一套可应用于ctDNA突变检测的流程，CAPP-Seq。CAPP-Seq先对已发表的肺癌大样本测序数据进行分析，建立覆盖大部分非小细胞肺癌突变的基因集，该基因集覆盖96%以上的非小细胞肺癌。

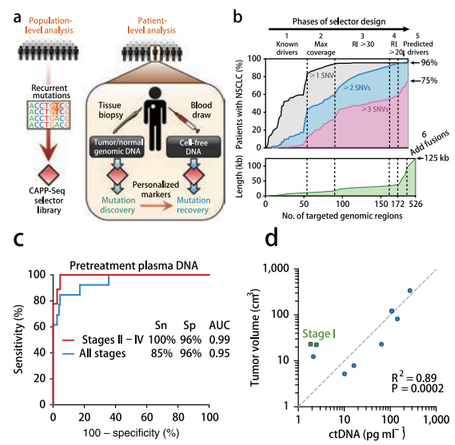


图10 CAPP-seq方法设计图示及检测灵敏度和特异性

a-b. 该研究设计selector覆盖大部分的非小细胞肺癌驱动基因，然后对90个非小细胞肺癌（NSCLC）样本，根据先前设计的selector进行目标区域捕获及高深度测序；c.运用该套方法，在II到IV期的NSCLC中检测到肿瘤突变信号的敏感性和特异性达到100%和96%；d. 发现ctDNA的水平与肿瘤体积显著相关。

案例二：ctDNA的全外显子测序案例

2017年3月，华大基因，国家基因库和复旦大学共同开发了克服肝癌瘤内异质性的测序方法，以检测肝癌的潜在治疗靶标。通过对肝癌进行多点取样，进行外显子测序，从中挑选1200个高质量的突变位点设计目标区域测序芯片，在大样本中进行目标区域捕获测序，然后对肺癌组织全外显子测序，目标区域捕获测序和ctDNA全外显子测序，目标区域捕获测序进行比较。发现通过ctDNA测序可以发现多位点取样外显子测序的64.7%的突变位点，更高深度的目标区域捕获测序可发现组织中83.9%的突变位点。最终在70例肝癌病人中发现了38.6%的具有潜在治疗价值的突变位点。说明了通过ctDNA测序可以成功的克服肝癌瘤内异质性对突变检测结果的影响，成功地检测到肝癌潜在治疗位点[15]。

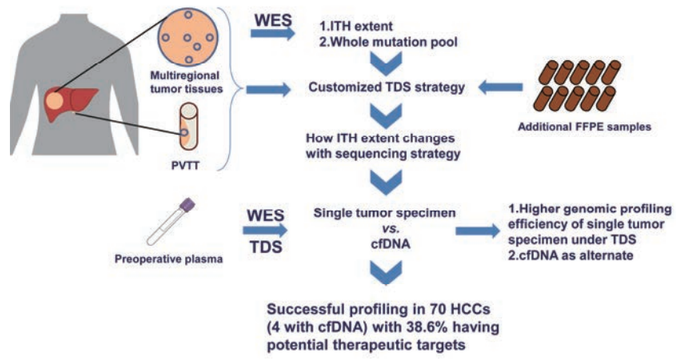


图11 ctDNA加入UID检测原理及检测灵敏度[3]

样品: 5例肝癌病人、多病灶取样（共32个样品），ctDNA样品。

研究方法:

1. 多病灶组织样品-WES（~211X），挑选位点设计目标区域；
2. 组织样本-目标区域测序（~681X）；
3. ctDNA样品-全外显子测序（~226X）和目标区域测序（~1806X）。

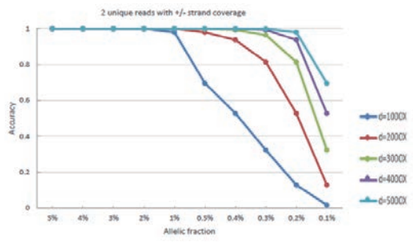
可能存在的风险

ctDNA在病人体内含量与肿瘤分期、肿瘤类型有关，因此可能存在检测不到肿瘤相关突变的情况。另外存在背景噪音干扰，需要通过加大测序深度来增加检测极低频突变的灵敏度。技术上可以通过建库时加UMI（分子标签）的方法，提高检测的灵敏度。

常见问题

1. ctDNA推荐测序深度?

去重后5000x测序深度检测0.1%频率的灵敏度约为70%，去重后4000x测序深度检测0.2%频率的灵敏度约为95%，而cfDNA中SNV突变频率中位值在0.5%，因此推荐去重后的测序深度在2000x以上。



2. 华大是否可提供ctDNA提取?

是，提取量根据ctDNA含量不同有变化，大约是在20ng以上。

3. 华大可提供哪些肿瘤高级分析?

序号	分析条目	分析内容介绍
0	过滤、比对	标准分析
1	SNV & INDEL 检测和注释	标准分析
2	超突变分类	根据 TMB, MSI 状态和 MMR 基因突变状态判断肿瘤是否超突变
3	生殖系突变	筛选肿瘤相关的生殖系突变
4	显著突变基因	1. 识别相对于背景突变率的高频突变基因 2. 识别具有功能偏向性的显著突变基因 3. 识别热点突变基因及突变分布图 4. mutation landscape 绘制
5	药靶注释	使用 CIViC 专业数据库鉴定特定体细胞突变对应的靶向药物，预测对靶向治疗的反应
6	突变特征	1. 提取突变特征并比较不同样品里各种突变特征的贡献度 2. 与机制已知的突变特征进行比较 3. 寻找突变特征活性相关的关键基因突变
7	突变链非对称性（仅WGS）	1. 计算每个肿瘤转录链、复制链偏向性及对应碱基型 2. 比较不同组别肿瘤的转录复制偏向性；
8	新抗原预测	异常蛋白/肽序列检测，并预测与HLA的结合亲和力
9	拷贝数变异	1. 绝对拷贝数变异 2. 拷贝数中性杂合性缺失 3. 得到 allele-specific 拷贝数，分析等位基因不平衡
10	显著拷贝数变异	得到显著扩增或缺失的区域及对应基因
11	克隆进化	1. 计算每个肿瘤体细胞突变的主克隆和亚克隆比例 2. 估计肿瘤样品的纯度和倍性 3. 推测肿瘤关键基因如驱动基因和抑癌基因在肿瘤发生发展过程中扮演的角色 4. 对多位点取样病人构建肿瘤进化模型
12	突变网络	1. 突变互斥网络 2. 突变共生网络
13	分子分型	1. 根据点突变、拷贝数变异等分子特征对肿瘤进行分型 2. 结合临床预后数据比较不同组别生存差异
14	TCGA 注释	提供 TCGA 数据库里感兴趣基因的突变及拷贝数变异频率
15	结构变异（仅 WGS）	1. 识别每个样品里的结构变异及相应机制 2. 绘制每个样品的 sv,cnv,snv circos 图
16	SV 特征（仅 WGS）	提取结构变异特征，并解析其代表着不同的潜在机制

华大优势

更全面的肿瘤分析流程：分析内容不仅限于体细胞突变，还包括CNV、SV、突变特征、突变网络、肿瘤新抗原识别、分子分型等分析。

全新开发的ctDNA分析流程，灵敏度高、准确度高，更适合极低频率突变的检测：

- 突 变检测的灵敏度高：能够检出VAF (突变等位基因频率)低至0.2%的点突变，其中VAF为1%的点突变检出率为100%；
- 突变检测的准确率高：标准品技术重复显示检出突变一致性至少70%以上。

专业的肿瘤分析团队：团队具有丰富的肿瘤基因组学分析经验，在国际顶级期刊发表文章30余篇，依托于华大基因在肿瘤领域长期积累的经验，为全行业提供方案设计、测序服务、数据分析、平台建设和技术优化等全面的解决方案。

参考文献

[1] Mandel P. Les acides nucleiques du plasma sanguin chez l'homme. CR Acad Sci Paris. 1948; 142: 241-3.

[2] Leon S, Shapiro B, Sklaroff D, Yaros M. Free DNA in the serum of cancer patients and the effect of therapy. Cancer research. 1977; 37: 646-50.

[3] Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology. 1989; 46: 318-22.

[4] Vasioukhin V, Anker P, Maurice P, Lyautey J, Lederrey C, Stroun M. Point mutations of the N‐ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. British journal of haematology. 1994; 86: 774-9.

[5] Sorenson GD, Pribish DM, Valone FH, Memoli VA, Bzik DJ, Yao S-L. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiology and Prevention Biomarkers. 1994; 3: 67-71.

[6] Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nature cell biology. 2007; 9: 654.

[7] Properzi F, Logozzi M, Fais S. Exosomes: the future of biomarkers in medicine. Biomarkers. 2013; 7: 769-78.

[8] Thakur BK, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B, Zheng Y, Hoshino A, Brazier H, Xiang J. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell research. 2014; 24: 766.

[9] Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nature reviews Cancer. 2011; 11: 426.

[10] Akca H, Demiray A, Yaren A, Bir F, Koseler A, Iwakawa R, Bagci G, Yokota J. Utility of serum DNA and pyrosequencing for the detection of EGFR mutations in non-small cell lung cancer. Cancer genetics. 2013; 206: 73-80.

[11] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016; 164: 57-68.

[12] Underhill H R, Kitzman J O, Hellwig S, et al. Fragment Length of Circulating Tumor DNA[J]. PLoS Genet, 2016, 12(7): e1006162.

[13] Kinde I, Wu J, Papadopoulos N, et al. Detection and quantification of rare mutations with massively parallel sequencing[J]. Proceedings of the National Academy of Sciences, 2011, 108(23): 9530-9535.

[14] Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature medicine. 2014; 20: 548-54.

[15] Huang A, Zhao X, Yang X-R, Li F-Q, Zhou X-L, Wu K, Zhang X, Sun Q-M, Cao Y, Zhu H-M. Circumventing intratumoral heterogeneity to identify potential therapeutic targets in hepatocellular carcinoma. Journal of hepatology. 2017.

外泌体非编码RNA疾病生物标记物研究方案

研究背景

生物标记研究背景

生物学标记研究起源于1950年，直至1998年，NIH对生物学标记进行标准定义：生物学标记是一种被客观测量和评价的特征，可作为正常的生物过程、致病过程或药物反应的指标，生物学标记具有敏感性、专一性、稳定性、可重复性等特点。

寻找新的生物标记物用于疾病的早期诊断和预后观察等更显得十分迫切。由于基于组织的采样偏差，往往导致其分子生物学特征仅仅是真实情况的掠影，并且重复性差，所以寻找稳定的、与对应疾病高度匹配的循环系统中的有效生物标记物，对疾病分子病理学和临床治疗研究具有重要的意义。

外泌体研究背景

2013年，外泌体因诺贝尔医学奖而被众人知晓，也因其作为生命信息传递者，在体液中广泛存在及易获得性等特点被誉为液体活检“新贵”，成为疾病的精确诊断和治疗及预后研究的热点，逐渐从小众研究走进大众视野。

外泌体如何定义的呢？外泌体（Exosomes）是从细胞质膜上形成、具有内吞作用的膜泡，属于细胞外囊泡（EVs）的一种，可以从管腔内释放到细胞外，直径在30-100nm。血浆、尿液、唾液、腹水、脑脊液、羊水等多种体液中均有外泌体的存在。外泌体会选择性地包容细胞内蛋白质、DNA、信使RNA、非编码RNA等多种活性物质，体内成像证实外泌体体内传递的广泛性[1]，在迁移过程中传递遗传信息，不仅可以传递激活免疫反应的脂多糖[2]，引起宿主贫血症的蛋白质[3]，还可以传递抗药作用的lncRNA[4]，调控骨重塑机制的miRNA[5]等。在疾病研究领域，外泌体参与多个系统的病理生理过程，如凝血、血管渗漏和基质受体细胞的重编程等。在临床应用上，可作为疾病进展的生物标记物和新的治疗靶标之一，用于预测和预防疾病等[6]。已研究发现外泌体的miRNA、lncRNA可作为多种疾病诊断和预后的生物标记物，应用前景可观。外泌体非编码RNA作为biomarker研究进展的发表文章，呈现逐年上升趋势。

所以如何合理设计实验方案，实现疾病发生发展过程中的外泌体非编码RNA biomarker鉴定研究，将成为疾病研究液体活检，实现疾病早筛的重要基础。

方案设计

研究目的

通过外泌体包容miRNA/lncRNA鉴定和表达量研究，以及与病患预后关联分析研究，挖掘到可用于疾病初筛和预后监控的潜在biomarker。

研究思路

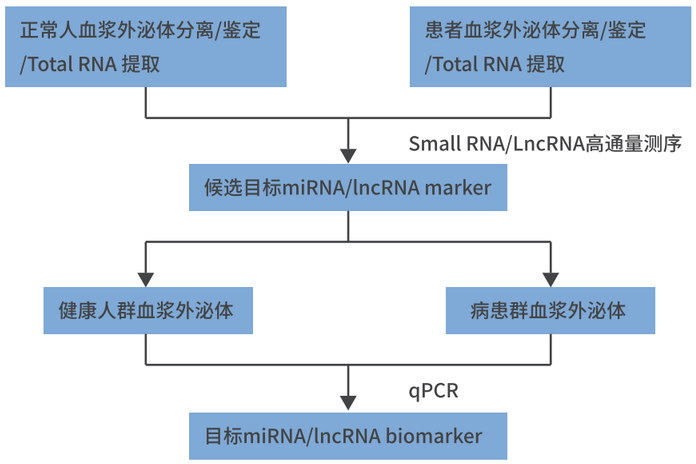


图1 方案研究思路

样本选择

选择10对左右疾病患者和对照的血浆样本，进行外泌体分离、鉴定、Total RNA提取。

文库构建

UMI Small RNA文库构建，SE50测序，推荐20M clean Reads。

LncRNA文库构建，PE100测序，推荐10G clean data。

鉴定筛选目标非编码RNA

a) 非编码RNA鉴定

根据数据库信息比对，鉴定表达的lncRNA、miRNA信息，确定其在基因组位置信息。

b) 非编码RNA表达及差异分析

与正常组织比对后，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或前15的非编码RNA进行目的性分析。

c) 非编码RNA调控机制研究

将非编码RNA与基因进行靶向分析，结合生物学现象，确定其调控作用机制。

功能验证

a) 定量验证：qPCR/茎环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

大样本量验证

a) 利用qPCR结果及表型结合验证

b) 后期大样本量验证可选择100对及以上（实验应当包含足够数量的样品，才能充分代表每种疾病亚型，以建立统计学可信度。多名患者中共有非编码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。）

方案设计注意事项

样本选择方面：

前期表型及相应生化指标等测量要精准；

患者样本-需要严格区分疾病亚型，以便指导血液采集；

设置生物学重复；

血清血浆、细胞上清液样本收集建议参考推荐送样说明；

数据分析及验证方面：

推荐UMI Small RNA测序，提高检测灵敏度；

必须在大的、独立的群体中验证。

应用案例

案例一：血浆外泌体miRNA测序筛选胃癌转移的生物标记物[7]

发表期刊：*Carcinogenesis*

研究概要：

胃癌是最致命的恶性肿瘤之一，复发率和死亡率高。大多数患者在疾病晚期才被诊断出患有胃癌。因此，迫切需要开发胃癌诊断技术以及胃癌转移的新指标。由多种细胞类型分泌的外泌体在细胞间通讯中起关键作用，有希望成为胃癌的诊断生物标志物。

本研究首次利用高通量测序技术从胃癌患者血浆中提取的外泌体中鉴定small RNA谱图，重点研究与转移相关的生物标记物。通过Western印迹，透射电子显微镜（TEM）和纳米颗粒跟踪分析（NTA）对外泌体进行表征。通过生物信息学分析，提出了三种候选物作为胃癌转移的生物标记物，即miR-10b-5p用于胃癌的淋巴结转移，miR-101-3p用于胃癌卵巢转移，miR-143-5p用于胃癌肝转移。并利用RT-qPCR加以验证。综上，成功地从胃癌患者血浆中分离和纯化了外泌体，并鉴定了三种潜在的外泌体miRNA标记物，用来区分患有各种转移的胃癌患者。

研究思路：

图2 研究思路

研究结果：

1、利用维恩图筛选每个组别中共有的miRNA。

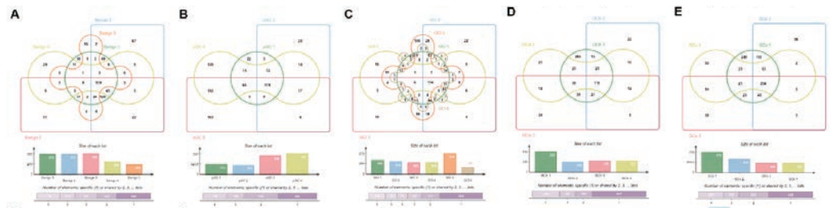


图3 每组中miRNA表达维恩图

A.健康组中miRNA表达维恩图；B. 原发性胃癌组中miRNA表达维恩图；C.胃癌肝转移组中miRNA表达维恩图；D. 胃癌淋巴结转移组中miRNA表达维恩图；E. 胃癌卵巢转移组中miRNA表达维恩图

2、筛选原发与转移组中差异表达的miRNA，选择上下调差异最大的miRNA进行后续实验（p<0.05，FC=1）。

图4差异表达小RNA（DES）的火山图

A.原发性胃癌和肝转移胃癌间的DES火山图；B.原发性胃癌和淋巴结转移胃癌间的DES火山图；C.原发性胃癌和卵巢转移胃癌间的DES火山图

3、筛选可能的候选生物标记物，miR-10b-5p用于胃癌的淋巴结转移，miR-101-3p用于胃癌卵巢转移，miR-143-5p用于胃癌肝转移。收集108名胃癌患者血液，利用qPCR进行标记物扩大验证。

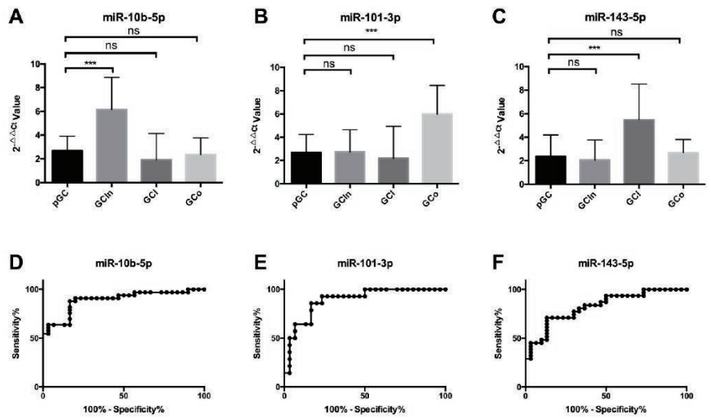


图5 胃癌转移候选标记的miRNA表达

A.miR-10b-5p在原发性胃癌，肝转移胃癌，淋巴结转移胃癌和卵巢转移胃癌中的miRNA表达；B.miR-101b-3p在原发性胃癌，肝转移胃癌，淋巴结转移胃癌和卵巢转移胃癌中的miRNA表达；C.miR-143b-5p在原发性胃癌，肝转移胃癌，淋巴结转移胃癌和卵巢转移胃癌中的miRNA表达。D-F.选定作为生物标记的外泌体miRNA的ROC曲线测定。

案例二：血浆外泌体miRNA-122-5p和miR-300-3p作为大鼠短暂性脑缺血的潜在标记物[8]

发表期刊：*Frontiers in Aging Neuroscience*

研究概要：

利用溶栓时间窗较难判别缺血性中风中的短暂性脑缺血发作（TIA），最新的成像技术既复杂又昂贵。因此考虑研究TIA的血浆标志物，且外泌体衍生的miRNA标记物未知。脑动脉闭塞（MCAo）5分钟、10分钟和2小时使大鼠局部脑缺血。通过深度测序和qRT-PCR鉴定脑脊液（CSF）和血浆外泌体miRNA的趋势一致。与对照和5分钟血浆相比，10分钟缺血大鼠中血浆外泌体rno-miR-122-5p显著下调。与对照组，10分钟和2小时大鼠相比，5分钟缺血大鼠的血浆外泌体rno-miR-300-3p显著上调。利用ROC生存曲线分析验证，评估miRNA对大鼠TIA诊断的准确性。因此血浆外泌体中rno-miR-122-5p和rno-miR-300-3p可能是基于血液的TIA生物标志物。

研究思路：

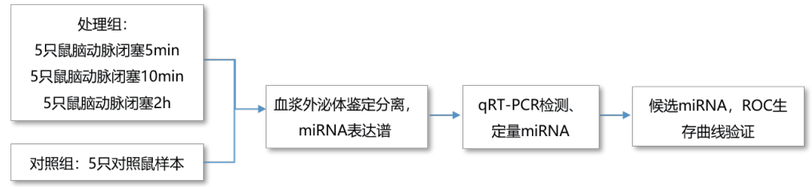


图6 研究思路

研究结果：

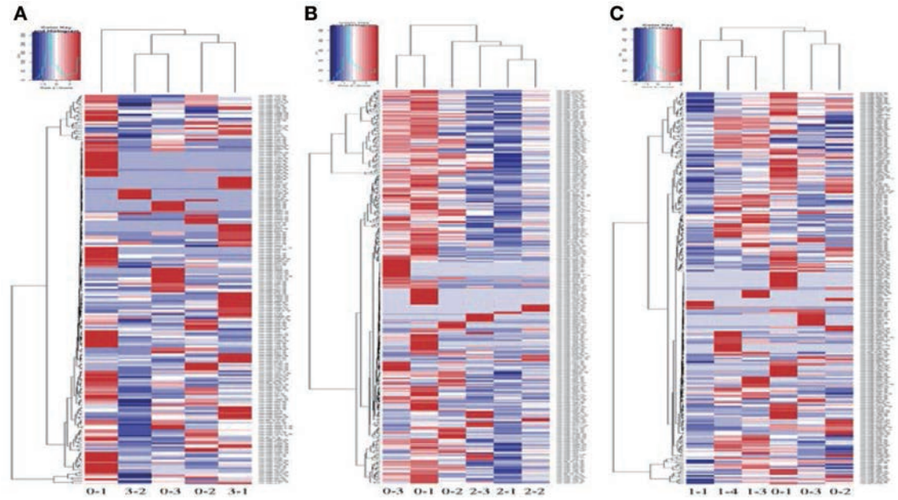


图7 差异表达的miRNAs聚类热图

A.脑动脉闭塞5分钟；B.脑动脉闭塞10分钟；C.脑动脉闭塞2小时

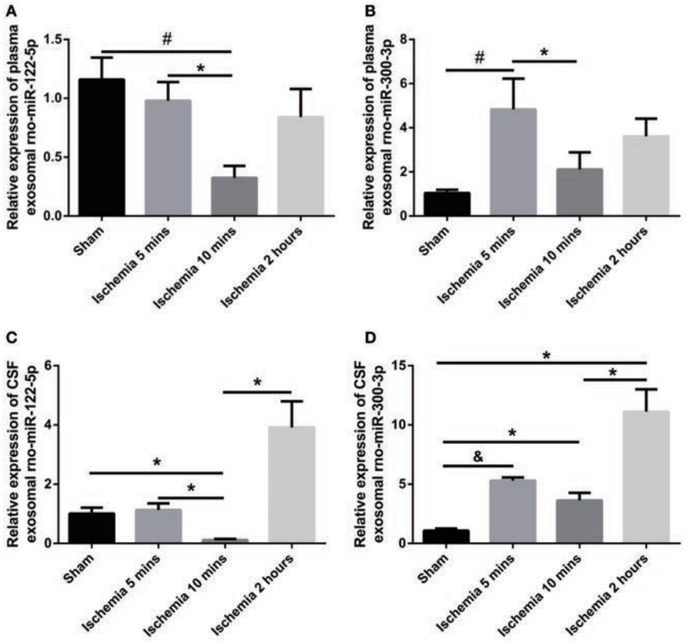


图8 脑动脉闭塞处理和对照组中rno-miR-122-5p、rno-miR-300-3p表达水平比较

AB.血浆外泌体中rno-miR-122-5p和rno-miR-300-3p的相对表达；
CD.脑脊液外泌体中rno-miR-122-5p和rno-miR-300-3p的相对表达。

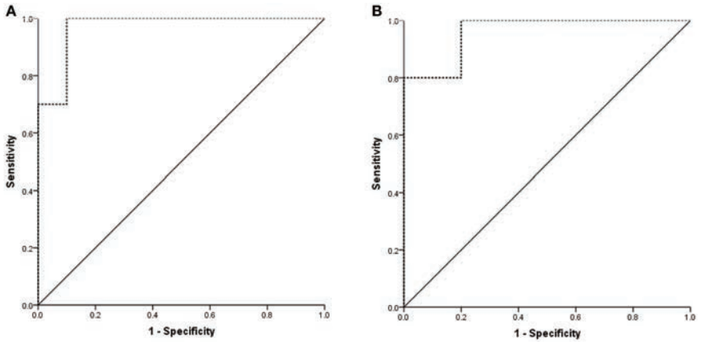


图9 血浆外泌体rno-miR-122-5p和rno-miR-300-3p对大鼠的诊断价值

可能存在的风险

1）外泌体分离鉴定结果出现假阳性，包含其他外囊泡。

2）外泌体提取的Total RNA 中小片段的RNA含量较高，经高通量测序检测到的miRNA有可能是mRNA和LncRNA降解片段，结果存在假阳性。

3）在样品选择的过程中，如若未注意患者在不同时期的特征，因外泌体分泌等具有时空性，所以需要采集时间梯度较密集的原则进行选取样本，使得检测分析结果具有代表性。

4）在样本准备过程中，因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存在。

5）初期筛选biomarker的样本数建议选取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。

6）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。

常见问题

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。

2、血清血浆哪种样本类型更适合分离外泌体？

答：目前文献中，多数是研究血浆外泌体RNA；而且处理血液样本时，血小板在凝血过程中会分泌大量外泌体到血清。所以研究非血小板相关疾病，建议优先选择血浆样本。但针对不同研究目的，还请结合实际需求选择样本类型。

3、血浆样本一般建议如何收集送样？

答：一般建议选取EDTA抗凝采血管或柠檬酸盐抗凝采血管（不可用肝素管），建议采取10mL全血，进行两步离心法操作处理，送4管（1.2ml）已处理的血浆即可。

4、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，可选择高表达的排名TOP10进行验证。

华大优势

服务全面。提供外泌体分离、鉴定、提取、Small RNA测序、LncRNA测序一站式解决方案。分离、鉴定方法多种可选。鉴定包括透射电镜形态学鉴定、粒径分布分析、Westen Blot 蛋白标志物鉴定。

经验丰富。适用于血浆、血清等体液样品和细胞培养上清液样品。已执行上千个项目，发表多篇文章。

技术领先。UMI Small RNA建库，定量精准，1ng低起始量，更多有效数据，成功率高。

分析无忧。独特的Dr.Tom多组学数据挖掘系统交付。数据图表循环挖掘，多维度结果图展示，10大注释数据库，12种分析小工具，多组学关联分析，互作网络可视化，随时更新文献信息，查基因得文献，便于文章撰写。

参考文献

[1]Zomer A, Maynard C, et al. In Vivo imaging reveals extracellular vesicle-mediated phenocopying of metastatic behavior. *Cell*. 2015 May 21;161(5):1046-57.

[2] Vanaja S K, Russo A J, et al. Bacterial Outer Membrane Vesicles Mediate Cytosolic Localization of LPS and Caspase-11 Activation. *Cell*. 2016 May 19;165(5):1106-19.

[3] Szempruch, A. J., et al. (2016). Extracellular Vesicles from Trypanosoma brucei Mediate Virulence Factor Transfer and Cause Host Anemia. *Cell*. 2016 Jan 14;164(1-2):246-57.

[4] Qu L et al. Exosome-Transmitted lncARSR Promotes Sunitinib Resistance in Renal Cancer by Acting as a Competing Endogenous RNA. *Cancer Cell*. 2016 May 9;29(5):653-68.

[5] Xie, Y., et al. The roles of bone-derived exosomes and exosomal microRNAs in regulating bone remodelling. *J Cell Mol Med*. 2016 Nov 23.

[6] Becker A, Thakur BK, et al. Extracellular Vesicles in Cancer: Cell-to-Cell Mediators of Metastasis. *Cancer Cell*. 2016 Dec 12;30(6):836-848.

[7] Zhang Yingyi,Han Ting,Feng Dan et al. Screening of non-invasive miRNA biomarker candidates for metastasis of gastric cancer by small RNA sequencing of plasma exosomes.[J] .*Carcinogenesis*, 2019.

[8] Li Dong-Bin,Liu Jing-Li,Wang Wei et al. Plasma Exosomal miRNA-122-5p and miR-300-3p as Potential Markers for Transient Ischaemic Attack in Rats.[J] .*Front Aging Neurosci*, 2018, 10: 24.

肿瘤发生的关键转录因子及其靶基因多组学研究方案

研究背景

原癌基因的激活和抑癌基因的失活主要是因为基因变异，从而导致了癌症的发生。真核生物基因表达是一个复杂而有序的过程，它是众多反式作用因子和顺式作用元件之间相互作用的结果，反式作用因子是指能直接或间接识别和结合在顺式作用元件上，调控靶基因表达的蛋白质因子，一般也称为转录因子（transcriptional factor,TF），转录因子结合位点((Transcription factor binding site, TFBS)是与转录因子结合的DNA 序列。确定TFBS是理解转录调控机制，建立转录调控网络的关键问题。转录因子结合位点（Transcription factor binding site，TFBS）是与转录因子结合的DNA片断，长度通常在5~20 bp范围内，一个转录因子往往同时调控若干个基因，而它在不同基因上的结合位点具有一定的保守性，又不完全相同，ChIP-Seq应用于转录因子结合位点研究和组蛋白修饰研究，一次ChIP-Seq能找到成千上万个结合位点，而哪些基因才是它真正作用的区域?它的生物学功能又是什么?类似的组蛋白的各种翻译后修饰（即甲基化，乙酰化，泛素化等）发生在氨基酸位点 。许多组蛋白修饰能够调节染色质结构中重要的功能性变化，并通过直接地改变染色质结构、动态变化或通过招募组蛋白修饰蛋白或核小体改构复合体来实现 。组蛋白翻译后修饰已经被发现能影响许多基于染色质的反应，包括转录、异染色质的基因沉默和基因组稳定性 。由于能影响到整个转录程序，与基因表达相关的修饰尤其具有特殊意义。在多种体外模型中，组蛋白翻译后修饰代谢途径的缺陷与基因表达失调有关。这在某些情况下也与人类疾病相关，并已在免疫缺陷和各种人类癌症中得到验证 。因此，转录因子结合位点与组蛋白修饰具体参与肿瘤变异调控哪一些基因，它参与的生物学功能及分子通路到底有哪些，我们可联合转录组数据加以佐证。

方案设计

主要结合表观组学数据和转录组学数据，对case和control进行数据收集，并在特定变量下对case和control进行差异表观修饰和差异表达量进行统计、聚类、分组，对显著差异基因进行关联分析，筛选出关键目的基因和关键蛋白，并进行相关验证工作，具体方案示意图如下：

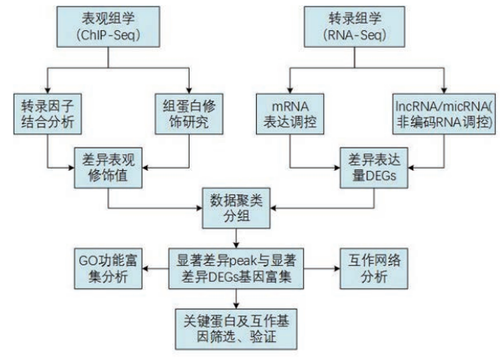


图1 方案研究思路

A. 样本选择

1. 建议选择成对的癌/癌旁组织样品，样本数不少于5对，每组至少2-3个生物学重复；
2. 分别对两个样本在相同实验条件下进行ChIP-Seq和RNA-Seq；
3. ChIP-Seq推荐使用input对照去除背景噪音；
4. ChIP-Seq推荐测序数据量不低于20M clean reads，RNA-Seq推荐测序数据量不低于20M clean reads。

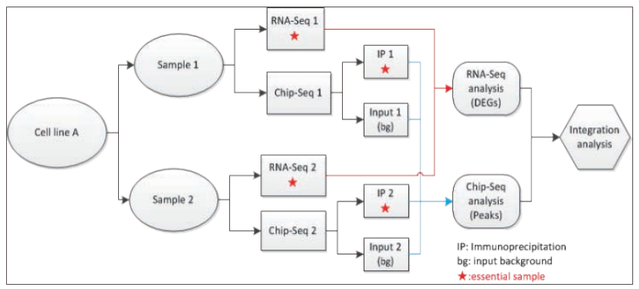


图2 样品选择建议

IP, Immunoprecipitation，免疫共沉淀

B. 采用的技术

采用特异性抗体对目的蛋白进行免疫沉淀后，分离与其结合的基因组DNA片段，再通过高通量测序与数据分析，在全基因组范围内寻找目的蛋白的DNA结合位点，并且可以基于多个样品进行差异比较。染色质免疫共沉淀（ChIP）是在体内环境中研究蛋白质与DNA相互作用的经典实验方法，广泛应用于组蛋白修饰、特定转录因子的基因调控作用等相关领域。随着新一代测序技术的发展和成熟，染色质免疫沉淀实验与高通量测序的整合——Chromatin Immunoprecipitation Sequencing（ChIP Sequencing），可在全基因组范围对蛋白结合位点进行高效而准确的筛选与鉴定，同时也为研究的深入开展打下基础；RNA-Seq是对某一物种或特定细胞在某一功能状态下产生的mRNA进行高通量测序，可以提供定量分析，检测基因表达水平差异。

C. 测序参数

建议ChIP-Seq推荐测序数据量不低于20M clean reads，RNA-Seq推荐测序数据量不低于20M clean reads。

D. 分析结果

1. 不同表达水平基因表观修饰值分析

通过与gene及功能区进行定位，可以把基因分为body区有 Peak Peak的基因、promoter区有 Peak Peak的基因、无Peak Peak基因。分析这3类基因在RNA-Seq中的表达水平，可以研究蛋白对基因表达的调控趋势。

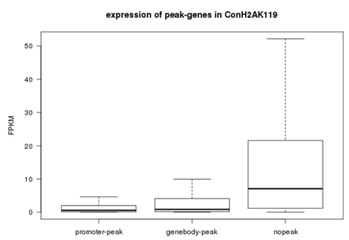


图3 不同表达水平基因表观修饰值分布图

2. 不同表达水平基因表观修饰值分布

把基因按表达水平进行分类，包括沉默基因（silence, fpkm=0）、低表达基因（low, 0 < fpkm < 10）、高表达基因（high, fpkm > 10），计算这3类基因在 ChIP-Seq数据中reads的丰度信息，以表观修饰值（epI-level）表示，值越大表示该基因越有可能是蛋白结合区且作用能力越强。

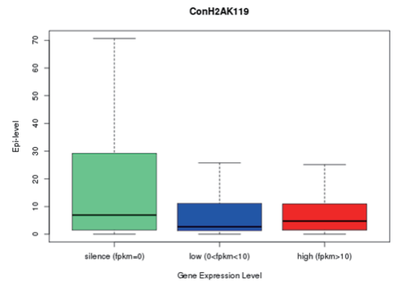


图4 不同表达水平基因表观修饰值分布图

3. Peak信号强度与基因表达水平关系

蛋白与DNA的结合强度不同会对基因造成不同程度的调控作用。基因信号由低到高排序（tag数或累积高度），计算各个gene的表达量分布。

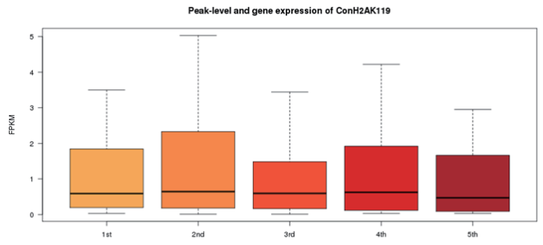


图5 Peak信号强度与基因表达水平关系图

4. 不同表达水平gene在基因上下游2k和body区的ChIP-Seq reads密度

按表达量水平高低把基因分成3类（High: FPKM > 10; medial：1 <= FPKM <= 10; Low: FPKM < 1），计算各样本在基因上下游及body区的 ChIP-Seq reads密度分布趋势。

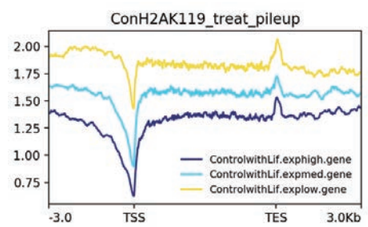


图6 差异表达基因上下游2k和body区的ChIP-Seq reads密度图

5. 基于PPS算法的DEGs和diffpeak分布关系

按为研究转录因子（或组蛋白修饰）对基因的调控关系，我们同时进行了一个样本的Chip-Seq实验和RNA-Seq实验，分别通过测序、计算找到该蛋白在全基因组水平上的结合热点（peak区域），以及实验前后样本转录水平的变化值（差异表达基因，DEGs），再通过联合定义这些peak与DEGs的相关系数，从而发现该蛋白对样本转录水平的影响，依此来得到转录因子（组蛋白修饰）与基因表达之间的。

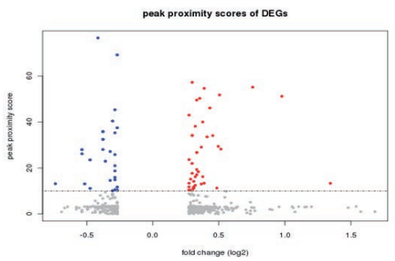


图7 peak proximity scores与DEGs表达差异值关系分布图

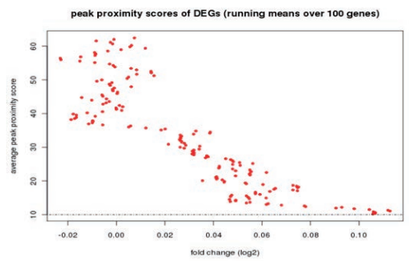


图8 peak proximity scores与DEGs表达差异值趋势图

6. 不同表达水平且出现高表观修饰值基因GO分析及经典分子通路聚类分析

为探索蛋白结合在基因表达过程中起到的作用，我们对上述3类不同表达水平且出现高表观修饰值（>10）的基因进行分析。

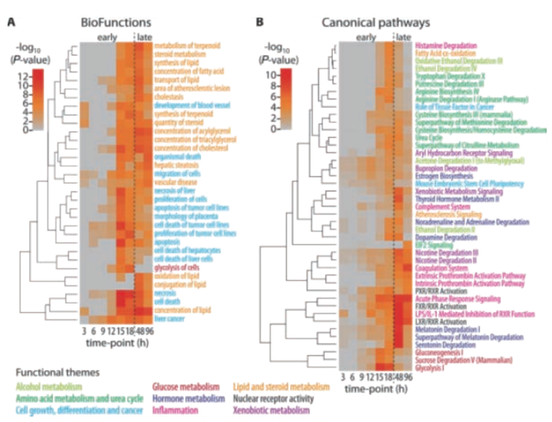


图9 关联分析差异显著富集基因GO分析及经典分子通路聚类分析

E. 项目执行周期

样品检测合格后，建库+测序+标准信息分析约30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。

F. 预期的结果

利用表观遗传学和转录组学技术手段，借助高通量测序平台，通过对不同处理、不同时间点的组织之间的比较，从基因组范围内的转录因子富集度变化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，找到导致肿瘤发生的关键表达蛋白及目的基因，为疾病的治疗提供有效靶标。

我们期望通过case与control的多组学分析，找到有关键基因，并通过对这些基因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。

G. 辅助研究策略

可以通过ATAC-Seq前期对显著富集的转录因子进行初步摸索，再使用ChIP-Seq技术对特定转录因子进行确认，结合基因敲除，腺病毒感染进行功能解读，从而辅助发掘与肿瘤发生、发展以及治疗相关的功能基因。

H. 后期验证手段

分析得到的差异表观修饰值/差异DEGs可以利用Q-PCR进行定量验证，从而获得相对准确的表观修饰值与基因表达量，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。

应用案例

案例一：肝细胞中TCF7L2全基因组靶基因调控机制研究[1]

研究背景：TCF7L2在肝脏中代谢通路中具有重要作用，之前的研究发现TCF7L2在成千上万个基因区域有富集，但具体哪一个才是重要的结合位点有待挖掘。

研究结果：TCF7L2沉默后实时RNA-Seq检测3-96h有406个差异表达基因，进一步结合ChIP-Seq peak相关系数PPS大于10的直接调控Gene有149个，确定TCF7L2参与九大通路调控，其中包括脂质，碳水化合物和氨基酸以及尿素代谢途径等。

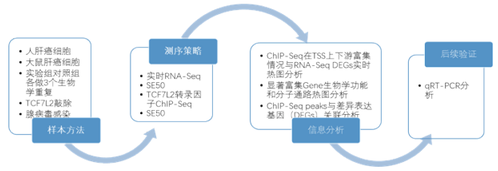


图10 文章多组研究思路

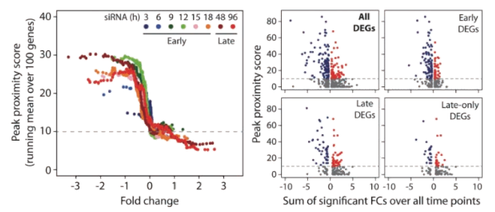


图11 TCF7L2 peak proximity scores与DEGs表达差异值趋势图及不同时期关系分布图

案例二：FOXN3-NEAT1-SIN3A 复合体在乳腺癌发生和转移中的功能研究[2]

案例描述：乳腺癌是人类常见的一种恶性肿瘤，据统计，全球每年新发乳腺癌高达120万人，是女性第一高发的肿瘤，近年来在我国乳腺癌的发病率明显增高，而乳腺癌转移是导致患者死亡的主要原因，乳腺癌转移的分子机理目前还不完全清楚，在以往的研究中观察到FOXN3的在其他恶性肿瘤中表达失调的现象，但FOXN3在恶性肿瘤（包括乳腺癌）发生中的作用及分子机理仍有待研究，它的功能又需要哪些分子共同作用。

研究技术：（BGISEQ-500 ChIP-Seq）染色质免疫共沉淀测序、（iRIP-Seq）RNA免疫共沉淀测序、lncRNA表达谱芯片、CHART-Seq、快速蛋白质液相色谱法（FPLC）等。

研究成果：

1. 揭示了FOXN3-NEAT1-SIN3A阻遏物复合体的存在方式；
2. 全基因组范围内鉴定了FOXN3-NEAT1-SIN3A阻遏物复合体的目标基因；
3. FOXN3-NEAT1-SIN3A阻遏复合体促进乳腺癌细胞的EMT转化和侵袭；
4. FOXN3-NEAT1-SIN3A阻遏物复合体促进乳腺癌的转移；
5. FOXN3和NEAT1在乳腺癌中上调，其高水平与较高的肿瘤分级和较差的存活率相关。

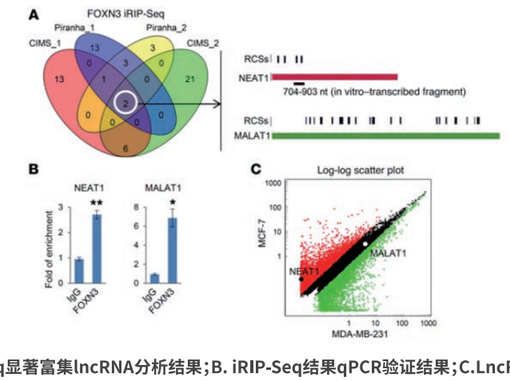


图12 A. FOXN3 iRIP-Seq显著富集lncRNA分析结果；B. iRIP-Seq结果qPCR验证结果；C.LncRNA表达芯片检测散点图

可能存在的风险

在项目实施过程中，可能存在由于样本群体过小、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ChIP显著peak关联的显著差异表达基因，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。

常见问题

1. ChIP-Seq的对照中input和IgG有什么不同？

答：Input对照：在进行免疫沉淀前，需要取一部分断裂后的染色质做Input对照。Input是断裂后的基因组DNA，不加抗体做富集，但是需要与沉淀后的样品DNA一起经过逆转交联，DNA纯化，以及最后的PCR或其他方法检测。Input对照不仅可以验证染色质断裂的效果，还可以根据Input中的靶序列的含量以及染色质沉淀中的靶序列的含量，按照取样比例换算出ChIP的效率，所以Input对照是ChIP实验必不可少的步骤。

阴性对照：用普通的IgG做为抗体（目的蛋白抗体宿主的IgG或血清）。理论上不会ChIP下来任何DNA片段，因此作为阴性对照，但是由于非特异结合，或者实验过程中，没发生结合的DNA清除不完全，可能也会出现条带，如果非常明显，那就证明实验过程有待改进。

2. 哪些因素会影响ChIP-Seq的结果？

答：抗体的质量与特异性、需要富集的目标区域在基因组上的比例、ChIP的实验操作、DNA片段长度范围等都会影响ChIP-Seq的结果。

3. ChIP-Seq与RNA-Seq 测20M的数据量是否足够？

答：从以往的项目经验中，一般是足够的，但也有个别项目，因样本特殊，出现测序数据不饱和的情况，可通过加测补充数据。

华大优势

专项开发：针对多组联合分析项目进行专项开发，建立多组学专项数据库平台，分析模块化，客户可根据感兴趣的产品或基因选择关联模式。

应用灵活：不仅可对表观组学与转录组学开展关联分析，表观组与表观组，转录组与转录组，表观组与基因组等均可灵活切换，根据客户需求选择相关模块展开。

平台全面：全面的测序平台和科研服务产品，保持了充足的数据来源和全面的分析视角。

技术领先：拥有国际领先水平的信息分析团队支撑。

经验丰富：多年的标准化及定制化多组项目执行经验，拥有丰富的多组流程及分析人才。

定制服务：也可根据客户需求进行内容量身定制分析。

完美售后：信息分析人员一对一服务，售后无忧。

参考文献

[1] Luke N., et al. (2014) The mechanisms of genome-wide target gene regulation by TCF7L2 in liver cells. Nucleic Acids Research, 2014, Vol. 42, No. 22 13647.

[2] Yongfeng Shang,Yu Zhang,et al.The FOXN3-NEAT1-SIN3A repressor complex promotes progression of hormonally responsive breast cancer. J Clin Invest. 2017 Aug.

自身免疫性疾病
免疫组库研究方案

研究背景

自身免疫性疾病（autoimmune disease, AID）是指由于某些原因造成免疫系统对自身成分的免疫耐受减低或破坏，致使自身抗体或（和）致敏淋巴细胞损伤自身器官组织而引起的疾病。临床上讲AID分为器官特异性和全身性两类，前者常见的有强直性脊柱炎、I型糖尿病、炎症性肠病、多发性硬化症、自身免疫性肝病等；后者常见的有系统性红斑狼疮、硬皮病、干燥综合症、类风湿性关节炎等。目前绝大多数AID的病因和发病机制尚不清楚，大体存在以下几个方面[1]：

（1）遗传因素是AID发病的基础，例如遗传易感性，AID具有家族聚集性和种族差异；表观遗传调控，部分DNA的CpG岛甲基化增强可能参与I型糖尿病的发病，T、B淋巴细胞的多种基因的低甲基化参与了红斑狼疮的发生发展等，miRNA、环状RNA和lncRNA也逐渐受到关注；

（2）环境因素是AID发病的诱因，例如吸烟与类风湿性关节炎、红斑狼疮发生有关，食物中碘摄入过多，诱发自身免疫性甲状腺疾病，很多病原体如大肠杆菌、EB病毒、巨细胞病毒等感染与AID发生密切相关；

（3）免疫系统是AID发生发展的直接参与者，这种异常主要包括T细胞免疫异常、B细胞免疫异常和固有免疫异常。例如在类风湿性关节炎和多发性硬化症患者中，发现特异性的TCRAV和TCRBV片段的扩张，并发现在受累部位中CD4+和CD8+ T细胞的寡克隆扩张。

免疫组库技术（Immune Repertoire Sequencig）是利用高通量测序，对TCR和BCR的克隆特征和频率进行定量研究，通过该技术可以清楚的了解自身免疫性疾病病人的T细胞和B细胞的异常表达特征，可用于疾病早诊、患病风险预测和疗效监控等。

什么是免疫组库？

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。

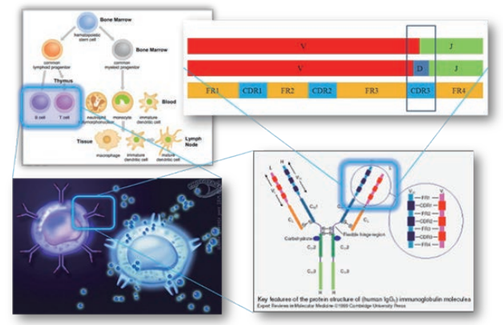


图1 免疫组库研究内容

以B细胞为例，B细胞表面有BCR（B细胞受体），即Y字形的抗体。BCR顶端的区域是CDR区域（抗原互补决定区），分别由V、D、J基因编码，其中CDR1和CDR2是由V基因编码，CDR3是由V(D)J基因编码。免疫组库是通过对编码CDR3/CDR的V(D)J基因进行测序，通过基因频率反映B细胞克隆多样性。

TCR/BCR CDR3多样性是如何实现的？

1、V(D)J recombination重排；

2、V-D和D-J间随机插入碱基；

3、抗体常发生体细胞突变。

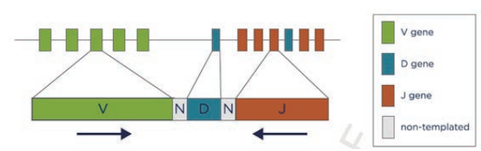


图2 CDR3编码基因的多样性是由V(D)J基因重排加上随机插入碱基产生的

免疫组库在肿瘤领域的研究进展

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等等；医学应用上，疫苗研发评价、药物研发、疾病诊断、器官和肝细胞移植等，发表文章呈逐年上升趋势。

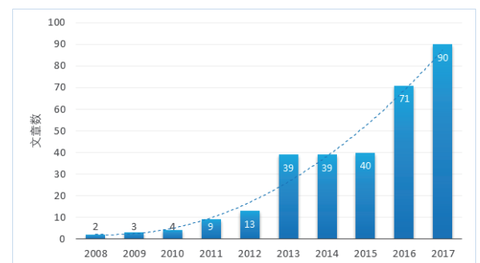


图3 免疫组库已发表文章统计（IF>5文章不完全统计）

在自身免疫性疾病方面，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[3]。Robert Winchester等发现EBV引发的TCR克隆在多发性硬化症病人的脑脊液（CSF）中富集[4]。赫尔辛基大学研究人员对82例新诊断未治疗的类风湿性关节炎病人和20例健康人进行TCR测序、外显子和目标区域测序，在CD8+ T细胞中发现30个新的体细胞突变，RNA-Seq发现这些基因在CD8+ T细胞中高表达，其功能与免疫调控、细胞增殖有关，因此推测CD8+ T细胞的体细胞突变与自身免疫性疾病发病有关。TCR测序发现病人的显著克隆群体在治疗前后保持不变[5]。临床上对类风湿性关节炎的评判指标只有20%的预测准确性，研究人员利用BCR测序数据，建立了患病风险评价模型，该模型预测含有5个及以上BCR主克隆的个体，36个月内发病的风险为83%，而小于5个主克隆的个体发病风险为13%[6]。

表1 已发表自身免疫性疾病和炎症相关疾病

自免病和炎症相关疾病
系统性红斑狼疮SLE
I型糖尿病
类风湿性关节炎
强直性脊柱炎
顽固性乳糜泻
多发性硬化症
白血病（LGL）合并类风湿性关节炎（RA）
自身免疫性糖尿病
Rasmussen脑炎
WAS综合症
多发性肌炎
非病毒相关的肝病
特应性皮炎
胃炎、胃淋巴瘤
系统性硬化症伴有肺动脉高压
原发性硬化性胆囊炎伴炎症性肠病
肺炎

方案设计

A. 研究目的（临床切入点）

1、从不同亚型的角度入手，比如A型和B型临床上很难区分，而且治疗手段不一样，如果有不同的TCR或BCR的克隆特征可以区分，就可以有临床意义。

2、对比不同部位的克隆特征，例如外周血或其他可能取样的部位；对比差异，寻找外周血中能够预测疾病发展或者预后相关的克隆特征（例如V基因、V-J基因的突变特征）。

3、从不同时间点入手，例如治疗前后取样，对比克隆频率变化，找到与疗效相关的biomarker。

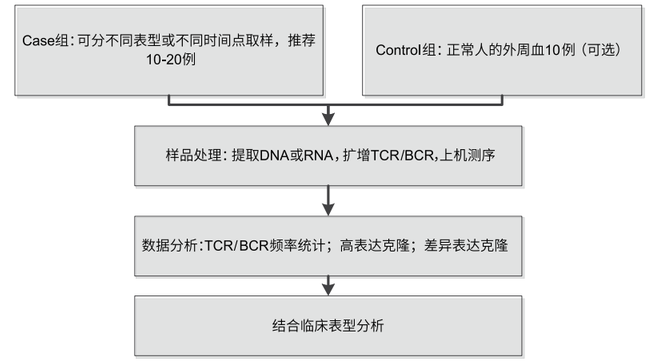


图4 研究思路简图

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，例如不同疾病亚型，或者临床上易混淆疾病，寻找不同疾病诊断的biomarker。

B. 样品选择

每个group 10-20例，分多点取样（时间点、外周血和病灶、不同亚型），对照样品外周血样品10-20例，总计50-100个样品。样品类型根据疾病特征，选择外周血、炎症组织等。

C. 实验技术

多重PCR在BCR或TCR的位于CDR3区两端的V、J基因保守区域设计PCR引物，通过多重PCR扩增得到互补决定区CDR3区域，扩增产物用于后续高通量测序PE151（数据量推荐1Gb raw data）。如果模板为RNA，则需要先进行反转录得到cDNA，再进行多重PCR。

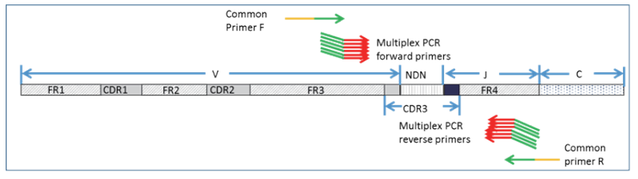


图5 多重PCR示意图

D. 信息分析流程

1）测序所得的数据称为raw reads或raw data，随后要对raw reads进行质控(QC)，以确定测序数据是否适用于后续分析；2）经过滤得到的clean reads比对到参考序列，对于比对上的reads，做下一步的组装，得到具体的功能区域，例如CDR3区（clones）；3）碱基质量符合要求的克隆序列会作为核心克隆（core clonotype），存在一个以上质量值较差碱基的克隆会以核心克隆作参考二次比对和校正；4）然后，对相差一个碱基的克隆，进行层次聚类，每个分支间仅有一个碱基差别（mismatch），依次聚类下去，克隆频率低的克隆会合并到上一分支，最终保留最顶端的head序列；5）将上述得到的克隆序列再次比对到V，D，J和C参考序列，最终得到的统计文件包含了克隆序列、氨基酸残基序列、克隆数量、克隆频率，V/J基因组合等信息。后续可以根据这些信息做克隆分布、基因重组、多样性分析等深入挖掘。

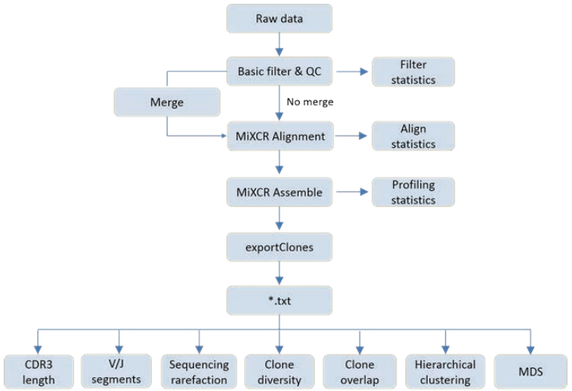


图6 信息分析流程

E. 信息分析内容

1. 基本数据统计

数据过滤，对原始数据进行去除接头污染及低质量reads的处理

数据搭建，数据拼接，消除测序背景及有效数据构建

数据统计，数据产出统计及测序数据的成分和质量评估

2. 数据比对分析

比对分析，与数据库V/D/J基因片段比对

比对结果统计

3. 克隆序列特征注释

CDR3区核酸序列和氨基酸序列

鉴定无效序列（包含终止密码子，超出结构范围）

鉴定单碱基突变（替换、删除、插入）（for BCR）

4. 单样品克隆群体特征分析

CDR3序列长度分布

V/J基因频率分布

V-J基因组合频率分布（3D,Circos）

克隆群体结构分析（频率分布，D50曲线，甜甜圈图）

5. 样品间比较分析

测序饱和度分析

克隆多样性分析(辛普森系数、香农威纳系数等)

样品间共有克隆分析

聚类分析（层次聚类，MDS聚类）

组间差异分析

F. 部分分析结果展示

1. V-J基因频率

针对测序数据结果序列，使用IMGT数据库进行比对，鉴定出V、D、J基因，并对样本中所有克隆的V基因、J基因、V-J基因组合形式进行了统计，以每种克隆reads数计算权重，V-J组合结果以3D和Circos图分别展示。

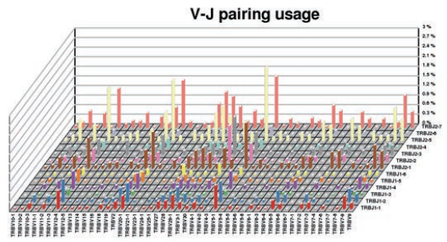


图7 V-J基因组合频率3D柱状图

平面上分别为V基因、J基因。柱子的高度代表一种V-J组合的频率。

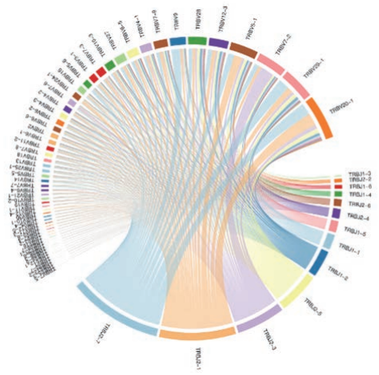


图8 V-J基因组合频率Circos图

每个颜色块代表一种基因，颜色块越宽，频率越高。色块间的连线代表一种V-J基因组合方式。

2. 克隆多样性分析

克隆多样性统计，是不同于V-J基因频率的统计。V-J基因会存在SNP（BCR存在超突变）、随机碱基插入等，增加了克隆的多样性。

样品克隆频率分布图直观反映每个样本中所有克隆类型频率分布情况，D50是近年来引入反映样本克隆群体结构的一个指标，值越低，反映克隆多样性越低，值越大，克隆多样性越高。

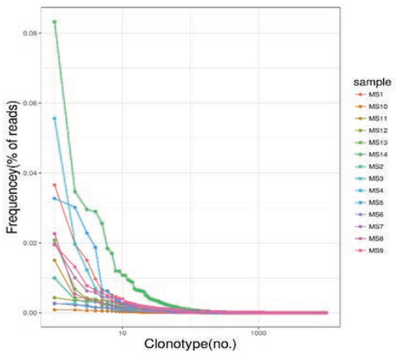


图9 克隆频率分布图

横纵坐标分别为克隆数和克隆频率。

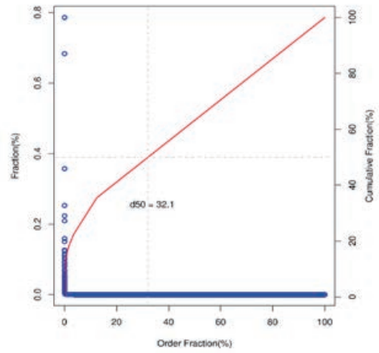


图10 D50曲线

X轴表示样本克隆组成累积百分比，左侧Y轴表示单个类型克隆频率，右侧Y轴表示克隆频率累计百分比。每个点表示单个克隆具体的频率，曲线为所有克隆的累计分布。其中的D50为累计频率达到50%时的克隆所在位置。

3. 组间差异分析

分组比较克隆多样性及top20高频表达V基因频率。

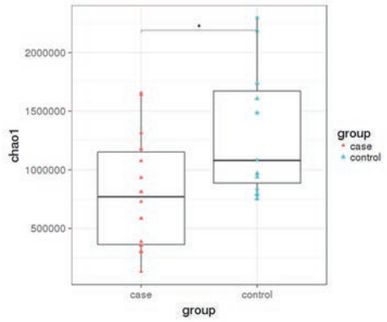


图11 克隆多样性箱线图

每个箱线代表一个group，每个箱线图对应五个统计量（自上而下分别为最大值，上四分位数，中值，下四分位数和最小值）。使用Student's t-Test进行差异显著性检验，其中ns表示差异不显著（P>0.05）；*表示有统计学差异（P<0.05）；**表示有显著统计学差异（P<0.01）；***表示有极其显著的统计学差异（P<0.001）。

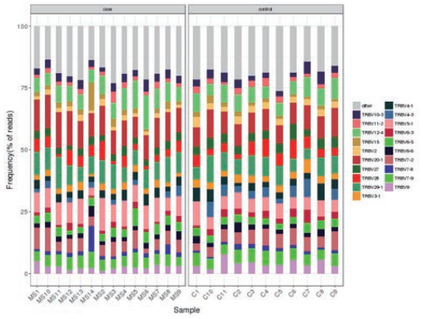


图12 Top20 V基因频率组间分布柱状图

X轴表示样品编号，Y轴表示重组结果中各基因的使用频率。

应用案例

案例一：杀伤T细胞的突变积累可能与类风湿性关节炎发病有关[5]

发表期刊：Nature Communications

影响因子：12.12

发表时间：2017年6月

研究目的：大颗粒淋巴细胞白血病（LGL）常合并发生类风湿性关节炎（RA），研究发现LGL病人的CD8+ T细胞克隆存在STAT3突变，而拥有该突变的LGL病人更容易并发RA（43% vs. 6%）。因此推测，发生体细胞突变的CD8+ T细胞可能与RA发病有关。

研究样本：82例新诊断未治疗的RA病人，20例健康人，取外周血。

研究结果：首次采用CD4+ T细胞做对照，过滤生殖系突变（germline mutations），对分选的T细胞进行目标区域测序、外显子测序和TCR β链CDR3测序。发现，大约20%（5/25）病人的CD8+ T细胞有体细胞突变（somatic mutations），而CD4+ T细胞没有，RNA-Seq确认了突变基因在CD8+ T细胞中高表达。功能分析发现，这些基因突变与免疫调控、细胞增殖有关，因此推测，发生体细胞突变的CD8+ T细胞可能与RA及其他自身免疫性疾病的发病有关。TCRβ链CDR3测序发现显著克隆在免疫治疗前、后保持不变。

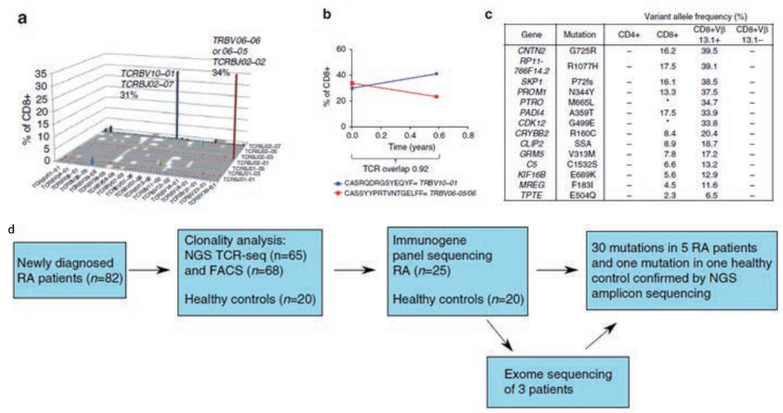


图13 病人1中CD8+ T细胞的克隆分布及突变。

(a) 利用TCR测序获得的CD8+ T细胞克隆分布。(b) 病人1中存在两个高频克隆，经过免疫抑制治疗后仍高表达。(c) 利用扩增子测序，在流式分选后的亚群中进行突变验证。(d) 本文研究思路图。

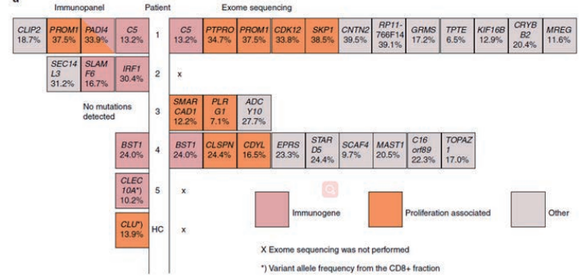


图14 5个RA病人和1个健康对照的CD8+ T细胞中发现的体细胞突变

粉色为免疫相关基因，橘色为细胞增殖相关基因，灰色代表其他基因。采用的技术为目标区域测序、外显子测序。HC为健康对照。

案例二：外周血B细胞主克隆可预测类风湿性关节炎发病风险[6]

发表期刊：Annals of the rheumatic diseases

影响因子：12.81

发表日期：2017年8月

临床意义：

血清学阳性的RA，患者在发病前就已经出现自身抗体的变化，但目前临床上对RA发病预测的有效率低（~20%）。临床发现，RA病人的滑膜有B细胞和浆细胞的渗透，B细胞清除治疗有效，这些都说明，B细胞和浆细胞与RA发病有关。本文通过对血液样品进行BCR测序，用BCR预测RA高风险人群，有助于早期治疗。

样品选取：

Test cohort：65个自身抗体阳性的RA高风险个体，随机选取21个RA高风险个体，分为高风险发病组（11个），高风险未发病组（10个），分别取外周血和滑膜组织，RNA样品。

Control：10个自身抗体阴性的健康个体

Validation cohort：50个高风险个体，跟踪36个月

注：未发病组跟踪平均69个月（5年），发病组平均15个月

分析结果：

1. 主克隆在RA发病之前会在外周血中检测到，而在未发病人群中检测不到

3个组间克隆差异分析发现：11个发病的高风险个体在发病之前，血液中均检测到多个dominant BCR clones（Dominant clones表示占比超过0.5%的克隆）；10个未发病的高风险个体、10个对照中，血液中均未发现dominant BCR clones。

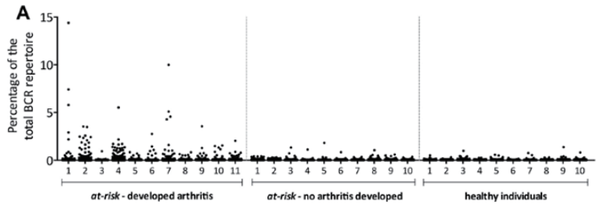


图15 外周血样品中不同组间的dominant克隆分布

2. 鉴定预测发病风险的biomarker

选择36个月发病的临床个体，65例；验证群体，50例。

ROC曲线评估主克隆数量、主克隆整体、the impact of the single most dominant clone，预测疾病风险的效率，最终建立了一个预测模型：>=5 dominant BCR clones为“BCR-clone positive”，<5 dominant BCR clones为“BCR-clone negative”。5个及以上BCR主克隆预测36个月内发病的风险为83%，而小于5个主克隆的发病风险为13%。

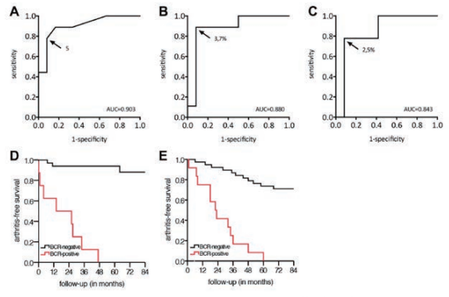


图16 ROC和Kaplan-Meier曲线分析

A-C为ROC曲线，AUC越大说明biomarker分类效果越好。D-E为Kaplan-Meier curve，D为测试群体，E为验证群体。

案例三：乳糜泻TCR特征[7]

发表期刊：Gut

影响因子：16.66

发表日期：2017年2月

临床意义：

>95%的乳糜泻（CD）接受无麸质饮食后会出现好转，但对顽固性乳糜泻（RCD）无效。RCD分为I和II型，I型通过免疫抑制可以控制，II型50%五年内会发展为EATL。临床上针对RCD I和II型的治疗策略是不同的，因此**临床诊断**非常重要。

样品选取：

十二指肠粘膜样品共47个，健康对照(n=9), active coeliacs (n=10), 无麸质饮食后好转病人 (n=9), RCD type I (n=8), RCD type II (n=8) and unclassified Marsh I cases (n=3)。

注：EATL（肠病型T细胞淋巴瘤）Marsh通过描述一系列异常变化说明小肠活检病理改变特点，制定了广为人知的Marsh标准（Marsh I 可见浸润灶：上皮内淋巴细胞增多）。

研究结果：

II型RCD频率最高的克隆（CDR3相同的reads算同一种克隆），平均占42.6%，而control和RCD I组中只有6.8%和6.7%。多次内窥镜检查发现II型RCD病人的克隆多年保持稳定，并且没有EATL的临床症状。II型RCD病人的克隆是unique的，病人间没有共有克隆。

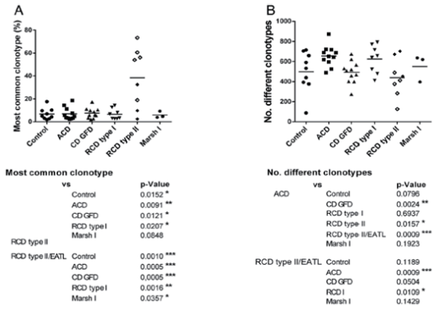


图17 乳糜泻病人中TCRB多样性

A.所有样品中最常见克隆频率，横线代表平均值，RCDII进展到EATL的病人与其他病人组相比，常见克隆频率显著偏高。B.病人克隆多样性。每个点代表每个样品unique的克隆，ACD高于CD GFD和RCDII /EATL group，RCDII/EATL group的unique克隆比RCDI显著偏低。

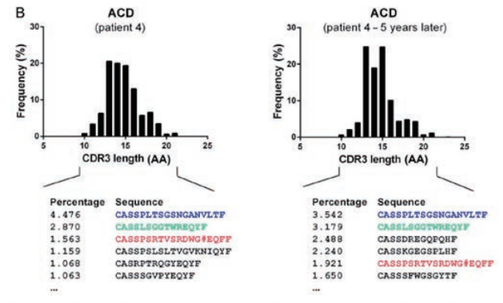


图18 间隔5年，同一个病人的TCRB保持稳定，彩色标记为相同的克隆

可能存在的风险

项目设计不合理或样品数太少，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，设计合理的疾病和对照组，并确定疾病与T细胞还是B细胞有关。

常见问题

Q1：TCR和BCR各条链编码基因的区别?推荐哪条链?

TCR Beta链和BCR重链是由V、D、J基因编码的，而TCR alpha链和BCR轻链是由V、J基因编码的。从发表文章来看，研究TCR beta链和BCR重链的比较多。

Q2：免疫组库测序可以区分IgG、IgM、IgD、IgE吗?

免疫球蛋白的亚型，通过C区序列可以区分，华大有相应的引物，但必须是RNA样品。利用多重PCR的方法，利用V区-C区的引物，用约30bp的序列区分。产物长度大约是在200-300bp。

Q3：免疫组库的测序深度?能得到多少序列?

分析数据显示，数据量的增加，主要影响低频克隆，并且这些克隆的排序在一千多到九千不等，而研究往往只关注top100的克隆和疾病的关系，所以推荐起始数据量1G raw data。但如果客户想关注更多低频克隆，可以加大测序数据量。

备注：表格中的样品（H1-H4，P1-P8）原始数据量为2-3G，截掉一半数据量为1-1.5G，表格中highest uniq_rank这一项表示unique 的克隆中频率最高的那个克隆在原始数据克隆中的排名。

Sample	origion	cut	overlap	uniq_by_origin	highest uniq rank
H1	26654	15662	15213	11441	2108
H2	17295	9975	9733	7562	1293
H3	27516	16301	15940	11576	2305
H4	25866	15153	14781	11085	2137
P1	28436	16748	16305	12131	3343
P2	27116	14921	14583	12533	2047
P3	25498	14453	14180	11318	2355
P4	26675	15407	15045	11630	2339
P5	55631	32664	31787	23844	5973
P6	46524	30249	28859	17665	8306
P7	63479	40190	38575	24904	8950
P8	49508	30062	29037	20471	5366

Q4：做免疫组库测序，用基因组DNA做模板好，还是RNA好?

A6：DNA水平侧重于研究基因重组信息，RNA水平侧重于研究基因的表达状态。使用DNA和RNA做模板各有优缺点:

gDNA的优点是：

1）因为每个基因只有两个拷贝，因此可准确地反映免疫细胞受体的克隆数；

2）DNA更稳定，易储存。

缺点是：

1）由于copy数不高，模板含量低，因此可能需要更多的样品；

2）由于J区和C区之间有很大的intron区，受测序长度的局限，缺少特异性扩增引物来扩增CDR全长。

RNA的优点是：

1）J区和C区之间无intron，可用C区进行引物设计，扩增全长CDR；

2）由于表达丰富，模板含量高，样品消耗量少。

缺点是：

1）免疫细胞受体的克隆性受到mRNA表达高低的影响，不能客观地反应本身的克隆数；

2）RNA不如DNA稳定，样品保存和操作要求较高。

华大优势

1. 丰富的项目经验：已完成包括肿瘤、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。

2. 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。

3. 扩增偏好性低：采用两步法建库，并优化引物配比，将扩增偏好性降低约70%。

4. 可重复性高：同一样本建库两次克隆一致性高。

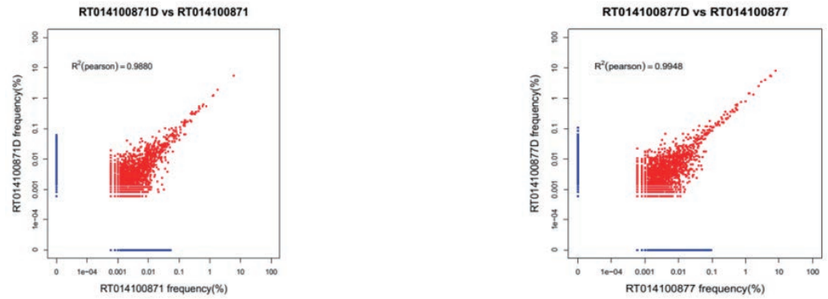


图19 同一样本实验重复性评估(Hiseq)

5. 更丰富的分析结果：新增多种结果统计图表；新增V/J基因频率分布统计、多样品聚类分析、共性分析、差异分析；新增四种多样性评估指标。

6. 更友好的xbio结题报告展现形式：全新升级的结题报告界面友好，对分析方法及结果解释详尽，图表按照发表文章要求展示，让您一目了然。

7. 更真实的克隆定量信息：将低质量reads与高质量克隆进行比对，挽回重要数据，让克隆定量信息不丢失。

8. 更强的纠错能力和错配处理能力：利用多层聚类方法，纠正PCR和测序引入的错误；错配处理能力提升，更适合分析BCR的高突变区域。

华大已发表文章

研究内容	发表时间	发表期刊	影响因子	文献标题
肝癌亚型差异分析	2015.04	*Oncoimmunology*	7.72	Identification of characteristic TRB V usage in HBVassociated HCC by using differential expression profiling analysis
分析软件	2015.08	*Genetics*	4.56	IMonitor: a robust pipeline for TCR and BCR repertoire analysis
骆驼	2016.09	*PLoS One*	2.81	Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies
实验方法学	2016.03	*PLoS One*	2.81	Systematic Comparative Evaluation of Methods for Investigating the TCR β Repertoire.
肝癌	2015.07	*Cancer Letters*	6.38	Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma
原发性胆汁性胆管炎	2016.09	Journal of immunology	4.86	Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis
微小残留病	2016.10	Frontiers in Immunology	6.43	Minimal Residual Disease Detection and Evolved IGH Clones Analysis in Acute B Lymphoblastic Leukemia Using IGH Deep Sequencing
预测 V/J 基因软件	2016.11	Frontiers in Immunology	6.43	IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data
乳腺癌、癌旁和淋巴结的 TCR 分析	2017.02	Cancer Immunology Research	8.28	The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues
结肠腺瘤和结肠癌浸润淋巴细胞	2017.05	Journal of immunology	4.86	Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma
免疫缺陷	2015.12	Human Molecular Genetics	5.99	DCLRE1C (ARTEMIS) mutations causing phenotypes ranging from atypical severe combined immunodeficiency to mere antibody deficiency

参考文献

[1] 曹雪涛. 免疫学前沿进展[J]. 中国免疫学杂志, 2010, 8: 001.

[2] Kirsch I, Vignali M, Robins H. T‐cell receptor profiling in cancer[J]. Molecular oncology, 2015, 9(10): 2063-2070.

[3] 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.

[4] H.C. von Budingen, T.C. Kuo, M. Sirota, C.J. van Belle, L. Apeltsin, J. Glanville, et al., B cell exchange across the blood-brain barrier in multiple sclerosis, J. Clin. Invest. 122 (2012) 4533–4543.

[5] Savola P, Kelkka T, Rajala H L, et al. Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis[J]. Nature Communications, 2017, 8: 15869.

[6] Tak P P, Doorenspleet M E, de Hair M J H, et al. Dominant B cell receptor clones in peripheral blood predict onset of arthritis in individuals at risk for rheumatoid arthritis[J]. Annals of the rheumatic diseases, 2017: annrheumdis-2017-211351.

[7] Ritter J, Zimmermann K, Jöhrens K, et al. T-cell repertoires in refractory coeliac disease[J]. Gut, 2017: gutjnl-2016-311816.

感染类疾病免疫组库研究方案

研究背景

免疫是指“人体识别自身和排斥异己的能力”，这种能力包括机体抵御感染（医学上称免疫应答/抗感染抵抗力）、稳定调节人体“内环境”和监视异常细胞(如肿瘤细胞)，即免疫应答、免疫自稳和免疫监视三种功能。免疫紊乱是感染、过敏、炎症的主要原因。换言之，免疫应答过弱可出现反复感染或感染不易控制；免疫应答过强或免疫自稳失控时，出现过敏/变态反应或炎症反应失控；当免疫监视功能不足时，会发生肿瘤，免疫监视会引起器官移植后的排异反应。可见免疫功能和免疫力是双刃剑，最适度的免疫应答，动态的内环境稳定状态才是最佳免疫状态，免疫失衡会引起感染、过敏相关疾病。

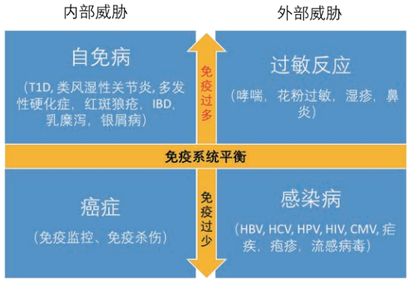


图1 疾病的发生与免疫系统失衡有关

当免疫过少感染病毒时，尽管固有免疫系统对控制病毒的早期复制很重要，但并不能完全清除病毒，病毒的清除依赖于后续的获得性免疫反应，即T/B细胞。DC（树突状细胞）是链接固有免疫和获得性免疫的桥梁，激活的DC吞噬病毒后向淋巴组织潜移，与此同时，开始表达共刺激分子，加工和提呈病毒抗原并分泌一些促炎症细胞因子，进入淋巴组织后，DC将诱导病毒特异性初始型T细胞激活、增殖和分化，最终形成效应性的CD4+和CD8+ T细胞；而在淋巴结的生发中心，在抗原和CD4+ T细胞的作用下，B细胞被激活并分泌抗体。效应性T细胞和抗体可通过血液循环运送到感染部位，CD4+ T细胞通过分泌细胞因子进一步发挥免疫调节效应，而CD8+ T细胞则通过诱导病毒感染的靶细胞凋亡，发挥效应功能。最终，在清除病毒后，机体还会建立针对该种病毒的特异的免疫记忆T/B细胞库[1]。

因此，暴露于同一种病原的不同病人，它们的记忆淋巴细胞是相同的或非常接近的，在二次感染的时候，记忆T/B细胞会对病原做出快速反应，因此通过免疫组库测序分析记忆淋巴细胞的组成，通过搜索与病毒相关的public TCR数据库，可以判断病人感染的病原，辅助临床诊断。

什么是免疫组库？

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。

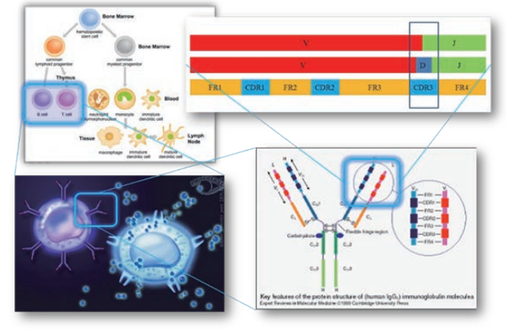


图2 免疫组库研究内容

以B细胞为例，B细胞表面有BCR（B细胞受体），即Y字形的抗体。BCR顶端的区域是CDR区域（抗原互补决定区），分别由V、D、J基因编码，其中CDR1和CDR2是由V基因编码，CDR3是由V(D)J基因编码。免疫组库是通过对编码CDR3/CDR的V(D)J基因进行测序，通过基因频率反映B细胞克隆多样性。

TCR/BCR CDR3多样性是如何实现的？

1、V(D)J recombination重排；

2、V-D和D-J间随机插入碱基；

3、抗体常发生体细胞突变。

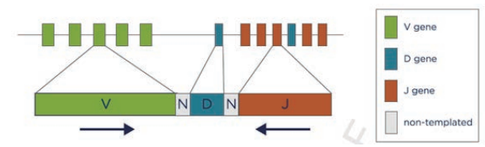


图3 CDR3编码基因的多样性是由V(D)J基因重排加上随机插入碱基产生的

免疫组库在感染类疾病方面的研究进展

免疫组库应用很广，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方面找到研究思路，例如自身免疫疾病、感染类疾病、癌症、HIV等等；医学应用上，疫苗研发评价、药物研发、疾病诊断、器官和肝细胞移植等，发表文章呈逐年上升趋势（影响因子5分以上的占到了80%以上，10分以上文章占到了近40%）。

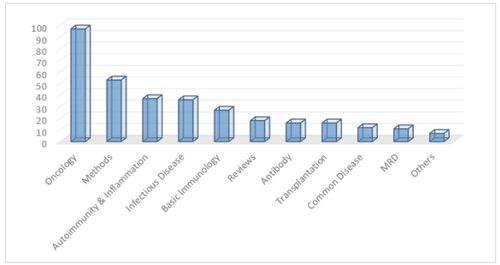


图4 免疫组库发表文章应用领域(IF>5文章不完全统计)

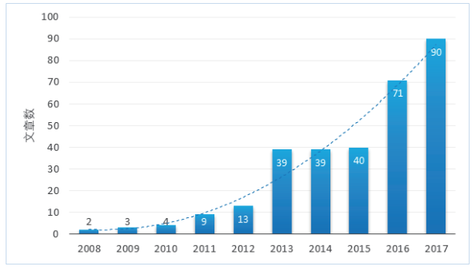


图5 免疫组库已发表文章统计(IF>5文章不完全统计)

细菌或病毒感染会刺激机体产生适应性免疫,研究发现,通过对比宫颈组织和外周血中HPV疫苗的有效性,发现疫苗引起宫颈位置的效应免疫反应,而对外周血的影响很小,因此外周血不适合评估疫苗的效果[2]。牛津大学的研究人员认为BCR高通量测序可用于评估感染后病人BCR库的多样性、动态变化及抗体进化过程[3]。Adaptive公司对681个样本(289个CMV阳性,352个CMV阴性,40个感染状态未知)进行TCR测序,发现TCRβ可用于评估CMV(巨细胞病毒)的感染状态,ROC曲线评估此方法的敏感度为0.90,特异性为0.89 [4]。德国杜伊斯堡大学的研究人员对22例混合性冷球蛋白血症(MC)阴性的HCV感染病人和7个健康人的外周血进行BCR测序,发现HCV感染干扰了B细胞的组成,刺激了IgM+ B细胞的增殖[5]。加拿大研究人员对5例病毒感染的病人进行定期跟踪,分析记忆T细胞和*de novo* T细胞在二次感染中的作用,结果发现二次感染后,大部分的HCV特异性的T细胞克隆来源于已存在体内的记忆T细胞库[6]。

表1免疫组库已发表文章研究的感染类疾病类型

感染类疾病
CMV
CMV、EBV
CMV、流感、水痘-带状疱疹病毒
H1N1
H7N9 感染
HBV
HCV 感染并发 MC 和 B 细胞淋巴瘤
HIV
HPV
Neonatal sepsis
Recombinant adeno-associated virus (rAAV)
蛋白结合 RNA-seq 流感
急性病毒感染
流感疫苗
疟疾感染
疱疹病毒
同卵双胞胎 HBV 感染
新生儿应对病毒感染

方案设计

A. 研究目标

感染对克隆组成及多态性的影响,选择不同细胞组群进行分析。

B. 样品选择

>20例,对照10例,多点、多组或多个细胞亚群进行跟踪。

C. 实验技术

多重PCR在BCR或TCR的位于CDR3区两端的V、J基因保守区域设计PCR引物,通过多重PCR扩增得到互补决定区CDR3区域,扩增产物用于后续高通量测序PE151(数据量推荐1Gb raw data)。如果模板为RNA,则需要先进行反转录得到cDNA,再进行多重PCR。

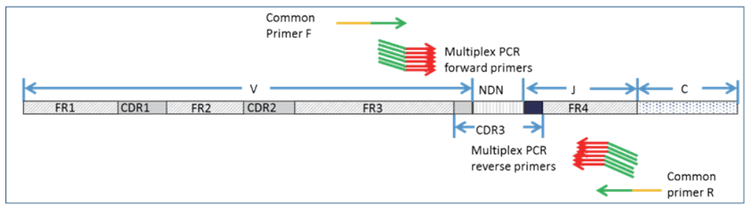


图5 多重PCR示意图

D. 信息分析流程

1）测序所得的数据称为raw reads或raw data，随后要对raw reads进行质控(QC)，以确定测序数据是否适用于后续分析；2）经过滤得到的clean reads比对到参考序列，对于比对上的reads，做下一步的组装，得到具体的功能区域，例如CDR3区（clones）；3）碱基质量符合要求的克隆序列会作为核心克隆（core clonotype），存在一个以上质量值较差碱基的克隆会以核心克隆作参考二次比对和校正；4）然后，对相差一个碱基的克隆，进行层次聚类，每个分支间仅有一个碱基差别（mismatch），依次聚类下去，克隆频率低的克隆会合并到上一分支，最终保留最顶端的head序列；4）将上述得到的克隆序列再次比对到V，D，J和C参考序列，最终得到的统计文件包含了克隆序列、氨基酸残基序列、克隆数量、克隆频率，V/J基因组合等信息。后续可以根据这些信息做克隆分布、基因重组、多样性分析等深入挖掘。

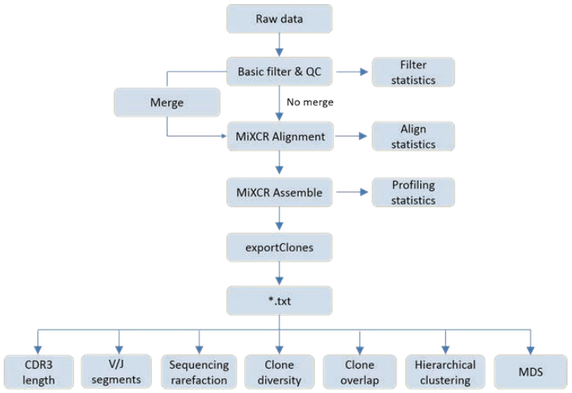


图6 信息分析流程

E. 信息分析内容

1. 基本数据统计

数据过滤：对原始数据进行去除接头污染及低质量reads的处理

数据搭建：数据拼接，消除测序背景及有效数据构建

数据统计：数据产出统计及测序数据的成分和质量评估

2. 数据比对分析

比对分析：与数据库V/D/J基因片段比对

比对结果统计

3. 克隆序列特征注释

CDR3区核酸序列和氨基酸序列

鉴定无效序列（包含终止密码子，超出结构范围）

鉴定单碱基突变（替换、删除、插入）（for BCR）

4. 单样品克隆群体特征分析

CDR3序列长度分布

V/J基因频率分布

V-J基因组合频率分布（3D,Circos）

克隆群体结构分析（频率分布，D50曲线，甜甜圈图）

5. 样品间比较分析

测序饱和度分析

克隆多样性分析(辛普森系数、香农威纳系数等)

样品间共有克隆分析

聚类分析（层次聚类，MDS聚类）

组间差异分析

F. 部分分析结果展示

1. V-J基因频率

针对测序数据结果序列，使用IMGT数据库进行比对，鉴定出V、D、J基因，并对样本中所有克隆的V基因、J基因、V-J基因组合形式进行了统计，以每种克隆reads数计算权重，V-J组合结果以3D和Circos图分别展示。

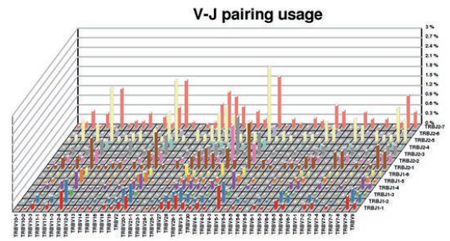


图7 V-J基因组合频率3D柱状图

平面上分别为V基因、J基因。柱子的高度代表一种V-J组合的频率。

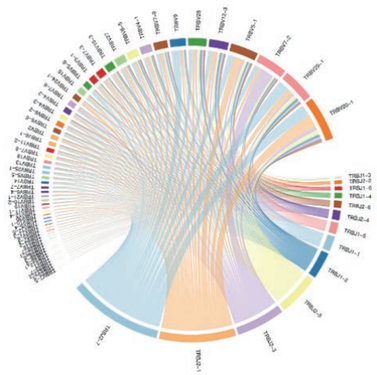


图8 V-J基因组合频率Circos图

每个颜色块代表一种基因，颜色块越宽，频率越高。色块间的连线代表一种V-J基因组合方式。

2. 克隆多样性分析

克隆多样性统计，是不同于V-J基因频率的统计。V-J基因会存在SNP（BCR存在超突变）、随机碱基插入等，增加了克隆的多样性。

样品克隆频率分布图直观反映每个样本中所有克隆类型频率分布情况，D50是近年来引入反映样本克隆群体结构的一个指标，值越小，反映克隆多样性越低，值越大，克隆多样性越高。

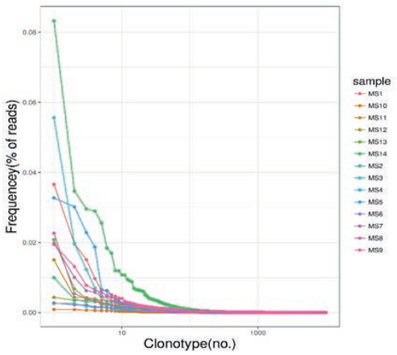


图9 克隆频率分布图

横纵坐标分别为克隆数和克隆频率。

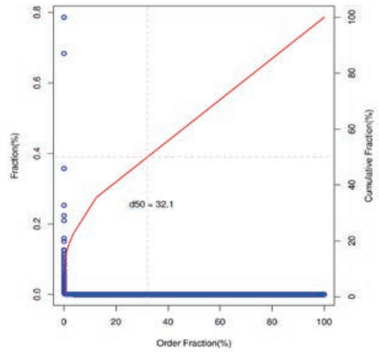


图10 D50曲线

X轴表示样本克隆组成累积百分比，左侧Y轴表示单个类型克隆频率，右侧Y轴表示克隆频率累计百分比。每个点表示单个克隆具体的频率，曲线为所有克隆的累计分布。其中的D50为累计频率达到50%时的克隆所在位置。

3. 组间差异分析

分组比较克隆多样性及top20高频表达V基因频率。

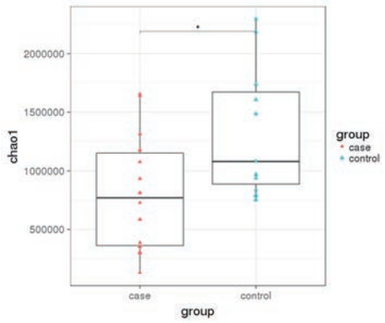


图11 克隆多样性箱线图

每个箱线代表一个group，每个箱线图对应五个统计量（自上而下分别为最大值，上四分位数，中值，下四分位数和最小值）。使用Student's t-Test进行差异显著性检验，其中ns表示差异不显著（P>0.05）；*表示有统计学差异（P<0.05）；**表示有显著统计学差异（P<0.01）；***表示有极其显著的统计学差异（P<0.001）。

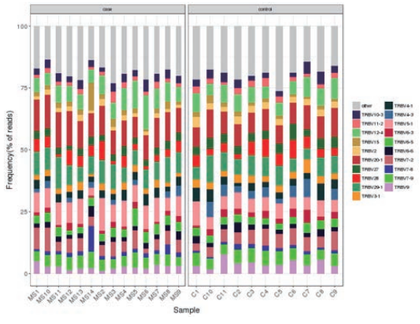


图12 Top20 V基因频率组间分布柱状图

X轴表示样品编号，Y轴表示重组结果中各基因的使用频率。

应用案例

案例一：免疫组库测序发现CMV感染特征[4]

发表期刊：Nature Genetics

影响因子：27.96

发表时间：2017年4月

研究目的：确定是否可以用TCR特征反映CMV（巨细胞病毒）感染状态。

研究样本：群体1：共计666个样本，其中289个已知CMV阳性，352个CMV阴性，40个感染状态未知；群体2：120个验证样本。

研究结果：建立了一种鉴定模型，可以用TCRβ序列鉴定CMV感染状态，并且在体外验证了三种TCRβ分子与CMV病毒结合，对15,601 TCRβ序列和61个HLA-A和HLA-B alleles进行关联分析，发现87个TCRβ序列与HLA正相关。

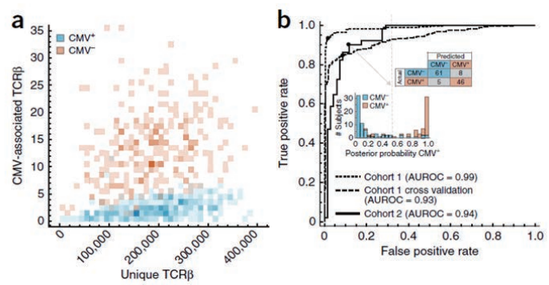


图13 TCRβ可预测CMV感染状态

(a) CMV+和CMV-样本中unique TCRβ分布散点图。(b) ROC曲线显示TCRβ作为CMV感染状态分类器的分类效果。其中虚线表示在群体1中的测试结果，短虚线代表群体1中的交叉验证效果，实线代表群体2中独立验证效果。每条线上的黑圈代表最大后验概率（MAP）的决定阈值。插入图表展示的是群体2中MAP分类效果，敏感度为0.90，特异性为0.89。

案例二：HCV病毒感染后BCR克隆变化[5]

期刊：Blood

影响因子：13.16

发表日期：2017年12月

研究目的：HCV感染病人，50%会发生混合性冷球蛋白血症（MC），一种单克隆B细胞扩增的疾病；10% MC病人会发展成B细胞淋巴瘤。本文目的是研究HCV感染对MC- 病人的B细胞克隆的影响及其机制。

研究样本：30例MC阴性的HCV感染病人，15个健康人，流式分离外周血B细胞，分析HCV对细胞组成的影响。其中22例HCV+病人和7例健康人进行BCR重链测序，DNA样品，分析不同细胞亚群的V基因和CDR3差异。

研究结果：HCV感染扰乱了B细胞组成，产生了很多大的B细胞克隆，尤其对IgM+ 记忆B细胞。

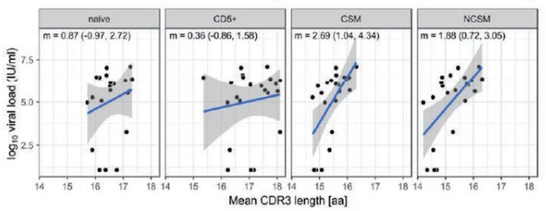


图14 记忆B细胞中CDR3长度与病毒负荷的对数相关

说明HCV感染促进了记忆B细胞长重链（long heavy chain）的产生

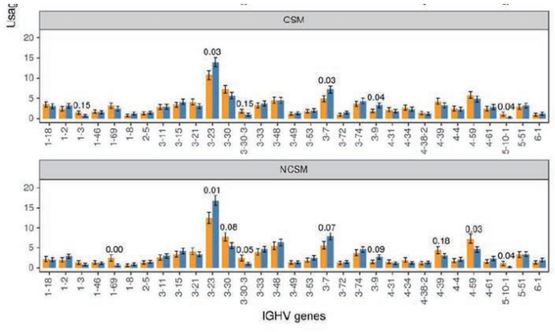


图15 不同细胞亚群V基因用量

其他细胞群没差异，只有IgM+记忆B细胞在IGHV1-69, IGHV3-30, IGHV4-59 usage提高，而这些V基因是参与自免病或HCV相关淋巴瘤的基因。注：naïve，初始B细胞；mature CD5+，成熟CD5+ B细胞；NCSM，类型未转换的记忆B细胞（即IgM+记忆B细胞）；CSM，类型转换的记忆B细胞。

可能存在的风险

项目设计不合理或样品数太少，可能会导致克隆表达差异不显著。因此，在项目执行前，需要充分了解疾病的背景，设计合理的疾病组和对照组，并确定疾病与T细胞还是B细胞有关。

常见问题

Q1：TCR和BCR各条链编码基因的区别?推荐哪条链?

TCR Beta链和BCR重链是由V、D、J基因编码的，而TCR alpha链和BCR轻链是由V、J基因编码的。从发表文章来看，研究TCR beta链和BCR重链的比较多。

Q2：免疫组库测序可以区分IgG、IgM、IgD、IgE吗?

免疫球蛋白的亚型，通过C区序列可以区分，华大有相应的引物，但必须是RNA样品。利用多重PCR的方法，利用V区-C区的引物，用约30bp的序列区分。产物长度大约是在200-300bp。

Q3：免疫组库的测序深度?能得到多少序列?

分析数据显示，数据量的增加，主要影响低频克隆，并且这些克隆的排序在一千多到九千不等，而研究往往只关注top100的克隆和疾病的关系，所以推荐起始数据量1G raw data。但如果客户想关注更多低频克隆，可以加大测序数据量。

备注：表格中的样品（H1-H4，P1-P8）原始数据量为2-3G，截掉一半数据量为1-1.5G，表格中highest uniq_rank这一项表示unique 的克隆中频率最高的那个克隆在原始数据克隆中的排名。

Sample	origion	cut	overlap	uniq_by_origin	highest uniq rank
H1	26654	15662	15213	11441	2108
H2	17295	9975	9733	7562	1293
H3	27516	16301	15940	11576	2305
H4	25866	15153	14781	11085	2137
P1	28436	16748	16305	12131	3343
P2	27116	14921	14583	12533	2047
P3	25498	14453	14180	11318	2355
P4	26675	15407	15045	11630	2339
P5	55631	32664	31787	23844	5973
P6	46524	30249	28859	17665	8306
P7	63479	40190	38575	24904	8950
P8	49508	30062	29037	20471	5366

Q4：做免疫组库测序，用基因组DNA做模板好，还是RNA好?

A6：DNA水平侧重于研究基因重组信息，RNA水平侧重于研究基因的表达状态。使用DNA和RNA做模板各有优缺点：

DNA的优点是：

1）因为每个基因只有两个拷贝，因此可准确地反映免疫细胞受体的克隆数；

2）DNA更稳定，易储存。

缺点是：

1）由于copy数不高，模板含量低，因此可能需要更多的样品；

2）由于J区和C区之间有很大的intron区，受测序长度的局限，缺少特异性扩增引物来扩增CDR全长。

RNA的优点是：

1）J区和C区之间无intron，可用C区进行引物设计，扩增全长CDR；

2）由于表达丰富，模板含量高，样品消耗量少。

缺点是：

1）免疫细胞受体的克隆性受到mRNA表达高低的影响，不能客观地反本身的克隆数；

2）RNA不如DNA稳定，样品保存和操作要求较高。

华大优势

1. 丰富的项目经验：已完成包括肿瘤、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。

2. 优化引物设计：完成了多重PCR引物的优化，更精确地反映免疫组库克隆情况。其中部分引物设计已申请专利。

3. 扩增偏好性低：采用两步法建库，并优化引物配比，将扩增偏好性降低约70%。

4. 可重复性高：同一样本建库两次克隆一致性高。

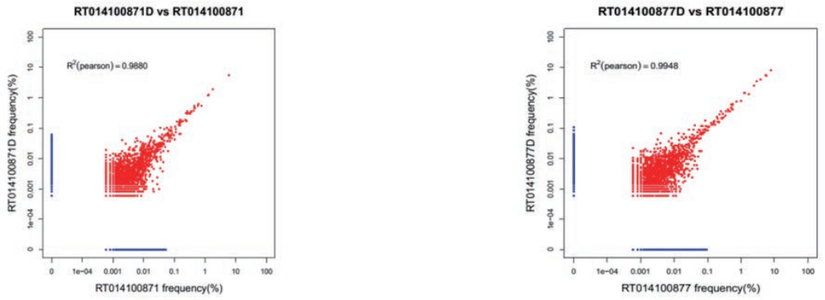


图16 同一样本实验重复性评估(Hiseq)

5. 更丰富的分析结果：新增多种结果统计图表；新增V/J基因频率分布统计、多样品聚类分析、共性分析、差异分析；新增四种多样性评估指标。

6. 更友好的xbio结题报告展现形式：全新升级的结题报告界面友好，对分析方法及结果解释详尽，图表按照文章发表要求展示，让您一目了然。

7. 更真实的克隆定量信息：将低质量reads与高质量克隆进行比对，挽回重要数据，让克隆定量信息不丢失。

8. 更强的纠错能力和错配处理能力：利用多层聚类方法，纠正PCR和测序引入的错误；错配处理能力提升，更适合分析BCR的高突变区域。

华大已发表文章

研究内容	发表时间	发表期刊	影响因子	文献标题
肝癌亚型差异分析	2015.04	*Oncoimmunology*	7.72	Identification of characteristic TRB V usage in HBVassociated HCC by using differential expression profiling analysis
分析软件	2015.08	*Genetics*	4.56	IMonitor: a robust pipeline for TCR and BCR repertoire analysis
骆驼	2016.09	*PLoS One*	2.81	Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies
实验方法学	2016.03	*PLoS One*	2.81	Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire.
肝癌	2015.07	*Cancer Letters*	6.38	Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma
原发性胆汁性胆管炎	2016.09	Journal of immunology	4.86	Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis
微小残留病	2016.10	Frontiers in Immunology	6.43	Minimal Residual Disease Detection and Evolved IGH Clones Analysis in Acute B Lymphoblastic Leukemia Using IGH Deep Sequencing
预测 V/J 基因软件	2016.11	Frontiers in Immunology	6.43	IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data
乳腺癌、癌旁和淋巴结的 TCR 分析	2017.02	Cancer Immunology Research	8.28	The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues
结肠腺瘤和结肠癌浸润淋巴细胞	2017.05	Journal of immunology	4.86	Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma
免疫缺陷	2015.12	Human Molecular Genetics	5.99	DCLRE1C (ARTEMIS) mutations causing phenotypes ranging from atypical severe combined immunodeficiency to mere antibody deficiency

参考文献

[1] 曹雪涛. 免疫学前沿进展[J]. 中国免疫学杂志, 2010, 8: 001.

[2] Maldonado L, Teague J E, Morrow M P, et al. Intramuscular therapeutic vaccination targeting HPV16 induces T cell responses that localize in mucosal lesions[J]. Science translational medicine, 2014, 6(221): 221ra13-221ra13.

[3] Hoehn K B, Fowler A, Lunter G, et al. The diversity and molecular evolution of B-cell receptors during infection[J]. Molecular biology and evolution, 2016, 33(5): 1147-1157.

[4] Emerson R O, DeWitt W S, Vignali M, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire[J]. Nature genetics, 2017, 49(5): 659.

[5] Tucci F A, Kitanovski S, Johansson P, et al. Biased IGH VDJ gene repertoire and clonal expansions in B cells of chronically hepatitis C virus–infected individuals[J]. Blood, 2018, 131(5): 546-557.

[6] Abdel-Hakeem M S, Boisvert M, Bruneau J, et al. Selective expansion of high functional avidity memory CD8 T cell clonotypes during hepatitis C virus reinfection and clearance[J]. PLoS pathogens, 2017, 13(2): e1006191.

复杂疾病*De novo*突变研究方案

研究背景

*De novo*突变又称为新生突变或从头突变，是基于家系的，首次发生在家系成员中，来源于父母生殖细胞（精子或卵子）中的突变或受精卵中的体细胞突变，即在一个家系里子代具有但父母不具有的突变。由于此突变在进化上未受到选择，发生在基因组上一些关键位置上的此类突变可能造成严重的影响，导致疾病的发生。

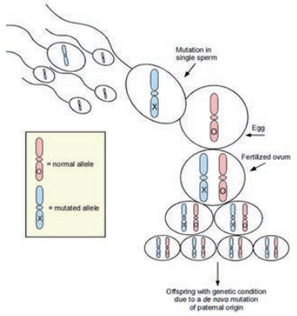


图1 *De novo* mutation定义

*De novo*突变可分为*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SV（Structural variants），不同*de novo*突变类型的突变率存在差异，根据文献报道显示：目前人类胚系*de novo* SNVs的平均突变率约为$1.18X10^{-8}$，与父母相比较，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，因为其长度更长，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
***De novo* mutations in regulatory elements in neurodevelopmental disorders**	2018.3	nature	神经发育障碍	对 8000 个患者及其父母三类假定的调控元件进行了靶向测序	神经发育障碍常常伴随着胎儿大脑中高度保守性激活元件中新生突变的特异性富集，并据此估算出了此种调节元件突变对疾病的贡献率。	40.137
Prevalence and architecture of *de novo* mutations in developmental disorders	2017.1	nature	发育障碍	对 4293 个发育障碍家系进行外显子组测序，并将这些数据与来自另外 3287 个具有相似障碍的个体的数据进行汇总分析	约 42%队列中携带编码序列中的致病性 DNM;这些 DNM 中大约一半破坏基因功能，导致蛋白质功能改变。由 DNMs 引起的发育障碍平均发病率新生儿中为 1/213–1/448。鉴于目前的全球人口统计数据，这相当于每年有近 40 万名患儿出生。	40.137
Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population	2016.3	nature genetics	自闭症	文章对大于 38000 个 ASD 数据进行自闭症谱系障碍和普通人 ASD 相关性状的关联分析	ASD 风险基因包括遗传性突变和 *de novo* 突变也广泛存在普通人中	29.352
CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders	2016.5	Mol Psychiatry	自闭症	对1534名患者和8936个对照进行 SNP 芯片测序检测 CNV，并对 212 个患者和 217 个对照进行 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
***De novo* loss- or gain-of-function mutations inKCNA2 cause epileptic encephalopathy**	2015.3	nature genetics	癫痫	首先用 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
The contribution of *de novo* coding mutations to autism spectrum disorder	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
Biological insights from 108 schizophrenia associated genetic loci	2014.6	nature	精神分裂症	研究人员对 36989 名患者和 113075 名正常人进行了关联分析	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 overlap	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**	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**	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**	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	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 REV3L cause Möbius syndrome**	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

2017年9月，《Nature》报道了迄今为止最大规模的人类自发性基因突变研究。该研究显示，父母年龄越大，尤其是父亲年龄越大，子女的*de novo* mutation发生率越高[3]。这项对人类基因组序列多样性突变过程的分析，对未来医学研究至关重要。越来越多的*de novo*突变研究倾向于基于大队列的群体研究，而不再仅仅局限于少数的家系样本，这在医学、遗传学和演化学的研究中非常重要。

表3 大型队列*de novo*突变研究进展

文章	时间	杂志	研究方向	方法	主要结果	影响因子
Germline *de novo* mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence.	2018.4	Nat Genet	群体研究	1291个trios进行全基因组测序	母亲来源等位基因上的聚类数量与母亲年龄呈正相关，并且与父源的聚类相比，这些聚类的个体突变更多，相互距离更远。母系突变聚类与双链断裂（DSBs）的过程相关，表明，DSB诱导的突变在整个卵母细胞衰老过程中积累，最终作为形成母系突变簇的机制。	27.959
Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland	2017.9	nature	群体研究	分析了来自14,688名冰岛人的平均覆盖率为35×的全基因组测序（WGS）数据	母亲 *de novo* 突变的数量每年增加0.37，是父亲每年1.51的四分之一，且突变的类型随着年龄的增加而显着改变。	40.137

方案设计

A. 整体思路

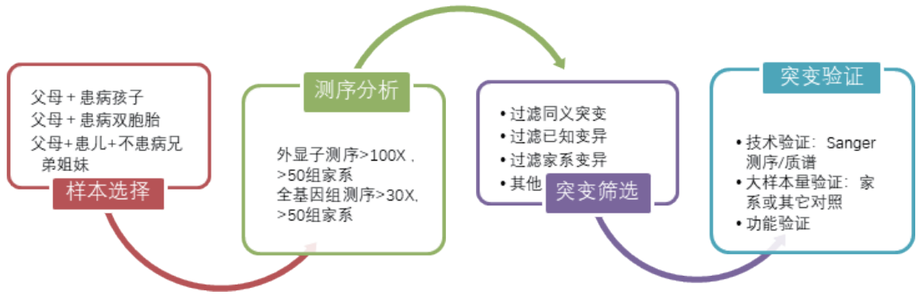


图2 *De novo*突变的设计方案

B. 样本选择建议

本研究方案中疾病样本选择比较特殊，建议选择>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。

C. 采用的技术

全基因组重测序或者外显子测序。*De novo* SNV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未能找到的*de novo*突变且验证率更高。

D. 测序参数

全基因组重测序>30X；外显子测序>100X

E. 预期的结果

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。

F. 信息分析流程图

全基因组重测序可以分析SNV，InDel，CNV，SV；外显子重测序只可分析SNV，InDel。

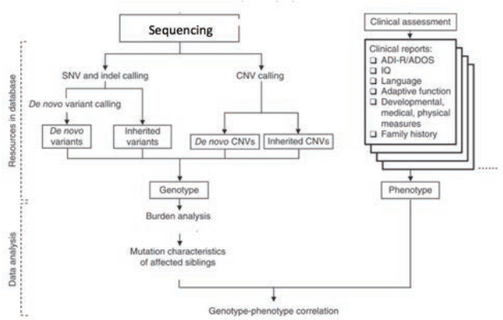


图3 大型队列*de novo*突变研究进展

G. 样本选择建议

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。

H. 采用的技术

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。

应用案例

案例一：父母生育子女的年龄对后代的影响[3]

研究者采集了14,688名冰岛人的血液或口腔细胞样品，并对样品中的遗传物质进行了全基因组测序分析。该研究表明，从基因组水平上来讲，父母的年龄越大，则其子女的新生突变频率就越高。相比目前来说，父亲的生育年龄很大程度上决定了后代的新生突变，而且后代中的这种新生突变数的增加与精神分裂症以及自闭症的发生概率具有一定的相关性。

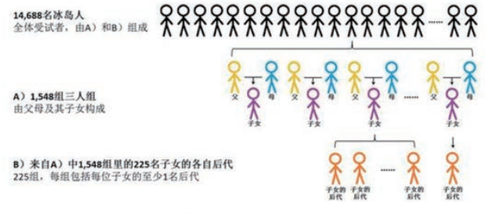


图4 研究样本组成

1. 随父母生育年龄变化，源自父系的新生突变年增长数量是源自母系的四倍

测序识别出了包括SNP、InDel在内的共108,788个新生突变，即受试的1548个家族的“基因平均突变率”为70.3个/家族。从新生突变数量角度，研究者发现，以225组祖孙三代样本作为研究群体时，源自母亲的新生突变数量随母亲年龄的增长每年增加0.37个，而源自父亲的新生突变数量是前者的四倍，随父亲年龄的增长每年增加1.51个。

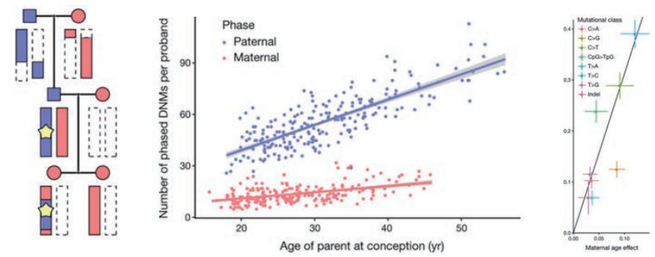


图5左图｜判断某突变为新生突变的依据：

中图｜后代基因组中新生突变数量与父/母生子时年龄的函数关系横轴

右图｜不同突变类型各自的父方（纵轴）/母方（横轴）遗传随年龄增长的变化率的对比

2. 染色体上的聚类突变随母方生育年龄变化产生了更大的变化幅度

与父亲相比，母亲年龄的增长会让聚类新生突变产生更大的增幅。人类8号染色体短臂起始端一个大小为20 兆（Mb）碱基的区域中，来自母亲的新生突变的出现密度是全部染色体中出现新生突变的平均密度的4.5倍。在C>G突变密集的染色体区域内，有56.5%的C>G新生突变簇来自母亲，相应来自父亲的仅为8.8%，印证了此种突变通常源于母亲一方遗传的说法。来自母亲的变异生成的基因簇片段的长度整体长于来自父亲的变异生成的基因簇。

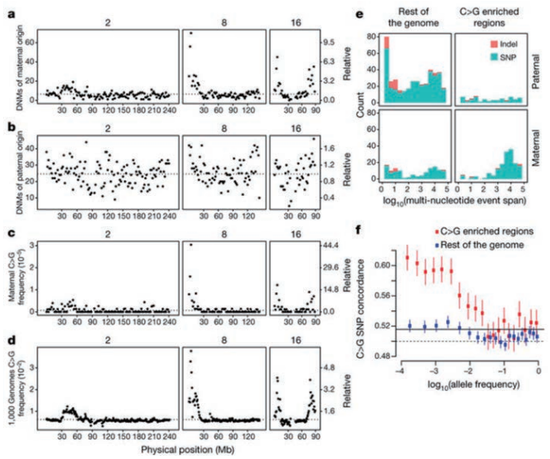


图6 以人类2号、8号（本研究中主要关注的染色体）和16号染色体为例，展示新生突变在染色体上的分布

3. 与古人类和灵长类物种基因组的对比

现代人与古人类基因组的对比表明古人类基因组中相应的染色体上也具有类似的C>G突变趋势。这种突变过程可能已经在人类基因组中发生并持续了数千年。人类与黑猩猩共享了很多基因组内区域大量的C>G突变趋势，大猩猩次之。相比之下，红毛猩猩的基因组中无明显此变化趋势。

案例二：发育障碍中新生突变的流行和结构 [4]

文章对对4,293个发育障碍家系进行外显子组测序，其中大多数患儿是唯一受影响的家庭成员，并将这些数据与来自另外3,287个具有相似障碍的个体的数据进行汇总分析。鉴定了94个具有破坏性DNM的基因，其中14个基因之前在发育障碍症研究中p值不显著。

表4 在之前发育障碍研究中与疾病相关性不显著的14个基因

Table 1 | Genes achieving genome-wide significant statistical evidence without previous compelling evidence for association with DDs

Gene	Missense	PTV	*P* value	Test	Clustering
CDK13	10	1	3.2×10^{-19}	DDD	Yes
GNAI1	7 (1)	1	2.1×10^{-13}	DDD	No
CSNK2A1	7	0	1.4×10^{-12}	DDD	Yes
PPM1D	0	5 (1)	6.3×10^{-12}	Meta	No
CNOT3	5	2 (1)	5.2×10^{-11}	DDD	Yes
MSL3	0	4	2.2×10^{-10}	DDD	No
KCNQ3	4 (3)	0	3.4×10^{-10}	Meta	Yes
ZBTB18	1 (1)	4	1.4×10^{-9}	DDD	No
PUF60	4 (1)	3	2.6×10^{-9}	DDD	No
TCF20	1	5	2.7×10^{-9}	DDD	No
KMT5B	0 (2)	2 (3)	2.9×10^{-9}	Meta	No
CHD4	8 (1)	1	7.6×10^{-9}	DDD	No
SET	0	3	1.2×10^{-7}	DDD	No
QRICH1	0	3 (1)	3.6×10^{-7}	Meta	No

从同一基因中DNMs个体的临床照片中产生整合面部图像，并且在这里显示了六个最显着相关的基因，每张脸都是由超过十个孩子的临床照片产生的，表征了这些疾病的表型多样性。

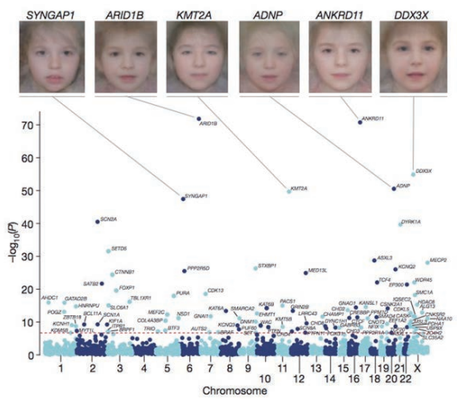


图7 曼哈顿所有测试基因的综合P值图，红色虚线表示全基因组显着性的阈值（$P<7\times10^{-7}$）

估计42%的队列中携带编码序列中的致病性DNM，这些DNM中大约一半破坏基因功能，导致蛋白质功能改变。由DNMs引起的发育障碍平均发病率新生儿中为1/213-1/448。鉴于目前的全球人口统计数据，这相当于每年有近40万名由DNM导致的DD儿童出生。

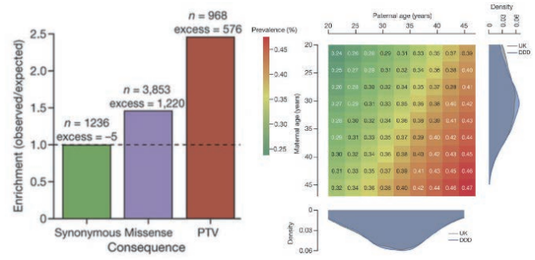


图8 同义DNM的数量归一化和显性DNMs导致DD的患病率

案例三：PDE10A中的*De Novo*突变引起儿童发作性舞蹈症 [5]

研究人员对3个患者和其中患者1和2的健康父母进行全外显子测序，平均测序深度91X。鉴于表型的零星发生，变体的过滤集中于*de novo*显性或隐性突变，个体1和个体2的基于家族的测序方法直接表明PDE10A *de novo*突变的发生，患者3的父母已经死亡，但六个未受影响的兄弟姐妹的DNA可用于测试，并且他们中没有一个携带此突变，通过分析发现在PDE10A上的一个显性*de novo*突变。

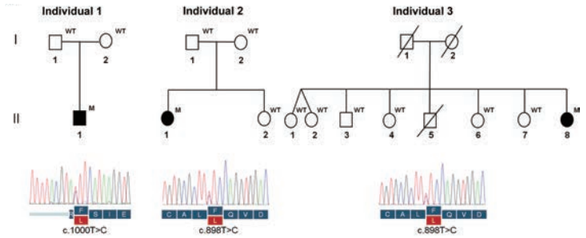


图9 样本家系图

在所分析的10个脑区域中PDE10A表达的变化，壳核中的表达比任何其他区域中的表达更高。PDE10A在下图（B）矢状和（C）冠状切片中的小鼠脑中的表达。PDE10A在纹状体和嗅结核中非常高且有选择性地表达，壳核异常表达A与图BC中的数据一致。

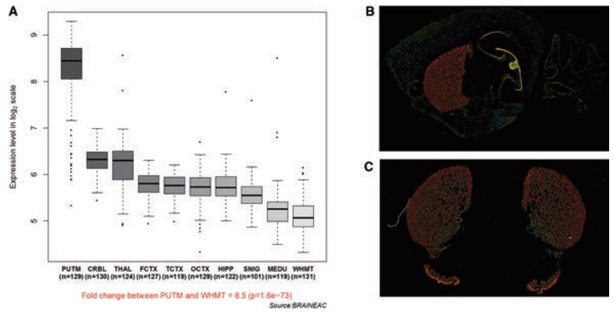


图 10

图A：缩写如下：PUTM，putamen; FCTX，额叶皮质; TCTX，颞叶皮层; OCTX，枕叶皮层; HIPP，海马; SNIG，黑质; MEDU，髓质（特别是下橄榄核）; WHMT，小叶内白质; THAL，丘脑; CRBL，小脑皮质;和N，分析每个大脑区域的样本数量。

图BC：这些基因在正常成年人脑中的区域表达，表达强度范围从低（蓝色）到中等（绿色，黄色）到高（红色）强度。

可能存在的风险

在项目实施过程中，可能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。

常见问题

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。

华大优势

华大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[6]。信息分析方法的升级，可以获

得传统过滤法未能找到的*de novo*突变且验证率更高。*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。

表5 *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。

华大在*de novo* mutation方向的研究有着6+年的经验，共发表文章高达12+篇，累计影响因子高达130+，发表文章详细列表详见表6所示。

表6 华大历年*de novo* mutation发表文章列表

文章	时间	杂志	研究疾病	方法	主要结果
Two *de novo* variations identified by massively parallel sequencing in 13 Chinese families with children diagnosed with autism spectrum disorder	2018.1	Clinica Chimica Acta	自闭症	对 13 个中国 ASD trios 家系进行低深度 WGS+目标区域测序	新发现的 *DEAF1* 基因和与 ASD 相关的 *AADAT* 基因新发错义突变 c.95 C> T 可能是探索该病的病因的重要线索。
***De novo* Paternal FBN1 Mutation Detected in Embryos Before Implantation**	2017.6	Medical Science Monitor.	马凡氏综合征	通过对胚胎中的 FBN1 基因进行有针对性的目标区域测序，Sanger 测序用于确认致病突变。	*FBN1* 中的 *de novo* 突变是在一名中国 MFS 患者中发现的。没有突变的胚胎被 PGD 鉴定并导致成功的怀孕。
Genome-wide characteristics of *de novo* mutations in autism	2016.8	Nature Partner Journals genomic medicine.	自闭症	对 200 个 ASD trios 家系进行全基因组重测序	这些个体在 ASD 风险和表观遗传学基因 *DNMT3A* 和 *ADNP* 中携带已鉴定的 DNM。结果强调非编码变异对 ASD 的病因学的贡献。
Case report of a Li-Fraumeni syndrome-like phenotype with a *de novo* mutation in CHEK2	2016.7	Medicine (Baltimore).	Li Fraumeni 综合征	使用外显子组测序对来自甲状腺，肺和皮肤肿瘤以及正常甲状腺组织的可用 DNA 样品进行测序。	*CHEK2* 是一种新的有害的种系突变它是一种 Li-Fraumeni 综合征致病基因。并证实了这种罕见疾病中新生突变的重要贡献。
Genome-wide patterns and properties of *de novo* mutations in humans	2015.7	Nature Genetics.	无	分析了来自 250 个家族的全基因组的 11,020 个从头突变。	老年父亲的后代中的 *de novo* 突变不仅更多，而且在早期复制的 genic 区域也更频繁地发生。提供了用于医学和群体遗传学应用的全基因组突变率图谱
***De novo* Heterozygous Mutations in SMC3 Cause a Range of Cornelia de Lange Syndrome-Overlapping Phenotypes**	2015.3	Human Mutation.	Cornelia de Lange 综合征	16 个患者进行外显子	*de novo SMC3* 突变占似 CdLS 表型的～1%-2%
Detection and phasing of single base *de novo* mutations in biopsies from human in vitro fertilized embryos by advanced whole-	2015.2	Genome Research.	无	对来自两个囊胚期胚胎的 3 个 5 至 10 细胞活检进行了 LFR WGS 测序，同时还对父母	这是第一个证明 LFR 的 WGS 可以用来准确鉴定这些从头突变，比先前发表的 10 个细

文章	时间	杂志	研究疾病	方法	主要结果
Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios	2015.1	Nature Communications.	无	10 个丹麦人 trios 进行 50X 的全基因组测序，并用从头组装的方法来 *de novo* mutation 突变率	使用概率方法预估 *de novo* SNV 和 indels 突变率为 1.27e-8 和 1.5e-9
TUBB4A de novo mutations cause isolated hypomyelination	2014.8	Neurology.	髓鞘形成减少	5 例患者 trios 家系进行外显子测序，核磁共振和临床信息进行了复核	发现新的 *TUBB4A* 突变
De novo mutation in ATP6V1B2 impairs lysosome acidification and causes dominant deafness-onychodystrophy syndrome	2014.6	Cell Research.	DDOD 综合征	对 2 个家系进行全外显子测序	在两个先证者中验证了 *ATP6V1B2* 中的相同杂合突变。在另一个 DDOD 家族中通过 Sanger 测序进一步证实了该结果
Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing	2013.8	Am J Hum Genet	自闭症	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	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
Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation	2012.12	Cell.	自闭症	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	87%建出的位点确定为 DNMs，确定了 forestDNM 方法的准确性。父亲的年龄越大，后代患自闭症的风险也越大

参考文献

[1] Veltman JA, Brunner HG. *De novo* mutations in human genetic disease. Nature Reviews Genetics,2012,13:565-575.

[2] O'Roak BJ, Deriziotis P. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nat Genet. 2011 Jun;43(6):585-9.

[3] Jónsson H, Sulem P, Kehr B, et al. Parental influence on human germline *de novo* mutations in 1,548 trios from, Iceland[J]. Nature, 2017, 549(7673):519.

[4] Mcrae J F, Clayton S, Fitzgerald T W, et al. Prevalence and architecture of *de novo* mutations in developmental disorders[J]. Nature, 2017, 542(7642):433-438.

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

[6] Michaelson J J, Shi Y, Gujral M, et al. Whole Genome Sequencing in Autism Identifies Hotspots for *De Novo* Germline Mutation[C]. International Meeting for Autism Research. 2013:1431–1442.

疾病相关的肠道菌群多组学研究方案

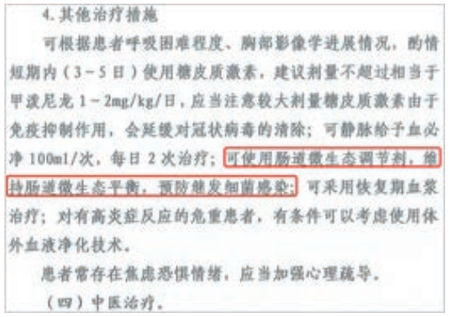
研究背景

肠道菌群人类消化道中居住的大量微生物的统称，肠道菌群是一个非常庞大的群体，在一个健康的成年人体内，肠道菌群总重量可达1.5千克，其中包含的细胞数量高达10^{14n}个，是人体细胞数目的10倍；目前已发现的肠道菌群物种数目多达4万种，其中99%以上都是细菌，而且绝大多数是厌氧菌；其中的基因数目远远超过了人体基因数目的100倍。

近年来与肠道菌群相关的疾病研究迅速发展，据报道有近90%的疾病都有可能与肠道菌群相关。如消化道类疾病、代谢性疾病，肝脏类疾病，免疫类疾病，肺部疾病，精神类疾病，感染类疾病等等。

2019新冠肺炎疫情受到全球范围的广泛关注。每个人一生中可能受到150种以上的病原体感染，其中大多数感染并不引发疾病。机体遭病原体侵袭后是否发病，一方面与病原体致病性的强弱和侵入数量的多寡有关，另一方面与其自身免疫力有关。此外，人体微生物在这个过程中也发挥着重要的作用。

目前国家卫健委发布的《新型冠状病毒感染的肺炎诊疗方案（试行第五版）》，在重型、危重型病例的治疗中也提及了“可使用肠道微生态调节剂，维持肠道微生态平衡，预防继发细菌感染”。2016年李兰娟院士团队发现，对H7N9感染的病人辅助实施微生态制剂的治疗，能有效的防止H7N9病毒导致的微生态失衡的继发的细菌感染。


4. 其他治疗措施

可根据患者呼吸困难程度、胸部影像学进展情况，酌情短期内（3－5日）使用糖皮质激素，建议剂量不超过相当于甲泼尼龙 1－2mg/kg/日，应当注意较大剂量糖皮质激素由于免疫抑制作用，会延缓对冠状病毒的清除；可静脉给予血必净 100ml/次，每日 2 次治疗；可使用肠道微生态调节剂，维持肠道微生态平衡，预防继发细菌感染；可采用恢复期血浆治疗；对有高炎症反应的危重患者，有条件可以考虑使用体外血液净化技术。

患者常存在焦虑恐惧情绪，应当加强心理疏导。

（四）中医治疗。

肠道菌群的结构会受到很多因素的影响，包括宿主遗传因素、地理因素、饮食、生活习惯、健康状况、药物使用等等。随着年龄的增加，老年人肠道中产短链脂肪酸的双岐杆菌、罗斯氏菌和粪杆菌等细菌会减少，而耐氧菌和致病菌会增加。这些变化会导致菌群紊乱，并与多种老年性疾病的高度相关，如高血压，糖尿病，心血管等疾病。而肠道是机体最大的代谢器官和免疫器官，肠道菌群的紊乱会加剧老人肠黏膜免疫的功能降低。这可能也是新冠病毒感染死亡病例多见于患有基础病的中老年人的原因之一。

疫情当前，除了调整饮食结构，保持健康的饮食习惯以外，适当的增加富含膳食纤维的食物如杂粮、粗粮、菌菇类食物、红薯、山芋、玉米粉、荞麦、燕麦粉等的摄入，可以促进有益菌的生长，抑制条件致病菌的繁殖，还可按医生建议，适当补充益生菌/益生元制品巩固微生态平衡，从而提高机体自身的免疫力。当前也有报道发现某些中药可能有利于新型冠状病毒感染的病人恢复健康，推测很可能部分原因是中药的某些活性成分巩固了肠道微生态平衡，通过肠-肺轴提高了机体自身的免疫力。

肠道菌群并不是一个孤立的存在，作为一个长期以来被忽略的“器官”，肠道菌群与我们身体的其他器官（包括脑、肾、肝、肺、心、皮肤等）有着密切的联系，参与人体生命活动的各个途径，包括脂肪储存、血管生成、免疫系统的发育和成熟、某些维生素和非必需氨基酸的合成、药物代谢、神经调节、食物消化和营养吸收、病原菌抵抗、上皮损伤修复、骨骼生成及骨代谢等。这些功能及其对应的pathway可以为肠道菌群相关机制的研究提供依据。

宏基因组以环境中所有微生物基因组为研究对象，通过对环境样品中的全基因组DNA进行高通量测序，获得单个样品的饱和数据量，基因组成及功能，特定环境相关的代谢通路等分析，从而进一步发掘和研究具有应用价值的基因及环境中微生物群落内部、微生物与环境间的相互关系。可为环境中微生物的研究、开发和利用提供基因资源库。

代谢组学(metabonomics/metabolomics)：是继基因组学和蛋白质组学之后新发展起来的学科，研究对象主要是生物体内1000Da以内的小分子物质，如有机酸、氨基酸、核苷酸、糖、脂质、维生素等。通过对生物体内的小分子代谢物进行定性定量分析，寻找代谢物与生理病理变化的关系。由于能够更为真实和直观地反应机体对于刺激的综合响应，代谢组学已经在疾病诊断、临床分型、预后评估、标志物筛选、发病机制研究、药效研究、个性化治疗等多个领域得到了广泛应用。

代谢组学是最接近表型的组学，可作为连通菌群和临床表型的重要桥梁，使得过去相互独立的菌落数据和表型数据获得了有效结合，可形成临床或动物表型-菌群-代谢物-生理病理机制闭环的研究思路。

研究目的

本方案适用于研究疾病（包括感染类疾病）和人体微生物的相关性。针对肠道样本进行宏基因组学和代谢组学研究，分析不同疾病程度样本间基因、功能、物种、代谢物间的差异；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。

方案设计

1. 样本选择

1） 推荐分组及样本量

分组：

a) 疾病组VS对照组

b) 对照组、疾病早期、疾病晚期

c) 对照组、疾病组、疾病治疗组

d) 在以上分组基础上增加其他分组，如不同治疗方法/时期、疾病不同阶段/不同型别、多种样本类型等。

样本量：至少2组，每组30个样本以上

2）样本类型

宏基因组样本：粪便样本、肠道内容物等

粪便样品采集建议：样品采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，-80°C保存，干冰运输。如果无法实现冷冻保存，常温粪便样本保存试剂盒，常温保存、运输。

代谢组样本：血浆、血清、粪便、肠道内容物等

代谢组样本量要求：血浆、血清 ≥250 μL/例；粪便、肠道内容物 ≥200 mg/例。

3）对采样人群进行饮食及生活习惯调查。

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

2. 研究方法

1）宏基因组：DNBSEQ平台，小片段文库，推荐测序10G clean data/样本。

2）代谢组学：非靶向代谢组学，脂质组学，靶向代谢组学（可根据研究目的选择合适的代谢组学产品）

靶向代谢组学包括氨基酸检测、胆汁酸检测、短链脂肪酸检测、氧化三甲胺（TMAO）及其相关代谢物检测，可根据研究目的选择合适的靶向代谢组学检测产品。

3. 分析内容

【宏基因组分析】

通过物种、基因、功能层面的分析寻找和疾病相关的marker及相关作用通路，研究微生物与疾病的相互关系，挖掘微生物在疾病发生中的作用机制、构建疾病预测模型等。

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

【代谢组组学分析】

非靶向代谢组学/脂质组学：通过一系列的统计分析找出不同处理条件下差异表达的代谢物，并对代谢物及差异代谢物进行鉴定，并进行pathway分析。

靶向代谢组学：通过标准品建立标准曲线，对样本中的目标代谢物进行绝对定量。

【宏基因组、代谢组与表型关联分析】

由于宏基因组和代谢组数据具有高维度和复杂性等特征，直接将两组数据进行整合分析具有很大挑战。因此我们设计的关联分析方案基于降维的思想，将代谢组数据和微生物物种通过数据驱动聚类（Clustering of co-abundant metabolites或Binning co-abundant genes）法降维，微生物组功能组成基于知识驱动如KEGG 层级功能模块分类法降维，筛选与表型显著相关的数据特征进行跨组学关联分析（cross-omics association analysis），最后鉴定与表型显著相关的KEGG功能模块的驱动物种。

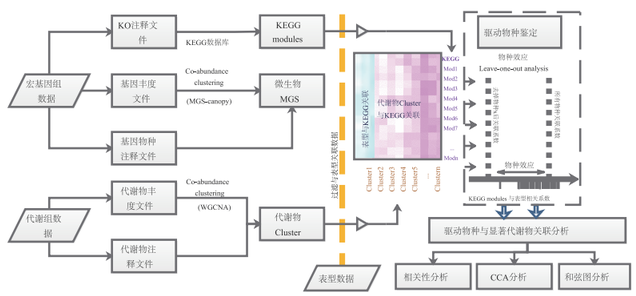


图1 宏基因组代谢组关联分析流程图

4. 部分信息分析结果展示

1) 差异物种/KO丰度热图

差异物种/KO丰度热图是以颜色梯度代表物种相对丰度大小，并根据物种/功能或样品丰度相似性进行聚类的一种图形展示方式。聚类结果加上样品的处理或取样环境分组信息，可以直观展示相同处理或相似环境样品的聚类情况，并反映样品的群落组成/功能组成相似性和差异性。

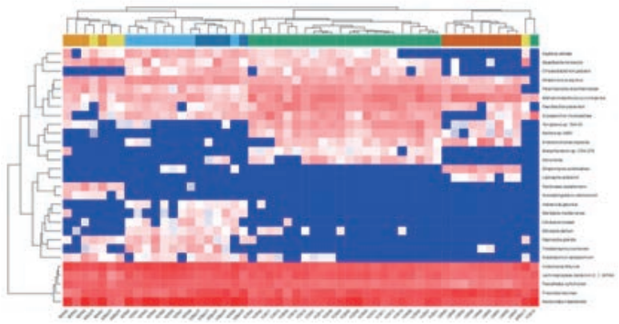


图2 差异物种丰度热图

2) pathway差异分析

通路的差异很难通过单独的KO差异去反映整体变化。ReporterScore方法把涉及 某一通路的所有KO进行统计检验，用整体KO的累计趋势去反映该通路的变化。

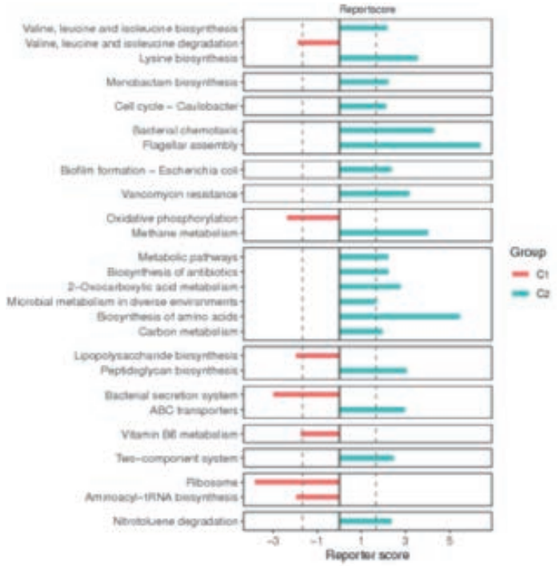


图3 通路差异图

3) CCA分析

CCA是基于环境因子约束的直接梯度分析，将排序分析与多元回归分析相结合，每 一步计算与环境因子进行回归。CCA又称多元直接梯度分析，基于对应分析CA 发展 而来，主要用于分析物种或功能与临床指标之间关系。

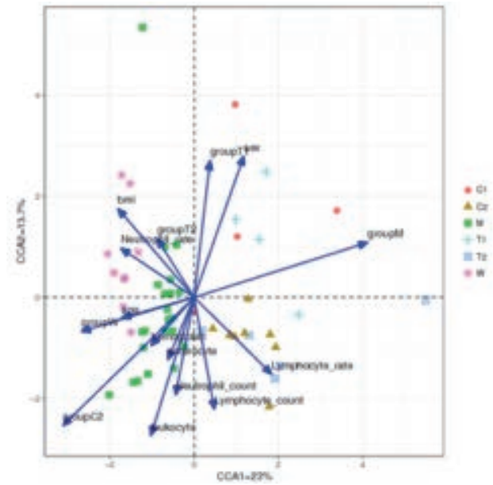


图4 物种CCA图

图中点表示样本，不同颜色或形状表示不同环境或条件下的分组;箭头表示环境因子，箭头连线的长度 代表某个环境因子与群落分布和种类分布间相关程度的大小，连线越长，说明相关性越大，反之越小; 灰色箭头示不同的物种;物种与环境因子之间的夹角代表物种与环境因子间的正、负相关关系(锐角: 正相关;钝角:负相关;直角:无相关性);由不同的样本向各环境因子做垂线，投影点越相近说明样 本间该环境因子属性值越相似，即环境因子对样品的影响程度相当;由不同的样本向各物种做垂线，投 影点的相对位置可以代表该物种在这些样方中多度值的排序情况，在反向延长线上则代表小于平均多度 值，反之则大于平均值;环境因子之间(或物种之间)的夹角反映它们的相关性;样方之间的距离代表它 们的之间物种组成的相似性，距离越近，相似度越高。

4) 差异物种Spearman热图分析

根据秩和检验得到某一特定水平的差异物种，通过R软件的绘制优势物种与表型之间Spearman相关性热图。将差异物种与临床指标进行关联分析，用于发现优势物种与表型之间的重要模式与关系。

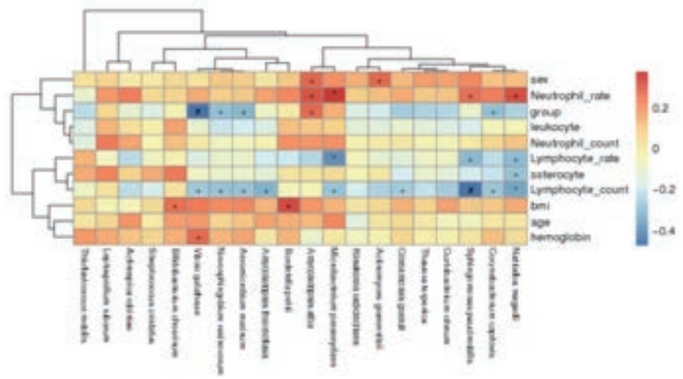


图5 差异物种Spearman热图

图中展示所有分类水平上差异物种与表型之间的相关性，颜色越深，表明物种和表型之间的相关性越强。

5) 代谢物与微生物模块跨组学相关性分析

将代谢物与微生物进行spearman相关性分析。对于有表型数据的宏基因组数据，则会将与表型显著关联的代谢物clusters和与表型显著关联的KEGG 功能模块进行跨组学关联，揭示肠道微生物与代谢物之间的相互作用关系。如图6所示，

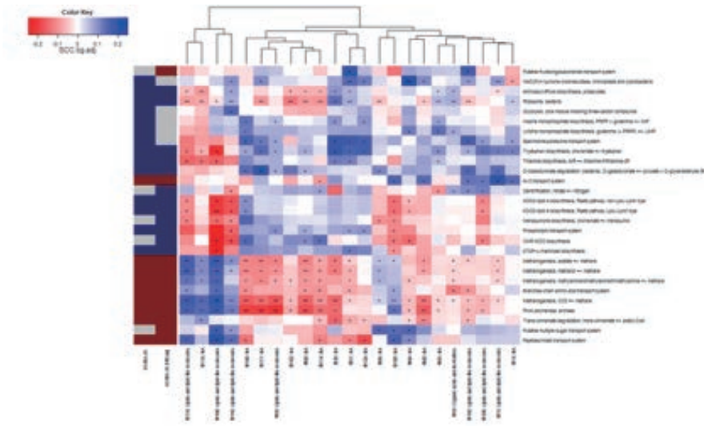


图6微生物KEGG功能模块与表型特征、代谢物相关性热图

左边的panel代表微生物KEGG功能模块与表型特征显著关联，右边的panel代表微生物KEGG功能模块与代谢物clusters显著关联。+号代表FDR<0.1，*号代表FDR<0.01; **代表 FDR<0.001。

6) 关联代谢物-微生物CCA分析

基于代谢物与微生物spearman相关性分析的结果，筛选出相关性显著的关联代谢物-微生物进一步做CCA分析，如图7所示。

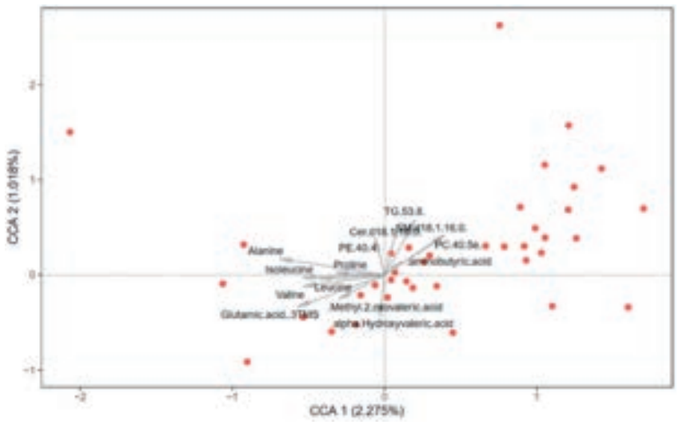


图7 驱动物种和显著代谢物CCA图

图中每个点代表一个驱动物种，每个箭头代表一个代谢物，箭头的连线长度代表某个代谢物与微生物种类分布间相关性程度的大小，连线越长，相关性越大。反之越小。

7) 关联代谢物-微生物网络图

基于前面的所有代谢物和所有微生物的spearman相关性分析和CCA分析，取两种关联分析结果中相关性大于0.5，Pvalue<0.05的代谢物/微生物的交集，用网络图展示：

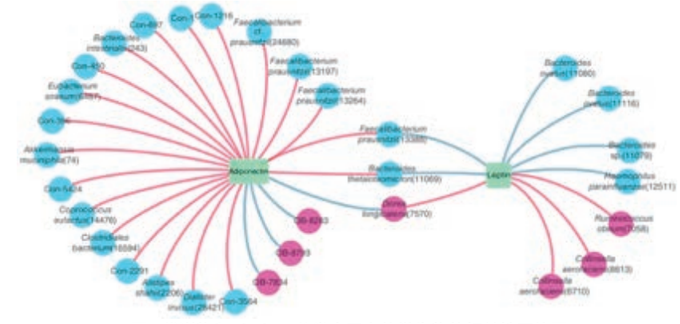


图8 代谢物和Phylum分类水平上的微生物的相关性网络图

方型为代谢物；圆形为微生物，红色和蓝色分别代表在不同组别富集的物种；红色线条代表正相关，蓝色线条为负相关。

5. 项目执行周期

样品检测合格后，建库+测序+标准信息分析（50个样本以内）：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。

6. 预期结果

1) 发现疾病与肠道菌群的关联/因果关系

2) 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。

3) 明确人体微生物在疾病发生发展中的作用及作用机制。

4) 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因素。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。

7. 后期验证手段

1) 扩大群体验证（Additional cohort）

选择新的群体中验证研究结果，以减少技术本身造成的误差。由于人体微生物组在不同人群中差异较大，这对后续的验证工作带来了较大困难；不同的群体由于地理位置、饮食习惯、BMI及年龄的不同往往会导致验证结果的差异。

2) 时间纵向研究（Longitudinal cohorts ）

对同一人群进行不同时间点研究，分析疾病发生发展过程中人体微生物和疾病的相互关系，发现和疾病发生具有因果关系的物种/基因/功能。也可以用于比较干预前后人体微生物的改变（如饮食干预、药物治疗和菌群移植）。

3) 多组学交叉验证

如宏转录组学、宏蛋白组学以及代谢组学等交叉验证。

4) 动物实验

动物实验可以通过严格的实验设计，控制环境因素及宿主因素对结果的影响，有利于大量生理生化指标的检测。常用模型：（伪）无菌小鼠/大鼠，特定病原缺陷的小鼠。

5) 体外功能验证

通过深入的体外功能研究明确meta研究中得到的与疾病相关的marker的作用机制。

常见方法：特定菌株的微生物学研究，菌种间相互作用研究，菌群和宿主（分子、细胞或组织）的相互作用研究。

研究案例

案例一：肠道与心血管疾病

Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk.

发表期刊：*Circulation Research*

影响因子：15.211

发表时间：2019年4月

研究背景

虽然已有研究报道了 心血管病（CVD） 肠道微生物的研究，但是 CVD 中肠道微生物功能和饮食 - 微生物 - 代谢 - 免疫相互作用的研究知之甚少，因此需要通过多组学基于系统生物学的方法去探究 CVD 中饮食 - 微生物 - 代谢 - 免疫相互作用。

样本来源及方法

LifeLines-DEEP队列（LLD）：本研究中 LLD 队列是荷兰 LifeLines 队列中的子队列，剔除了 57 名服用抗生素或降脂药物和 11 名非禁食受试者的参与者，剩下的 978 名受试者（411 名男性和 567 名女性）的粪便宏基因组、血清代谢组和详细的饮食、12 种炎症标志物和 5 种粪便的短链脂肪酸（SCFAs）用于分析。

300 肥胖队列 (300-OB)：纳入了体重指数（BMI）> 27 kg / m2，年龄在 55 至 80 岁之间的受试者，进行了详细的心脏代谢表型分析，包括评估颈动脉斑块和测量皮下和内脏脂肪组织和肝脏脂肪含量。队列纳入要求比较严格，排除各种手术史、用药史和近期心脏事件。

结果与分析

1. 经过质控和估计 ~2.2% 缺失值后，确定了 188 个微生物物种、562 个细菌代谢途径和 231 个代谢性状进行后续的关联分析。在校正年龄，性别和 BMI 后，代谢物分别与物种和代谢途径在两个队列的关联情况如下：

	LLD	300-OB
微生物 代谢物	210 个代谢物与12个物种存有64个关联关系	42个代谢物与微生物物种有一种关联
微生物 pathway	210 个代谢物与308个pathway 存在4,135个关联关系	42个代谢物与19个pathway存有105个关联关系

其中，300-OB 关联数较少可能原因是受到样品数的影响，但结果显示 LLD 中鉴定的微生物因子通常对 300-OB 的代谢变异具有较低的预测值，进一步比较了两个队列的排前发关联强度和方向，即便 300-OB 样品数少，但是存在一些 LLD 没有相关关联，特别是脂蛋白亚类，同时有些关联在两个队列相关方向是相反的。这些结论表明基于人群和肥胖群体之间的微生物关联存在一些强有力的差异。

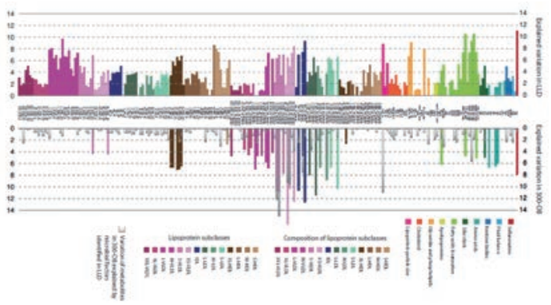


图9 肠道微生物组解释的血浆代谢物变化

2. 为了探索血浆代谢物相关的微生物因子是否与心血管疾病临床表现相关，作者重点关注在2 个队列中相关系数排前的物种和 pathway，并计算它们与 26 个心脏临床表型的相关性。结果显示，在 300-OB 队列中，高丰度的细菌L-甲硫氨酸生物合成与斑块存在和最大狭窄程度显著相关；瘤胃球菌 sp_5_1_39BFAA 与肝脏脂肪含量（Liver-fat）呈正相关。虽然已有文献报道肠道微生物组衍生的 TMAO 会增加心血管疾病的风险。但作者观察到本研究中，TMAO 的血浆水平与内脏脂肪呈正相关，但与动脉粥样硬化表型和肝脂肪不相关，也与代谢物相关物种和细菌通路不相关。

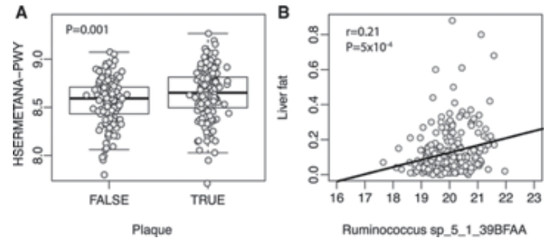


图10 300-OB队列中肠道微生物通路和物种与临床结果的关联

3. 为了评估人群队列中 CVD 病情发展的个体代谢风险，作者使用 33 个确定与未来 CVD 事件相关但又不和已知风险因素（年龄、性别、BMI 和吸烟状况）相关的代谢物 biomarkers 基于加权风险模型构建个体的 CVD 代谢风险评分（MRS）。在 LLD 队列中，发现了 48 个微生物 pathway 与 MRS 关联。另外还鉴定了与 MRS 关联的 pathway 的驱动物种（drive species），结果表明顶级分类群对 MRS 相关的微生物 pathway 贡献差异很大，相关系数介于 0.26 到 0.89 之间，平均值为 0.60。这表明一些 pathway 是由一种占主导地位的微生物驱动，而其他 pathway 可能由许多不同的微生物驱动。

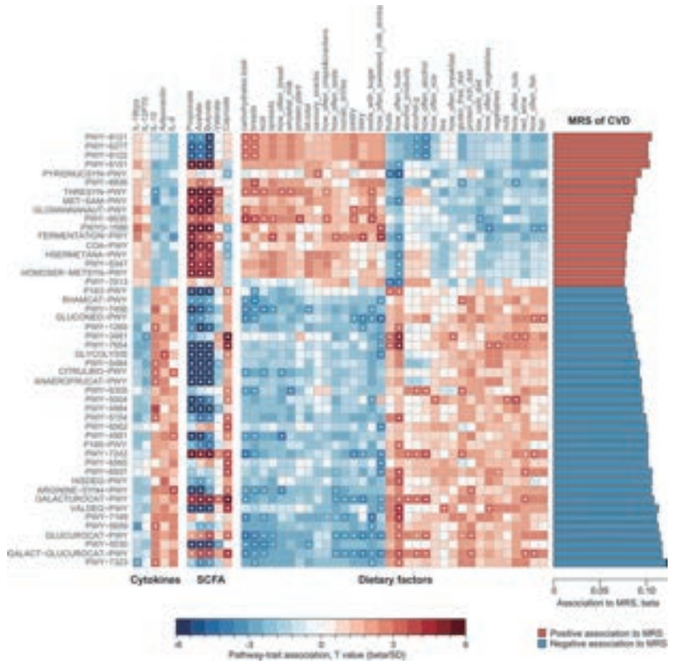


图11 48个生物通路与CVD MRS、粪便SCFAs、细胞因子和饮食因素的关联

4. 作者还深入探讨了宿主-微生物-饮食相互作用对代谢和炎症的影响，将与 MRS 相关的 48 个细菌 pathway 与血浆水平的 12 种细胞因子和 78 个饮食因子进行系统整合分析，结果发现12 个pathway 和 5 种细胞因子存有 14 种显著关联。已知饮食会影响代谢，CVD 风险和肠道微生物，结果发现78 个饮食因素中有 34 个与 MRS 相关的微生物通路有关。MRS 与饮食关联发现，水果和蔬菜的摄入量与 48个 MRS 相关的 pathway 有关，对于这个现象可能的解释机制是细菌通过发酵膳食纤维产生 SCFAs 以影响 CVD。于是作者测定了粪便中 5 种 SCFAs 浓度，结果表明这 5 种 SCFAs 与大部分的 MRS 相关的 pathway 显著关联，进一步分析了血浆中的 SCFA 浓度最高的乙酸盐，正如预期一样，血浆乙酸盐的浓度与低 MRS 和 29 个 MRS 相关的 pathway 相关。

案例二：肠道菌群代谢与HIV感染炎症之间的相互作用

Interplay between gut microbiota metabolism and inflammation in HIV infection

发表期刊：*The ISME Journal*

影响因子：9.520

发表时间：2018年4月

HIV感染导致肠道相关淋巴组织的破坏，使得肠道菌群的组成发生变化。但目前还缺乏关于HIV感染如何影响微生物与宿主之间相互作用的研究。本文结合宏基因组和宏转录组数据来研究HIV相关微生物的功能修饰。结果表明，HIV相关微生物一方面很好地适应炎症环境，例如抗氧化应激反应通路的高表达、抗炎反应过程的低表达，另一方面促进肠道炎症的发生发展。本文通过共发生网络和代谢网络，分析了群落结构中物种标记物和代谢标记物的相关性。通过贝叶斯网络，发现了维持群落结构稳定性的关键通路。此外，确定了群落中各物种对代谢活动的贡献，以及各物种和宿主健康的相互作用。

研究对象：病毒血症患者12例（VU=12），免疫应答者18例（IR =18），免疫无应答者9例（INR=9），以及未感染HIV的健康对照15例（HIV-=15）。

研究方法：宏基因组、宏转录组、代谢组学

主要结果：

1）HIV+组和HIV-组样本中KO基因含量的差异性很大。HIV+样本菌群表现为促炎症反应通路（ko00540、ko05111、ko05120）的增加，抗氧化应激反应通路（ko00250、ko00908）的增加，而信号转导和膜转运通路的减少。HIV+样本菌群表现为Prevotella、Acidaminococcus、Streptococcus等菌属的增加，而Bacteroides、Bifidobacterium、Akkermansia、Odoribacter、Alistipes等菌属的减少。

2）宏转录组数据结果表明，HIV+组和HIV-组样本中RNA-KO基因含量的差异性很大。HIV+样本菌群表现为应激反应通路（ko04141、ko00521、ko00730、ko00053）的转录本丰度增加，而抗炎症代谢过程（ko00650、ko00640、ko00071）的丰度减少。基于宏转录组数据的物种注释结果，HIV+样本菌群表现为Prevotella、Acidaminococcus、Coprobacillus、Streptococcus等菌属的增加。这些被认为是转录活跃细菌。

3）使用广义线性模型GLM分析，发现HIV+组中细菌物种标记物与通路标记物为正相关。发现物种标志物提供了涉及其相关代谢通路的基因，而且普氏菌属各个菌种都提供了涉及HIV相关通路的基因，表明普氏菌属在HIV发病过程中的重要作用。

4）物种关系网络图产生了20个模块，各模块都包含3个以上物种，最大模块包含34个物种。大多数模块的主导细菌为厚壁菌门与拟杆菌门，少数模块的主导细菌为放线菌门和变形菌门。结果还发现，如果模块中包含双歧杆菌属（放线菌门），则该模块中不存在变形菌门，推测放线菌门和变形菌门有竞争关系。

可能存在的风险

人体微生物与对应疾病无显著的关联；宏基因组和代谢组检测结果矛盾，得到的biomarker验证失败。由于肠道菌群会受到各种因素的影响，为了保证分析结果的可靠性，我们在群体选择过程中需要严格进行相关条件控制。如果有的条件难以控制，比如年龄或BMI等，请务必如实详细记录对应的信息，这样在后续分析中可能通过相应的分析方法一定程度上降低这些因素的影响。

华大优势

- **宏基因组学优势：测序准确性高-**DNBSEQ平台滚环扩增构建DNB测序文库，PCR扩增错误不会累积，高保真序列信息；**Duplication率低-**DNBSEQ平台Duplication率低，同样的数据量有效数据多出3%-17%；**无index hopping担忧-**DNBSEQ平台无index hopping担忧，结果更可靠；**样本需求量低-**常规宏基因组建库建议样本量在500ng以上，样本量需求低于同行其他公司要求；对于样本获取困难的样本，也可以选择微量建库，样本量可低至几ng。
- **代谢组学优势：先进的检测平台—**检测仪器包括Thermo Q-Exactive、Waters XEVO-G2XS-QTOF、Sciex 5500/6500+等；**强大的数据库-**使用的数据库包括自建库、mzCloud、LipidSearch等，鉴定能最大化，提供更多更可靠的鉴定结果；强大的信息分析团队-自主开发代谢组学分析软件包metaX用于项目分析。

性价比高：自主检测平台，成本可控。

经验丰富：承担MetaHIT等多个大型国际合作项目，有丰富的项目经验，已发表文章100+，其中CNS系列文章26篇。

合作模式：已合作发表多篇高水平文章。提供切实可行的项目方案，兼顾商业合作、科研合作优势。

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。

2. 多样本间表达模式聚类

为了找寻到比较组间的蛋白表达变化规律，可通过多样本间表达模式聚类快速实现。多样本间表达模式聚类一般涉及到比较组和差异蛋白两个维度的聚类分析：通过对比较组进行聚类，可快速找到疾病发生发展或药效作用机理的规律；通过对差异蛋白的聚类，可对样本内蛋白表达量变化进行归类，以期获得具有一致表达变化规律的蛋白簇。

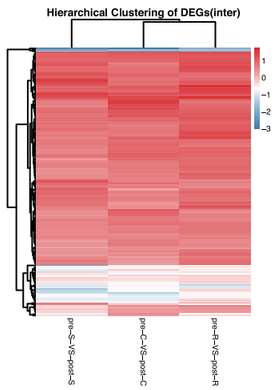


图4 差异蛋白的表达聚类图

该图X轴为比较组名称，Y轴为差异蛋白名称。这里是用Euclidean距离和系统聚类方法(Hierarchical Cluster)来对差异蛋白进行聚类

3. 时间序列分析（仅限蛋白DIA定量分析产品）

疾病发生发展过程，或用药周期中，样本蛋白表达丰度变化规律是科研工作者需要深入开展研究的领域，对于疾病早诊、药效评估有着深远意义。时间序列分析，可以帮助科研工作者迅速找到样品中不同时间点表达模式一致的蛋白簇。

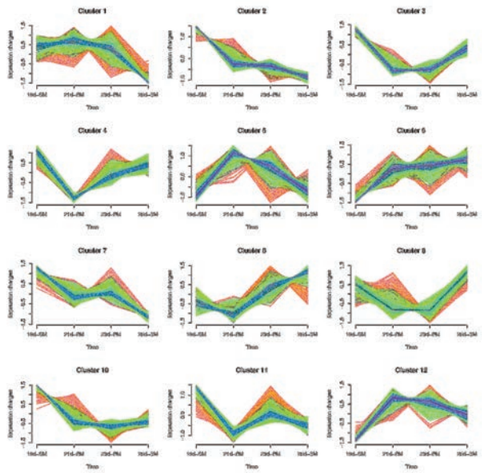


图5 时间序列分析图

X轴代表各个时间点，Y轴代表聚类到一起的蛋白簇均一化后的表达量值。

4. 差异表达蛋白互作分析

蛋白之间通常通过相互作用结合成复合物之后行使相应的功能，所以研究核心蛋白及其互作蛋白至关重要。差异表达蛋白互作分析，通过与STRING蛋白互作数据库比对，对差异表达蛋白进行互作分析，取可信度排名前100的互作关系关系绘制网络互作图：

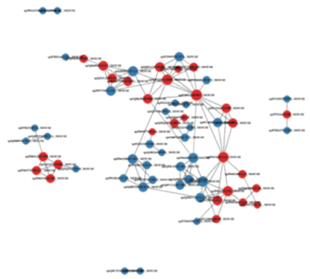


图6 差异蛋白互作关系网络图

图中红色表示上调蛋白，蓝色表示下调蛋白，圈的大小表示关系密集程度。

5. 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结果进行比较后，进一步缩小蛋白标志物的范围。

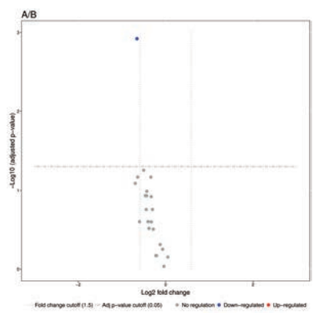


图7 目标蛋白的相对定量分布

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倍差异且校正p值<0.05的条件下认为是差异蛋白。

D. 项目执行周期

样品提取检测合格后，蛋白iTRAQ定量分析标准周期约20个工作日完成；蛋白DIA定量分析标准周期约40个工作日完成；蛋白定量MRM技术标准周期约40个工作日完成。实际项目完成时间由所选样本数、上机情况以及信息分析条款决定。

E. 预期的结果

本方案期望结合多种蛋白质组学定量技术手段，借助高通量质谱平台，通过对疾病和正常样本（或其他相关疾病样本）、用药前后样本、服用不同药物样本之间的比较，寻找特异性好、分辨度高、易于取样的疾病诊断蛋白生物标志物，或药效相关生物标志物，用于疾病早诊及药效评估工作。

辅助研究策略（可选）

可以通过免疫组化技术，验证疾病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。

应用案例

案例一：寻找结直肠癌发生发展过程中的蛋白生物标志物-BGI[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3个蛋白标志物在16组临床血清样本中进行验证，其中2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。

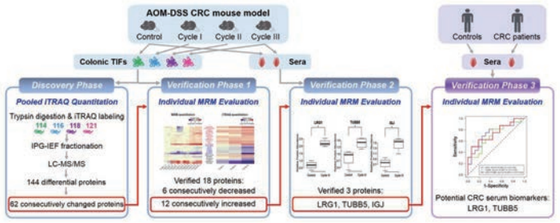


图8 实验整体思路

案例二：寻找胰腺癌早期诊断及确诊的蛋白生物标志物

1. 第一项关于胰腺癌的研究[2]，主要是寻找胰腺癌早期诊断标志物。文章采用iTRAQ技术对乙酰胆管腺癌（PDAC）病人的血清样本进行，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含量低于糖尿病相关。TSP-1可以和CA19-9一起作为早期诊断PDAC病的标志物。

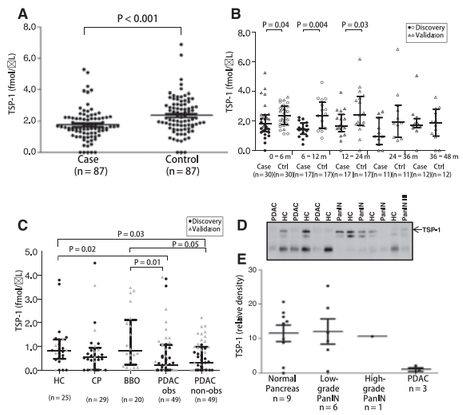


图9 MRM方法对346个样本中的TSP-1蛋白进行验证

2. 第二项关于胰腺癌的研究[3]，目的是找到新型胰腺癌确诊标志物。文章采用iTRAQ-2DLC-MS/MS技术联合目标蛋白定量P-MRM和SID-MRM技术，分析健康人群（NC）、良性疾病（BD）和胰腺癌患者（PC）三种人群约100例血清样本，鉴定到超过一千种的蛋白类型，证实其中142个蛋白表达量发生变化，最后选取4个蛋白进行绝对定量。新发现的生物标志物组合包含APOE、ITIH3、APOA1、APOL1，与原有的CA19-9标志物相比，在诊断PC方面，在敏感性（95%）和特异性（94.1%）方面都有显著的统计学改善。

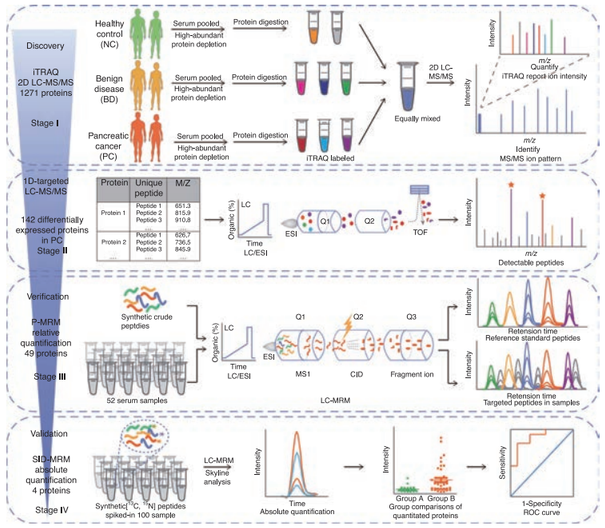


图10 基于质谱方法寻找胰腺癌全新蛋白标志物的整体流程

案例三：寻找成人发病型糖尿病（MODY）血清蛋白标志物-BGI[4]

MODY是一种单基因遗传型糖尿病。本文采集了一个维吾尔族MODY家庭三代人的血清样本，通过非靶向蛋白组定量技术与靶向蛋白组定量技术联合，初步发现了32个血清蛋白与MODY相关。进一步的验证实验发现，MODY患者中有12个候选的显著变化的蛋白标志物。这12种候选蛋白标志物，在一型糖尿病、二型糖尿病、MODY以及健康人群中进行测试的结果显示，与MODY明确相关的特异性标志物是SERPINA7, APOC4, LPA, C6, 和F5。

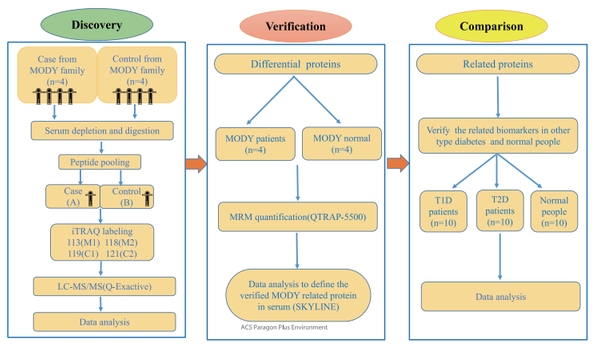


图11 本研究寻找MODY蛋白标志物的整体流程

案例四：二型糖尿病两种药物药效评估-BGI[5]

本研究选取服用两种不同药物消渴丸和格列本脲前后、是否出现低血糖症的二型糖尿病患者共32例，获得血清后，采用iTRAQ研究蛋白质组表达量变化。发现服用消渴丸后未出现和出现低血糖患者血液中存在25和21个差异表达蛋白、服用格列本脲后存在24和25个差异蛋白；服用同种药物有无低血糖症患者血液中差异表达蛋白的交集、比服用不同药物的多。推测参与两种抗糖尿病药物作用机制的血清蛋白可以作为评估疗效的biomarker、且对深入研究作用机制有着至关重要的作用。

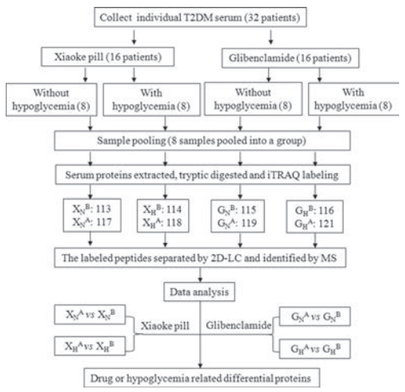


图12 本文整体实验设计

案例五：寻找肺癌病人蛋白标志物[6]

近期在仪器和生物信息分析流程上的技术进步，使得我们关注到一种新的定量质谱平台采集模式——数据非依赖型采集（DIA），以及目标分析法——平行反应监测（PRM），作为成熟方法数据依赖性采集（DDA）及多反应监测（MRM）的变换方式。这些工具可以用于监测肺癌及其他恶性癌症的信号扰动，支撑高效的激酶抑制改变，提供研究治疗抗性机制和药物改造的新契机。为了检验不同的质谱平台采集模式的有效性，ATP激酶定量伴随着后续的抑制治疗法通过四个不同的方法进行：LC-MS/MS-DDA/DIA，LC-MRM/PRM。在发现数据集中，DIA相比于DDA 增加了21%的激酶鉴定数并减少了缺失值，同时在这个范畴内，MRM和PRM相较于抑制治疗法都显示出更有效的鉴定率，显示出先验实验以及定量蛋白质组数据集的价值。

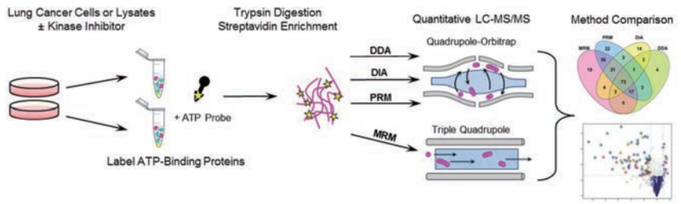


图13 本文整体实验设计

案例六：通过DIA-MS联合MRM方法找寻致关节炎的相关蛋白标志物[7]

人HLA-B27蛋白的表达与强直性脊柱炎及其他脊椎关节病有强相关性。为了检测HLA-B27肽段亚型中更多的细微量的变化，本研究采用数据非依赖型采集（DIA）和多反应检测技术（MRM）相结合的方法，通过8种常见的HLA-B27同种抗免疫球蛋白（HLA-B*27:02-HLA-B*27:09），定量了1646条HLA-B27的肽段丰度。本文参考这8种HKA-B27同种抗免疫球蛋白，采用Kmeans聚类的方法将相同等位基因肽段分组，使得我们能够用最严格的结合特征发现每一种HLZ-B27亚型，从而进一步细化

已有的一致的结果模块。而且，本文对于定量数据库的深入分析发现位于HLA-B*27:06和HLA-B*27:09的26条肽段，与疾病相关的HLA-B27亚型相比，具有更低的表达丰度。

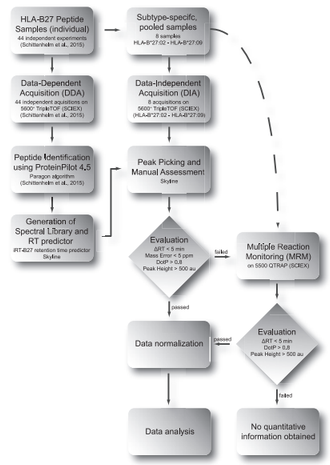


图14 本文整体实验设计

可能存在的风险

在项目实施过程中，由于某些蛋白的丰度低，蛋白iTRAQ/DIA定量和MRM技术方法存在一定的差异，可能存在用iTRAQ/DIA初步筛选出的差异蛋白，用MRM方法验证不到的情况，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。

常见问题

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。谈到蛋白质全局变化，那么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的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。

2. 只通过蛋白iTRAQ/DIA定量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iTRAQ/DIA定量技术找到的差异蛋白，一般还需要通过其他技术进行验证后，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。

华大优势

成熟的非靶向蛋白定量和靶向MRM技术，项目经验达千个以上[8,9]，涉及人类、动植物、微生物等多个研究领域；

完善的非靶向蛋白定量和靶向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[1,10]。

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 Proteomics. 2016 Jan 30.

[2] Jenkinson C, Elliott VL, 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. Clin Cancer Res. 2016 Apr 1.

[3] Liu X, Zheng W, Wang W, et al. A new panel of pancreatic cancer biomarkers discovered using a mass spectrometry-based pipeline. Br J Cancer. 2018 Mar 20.

[4] Tuerxunyiming M, Xian F, Zi J, et al. Quantitative Evaluation of Serum Proteins Uncovers a Protein Signature Related to Maturity-Onset Diabetes of the Young (MODY). J Proteome Res. 2018 Jan 5.

[5] Zhang X, Sun H, Paul SK, et al. The serum protein responses to treatment with Xiaoke Pill and Glibenclamide in type 2 diabetes patients. Clin Proteomics. 2017 May 17.

[6] Hoffman MA,Fang B,Haura EB, et al. Comparison of Quantitative Mass Spectrometry Platforms for Monitoring Kinase ATP Probe Uptake in Lung Cancer. J Proteome Res. 2018 Jan 5

[7] Schittenhelm RB,Sivaneswaran S,Lim Kam Sian TC, et al. Human Leukocyte Antigen (HLA) B27 Allotype-Specific Binding and Candidate Arthritogenic Peptides Revealed through Heuristic Clustering of Data-independent Acquisition Mass Spectrometry (DIA-MS) Data. Mol Cell Proteomics. 2016 Jun.

[8] 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. PLoS One. 2015 Oct 8.

[9] Liu L, Li G, Sun P, et al. Experimental verification and molecular basis of active immunization against fungal pathogens in termites. Sci Rep. 2015 Oct 13.

[10] 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. PLoS One. 2013 Sep 18

药物作用机制的基因表达研究方案

研究背景

药物发展的两个目标是有效性和安全性，也就是了解某药的作用和副作用。研究药物的作用（例如减轻疼痛、降低血压、降低血浆胆固醇水平），还需要研究药物在什么部位和怎样发挥作用（即作用机制）。虽然药物作用比较容易显现，但其作用部位和机制不可能很快弄清楚。例如，阿片和吗啡用于镇痛和治疗抑郁已有几百年了，但仅仅是不久前才发现与镇痛欣快有关的大脑结构和脑化学成分。并且也要关注药物间的相互作用，当两种药物同时使用(药-药相互作用)或食用某些食物(药-食物相互作用)，该药的作用可被其影响，称为药物的相互作用。尽管联合用药有时是有益的，但多数时候是无益的甚至是有害的。

另一方面，虽然药物可治疗疾病，但也伴随出现一些副作用或不良反应，或者是产生耐药性，如何克服耐药性是一个研究重点。耐药性是指微生物、寄生虫以及肿瘤细胞对于化疗药物作用的耐受性，耐药性一旦产生，药物的化疗作用就明显下降。耐药性根据其发生原因可分为获得耐药性和天然耐药性。

即使弄清楚了药物的作用机制，药物如何高效生产也是一个挑战。例如青蒿素，是众所周知治疗疟疾的特效药，并且其在红斑狼疮、糖尿病、恶性肿瘤等疾病治疗上的应用价值逐步被发现，使青蒿素需求量在全球范围内剧增。因此提高青蒿中青蒿素含量，进而培养高产品种，是现在的全球研究的热点。

从小群体或个体角度看，同一种药，不同人群或个人服用后对药物反应程度可能不同，许多因素可影响药物的作用，如受遗传因素、体重差异、不同年龄阶段（如新生儿和老年人对药物的代谢慢于儿童和青年人）、疾病（如肝肾病患者对药物的清除比正常人困难）等因素影响。

mRNA测序技术可以更好地在分子水平了解患者服用某药物后的表达层面上的反应，通常的方法是研究在实验组与对照组的蛋白编码mRNA水平上的差异基因表达（DGE）变化，揭示药物发挥作用、不良反应和耐药性的分子机制；也可以深入挖掘药物合成的调控网络，用于研究药物高效生产的分子机制，培育高产品种；以及目标人群及个体的用药指导。在药物研发和临床研究中就需要快速评估成千上万份样本中的基因表达变化，mRNA测序技术可以帮助研发者更快速高效地进行药物筛选和研究。

方案设计

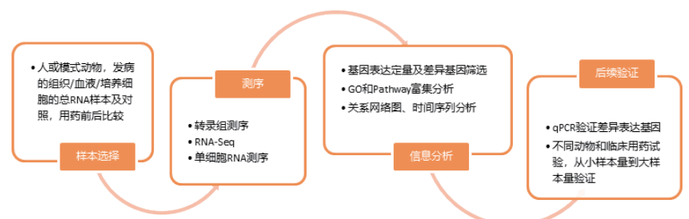


图1 实验流程

A. 样本选择建议

1. 设计不同的实验条件，例如正常样本和患病样本，加入药物前后；应设阴性和阳性对照组，阴性对照为不含药物的空白基质，阳性对照为已被证明有效的且给药途径相同的药物。

2. 变量设计一般考虑在加入药物时，可设置不同浓度，来研究最合适的药物剂量。一般应设三个剂量组以体现量效关系。如果药物作用于多个组织器官，也可考虑不同组织器官的细胞。此外可以采集处理不同的时间点、在不同模式动物中，在不同性别、年龄的样本中试验，以及不同药物的疗效，比较哪种药物疗效最佳。

3. 初步试验中建议3个以上的生物学重复；后期试验需至少几十个至上百个样本大样本量验证。

4. 可结合表观组学、蛋白组学等其他研究技术，注意尽量采用相同样品。

B. 采用的技术

采用转录组测序或者RNA-Seq测序技术，通过样本间的比较及筛选，寻找差异表达的基因，并对差异表达的基因进行GO和Pathway的富集分析等。

C. 测序参数

建议转录组项目每个样本6G-10G clean data，RNA-Seq每个样本20Mb clean reads。

D. 分析结果

1. 差异表达基因分析

根据各个样品基因表达水平数据，我们可以检测样品（或者样品组）之间的差异表达基因。对于设置生物学重复的实验，我们可以采用DEGseq、DEseq2、EBseq、NOIseq进行组间样品基因差异表达分析，从而比较处理组与对照组，得到上下调基因个数。对于无生物学重复样品，则采用PossionDis进行基因差异表达分析。差异表达基因数量、比较组之间共有和特有的差异基因、及差异表达基因层次聚类都可以用不同形式的图片进行展示。

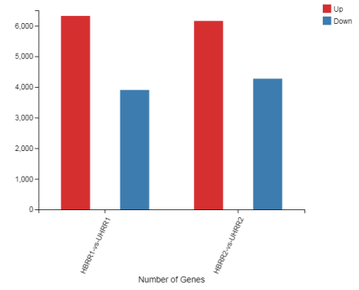


图2差异表达基因数量统计图

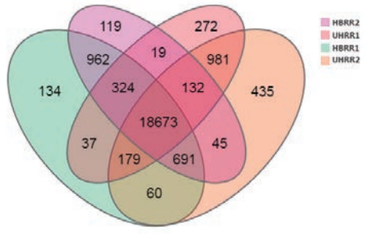


图3 样本特异表达及共有表达基因Venn图

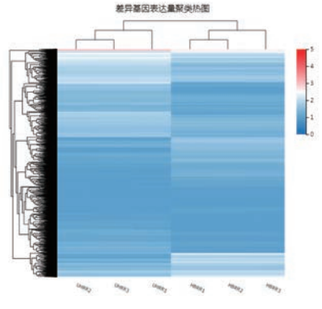


图4 差异表达基因层次聚类热图

2. 差异表达基因GO分析

根据差异基因检测结果，我们对其进行GO功能分类以及富集分析。GO分为分子功能、细胞组分和生物过程三大功能类，我们将对三大功能类单独进行进一步的分类以及富集分析。差异基因GO功能分类与富集结果见图5。

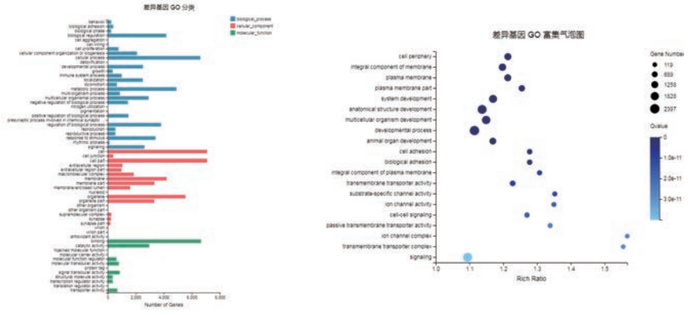


图5 差异基因GO功能分类与富集结果

3. 差异表达基因Pathway分析

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。根据差异基因检测结果，我们对其进行KEGG生物通路分类以及富集分析。差异基因Pathway分类与富集结果见图6。

图6 差异基因Pathway分类与富集结果

4. 蛋白互作网络分析

蛋白之间通常通过相互作用结合成复合物之后行使相应的功能。通过PPI分析，具有相互作用的 DEG 通常具有相似的功能。根据 STRING蛋白互作数据库，对每组差异表达基因进行蛋白互作分析，互作关系结果如下表。

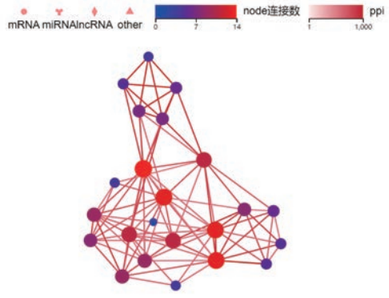


图7蛋白互作网络分析

5. GSEA富集

GSEA和常规富集分析不同，无需做差异分析，直接检测所有基因的表达量即可找到两组间有一致性差异的感兴趣的通路，能找到那些差异不很明显但是基因差异趋势很一致的功能基因集。

使用 Broad Institute 开发的基因集富集分析（GSEA）算法进行分析，帮助识别表达量变化最明显的基因集中在哪些通路、生物学过程或功能组份等。该分析将根据您随后选定的样本，使用所有表达量大于0的基因进行分析。

Dr.Tom现已上线增强版GSEA富集分析工具，多个物种都已经准备好，领头基因集也可以一键获取。

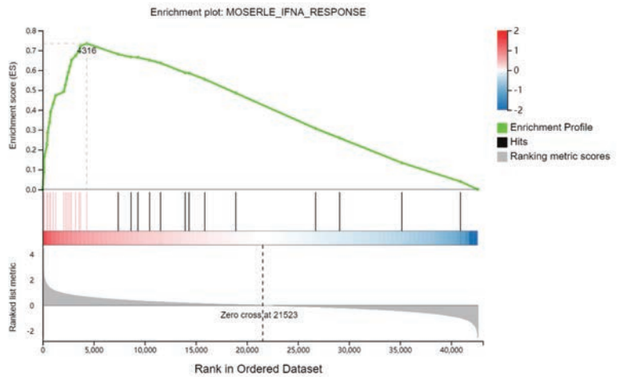


图8 GSEA富集分析图

A. 项目执行周期

样品检测合格后，建库+测序+标准信息分析：转录组测序为24个工作日，RNA-Seq为15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。

B. 预期的结果

利用RNA研究手段，借助高通量测序平台，通过对用药后基因表达变化的分析，从基因表达层面发现药物发挥作用、不良反应和耐药性的分子机制，从而开发新药、提供用药指导等；或者通过研究药物高效生产的分子机制等进行深入挖掘，从而培养高产品种等。

C. 后期验证手段

挑选目标或者候选基因，通过qPCR进行验证，验证测序结果是否和qPCR结果一致。然后从不同动物实验和临床实验进行验证，样品量从少量样本到上百个的大量样本验证。

应用案例

案例一：干扰素治疗乙肝的分子机制[1]

发表期刊：*Cell*

影响因子：30.41

发表时间：2017年7月

全球有乙肝病毒（HBV）感染而导致的乙肝患者超过3.5亿，其中约1/3在中国。除了接种疫苗，目前广泛使用抗病毒天然免疫细胞因子IFNα（I型干扰素）治疗，其分子机制尚不清楚。本文目的为研究抗病毒天然免疫细胞因子IFNα治疗乙肝的分子机制。

研究样本：

高通量小RNA干扰筛选体系筛选、肝细胞特异性敲除基因的小鼠模型、8个人细胞样品RNA-Seq（4组，每组2个生物学重复）

研究结果（RNA-seq部分）：

图9 RNA-Seq研究思路

1. ETD2经SET域（包含片段）促进STAT1活化，从而提高干扰素诱导的抗病毒功能

HepG2细胞中敲除SETD2，产生出3个SETD2-KO细胞系，呈现出SETD2表达缺失和总的H3K36me3水平下降（图10F），与对照HepG2相比，都表现出干扰素诱导的STAT1磷酸化作用受损（图10G）。观察干扰素处理的SETD2-KO HepG2细胞，IFNAR1/R2或 STAT1上游信号分子的表达没有显著差异。通过RNA-Seq分析，发现一系列ISGs响应干扰素刺激，在SETD2-KO细胞比HepG2 WT细胞表达更低（图10H）。qRT-PCR分析确定干扰素刺激后在SETD2-KO细胞中两个ISG表达降低，包括ISG15 和MX2（图10I）。所以，在干扰素刺激下，SETD2提高STAT1磷酸化作用和ISG表达。

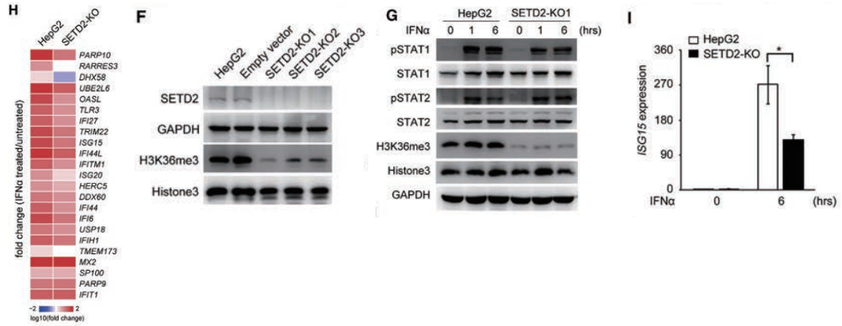


图10 SETD2促进干扰素诱导的STAT1 磷酸化和 ISGs 表达

2. SETD2选择性促进H3K36me3修饰的ISGs表达

利用RNA-Seq，在SETD2-KO细胞、STAT1-K525A-Re细胞、STAT1-KO细胞以及对照的HepG2 细胞在干扰素刺激下6小时（图11A）。干扰素刺激后对照HepG2细胞中222个基因被诱导，SETD2-KO 细胞、STAT1-K525A-Re细胞和STAT1-KO 细胞，与对照HepG2细胞相比，分别有89、128和180个基因显著性下调（图11B）。GO富集分析显示SETD2-KO 细胞和STAT1-K525A-Re细胞中表达更低的基因被归类于病毒的防御反应或I型干扰素通路相关基因（图11C）。STAT1-KO细胞在干扰素刺激下下调的180个基因，其中有40%的基因在SETD2-KO细胞在干扰素刺激下也表达更低，包括SETD2促进一系列的STAT1依赖的ISG（图11D）。另外，在干扰素刺激后的SETD2-KO 细胞和STAT1-K525A-Re细胞中，75个基因表达更低，包括有抗病毒功能的ISGs，比如ISG15、ISG20、MX2等（图11D和11E）。进一步确定了SETD2介导的ISGs调控是在很大程度上依赖于在介导STAT1 K525单甲基化作用中它起到的作用。

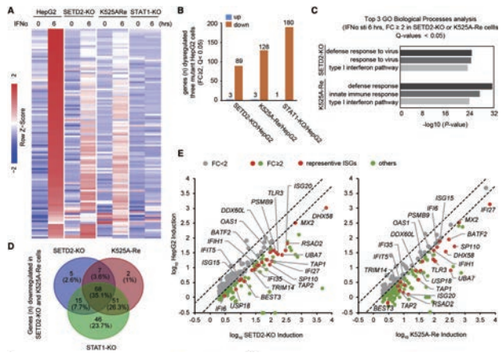


图11 SETD2选择性催化一些ISGs的 H3K36me3 修饰

案例二：药物青蒿素合成的腺毛发育调控机制[2]

发表期刊：*New Phytologist*

影响因子：7.330

发表日期：2018年1月

青蒿是我国传统中草药，因其能合成用于治疗疟疾的特效成分青蒿素而闻名全球。因发现具有抗疟作用的青蒿素，挽救了数百万人的生命，屠呦呦教授获得了2015年诺贝尔生理或医学奖。此外，随着青蒿素及其衍生物在红斑狼疮、糖尿病、恶性肿瘤等疾病治疗上的应用价值逐步被发现，使青蒿素需求量在全球范围内剧增。因此提高青蒿中青蒿素含量成为全球研究的热点。青蒿素主要在青蒿叶片表面的分泌型腺毛中合成和积累，但腺毛发育调控机理仍然不明确。

上海交通大学唐克轩教授研究团队以青蒿为模式，研究植物分泌型腺毛发育机制，先后在New Phytologist 上报道了两个青蒿腺毛发育的正向调控因子AaHD1和AaMIXTA1。

此次又发表在New Phytologist上，发现一个正向调控AaHD1基因表达的HD-ZIP IV转录因子AaHD8。发现AaHD8与AaMIXTA1能够相互作用形成复合体，通过结合AaHD1以及多个蜡质角质合成酶基因启动子上的L1-box，协同调控这些基因的表达，从而促进角质层合成和腺毛的起始发育（图12）。

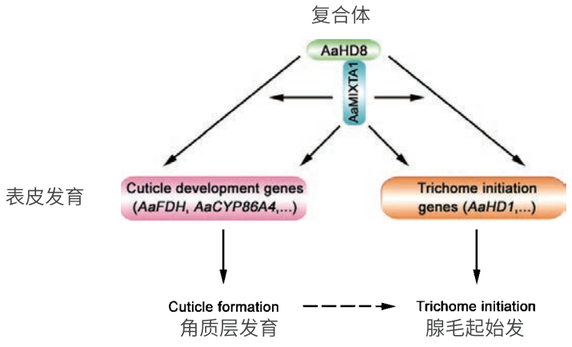


图12 青蒿角质层合成和腺毛的起始发育机制图

首先从表型上，通过AaHD8 RNAi干扰植株转录组数据和生理生化实验分析两个角度研究角质层发育。

然后采集5个月大小的WT和AaHD8 RNAi植株的叶子，利用BGISEQ-500 RNA-Seq技术比较AaHD8沉默株系和WT植株的转录组，鉴定出38个拟南芥同系物的酶参与角质层发育，其中的19个酶在沉默系中显著性地下调表达，与WT相比。在这19个基因中，有9个在启动子上包含至少一个L1-box 或者 HD-ZIP VI 结合位点。这其中有4个是蜡质合成必须的，4个是角质积累必须的，1个是蜡质和角质合成都参与的（表1）。qRT-PCR证实了这9个基因的表达（图13 d, e）。后面通过酵母单杂试验证明这9个角质层合成酶是做为AaHD8直接靶标，及结合区域。

表1 RNA-Seq检测出的角质层合成相关基因(仅截取部分基因展示)

Table S4 The genes related to cuticle biosynthesis in RNA-seq assays

Gene name	Characterized Ortholog	Function	wild-type		iAaHD8		Fold change	P-value	L1-box/ HZBS
			average	SE	average	SE			
AaCYP86A4	*CYP86A7 AT2G45970*	C	33.37	0.16	28.56	0.82	0.85	0.0354	Y
AaFDH	*FDH At2G26250*	W/C	115.85	5.21	80.41	2.53	0.69	0.0202	Y

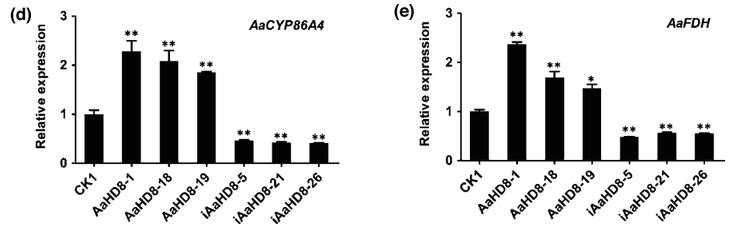


图13 qRT-PCR验证了9个角质层合成酶基因的表达在沉默系中减少

案例三：克服肺腺癌EGFR-TKIs靶向治疗耐药的新机制[3]

发表期刊：*Theranostics*

影响因子：8.766

发表日期：2018年2月

本研究首次发现抑制磷酸甘油酸脱氢酶（PHGDH）能显著抑制肺腺癌耐药细胞增殖，诱导细胞凋亡，逆转耐药。

Erlotinib等表皮生长因子受体酪氨酸激酶抑制剂（EGFR-TKIs）靶向治疗敏感性肺腺癌疗效显著，但易耐药复发。已知耐药机制有靶基因突变、旁/下游通路代偿激活等，代谢重编程调控机制也引起越来越多的关注。磷酸甘油酸脱氢酶（PHGDH）是丝氨酸合成途径限速酶，与多种肿瘤的发生发展密切相关，但是在肿瘤耐药中的作用及其机制还没有研究清楚。

研究人员选取含有EGFR突变（Del19）的肺腺癌细胞系PC9和HCC827，构建了耐Erlotinib细胞模型，RNA-Seq检测发现耐药细胞的PHGDH明显上调(图14)；LC-MS/MS靶向氨基酸代谢组学方法检测细胞培养液中20种游离氨基酸的消耗量，丝氨酸是耐药细胞中消耗量最多的，细胞内和分泌到细胞外的丝氨酸含量也大大增加，14C示踪靶向葡萄糖代谢流分析结果也证实，细胞内葡萄糖来源的丝氨酸含量显著上升；体外细胞实验和体内动物实验结果显示，抑制PHGDH能抑制耐药细胞增殖，诱导细胞凋亡，逆转耐药；反之，过表达PHGDH能促进肺腺癌对Erlotinib耐药性产生；深入研究其机制发现，抑制PHGDH能引起耐药细胞中DNA损伤标志物γH2AX显著上升，较亲本株更为显著，活性氧抑制剂NAC则能逆转其效应；通过检测细胞内还原型谷胱甘肽（GSH）和氧化型谷胱甘肽（GSSG）含量比值发现，抑制PHGDH能明显降低GSH/GSSH比值，揭示抑制PHGDH引起耐药细胞中DNA损伤与PHGDH通过丝氨酸合成途径调节GSH/GSSG含量，调控细胞内活性氧水平有关(图14)。

本研究揭示PHGDH作为肺腺癌EGFR-TKIs耐药的潜在靶标，为PHGDH抑制剂联合EGFR-TKIs治疗克服单用EGFR-TKIs引起的耐药作用提供实验理论依据。

耐药细胞的PHGDH明显上调。RNA-Seq检测发现PC9ER4耐药细胞相比于亲本细胞的PHGDH明显上调(图14C)，然后使用qRT-PCR和免疫印迹法在mRNA水平与蛋白水平上，证实其他耐药细胞中PHGDH也同样上调的结果（图14D E）。

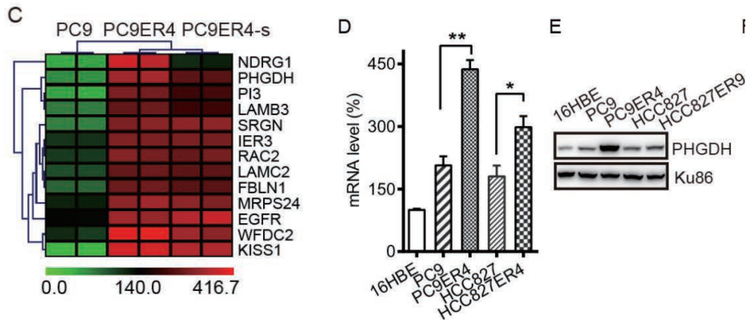


图14 耐药细胞的PHGDH明显上调

PHGDH促进Erlotinib耐药性产生是通过调节与DNA损伤修复与核苷酸代谢有关的转录本实现。通过RNA-Seq检测发现抑制PHGDH引起耐药细胞中1011个基因上调表达（图15A），KEGG通路分析发现1011个基因与多个DNA损失修复和核苷酸代谢密切相关。siPHGDH#4转染72小时后γH2AX在耐药细胞中极端提高，这说明PHGDH抑制引起耐药细胞广泛的DNA损失（图15C）。siPHGDH#4和siPHGDH#5处理后，γH2AX量也在耐药细胞中显著性提高（图15D）。

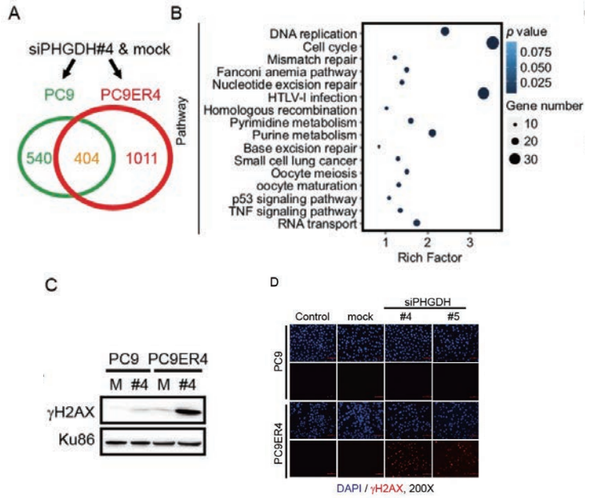


图15 PHGDH促进Erlotinib耐药性产生是通过调节与DNA损伤修复与核苷酸代谢有关的转录本实现

案例四：chidamide和MI-3协同治疗混合系白血病（MLL）基因重组的预后不良机制[4]

发表期刊：*Clinical Epigenetics*

影响因子：5.496

发表日期：2019年10月

研究目的：研究chidamide和MI-3（下面简称为C和M）两种抑制剂，协同治疗混合系白血病（MLL）基因重组的预后不良的机制。

实验设计：24个样品=4组×3个重复×2个时间C和M，处理MLL重组的白血病细胞，加M单药组、加C单药组、加M+C联合组及对照，共4组，每组3个重复，处理24和48小时。华大DNBSEQ平台转录组测序。

Dr. Tom数据挖掘：1、想找处理后有哪些变化的通路？从KEGG富集入手。KEGG富集分析显示C和M联合时，最显著改变的通路，有细胞周期、DNA复制和修复通路。

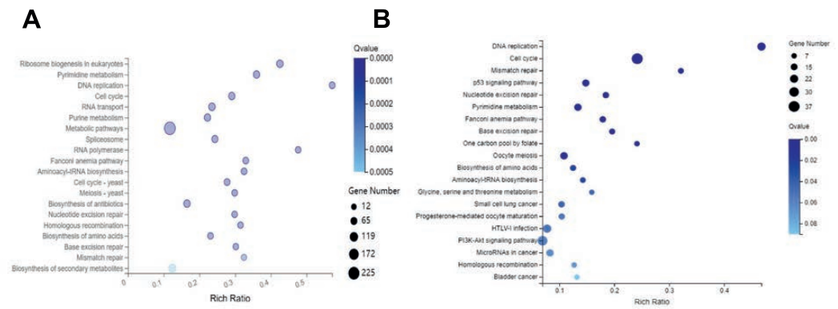


图16 A和B分别是处理24、48小时KEGG富集分析

2、想知道哪种处理效果更明显，引起上面的通路变化？小工具做GSEA。GSEA分析进一步显示这些改变绝大多数源于C，而非M（图17是C处理）。

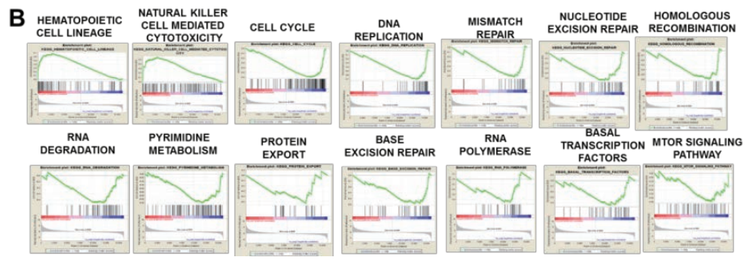


图17 C处理下的GSEA分析

3、想找各处理下差异表达基因交集？从维恩图入手。不同处理下差异基因维恩图，有635个基因在三种处理下都表达（图18C）。

这些基因在各处理下都是一致上调/下调吗？点击图中心，获取基因集，小工具做热图。59个基因表现不同趋势（图18D方框所示），M处理下调，C处理或M+C处理上调。

把上面不同趋势的基因再筛出来？图上框选，获取基因集，小工具做热图。热图（图18E），与第2点结果相印证。

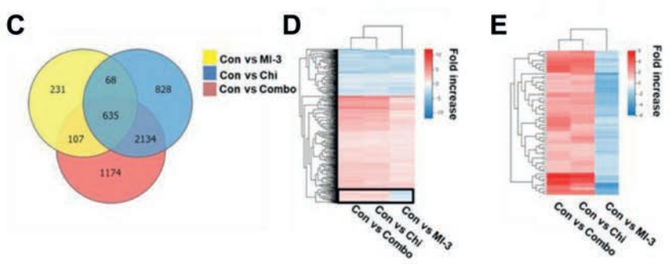


图18 C三种处理在48小时差异基因维恩图，D共有基因聚类热图，E各处理下表达趋势不同的59个基因聚类热图

4、想知道上一步获得基因有什么功能?选小工具分别做GO和KEGG富集。GO和KEGG分析（图19D和E）发现这些基因涉及关键生存信号通路，炎症反应必需的细胞因子通路。

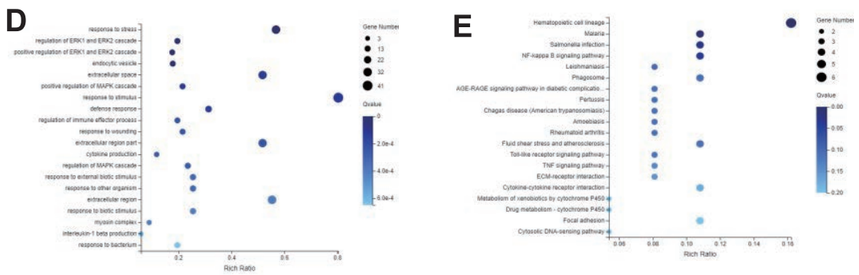


图19 D和E分别是59个基因的GO和KEGG分析

5、想再从中继续筛选寻找关键基因?表格扩展列勾选不同组表达量，调整筛选条件，选小工具做热图。继续设置表格筛选条件：同时至少满足单药组2倍下调，联合组4倍下调，得到4个基因做热图（图20A），与DNA损伤应答有关。

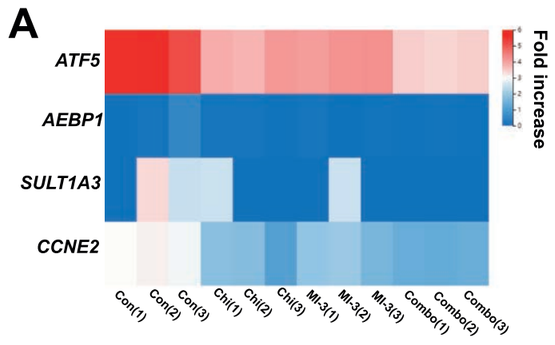


图20 4个关键基因聚类热图

案例五：黄多糖增强高脂高糖饮食小鼠的胰岛素敏感性[5]

发表期刊：*Carbohydrate Polymers*

影响因子：5.158

发表日期：2018年11月

研究概述：桑黄多糖治疗糖尿病效果显著，但其机制尚不清楚。研究者基于高糖高脂饮食的小鼠模型，在BGISEQ平台完成转录组测序，通过层层深入的数据分析挖掘，构建出关键代谢通路及代谢网络，最终锁定一种肠道细菌，阐明了桑黄多糖增强胰岛素敏感性的机制。

实验设计：3组小鼠肝脏样本，每组3个重复；对照组：常规饮食16周；HFD组：高糖高脂饮食12周服用盐水4周；PLP组：高糖高脂饮食12周服用桑黄多糖4周；

1、RNA测序研究

处理3组小鼠肝脏样本，进行生化指标数据采集，其中HFD和PLP两组（每组3个重复样本）基于DNBSEQ平台完成6个样本的RNA-Seq测序。

分析转录组数据，相比于HFD组，在PLP组中1110个基因表达量上调，221个基因表达量下调。通过KEGG代谢通路富集分析，发现差异表达基因主要富集在信号转导通路，而其中的FOXO通路、钙代谢通路等正好与研究者一直关注的性状相关（图21）。

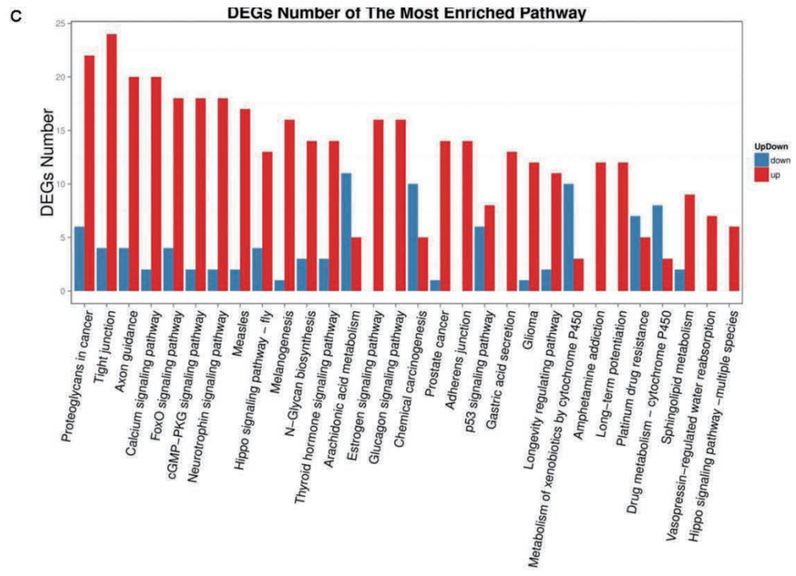


图21 差异表达基因KEGG分析，主要富集在FOXO通路、钙通路（图中红色方框）等

2、桑黄多糖作用于肝脏FOXO代谢通路

聚焦差异表达基因富集分析捕获的FOXO通路。FOXO蛋白作为胰岛素代谢通路上的重要分子，对调控机体血糖的平衡有重要作用。但该通路当中各基因上下调关系未知，研究者通过分析KEGG通路图，发现该通路上的关键基因总体上都是上调（图22D），后续qPCR验证与测序结果一致（图22E）。

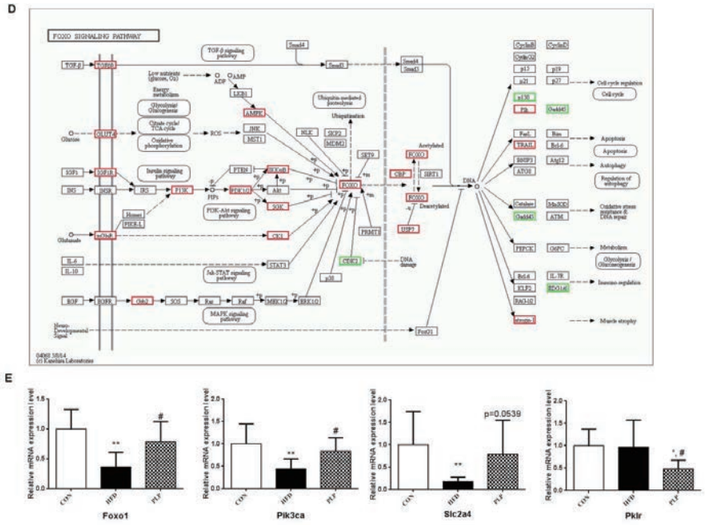


图22 小鼠肝脏转录组测序，高脂高糖组vs高脂高糖+桑黄多糖组，（D）FOXO通路的基因基本都上调，（E）qPCR验证与测序结果一致

3、钙信号通路

研究者还关注差异表达基因富集通路当中所包含的钙信号通路。肝脏细胞的钙平衡，受磷脂酰胆碱与磷脂酰乙醇胺的比例（PC/PE）调控研究。发现，PC/PE比例在高糖高脂组降低，服用桑黄多糖后恢复（图23）。

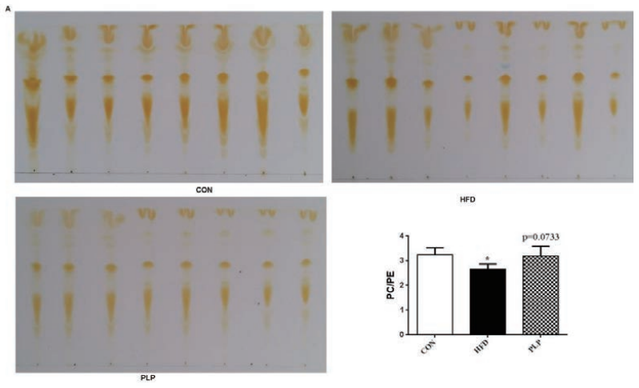


图23 高糖高脂饮食降低了PC/PE 比例，而桑黄多糖能使比例恢复正常水平

4、PC/PE比例的调控机制分析

研究者首先分析PEMT相关基因表达。qPCR实验显示，Pemt（表达PEMT的基因）在高糖高脂组中表达量高于常规饮食组，服用桑黄多糖后，Pemt表达量下降。

随后分析SAM基因表达量变化情况。回归转录组数据，发现有两个差异表达基因参与SAM合成：Mat1a和Mat2b。高糖高脂+桑黄多糖组Mat1a表达量上调，但Mat2b表达量下调。SAM的浓度在组间没有显著差异，但其反式甲基反应副产物s-腺苷高半胱氨酸（SAH）在高糖高脂组升高，在桑黄多糖组恢复，从SAM/SAH比例看，高糖高脂组比常规组低，桑黄多糖服用组与常规组持平。SAH经代谢转化为高半胱氨酸（HCY），但HCY在三组间并无显著差异（图24）。

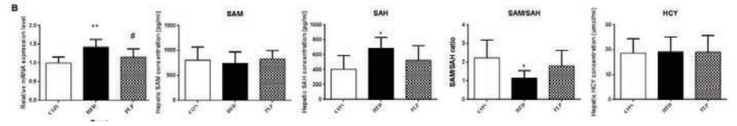


图24 SAH在高糖高脂组升高，服用桑黄多糖后恢复，从SAM/SAH比率看，高糖高脂组比常规组低，在桑黄多糖组与常规组持平。HCY在三组间并无显著差异。

5、桑黄多糖对VB12的影响机制

进一步挖掘转录组测序数据，发现桑黄多糖摄入改变了3个参与维生素B12运输和代谢相关基因的表达（从KEGG和GO代谢路径分析可得）；再查看维生素B12的表型数据，发现高糖高脂组血浆维生素B12水平相比于常规组显著降低，服用桑黄多糖后改善（图25）。

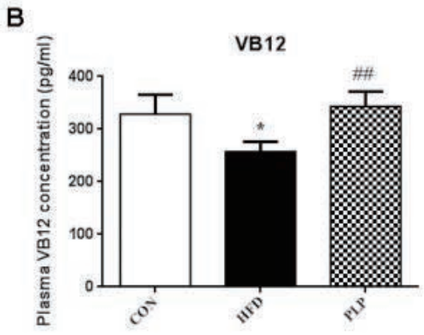


图25 高糖高脂组血浆维生素B12水平相比于常规组显著降低，服用桑黄多糖后改善

6、桑黄多糖对肠道菌群的影响机制

维生素B12参与HCY合成甲硫氨酸的代谢路径，是甲硫氨酸合成催化酶的辅因子。回归转录组数据，发现桑黄多糖处理提高了甲硫氨酸合成还原酶基因（Mtrr）表达，MTRR能恢复甲硫氨酸合成酶活性，促进氧化态维生素B12还原再生。

动物自身不能合成维生素B12，但肠道菌群可以。为了弄清桑黄多糖是否通过改变肠道菌群来影响维生素B12代谢，研究者采用PCR-DGGE实验，在高糖高脂饮食+桑黄多糖服用组，发现了一个特别的细菌——卟啉单胞菌丰度的改变。

这个菌是不是真的影响了维生素B12水平?钴啉酸a,c-二酰胺合酶（cbiA）是参与维生素B12合成的一个关键酶，检测编码该酶的DNA在肠道菌群中的相对水平，发现服用桑黄多糖后，该DNA含量更高，这促进了维生素B12血浆浓度的恢复（图26）。

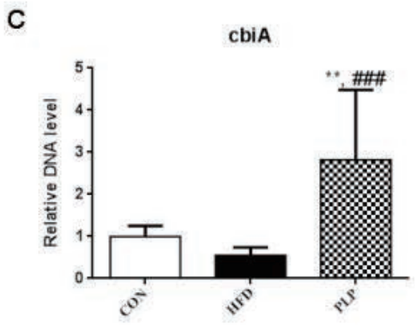


图26 服用桑黄多糖显著提升编码cbiA的DNA相对水平

研究结论：

总之，两条代谢通路分析结果交汇到了一起，研究者据此推断，桑黄多糖的摄入影响了肠道菌群构成，特异性地促进了肠道细菌——卟啉单胞菌的增殖，促进了维生素B12合成，进而调节肝脏的甲硫氨酸代谢，使得磷脂酰胆碱与磷脂酰乙醇胺的比例恢复平衡。

可能存在的风险

做完转录组/RNA-Seq以后，可能有部分基因测序的结果和qRT-PCR的结果不一致，因为两种方法本身存在差异，出现不一致情况也是正常的。除了方法本身差异，需要从下面几个方面分析可能的问题：

1. 要保证测序时所用样品同RT-PCR实验中所用需是同一批材料、处理条件一致；

2. 尽量选取表达量高的基因进行验证，差异倍数在5~10倍的基因更合适；

3. 考虑客户qPCR实验的实验方案、引物序列及原始结果。比如客户设计探针是否考虑多转录本情况，RNA-Seq是对转录本定量。如该基因对应多个转录本，则可能有偏差；

4. 该基因是否存在新的可变剪切；

5. 客户指定的数据库质量不好，大部分参考序列不是全长的。这样比对上某一转录本的reads也会偏少。可以去NCBI下载全长的再做后续验证。客户提供的数据库冗余较高，当证的基因被冗余的reads比对上，这样的reads分析的时候是去掉的；

6.检查有无错误，比如样品对比关系反了。

常见问题

1. 是否需要生物学重复?重复几次?

答：建议3次以上的生物学重复更好。2011年7月Hansen[7]发表的文章表明生物学差异是基因自身表达的特性，与检测技术的选择以及数据处理的方式无关。如果不设生物学重复，高影响因子的杂志可能会因此而拒稿。不建议2个生物重复，推荐3个以上重复设置，原则上越多越好。不建议2个生物学重复是因为，如果两者结果不一致，无法确定以哪个数据为参考。3个生物学重复，如果出现1个结果不一致的，可以取另外2个的结果。

2. 测序后有何验证方法?

答：实验验证的方法最常见的是通过实时荧光定量 PCR（qRT-PCR）技术来验证测序结果。还有FISH（原位荧光杂交）、微阵列芯片技术、Northern blot等。功能验证一般是基因敲除、敲低或过表达，转基因等。

华大优势

高品质的自主测序平台：DNBSEQ测序仪是华大基因自主研发的高通量测序系统，提供多个高性价比测序服务，优质、稳定、高效，无index hopping风险。从2016年11月以来该平台连续发表多篇高水平文章，《Nature》、《Cell》、《Immunity》等。

样品起始量更低：常规建库，人鼠样品200ng起，其他物种样品1ug起。

全面的样品类型：提供单细胞、FFPE、微量等特殊样品服务，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，完美匹配PE150的长读长，转录本组装长度更长，可变剪接、基因融合鉴定更敏感更准确；

交互式的报告系统：Dr. Tom交互式系统自由、便捷、深度地进行数据挖掘，自由做个性化分析。

参考文献

[1] Chen K, Liu J, et al. Methyltransferase SETD2-Mediated Methylation of STAT1 Is Critical for Interferon Antiviral Activity. *Cell*. 2017 Jul 27;170(3):492-506.e14.

[2] Yan T, Li L, Xie L, et al. A novel HD-ZIP IV/MIXTA complex promotes glandular trichome initiation and cuticle development in Artemisia annua. *New Phytologist* . 2018 Apr;218(2):567-578. doi: 10.1111/nph.15005.

[3] Dong JK , Lei HM , Liang Q, et al. Overcoming erlotinib resistance in EGFR mutation-positive lung adenocarcinomas through repression of phosphoglycerate dehydrogenase. *Theranostics*. 2018; 8(7):1808-1823.

[4] Ye Jing, Zha Jie, Shi Yuanfei et al. Co-inhibition of HDAC and MLL-menin interaction targets MLL-rearranged acute myeloid leukemia cells via disruption of DNA damage checkpoint and DNA repair. *Clinical Epigenetics*, 2019, 11: 137.

[5] Feng H, Zhang S, et al. Polysaccharides extracted from Phellinus linteus ameliorate high-fat high-fructose diet induced insulin resistance in mice. *Carbohydrate Polymers*, 2018, Nov 15; 200: 144-153.

[6] Kasper D Hansen, Zhijin Wu, et al. Sequencing technology does not eliminate biological variability. *Nat Biotechnol*. 2011. 29(7): 572–573.

基于高通量测序的种群特征研究方案

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。

方案设计

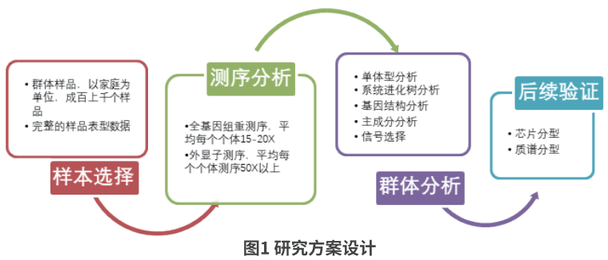


图1 研究方案设计

A. 样本选择建议

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；

B. 采用的技术

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。

C. 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。

D. 分析结果

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

E. 部分分析结果图展示

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。

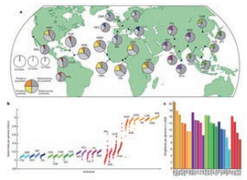


图2 群体样品选择和群体SNP统计结果[9]

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。

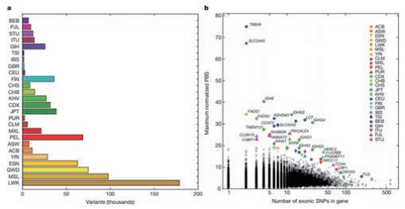


图3 群体分化[9]

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。

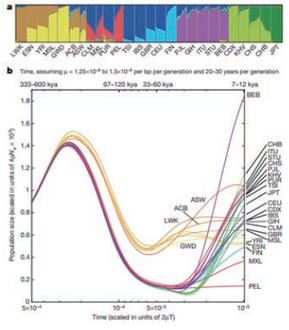


图4 群体结构和人口统计[9]

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

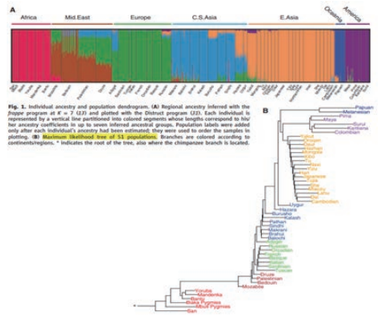


图5 个体血统分析和群体系统数构建[10]

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

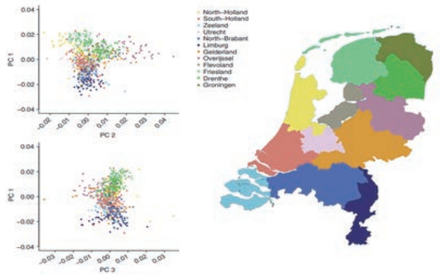


图6 群体主成分分析[11]

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。

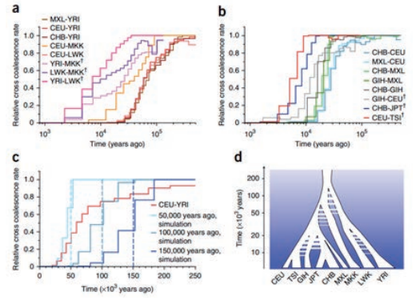


图7 群体间的遗传分离[12]

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

F. 项目执行周期

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。

G. 预期的结果

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。

H. 后期验证手段

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。

疾病诊断及药效评估的蛋白标志物研究方案

研究背景

疾病的早期诊断和确诊以及药效的评估对于患者至关重要，直接影响到患者的治疗和康复效果。现阶段在精准医疗的大背景下，蛋白质组学技术也得到了飞速发展，为寻找疾病诊断药效评估相关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蛋白生物标志物的高通量筛选，做到对疾病的早期诊断和确诊，药效的评估及施药方案的调整，为医务工作者开展切实可行的疾病预防和治疗工作提供重要的指导。

方案设计

A. 研究目的

筛选疾病诊断及药效评估相关的差异表达蛋白，用于临床疾病早期诊断、药物选择。

B. 研究思路

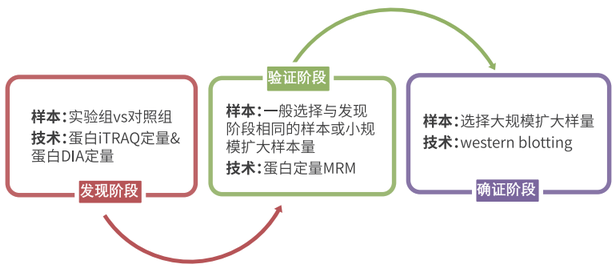


图1 研究思路

1. 发现阶段

本阶段旨在初步筛选疾病诊断及药效评估相关的差异表达蛋白，得到候选的蛋白标志物。

样本选择：

疾病vs正常样本；药物处理前vs药物处理后样本；不同药物处理之间样本。组织200mg，血液500μL，细胞3-5×10^6个。本阶段可使用混合样本。

疾病研究可根据实验目的选取疾病发生不同时段的样本，与未发病或其他相关疾病样本进行比较分析；药效评估可根据实验目的选取药物处理前后，或相关药物处理之间的样本进行比较分析。

技术介绍：

(1)蛋白iTRAQ定量技术，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。

(2)蛋白DIA定量技术，与蛋白iTRAQ定量技术类似，是定量蛋白质组学常用的高通量筛选技术之一。不同的是，该技术采用数据非依赖型采集模式（data independent acquisition，DIA）的质谱采集方式对样本中的蛋白进行分析，同时比较的样本数不受标签数量限制，更适合开展大规模样本的蛋白质组定量分析。

2. 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，得到经过验证的可靠的蛋白标志物。

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，验证阶段也同样适用组织样本）。

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。

3. 确证环节

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（≥100个）中进行广泛的确证，最终得到可信的并且易于获取的疾病诊断及药效评估蛋白标志物。

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，通过分析着色的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。

C. 信息分析结果

1. 蛋白定量结果——样本间差异表达蛋白信息统计

在发现阶段，最基本的结果是要获得关于蛋白的定量数据。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。

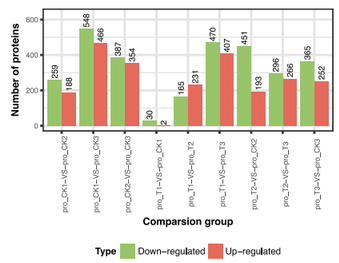


图2 差异蛋白柱形图

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。

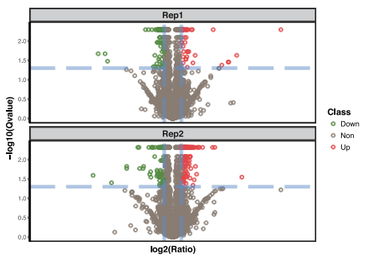


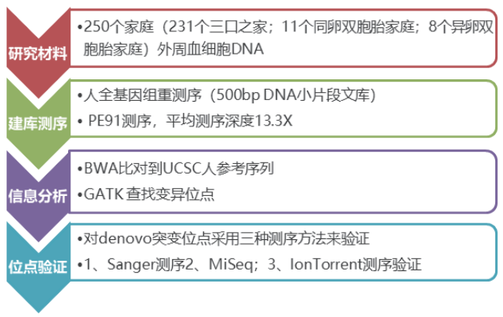
图3 显著差异蛋白火山图

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。

应用案例

案例一：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，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地研究。



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)

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。

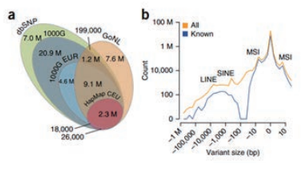


图8 单碱基变异和结构变异的发现

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。

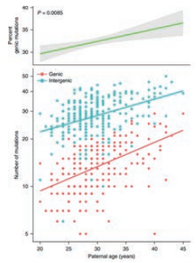


图9 与父亲年龄相关的子代基因中*De novo*突变分布

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：

- 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。
- 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。
- 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。

可能存在的风险

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。

华大已发表文章

研究领域	文章名	期刊/杂志
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)
	Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture	Nature Genetics. (2015)

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、数据交付为一体的平台，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。

参考文献

[1] Xin Yi, Yu Liang, e.al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329, 75-78 (2010).

[2] Morten Rasmussen, Yingrui Li. et al. Ancient human genome sequence of anextinct Palaeo-Eskimo. Nature 463, 757-762 (2010).

[3] Morten Rasmussen, Xiaosen Guo, et al. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94-98 (2011).

[4] David Reich, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053-1060 (2010).

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

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

[7] Dhandapany, P. S. et al. A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. Nature Genet.41, 187–191 (2009).

[8] Noah A. Rosenberg, et al. Genome-wide association studies in diverse populations. Nature Reviews Genetics 11, 356-366(2010).

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

[10] Jun Z. Li, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 319, 1100-1104 (2008).

[11] Dorretl Boomsma, Cisca Wijmenga, Eline P Slagboom, et al. The Genome of the Netherlands: design, and project goals. Eur J Hum Genet., 22(2):221-227(2014).

[12] Stephan Schiffels & Richard Durbin. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925 (2014).

[13] Genomes of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825 (2014).

[14] Genomes of the Netherlands Consortium. Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. 47, 822–829 (2015).

[15] Jun Wang, Wei Wang, et al. The diploid genome sequence of an Asian individual. Nature 456,60-65 (2010).

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

[17] Elisabeth M van Leeuwen, Lennart C Karssen, et al. Genome of the Netherlands opulation-specific imputations identify an ABCA6 variant associated with cholesterol levels. Nature Communications. (2015).

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

[19] Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81 (2015).

[20] Laura Clarke, Xiangqun Zheng-Bradley, et al. The 1000 Genomes Project: data management and community access. Nature Methods 9, 1-4 (2012).

[21] Jie Huang, Bryan Howie, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nature Communications. (2015).

[22] The UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature 526, 82-88 (2015).

[23] The UK10K Consortium. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112-116 (2015).

[24] Magdalena Zoledziewska, Carlo Sidore, et al. Height-reducing variants and selection for short stature in Sardinia. Nature Genetics 47,1352-1356 (2015).

孟德尔遗传病基因组测序研究方案

研究背景

单基因病是由1对等位基因控制的疾病或病理性状，主要由遗传因素决定，遵循孟德尔遗传规律。发病率低，家族遗传，外显率较高，通常比较严重，发病慢，而且是渐进性的。通常分为常染色体显性、常染色体隐性、X伴性显性、X伴性隐性、Y伴性遗传等几类。根据OMIM数据库，截止到2018年6月，共有2.4万多个基因被报道与已知疾病或病理性状相关。

表1 OMIM数据库与单基因病种类

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	15,106	729	49	35	15,919
Gene and phenotype, combined +	50	0	0	2	52
Phenotype description, molecular basis known #	4,911	325	4	31	5,271
Phenotype description or locus, molecular basis unknown %	1,454	124	4	0	1,582
Other, mainly phenotypes with suspected mendelian basis	1,657	105	3	0	1,765
Totals	23,178	1,283	60	68	24,589

在单基因病的致病基因定位和克隆研究中,通常采用的家系连锁分析法及定位克隆技术使得超过3000多种单基因疾病的遗传病因得以阐明，但人们仍剩余的（超过一半）单基因疾病的致病基因缺乏了解。同时如果患病亲属的人数有限，或不外显、外显不全，或基因突变是自发产生的，则连锁分析多半失效，这是单基因疾病分子遗传学研究中常常难以克服的"瓶颈"。

全外显子组测序通过对基因组1%的序列测序，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，具有所需样本数量少、低费用、高通量的优势和特点，是经济有效地挖掘与蛋白质功能相关的遗传变异的方法，为找出治病基因和探索疾病发生发展机制提供重要依据。其研究思路根据样本情况不同，可分为家系样本和散发样本研究 。在费用足够的情况下，可以采用全基因组测序或者多组学的方法揭示基因与疾病之间的关系，加快鉴定人类疾病基因的进程。

方案设计

2016年有文章统计了2013~2015年底发表的孟德尔遗传疾病致病基因相关文献，共计492篇，其中76%为Exome相关测序。DOI:10.1002/mgg3.221,PMID:27468415

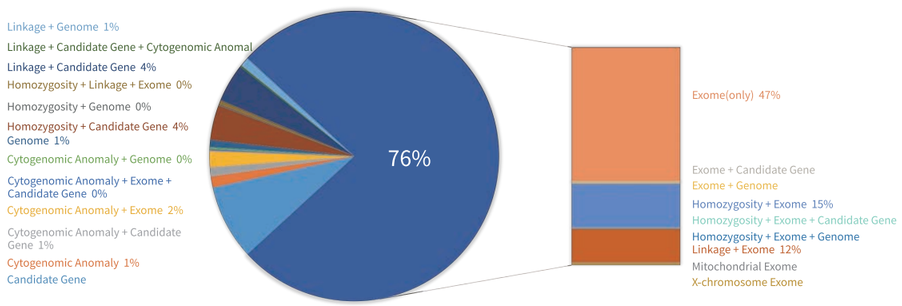


图1 NGS技术在孟德尔遗传疾病研究中的应用

华大基于家系的单基因疾病研究策略如下图所示，包括五个主要阶段，前期准备，建库测序，信息分析，筛选解读和验证，每个阶段的工作都非常重要。接下来将分别介绍每个阶段的具体内容和注意事项。

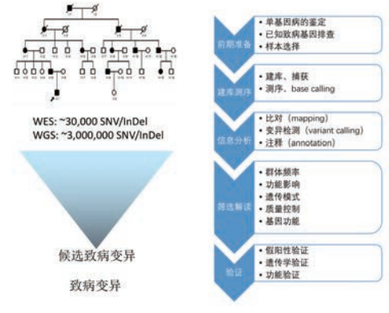


图2 华大基于家系的单基因疾病研究策略

A. 前期准备

前期准备指为开展NGS测序分析前进行的一系列准备和调研工作，目的是充分了解家系和疾病相关信息，制定合理的研究策略。可以从以下方面对重要信息进行整理：

1. 单基因疾病鉴定

该部分要求根据患者家系情况绘制系谱图和临床表征，并运用遗传学基本原理对系谱进行分析，观察疾病的遗传规律，判断遗传病的遗传方式，结合文献检索和发病率进行鉴定。

2. 疾病调研

该部分要求全面的了解所关注的疾病，包括疾病的定义、流行病学数据（发病率）、疾病分类（亚型）、已知基因和变异、机制与功能、未解决的问题等，调研的信息来源可以是疾病数据库OMIM和致病变异数据库ClinVar/HGMD/LOVD和文献；这些信息对研究策略的选择，致病变异的筛选和解读非常重要。

3. 已知基因排查

基于调研的结果对已知致病基因进行排查，在排查完已知致病基因后，如果没有找到致病基因再进行测序研究，这样可以避免经费和人工的浪费，也更可能获得新的科学发现。通常可采用以下策略进行排查。

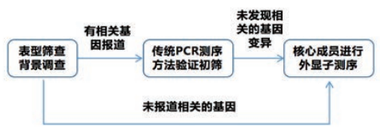


图3 已知致病基因的排查

常规PCR+Sanger测序：适用于已报道致病基因少且小，否则成本、成功率、时间不划算.

连锁定位：适用于有多个致病基因、区域报道的疾病；要求有较大的家系，定位到一个新的区域，相当于排除了已知的所有信息；若定位到一个区域，相当于排除了其他的区域信息；

直接测1个样本外显子序列：选择捕获平台，主要参考所选用平台设计用的数据库是否对要排查的基因有很好的覆盖，少数无法覆盖的还是需要用常规PCR+Sanger测序补洞。

需要说明的是，排查工作很重要，是否进行排查还需要考虑实际情况，如果已知基因数目很多，经费和人工投入很大，也可以选择不做排查。

4. 测序样本选择

一般通用原则是（1）优先选择病例；（2）病例之间的亲缘关系尽量远；（3）病例尽量选择表型严重、症状典型、一致的患者，

如先证者；(4) 对照尽量选择和所有病例都有亲缘关系避免病例的表型正常的兄妹作为对照，避免病例的表型正常的子代作为对照，对照尽量选择父辈或祖辈。

- 以上样本选择策略均是在样本数充足、研究经费充裕以及所选样本DNA可获得的情况下制定的，在实际操作时，则应根据实际情况，具体问题具体分析；
- 以上样本选择策略均只考虑单个家系内进行研究，如果要不同家系混合分析，需考虑疾病的异质性问题，尽量将表型一致的家系样本放到一起分析；
- 散发样本如果考虑de novo突变，建议选择患者及其父母进行测序；

表2 不同遗传模式的样品选择

遗传方式	家系/疾病特点	样本选择建议
AD（常显）	通常情况	2~4 case，1~2 个 control
AD（常显）	De novo	先证者，先证者父母（表型正常）
AD（常显）	如果有定位区	可选择更少的样本，有时只需要 1 个 case
AR（常隐）	通常情况	1~3 个 case，1~2 个 control
AR（常隐）	复合杂合致病	建议选择较多样本，如 2~3 个 case，1~2 个 control
AR（常隐）	近亲婚配，考虑纯合致病	通过纯合子定位快速锁定到较小的区域，通常可以选择更少的样本：如 1~2 个 case、0~1 个 control
XL	通常情况	1~3 个 case
XL	有定位区	1~2 个 case

- 如果能够确定为X连锁显性遗传病，则可以只选1-3个case。
- 优先选择患病男性作为case。原因：男性患者为X染色体的半合子，测序时X染色体上的SNP都是纯合SNP。而纯合SNP的假阳性率相比杂合SNP要低。
- 如果做过连锁定位，则可以只选1-2个case。

B. 研究策略

捕获平台主要指全外显子捕获平台，通常有Agilent、BGI、NimbleGen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。

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。华大单病流程整合多种比对、变异检测工具，提供高质量的变异结果；适用于各种孟德尔遗传模式的家系分析，包括染色体显性遗传（AD），常染色体隐性遗传（AR），X染色体连锁遗传（XL）。

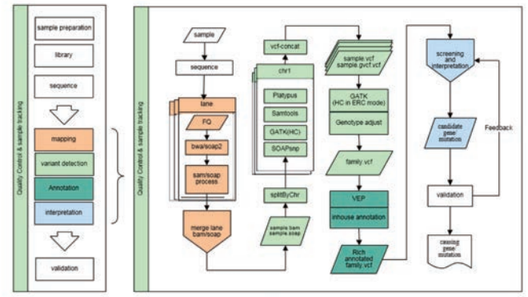


图4 标准信息分析流程图

数据筛选

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。

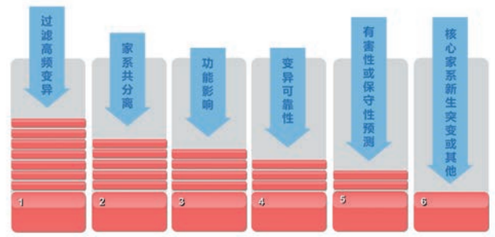


图5 数据筛选（过滤）过程

(1)对于群体罕见突变，通常推荐使用的低频阈值为小于等于0.005，即变异后的等位基因在群体中的携带率小于等于千分之五。需要注意的是，不同的疾病的发病率高低不同，可以基于调研的结果进行调整。设置为0.005时，如果是显性遗传，则发病率约为5/1,000；如果是隐性遗传，则发病率约为0.005的平方，即2.5/100,000。

(2)考虑单基因疾病患者临床表征通常很严重，常会致死、致畸、致残，倾向于认为致病变异严重影响基因的转录或翻译过程，所以在筛选单基因疾病候选变异时，可以优先关注功能影响严重的变异，如transcript_ablation，splice_acceptor_-variant，splice_donor_variant，stop_gained，frameshift_variant，stop_lost，missense_variant等。

(3)考虑单基因疾病的致病变异倾向于直接影响基因的表达，在一个家系内通常假设疾病完全外显，基因型和表型在家系中呈现共分离，即家系成员中的患者都携带致病基因型，正常人都不携带致病基因型。

C. 项目执行周期

48个样品内，从检测合格后，建库+测序+标准信息分析：约15-18个工作日。

D. 预期的结果

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。

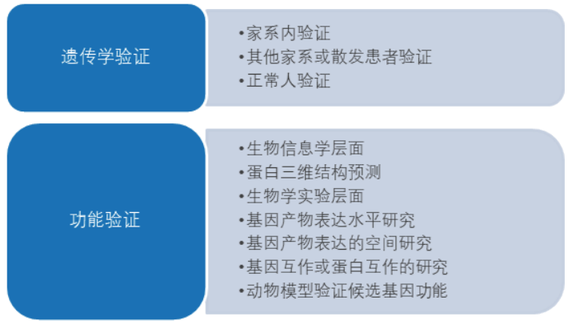


图6 后期的验证方法

E. 关键点把控

前期准备中文献的调研一定要充足，家系遗传信息尽量搜集；

选样个体要为确定是否为患病个体，以防止遗漏重要信息；

如果想要做CNV，建议选择对应的control开展项目。

F. 代表案例

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

应用案例

案例一：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。

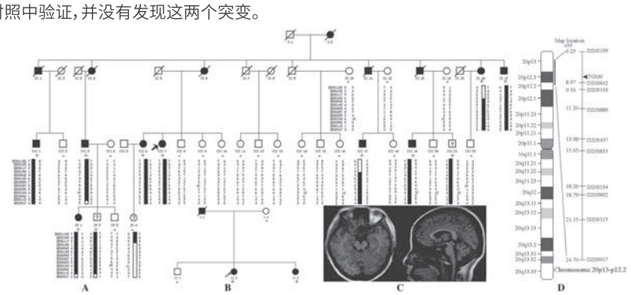


图7 小脑共济失调家系图，先证者大脑核磁共振图以及20号染色体的比对区域

案例二：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。

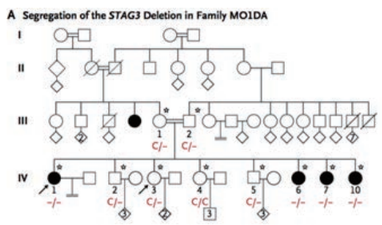


图8 MO1DA家系图谱

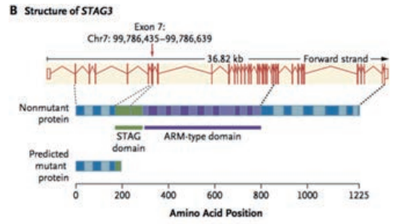


图9 STAG3基因结构图

案例三：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。

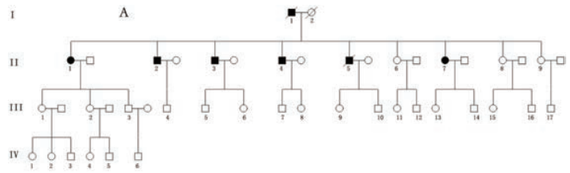


图10 RP四代家系

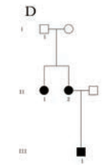


图11 用来验证的独立RP三代家系

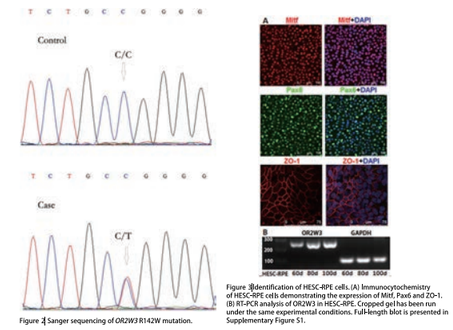


图12 sanger验证图，RT－PCR发现OR2W3基因在HESC－RPE中表达

可能存在的风险

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。

常见问题

1. 什么是家系共分离？

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。

2. 为什么隐性遗传模式需要筛选复合杂合变异？

隐性遗传指父母双方表型正常，携带隐性致病位点，遗传给子代，导致患病。复合杂合变异指一对等位基因不同位点发生突变。假设双亲基因型为Aa1，Aa2,子代复合杂合变异为a1a2.对于近亲结婚的双亲，筛选春和变异，对于非近亲结婚的家系，一般筛选复合杂合变异位点。

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BGI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。

参考文献

[1] Jun Ling W, Xu Y, Kun X, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing.[J]. Brain A Journal of Neurology, 2010, 133(Pt12):3510-3518.

[2] Caburet S, Arboleda VA, Llano E, et al. Mutant cohesin in premature ovarian failure. N Engl J Med. 2014 Mar 6;370 (10):943-9.

[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, 5:1-6.

[4] 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.

[5] Wang J, Zhang W, Jiang H, et al. Mutations in HFM1 in recessive primary ovarian insufficiency[J]. New England Journal of Medicine, 2014, 370(10): 972-974.

肿瘤融合基因研究方案

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，通过模拟数据和真实数据对这些软件进行测试，总体来说我们自主开发的SOAPfuse软件更有优势。

本方案主要利用华大自主研发的融合基因检测软件SOAPfuse[6]对样本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。

方案设计

A. 方案流程图

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 基因融合检测方案流程图

B. 样本要求

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10 -1000 ng/μL，RIN≥7.0，28S/18S≥1.0；

C. 技术流程

1. 文库构建

构建小片段文库，具体流程如下：

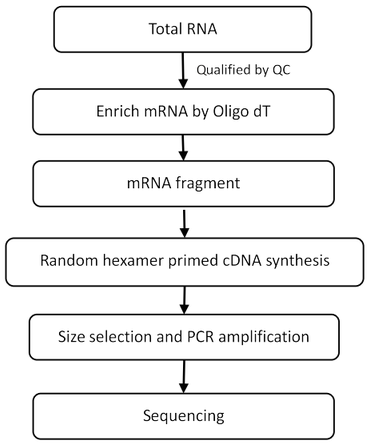


图2 转录组建库流程图

2. 测序策略

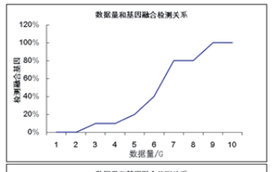
利用BGISEQ或Illumina平台进行转录组测序，应用华大自主研发的SOAPfuse对融合基因进行检测。

测序平台：BGISEQ或Illumina平台

测序策略：PE100，PE150

测序数据量：建议8G clean data。根据图2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。一般推荐10G clean data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

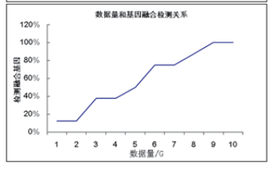


图3 测序数据量与融合基因检测饱和度模拟数据分析

D. 数据分析

首先对DNBSEQ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SOAPaligner/SOAP2[7]将clean reads比对到参考序列。比对完，通过统计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SOAPfuse软件进行融合基因的检测，另外对于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、注释、可变剪接等分析。

SOAPfuse分析软件是由华大基因自主开发的一款能够快速检测融合基因的软件。该算法首先通过比对到基因组和转录本中双末端(paired-end)关系的序列寻找候选的基因融合，然后采用改进的局部穷举算法，构建包含融合位点序列的文库，再通过一系列精细的过滤策略，在尽量保留真实融合的情况下过滤掉其中假阳性的基因融合模拟数据和真实验证数据的综合测评表明，SOAPfuse与其他方法相比准确率更高、灵敏度更强、精度更高、资源消耗大大减少。此外，该算法还具有融合断点预测和可视化功能。这些功能能够极大提高基因融合的检测效率，大力推动疾病尤其是肿瘤的研究，这对临床分子分型和肿瘤新药的开发具有重要意义。下图为SOAPfuse分析流程。

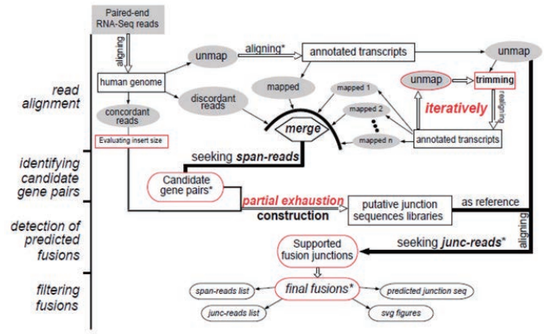


图4 SOAPfuse分析流程

E. 预期结果

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。

1. 结果展示-融合基因的Circos图展示

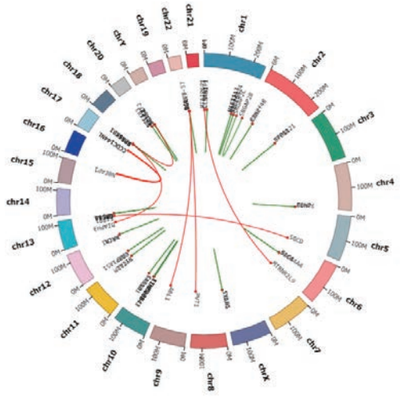


图5 融合基因的Circos图展示

2. 结果展示- SOAPfuse融合图片展示

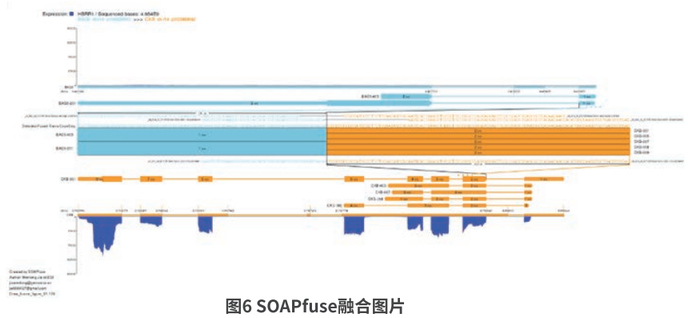


图6 SOAPfuse融合图片

F. 后期验证手段

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。

G 后期验证手段

样品检测合格后，项目周期约24个工作日，可完成全部数据交付。

应用案例

案例一：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。

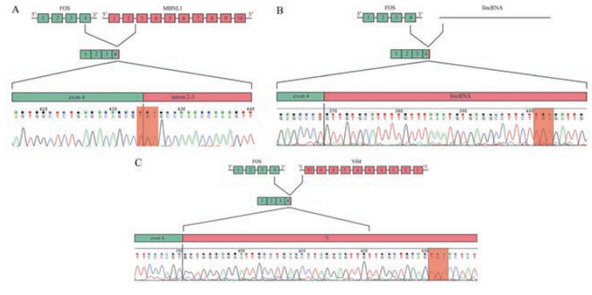


图7 FOS的三种融合基因

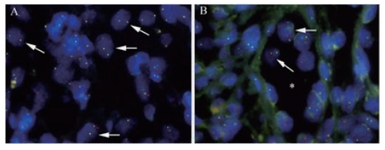


图8 FOS融合基因的验证

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。

案例二：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1 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*

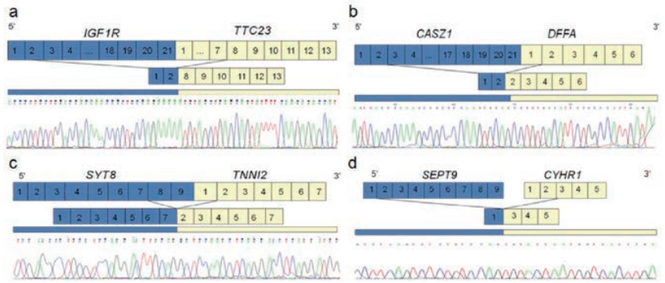


图9 四个新融合基因的结构

案例三：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。

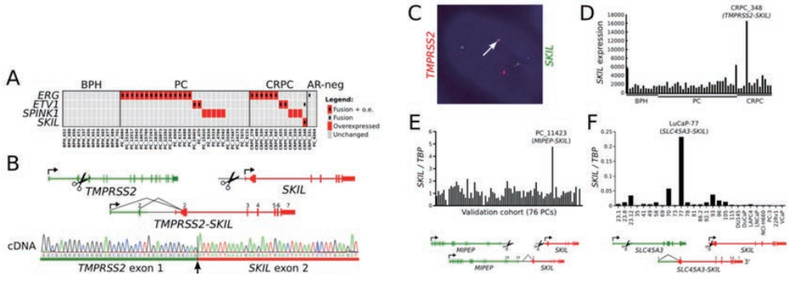


图10 融合基因的检测及验证

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。

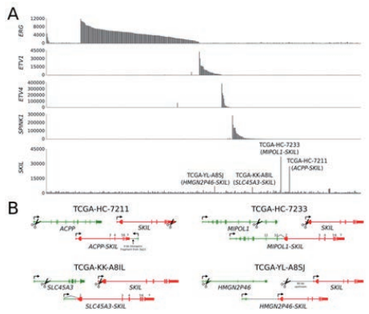


图11 四例SKIL参与的基因融合

常见问题

1. 融合基因检测一般推荐多大的数据量？

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10 G数据量。

2. 样本量应该用多少？

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。

3. 融合基因的发生机制以及研究应用有哪些？

基因融合是指两个基因的全部或一部分的序列相互融合为一个新的基因的过程。形成融合基因的机制较为复杂：包括染色体易位、中间缺失、染色体倒置以及反式剪切等多种机制。目前基因融合主要集中在癌症的研究中，比如，白血病、前列腺癌、乳腺癌等。研究表明融合基因与癌症的发生发展有很大关联，参与融合的基因常常是一些原癌基因。融合的结果多样：可能会产生新的融合蛋白，它兼具融合的两部分基因的功能或者具有新功能；也有可能某原癌基因与其他基因的强启动子融合导致该原癌基因高表达，等等。很多研究结果均表明融合基因可以作为癌症特有的分子标签和潜在的药物靶点，在癌症的临床诊断和治疗中有重要的意义。

某些特定类型的染色体畸变及引起的融合基因，常用于癌症诊断，以提高精确度。染色体显带分析、荧光原位杂交(fluorescence In Situ hybridization，FISH)、逆转录聚合酶链反应(RT- PCR)是诊断实验室常用的方法，但由于癌症基因组本身的复杂性，这些方法仍有明显缺陷。目前，随着高通量测序技术的发展，为融合基因的检测带来了更为准确、有效的方法。

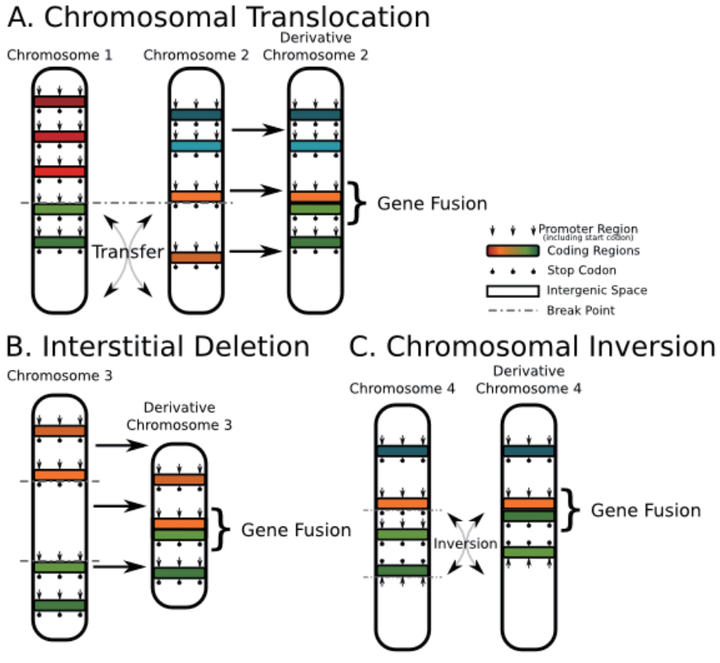


图12 融合基因类型

融合基因形成的原因有：A)染色体易位；B)中间缺失；C)染色体倒置。

华大优势

高品质的自主测序平台：DNBSEQ测序仪是华大基因自主研发的高通量测序系统，提供多个高性价比测序服务，优质、稳定、高效，低dup无需人为干预，无index hopping风险。从2016年11月以来该平台连续发表多篇高水平文章，包括Nature、Science、Cell等顶级期刊；

样品起始量更低：常规建库，人鼠样品200ng起，其他物种样品1ug起，可微量定制化建库，低至200pg；可单细胞定制化建库，低至单个细胞；

全面的样品类型：提供单细胞、FFPE、微量等特殊样品服务，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，完美匹配PE150的长读长，转录本组装长度更长，可变剪接、基因融合鉴定更敏感更准确；

Dr. Tom系统解决个性化难题：无需生信分析基础，只需鼠标一键点击操作，随时随地即兴交互，玩转个性化分析，只需任意一种RNA测序数据，就可给您多组学关联信息，多数据库联合分析，多维度数据展示，循环挖掘数据，在成千上万的候选基因中轻松锁定解释生物学问题的核心基因。

参考文献

[1] Tkachuk D, Westbrook C, Andreeff M, Donlon T, Cleary M, Suryanarayan K, Homge M, Redner A, Gray J, Pinkel D. Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. Science (New York, NY) 1990, 250:559.

[2] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, Varambally S, Cao X, Tchinda J, Kuefer R. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005, 310:644-648.

[3] Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007, 448:595-599.

[4] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S-i, Watanabe H, Kurashina K, Hatanaka H. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature 2007, 448:561-566.

[5] Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nature genetics 2011, 43:964-968.

[6] Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, Yu Y, Zhu D, Nickerson ML, Wan S. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome biology 2013, 14:R12.

[7] Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009, 25:1966-1967.

[8] Van Ijzendoorn D G, 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 V, Tanas A S, Kanygina A, et al. Novel fusion transcripts in bladder cancer identified by RNA-Seq[J]. Cancer Letters, 2016.

[10] Annala M, Kivinummi K, Tuominen J, et al. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer[J]. Oncotarget, 2015, 6(8): 6235-6250.

病毒感染相关的转录组与蛋白质组关联分析研究方案

研究背景

病毒感染类疾病的早期筛查、确诊以及预后和药效的评估对于患者至关重要，直接影响到患者的治疗和康复效果。现阶段在精准医疗的大背景下，转录组学和蛋白质组学技术的也得到了飞速发展，为寻找相关标志物和研究病原菌与侵染宿主机制提供了有利条件。通过同步检测mRNA和蛋白质的表达量并进行联合分析，科研工作者可以实现蛋白生物标志物的高通量筛选，做到对疾病的早期诊断和确诊，药效的评估及施药方案的调整，为医务工作者开展切实可行的疾病预防和治疗工作提供重要的指导。

生命体是一个多层次，多功能的复杂结构体系，从DNA、RNA、蛋白质到代谢物的过程中涉及到一整套精细的表达调控机制，如转录调控、转录后调控、翻译调控、翻译后调控等。高通量技术的发展积累了大量的组学数据，这使得由精细的分解研究转向系统的整体研究成为可能[1]。整合多组学数据能够实现对生物系统的全面了解，建立有效指示表型的模型，揭示重要的生物标志物。

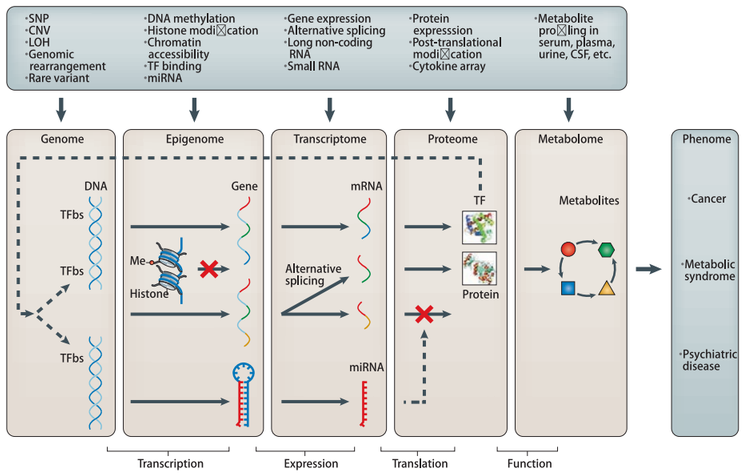


图1 系统生物学之基因组，表观组，转录组，蛋白组，代谢组与表型的关系[1]

该图从系统生物学角度描述了基因组，表观组，转录组，蛋白组，代谢组学的研究对象和能够获得的主要结论，以及组学和表型之间的关系。

蛋白质是生命功能的执行者，其含量的变化在生物体的生长发育[2]、环境应激[3,4]、疾病发生发展[5]等过程中发挥着重要的作用，对于蛋白质的表达量进行深入研究是十分重要而又关键的。蛋白质组学包含基因组和转录组所不曾有的功能性相关信息：

- 基因组基因的表达呈现时空和丰度高低的特征；
- 许多蛋白质的修饰形式具备特定的生物学功能；
- 大多数蛋白质可形成具有功能性的复合物，诸如蛋白质/蛋白质、蛋白质/核酸、蛋白质/脂类等。

mRNA的表达量是影响蛋白表达最为重要和直接的因素，通过分析蛋白质组与转录组的关联性，可以系统全面地了解生物体内基因表达调控途径[6]。

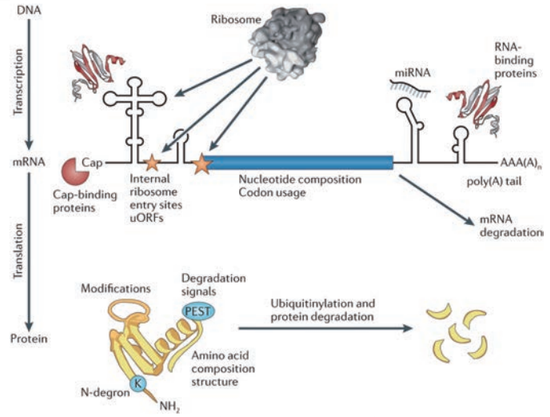


图2 转录调控和蛋白翻译后修饰模式[6]

该图展示了蛋白质的丰度取决于RNA和蛋白产生与降解之间的平衡，图中上半部分主要解释转录和翻译过程的合成及稳定状态，下半部分主要阐述蛋白降解。

因此，要全面开展病毒感染相关的精准医疗研究，同步检测mRNA和蛋白质的表达量并进行联合分析已成为当前研究的必然趋势。本方案重点介绍蛋白质组和转录组关联分析的方法，以及涉及到的分析内容，旨在帮助广大科研工作者进行病毒感染类疾病的早期筛查、确诊以及预后和药效的评估。

方案设计

2.1 拟解决的关键科学问题

用于解决病毒感染相关疾病的机制调控、早筛及预后的生物标志物筛选、药物药效评估等。

2.2 拟采取的研究方案

2.2.1 整体思路图

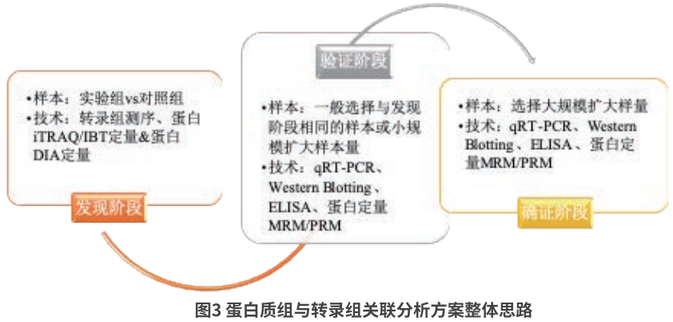


图3 蛋白质组与转录组关联分析方案整体思路

2.2.2 发现阶段

本阶段旨在初步筛选疾病诊断及药效评估相关的差异表达mRNA和蛋白，得到候选的生物标志物。

样本选择：

疾病vs正常样本；药物处理前vs药物处理后样本；不同药物处理之间样本。

转录组：总RNA 200ng，组织≥30mg，≥ 1 mL全血收集的淋巴细胞or ≥ 1 mL收集管保存的全血，细胞≥2×10^5个。每组至少2个生物重复，推荐3个以上的生物重复。推荐6Gb或10Gb clean data，PE100/150。

蛋白质组：组织≥10mg，血液≥100μL，细胞3-5×10^6个。每组至少2个生物重复，推荐3个以上的生物重复。本阶段可使用混合样本。

疾病研究可根据实验目的选取疾病发生不同时段的样本，与未发病或其他相关疾病样本进行比较分析；药效评估可根据实验目的选取药物处理前后，或相关药物处理之间的样本进行比较分析。转录组和蛋白组学的样本选择尽量保持一致，即遵循同时、同类型、同部位取样的原则。推荐转录组和蛋白组的取样一一对应，如果不能保证样本和生物重复完全一致，则有可能出现不同时期不同样本的转录组和蛋白组相关性系数整体偏低的情况。

技术介绍：

1）转录组测序技术，对某一物种的特定组织或细胞在某个时期某一功能状态下产生的mRNA进行测序，可以提供定量分析，检测基因表达水平差异；又可以提供结构序列分析，识别可变剪切位点、基因融合等；而且不依赖于参考基因组。

2）蛋白iTRAQ/IBT定量技术，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8-10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。

3）蛋白DIA定量技术，与蛋白iTRAQ定量技术类似，是定量蛋白质组学常用的高通量筛选技术之一。不同的是，该技术采用数据非依赖型采集模式（data independent acquisition，DIA）的质谱采集方式对样本中的蛋白进行分析，同时比较的样本数不受标签数量限制，更适合开展大规模样本的蛋白质组定量分析。

2.2.3 验证阶段

本阶段旨在缩小发现阶段筛选出的候选生物标志物范围，得到经过验证的可靠的生物标志物。

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，验证阶段也同样适用组织样本）。

技术介绍：

1）qRT-PCR技术，实时荧光定量 PCR 是一种应用广泛的研究基因表达模式的方法，尤其是在样品量少或者说非常珍贵的时候。可以有目标地分析检测转录组测序得到的差异表达基因，验证转录组测序结果的可靠性。

2）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，通过分析着色的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。

3）ELISA（酶联免疫吸附测定enzyme linked immunosorbent assay）指将可溶性的抗原或抗体结合到聚苯乙烯等固相载体上，利用抗原抗体特异性结合进行免疫反应的定性和定量检测方法。

4）蛋白定量MRM（多重反应监测, Multipel Reaction Monitoring）技术，是指利用基于质谱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。

5）蛋白定量PRM（平行反应监测, Parallel Reaction Monitoring）技术，与MRM类似，是一种基于高分辨、高精度质谱的靶向定量技术，能够对目标蛋白质/肽段（以及含有修饰的肽段）进行选择性检测，从而实现对目标蛋白质/肽段的精准定量。

2.2.4 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（≥100个）中进行广泛的确证，最终得到可信的并且易于获取的疾病诊断及药效评估蛋白标志物。

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。

2.3 结果展示

2.3.1 鉴定关联分析

利用转录组数据库构建蛋白质数据库，提高蛋白鉴定数。

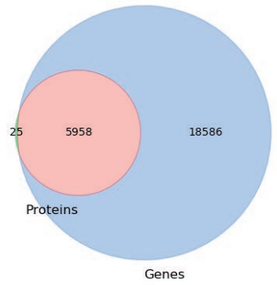


图4 定性关联韦恩图

该图绿色代表转录组鉴定到的基因个数，红色代表蛋白质组鉴定到的蛋白个数，两个圆圈重叠的部分即为转录组与蛋白质组共同鉴定到的基因个数。

对所有鉴定到的蛋白和mRNA进行关联，一般情况下，转录组鉴定到的基因表达情况的覆盖度高于蛋白质组，故利用转录组数据库构建蛋白质数据库，可提高蛋白鉴定数。

通过鉴定关联分析韦恩图，可从整体上分析鉴定到的mRNA和蛋白质的情况。

2.3.2 关联与非关联表达量分布

基于转录组数据，对基因分为关联与非关联，展示两类基因的表达量分布，从而指导后续的转录后调控机制分析。

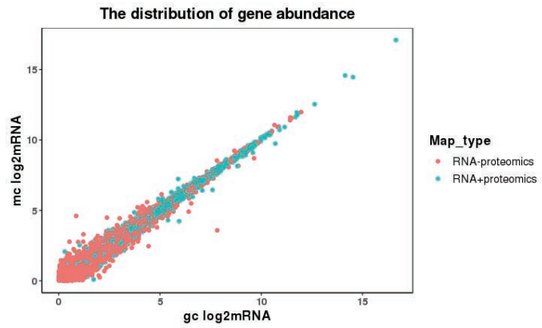


图5 关联与非关联表达量散点图

横坐标为对照组的表达量，纵坐标为实验组的表达量。其中红点为关联上的基因，蓝点为没有关联上的基因

2.3.3 定量关联分析

对于转录组和蛋白质组数据在表达量的层面上进行关联，将两个组学层面表达量趋势和相关性进行分析，对表达情况进行聚类，并对于表达趋势一致或者不同的基因进行深入分析。

首先，按照蛋白和mRNA表达量的变化将所有关联到的基因分成5类，以便细致描绘基因表达调控模式：

A. 蛋白和mRNA表达趋势相同－DEPs_DEGs_Same Trend

B. 蛋白和mRNA表达趋势相反－DEPs_DEGs_Opposite

C. 蛋白表达有差异，mRNA表达无差异－DEPs_NDEGs

D. 蛋白表达无差异，mRNA表达有差异－NDEPs_DEGs

E. 蛋白表达和mRNA表达均无差异－NDEPs_NDEGs

对5类表达关联类型细分，有助于验证表达一致性（A-正相关），补充（C/D/E-仅蛋白或RNA差异或无差异）、揭示特殊（B-负相关）的生物调控和代谢机制。

然后，为了深入了解各种表达情况的相关性趋势和相关性系数，需要对5类情况分别进行分析：

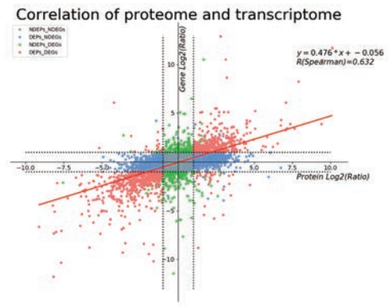


图6 所有定量蛋白质和基因的表达量关联图

横坐标为蛋白质的表达量，纵坐标为基因的表达量。灰点表示mRNA和蛋白都无显著差异；蓝点表示mRNA无显著差异，而蛋白显著差异；绿点表示mRNA显著差异，而蛋白无显著差异；红点表示mRNA和蛋白都显著差异

在生物标志物和生理机制的研究中，最为关注的是转录组和蛋白质组表达趋势一致的情况。表达情况正相关，有利于说明关键基因的表达情况在两个组学层面都得到了验证；对于转录组和蛋白质组表达趋势相反的情况，一般是为了说明一些特殊的抑制调控方式，可根据具体的蛋白功能进行分析。另外，单一组学表达发生变化的情况，可通过后续的功能关联分析，寻找基因上下调的关系进行调控关键基因的深入挖掘。

2.3.4 功能关联分析

通过两个组学的鉴定、定量层面的关联，可以分析基因表达产物mRNA和蛋白一对一的关联方式，但对于某一类基因或者具有上下游调控关系的基因，仅通过一对一的关联方式无法进一步分析，需要通过功能和代谢通路的分析进行解释和分析。

对差异表达基因和差异表达蛋白在GO条目/Pathway上的注释及富集情况进行分析，并将相同的GO功能条目/Pathway中注释到的基因和蛋白进行整合分析，同一GO条目/Pathway上的基因和蛋白在功能上相似，对环境因素的反馈可能存在共调控或是共表达等情况，有利于从基因集的层面研究基因表达调控。(如图8)；

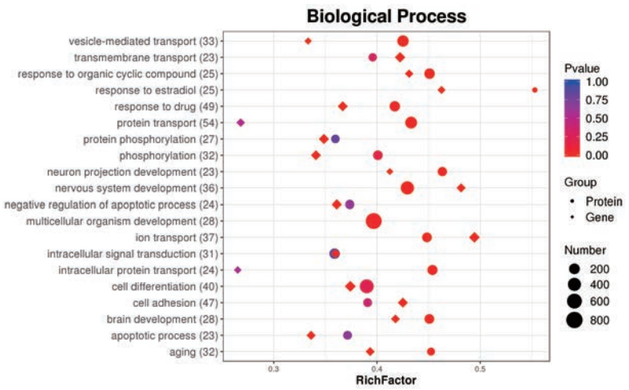


图7 GO关联数量及富集因子分析四维气泡图

Y轴为各GO条目，括号中统计了该条目所关联的蛋白数量，气泡大小代表富集数量，颜色代表富集因子，圆圈代表转录组（蛋白质组）P值大于等于0.05，钻石形代表转录组（蛋白质组）P值小于0.05

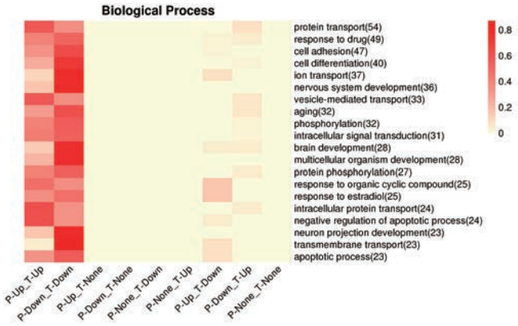


图8 GO条目关联分类概率热图

Y轴为各GO条目，括号为该条目的关联数量；X轴为不同差异类型分布情况。热图颜色代表该类型在该GO条目关联上的蛋白中所占的比例

根据蛋白质组和转录组差异表达基因的分析结果，对其进行KEGG生物通路分类以及富集分析，最终将蛋白质组和转录组差异表达基因的信息汇总在一张通路图中，分别显示mRNA上下调，蛋白上下调，mRNA和蛋白同时改变或者一方改变，mRNA和蛋白同时不变等情况，直观展示一条通路中的所有转录组和蛋白质组鉴定和定量到的数据，更加方便地展示关键调控基因。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。

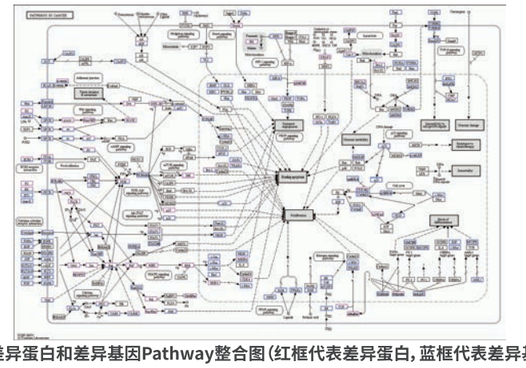


图9 差异蛋白和差异基因Pathway整合图（红框代表差异蛋白，蓝框代表差异基因）

2.3.5 转录因子关联分析

转录因子在生命体的生长发育及其对外界环境的反应中起着重要的调控作用，对转录因子进行的数量及表达量分布进行统计分析。

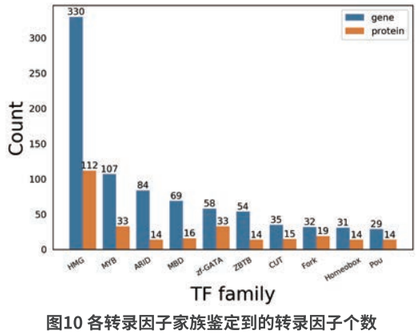


图10 各转录因子家族鉴定到的转录因子个数

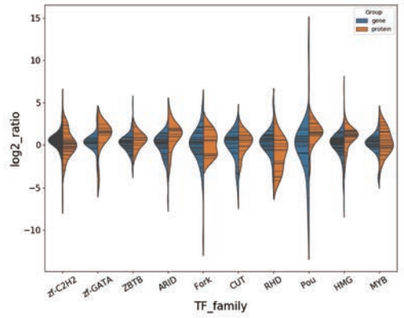


图11 各转录因子家族在基因及蛋白层面的表达量分布图

2.3.6 多组学网络关联分析

针对上图中的所有差异蛋白，通过与STRING蛋白互作数据库比对，绘制网络图。

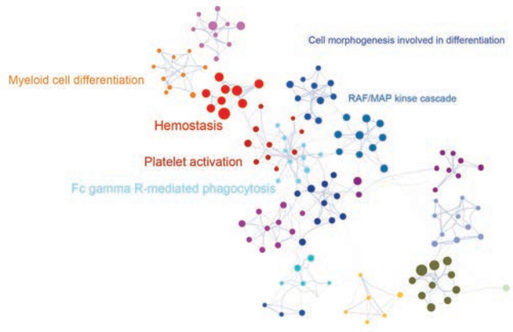


图12 差异蛋白和差异基因网络互作图

2.4项目执行周期

转录组分析周期：样品检测合格后，建库+测序+标准信息分析：约24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。

蛋白组分析周期：样品检测合格后，蛋白定量iTRAQ/IBT技术标准周期约26个自然日（≤一批次）完成；蛋白定量DIA技术标准周期约42个自然日（≤25个样本）完成。如满足极致交付条件，最快可于13天内完成，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。

关联分析周期：数据分析，约12个工作日完成。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。

项目总时间：约40个工作日完成。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。

应用案例

案例一：人巨细胞病毒感染期间宿主蛋白质的稳定性分析[7]

人巨细胞病毒（HCMV）是一种临床上广泛存在的病原体，一旦感染则终身潜伏，当机体免疫力低下时，病毒激活则会导致多种疾病。HCMV 具有多种免疫逃避策略，包括促进宿主抗病毒限制因子（ARFs）的降解。在 HCMV 感染早期会通过蛋白酶体和溶酶体激活宿主蛋白降解途径，基于此，本研究使用**蛋白质组**和**转录组**的方法识别与先天免疫功能相关的关键蛋白，鉴定到35种富集于抗病毒限制因子的蛋白，最后通过一种病毒突变子 panel 预测到靶向250多种人类蛋白的病毒基因。本研究的方法和数据预示了先天抗病毒免疫分子的重要性，能够进一步识别 HCMV或其他病毒靶向的宿主先天免疫途径。

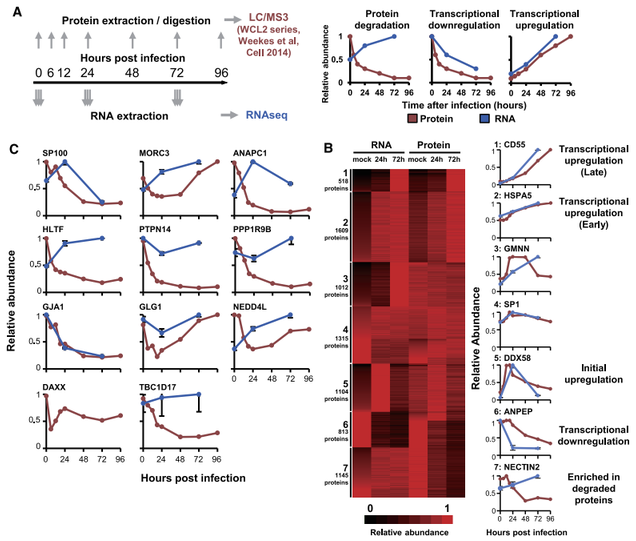
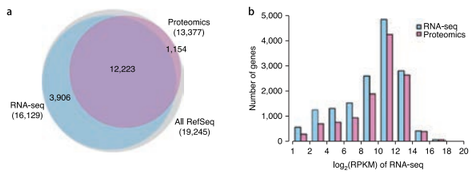


图13 通过转录本和蛋白质丰度关联分析明确宿主蛋白调控机制

a.实验流程示意图；b.7516个蛋白和转录本的聚类分析，右边举例说明7种表达模式；c.部分基因的转录本和蛋白质丰度检测结果。

案例二：人类抗艾滋病病毒蛋白PSGL-1及其调控机制[8]

HIV通过攻击患者体内CD4+T淋巴细胞，使患者丧失免疫功能、易患各种并发疾病，最后导致死亡。深入研究HIV-宿主细胞互作可优化现有抗艾滋病毒药物靶点为艾滋病患者带来福音。来自清华大学、复旦大学和美国乔治梅森大学的研究人员，通过iTRAQ定量蛋白质组学、转录组学结合传统病毒学研究策略，从人类细胞中筛选出一种新型抗HIV蛋白PSGL-1（P-selectin glycoprotein ligand 1）。PSGL-1具有多重抗病毒功能，包括抑制病毒DNA复制、抑制新生病毒颗粒的二次感染。HIV病毒通过其附属蛋白Vpu与PSGL-1结合并促进PSGL-1的降解，从而逃逸PSGL-1的抗病毒功能。研究表明，Vpu和PSGL-1的结合抑制剂可为抗艾滋病新药研发提供了新作用途径和理论基础，具有极大的临床应用潜能。



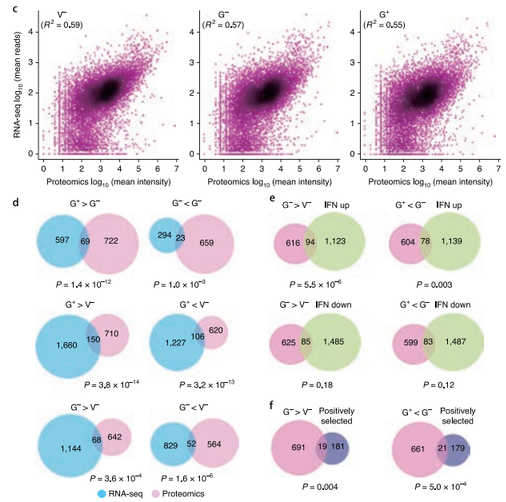


图14 蛋白质组和转录组关联分析比较

a.蛋白质与转录本的韦恩图；b.基于RNA-Seq数据定量的转录本和蛋白质分布；c.蛋白组和转录组相关性分析。d-f. 典型细胞系的蛋白与转录组维恩图展示

案例三：多组学技术研究埃博拉病毒发病机制及组合生物标志物的发现[9]

2013-2016年西非埃博拉病毒病（EVD）疫情是迄今为止最具破坏性的人类埃博拉疫情，2018年5月，埃博拉病毒又突然在非洲中部国家刚果（金）爆发，迄今为止，已有超过1600人因此死亡。研究团队收集了11个EVD感染幸存者，从初步诊断至康复的一系列样本共计29例、9个EVD致死患者死亡前样本以及10个健康的志愿者对照样本。基于**转录组学、蛋白质组学、代谢组学及脂质组学平台**，发现血浆游离氨基酸（PFAAs）、葡萄糖、果糖、二酰基甘油磷酸甘油、单酰基甘油磷酸丝氨酸（PS）、神经酰胺、可溶性VSIG4、骨髓细胞趋化因子受体（CCR1和CCRL2）等的变化，并由此推断肠组织损伤、T细胞激活受损、炎症、胰腺组织损伤和胰酶释放等可能在EVD发病机制中的作用，研究结果有助于改善高危患者的预后。

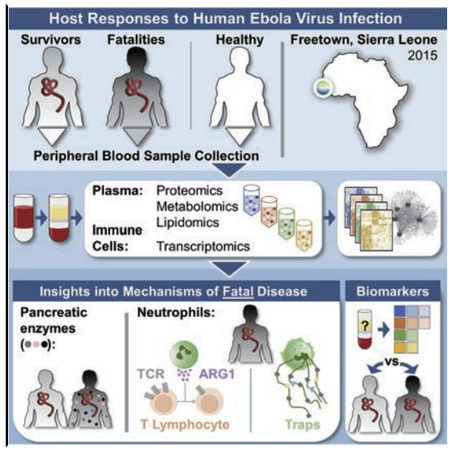


图15 实验整体流程图

取患者和健康人的血液样本和免疫细胞，分别开展蛋白质组、代谢组、脂质组和转录组分析，寻找biomarks。

可能存在的风险

蛋白质组和转录组关联分析后，发现蛋白质组和转录组相关性不高，约为27%~40%[10]。这种情况是比较常见的，主要原因是转录和翻译的速率不同，mRNA和蛋白质的半衰期也不同，mRNA和蛋白质的丰度并非总是线性相关。研究者可通过代谢通路分析，研究mRNA和蛋白质的上下调关系，分析关键的调控基因，深入研究代谢通路的精细调控关系，结合基因组层面的突变信息，寻找与表型相关的完整代谢调控通路。

常见问题

1、蛋白质组和转录组生物重复是否需要一致?

答：是的，推荐蛋白质组和转录组选取的生物重复一致。如果项目设计无法保证二者的生物重复完全一致，也可以开展关联分析，需要分别提供两个组学层面的相关信息，如鉴定定量到的差异表达基因，表达量，p-value等。

2、非华大的转录组数据或者蛋白质组数据是否能用华大的分析流程进行关联分析?

答：可以，其他公司的转录组数据有基因序列文件和转录组定量文件可进行关联分析；蛋白质组数据需要蛋白质组鉴定序列文件、蛋白质组定量文件可进行关联分析。

3、转录组和蛋白质组相关性不高，应该怎样开展解释和后续分析?

答：需要区分具体是哪种类型的相关性系数不高，建议按照常见的几大类开展解释和后续分析：

1）蛋白和mRNA表达趋势相同的相关性不高：说明表达一致性不高，这种情况很常见，主要原因是转录和翻译的速率不同，mRNA和蛋白质的半衰期也不同，mRNA和蛋白质的丰度并非总是线性相关；后续可通过其他实验方式进行关键基因表达产物的验证；

2）蛋白和mRNA表达趋势相反的相关性不高：此类表达情况主要说明一些特殊的调控关系，如反馈抑制调节方式的蛋白，相关性系数并无高低好坏的判定；后续分析可根据差异富集的关键代谢通路，逐一查看相关的蛋白和mRNA的表达情况，帮助分析上下游基因的代谢调控关系；

3）仅单一组学表达有差异，或者两组学层面均无差异的情况相关性不高：此类表达情况主要是前两种情况的补充，相关性系数也并无高低好坏的判定；后续分析也是需要结合代谢通路的上下游调控关系进行代谢调控关键因子的分析，避免基因表达产物mRNA和蛋白一对一的相关性关系，拓展到更大范围的对通路或者功能条目的影响中分析将会获得更多可能的结论。

华大优势

最强悍的数据兼容性：无障碍兼容多种蛋白质组和转录组数据类型，测序、芯片等具有定量信息的数据均可分析，推荐采用蛋白定量iTRAQ、IBT、DIA和RNA-Resequencing数据进行关联分析；

更细致的关联分类：按照蛋白组和转录组基因表达相关性趋势将数据细分为5大类，对于每类相关联的结果提供更完整的相关性分析和功能注释信息；

完美的通路关联方式：借助最新版KEGG数据库丰富的通路信息，将原本的蛋白组和转录组鉴定、定量层面的关联拓展到功能和代谢通路的关联，对分析关键基因对上下游相关蛋白的调控作用，以及迅速找到两个组学层面差异富集的关联代谢通路起到决定性作用；

华大一贯的高品质体验：完善的多组学研究技术，可提供一站式解决方案及相关技术服务；提供高质量的分析结果，报告结构科学、图片清晰可直接用于文章发表，有效提升报告阅读体验。

参考文献

1. Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, et al. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*. 13 January 2015.

2. Wang XC, Li Q, Jin X, et al. Quantitative proteomics and transcriptomics reveal key metabolic processes associated with cotton fiber initiation. *J Proteomics*. 2015 Jan 30; 114:16-27.

3. Trevisan S, Manoli A, Ravazzolo L, et al. Nitrate sensing by the maize root apex transition zone: a merged transcriptomic and proteomic survey. *J Exp Bot*. 2015 Apr 23.

4. Yang N, Xie W, Yang X, et al. Transcriptomic and proteomic responses of sweetpotato whitefly, Bemisia tabaci, to thiamethoxam. *PLoS One*. 2013 May 9;8(5): e61820.

5. Chen Q, Guo W, Feng L, et al. Transcriptome and proteome analysis of Eucalyptus infected with Calonectria pseudoreteaudii. *J Proteomics*. 2015 Feb 6; 115:117-31.

6. Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*. 13 March 2012.

7. Nightingale K, Lin KM,Ravenhill BJ, Davies C1, Nobre L, et al. High-Definition Analysis of HostProtein Stability during Human Cytomegalovirus Infection Reveals AntiviralFactors and Viral Evasion Mechanisms [J]. *Cell Host Microbe*. 2018 Sep12;24(3):447-460.

8. Liu Y, Fu Y, Wang Q, et al. Proteomic profiling of HIV-1 infection of human CD4+ T cells identifies PSGL-1 as an HIV restriction factor. *Nat Microbiol*, 2019 May;4(5):813-825.

9. Amie J, Eisfeld, Peter J, et al. Multi-platform' Omics anlysis of human Ebola virus disease pathogenesis. *Cell Host Microbe*. 2017 Dec 22,817-829.

10. Muers M. Gene expression: Transcriptome to proteome and back to genome. *Nat Rev Genet*. 2011 Jun 28; 12(8):518.

肿瘤融合基因研究方案

研究背景

细胞是生命的单位，目前大部分的基因检测均是从组织中抽提DNA 来进行测序，得到的实验结果往往是细胞群体中信号表达的均值，是对细胞群体进行整体表征，或者只代表其中在数量上占优势的细胞信息，单个细胞独有的细胞特性往往被忽略。

而大量研究发现在同一器官或组织的相同类型细胞也表现出显着的异质性，每个细胞都有其独特的表达模式。例如实体瘤样本的总RNA，一半以上来源于非癌细胞（成纤维细胞、淋巴细胞、巨噬细胞等），使得癌细胞的信号可能被隐藏。因此，采用均值对单个细胞进行表征是不合适的，可能会丢失许多关键信息。

另一方面，传统高通量测序方法，难以应用在对自然界中难培养的微生物的研究、罕见循环肿瘤细胞的转录组分析、胚胎发生最早期的分化特征研究、肿瘤的非均质性和微进化研究等精确程度较高的研究领域。随着细胞分选和测序技术进步，单细胞测序技术应运而生。

2011年，《自然方法》杂志（Nature Methods）将单细胞测序列为年度值得期待的技术之一，2013年，《科学》杂志（Science）将单细胞测序列为年度最值得关注的六大领域榜首。随着单细胞测序技术发展，近几年出现大规模单细胞分选平台：2015年WaferGen推出ICELL8 Single-Cell System，每次运行可分离500-1000个细胞，大幅降低肿瘤单细胞转录组测序研究的成本，在肿瘤单细胞精准研究有很大的应用前景。2016年，10x Genomics推出Single Cell 3' Solution，2019年，华大智造推出DNBelab C4便携式单细胞分选仪，可以实现高通量分选单细胞，特别适合大量单细胞的精准研究，助力疾病分子致病机理挖掘。

方案设计

A. 研究目的

组织的功能需要借助不同细胞之间的相互协作才得以实现，这其中包括上皮细胞，免疫细胞，基质细胞等等，而任何一类细胞的功能失调都可能导致疾病的发生。

通过高清单细胞图谱绘制，可以探究了组织成分变化和不同细胞类型之间相互作用是如何影响疾病的，并且确定了与疾病相关的细胞和通路基因在这其中扮演的角色，以及不同细胞之间是如何传递信号、协同合作的；每种细胞产生、释放的信号分子，接受怎样的信号传入，进一步探索单细胞级别分子生物学进程中可能病因。

A. 样本选择

样本类型：

不同时间点

药物处理 vs 对照

患病样本 vs 对照

生物学重复：

可选择3对及以上

C. 技术方法

单细胞悬液制备——流式细胞分选目的细胞（可选）——高通量单细胞分选—— 单细胞转录组建库——高通量测序

D. 分析内容

- 基因表达鉴定
- 细胞类型鉴定
- 细胞类型差异分析
- 特异性细胞类型功能分析
- 特异性细胞代谢通路分析
- 特异细胞类型互作功能分析
- 目标疾病类型maker鉴定

E. 项目执行周期

BGISEQ平台测序，实验+信息分析标准分析周期——35个自然日。

F. 预期结果

通过特异性细胞类群鉴定及比比较分析，可以实现对照组间差异分析，以及结合特异性细胞亚群基因功能分析，可以用于解释疾病发生相关分子生物学相关机制。

G. 后期验证手段

验证手段主要有以下几种：

a. 流式细胞分选——验证标记marker基因

b. qPCR—— 验证标记基因定量

c. RNA fish试验——验证marker基因，以及对应的细胞群在组织中的空间位置

d. 免疫荧光试验——验证marker基因，观察marker基因对应的细胞群在组织中的空间定位，不仅如此还能观察蛋白在细胞内的定位，根据不同的细胞器定位进行后续的功能研究提供线索。

e. 谱系示踪技术——验证细胞群的演化关系谱系示踪。

应用案例

案例一：单细胞技术研究正常和非酒精性脂肪性肝炎（NASH）小鼠肝脏细胞间信号传递 Landscape of intercellular crosstalk in healthy and NASH liver revealed by single-cell secretome gene analysis

发表期刊：Molecular Cell

影响因子：14.548

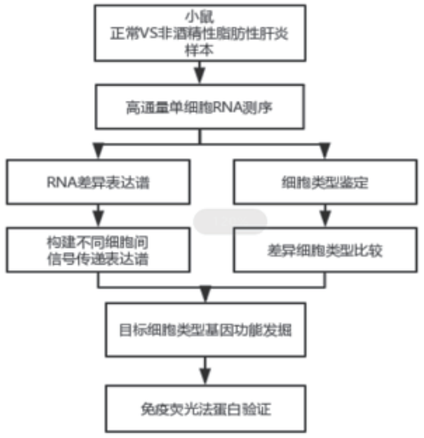
发表时间：2019年8月8日

研究背景：NASH是非酒精性脂肪性肝病的进展和极端形式，在NASH基础上发生肝硬化、肝细胞肝癌的风险明显增加。目前已知，NASH与肥胖、代谢综合征密切相关，然而驱动NASH发生的始动因素和具体机制尚不十分清楚。该项研究利用单细胞测序这一新兴技术对NASH的发病机制进行了新的探索。

样本来源：三对正常和非酒精性脂肪性肝炎（NASH）小鼠肝脏

技术方案：采用高通量单细胞分选模式，对6个样本单细胞悬液进行细胞分选，然后进行转录组建库、测序及分析。

分析思路：



主要结果展示：

1、细胞图谱绘制：

研究人员首先通过对肝脏细胞进行单细胞RNA测序，获得每种细胞的基因表达谱信息。在此基础上，进一步区分每种细胞组分的分泌蛋白表达谱以及受体表达谱，在此基础上构建了肝脏不同细胞间信号传递的高分辨图谱。

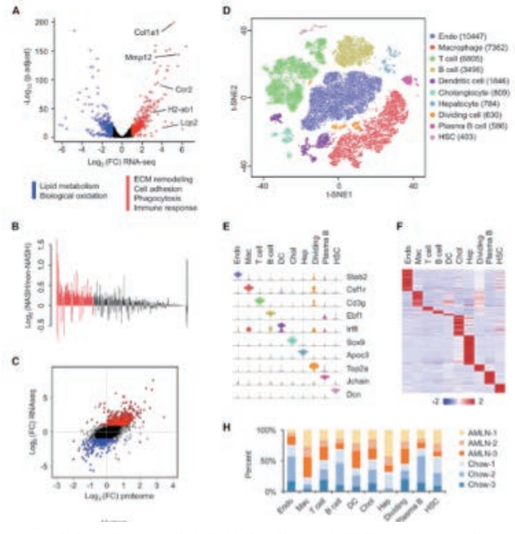


图1.从健康和NASH小鼠肝脏分离的NPC的单细胞RNA-Seq分析

RNA表达火山图、单细胞分类图，
以及特异单细胞类型分泌蛋白表达谱以及受体表达谱

2、差异细胞类群比较

进一步比较对照肝脏和NASH小鼠模型，研究人员发现，在NASH进程中不仅细胞数目发生了改变，每种细胞的特性也发生了改变。最为显著的变化是健康肝脏中Trem2阳性的巨噬细胞数目极少，而NASH时这一比例显著升高。这一标记基因在NASH患者肝脏中表达也是升高的。利用这一特性，研究人员检测外周血中Trem2阳性细胞产生的分泌蛋白，分析其与NASH进展的关系，以期寻找新的NASH血清学标记物。

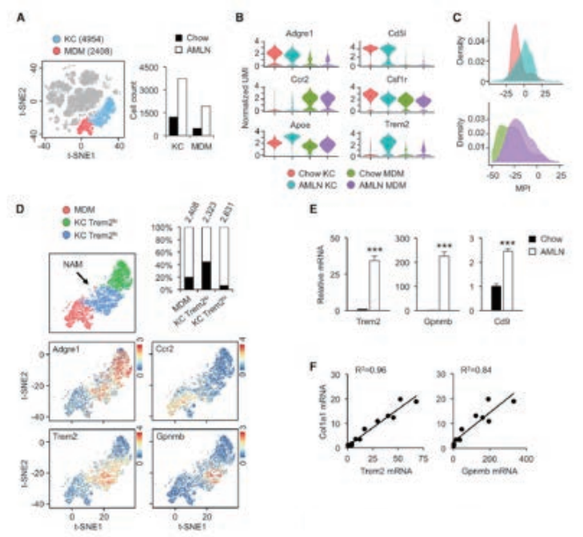


图2 NASH相关的巨噬细胞在肝脏中的出现

(A) 驻留在组织中的库普弗细胞 (KC，蓝色) 和单核细胞衍生的巨噬细胞 (MDM，红色) 的插图。来自松鼠和AMLN小鼠肝脏的总细胞计数每个子集群都显示在右侧 (n = 3)。

(B) 归一化UMI的小提琴图，显示标记基因表达的分布。

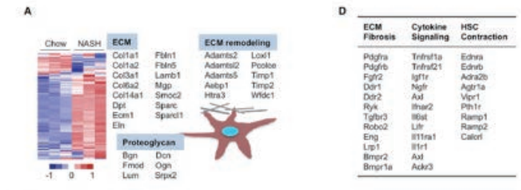
(C) 肝巨噬细胞的巨噬细胞极化指数的直方图。细胞类型和饮食的颜色如 (B) 所示。

(D) t-SNE图说明了以低 (绿色) 和高 (蓝色) Trem2 mRNA表达为标志的KC的亚群。松饼 (实心) 和显示了每个亚群的AMLN (开放) 巨噬细胞和总细胞数。标记基因表达的特征图显示在底部。

(E) 喂养6个月或AMLN日粮的小鼠中NAM标记基因的全肝qPCR分析 (n = 4)。

3、特异性细胞类群功能研究：

发现是肝脏星型细胞 (HSC) 在活跃的信号传递中功能。以往概念中，肝脏星型细胞的主要是产生细胞外基质成分、参与纤维化的发生，该项研究发现，肝脏星型细胞除了产生细胞外基质成分以外，能分泌众多的信号蛋白分子，称为“Stellakines”。同时，肝脏星型细胞表面亦有众多的受体，能接受胞外信号，调控其包括收缩活性在内的生理功能。



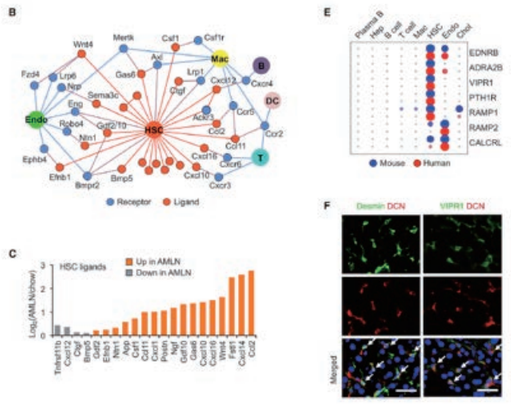


图3 肝星状细胞HSC信号互作网络图

(A) 使用来自中国松鼠和AMLN小鼠的肝脏RNA-seq数据，对HSC富集的分泌相关基因热图

(B) HSC分泌组配体受体配对。在HSC簇中表现出> 3倍富集表达的配体显示为橙色，其已知受体显示为蓝色，当基于scRNA-seq数据集在至少一个簇（归一化UMI> 1.0）中观察到受体表达时，

(C) 在NASH中恒星素基因表达的调节。使用来自chow和AMLN肝RNA-seq数据集的平均表达值。

(D) 富含HSC的膜受体。

(E) 小鼠和人类肝细胞中受体基因表达的点图。

(F) 对冷冻的肝脏切片进行免疫荧光染色。细胞核用DAPI（蓝色）染色。箭头表示共定位HSC中蛋白质表达的比例（比例尺，50 mm）。

案例二：单细胞技术研究结肠炎过程中人结肠细胞内和细胞间重排

Intra- and Inter-cellular Rewiring of the HumanColon during Ulcerative Colitis

发表期刊：Cell

影响因子：36.216

发表时间：2019年7月25日

研究背景：溃疡性结肠炎是炎症性肠病(inflammatory bowel disease, IBD)的一种。目前已知的IBD致病风险因子所涉及的通路包括，先天及获得性免疫，胃肠道屏障以及病原体感受和响应系统。事实上，尽管GWASs(Genome-wide association studies)分析已经找到了其中很多的致病风险因子，但是没有搞清楚的是，这些致病因子究竟和哪些细胞种类和信号通路相关

样本来源： UC患者 vs 正常人（18 vs 12）

三组样本： 经不同药物治疗后临近溃疡处的正常组织(non-inflamed)和溃疡或发炎组织(inflamed)，与健康人群的组织（healthy）（分两批采样测序合并数据分析）

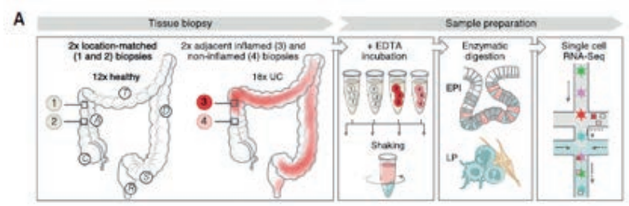
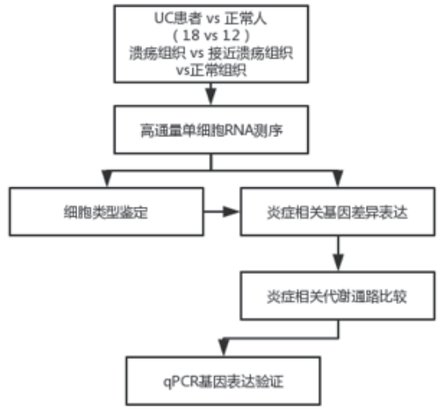


图4 样本取样及单细胞悬液准备和实验过程

技术方案：

采用高通量单细胞分选模式，制备单细胞悬液进行细胞分选，然后进行转录组建库、测序及分析。

分析思路：



主要结果展示：

1、细胞类型鉴定：

本次获得366,650个单细胞转录组数据，并对鉴定到的细胞进行分类：一共鉴定到51种细胞，主要分为上皮细胞(epithelial)，基质细胞(stromal)，免疫细胞(immune cell)三大类

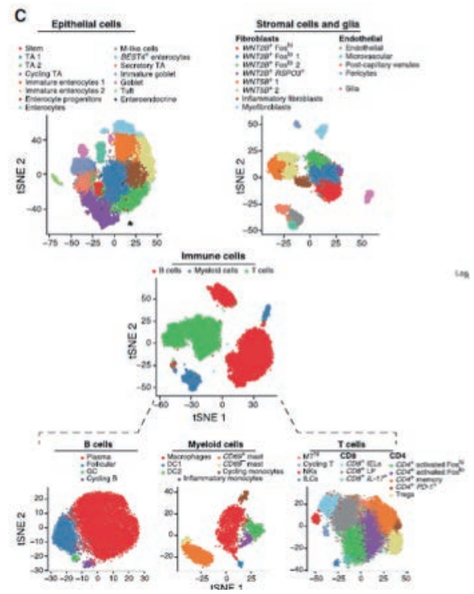


图5 BEST4 +上皮细胞和RSPO3 +的表征健康结肠中的成纤维细胞

2、差异细胞比较：炎症相关的成纤维细胞亚群是UC Colon独有

皱褶细胞（Microfold-like）-M-like细胞(黏膜免疫系统中一种特化的抗原转运细胞，散布于肠道黏膜上皮细胞间)在正常结肠组织几乎检测不到，但在炎症结肠组织中却显著增加，并且在细胞间的相互作用中扮演重要角色。

从健康组织到UC患者的未发炎组织再到发炎组织，CD8+IL-17+T细胞和Tregs细胞数量均显著上升，并且是细胞因子IL-17和TNF的主要来源。同时作者发现，TNF+ Treg细胞可能是决定IBD病程发展以及导致TNF抗体抵抗的一个重要细胞类型，并且对于CD8+ T细胞的可塑性调节也很关键。

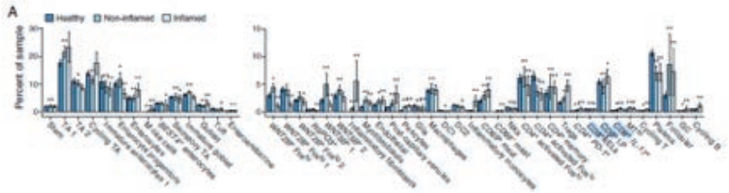


图6 健康组织到UC患者的未发炎组织再到发炎组织特异性因子表达差异分析

3、炎症相关基因表达差异：

作者还发现IAFs特异性的表达IL-11，一个已经报道在小鼠体内可以调控纤维化的细胞因子，并且很有可能在人里发挥同样的功能。更有趣的是，IAFs不仅表达很多CAFs的标志物，并且很多IAFs的标志物在结肠癌中也是高表达的。这提示，IAFs在结肠的炎癌转化过程中可能扮演重要角色。

4、特异性代谢通路研究

在某些未知的髓细胞和基质细胞中，OSM(Oncostatin M,一种多效性的细胞因子,属于IL-6组的细胞因子)信号通路也可能是导致TNF抗体抵抗的诱因之一。因为作者发现OSM与TNF的功能十分类似，因此一种可能的解释是OSM可以协同增强TNF的作用，从而导致TNF抗体抵抗。同时，作者发现对于TNF抗体没有响应的组织样本中，一类炎症相关的成纤维细胞(inflammation associated fibroblasts，IAFs)高度富集，这意味，IAFs可以作为一个非常理想的生物标志物用于评估IBD对于TNF抗体的耐药性。

案例二：衰老影响CD4 T 细胞分化和功能表型

Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes

发表期刊：Science Advances

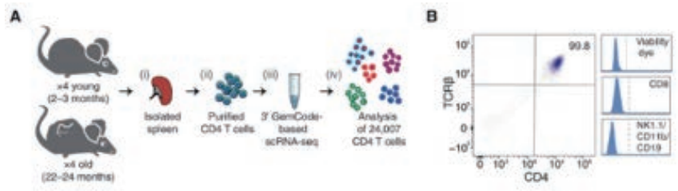
影响因子：12.804

发表时间：2019年8月21日

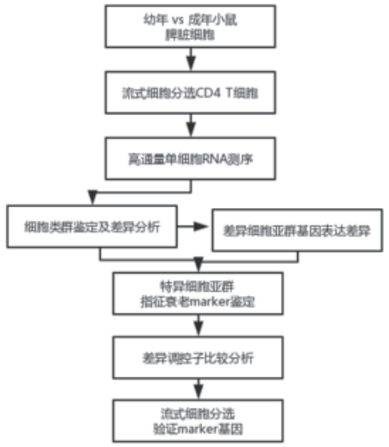
研究背景：衰老的关键标志之一是免疫系统的恶化，使老年人更易于感染，慢性炎症性疾病和疫苗接种失败。在衰老中观察到的显着变化与CD4 T细胞的组成和功能有关，CD4 T细胞是适应性免疫应答的主要协调者。在年轻的啮齿动物和人类中，CD4 T细胞包含高频率的幼稚细胞，反映了免疫系统遇到新抗原能够有效对其应答并产生免疫记忆的能力。随着年龄的增长，免疫记忆随着高度分化的存储单元的积累而收缩，这些存储单元通常显示出失调的特性。总体而言，衰老与CD4 T细胞区室结构的改变有关，在此积累的功能障碍亚群会导致免疫衰竭。尽管这些研究提供了对衰老中发生的功能改变的见解，但它们主要依赖于先前描述的CD4 T细胞亚群的标志物或少量细胞，这些标志物可能掩盖了更复杂的细胞组成。因此，需要对衰老的CD4 T细胞群体中的细胞亚群的组织进行无监督的大规模研究。

样本来源： 2-3个月的小鼠（n = 4） vs 22-24个月（n = 4）的小鼠（C57BL/6）：分别提取脾脏细胞

技术方案：使用流式细胞仪富集CD4 T细胞，采用高通量单细胞分选模式，制备单细胞悬液进行细胞分选，然后进行转录组建库、测序及分析。



分析思路：



主要结果展示：

1、共获得13186个幼年小鼠和10821个成年小鼠CD4 T细胞。对不同群体的单细胞进行基因表达的分析，发现小鼠年龄的不同确实造成了CD4 T细胞分群的不同，共鉴定出7个不同的亚群包括细胞毒性亚群，效应记忆性T细胞亚群，耗竭性细胞亚群，激活性调节性T细胞亚群，调节性T细胞亚群，Isg15表达的调节性T细胞亚群和Naïve 细胞亚群。其中Naïve亚群主要出现在成年小鼠中，调节性T细胞亚群比例均等，激活性调节性T细胞，耗竭性T细胞和细胞毒性T细胞（这三类细胞缩写为RECs）却在所有老年小鼠细胞中均高度富集。

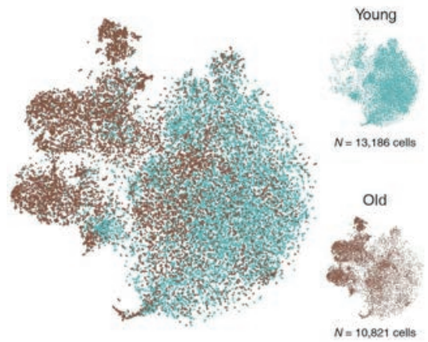


图7 幼年和成年小鼠单细胞图谱（分别来自年轻（绿松石）小鼠和年老（棕色）小鼠）

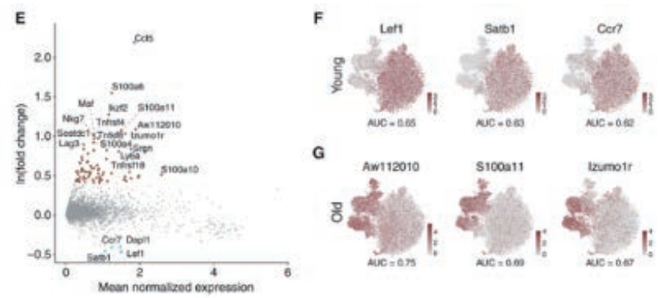


图8 标记幼年和成年T细胞类型的marker基因

2、发现了27个高信号的调节因子，根据调节子进行的亚群分类和10xGemcode系统得到的结果一致，Prdm1调节子暗示了RECs的耗竭水平；此外，作者还发现了激活性调节性T细胞展现了Foxp3, nuclear factor kB (NF-kB) TF family,Irf4, 和Maf活性，暗示了与调节性T细胞的激活相关。而细胞毒性T细胞主要被细胞毒性和辅助性1型T细胞转录因子调节，展示了极高的分泌促炎因子的特性。耗竭过程相关的Nfatc1 调节子出现在耗竭性T细胞群体中。

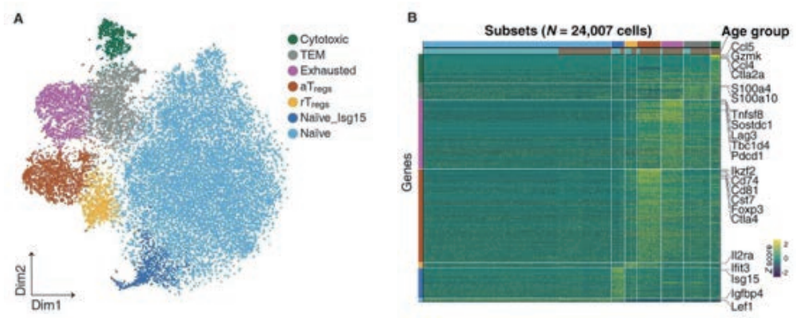


图9 通过调节因子进行单细胞分群，以及各亚群maker基因热图分析

3、根据不同年龄段细胞的差异表达分析确定了已经报道的基因和炎症程度的相关性，并确定了这些基因可以作为老化的CD4 T细胞的标记基因。并利用流式分选进行验证。

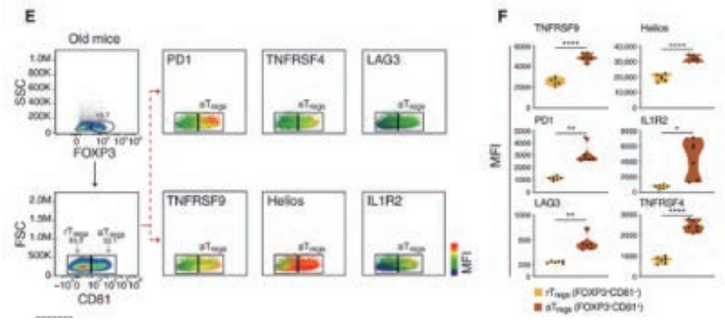


图10 流式细胞分选验证标记基因结果图

可能存在的风险

1、细胞悬液风险：单细胞悬液因取样问题或样本本身问题，导致单细胞活率低，进而导致高通量单细胞分选得到的有效单细胞数少，建议严格参照送样取样建议。

2、特殊样本类型风险：单细胞体积大于40μm，容易堵塞分选孔道，导致油包水结构生成失败，致使实验失败，建议选取合适细胞大小，或者采用细胞核分选方式研究。

3、基因数鉴定偏少：测序数量量不足，导致平均单个细胞reads数覆盖度偏度，导致细胞基因中位数偏少，建议增加测序数据量。

4、流式细胞分选验证风险：抗体的合适选择影响marker基因的验证。

5、重复样本的选择：建议基于一定的生理生化重复试验基础上，指导单细胞取样。

华大优势

- 2倍基因数*，探索更多可能：自研高效mRNA捕获微珠系统+高效破乳回收系统+高深度测序，助力发现更多稀有细胞亚群，深入分析细胞功能；
- “0”单细胞数据拆分错误#：搭载DNBSEQTM测序系统，不惧样本间单个细胞数据弄混；
- 现选现测，一站式服务，呵护你的样品：独家流式分选服务，一站式实现从组织到数据，避免细胞反复冻融；
- 样本不挑剔：除常规活细胞悬液外，珍稀液氮速冻样本皆可实现单细胞水平研究。
- 高、低通量单细胞研究产品一网打尽：高通量10x Genomics单细胞免疫组库/RNA-Seq/ATAC-seq产品；DNBelab C4 高通量单细胞RNA-Seq产品；单管单细胞DNA/RNA/甲基化；微量单细胞研究；

参考文献

[1] Xiong X, Kuang H, Ansari S, et al. Landscape of intercellular crosstalk in healthy and NASH liver revealed by single-cell secretome gene analysis[J]. *Molecular cell*, 2019, 75(3): 644-660. e5.

[2] Smillie C S, Biton M, Ordovas-Montanes J, et al. Intra-and inter-cellular rewiring of the human colon during ulcerative colitis[J]. *Cell*, 2019, 178(3): 714-730. e22.

[3] Elyahu Y, Hekselman I, Eizenberg-Magar I, et al. Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes[J]. *Science advances*, 2019, 5(8): eaaw8330.