DOI: 10.13376/j.cbls/2021025

文章编号：1004-0374(2021)02-0223-08

基于蛋白质基因组学发现的新编码序列在疾病中的研究进展

蒙书红[1,2]，张　瑶[2*]，徐　平[2*]

(1 河北大学生命科学学院，保定 071002；2 军事科学院军事医学研究院生命组学研究所，国家蛋白质科学中心·北京，北京蛋白质组研究中心，蛋白质组学国家重点实验室，中国医学科学院蛋白质组学与新药研发新技术创新单元，北京 102206)

摘　要：非编码序列在不同物种全基因组中占了很大的比例，在人类基因组中高达98%以上。近年来，越来越多的研究发现，新编码序列不仅参与调控发育、分化、增殖与凋亡等生命活动，而且在疾病的发生发展等过程中也扮演了重要的角色。现对基于蛋白质基因组学发现的新编码序列在疾病中的鉴定及应用进行综述，有望为疾病预测、分型、诊断、治疗、疗效判断及预后的新型生物标志物和潜在治疗靶点研究提供丰富源泉。

关键词：非编码序列；蛋白质基因组学；新编码序列；疾病

中图分类号：Q503；Q51　　**文献标志码**：A

Progress on novel coding sequences by proteogenomics in diseases research

MENG Shu-Hong[1,2], ZHANG Yao[2*], XU Ping[2*]

(1 College of Life Sciences, Hebei University, Baoding 071002, China; 2 Research Unit of Proteomics & Research and Development of New Drug of Chinese Academy of Medical Sciences, State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences·Beijing, Institute of Lifeomics, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 102206, China)

Abstract: Non-coding sequences account for a large proportion in whole genome of different species, especially in human genome which is up to 98%. In recent years, an increasing number of studies have shown that novel coding sequences not only participate in life activities, such as the regulation of development, differentiation, proliferation, and apoptosis, and so on, but also play a significant role in occurrence and development of diseases. This paper reviewed the discovery and application of novel coding sequences based on proteogenomics in various diseases, which will provide rich resources for researching novel biomarkers and potential therapeutic targets of disease prediction, classification, diagnosis, treatment, curative effect judgment and prognosis.

Key words: non-coding sequences; proteogenomics; novel coding sequences; diseases

非编码序列是指基因组中不转录为RNA或转录后不翻译为蛋白质的序列。不同物种的基因组包含的编码序列比例差异比较大，且随着进化由低等到高等，由简单到复杂，编码序列占比越来越小，而非编码序列占比却越来越高，如大肠杆菌(*Escherichia coli*)中非编码序列约11% [1]，酿酒酵母(*Saccharomyces cerevisiae*)中约29%，拟南芥中约71% [2]。2001年，人类基因组测序发现人类蛋白质编码序列仅占全基因组的1.1%~1.4% [3]，非编码序列高达98%以上。目前对编码序列的生物功能研究已取得较大的进展，但对非编码序列的生物功能所知甚少。近10多

收稿日期：2020-06-24；**修回日期**：2020-08-11

基金项目：国家自然科学基金项目(31901037)；“艾滋病和病毒性肝炎等重大传染病防治”科技重大专项(2018ZX10302302001003)

***通信作者**：E-mail: xupingghy@gmail.com (徐平)；zhangyaowsw@163.com (张瑶)

年来，二代测序技术以及质谱技术的发展促进了基因组学、转录组学和蛋白质组学的快速发展，发现部分非编码序列含有特殊的功能性元件，在生命活动中也发挥了重要的生物学功能。Jaffe等[4]在肺炎支原体(*Mycoplasma pneumoniae*)的研究中首次提出了蛋白质基因组学(proteogenomics)的概念。随后，蛋白质基因组学技术逐渐被广泛地应用于物种全基因组重注释工作中。近5年应用于疾病相关研究中，为人类重新认识非编码序列的功能、机理、遗传、进化等提供了契机。

1 非编码序列的种类及功能

非编码序列可分为非编码DNA (non-coding DNA, ncDNA)和非编码RNA (non-coding RNA, ncRNA)。ncDNA至少包含顺式调控元件、内含子及可转录为ncRNA的DNA序列。顺式调控元件，如启动子、增强子、沉默子、绝缘子等；ncRNA大体上有：rRNA (ribosomal RNA)、tRNA (transfer RNA)、snoRNA (small nucleolar RNA)、snRNA (small nuclear RNA)、lncRNA (long non-coding RNA)、miRNA (microRNA)、siRNA (small interfering RNA)、circRNA (circular RNA)、piRNA (piwi-interacting RNA)、5′-UTR (untranslated regions)、3′-UTR等。LncRNA [5]长度大于200 nt，主要分为反义长链非编码RNA (antisense long non-coding RNA)、长内含子(intronic)、lincRNA (long intergenic non-coding RNA)、Pseudogene、eRNA (enhancer RNA) 等。

1.1 ncDNA的功能

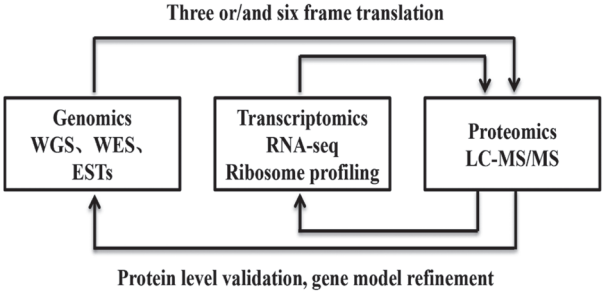
ncDNA涉及大量的DNA剪切和剪接、转座子重组、基因重排、基因稳定、基因调控、ncRNA的形成等。Frokjaer-Jensen等[6]发现了普遍存在于秀丽隐杆线虫(*Caenorhabditis elegans*)生殖细胞中以10 bp的周期性A_n/T_n簇为特征的ncDNA序列。当外源基因导入生殖细胞时，该序列可防止内源基因被随机沉默；而当含有该序列的合成基因导入生殖细胞时，也可防止细胞识别而被沉默。Zhang等[7]发现了近200个与癌症相关的可使基因表达异常的非编码序列突变，其中一个非编码突变能激活*DAAM1*基因，使得肿瘤细胞更具侵袭性。Morton等[8]重点分析了致癌基因表皮生长因子受体(epidermal growth factor receptor, EGFR)基因，发现*EGFR*在胶质母细胞瘤中大量拷贝时倾向于成环状DNA的形式，且与染色体分离独立存在。环状DNA中包含几千碱基的ncDNA，含数10种增强子以及其他调控元件。这些额外的ncDNA使得*EGFR*异常急剧扩增，表明染色体外ncDNA可促进癌症的发生发展。Takeda等[9]发现，早期阶段的前列腺癌雄激素受体(androgen receptor, AR)基因上游的增强子被沉默，而在治疗抵抗性转移性前列腺癌中，*AR*上游增强子的乙酰化被激活，使得*AR*和增强子大量复制，揭示ncDNA导致晚期前列腺癌治疗抵抗性的分子机制。

1.2 ncDNA的功能

非编码序列的双链RNA干扰现象于2006年被授予诺贝尔生理学或医学奖。随后大量研究表明，ncRNA通过与DNA、RNA、蛋白质结合或编码小肽参与胚胎发育、干细胞维持、细胞增殖与分化、代谢、信号转导、免疫应答、衰老、凋亡、疾病的发生发展等几乎所有生理或病理过程的基因表达调控，发挥重要生物学功能。ncRNA在老年性痴呆症、糖尿病、心肌肥大、艾滋病、肺炎、白血病、皮肤病、胸腺癌、肺癌、直肠癌、皮肤癌等疾病的调控中发挥了重要作用，意味着ncRNA可能在药物研发、疾病诊断、治疗等方面具有潜在的研究及应用价值。

2 蛋白质基因组学在发现非编码序列编码蛋白质中的作用

近年来，随着高通量测序技术的成熟，全球已步入了大数据时代，积累了大量的基因组、转录组和蛋白质组学数据。蛋白质是生命活动功能的直接承担者。基于蛋白质组大数据集的多维组学数据分析策略可为疾病发生、发展调控新机制的揭示提供新见解(图1) [10]。在这个数据分析流程中，三框或六框蛋白质翻译数据库的深度解析，不仅可比较正常模型组与病理组样本已知编码蛋白质的差异表



WGS (whole genome sequencing)：全基因组测序；WES (whole exome sequencing)：全外显子测序；ESTs (expressed sequence tags)：表达序列标签。

图1 蛋白质基因组学研究流程图[10]

达，而且还可发现人类原先所认识的非编码区序列的编码能力及其在疾病发生、发展及预后过程中的调控规律，为人们全面认识疾病类型及过程提供了很好的技术支撑。

3　蛋白质基因组学在基因组重注释研究中的应用

基于常规算法对全基因组测序结果进行注释，往往会因人类对基因结构认识的不足而将特殊基因错误归类为非编码序列，导致漏注释。如非经典开放阅读框[11-12]的发现提示起始密码子并非局限于ATG，还包括CTG、GTG、TTG、ACG、ATT等非经典起始密码子；因可能的开放读码框太短也会在注释中遗漏很多小蛋白质和多肽。

2004年至今，蛋白质基因组学技术一直广泛运用于基因组重注释、已注释编码基因的验证，已注释编码基因边界的校正、外显子边界的校正、新外显子和新可变剪接体的发现等。其应用包括完善多种微生物[13-17]、原生生物[18]、昆虫[19-21]、拟南芥[22-23]、毛竹[24]、水稻[25]、斑马鱼[26]、老鼠[27-28]，甚至人类[29]的注释基因组。随着质谱技术的成熟和发展，基于蛋白质基因组学技术发现的许多非编码序列中有很多小的开放读码框 ，可翻译成小肽或小蛋白质，见表1 [30-39]。

4　蛋白质基因组学在疾病发生发展分子机制研究中的应用

近年来，蛋白质基因组学逐渐应用于疾病发生发展的分子机制研究中，已在肿瘤发生、发展、分型分子标志物、肿瘤抗原研发和精准医疗中得到了越来越多的应用[40-42]，为这些疾病的分子机制研究提供了新的视角，补充了人类对非编码序列在疾病中发生发展的新认识。此外，蛋白质基因组学也在心血管疾病、传染性疾病、炎症、药效评价、新药开发等多领域得到了应用[43]。

4.1　蛋白质基因组学发现的新编码序列在疾病发生发展中的调控作用

Dou等[44]对95个患者的子宫内膜癌样本进行了包括DNA、RNA、蛋白质及其翻译后修饰组等多组学研究，发现circRNA在上皮间质转化中具有调控作用。Hosono 等[45]发现新型超保守*THOR* (testis-associated highly-conserved oncogenic long non-coding RNA)在睾丸组织和黑色素瘤、肺癌等肿瘤组织中显著高表达，影响减数分裂和促进肿瘤的发生。*THOR* RNA pulldown样品质谱鉴定到*THOR*通过直接增强IGF2BP (insulin-like growth factor 2 mRNA-binding protein)与下游基因mRNA的结合提高目的基因mRNA的稳定性，从而上调其表达水平。Trembinski等[46]发现，保守ncRNA *Sarrah*可抗细胞衰老、凋亡，是心肌细胞存活的调节剂，可用于治疗心血管疾病。Yang等[47]发现，*CDR1as* (cerebellar-degeneration-related protein 1 antisense RNA)在肝癌、乳腺癌、宫颈癌细胞系中表达量显著降低。该基因可通过miR-7靶向调控EGFR的信号表达，促进肝癌细胞的增殖。

4.2　蛋白质基因组学发现疾病诊断的生物标志物

生物标志物[48]是可作为正常生物学过程、病理过程或治疗干预药理学反应的指示因子，是疾病检测、分类、监测、检验临床治疗效果和预后的有力工具。蛋白质基因组学技术所鉴定的多个具有编码

表1 基于蛋白质基因组学发现的新编码序列

研究对象	发现非编码序列的类型	数量	参考文献
A431 cells	Pseudogene、Intergenic、5′-UTR、3′-UTR、Intronic、ncRNA	329	[30]
Normal human tissues	Pseudogene、Intergenic、5′-UTR、3′-UTR、Intronic、ncRNA	149	[30]
Breast tumor	Pseudogene、Intergenic、5′-UTR、3′-UTR、Intronic、ncRNA	356	[31]
Human B cells	Pseudogene、LincRNA、Annotated antisense、Unannotated intergenic region	17	[32]
Human umbilical vein endothelial cells	Annotated 5′-UTR region	7	[33]
Human cell lines	Long noncoding RNA	308	[34]
H1299 cells	5′-UTR; 3′-UTR	61	[35]
Haploid cell lines	Long noncoding RNA	2	[36]
Mycobacterium tuberculosis	Pseudogene	10	[37]
Leishmania major	Nong coding RNA	8	[38]
Yeast	Long noncoding RNA	1	[39]

能力的非编码序列具有作为疾病生物标志物的潜力，可为疾病的临床诊断、治疗、预后提供潜在靶点，见表2 [49-54]。

表2 基于蛋白质基因组学发现的候选生物标志物

非编码序列	疾病类型	功能	参考文献
HOXB-AS3	结肠癌	抑癌	[49]
circPPP1R12A	结肠癌	促癌	[50]
circ-FBXW7	胶质母细胞瘤	抑癌	[51]
LINC-PINT	恶性胶质瘤	抑癌	[52]
LINC00961	多种癌	抑癌	[53]
LINC00266-1	多种癌	促癌	[54]

4.2.1 *HOXB-AS3*

Huang等[49]发现，lncRNA *HOXB-AS3* (HOXB cluster antisense RNA3)可编码长度为53个氨基酸的保守小肽。在高度迁移的结肠癌细胞中，*HOXB-AS3*显著下调；沉默*HOXB-AS3*会促进结肠癌的形成、生长和侵袭转移等，这可能是由于该基因编码的小肽通过竞争性结合hnRNP A1，阻断hnRNP A1所介导的丙酮酸激酶PKM (pyruvate kinase M)的剪接能力，从而减少PKM2剪切体的形成，抑制结肠癌细胞中糖代谢重编程以及结肠癌细胞系生长。同时，HOXB-AS3还是内源性抗癌肽，其表达水平低的结肠癌患者预后较差。

4.2.2 *circPPP1R12A*

Zheng等[50]通过circRNA基因芯片筛选了10例人类结肠癌组织和癌旁组织中circRNA的表达谱，发现*circPPP1R12A*表达上调最显著。*circPPP1R12A*高表达的患者总生存期较短。Western 印迹和LC-MS/MS证明*circPPP1R12A*可编码由73个氨基酸组成的蛋白质。体内外实验发现该蛋白质通过激活Hippo-YAP信号通路促进结肠癌细胞的生长和转移。

4.2.3 *circ-FBXW7*

Yang等[51]对10个胶质母细胞瘤样本进行深度测序，发现 *circ-FBXW7*在正常人脑中大量表达，但在癌组织中显著低表达。进一步分析发现，*circ-FBXW7*存在内部核糖体进入位点，由此编码185个氨基酸组成的蛋白质。体外和动物实验证明circ-FBXW7可下调*c-Myc*的表达，稳定FBXW7，可引起细胞周期阻滞，抑制癌细胞生长、增殖。

4.2.4 *LINC-PINT*

Zhang等[52]对人类正常和胶质母细胞瘤的细胞进行高通量测序，结合免疫沉淀和LC-MS/MS分析，发现转录自*LINC-PINT* 2号外显子的*circPINT*可编码由87个氨基酸组成的PINT87aa。该蛋白质可通过与PAF1蛋白复合物结合，影响下游基因mRNA的延长，从而抑制胶质瘤的生长。

4.2.5 *LINC00961*

Matsumoto等[53]整合转录组学、翻译组学、蛋白质组学证明*LINC00961*可翻译90个氨基酸的新型多肽SPAR (small regulatory polypeptide of amino acid response)。该多肽在人和小鼠中保守，在一些组织如心脏、肺和骨骼肌肉中特异高表达。通过定位于溶酶体上，SPAR与溶酶体V型ATP酶相互作用，抑制mTORC1 (the mechanistic target of rapamycin complex)的活性，特异性阻断mTORC1感应氨基酸刺激的能力。SPAR在急性损伤的骨骼肌中下调，可降低mTORC1的活性，促进肌肉再生，表明SPAR响应损伤后以组织特异性方式精细调节mTORC1的活性，调控肌肉再生以及修复肌肉损伤。*LINC00961*在肺癌[55-56]、神经胶质瘤[57]、肾细胞癌[58]、肝癌[59]、口腔鳞状细胞癌[60-61]、黑色素瘤[62]、冠心病[63]、结肠癌[64]中均具有抑癌作用，可以诱导癌细胞凋亡，抑制癌细胞增殖、迁移及侵袭。

4.2.6 *LINC00266-1*

Zhu等[54]发现lncRNA *LINC00266-1*编码71个氨基酸的天然内源性多肽RBRP (RNA-binding regulatory peptide)，广泛分布于多种细胞和组织，并在结直肠癌、乳腺癌、卵巢癌和鼻咽癌等多种癌组织细胞中高表达，且表达量的高低与癌症患者恶性程度高低成正相关。在分子机制上，RBRP通过与*IGF2BP1*结合，识别如*c-Myc* mRNA的m^6A (N^6-methyladenosine)，同时募集RNA，稳定分子所结合的靶基因mRNA，以维持*c-Myc* mRNA的稳定表达，促使肿瘤的发生发展。因此，RBRP可能是新的肿瘤标志物和潜在的抗癌药物靶标。

4.3 蛋白质基因组学发现的新编码产物在微生物感染免疫中的调控作用

Jackson等[65]报道，核糖体与之前被注释为非蛋白质编码RNA相关联，发现了一系列短的、非ATG起始的ORF编码的蛋白质。利用Western 印迹和质谱蛋白质组学技术证明lncRNA *Aw112010*在应对细菌感染的先天免疫应答中可以编码蛋白质。该蛋白质在细菌感染和结肠炎期间协调黏膜免疫中必不可少，论证了非经典 ORFs翻译的蛋白质对宿主防御和炎症性疾病起关键免疫作用。

传染性肠胃炎病毒(transmissible gastroenteritis virus, TGEV)可引起二周龄幼猪患胃肠炎，具有高

度传染性和致命性。Song等[66]发现猪感染TGEV后，miRNA表达谱改变，差异上调表达的*miR-4331*通过靶向CDCA7 (cell division cycle-associated protein 7)间接抑制TGEV亚基因7的转录。随后，LC-MS/MS、RT-PCR、Western 印迹的结果表明，随着*miR-4331*含量的增高，线粒体蛋白IL1RAP (interleukin-1 receptor accessory protein)和RELA (v-rel reticuloen-dotheliosis viral oncogene homolog A)的蛋白质丰度增大，揭示了*miR-4331*通过抑制靶基因*RB1* (retinoblastoma 1)表达，上调IL1RAP并激活p38 MAPK通路，而调控TGEV诱导的线粒体损伤[67]。

4.4　蛋白质基因组学发现的新编码产物作为癌症特异性抗原在肿瘤特异性免疫治疗中的应用

癌细胞存在大量突变，这与正常细胞有诸多不同，有可能在癌细胞表面形成特异性抗原，从而激活免疫系统，攻击癌细胞。以往的肿瘤抗原研究重点在已知的基因组编码序列上，随着蛋白质基因组学技术的发展，陆续有研究发现许多所谓的非编码序列也能够编码和产生大量肿瘤特异性抗原[68-69]。

Johansson等[31]依据PAM50分类，分别选取5个亚型，对每个亚型的9个乳腺组织样本进行多层次系统性检测，鉴定到了对应于基因组非编码区的数百个多肽，其中有30%预测可以与主要组织相容性复合体(major histocompatibility complex, MHC) I类分子结合。通过合成肽质谱验证了61种，靶向技术证明只存在于肿瘤组织而非正常组织，可以作为潜在的新型肿瘤特异性免疫治疗靶标。在该研究中，*lnc-AKAP14-1:3*和*lncCXorf36-3:1*扩增得到患者特异性候选免疫靶点，分别在一个肿瘤和两个肿瘤中表达量升高。

Laumont 等[32]对人类B细胞富集的MHC I类分子相关肽(MHC class I-associated peptides, MAP)进行鉴定，发现这些鉴定的蛋白质大约有10%来源于基因组非编码序列或外显子翻译框外，表明非编码区是肿瘤特异性抗原的主要来源。在2种鼠类癌细胞系(CT26和EL4)和7种人类原发性肿瘤(4种B系急性淋巴细胞白血病和3种肺癌)中发现了40种肿瘤特异性抗原，大部分来源于非突变导致的异常表达的转录物[70]。在这些转录和翻译产物中，约90%来自非编码区，具有肿瘤的共性，而且其中一些抗原对癌细胞还具有特异性。由于是非注释编码基因产物，这些肿瘤特异性抗原利用标准的外显子测序方法可能会被遗漏。这些肿瘤特异性抗原制成的肿瘤疫苗在小鼠中被证明有效，有些抗原甚至能够对小鼠产生终身保护作用。Zhao等[71]用蛋白质基因组学技术在23个高级别浆液性卵巢癌样本中检测到了103个特异性抗原，其中91种是异常表达特异性抗原，12种是突变特异性抗原(7种来源于非编码序列)。同样的样品，利用传统的突变外显子测序只能发现其中的3种，充分证明了蛋白质基因组学在鉴定这些肿瘤特异性非编码区编码蛋白中的高效性。

5　小结与展望

蛋白质基因组学为人们在基因组和转录组基础上，从蛋白质组水平全面重新认识非编码序列提供了技术支撑，在认识生物体生长、发育、遗传变异以及疾病发生发展方面具有重要意义。虽然基因组非编码序列数据庞大、序列类型复杂、分析难度较大，但是随着蛋白质基因组学技术的不断发展，越来越多的非编码序列的特征、功能和作用机制逐渐被人类认知，其在疾病发生发展中重要性也得到了进一步揭示。这为疾病的分型、治疗靶点的挖掘、疾病发生发展预期、耐药机制揭示、个性化精准医疗及疗效判断和预后均提供了新的方向，为新药开发、疫苗研发提供了新思路和新技术，并产生新理论。目前已发现的功能性非编码序列仅仅是冰山一角，对于非编码序列的认识也还只是一个起始阶段，解密非编码序列工作任重而道远，有待全球工作者的协同努力。

我们课题组经8年多的努力，建立了精准蛋白质基因组学技术体系，已成功地应用于酿酒酵母[72]、结核分枝杆菌[73] (*Mycobacterium tuberculosis*)及耻垢分枝杆菌(*Mycobacterium smegmatis*)中，且均发现了一系列数据库中尚未注释的新基因，这些新发现基因的潜在生物学功能有待我们进一步探索。此外，课题组还将蛋白质基因组学技术成功地应用于人类染色体蛋白质组计划(chromosome-centric human proteome project, C-HPP)中，自2015年从睾丸组织中共鉴定到400多个“漏检蛋白”，基于质谱技术为其找到高可信的蛋白质组学证据，为人类基因组的重新绘制及其进一步功能研究提供了基础[74-79]。

[参　考　文　献]

[1] 林昊. 大肠杆菌启动子特征参数的统计分析. 生物信息学, 2009, 7: 37-9
[2] 黄小庆, 李丹丹, 吴娟. 植物长链非编码RNA研究进展. 遗传, 2015: 344-59
[3] McPherson JD, Marra M, Hillier L, et al. A physical map

of the human genome. Nature, 2001, 409: 934-41
[4] Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics, 2004, 4: 59-77
[5] Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics, 2013, 193: 651-69
[6] Frokjaer-Jensen C, Jain N, Hansen L, et al. An abundant class of non-coding DNA can prevent stochastic gene silencing in the *C. elegans* germline. Cell, 2016, 166: 343-57
[7] Zhang W, Bojorquez-Gomez A, Velez DO, et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat Genet, 2018, 50: 613-20
[8] Morton AR, Dogan-Artun N, Faber ZJ, et al. Functional enhancers shape extrachromosomal oncogene amplifications. Cell, 2019, 179: 1330-41.e13
[9] Takeda DY, Spisak S, Seo JH, et al. A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. Cell, 2018, 174: 422-32.e13
[10] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods, 2014, 11: 1114-25
[11] Chen J, Brunner AD, Cogan JZ, et al. Pervasive functional translation of noncanonical human open reading frames. Science, 2020, 367: 1140-6
[12] Na CH, Barbhuiya MA, Kim MS, et al. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res, 2018, 28: 25-36
[13] Kelkar DS, Kumar D, Kumar P, et al. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. Mol Cell Proteomics, 2011, 10: M111. 011627
[14] Potgieter MG, Nakedi KC, Ambler JM, et al. Proteogenomic analysis of *Mycobacterium smegmatis* using high resolution mass spectrometry. Front Microbiol, 2016, 7: 427
[15] Prasad TS, Harsha HC, Keerthikumar S, et al. Proteogenomic analysis of *Candida glabrata* using high resolution mass spectrometry. J Proteome Res, 2012, 11: 247-60
[16] Nagarajha Selvan LD, Kaviyil JE, Nirujogi RS, et al. Proteogenomic analysis of pathogenic yeast *Cryptococcus neoformans* using high resolution mass spectrometry. Clin Proteomics, 2014, 11: 5
[17] Venter E, Smith RD, Payne SH. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One, 2011, 6: e27587
[18] Pawar H, Sahasrabuddhe NA, Renuse S, et al. A proteogenomic approach to map the proteome of an unsequenced pathogen —— *Leishmania donovani*. Proteomics, 2012, 12: 832-44
[19] Chaerkady R, Kelkar DS, Muthusamy B, et al. A proteogenomic analysis of *Anopheles gambiae* using high-resolution Fourier transform mass spectrometry. Genome Res, 2011, 21: 1872-81
[20] Prasad TS, Mohanty AK, Kumar M, et al. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res, 2017, 27: 133-44
[21] Ye X, Tang X, Wang X, et al. Improving silkworm genome annotation using a proteogenomics approach. J Proteome Res, 2019, 18: 3009-19
[22] Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics, 2010, 73: 2124-35
[23] Castellana NE, Payne SH, Shen Z, et al. Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc Natl Acad Sci USA, 2008, 105: 21034-8
[24] Yu X, Wang Y, Kohnen MV, et al. Large scale profiling of protein isoforms using label-free quantitative proteomics revealed the regulation of nonsense-mediated decay in moso bamboo (*Phyllostachys edulis*). Cells, 2019, 8: 744
[25] Chen MX, Zhu FY, Gao B, et al. Full-length transcript-based proteogenomics of rice improves its genome and proteome annotation. Plant Physiol, 2020, 182: 1510-26
[26] Kelkar DS, Provost E, Chaerkady R, et al. Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. Mol Cell Proteomics, 2014, 13: 3184-98
[27] Brosch M, Saunders GI, Frankish A, et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. Genome Res, 2011, 21: 756-67
[28] Kumar D, Yadav AK, Jia X, et al. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. Mol Cell Proteomics, 2016, 15: 329-39
[29] Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature, 2014, 509: 575-81
[30] Zhu Y, Orre LM, Johansson HJ, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun, 2018, 9: 903
[31] Johansson HJ, Socciarelli F, Vacanti NM, et al. Breast cancer quantitative proteome and proteogenomic landscape. Nat Commun, 2019, 10: 1600
[32] Laumont CM, Daouda T, Laverdure JP, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun, 2016, 7: 10238
[33] Madugundu AK, Na CH, Nirujogi RS, et al. Integrated transcriptomic and proteomic analysis of primary human umbilical vein endothelial cells. Proteomics, 2019, 19: e1800315
[34] Lu S, Zhang J, Lian X, et al. A hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res, 2019, 47: 8111-25
[35] Choi S, Ju S, Lee J, et al. Proteogenomic approach to UTR peptide identification. J Proteome Res, 2020, 19: 212-20
[36] Lee SE, Song J, Bosl K, et al. Proteogenomic analysis to identify missing proteins from haploid cell lines. Proteomics, 2018, 18: e1700386
[37] Bespyatykh J, Smolyakov A, Guliaev A, et al. Proteogenomic analysis of *Mycobacterium tuberculosis* Beijing B0/W148 cluster strains. J Proteomics, 2019, 192: 18-26

[38] Pawar H, Pai K, Patole MS. A novel protein coding potential of long intergenic non-coding RNAs (lincRNAs) in the kinetoplastid protozoan parasite *Leishmania major*. Acta Trop, 2017, 167: 21-5
[39] Yagoub D, Tay AP, Chen Z, et al. Proteogenomic discovery of a small, novel protein in yeast reveals a strategy for the detection of unannotated short open reading frames. J Proteome Res, 2015, 14: 5038-47
[40] Low TY, Mohtar MA, Ang MY, et al. Connecting proteomics to next-generation sequencing: proteogenomics and its current applications in biology. Proteomics, 2019, 19: e1800235
[41] Ang MY, Low TY, Lee PY, et al. Proteogenomics: from next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. Clin Chim Acta, 2019, 498: 38-46
[42] Gao Q, Zhu H, Dong L, et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell, 2019, 179: 561-77.e22
[43] Zhang B, Whiteaker JR, Hoofnagle AN, et al. Clinical potential of mass spectrometry-based proteogenomics. Nat Rev Clin Oncol, 2019, 16: 256-68
[44] Dou Y, Kawaler EA, Cui Zhou D, et al. Proteogenomic characterization of endometrial carcinoma. Cell, 2020, 180: 729-48.e26
[45] Hosono Y, Niknafs YS, Prensner JR, et al. Oncogenic role of THOR, a conserved cancer/testis long non-coding RNA. Cell, 2017, 171: 1559-72.e20
[46] Trembinski DJ, Bink DI, Theodorou K, et al. Aging-regulated anti-apoptotic long non-coding RNA *Sarrah* augments recovery from acute myocardial infarction. Nat Commun, 2020, 11: 2039
[47] Yang X, Xiong Q, Wu Y, et al. Quantitative proteomics reveals the regulatory networks of circular RNA *CDR1as* in hepatocellular carcinoma cells. J Proteome Res, 2017, 16: 3891-902
[48] 李爱玲, 宋健. 生物标志物分类及其在临床医学中的应用. 中国药理学与毒理学杂志, 2015, 29: 7-13
[49] Huang JZ, Chen M, Chen D, et al. A peptide encoded by a putative lncRNA *HOXB-AS3* suppresses colon cancer growth. Mol Cell, 2017, 68: 171-84.e6
[50] Zheng X, Chen L, Zhou Y, et al. A novel protein encoded by a circular RNA *circPPP1R12A* promotes tumor pathogenesis and metastasis of colon cancer via Hippo-YAP signaling. Mol Cancer, 2019, 18: 47
[51] Yang Y, Gao X, Zhang M, et al. Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. J Natl Cancer Inst, 2018, 110: 304-15
[52] Zhang M, Zhao K, Xu X, et al. A peptide encoded by circular form of *LINC-PINT* suppresses oncogenic transcriptional elongation in glioblastoma. Nat Commun, 2018, 9: 4475
[53] Matsumoto A, Pasut A, Matsumoto M, et al. mTORC1 and muscle regeneration are regulated by the *LINC00961*-encoded SPAR polypeptide. Nature, 2017, 541: 228-32
[54] Zhu S, Wang JZ, Chen D, et al. An oncopeptide regulates m^6A recognition by the m^6A reader IGF2BP1 and tumorigenesis. Nat Commun, 2020, 11: 1685
[55] Jiang B, Liu J, Zhang YH, et al. Long noncoding RNA *LINC00961* inhibits cell invasion and metastasis in human non-small cell lung cancer. Biomed Pharmacother, 2018, 97: 1311-8
[56] Huang Z, Lei W, Tan J, et al. Long noncoding RNA *LINC00961* inhibits cell proliferation and induces cell apoptosis in human non-small cell lung cancer. J Cell Biochem, 2018, 119: 9072-80
[57] Lu XW, Xu N, Zheng YG, et al. Increased expression of long noncoding RNA *LINC00961* suppresses glioma metastasis and correlates with favorable prognosis. Eur Rev Med Pharmacol Sci, 2018, 22: 4917-24
[58] Chen D, Zhu M, Su H, et al. *LINC00961* restrains cancer progression via modulating epithelial-mesenchymal transition in renal cell carcinoma. J Cell Physiol, 2019, 234: 7257-65
[59] Yin J, Liu Q, Chen C, et al. Small regulatory polypeptide of amino acid response negatively relates to poor prognosis and controls hepatocellular carcinoma progression via regulating microRNA-5581-3p/human cardiolipin synthase 1. J Cell Physiol, 2019, 234: 17589-99
[60] Zhang L, Shao L, Hu Y. Long noncoding RNA LINC00961 inhibited cell proliferation and invasion through regulating the Wnt/β-catenin signaling pathway in tongue squamous cell carcinoma. J Cell Biochem, 2019, 120: 12429-35
[61] Pan LN and Sun YR. *LINC00961* suppresses cell proliferation and induces cell apoptosis in oral squamous cell carcinoma. Eur Rev Med Pharmacol Sci, 2019, 23: 3358-65
[62] Mu X, Mou KH, Ge R, et al. *Linc00961* inhibits the proliferation and invasion of skin melanoma by targeting the miR367/PTEN axis. Int J Oncol, 2019, 55: 708-20
[63] Wu CT, Liu S, Tang M. Downregulation of linc00961 contributes to promote proliferation and inhibit apoptosis of vascular smooth muscle cell by sponging miR-367 in patients with coronary heart disease. Eur Rev Med Pharmacol Sci, 2019, 23: 8540-50
[64] Wu H, Dai Y, Zhang D, et al. *LINC00961* inhibits the migration and invasion of colon cancer cells by sponging miR-223-3p and targeting SOX11. Cancer Med, 2020, 9: 2514-23
[65] Jackson R, Kroehling L, Khitun A, et al. The translation of non-canonical open reading frames controls mucosal immunity. Nature, 2018, 564: 434-8
[66] Song X, Zhao X, Huang Y, et al. Transmissible gastroenteritis virus (TGEV) infection alters the expression of cellular microRNA species that affect transcription of TGEV gene 7. Int J Biol Sci, 2015, 11: 913-22
[67] Zhao X, Bai X, Guan L, et al. *microRNA-4331* promotes transmissible gastroenteritis virus (TGEV)-induced mitochondrial damage via targeting RB1, upregulating interleukin-1 receptor accessory protein (IL1RAP), and activating p38 MAPK pathway *in vitro*. Mol Cell Proteomics, 2018, 17: 190-204
[68] Laumont CM, Perreault C. Exploiting non-canonical

translation to identify new targets for T cell-based cancer immunotherapy. Cell Mol Life Sci, 2018, 75: 607-21

[69] Kanaseki T, Tokita S, Torigoe T. Proteogenomic discovery of cancer antigens: neoantigens and beyond. Pathol Int, 2019, 69: 511-8

[70] Laumont CM, Vincent K, Hesnard L, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med, 2018, 10: eaau5516

[71] Zhao Q, Laverdure JP, Lanoix J, et al. Proteogenomics uncovers a vast repertoire of shared tumor-specific antigens in ovarian cancer. Cancer Immunol Res, 2020, 8: 544-55

[72] He C, Jia C, Zhang Y, et al. Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. J Proteome Res, 2018, 17: 2335-44

[73] 赵加玲, 武舒佳, 王红, 等. 结核分枝杆菌H37Rv新基因Rv2742克隆表达及纯化. 生物工程学报, 2019, 35: 1771-86

[74] Sun J, Shi J, Wang Y, et al. Open-pFind enhances the identification of missing proteins from human testis tissue. J Proteome Res, 2019, 18: 4189-96

[75] Sun J, Shi J, Wang Y, et al. Multiproteases combined with high-pH reverse-phase separation strategy verified fourteen missing proteins in human testis tissue. J Proteome Res, 2018, 17: 4171-7

[76] Wang Y, Chen Y, Zhang Y, et al. Multi-protease strategy identifies three PE2 missing proteins in human testis tissue. J Proteome Res, 2017, 16: 4352-63

[77] Zhao M, Wei W, Cheng L, et al. Searching missing proteins based on the optimization of membrane protein enrichment and digestion process. J Proteome Res, 2016, 15: 4020-9

[78] Wei W, Luo W, Wu F, et al. Deep coverage proteomics identifies more low-abundance missing proteins in human testis tissue with Q-exactive HF mass spectrometer. J Proteome Res, 2016, 15: 3988-97

[79] Zhang Y, Li Q, Wu F, et al. Tissue-based proteogenomics reveals that human testis endows plentiful missing proteins. J Proteome Res, 2015, 14: 3583-94