**泛人类质谱谱图库DPHL的构建和应用**

**要点介绍**

• 研究问题：

  人体多种组织DIA谱库的构建及应用

• 研究方法：

  通过Docker技术封装并创建新的建库流程，使用该流程建立DIA谱库，通过分析临床队列验证谱库的实用性

• 主要结论1：

  建立了开源的DIA谱库的流程

• 主要结论2：

  构建了谱库DPHL

• 主要结论3：

  通过对前列腺癌和弥漫大B细胞淋巴瘤队列的DIA和PRM数据分析验证了DPHL可以作为DIA的谱库分析临床样本

• 数据链接：

  原始质谱文件及谱库：  
https://www.iprox.org/page/project.html?id=IPX0001400000  
  DPHL Docker镜像：  
docker pull guomics/dphl:v02

**背景**

近年来，高通量基因组和蛋白质组等多种组学技术的快速发展和广泛应用，促进了各种疾病生物标志物的发现，有力地推动着癌症精准医学的成熟。SWATH/DIA是瑞士苏黎世联邦理工学院的Ruedi Aebersold博士及其团队与SCIEX于2012年联合推出的一种数据独立采集技术，它将鸟枪蛋白质组学的蛋白量鉴定能力与SRM的高精度数据分析相结合，进行快速的蛋白质组定量研究。近年来，DIA技术在多个质谱技术平台（包括Orbitrap）得以广泛应用。SWATH/DIA的数据分析可以使用OpenSWATH，DIA-Umpire，Group-DIA，Skyline和Spectronaut等软件。这些软件大多都需要一个先验的谱图库。尽管蛋白质组学的研究取得了这些进展，但由于建立质谱谱库非商业工具的缺乏，到目前为止还没有一个完整的基于Orbitrap的人体多种组织的谱库，更没有一套生物信息学的流程可以分析DIA数据，并采用PRM-MS来做验证。

Spectronaut是业界广泛使用的用于分析DIA数据和建立谱库的软件，遗憾的是，Spectronaut只提供商业版本，费用昂贵。而目前免费的建库方法安装配置过程复杂，并且过程繁多。为此，本文开发了一套开源的从Orbitrap质谱数据中构建谱库的计算方法和流程，并利用这个流程从16种不同的人类癌症类型中生成了一个综合的泛人类质谱谱图库（DPHL）。

**谱库建立**

**谱库流程**

Docker作为一个开源的轻量虚拟化软件，使用者可以打包自己的一系列软件以及依赖包到一个可移植的镜像中，非常方便的发布到Mac， Linux或Windows 机器上，实现虚拟化应用。我们通过Docker封装了建立谱库的工作流程。DPHL的构建过程如**图1**：A.鸟枪法蛋白质组学数据采集示意图。括号中的数字表示每种组织类型的DDA文件的数量。B.-E.用于建立DIA谱库的计算流程。B.使用pFind从DDA原始文件中进行蛋白鉴定和iRT检测。C.使用SiRT建立谱库。D.建立CiRT并通过CiRT建立谱库。E.生成最终的谱图库。

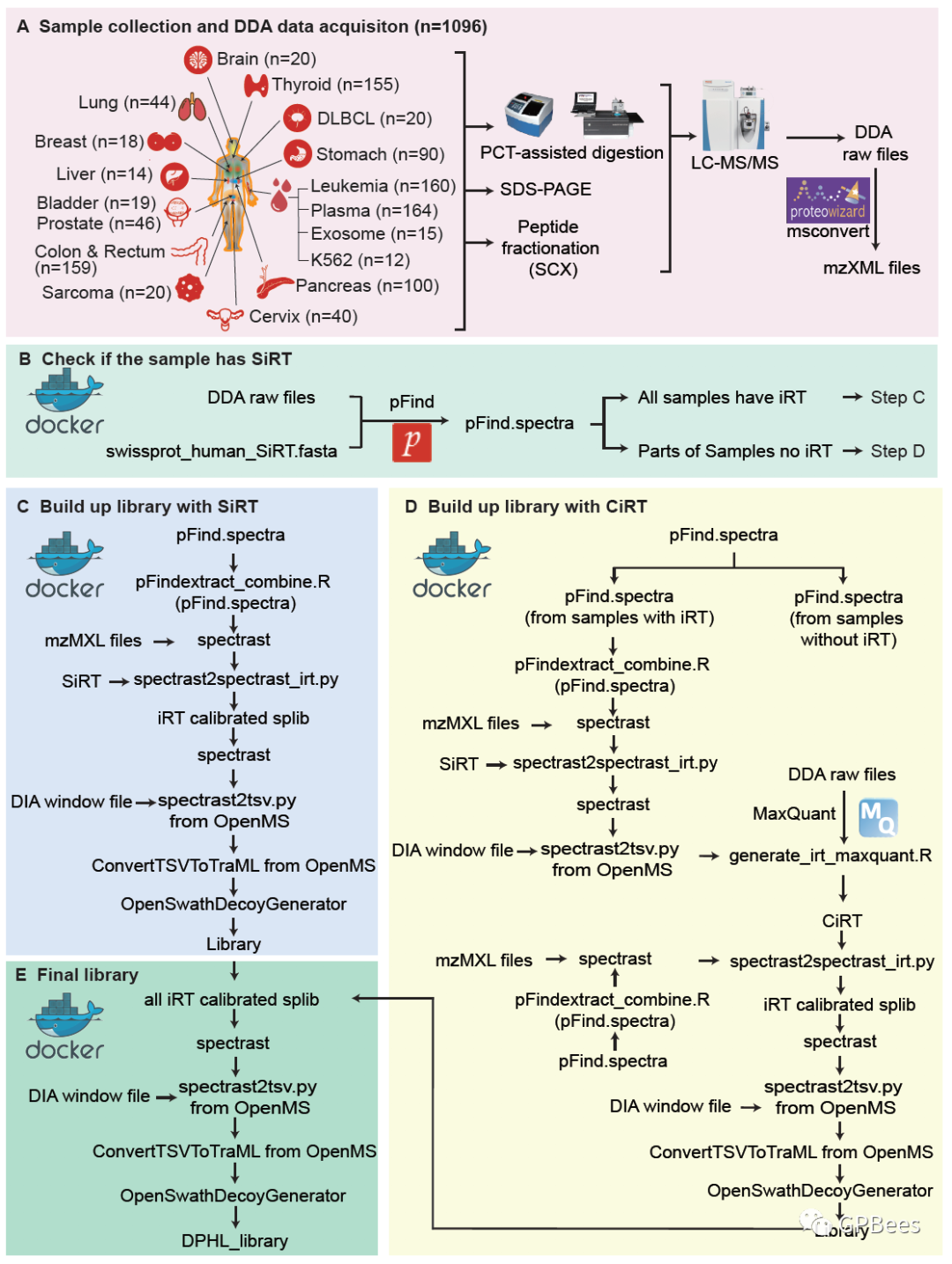


图1 DPHL谱库构建流程

**谱库分析**

对生成的DPHL库的内容进行了肽和蛋白分析，并与SWATH的PHL文库进行了比较。DPHL库包含359,627个母离子、242,476个唯一肽序列、14,782个蛋白质群和10,943个含唯一肽段的蛋白。DPHL和PHL共同拥有9241种蛋白,分别占DPHL总蛋白数量的84.4%，和PHL总蛋白数量的89.5%；与PHL的SWATH库相比，DPHL文库包含更多的母离子、肽序列和蛋白（如**图2**A所示）。血浆的蛋白质组因高丰度蛋白的影响而同其他组织差异较大。而从文章分析的其他组织中鉴定到的共同蛋白近2000个。

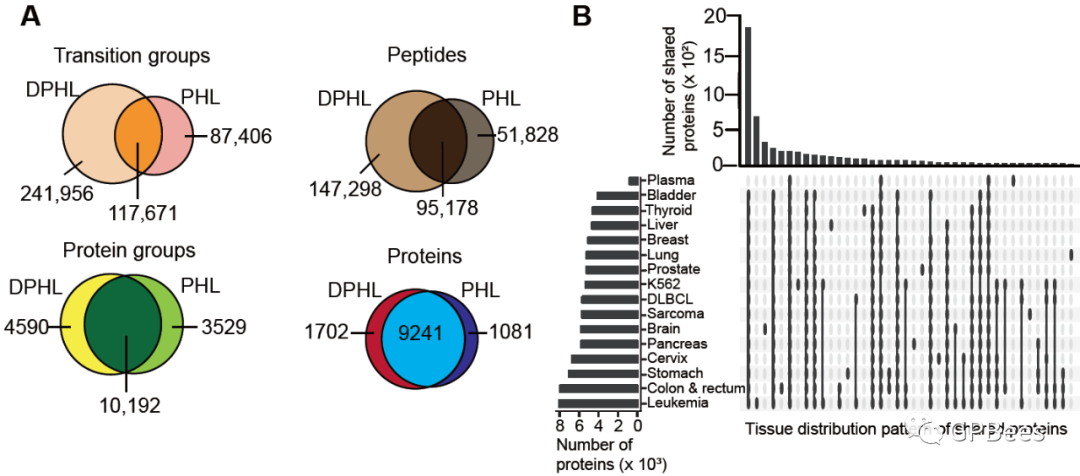


图2 DPHL和PHL的比较

**临床队列应用**

**案例1**采集了17例患者的前列腺组织样本，其中包括8例前列腺癌（PCa）和9例良性前列腺增生（BPH），采用DIA模式下的QE-HF质谱进行分析。利用CiRT策略在DPHL库中搜索出4785个蛋白质群，鉴定到37,581个多肽前体离子（**图3**A）。图3B显示，在所检测的样品中，质谱对蛋白的鉴定具有高度的可重现性。使用SiRT和CiRT策略的对蛋白的鉴定能力相当（图3C）。t-SNE图显示，通过CiRT和SiRT两种策略分析的数据，可以清楚地区分PCa和BPH（图3D）。

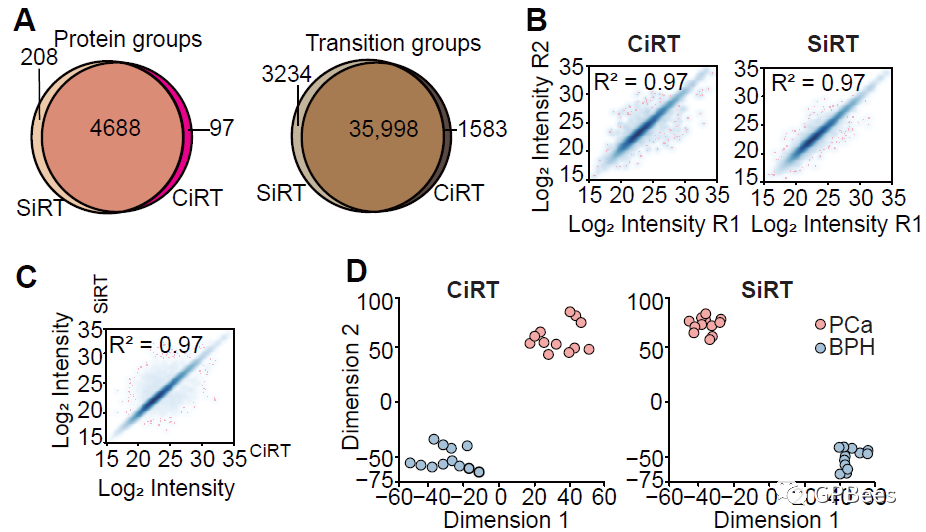


图3 前列腺组织DIA数据分析

**案例2**  
血浆因其获取非常方便，已广泛应用于临床诊断。在此，我们应用DIA-MS和DPHL分析弥漫性大 B 细胞淋巴瘤（diffuse large B cell lymphoma，DLBCL）患者的血浆样本。血浆样本取自19例DLBCL患者和18例健康人。胰蛋白酶酶解血浆样品，产生的肽使用20分钟色谱梯度分离，并在QE-HF仪器上用DIA-MS测量，**图4**A体现了方法较好的技术重复性。通过对定量的蛋白的t-SNE分析，DLBCL样品与健康对照样品有明显的区别（图4B)，说明通过DPHL的工作流程可以根据血浆蛋白质组将DLBCL患者与健康个体区分开来。

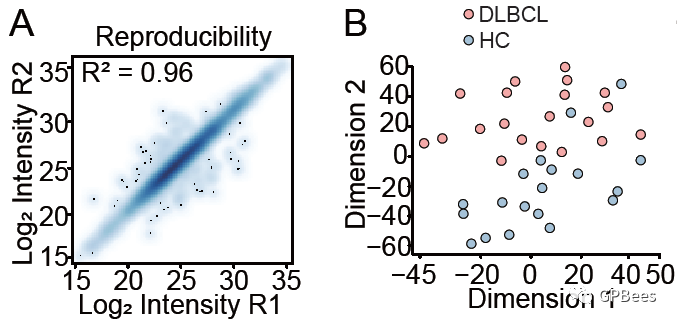


图4 DLBCL的DIA数据分析

文章进一步针对以上两个案例进行了蛋白标志物的筛选，发现了前列腺癌的3个蛋白候选标志物和DLBCL的2个蛋白候选标志物，利用该计算流程组建了一个泛人类谱库，并使用PRM技术进行了标志物的验证，展示了DPHL对生物标记物的探索、筛选和验证的全面支持。

**总结和讨论**

在这项研究中研发了一个适用于靶向蛋白质组学的建库流程，为Orbitrap仪器上的DIA和PRM分析生成谱库。并利用该计算流程组建了一个泛人类谱库DPHL，可用于DIA-MS鉴定和验证临床样本中的蛋白生物标志物。DPHL文库拥有超过370,000个多肽前体离子和10,000多个含有唯一肽的SwissProt蛋白，是迄今为止建立的最全的人类组织DIA库。此外，相对于PHL主要用于细胞系和合成肽的SWATH数据，DPHL则是专为各种临床样品（包括组织和血浆）的生物标记物的探索和验证而构建的。DPHL将继续从更多类型的人类肿瘤中合并更多的DDA文件，创建反映肿瘤异质性的综合资源，使生物标志物的发现成为人类蛋白质组癌症项目的一项使命。DPHL不仅适用于OpenSWATH等开源SWATH/DIA分析工具，还适用于Spectronaut和Skyline等其他工具。

文章编译来源：Tiansheng Zhu, Yi Zhu, Yue Xuan, Huanhuan Gao, Xue Cai, Sander R. Piersma, et al. DPHL: A DIA Pan-human Protein Mass Spectrometry Library for Robust Biomarker Discovery. *Genomics Proteomics Bioinformatics* 2020;18(2). 引用请参考以上格式，英文全文详见  
https://www.sciencedirect.com/science/article/pii/S1672022920300930。  
  
该项目由西湖大学生命科学学院蛋白质组大数据实验室联合国内外多个实验室和医疗机构共同完成。西湖大学特聘研究员郭天南博士为通讯作者，副研究员朱怡博士为共同通讯作者，朱天生、朱怡和轩玥为共同第一作者。该研究得到了国家基金委、浙江省杰青、杭州市科技发展计划项目等多项资金的资助。更多细节请参考www.guomics.com.