遗传 Hereditas (Beijing) 2021 年 8 月, 43(8): 737－746
www.chinagene.cn

综 述

小开放阅读框编码微肽的研究进展

陈相颖，李梦玮，王颖，陈权，徐寒梅

中国药科大学江苏省合成多肽药物发现与评价工程中心，南京 211198

摘要： 已有的研究表明，生命体中存在着大量的非编码 RNA (non-coding RNA, ncRNA)，先前被错误注释为 ncRNA 的分子序列中实际上包含小的开放阅读框(short open reading frame, sORF)，部分 sORF 可转录并翻译成进化保守的微肽(micropeptide)，这些 sORF 由于序列较短和研究技术手段的限制而被忽略。迄今为止，已在生命体中发现一些 sORF 编码的功能各异的微肽，它们对生命活动的调控起着重要作用。本文对近年来发现的功能性微肽进行综述，介绍了本课题组发现新型微肽 MIAC (micropeptide inhibiting actin cytoskeleton)的过程，同时总结了研究潜在微肽的相关技术，以期为研究人员利用相关技术发现新微肽提供借鉴和参考。

关键词： 非编码 RNA；小开放阅读框；微肽

Progress on sORF-encoded micropeptides

Xiangying Chen, Mengwei Li, Ying Wang, Quan Chen, Hanmei Xu

Engineering Research Center of Synthetic Peptide Drug Discovery and Evaluation of Jiangsu Province, China Pharmaceutical University, Nanjing 211198, China

Abstract: Existing research has shown that there are a large amount of non-coding RNAs (ncRNAs) in organisms. Short open reading frames (sORFs) abundantly exist in molecular sequences inaccurately annotated as ncRNAs. Several sORFs can be transcribed and translated into evolutionarily conserved micropeptides, which were ignored in previous studies due to short sequence lengths and the limitations of research techniques. To date, sORF-encoded micropeptides with various functions have been found to play important roles in regulating vital biological activities. This article reviews the functional micropeptides which have been found in recent years, introduces the new micropeptide designated as MIAC that we have discovered and describes the related technologies for mining potential micropeptides, thereby providing insights and references for new micropeptide discovery for researchers.

Keywords: non-coding RNA; small open reading frames; micropeptides

收稿日期: 2021−05−08; **修回日期:** 2021−07−06

基金项目: 中国药科大学天然药物活性组分与药效国家重点实验(编号：SKLNMZZCX201821, SKLNMZZ202028)，国家科技重大新药开发项目(编号：2019ZX09301124, 2019ZX09201001, 2019ZX09301-110)和中国博士后科学基金资助项目(编号：2017M621884, 2020M681787)资助[Supported by the Project Program of State Key Laboratory of Natural Medicines (Nos. SKLNMZZCX201821, SKLNMZZ202028), the National Science and Technology Major Projects of New Drugs (Nos. 2019ZX09301124, 2019ZX09201001, 2019ZX09301-110) and China Postdoctoral Science Foundation (Nos. 2017M621884, 2020M681787)]

作者简介: 陈相颖，在读硕士研究生，专业方向：微生物与生化药学。E-mail: cxy000111@qq.com

通讯作者: 徐寒梅，博士，教授，研究方向：多肽类药物研究与开发。E-mail: 13913925346@126.com

DOI: 10.16288/j.yczz.21-167

网络出版时间: 2021/7/23 16:26:21

URI: https://kns.cnki.net/kcms/detail/11.1913.R.20210723.0937.002.html

随着科学技术的发展，人们对于生物复杂性有了更进一步的认识。中心法则指出，遗传信息通常经历由脱氧核苷酸转录到核糖核苷酸再翻译为蛋白质的过程[1]。人类基因组计划证明，人类基因组的3/4 能够被转录，但只有约 1.5%的基因具有编码蛋白的能力[2]。这就引发了人们对基因组中剩余的大量非蛋白编码基因的思考，这些非蛋白编码基因是否包含更多的遗传信息。DNA 元件百科全书(encyclopedia of DNA elements, ENCODE)的数据表明，80%的基因具有特定的生物学功能，而大部分基因处于非蛋白质编码区域，这部分基因转录产生大量的 ncRNA[3]。随着高通量测序技术的发展，越来越多的实验证明 ncRNA 分子序列上含有小的开放阅读框序列(short open reading frame, sORF)，可编码小于 100 个氨基酸的微小蛋白，被人们称为微肽(micropeptide)，加工修饰后的微肽可通过与其他蛋白相互作用而发挥其生理或病生理的作用[4]。研究表明，包括果蝇(*Drosophila melanogaster*)、小鼠(*Mus musculus*)、人(*Homo spaiens*)在内的许多动物基因组中包含数百万个 sORF，其中一些具有关键的生理或病生理功能，如钙离子稳态、代谢、成肌细胞融合和肌肉发育、胚胎发育、物质降解、癌症等[5~24]。本文主要介绍了近年来发现的功能性微肽、本课题组发现新型微肽 MIAC (micropeptide inhibiting actin cytoskeleton)的过程以及研究潜在微肽的技术，期望为进行相关微肽研究的科研人员提供新思路。

1 sORF 和微肽简介

开放阅读框(open reading frame, ORF)最初被定义为起始密码子与终止密码子间的潜在翻译序列[25]。可翻译的 ORF 通常是指 mRNA 上的编码序列(coding sequences, CDS)，该序列翻译产生具有生物学功能的蛋白质[26]。由于 ORF 编码蛋白质的可能性随着其长度的增加而增加，查找 ORF 的算法大多都以 300 个密码子或 100 个氨基酸为阈值作为最短的检测长度[27]。sORF 在序列长度上区别于 ORF，理论上 sORF 的大小可以从最低限制的 2 个密码子到 100 个密码子，sORF 由于其极短的长度在最初被认为是非编码的[28]。最近研究发现，真核基因组中存在数百万个 sORF 序列，并且有些 sORF 序列可以定位到转录本，这部分 sORF 具有编码并翻译产生蛋白的能力[28,29]。因此，微肽被定义为长度小于 100 个氨基酸的蛋白质。

根据果蝇和哺乳动物中 sORF 的位置、大小、保守性和翻译方式等特性，sORF 可分为五类(图 1)：基因间 ORF (intergenic ORF)、上游 ORF (upstream ORF, uORF)、长非编码 ORF (long non-coding ORF, lncORF)、短编码序列(short coding sequence, short CDS)和短同工型 ORF (short isoform ORF)。其中，基因间 ORF 占 sORF 的 96%，但其并不会进行转录和翻译。uORF 位于 5′端非翻译区(5′ untranslated region, 5′ UTR)，具有较低效率的翻译功能并能调节转录本中下游的 ORF。lncORF 存在于 lncRNA (long non-coding RNA)中，与 uORF 相似具有较低的翻译效率。最近发现，几种 lncRNA 可编码和翻译为具有生物学功能的微肽，并且在进化中高度保守。短 CDS 是在单顺反子转录本中发现的，具有与 ORF 类似的翻译效率，在果蝇和哺乳动物中存在着数百种短 CDS。短同工型 ORF 是 sORF 中占比最少的一类，由 mRNA 的选择性剪接产生[4]。

关于微肽的翻译机制，目前有如下几种可能的解释：根据核糖体的扫描模型，mRNA 的 5′端帽子结构与核糖体 40S 小亚基结合以复合物形式向 3′端扫描，若遇起始密码子，核糖体 40S 小亚基便与 60S 大亚基形成 80S 核糖体，从而介导 5′UTR 中 sORF 的翻译，遇到终止密码子时翻译结束，大小亚基解离；而 40S 小亚基则继续向前扫描，当遇到 ORF 的起始密码子时重新结合核糖体 60S 大亚基。第二种可能的机制是只有部分 40S 小亚基结合在 5' UTR 中 sORF 的起始密码子处，另一部分继续向前扫描至 ORF 的起始密码子，这种机制被称为核糖体的泄漏扫描[30]。但上述两种机制只适用于 uORF 产生的微肽，关于其他类型微肽的翻译机制还有待研究，还有一种猜想是 RNA 编辑促成了微肽的翻译，即在转录后将 A-C/G/A-G 修改为 A-U-G[31]。未来可以通过基因敲除 RNA 编辑的关键酶来研究 RNA 编辑在 RNA 水平上产生 sORF 起始密码子的作用。

2 功能性微肽的发现

基于生物信息学及高通量测序技术对微肽的深

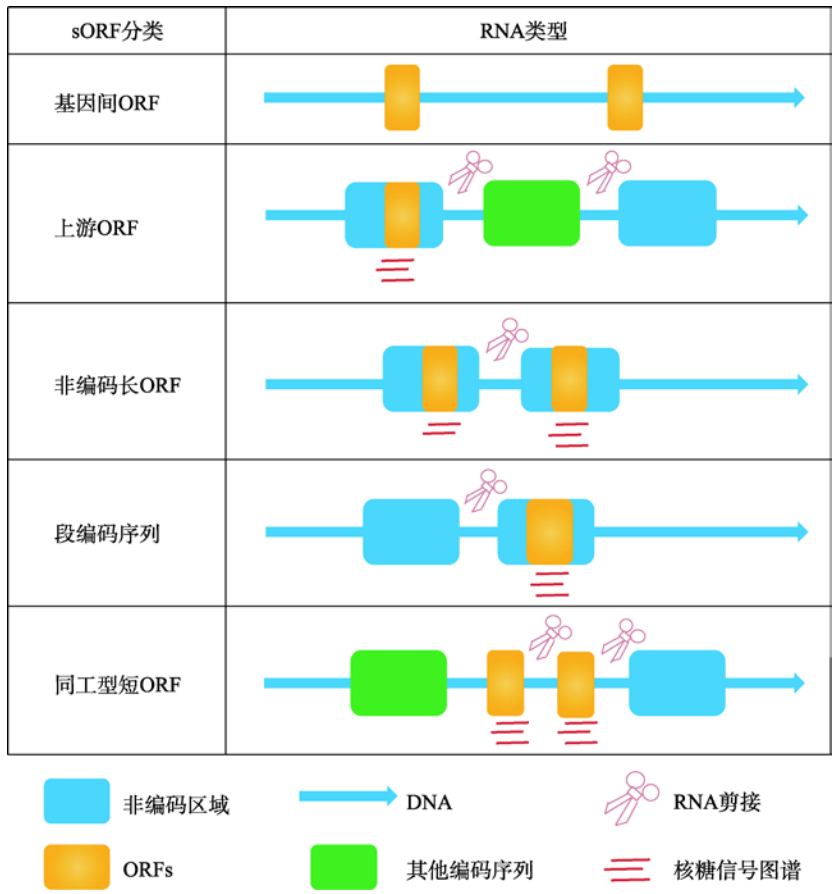

图 1　sORF 的分类

Fig. 1　sORF classification

入研究，越来越多的微肽被证明在生命活动的许多过程中起着重要的调节作用，包括钙离子稳态、代谢、成肌细胞融合和肌肉发育、胚胎发育、物质降解、癌症等(图 2)[5~24]。下面将对近年来功能性微肽的发现进行介绍，并将其总结在表 1 中。

2.1　钙离子稳态相关微肽

Ca^{2+}是肌肉收缩的主要调节因子，控制着肌肉的生长、代谢和病理重塑[32]。美国德克萨斯大学西南医学中心 Eric N. Olson 实验室的研究结果显示，微肽对于 Ca^{2+}稳态的调节起重要作用[5,6,8]。肌调素(myoregulin, MLN)是由骨骼肌特异性 lncRNA 编码的 46 个氨基酸的微肽，它可直接与肌浆网 Ca^{2+}-ATP 酶(sarcoplasmic reticulum Ca^{2+}-ATPase, SERCA)相互作用以降低 SERCA 对 Ca^{2+}的亲和力，从而减少 Ca^{2+}摄入肌浆网和肌细胞的收缩性[6]。因此，将 MLN 鉴定为骨骼肌中的 SERCA 抑制性微肽。相反，DWORF (dwarf open reading frame)可解除 SERCA 抑制性微肽的作用而增强肌浆网摄取 Ca^{2+}的能力，它是由心肌特异性 lncRNA 编码的 34 个氨基酸的微肽[8]。随后，Anderson 等[5]在非肌肉细胞中发现两种 SERCA 抑制性微肽 ELN (endoregulin)和 ALN (another-regulin)，这两种微肽具有与 MLN 相似的结构和功能，表明 Ca^{2+}相关微肽在不同的细胞类型中保守，Ca^{2+}稳态的调节对于许多细胞功能具有重要意义。

另外，Magny 等[7]在果蝇的 *pncr003:2L* 基因中也发现编码 SERCA 抑制性微肽的序列，该微肽可影响果蝇心肌中 Ca^{2+}的运输。跨物种的氨基酸相关序列分析表明，Ca^{2+}相关微肽的结构和功能在果蝇到脊椎动物中具有高度保守性，与其在 SERCA 中调节 Ca^{2+}摄取的生物学功能相关[5,7]。

2.2　线粒体代谢相关微肽

线粒体作为一种功能性细胞器，在新陈代谢及

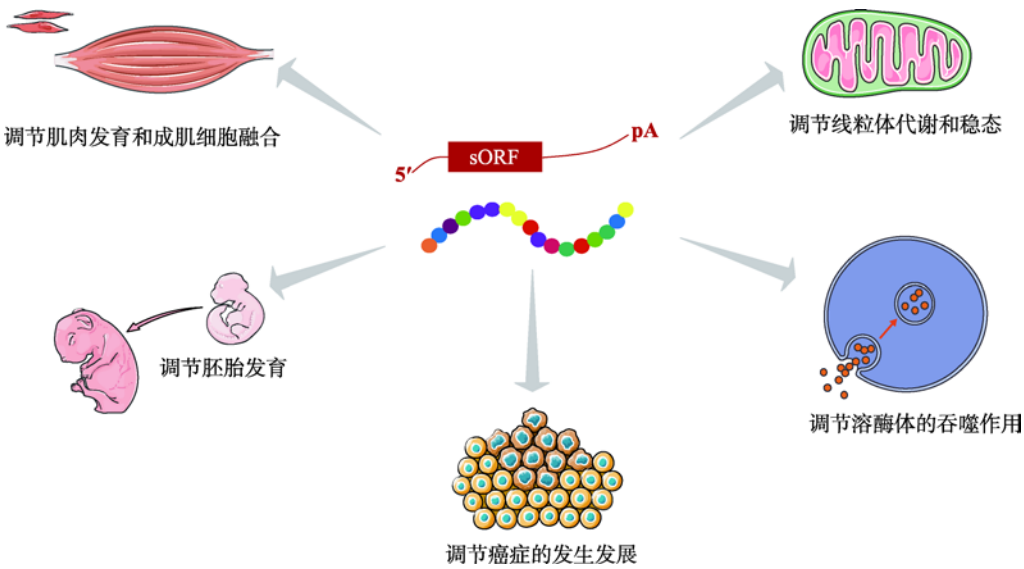

图 2　微肽的生理与病生理功能

Fig. 2　Physiological and Pathophysiological functions of micropeptides

表 1　功能性微肽的发现

Table 1　Discovery of functional micropeptides

基因	微肽	长度(氨基酸)	作用	参考文献
鼠 *AK009351*、人 *LINC00948*	MLN	46	抑制 SERCA，调节钙离子转运	[6]
1110017F19Rik/SMIM6	ELN	56	抑制 SERCA，调节钙离子转运	[5]
1810037I17Rik	ALN	65	抑制 SERCA，调节钙离子转运	[5]
鼠 *NONMMUG026737*、人 *LOC100507537*	DWORF	34	激活 SERCA，调节钙离子转运	[8]
pncr003:2L	Scl	28/29	调节钙离子转运，影响肌肉收缩	[7]
鼠 *1500011K16Rik*、人 *LINC00116*	MOXI	56	增强脂肪酸 β-氧化作用	[10]
LINC00116	Mtln	56	增强呼吸效率	[11]
12S rRNA	MOTS-c	16	调节胰岛素敏感性	[9]
LOC101929726	Minion	84	促进成肌细胞融合和肌肉发育	[12]
LOC101929726	Myomixer	84	促进成肌细胞融合和肌肉发育	[13]
LOC100506013	Toddler	54	激活 APJ/Apelin 受体促进胚胎发育	[14]
polished rice(*pri*)	Pri	11/32	促进胚胎发育中的表皮形成	[15]
Tarsal-less(*tal*)	Tal	11	控制基因表达和组织折叠	[16]
LINC00961	SPAR	90	抑制 mTORC1 和肌肉再生	[17]
hemotion	Hemotion	88	促进吞噬细胞吞噬作用	[18]
PIGBOS	PIGBOS	54	调节内质网应激反应	[19]
SMIM22	CASIMO1	83	促进乳腺癌	[20]
HOXB-AS3	HOXB-AS3	53	抑制结肠癌	[21]
LINC00998	SMIM30	59	促进肝癌	[22]
LINC00278	YY1BM	21	抑制食管鳞状细胞癌	[23]
AC025154.2	MIAC	51	抑制头颈鳞状细胞癌	[24]
LINC01420	NoBody	68	促进无义介导的 mRNA 衰变	[38]
MIR155HG	miPEP155(P155)	17	调节抗原呈递细胞的抗原转运和呈递	[45]

能量供应方面起着重要的作用，大量的研究表明线粒体 DNA 中也存在 sORFs[9,33]。Makarewich 等[10]在线粒体内膜中发现由 lncRNA 编码的微肽 MOXI (micropeptide regulator of β-oxidation)，MOXI 与催化长链脂肪酸氧化的线粒体三功能蛋白(mitochondrial trifunctional protein, MTP)结合，可增强脂肪酸的 β 氧化作用。Stein 等[11]在骨骼肌和心脏中发现由 lncRNA *LINC00116* 编码的线粒体跨膜蛋白 Mtln (mitoregulin)，Mtln 作为粘性分子可通过增强线粒体蛋白复合物的装配和稳定性从而提高线粒体的呼吸效率。此外，在线粒体中还发现由 *12S rRNA* 编码的 16 个氨基酸的微肽 MOTS-c (mitochondrial open reading frame of the 12S rRNA-c)，它可抑制叶酸循环及嘌呤核苷酸的从头合成途径而导致 AMPK (AMP-activated protein kinase)活化，从而调节胰岛素的敏感性[9]。这些结果表明，线粒体可通过微肽在细胞和机体水平上主动调控代谢稳态。

2.3　成肌细胞融合和肌肉发育相关微肽

骨骼肌的形成需要单核成肌细胞融合形成多核细胞肌管以产生收缩性肌纤维，Myomaker 是成肌细胞融合所需的肌肉特异性蛋白[34,35]。最近研究发现，有多种肌肉特异性的微肽在哺乳动物成肌细胞融合的过程中也起着关键作用[12,13]。Zhang 等[12]发现一种 sORF 编码的新微肽 Minion (microprotein inducer of fusion)，Minion 与 Myomaker 共表达可诱导细胞融合和细胞骨架的快速重排。Myomixer 是长为 84 个氨基酸的肌肉特异性微肽，可促进成肌细胞融合，Myomixer 与 Myomaker 结合还可诱导成纤维细胞间的融合及成纤维细胞和成肌细胞的融合[13]。因此，sORF 编码的微肽对于肌肉发育过程中的肌纤维形成具有重要调控作用。

2.4　胚胎发育相关微肽

Toddler 是在斑马鱼(*Danio rerio*)中发现的由 lncRNA *LOC100506013* 编码的长为 58 个氨基酸的微肽，研究发现它作为 APJ/Apelin 受体信号转导的激活剂，可促进原肠胚的形成[14]。先前的研究表明 APJ/Apelin 受体信号转导在心血管发育和生理调节等多种生物过程中发挥着重要作用[36]。Pauli 等[14]发现 Toddler 功能缺失的斑马鱼没有正常的心脏和血液循环，这些研究表明 Toddler 在早期胚胎发育过程中是不可或缺的。

另外，在果蝇中还发现与胚胎发育相关的微肽[15,16]。Kondo 等[15]在果蝇的上皮组织中发现 lncRNA *polished rice*(*pri*)实际上被转录成多顺反子 mRNA，可编码长为 11 或 32 个氨基酸的微肽(Pri)。Pri 通过调节 F-actin 在上皮形态的发生中起重要作用，而 Pri 功能的丧失可完全消除果蝇的表皮结构。Galindo 等[16]在果蝇中发现基因 *tarsal-less*(*tal*)对果蝇的胚胎发育和形态发生至关重要，*tal* 可翻译为短至 11 个氨基酸的微肽，控制着果蝇的基因表达和组织折叠。这些结果表明，极短的 sORF 具有翻译功能并在发育过程中具有重要调控作用。

2.5　物质降解相关微肽

近年来，在生物技术的驱动下，与物质降解作用相关的微肽也不断被发现，它们在废物和毒素的降解方面发挥着重要的作用[17~19]。SPAR (small regulatory polypeptide of amino acid response)是由 lncRNA *LINC00961* 编码的长为 90 个氨基酸的保守性微肽，定位于晚期内体及溶酶体。SPAR 与溶酶体表面 v-ATPase 复合物的四个亚基相互作用，负性调节 mTORC1 的活化而抑制肌肉再生[17]。

此外，Pueyo 等[18]在果蝇中发现一个组织特异性的 sORF 基因 *hemotion*，其编码果蝇巨噬细胞中长为 88 个氨基酸的跨膜微肽(Hemotin)。实验研究表明，Hemotin peptide 结合并抑制衔接蛋白 14-3-3ζ，进而促进磷脂酰肌醇的磷酸化而调节吞噬作用中的内体成熟。并且，研究人员在脊椎动物中还发现 Hemotin 的功能同源物 Stannin，表明这种吞噬作用的新型调节因子具有物种间保守性[18]。

未折叠蛋白反应(unfolded protein response, UPR)是真核细胞内质网(endoplasmic reticulum, ER)中的一个基本过程，在 ER 中只有正确组装和折叠的蛋白质才能分泌到胞外或展示在细胞表面，而不折叠的蛋白将被内质网相关蛋白所降解[37]。位于线粒体外膜的微肽 PIGBOS 与 ER 蛋白 CLCC1 结合从而调节内质网中的 UPR，而 PIGBOS 的缺失会导致 UPR 升高和细胞死亡[19]。由此可见，微肽对细胞器间的

通讯、体内的平衡以及细胞的存亡至关重要。

2.6 癌症相关微肽

微肽在癌症的发生发展中也具有重要的调控作用[20~24]。CASIMO1 (cancer-associated small integral membrane open reading frame 1)是第一个被发现具有致癌作用的功能性微肽，它与胆固醇合成的关键酶角鲨烯环氧化酶(squalene epoxidase, SQLE)相互作用从而调节癌细胞的代谢稳态，敲低 *CASIMO1* 可导致乳腺癌细胞的增殖减少[20]。在结肠癌(colon cancer, CRC)方面，Huang 等[21]发现由 lncRNA *HOXB-AS3* 编码的长为 53 个氨基酸的保守微肽(HOXB-AS3)的缺失是 CRC 代谢中的关键致癌因素，HOXB-AS3 能抑制结肠癌的生长。Pang 等[22]还发现由 lncRNA *LINC00998* 编码的 59 个氨基酸的微肽 SMIM30 可通过调节细胞增殖和迁移促进肝癌的发生发展。Wu 等[23]通过对 281 对男性食管鳞状细胞癌(esophageal squamous cell carcinoma, ESCC)组织样本中 lncRNA 的差异表达分析发现与癌旁组织相比，*LINC00278* 在 ESCC 组织中显著下调，进一步研究发现 *LINC00278* 编码微肽 YY1BM (Yin Yang 1 (YY1)-binding micropeptide)，YY1BM 可与雄激素受体(androgen receptor, AR)结合并下调 *eEF2K* 的表达从而导致癌细胞凋亡，由此可见，YY1BM 可作为一种潜在的抗癌微肽。另外，Li 等[24]证明 MIAC 能抑制头颈鳞状细胞癌(head and neck squamous cell carcinoma, HNSCC)的生长和转移。此外，还有一些微肽与癌症并不直接相关，例如 NoBody (*non*-annotated P-*body* dissociating polypeptide)是由 *LINC01420* 编码的 68 个氨基酸的微肽，它与 mRNA 脱帽蛋白相互作用，促进无义介导的 mRNA 衰变(nonsense mediated decay, NMD)，癌细胞可能利用此过程降解抑制肿瘤的 mRNA[38]。总而言之，这些新发现的转录本丰富了肿瘤调控分子，并为癌症的临床诊断和治疗提供新的潜在靶标。

3 微肽 MIAC 的发现

本课题组研究发现，lncRNA *AC025154.2* 可能编码长为 51 个氨基酸的内源性微肽[24]。在验证这段序列的翻译编码能力时，我们通过体外翻译实验和体内细胞构建实验加以证明，结果表明 lncRNA *AC025154.2* 能够编码一种新型微肽，我们将其命名为 MIAC。在微肽 MIAC 的功能研究中，我们构建了 MIAC 稳定过表达和敲除的 CAL27 细胞系，实验发现 MIAC 通过负调控癌细胞的增殖和转移而抑制 HNSCC 的发生发展。

接下来，我们对 MIAC 抑制 HNSCC 相关机制做了进一步的研究。通过质谱鉴定与 MIAC 相互作用的蛋白，并结合 50 个 HNSCC 临床样本和 50 个正常样本中蛋白的表达情况，最后聚焦于其中的这三种蛋白：水通道蛋白 2 (aquaporin 2, AQP2)、ITGB4 (integrin beta 4)和 SEPT2 (septin 2)。进一步的机制探究表明 MIAC 直接与 AQP2 相互作用，通过调控 SEPT2/ITGB4 抑制骨架蛋白重排，最终抑制 HNSCC 的生长和转移。由此可见，MIAC 在 HNSCC 中具有调控作用，为开发治疗 HNSCC 的药物提供新的研究方向，而 AQP2 作为 MIAC 的作用靶点对于研究 HNSCC 的药物同样存在重要意义。

为进一步探究 MIAC 的临床和治疗意义，我们分析 TCGA 数据库中 500 个 HNSCC 临床样本和 44 个正常样本中 MIAC 的相对表达情况，发现 MIAC 在 HNSCC 中呈下调趋势，并且 MIAC 表达水平的降低与 HNSCC 患者的整体生存率差呈正相关。我们进一步分析 94 对 HNSCC 临床样本中 MIAC 的相对表达量，分析结果也与数据库中的情况一致，相比于在正常样本中，MIAC 在 HNSCC 样本中的表达量降低。因此，MIAC 是由 lncRNA *AC025154.2* 编码的一种新型内源性微肽，对于 HNSCC 的发生发展起着重要的调控作用。

创新型药物作为自主研发和具有自主知识产权的药物，对于我国建设创新型国家的进一步发展具有重要意义。MIAC 作为 HNSCC 的调控分子，在创新型 HNSCC 药物的开发中具有重大的研究意义：MIAC 可作为潜在诊断标志物来制备诊断 HNSCC 的试剂盒，为 HNSCC 的诊断和预防提供新的途径；而 MIAC 作为调控 HNSCC 的小分子多肽，也可通过偶联化学药物的方式来提高治疗 HNSCC 药物的靶向性和稳定性。此外，MIAC 在其他肿瘤和疾病中的意义还有待探究。

4　研究潜在微肽的相关技术

当前的研究表明，在动物基因组中约 1.2%的 sORF 可被转录，其中只有约 1/3 能被翻译[4]。这些占比很小的功能性 sORF 理论上也可产生成千上万个未被表征的微肽，即使这些微肽中只有小部分具有生物活性，仍意味着可能存在数百甚至数千种有生物学功能的微肽。因此，当前面临的挑战是如何识别具有生物活性的 sORF 及其微肽。下面将总结介绍研究潜在微肽的相关技术，这些技术可用于鉴定可能编码微肽的 sORF。

4.1　生物信息学分析

生物信息学(bioinformatics)是利用生物数据来开发算法和软件的交叉学科，目前运用生物信息分析技术，基于保守序列可从非编码区域预测具有编码蛋白能力的 sORFs，生物信息分析技术还依据 sORFs 序列中的密码子含量和编码特征以区分 sORFs 编码区与非编码区[39]。我们可以利用生物信息数据库挖掘相关数据，如 ATCG、UCSC 等，而常用于预测 sORFs 的分析软件有 CPAT、ORFfinder、PhyloCSF 、uPEPperoni[40~44] 等 。Niu 等[45] 运用 ORFfinder 在人源 *MIR155HG* 基因中预测到一条 54 个碱基的 sORF，后续实验证明该 sORF 可编码长为 17 个氨基酸的功能性微肽 miPEP155。miPEP155 可调节抗原呈递细胞(antigen-presenting cells, APC)中的抗原转运和呈递，可作为自身免疫性疾病的候选药物[45]。

4.2　核糖体图谱分析

核糖体图谱分析(ribosome profiling)可用来识别具有翻译潜力的 sORF，该技术的原理是翻译核糖体可保护长为 20~30 个核苷酸的 mRNA 片段免受核酸酶的消化[46]。然而，Wilson 等[47]的研究表明某些 sORF 虽然与核糖体结合但并不进行翻译。于是，在 Ribo-Seq 的基础上改良而开发了多聚核糖体分析(Poly-Ribo-Seq)，使用这种技术可以分离由多个核糖体结合并被主动翻译的 mRNA，由此可将不进行翻译的单核糖体-mRNA 复合物区分开[48]。此外，Guttman 等[49]还开发了核糖体释放分数(ribosome release score, RRS)作为翻译的度量指标，相比于终止密码子下游的非编码区，编码区与核糖体具有更高的相关性，由此可区分编码转录本和非编码转录本。Chen 等[26]利用核糖体图谱分析发现了 3455 个非经典 CDS，其中的 96%是编码小于 100 个氨基酸的微肽。

4.3　质谱和蛋白质组学

最近基于质谱(mass spectrometry, MS)的蛋白质组学(proteomics)也用于发现和验证内源表达的微肽。该技术的基本原理是通过测量气态的离子化肽或蛋白质的质荷比来研究蛋白质的表达和相互作用，因此 MS 通过检测从 sORF 翻译的微肽，从而直接验证转录产物编码蛋白质的潜力[50]。基于 MS 的蛋白质组学在研究和鉴定新型微肽方面已取得实质性的进展，Chen 等[26]通过基于 MS 的 HLA-I 肽组学，发现 240 个微肽可被 HLA-I 提呈，表明这些肽会进入 HLA-I 呈递途径并可能拥有免疫原性。但 MS 在技术上仍然存在一定限制，样品制备过程中的消化酶决定微肽片段化的方式，片段过小不能产生足够的检测信号，片段过大则无法用于 MS 分析，小片段的微肽在样品制备过程中还存在丢失的可能[39]。因此，需要进一步结合核糖体图谱分析等其他分析方法以确定新型微肽的存在。

4.4　蛋白质基因组学

蛋白质基因组学(proteogenomics)是在基于蛋白质组学分析的基础上结合基因组学和转录组学的分析方法，通过追溯基因组和转录本中的蛋白质/微肽的预测序列，来鉴定基因的翻译和表达情况[51]。在蛋白质基因组学研究中，Slavoff 等[31]从人白血病细胞系 K562 细胞中发现了 86 个未报道过的微肽。

4.5　其他相关技术

为证实 sORF 是否具有编码蛋白产生微肽的能力，可以使用以下几种方法来进行验证。在理想的状态下，可以设计目的微肽的抗体并通过免疫组化或蛋白质印记来验证其特异性[52]。例如，Li 等[24]通过制备 MIAC 的单克隆抗体以检测 MIAC 的内源表达。对于不能产生抗体的目的微肽而言，也可采用 CRISPR/Cas9 基因编辑技术。该技术通过同源定向修复将 FLAG/MYC 或其他标签添加到预测的

sORF，从而产生融合蛋白，再通过检测融合蛋白以验证目的微肽的存在[52]。为确定 *CASIMO1* 转录本中的 sORF 是否翻译为微肽，Polycarpou-Schwarz 等[20]在 CASIMO1 编码序列的 C 端插入了一个 Flag 标签，并通过 anti-Flag 抗体检测到了 CASIMO1-Flag 的表达。此外，还可通过体外翻译来评估 sORF 编码蛋白的能力，通过多方面的验证以确定 sORF 是否具有编码能力。

5 结语与展望

大规模基因组测序的迅速发展促进人们对基因组的深入研究，揭示 sORF 序列的复杂性。微肽的发现使人们认识到这些重要小肽的生物学作用，它们在生命活动及疾病的发展进程中起着重要调控作用。微肽可以以配体或信号分子的形式发挥作用，也可与其他蛋白质相结合，通过遮蔽受体蛋白的关键位点或影响受体蛋白的活性从而发挥调控作用，如前所述的 HOXB-AS3 peptide[21]通过竞争性结合 hnRNP A1 中 RGG 基序的精氨酸残基，阻断精氨酸残基与丙酮酸激酶 M(pyruvate kinase M, PKM)的结合，从而抑制结肠癌细胞的葡萄糖代谢进程。SMIM30[22]与非受体酪氨酸激酶 SRC/YES1 结合，驱动其膜锚定和磷酸化，激活下游丝裂原活化蛋白激酶(mitogen-activated protein kinase, MAPK)信号通路，通过调节细胞增殖和迁移促进肝癌的发生发展。

然而，对于功能性微肽及其作用机制的探索仍处于起步状态，虽然已经存在许多挖掘未知微肽的生物技术，但由于微肽本身分子量小、表达丰度低等特点，这些生物技术的应用仍然存在局限性，生命体中仍有大量的微肽等待被发现。相信在未来的研究中，能克服检测障碍，进一步拓展和优化挖掘微肽的技术与方法。另一方面，还需要进行大量的工作以阐明微肽的生物学作用，并对其作用机制开展进一步的研究，以便应用于正常生理功能的探索及疾病的临床诊疗。

参考文献(References):

[1] Crick F. Central dogma of molecular biology. *Nature*, 1970, 227(5258): 561–563. [DOI]

[2] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue CH, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang HE, Wrobel J, Yu YB, Ruan XA, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan YJ, Wold B, Carninci P, Guigó R, Gingeras TR. Landscape of transcription in human cells. *Nature*, 2012, 489(7414): 101–108. [DOI]

[3] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489(7414): 57–74. [DOI]

[4] Couso JP, Patraquim P. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol*, 2017, 18(9): 575–589. [DOI]

[5] Anderson DM, Makarewich CA, Anderson KM, Shelton JM, Bezprozvannaya S, Bassel-Duby R, Olson EN. Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci Signal*, 2016, 9(457): ra119. [DOI]

[6] Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, Olson EN. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, 2015, 160(4): 595–606. [DOI]

[7] Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, Couso JP. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, 2013, 341(6150): 1116–1120. [DOI]

[8] Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu FF, Reese AL, McAnally JR, Chen XW, Kavalali ET, Cannon SC, Houser SR, Bassel-Duby R, Olson EN. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, 2016, 351(6270): 271–275. [DOI]

[9] Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan JX, Kim SJ, Mehta H, Hevener AL, de Cabo R, Cohen P. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and

insulin resistance. *Cell Metab*, 2015, 21(3): 443–454. [DOI]

[10] Makarewich CA, Baskin KK, Munir AZ, Bezprozvannaya S, Sharma G, Khemtong C, Shah AM, McAnally JR, Malloy CR, Szweda LI, Bassel-Duby R, Olson EN. MOXI is a mitochondrial micropeptide that enhances fatty acid beta-oxidation. *Cell Rep*, 2018, 23(13): 3701–3709. [DOI]

[11] Stein CS, Jadiya P, Zhang XM, McLendon JM, Abouassaly GM, Witmer NH, Anderson EJ, Elrod JW, Boudreau RL. Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell Rep*, 2018, 23(13): 3710–3720.e8. [DOI]

[12] Zhang Q, Vashisht AA, O'Rourke J, Corbel SY, Moran R, Romero A, Miraglia L, Zhang J, Durrant E, Schmedt C, Sampath SC, Sampath SC. The microprotein Minion controls cell fusion and muscle formation. *Nat Commun*, 2017, 8: 15664. [DOI]

[13] Bi PP, Ramirez-Martinez A, Li H, Cannavino J, McAnally JR, Shelton JM, Sánchez-Ortiz E, Bassel-Duby R, Olson EN. Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, 2017, 356(6335): 323–327. [DOI]

[14] Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, Mitchell A, Ma J, Dubrulle J, Reyon D, Tsai SQ, Joung JK, Saghatelian A, Schier AF. Toddler: an embryonic signal that promotes cell movement *via* Apelin receptors. *Science*, 2014, 343(6172): 1248636. [DOI]

[15] Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol*, 2007, 9(6): 660–665. [DOI]

[16] Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol*, 2007, 5(5): e106. [DOI]

[17] Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama KI, Clohessy JG, Pandolfi PP. mTORC1 and muscle regeneration are regulated by the *LINC00961*-encoded SPAR polypeptide. *Nature*, 2017, 541(7636): 228–232. [DOI]

[18] Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, Bishop SA, Couso JP. Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS Biol*, 2016, 14(3): e1002395. [DOI]

[19] Chu Q, Martinez TF, Novak SW, Donaldson CJ, Tan D, Vaughan JM, Chang TN, Diedrich JK, Andrade L, Kim A, Zhang T, Manor U, Saghatelian A. Regulation of the ER stress response by a mitochondrial microprotein. *Nat Commun*, 2019, 10(1): 4883. [DOI]

[20] Polycarpou-Schwarz M, Groß M, Mestdagh P, Schott J, Grund SE, Hildenbrand C, Rom J, Aulmann S, Sinn HP, Vandesompele J, Diederichs S. The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene*, 2018, 37(34): 4750–4768. [DOI]

[21] Huang JZ, Chen M, Chen D, Gao XC, Zhu S, Huang HY, Hu M, Zhu HF, Yan GR. A peptide encoded by a putative lncRNA *HOXB-AS3* suppresses colon cancer growth. *Mol Cell*, 2017, 68(1): 171–184.e6. [DOI]

[22] Pang YN, Liu ZY, Han H, Wang BL, Li W, Mao CB, Liu SR. Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *J Hepatol*, 2020, 73(5): 1155–1169. [DOI]

[23] Wu SQ, Zhang LY, Deng JQ, Guo BB, Li F, Wang YR, Wu R, Zhang SH, Lu JC, Zhou YF. A novel micropeptide encoded by Y-linked *LINC00278* links cigarette smoking and AR signaling in male esophageal squamous cell carcinoma. *Cancer Res*, 2020, 80(13): 2790–2803. [DOI]

[24] Li MW, Li X, Zhang YN, Wu HM, Zhou HZ, Ding X, Zhang XM, Jin XR, Wang Y, Yin XQ, Li CC, Yang PW, Xu HM. Micropeptide MIAC inhibits HNSCC progression by Interacting with aquaporin 2. *J Am Chem Soc*, 2020, 142(14): 6708–6716. [DOI]

[25] Sieber P, Platzer M, Schuster S. The definition of open reading frame revisited. *Trends Genet*, 2018, 34(3): 167–170. [DOI]

[26] Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, Weissman JS. Pervasive functional translation of noncanonical human open reading frames. *Science*, 2020, 367(6482): 1140–1146. [DOI]

[27] Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, Khan OM, Brewer JR, Skadow MH, Duizer C, Harman CCD, Chang L, Bielecki P, Solis AG, Steach HR, Slavoff S, Flavell RA. The translation of non-canonical open reading frames controls mucosal immunity. *Nature*, 2018, 564(7736): 434–438. [DOI]

[28] Orr MW, Mao YH, Storz G, Qian SB. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res*, 2020, 48(3): 1029–1042. [DOI]

[29] Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open reading frames in drosophila. *Genome Biol*, 2011, 12(11): R118. [DOI]

[30] Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet*, 2014, 15(3): 193–204. [DOI]

[31] Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*, 2013,

9(1): 59–64. [DOI]

[32] Dufresne SS, Dumont NA, Boulanger-Piette A, Fajardo VA, Gamu D, Kake-Guena SA, David RO, Bouchard P, Lavergne É, Penninger JM, Pape PC, Tupling AR, Frenette J. Muscle RANK is a key regulator of Ca^{2+} storage, SERCA activity, and function of fast-twitch skeletal muscles. *Am J Physiol Cell Physiol*, 2016, 310(8): C663–C672. [DOI]

[33] Gusic M, Prokisch H. ncRNAs: new players in mitochondrial health and disease? *Front Genet*, 2020, 11: 95. [DOI]

[34] Krauss RS, Joseph GA, Goel AJ. Keep your friends close: cell-cell contact and skeletal myogenesis. *Cold Spring Harb Perspect Biol*, 2017, 9(2): a029298. [DOI]

[35] Millay DP, Gamage DG, Quinn ME, Min YL, Mitani Y, Bassel-Duby R, Olson EN. Structure-function analysis of myomaker domains required for myoblast fusion. *Proc Natl Acad Sci USA*, 2016, 113(8): 2116–2121. [DOI]

[36] Read C, Nyimanu D, Williams TL, Huggins DJ, Sulentic P, Macrae RGC, Yang PR, Glen RC, Maguire JJ, Davenport AP. International union of basic and clinical pharmacology. CVII. structure and pharmacology of the Apelin receptor with a recommendation that Elabela/Toddler is a second endogenous peptide ligand. *Pharmacol Rev*, 2019, 71(4): 467–502. [DOI]

[37] Walter P, Ron D. The unfolded protein response: from stress pathway to homeostatic regulation. *Science*, 2011, 334(6059): 1081–1086. [DOI]

[38] D'Lima NG, Ma J, Winkler L, Chu Q, Loh KH, Corpuz EO, Budnik BA, Lykke-Andersen J, Saghatelian A, Slavoff SA. A human microprotein that interacts with the mRNA decapping complex. *Nat Chem Biol*, 2017, 13(2): 174–180. [DOI]

[39] Makarewich CA, Olson EN. Mining for micropeptides. *Trends Cell Biol*, 2017, 27(9): 685–696. [DOI]

[40] Li X, Li MW, Zhang YN, Xu HM. Common cancer genetic analysis methods and application study based on TCGA database. *Hereditas(Beijing)*, 2019, 41(3): 234–242.
李鑫, 李梦玮, 张依楠, 徐寒梅. 常用肿瘤基因分析方法及基于 TCGA 数据库的分析应用. 遗传, 2019, 41(3): 234–242. [DOI]

[41] Wang LG, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res*, 2013, 41(6): e74. [DOI]

[42] Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu SH. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics*, 2010, 26(3): 399–400. [DOI]

[43] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 2011, 27(13): i275–i282. [DOI]

[44] Skarshewski A, Stanton-Cook M, Huber T, Al Mansoori S, Smith R, Beatson SA, Rothnagel JA. uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics*, 2014, 15: 36. [DOI]

[45] Niu LM, Lou FZ, Sun Y, Sun LB, Cai XJ, Liu ZY, Zhou H, Wang H, Wang ZK, Bai J, Yin QQ, Zhang JX, Chen LJ, Peng DH, Xu ZY, Gao YY, Tang SB, Fan L, Wang HL. A micropeptide encoded by lncRNA *MIR155HG* suppresses autoimmune inflammation *via* modulating antigen presentation. *Sci Adv*, 2020, 6(21): eaaz2059. [DOI]

[46] Ingolia NT. Ribosome footprint profiling of translation throughout the genome. *Cell*, 2016, 165(1): 22–33. [DOI]

[47] Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol*, 2011, 3: 1245–1252. [DOI]

[48] Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso JP. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife*, 2014, 3: e03528. [DOI]

[49] Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, 2013, 154(1): 240–251. [DOI]

[50] Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*, 2016, 537(7620): 347–355. [DOI]

[51] Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: a growing field. *Mass Spectrom Rev*, 2017, 36(5): 584–599. [DOI]

[52] Housman G, Ulitsky I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim Biophys Acta*, 2016, 1859(1): 31–40. [DOI]

[53] Yang J, Meng XD, Pan JC, Jiang N, Zhou CW, Wu ZH, Gong ZH. CRISPR/Cas9-mediated noncoding RNA editing in human cancers. *RNA Biol*, 2018, 15(1): 35–43. [DOI]

(责任编委: 薛宇)