《基于质谱的定量蛋白质组技术规程》编制说明

一、工作简况

1、任务来源

本标准基于卫健委/科技部国家重点研发计划《医学生命组学数据质量控制关键技术研发与应用》（2018YFC0910200，2018年立项，2021年结题）和广东省重点领域研发计划《多组学整合技术研发及标准化组学数据质量控制技术的推广应用》（2019B020226001，2019年立项，2022年结题）两个项目的研究成果，本标准由深圳承启生物科技有限公司于2023年9月提出申报，根据中国国际科技促进会标准化工作委员会[2023]中科促标字第982号文件：关于开展《基于质谱的定量蛋白质组技术规程》团体标准立项通知，正式确立立项，本项目计划编号为：CI2023419。

2、目的和意义

质谱是目前蛋白质组学研究的主流技术，可以一次性分析复杂样品中上万种蛋白质。但质谱的重现性一直是一个很大的问题。同一个样品做两次质谱实验，鉴定到的蛋白质都有较大的不同，重现性（即两次都能鉴定到的蛋白数量占总鉴定数的百分比）一般只能达到65-85%。2017年《Nature Communications》上的一篇文章指出(Collins et al., Nature Communications)，只有约一半的蛋白质能在SWATH（Sequential Windowed Acquisition of all Theoretical fragment ions, SWATH）质谱技术中被重复鉴定到，而SWATH已经被认为是重现性相对较高的质谱技术了。《Science》曾组织专题研讨会讨论如何提高质谱的重现性，指出现有质谱的重现性远远不能满足生物标志物的要求。重现性问题如果不能得到很好的解决，那么基于质谱的生物标志物发现及临床应用将十分困难。对于定量蛋白质组学研究，因为样品制备流程、质谱平台和分析流程常常不一致，加大了蛋白质组学在精准医学中的应用难度；此外，Hela细胞蛋白酶解的标准品作为一种高度验

证的哺乳动物蛋白消化物，可用作复杂蛋白质组学样品的质谱分析的质控样品被人们广泛使用。然而Hela细胞因其存在超强的种内污染和种间污染而存在广泛争论。基于以上领域内存在的诸多问题，我们利用多组学技术手段评估并开发出一种或多种可作为蛋白质组标准品的细胞株，并以此制定蛋白质组标准操作流程供行业内流通使用。

目前国内外均无相关的标准流程，各蛋白质组学研究单位处于各自为战的状态。流程不通用，结果不可比。

本标准对样品采集、处理、运输、质控和检测以及数据产生、质量评估和数据分析全流程进行标准化和规范化。该标准的特点是：流程标准化，可应对各种类型样品蛋白质提取和制备需求，同时保持结果的重现性，对样品保存运输的要求低；开发出了具有极高蛋白质组稳定性且可稳定传代的蛋白质组标准品细胞株：MHCC97H，可用于监控实验流程或校准仪器，该标准品已申请国家发明专利，具有完整自主知识产权。

3、起草单位

本标准负责起草单位：深圳承启生物科技有限公司。

本标准参加起草单位：暨南大学、华南理工大学、北京师范大学、中国科学院大连化学物理研究所、福建农林大学、福建省妇幼保健院、上海易算生物科技有限公司、布鲁克（北京）科技有限公司。

本标准主要起草人：张弓、陈洋、余卓、卢少华、赵晶、李亚星、范小路、杜红丽、高友鹤、袁晓明、沈城频、王心睿、卢红、刘先明、杜萧贤。

参加单位工作分配：深圳承启生物科技有限公司负责标准方案的制定及网页发布；深圳承启生物科技有限公司、暨南大学、华南理工大学、中国科学院大连化学物理研究所、北京师范大学、福建农林大学、福建省妇幼保健院、上海易算生物科技有限公司和布鲁克生物科技有限公司参与了标准方案的实验测试和质量评估；暨南大学、深圳承启生物科技有限公司进行了标准起草和编制说明编写，其他参与单位对标准各个版本的全部内容提出

编写和修改意见。

4、起草过程

4.1 起草阶段

2019年1月至2019年6月，标准起草单位组织相关技术人员对标准项目进行了预研，课题组成员广泛收集了国内外基于质谱的定量蛋白质组技术的应用现状及技术发展趋势相关的标准、文献，了解了国内外相关技术动态，并且明确了工作思路和进程安排；

2019年7月至2022年4月，标准起草单位对标准草案进行实验验证和技术论证并确立了基于质谱的定量蛋白质组技术从采样、前处理、运输、质控和检测以及数据产生、质量评估和数据分析全流程，形成了《基于质谱的定量蛋白质组学技术规程》的初稿，并于2023年3月以网页形式向社会发布（https://translatome.net/Resources/omicsStandard/index.html）；

2022年8月，国家卫生健康委医药卫生科技发展研究中心组织专家对本标准流程初稿进行评审，参会专家做“基本完成主要相关考核指标”的评价，建议进一步加强相关数据采集与检测标准的认定，同时提出应加强应用示范的建议；

2022年9月至11月，编制组针对收集的意见，对标准进行了补充和完善。包括：1）增加了蛋白质提取质量检测标准；2）增加了使用MHCC97H标准品及标准数据集进行数据质量控制的内容；3）在暨南大学、华南理工大学、北京师范大学和布鲁克生物科技有限公司等单位宣传和推广本标准，充分交流并进一步收集相关单位的反馈意见；

2022年12月，向广东省科技创新监测研究中心组织的专家组进行总结报告，专家组考察了标准实施场地，审阅了相关资料并进行了咨询。项目开发技术已在多家机构推广应用，取得较好的评价和一定的经济效益；

本标准由深圳承启生物科技有限公司提出，由暨南大学和深圳承启生物科技有限公司主编，其他单位共同参与起草，于2023年9月提交团体标准《基于质谱的定量蛋白质组学

技术规程》的立项申请书，经中国国际科技促进会标准化工作委员会及相关专家技术审核，于2023年10月19日正式发布立项公告，并于同日在全国团体标准平台发布立项公示。

2024年1月，中国国际科技促进会标准化工作委员会完成了对本标准内容的审核并提出对格式重排和规范书写的意见，编制组进一步参考标准书写规范，修改和完善了标准文本内容，形成了《基于质谱的定量蛋白质组技术规程》团体标准征求意见稿和团体标准编制说明。

4.2 征求意见阶段

2024年2月，本标准由中国国际科促会标准化工作委员会，在全国团体标准信息平台面向社会进行公开征求意见。同时由编制组向相关单位进行定向征求意见。

4.3 审查阶段

4.4 报批阶段

二、标准编制原则、主要内容及其确定依据，修订标准时，还包括修订前后技术内容的对比

1、标准的编写原则

本标准的制定工作遵循“先进性、适用性、可操作性、实用性、规范性”的原则，依据 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》和 GB/T 20001.4-2015《标准编写规则 第 4 部分：试验方法标准》给出的规则编写。

2、标准主要内容

1）样品量要求；

2）样品采集与保存要求；

3）样品前处理要求；

4）各种裂解液主要特点、差异和选择建议；

5）样品运输要求；

6）SDS-PAGE 质量控制结果判定及解决方法；

7）非标记定量蛋白质组蛋白质酶解实验流程；

8）iTRAQ 标记定量蛋白质组蛋白质酶解及标记实验流程（以 8 标为例）；

9）多肽样品除盐实验流程；

10）质谱数据采集实验流程；

11）质谱数据质量控制要求。

3、制定本标准的依据

3.1 蛋白质组标准流程可极大提高大肠杆菌蛋白组的重复性

本项目前期已初步建立蛋白质组的制备分析全流程，并已研发获得极为稳定的蛋白质标准品，据此，我们利用前期研究成果进一步验证流程和标准品的稳健性。细菌中半数以上的蛋白质编码基因是由两个或两个以上的基因组成的多顺反子操纵子所组成。操纵子组织是否维持操纵子内基因的化学计量表达仍有争议。在本研究中，我们进行了一项无标记、与数据无关的采集超反应监测质谱(HRM-MS)实验，以定量指数期的大肠杆菌蛋白质组，并定量了 93.6%的胞质蛋白，分别涵盖了 BW25113 和 MG1655 菌株中 67.9%和 56.0%的翻译多顺反子操纵子。我们发现翻译调控在很大程度上有助于蛋白质组的复杂性:对于化学计量，较短的操纵子往往比较长的操纵子受到更严格的控制；主要编码复合物的操纵子比主要编码代谢途径的操纵子更严格地控制化学计量。基因间隔(一个操纵子中相邻基因之间的距离)可以作为化学计量的调节因子。催化效率可能是一个操纵子编码的酶的差异表达的驱动力。这些结果说明了操纵子调控的多面性:操纵子统一转录水平和基因特异性翻译水平。这种多水平调控通过优化代谢途径的生产力效率和维持不同类型的蛋白质复合物而有益于宿主。

为了评估 HRM-MS 方法的定量能力和再现性，我们应用自研标准品对质谱进行标准测

定及校准，使用上述已建立的蛋白质组标准流程对菌株 BW25113 和 MG1655 的大肠杆菌总可溶性蛋白进行了两次生物学重复实验。当使用先前的鉴定标准(至少一种独特肽，肽长度≥6 个氨基酸)时，我们的 HRM-MS 结果分别定量了这两种菌株中的 1951 和 1571 个蛋白质。两次生物复制鉴定了几乎相同的蛋白质，证明了高稳健性和再现性(图 1A)。两个重复对几乎相同的蛋白质进行了定量:仅在一个重复中对少数蛋白质进行了定量(图 1A)。在严格标准下为两个独特的肽和至少七个氨基酸的肽长度。即使在严格标准下，我们仍以高再现性定量了 1675 和 1252 个蛋白质(图 1B、C)。与之前通过其他方法定量的结果(APEX 为 1021 个蛋白质，iBAQ 为 1183 个蛋白质，emPAI 为 1138 个蛋白质)相比，定量的可溶性蛋白质数量几乎增加了一倍。

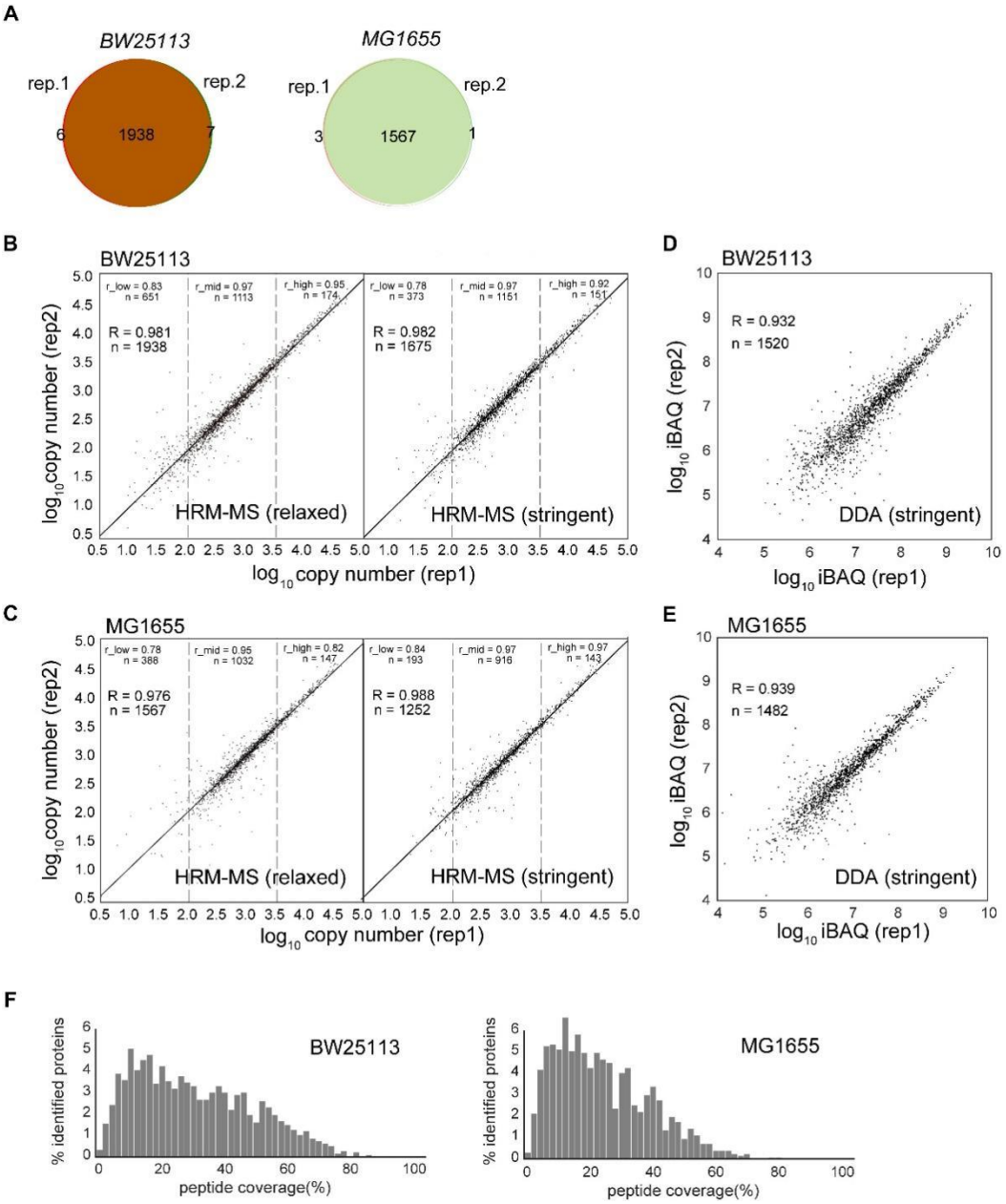

图 1.不同非标记定量方法获得的蛋白质丰度比较。

为了排除仪器的差异，我们在同一台 Orbitrap Fusion Lumos 仪器中对两种菌株进行了 DDA MS 实验。根据严格标准鉴定蛋白质，并使用 iBAQ 法进行定量。DDA MS 分别定量了 BW25113 和 MG1655 菌株的 1520 和 1482 个蛋白质，与 HRM-MS 实验相当。然而，两个生物学复制品的 iBAQ 定量的相关系数分别为 0.932 和 0.939(图 1D，E)，低于 HRM-MS(两个菌株的 R 分别为 0.982 和 0.988)。两个菌株中每个鉴定蛋白平均被 19.20 和 15.45 个肽覆盖，分别覆盖 31.56%和 24.87%的氨基酸序列(图 1F)，高于人类蛋白质组质谱实验的典型肽覆

盖(单个搜索引擎，覆盖高达 20%)。

以上研究发表在：Frontiers in Genetics (2019) 10:473.

3.2 翻译组可作为指导蛋白质组质量控制的独立数据集

到目前为止，在质谱蛋白质组鉴定的时候依然有超过一半的谱图无法被鉴定。开放式搜索（Open Search）策略很早就被提出用来解决这一问题。但是 Open Search 的搜索策略的鉴定结果可靠程度一直难以评估，也缺乏一套完整系统的使用手册。我们提出使用翻译组测序数据作为最小化库对 Open Search 的鉴定结果进行质控，主要的方法是引入可疑发现率（SDR）作为参考指标对 Open Search 策略和限制性搜索（Narrow window Search）策略的鉴定结果进行评价。结果发现，Open Search 策略确实可能引入更多的假阳性，但它确实能够有效提高谱图利用率，我们还提出将肽段 FDR 控制在 0.001 以下能够有效控制 Open Search 鉴定结果假阳性发生，并且给出了一套 Open Search 较为完整的使用规范，这对于促进容差搜索策略的发展有重要意义，并且能帮助我们更好地认识那些无法被鉴定和使用的谱图，揭示更多的翻译后修饰、单氨基酸多态性以及新型剪接变体等事件。

3.2.1 实现接近完整的翻译中基因鉴定率

所有 RNC-seq 都是以非常高的通量下测序，测序通量达到 144-188M 读段。假设一个典型的哺乳动物细胞含有 200，000 个 mRNA 分子，一个平均长度为 2.2 kb 且丰度为 1 拷贝/细胞的转录物相当于 2.27 RPKM。采用每种转录物 10 次阅读的量化标准，这种通量可以量化平均长度的转录物，丰度为 0.014 拷贝/细胞，表明翻译 mRNA 鉴定接近完成。事实上，我们发现 RNC-seq-量化的低丰度翻译 mRNAs 低至 0.001 RPKM(图 2A)。此外，在此处理量下，量化的基因数量接近饱和(图 2B)，再次表明翻译 mRNAs 的鉴定接近完整。

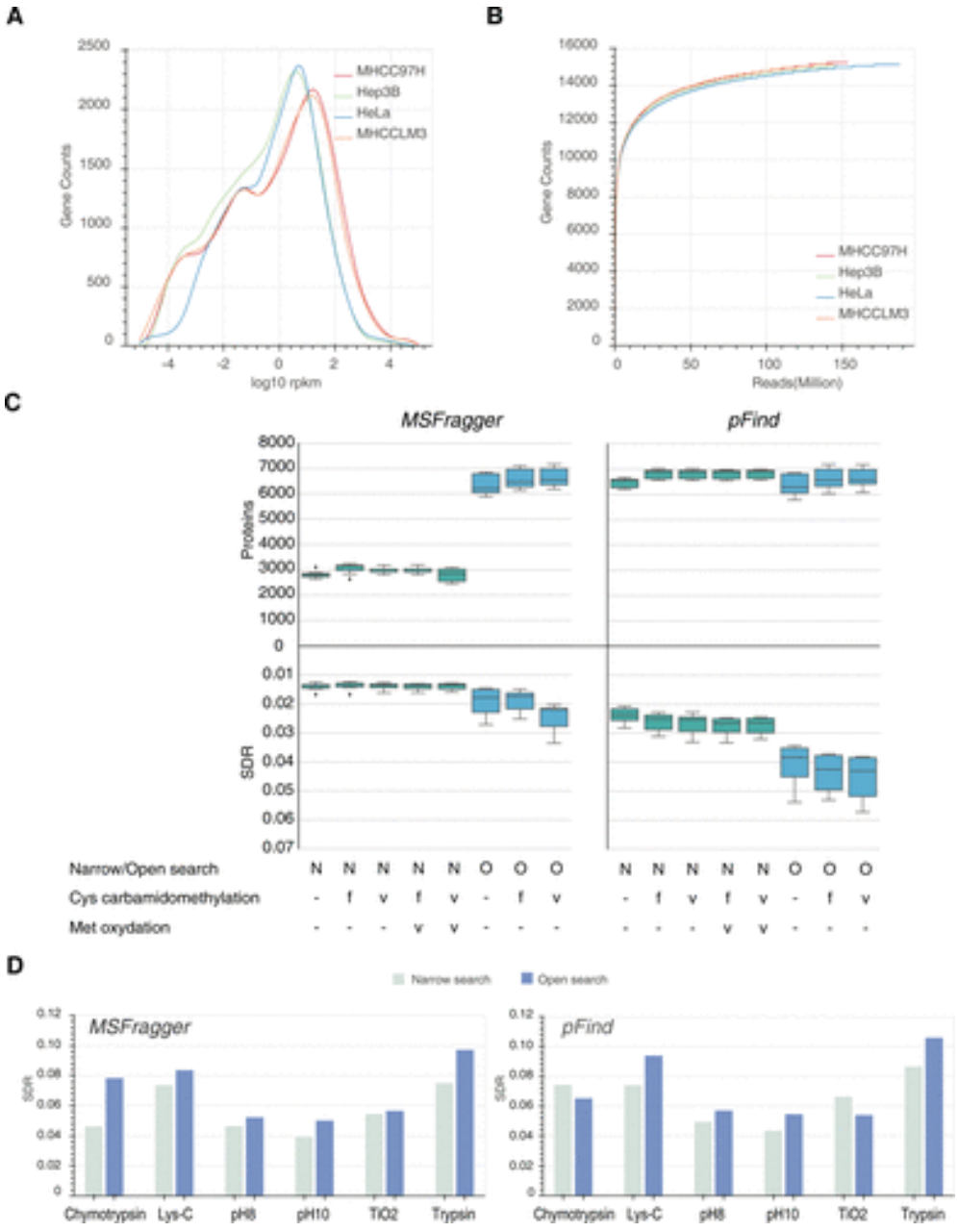

图 2. 不同设置下限制性搜索和开放搜索策略的 SDR 评估。

3.2.2 利用质量容差和多肽 FDR 可以控制开放搜索 SDR

不管是 Narrow Search 策略还是 Open Search 策略，在进行数据库搜索之前都会对氨基酸修饰进行预先的设置，包括固定修饰或者可变修饰。在通常情况下，氨基酸的修饰千变万化，无法准确地清楚所有的氨基酸修饰的具体情况，那么较为经典的做法是对氨基酸进行占比例较大的一些修饰进行预先的设置，例如半胱氨酸的烷基化修饰和蛋氨酸的氧化修饰以及肽段的 N 端乙酰化修饰等。理论上，增加可变修饰的种类将使得搜空间扩大，不

仅需要更多的计算也会导致鉴定结果的准确性的下降，因此，我们在这里只进行两种占比例较大的氨基酸修饰进行预先的设置，分别为半胱氨酸的烷基化修饰和蛋氨酸的氧化修饰。通过不同的修饰设置的组合对相同数据集进行 Narrow window Search（Precursor tolerance=10 ppm），我们发现对于一般的搜库软件如 pFind，增加可变修饰会使得鉴定结果的 SDR 提高，不设置任何修饰信息不会使得 SDR 增加（图 2C）。

开放搜索策略的关键是允许更大的母离子质量容差，以耐受由修饰引起的质量偏移。然而，这也将导致允许更多假阳性的副作用。允许较小的母离子容差在理论上应该会降低 SDR。当允许 100 - 800 Da 母离子质量容限时，我们验证了两种算法的这一假设:容限越低，SDR 越低(图 3A)。我们还注意到，当质量容差超过 400 Da 时，SDR 几乎饱和。相比之下，识别能力几乎与窗口大小无关(图 3B)。这些结果表明，当大部分预期修改已经包括在内时，较大的质量公差是没有用的。事实上，300 Da 的耐受性已经涵盖了 90%的预期修改，500 Da 的耐受性涵盖了 94%。38 - 70%的质量位移为+1 至+3，表明质量位移的主要来源是同位素偏移。

肽 FDR 过滤是用于控制鉴别质量的另一个因素。由于大的前体耐受性可能导致肽鉴定的较高误差，因此更严格的肽 FDR 应有助于控制质量。事实上，我们发现设置更严格的肽 FDR 标准显著降低了两种算法的 SDR(图 3C 和 D)。当将肽 FDR 设置为 0.001 时，无论将半胱氨酸氨基甲酸乙酯化设置为可变修饰还是固定修饰，均可将开放搜索 SDR 平均控制在窄窗口搜索水平(图 3C 和 D)。值得注意的是，算法之间存在差异:当将肽 FDR 设置为 0.002 时，pFind 开放搜索 SDR 降至窄搜索 SDR 以下，而 MSFragger 需要设置为 0.001。

两个搜索引擎获取的身份信息的一致性反映了身份识别质量的提高。在所有数据集中，肽 FDR < 0.001 时的一致性鉴定分数高于肽 FDR < 0.01 时的一致性鉴定分数(图 3E)。

以上研究发表在：Journal of Proteome Research (2018) 17(11), 3719-3729.

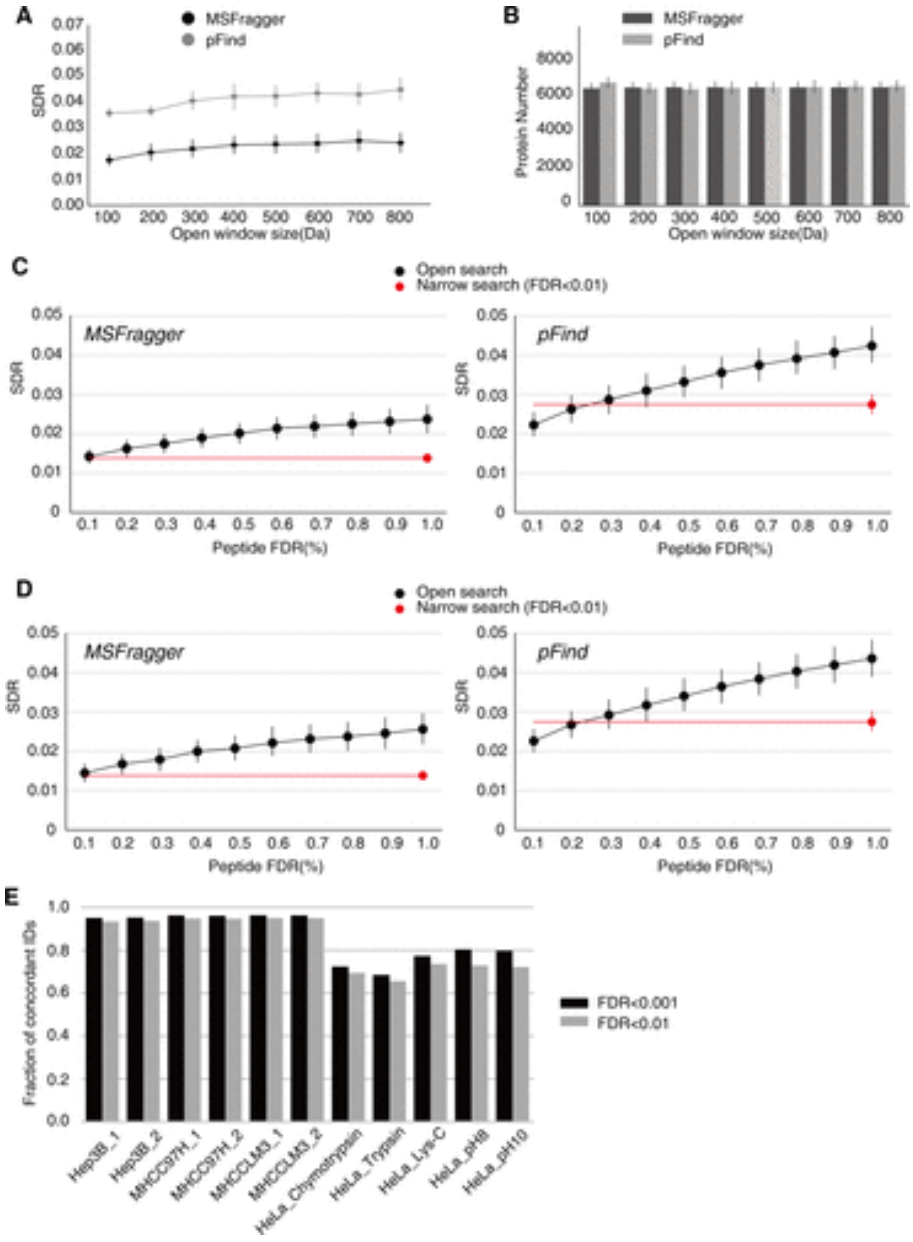

图 3. 使用不同因素控制 SDR。

3.3 利用翻译组发现隐藏的非编码基因来源蛋白质组

人类基因组上已知大约有 5 万个基因，其中约 2 万个被标注为可以表达蛋白质的“编码基因”，而另外 3 万个基因被标注为“非编码基因”(non-coding genes)。已有的报道中，除了部分非编码基因可以表达为小肽行使调控功能外，也有个别 lncRNA 被发现实际上能翻译成>50 氨基酸的蛋白质，例如 CLUU1，ESRG 等，问题是，如果这种情况不是个案而是普遍存在的现象，则确实存在部分“编码基因”被错误地标注成了“非编码基

因”，这将意味着人基因组需要被系统性地重新注释。

事实上，这一问题很早就被学界关注过。2013 年，Eric Lander 等人仅仅根据 Ribosome profiling 计算模型认为 lncRNA 不可能编码蛋白质。但仅仅一年后，该团队又发表文章，仅仅是调整了计算模型，认为 lncRNA 可能编码蛋白质，然而其始终拿不出蛋白质实验证据。2014 年，人类蛋白质组草图在 Nature 上发表，声称发现千余个 lncRNA 所编码的“新蛋白质”，但随后便被人类蛋白质组组织(HUPO)爆出其分析不合规范，存在大量的假阳性鉴定，在用较严格的标准进行质控后，这些所谓的“新蛋白质证据”几乎都被认定为假阳性。因此，如何避免蛋白质组学质谱技术的固有缺陷，提供独立的蛋白质编码信息源，就显得非常重要。在本项目中使用翻译组测序技术（RNC-seq），在肺癌细胞中发现了 1397 个有可能被翻译的“非编码 RNA”[16]。从 9 株人细胞系中共鉴定到约 4700 种 lncRNA 正在被翻译，且可能以经典翻译起始方式翻译出>50 氨基酸的蛋白质。利用目前公认的验证标准，我们提供了其中 314 个新蛋白质的证据。这些蛋白质是稳定存在的，并且有着明确的细胞定位，功能实验也证实它们以蛋白质形式（而非 RNA 形式）行使着明确的生物学功能。

3.3.1 人类细胞中含有大量正在翻译的长链非编码 RNA

我们先前已对人肺 HBE 细胞、A549 和 H1299 细胞、人结直肠癌 Caco-2 细胞、人肝癌 MHCCLM3、MHCC97H 和 Hep3B 细胞和人宫颈癌 HeLa 细胞中的全长翻译 mRNA 进行了测序。在此，我们进一步对 MHCCLM6 细胞进行了 mRNA 测序和全长翻译 mRNA-seq (RNC-seq)分析。总之，我们发现 1028-3330 lncRNA 与 9 个受试人类细胞系中的核糖体结合。其中，2969 个翻译型 lncRNAs 具有至少一个以 AUG 起始的规范开放阅读框(ORF)，可编码长度至少为 50 aa 的蛋白质(图 4A)。这意味着一个前所未有的显著的隐藏的人类蛋白质组。

我们的核糖体分析结果证实，这些翻译型 lncRNAs 由核糖体主动翻译，因为核糖体足迹跨越转录本，尤其是跨越预测的 CDS 区域(图 4B)。翻译的 lncRNAs 分布在整个人类基因

组的所有染色体中，几乎未检测到染色体富集(图 4C)，表明这些潜在的新编码基因在整个人类基因组中普遍存在。同时，在几乎所有 9 个受试细胞系中检测到约 1/3 的这些翻译 lncRNAs(≥0.1 RPKM 和≥10 次读取)，其余的表现出显著的细胞特异性表达(图 4D)，这在我们使用 RPKM = 1.0 作为阈值时也观察到(图 4E)。这些结果表明这些 lncRNAs 的翻译受到普遍调控。

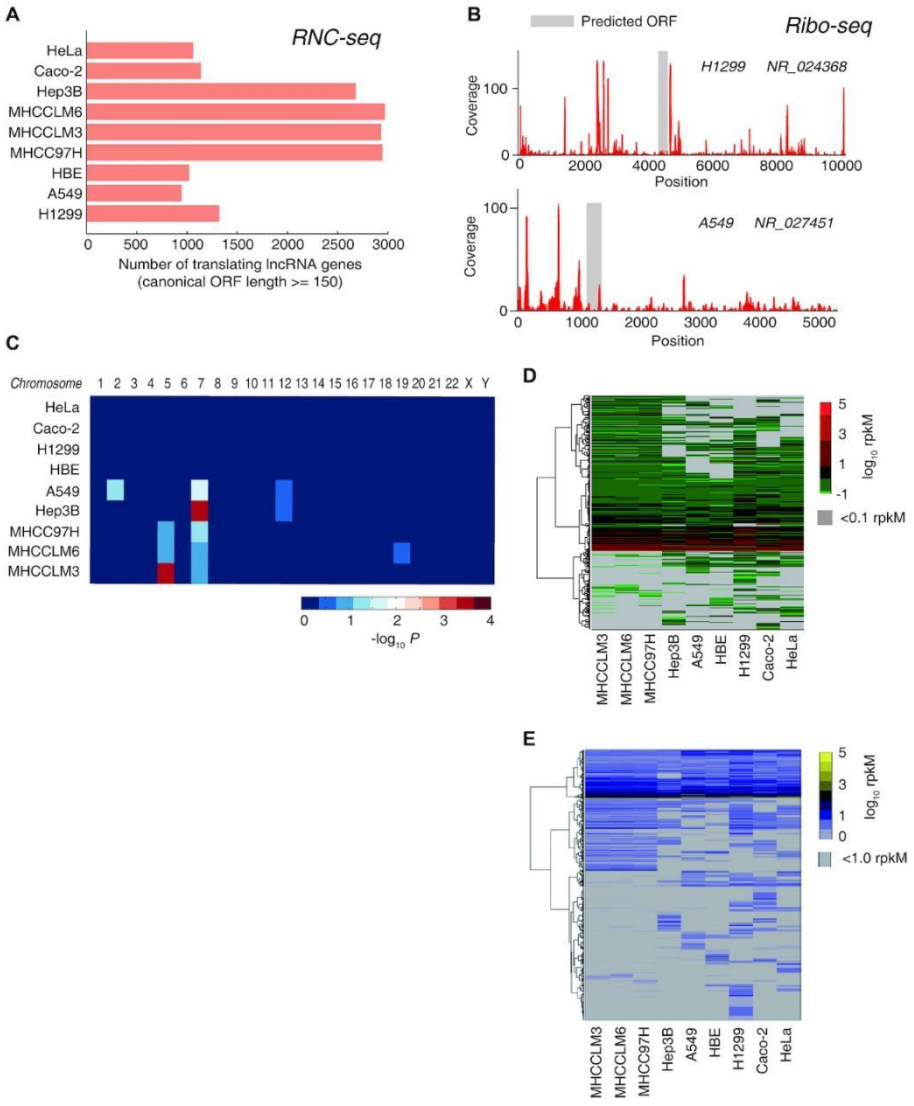

图 4. 翻译中的 lncRNA 可以编码蛋白质。

3.3.2 lncRNA 编码蛋白质组的蛋白质证据

接下来，我们旨在为这些翻译的 lncRNAs 编码的潜在新蛋白找到质谱证据。我们注意到，新蛋白的长度比已知蛋白短得多，即由人类蛋白质组项目(HPP)提出的蛋白证据水平 1

(PE1)蛋白(图 5A)。因此，我们对低分子量蛋白质(< 25 kDa)进行了 MS 分析，以补充对总蛋白质的鸟枪法蛋白质组学分析。我们认为根据以下标准进行鉴定是有把握的:(I)蛋白质水平 FDR <1%，(ii)肽长度至少 9 aa，(iii)如前所述，根据 Smith-Waterman 分析，至少一种独特的肽具有不超过两种可能的 aa 错配。因此，我们通过 MS 分析分别从总(216 种蛋白质)和低分子量(109 种蛋白质)蛋白质组分中鉴定出与 lncRNAs 编码的 308 种新蛋白质相对应的 372 种独特肽(图 5B)，只有 17 种重叠。308 个新蛋白的所有蛋白识别信息，包括蛋白名称、唯一肽序列、肽长、搜索引擎，均可在 S2 增补表中找到。可通过 iProX 数据库(注册号:IPX00076300)获取原始数据文件、参考数据库和搜索结果文件。

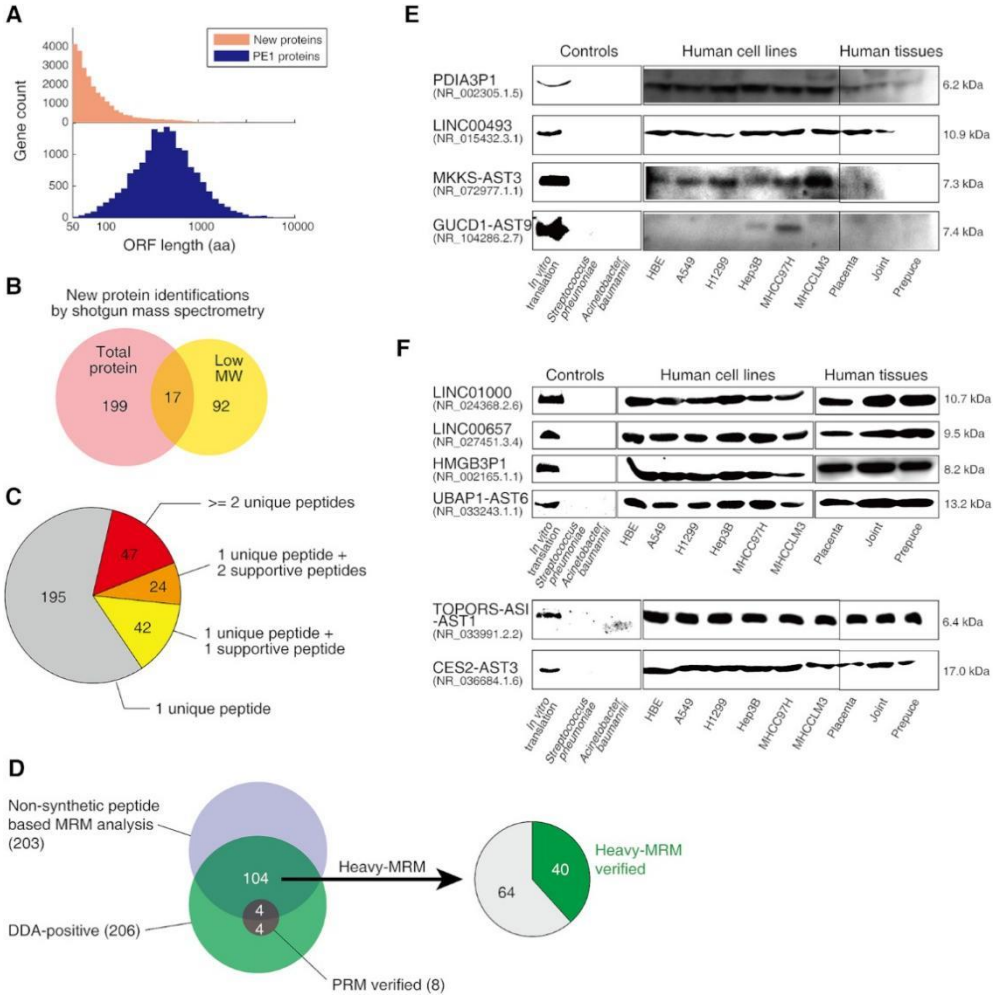

图 5.翻译中 lncRNA 的蛋白质证据。

我们发现，超过 36%的 MS 检测到的新蛋白由至少两种独立肽段(47 种蛋白质)或一种

独立肽段加上至少一种支持肽段(66 种蛋白质)证明。这里，支持肽段是由数据库搜索引擎分配给某个蛋白质鉴定的非唯一肽。此外，在 308 个 MS 鉴定的新蛋白中，有 261 个仅由一种独立肽段证明(图 5C)，这可以部分解释为翻译-mRNA 分析中大多数新蛋白的表达水平普遍较低(< 10 RPKM)(图 4D 和 E)以及长度较短(图 5A)。

我们参考了 HPP 指南 2.1 标准来表述特殊检测结果，该标准要求使用独立于数据的采集方法，包括多反应监测(MRM)和平行反应监测(PRM)。通常，MRM/PRM 用于选择某些母肽进行片段化和定量。可通过添加合成肽作为样品中这些肽的进一步定量的加样标准来辅助这一操作。对于常规 MRM(不含合成肽)，我们发现 184 种新蛋白有 203 种独特肽的证据(图 5D)，分别占所有已鉴定独特肽和新蛋白的 54.6%和 59.7%。

对于基于合成肽的 PRM 分析，在 372 个新的蛋白质编码肽中，我们合成了 313 个技术上允许的重标记标准肽。我们可以通过数据相关采集(DDA)分析确定 313 项标准中的 206 项(图 5D)。通过对这 206 种肽的 PRM 分析，我们只能验证 8 种肽的存在(图 5D)，表明需要优化 MRM。在 DDA 和常规 MRM 中均检测到 104 种重肽(图 2D)。由此，通过对重链 MRM 分析的逐一优化，我们可以进一步验证细胞裂解液中存在来自 38 种蛋白质的 40 种肽(图 5D)。这些肽的峰组的相对强度和洗脱时间均与合成参考肽匹配，表明根据 HPP 指南 2.1 标准对其存在进行了有把握的验证。

接下来，我们使用含有抗原表位的特异性肽制备了用于鸟枪法 MS 检测的四种新蛋白的多克隆抗体。免疫印迹结果显示，在各种人细胞系和人组织中，所有这些蛋白在预期分子量下均有清晰特异性条带(图 5E)，表明它们在体内以全长和稳定形式存在。为了进一步确保抗体的特异性，我们使用体外翻译的全长蛋白作为阳性对照，从细菌鲍曼不动杆菌和肺炎链球菌中提取的总蛋白作为阴性对照(图 5E)。此外，我们为在翻译 mRNA 分析中发现但未被 MS 检测到的另外 6 种新蛋白制备了多克隆抗体，并获得了它们可靠的免疫印迹证据(图 5F)。这些结果表明，由于 MS 的局限性，它只能检测到一部分新蛋白，并提示隐藏