基于“人造精子细胞”的基因组标签计划

常柏然, 李劲松*

中国科学院分子细胞科学卓越创新中心(生物化学与细胞生物学研究所), 上海 200031

* 联系人, E-mail: jsli@sibcb.ac.cn

收稿日期: 2020-01-13; 接受日期: 2020-02-19; 网络版发表日期: 2020-04-22

中国科学院战略性先导科技专项(批准号: XDB19010204)和上海市科学技术委员会科技创新行动计划项目(批准号: 17411954900, 19411951800)资助

摘要 蛋白质是一切生命活动的基础, 蛋白质研究特别是生命活动中蛋白质功能网络调控的揭示成为继人类基因组计划后广受关注的重要科学问题。然而, 大规模开展蛋白质研究缺乏一个高效、标准的平台, 构建基因组范围的蛋白质标签模式生物资源平台是解决这一问题的关键。虽然在酵母、线虫、果蝇等模式生物上已经实现了基因组范围的蛋白质标签文库的构建, 但是在小鼠个体水平上, 基因组范围的蛋白质标签的构建仍然困难重重。基因组标签计划(Genome tagging project, GTP)的提出是基于孤雄单倍体胚胎干细胞(又称为“人造精子细胞”)介导的半克隆技术和CRISPR-Cas9基因编辑技术的发展: 首先在“人造精子细胞”上利用CRISPR-Cas9技术对特定编码蛋白质基因进行标签序列的原位插入获得相应的细胞系, 进一步通过卵子注射即能获得携带标签的小鼠模型。因此, 该策略简单、高效、成本低, 可以用于基因组范围制备蛋白质标签小鼠。GTP的目标是给小鼠中的每个蛋白质带上标签, GTP的实施有望实现以一套标准化的方法来对每个蛋白质进行在体研究, 将促进生命科学的发展。

关键词 基因组标签计划, 蛋白质标签, “人造精子细胞”, 半克隆技术, CRISPR-Cas9

1990年启动, 耗时14年由6个国家合作完成的人类基因组计划(Human genome project, HGP)细致地绘制了人类的基因图谱, 提供了近30亿个碱基对和约2万~2.5万个蛋白质编码基因的信息^[1,2]。HGP的实施和完成不仅为生命科学研究提供了详细的遗传信息, 而且也基因组测序技术的发展带来了机遇, 使其在随后的岁月中得到了迅猛的发展, 极大降低了DNA测序的成本。HGP的完成标志着后基因组时代的到来, 而后基因组时代一项主要的任务就是要对蛋白质编码基因进行注释, 研究蛋白质的动态定位和在体功能。

随着生物信息学的发展, 人们挖掘出了越来越精

确的蛋白质编码基因的数量。在最新的一项研究中, Pertea等人^[3]提出, 存在20352个编码蛋白质的基因。同时, 随着测序技术和蛋白质组学技术的发展, 人们已经可以在RNA水平和蛋白质水平上形成基因转录和翻译的全局视图^[4-6]。然而, 要精准描述体内单个蛋白质的分子特性, 例如表达动力学、亚细胞定位和相互作用网络等, 仍非常困难。这其中一个主要的原因就是蛋白质研究依赖于抗体, 而抗体的制备存在种种问题, 如抗体制备成本高、难度大, 有很多蛋白质缺乏有效可靠的抗体; 不少蛋白质的抗体只能开展体外实验而不能用于体内实验; 有些蛋白质抗体可以用于生化和

引用格式: 常柏然, 李劲松. 基于“人造精子细胞”的基因组标签计划. 中国科学: 生命科学, 2020, 50: 549–558

Chang B R, Li J S. “Artificial spermatid”-based Genome tagging project (in Chinese). Sci Sin Vitae, 2020, 50: 549–558, doi: 10.1360/SSV-2020-0071

组化分析, 却不能用于染色质免疫共沉淀技术(chromatin immunoprecipitation sequencing, ChIP-seq)研究等. 没有抗体特异指示蛋白质, 许多实验手段就无法实现蛋白质的在体研究. 这些问题导致的最终结果是, 截至目前, 仍有超过50%的蛋白质从未被研究过^[7]. 另外, 不同蛋白质抗体各不相同, 导致蛋白质之间的比较研究、蛋白质相互作用网络研究、蛋白质与核酸的相互作用研究等受到了极大的限制. 这使得目前的研究模式常常是以单个或少数几个蛋白质为核心开展的, 往往难以描绘复杂生命过程的全部.

当然, 针对抗体问题也存在解决办法, 即标签技术. 蛋白质标签技术是指在目标蛋白质编码序列原位进行遗传标记, 带上一段编码标签蛋白质或多肽的DNA序列, 使其与目标蛋白质共同转录翻译产生融合蛋白质. 通过这一策略, 人们可以利用通用的标签蛋白质的抗体特异性识别标签, 从而实现对目标蛋白质的识别和研究. 目前, 常用的标签包括可用于在体观察的荧光蛋白质和用于研究目标蛋白质与生物大分子相互作用的亲和性标签. 早在1994年, Chalfie等人^[8]就将从小维多利亚水母(*Aequorea victoria*)中提取的绿色荧光蛋白质(green fluorescent protein, GFP)基因序列与目标蛋白质序列进行融合, 实现了目标蛋白质的在体可视化, 从而在活细胞中监测蛋白质的表达量以及蛋白质的亚细胞定位. 2年后, Cormack等人^[9]对GFP进行了遗传突变, 筛选出荧光亮度更强的GFP变种(enhanced green fluorescent protein, EGFP). EGFP目前已被广泛应用于活细胞蛋白质成像, 可用于蛋白质在体、实时、动态地示踪, 在生物学研究中发挥着重要的作用. 另外, 除EGFP外, 许多亮度更强的荧光蛋白质也不断涌现出来, 如mNeonGreen^[10]和mClover3^[11]. 然而, 荧光蛋白质这一类标签也存在缺陷, 如编码序列较长(如GFP的编码序列为714 bp), 给目标蛋白质基因原位插入带来了困难; 另外, 荧光蛋白质体积普遍比较大(例如所有GFP变体的蛋白分子量都在27 kD左右), 可能会影响目标蛋白质的结构、亚细胞定位和功能. 另一类亲和性标签通常比较小, 便于原位插入并可减小其对目标蛋白质结构和功能的影响, 如仅有9个氨基酸的HA标签^[12]. HA标签来自人类流感血凝素(hemagglutinin, HA), 对应于98~106位氨基酸, 是一种强免疫反应性表位. HA抗体制备相对简单, 且抗体特异性强, 可广泛用于目标蛋白质的分离、纯化、检测和跟踪.

在过去20多年中, 科学家使用标签技术解决了一部分蛋白质没有合适抗体的研究困境. 同时, 随着标签技术的不断发展和完善, 科学家提出设想: 如果在一个特定的模式生物中为所有编码蛋白质的基因都带上统一的标签, 如HA, 构建相应的标签文库, 就可以用同一种HA抗体来对所有蛋白质开展在体研究, 不但可以简化蛋白质研究的程序, 而且有望实现蛋白质研究的精准、标准化和高通量, 从而推动生命科学的发展.

1 基因组范围的蛋白质标签文库构建仅限于在低等模式生物中开展

(1) 酵母的基因组标签研究. 在模式生物里开展基因组范围的蛋白质标签研究已有多年的历史. 早期的基因原位插入外源DNA序列取决于DNA同源重组技术. 由于酵母的同源重组效率非常高, 且因其是单倍体的单细胞模式生物, 酵母也首先被用来进行基因组范围的蛋白质标记. 2002年, Gavin等人^[13]在单细胞酿酒酵母(*Saccharomyces cerevisiae*)蛋白质基因的3'端插入了串联亲和纯化标签(tandem-affinity purification TAP)(图1A)构建了包含1548株蛋白质标签酵母的文库; 进而利用TAP标签的抗体拉出蛋白质复合体进行质谱, 系统地揭示了酵母中的蛋白质复合体, 对酵母的蛋白质复合体进行了网络式的阐述. 这是首次利用基于同一抗体的技术在同一细胞类型中开展大规模内源蛋白质复合体研究的案例. 2003年, Huh等人^[14]构建了6029株蛋白质C端原位带有GFP标记的酵母(图1B), 其中4156株有明显的GFP信号, 因此可以通过活细胞中直接观测荧光信号来对蛋白质进行定位研究, 构建酵母中蛋白质的动态网络图谱. 同年, Ghaemmaghami等人^[15]在全基因组范围对酵母蛋白质编码基因C端插入TAP-tag, 标记了98%酵母基因组中的基因可读框, 对蛋白质的表达图谱进行了系统的分析. 2006年, Gavin等人^[16]和Krogan等人^[17]在酵母中对全基因组范围的基因可读框3'端插入TAP标签建立相应带有标签的酵母, 利用TAP的抗体并通过质谱分析纯化出蛋白质复合体, 系统性描绘了酵母蛋白质的相互作用网络.

(2) 线虫和果蝇的基因组标签研究. 在酵母上进行的这些基于标签技术的蛋白质表达、定位、相互作用网络研究主要得益于酵母高效的同源重组特性. 然而, 由于其他模式生物中的同源重组效率非常低, 因此不

图 1 酵母和果蝇中进行基因组标签研究的策略. A: 在酵母中, 使用同源重组方法将TAP标签插入至目标蛋白质的C端, 生成TAP标记的融合蛋白; B: 在酵母中, 使用同源重组方法将绿色荧光蛋白(GFP)插入至目标蛋白质的C端, 生成GFP标记的融合蛋白; C: 线虫中TransgeneOmics方法示意图. 将线虫的基因组切割成线性化的片段, 并连入BAC的载体中, 构成线虫的基因组文库, 之后对基因组文库进行同源重组, 在目标蛋白质的C端插入EGFP和flag, 最后将改造好的线虫基因组文库转入线虫体内, 生成表达标签蛋白质的线虫(网络版彩图)

Figure 1 Strategies for genome-wide tagging in yeast and *Drosophila*. A: In yeast, the TAP tag is inserted into the C-terminus of the target protein using a homologous recombination method to generate a TAP-tagged fusion protein; B: in yeast, the green fluorescent protein (GFP) is inserted into the C-terminus of the target protein using homologous recombination to generate a GFP-tagged fusion protein; C: the diagram of TransgeneOmics method for genome-wide tagging in *C. elegans*. The genome of the *C. elegans* is digested into linearized fragments and ligated into the vector of BAC to form the genomic BAC library of *C. elegans*. EGFP and flag are then inserted into the BAC library at the C-terminus of the target protein using homologous recombination. Finally, the tagged BAC library is transferred into *C. elegans* to derive transgenic worms expressing EGFP-tagged fusion protein (color online)

可能进行基于同源重组技术的标签文库的构建. 为了解决这一问题, 科学家们另辟蹊径, 利用细菌人工染色体(bacterial artificial chromosome, BAC)技术在其他模式生物上开展了基因组范围的标签文库构建. 1992年, Shizuya等人^[18]首次构建了细菌人工染色体载体, 该载体以大肠杆菌致育因子(F因子)为基础, 具有携带1 Mb插入DNA片段的潜能, 足以包含一个特定基因的完整序列, 包括其内在的调控序列. 因此, BAC载体被大规模地用来构建各种真核生物的基因组文库, 使得通过BAC转基因技术来研究基因组范围的蛋白质功能成为可能. 早在1997年, Janke等人^[19]构建出了秀丽隐杆线虫(*C. elegans*)的BAC转基因文库. 2012年, Sarov等人^[20]在此基础上对BAC文库中的插入DNA进行修改, 在蛋白质编码基因3'端插入GFP蛋白和3*Flag序列(图1C)(BAC TransgeneOmics系统), 并将这些基因注入线虫体内获得转基因线虫, 构建了基因组范围的标签蛋白质转基因线虫文库. 利用相似的转基因方法, 2017年, Reinke等人^[21]构建了3000株不同蛋白质带有

抗坏血酸过氧化物酶(ascorbate peroxidase, APX)标签的线虫品系, 并系统地研究了线虫的蛋白质互作网络. 此外, 2016年Sarov等人^[22]利用BAC TransgeneOmics系统构建了果蝇(*Drosophila*)的转基因蛋白质标签文库, 包含将近10000个蛋白质标记GFP的果蝇品系; 利用这一资源, 他们详细研究了果蝇体内蛋白质的定位、互作以及功能.

2 CRISPR-Cas9基因编辑技术和“人造精子细胞”技术使得构建小鼠基因组范围的蛋白质标签文库成为可能

尽管人们在低等模式生物中已取得了一些基因组范围蛋白质标签研究的进展, 但是在哺乳动物个体水平上用这种高通量的方法来标记蛋白质编码基因仍是非常困难的, 目前哺乳动物中大规模的蛋白质标记仍被限制在细胞层面. 2008年, Poser等人^[23]利用BAC TransgeneOmics系统在HeLa细胞中进行了基因组标签

文库的构建,并在小鼠胚胎干细胞中开展了蛋白质标签的初步尝试。然而,转基因的随机插入会导致部分转基因无法准确模拟蛋白质的生理表达状况。解决转基因带来的问题,可以通过基因打靶的策略在胚胎干细胞上内源性地标记蛋白质编码基因,再将这些细胞注入囊胚中以构建出嵌合体小鼠,进而通过嵌合小鼠生殖细胞将标签基因传递给下一代,获得标签小鼠^[24,25]。但是,嵌合体的形成和生殖细胞传递是该策略的限速步骤,往往需要1年左右时间构建一个标签小鼠,因而极大地阻碍了其大规模的运用。因此,构建基因组范围的标签小鼠文库仍是亟待解决的需求,是生物科学研究领域上的一大挑战。近年来,clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (CRISPR-Cas9)基因编辑技术和“人造精子细胞”介导的半克隆技术这两项生物技术的快速发展,使得在小鼠中开展基因组范围蛋白质标签研究成为可能。

(1) CRISPR-Cas9基因编辑技术。基因编辑技术是研究基因功能的关键技术,从早期的同源重组技术到现在蓬勃发展的人工核酸酶技术,它一直在推动着生物学的发展。近20年来,人工核酸酶介导的基因编辑技术日渐成熟,极大地改善了早期基于同源重组的基因打靶技术效率低的问题,使得基因编辑越来越容易。人工核酸酶由DNA结合域和核酸内切酶组成,它主要包括3种类型:锌指核酸酶(zinc finger nucleases, ZFN),转录激活因子样效应物核酸酶(transcription activator-like effector nucleases, TALEN)及CRISPR-Cas9技术^[26,27]。锌指核酸酶是由锌指蛋白质结构域和Fok I切割结构域人工融合而成,锌指蛋白质结构域能够识别并结合特异的DNA序列,Fok I在二聚化后产生核酸内切酶活性,造成双链断裂^[28]。TALEN技术是基于TALE蛋白质的一种技术,TALE蛋白质家族来自一类特殊的植物病原体——黄单胞菌属(*Xanthomonas*),可以识别并结合特异的DNA序列。基于这一特性,2011年,Li等人^[29]和Mahfouz等人^[30]分别改造了天然的TALE蛋白,将C端的转录激活因子替换为Fok I,将DNA靶向与DNA切割结合起来,实现了对基因组的定向编辑。TALEN和ZFN这两项技术实际操作技术难度大,因此后来逐渐被从原核生物II型CRISPR/Cas系统成功开发的CRISPR-Cas9系统所取代。Cas9核酸内切酶可与人工设计的可以识别特异DNA序列的single-

guide RNA(sgRNA)结合形成核糖核蛋白,并在sgRNA的引导下靶向基因组的特定序列。Cas9-sgRNA与基因组目标序列形成三元复合物,Cas9可以切割目标DNA序列的磷酸酯键形成双链断裂损伤(double strand breaks, DSBs),细胞通过非同源重组末端直接连接(non-homologous end joining, NHEJ)和同源重组修复(homology directed repair, HDR)机制修复所产生的双链断裂^[31]。NHEJ修复途径是在相关蛋白的介导下直接连接修复,在修复过程中会产生插入缺失突变,这一点也被用于进行蛋白质编码基因敲除^[32]。HDR修复主要是以内源或者外源供体DNA为模板对DSBs所在的序列进行修复,使用人工设计的单链寡聚核苷酸或者载体质粒可以对靶向序列进行辅助大片段敲除^[33]、构建点突变^[34,35]或者DNA序列的敲入^[36]。因此,人们可以非常方便地利用CRISPR-Cas9对细胞和胚胎的基因组进行编辑,构建出带有突变基因或者原位敲入外源标签序列的细胞或者胚胎^[34,37]。

(2) “人造精子细胞”介导的半克隆技术。“人造精子细胞”,又称为孤雄单倍体胚胎干细胞,是从孤雄单倍体囊胚中建立的单倍体胚胎干细胞系^[38,39]。孤雄单倍体胚胎是通过将精子注入到去核的卵母细胞中或在受精卵的原核期将雌原核去除这两种方法构建出来的,其可以在体外发育成囊胚,从中建立胚胎干细胞。由于单倍体细胞会发生自发二倍体化,因此需要通过流式细胞分选技术将单倍体细胞富集出来。之后,通过定期流式分选富集单倍体细胞的方法,就可以维持细胞的单倍体特性^[38-41]。有意思的是,这些携带精子遗传物质的单倍体细胞在注入到成熟的卵母细胞中后,能够支持重构胚胎的发育,并产生健康的小鼠。由于这些小鼠一半的遗传物质都是来源于注射的单倍体细胞(完全相同),而另一半遗传物质来源于不同的卵母细胞(各不相同),所以这类小鼠被称为半克隆小鼠(semi-cloned mice, SC mice)^[38]。相应地,孤雄单倍体胚胎干细胞注射入卵子的技术(intracytoplasmic AG-haESC injection, ICAHCI)也被称为半克隆技术(semi-cloning technology, SC technology)。然而,半克隆小鼠胚胎发育效率非常低,只有2%左右能发育成健康的小鼠,极大地限制了这一技术的应用。后续的研究显示,2个父源印记基因(即在父源基因组中不表达的基因)——*H19*和*Gtl2*,在孤雄单倍体胚胎干细胞和发育失败的半克隆胚胎中异常地高表达。究其原因,是因

为孤雄单倍体胚胎干细胞*H19*-DMR(differentially DNA methylated region, DMR)(调控*H19*表达的区域)和*IG*-DMR(调控*Gtl2*表达的区域)的甲基化在培养环境下逐渐丢失,所以在将这两个差异甲基化区域敲除以模拟其甲基化的状态后,半克隆小鼠的出生率提高至20%^[42,43]。重要的是,*H19*-DMR和*IG*-DMR敲除的孤雄单倍体胚胎干细胞在体外经过多轮基因编辑后,仍然可以高效支持半克隆胚胎的发育,这显示半克隆技术可以作为一个高效构建基因修饰小鼠的工具^[44-48]。因此,我们将携带*H19*-DMR和*IG*-DMR敲除的孤雄单倍体胚胎干细胞称为“人造精子细胞”。

(3) 基于“人造精子细胞”的基因组标签计划(Genome tagging project, GTP)。随着CRISPR-Cas9这一高效编辑工具的出现,涌现出了很多在早期胚胎中应用CRISPR-Cas9系统来一步法构建基因修饰动物模型的方法,如原核时期胞浆注射^[37,49]和2-细胞时期胞浆注射^[50-52],这极大地丰富了基因修饰小鼠的构建方法。然而,这类方法操作虽然相对简单,但是存在以下弊端。首先,这类方法只有在获得小鼠个体后才能进行基因型分析确定是否产生了需要的基因型,因此需要饲养更多的小鼠,这大大增加了获得基因编辑小鼠的成本。其次,虽然CRISPR-Cas9基因编辑效率高,但在受精卵阶段注射难以避免其在随后的发育阶段发挥基

因编辑功能,产生基因型嵌合的个体^[49,53],从而大大增加了基因型鉴定的复杂程度。第三,这种嵌合有可能存在于生殖细胞中,导致F1小鼠基因型的不确定性。第四,这类方法只适合相对简单的基因编辑,对于复杂的基因编辑,其效率则大大降低。因此,利用CRISPR-Cas9受精卵直接注射的方法开展大规模的标签小鼠制备,会极大增加操作的成本。相比之下,“人造精子细胞”介导的半克隆技术与CRISPR-Cas9技术结合,可以将遗传操作以一种可控且非随机的方式传递到所有子代中,降低了后续对基因型鉴定的要求^[35,54]。而且半克隆技术可以胜任多基因操作等复杂基因编辑的工作,只需对孤雄单倍体胚胎干细胞进行多基因的编辑,就可以一步得到需要的基因型小鼠^[55-57]。

正是基于“人造精子细胞”介导的半克隆技术和CRISPR-Cas9基因编辑技术,人们才可以在相对低成本的情况下,在一定的周期内实现基因组范围的蛋白质编码基因的标记,获得基因组范围的蛋白质标签小鼠文库^[58](图2)。基于“人造精子细胞”的基因组标签计划具有以下优势。首先,在“人造精子细胞”进行基因原位插入的编辑效率高,且可以在体外开展细胞系的基因型鉴定分析,因此所得到的标签半克隆小鼠的基因型是固定的。其次,标签序列的插入可能会对蛋白质的结构和功能产生影响,由于很多蛋白质在胚胎

图 2 基因组标签计划的示意图。利用“人造精子细胞”介导的半克隆技术和Cas9介导的基因编辑技术构建带有标签蛋白质的“人造精子细胞”,经筛选鉴定后,将“人造精子细胞”注入MII卵子中,并在2-cell时期移入母鼠输卵管内,发育成带有标签蛋白质的半克隆小鼠。人们可以利用这项资源,在细胞水平和在体水平上进行一系列的研究(网络版彩图)

Figure 2 The diagram of Genome tagging project (GTP) in mice. The “artificial spermatid”-mediated semi-cloning (SC) technology and CRISPR-Cas9-mediated genome editing technology are used to construct “artificial spermatids” with a endogenously tagged protein in liquid culture. The tagged “artificial spermatids” can be cryopreserved in liquid nitrogen and/or directly injected into MII oocytes to produce tagged semi-cloned mice in one step. The tagged “artificial spermatids” and mice can be used as the GTP platform for standard protein analyses through tag antibody-based assays (color online)

干细胞系中均有表达, 因此可以在“人造精子细胞”中开展标签插入位点的选择研究, 进而确保产生的小鼠蛋白质功能不受到影响。最后, “人造精子细胞”可以在体外维持, 甚至进行冷冻保存, 需要的时候可通过解冻复苏和卵子注射产生标签小鼠。总的来说, 先构建携带标签的“人造精子细胞”库, 再通过卵子注射一步产生标签小鼠的这一策略适合大规模开展蛋白质标签小鼠的制备, 这一策略使得我们可以在细胞水平上解决很多问题, 既节省了花费和精力, 也可以最大程度地保证小鼠中蛋白质标记的正确性。

3 GTP的提出和现状

2017年5月, 基因组标签计划这一想法在中国科学院分子细胞科学卓越创新中心(生物化学与细胞生物学研究所, SIBCB)举办的一场头脑风暴会议上被第一次提出, 引起了与会专家的共鸣。在这之后, SIBCB就一直大力推动基因组标签计划的实施和发展, 通过召开多场学术研讨会来探讨基因组标签计划的可行性与重要性, 并通过中国科学院战略性先导科技专项(B类)“细胞命运可塑性的分子基础与调控”(项目号XDB19010204)首席调控费拨付200万元用于支持开展50个蛋白质编码基因标签的尝试。上海市科学技术委员会对基因组标签计划高度重视, 在听取相关介绍后, 于2017年底通过启动创新行动计划项目《基于“人造精子”技术的重要基因标签研究》(项目号17411954900, 执行期2017.12~2019.6, 资助经费: 2000万), 对GTP项目给予初步资金支持, 以期在试点项目中完成500个重要蛋白质的标签细胞系和100个标签小鼠的制备。SIBCB为了更好地推动GTP的实施和项目的执行, 成立了GTP研发中心, 以探索建立相应的管理和运行规则。项目执行期间, GTP研发中心建立了一套实验标准化、流程化和信息化管理体系, 保障标签产品的标准化制备; 制定了一系列规章制度、培养了一批技术骨干; 建立中心网站(<http://www.sibcb.ac.cn/gtp/pro/index.jsp>)、知识产权保护条例、资源共享机制, 促进标签产品的全球共享。另外, 项目执行期内, GTP研发中心围绕标签技术及GTP在生物医学中的应用组织了多次小规模的大脑风暴, 为GTP的推动和实施集聚智慧。2019年5月23日, GTP研发中心在沪举办以“基因组标签计划(GTP)”为专题的香山科

学会议(第S47次学术研讨会), 来自中国科学院、国内高校、医院、上海市科学技术委员会和科技媒体等31家单位的50多位专家、学者出席了会议。会议围绕“GTP与基础生物学”和“GTP与精准医学”两个中心议题, 进行了广泛而深入的讨论, 就GTP实施的可行性、重要性以及潜在的应用前景达成了共识。

“人造精子细胞”可以用于携带不同的标签序列以开展不同目的的科学研究, 在第一个版本的基因组范围编码蛋白质基因标签制备中(GTP1.0), 基于以下3方面的考虑, GTP研发中心优先选择HA(人流感血凝素肽段, 9个氨基酸)标签: (i) HA序列短, 对蛋白质的分布和生物活性造成影响的可能性小; (ii) 短的DNA序列有利于基因原位插入; (iii) 基于HA标签的各种蛋白质研究方法成熟。另外, 由于蛋白质结构各异, 而其功能的发挥与其结构密不可分, 因此需要仔细分析标签序列的插入位置(N端、C端或者中间特定位点)以避免对蛋白质结构和功能产生影响。在GTP1.0中, 我们首先选择在C端插入, 若是C端不合适的话, 再对N端进行尝试。同时, 大部分蛋白质存在多个剪切体, 我们首先选择在最长的剪切体上进行标记。再者, CRISPR-Cas9可能会引入脱靶突变, 所以我们在设计sgRNA时会谨慎选择位点降低脱靶的可能性。基于以上的构建原则, 截至2019年7月1日, GTP研发中心制备了614个标签细胞系和141个标签小鼠品系; 通过基于HA抗体的Western Blotting分析发现, 在300余细胞系中能够检测到蛋白质的表达; 标签细胞系已提供给国内外87个研究组、标签小鼠已提供给国内外22个研究组; 一批利用过GTP研发中心产品或者服务的研究成果已经发表(标注课题号或者致谢GTP研发中心)^[43,54,55,57,59-67]。

4 GTP的展望

GTP研发中心顺利完成了GTP预研项目, 既证明“人造精子细胞”介导GTP的可行性, 又为进一步大规模开展标签产品的制备奠定了基础, 储备了技术和人才。在一定经费的支持下, 我们有信心在合理的时间范围内实现基因组范围蛋白质标签“人造精子细胞”系和小鼠的构建。随着GTP平台的建设和发展, 我们可以应用基于标签抗体(如HA抗体)的标准化实验方法进行特定蛋白质的在体、实时、动态研究, 不仅可以

解决不少蛋白质缺乏有效抗体导致的研究困境, 也为不同蛋白质之间的比较研究提供了标准的模式细胞和小鼠资源, 促进蛋白质研究的标准化. 同时, 围绕一个特定生物过程, 我们可以利用系列相关蛋白质的标签细胞或小鼠开展基于群体蛋白质的蛋白质表达图谱、细胞定位、调控网络等研究(蛋白质伙伴组, **protein patternome**). 这一研究方式将克服传统基于一个或几个蛋白质研究方式带来的局部性和片面性, 可以从更

全面的视角了解生命过程, 有望产生新的发现. 例如, 基于GTP平台, 我们可以开展胚胎发育的精准蛋白质图谱研究(**precision protein atlas project, PPAR**). 同时, 我们相信, 在针对所有蛋白质的标签细胞和小鼠构建尝试过程中, 一定会催生新的生物技术和研究方法, 如灵敏、可控的标签技术, 蛋白质研究新方法等, 而这些技术的产生也必将反过来促进生物医学研究发展的进程.

参考文献

- Collins F S, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*, 2003, 300: 286–290
- International Human Genome Sequencing C. CFInishing the euchromatic sequence of the human genome. *Nature*, 2004, 431: 931–945
- Pertea M, Shumate A, Pertea G, et al. CHES: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*, 2018, 19: 208
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10: 57–63
- Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, 2009, 6: 377–382
- Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*, 2016, 537: 347–355
- Kusnezow W, Hoheisel J D. Antibody microarrays: promises and problems. *BioTechniques*, 2002, 33: S14–S23
- Chalfie M, Tu Y, Euskirchen G, et al. Green fluorescent protein as a marker for gene expression. *Science*, 1994, 263: 802–805
- Cormack B P, Valdivia R H, Falkow S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 1996, 173: 33–38
- Shaner N C, Lambert G G, Chamma A, et al. A bright monomeric green fluorescent protein derived from *Branchiostoma lanceolatum*. *Nat Methods*, 2013, 10: 407–409
- Bajar B T, Wang E S, Lam A J, et al. Improving brightness and photostability of green and red fluorescent proteins for live cell imaging and FRET reporting. *Sci Rep*, 2016, 6: 20889
- Field J, Nikawa J, Broek D, et al. Purification of a RAS-responsive adenylyl cyclase complex from *Saccharomyces cerevisiae* by use of an epitope addition method. *Mol Cell Biol*, 1988, 8: 2159–2165
- Gavin A C, Bösch M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 2002, 415: 141–147
- Huh W K, Falvo J V, Gerke L C, et al. Global analysis of protein localization in budding yeast. *Nature*, 2003, 425: 686–691
- Ghaemmaghami S, Huh W K, Bower K, et al. Global analysis of protein expression in yeast. *Nature*, 2003, 425: 737–741
- Gavin A C, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006, 440: 631–636
- Krogan N J, Cagney G, Yu H, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 2006, 440: 637–643
- Shizuya H, Birren B, Kim U J, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA*, 1992, 89: 8794–8797
- Janke D L, Schein J E, Ha T, et al. Interpreting a sequenced genome: toward a cosmid transgenic library of *Caenorhabditis elegans*. *Genome Res*, 1997, 7: 974–985
- Sarov M, Murray J I, Schanze K, et al. A genome-scale resource for *in vivo* tag-based protein function exploration in *C. elegans*. *Cell*, 2012, 150: 855–866
- Reinke A W, Mak R, Troemel E R, et al. *In vivo* mapping of tissue- and subcellular-specific proteomes in *Caenorhabditis elegans*. *Sci Adv*, 2017, 3: e1602426
- Sarov M, Barz C, Jambor H, et al. A genome-wide resource for the analysis of protein localisation in *Drosophila*. *eLife*, 2016, 5: e12068
- Poser I, Sarov M, Hutchins J R A, et al. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat Methods*, 2008, 5: 409–415

- 24 Capecchi M R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. [Nat Rev Genet](#), 2005, 6: 507–512
- 25 Bouabe H, Okkenhaug K. Gene targeting in mice: a review. In: Bailer S, Lieber D, eds. *Virus-Host Interactions. Methods in Molecular Biology (Methods and Protocols)*. Totowa: Humana Press, 2013. 315–336
- 26 Yan M, Li J. The evolving CRISPR technology. [Protein Cell](#), 2019, 10: 783–786
- 27 Zhang X, Wang L, Liu M, et al. CRISPR/Cas9 system: a powerful technology for *in vivo* and *ex vivo* gene therapy. [Sci China Life Sci](#), 2017, 60: 468–475
- 28 Miller J C, Holmes M C, Wang J, et al. An improved zinc-finger nuclease architecture for highly specific genome editing. [Nat Biotechnol](#), 2007, 25: 778–785
- 29 Li T, Huang S, Jiang W Z, et al. TAL nucleases (TALENs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. [Nucleic Acids Res](#), 2011, 39: 359–372
- 30 Mahfouz M M, Li L, Shamimuzzaman M, et al. *De novo*-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. [Proc Natl Acad Sci USA](#), 2011, 108: 2623–2628
- 31 Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. [Mol Cell Biol](#), 1994, 14: 8096–8106
- 32 Heidenreich E. Non-homologous end joining as an important mutagenic process in cell cycle-arrested cells. [EMBO J](#), 2003, 22: 2274–2283
- 33 Kato T, Hara S, Goto Y, et al. Creation of mutant mice with megabase-sized deletions containing custom-designed breakpoints by means of the CRISPR/Cas9 system. [Sci Rep](#), 2017, 7: 59
- 34 Wu Y, Liang D, Wang Y, et al. Correction of a genetic disease in mouse via use of CRISPR-Cas9. [Cell Stem Cell](#), 2013, 13: 659–662
- 35 Wei L, Wang X, Yang S, et al. Efficient generation of the mouse model with a defined point mutation through haploid cell-mediated gene editing. [J Genets Genomics](#), 2017, 44: 461–463
- 36 Schumann K, Lin S, Boyer E, et al. Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. [Proc Natl Acad Sci USA](#), 2015, 112: 10437–10442
- 37 Yang H, Wang H, Shivalila C S, et al. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. [Cell](#), 2013, 154: 1370–1379
- 38 Yang H, Shi L, Wang B A, et al. Generation of genetically modified mice by oocyte injection of androgenetic haploid embryonic stem cells. [Cell](#), 2012, 149: 605–617
- 39 Li W, Shuai L, Wan H, et al. Androgenetic haploid embryonic stem cells produce live transgenic mice. [Nature](#), 2012, 490: 407–411
- 40 Leeb M, Wutz A. Derivation of haploid embryonic stem cells from mouse embryos. [Nature](#), 2011, 479: 131–134
- 41 Elling U, Taubenschmid J, Wirmsberger G, et al. Forward and reverse genetics through derivation of haploid mouse embryonic stem cells. [Cell Stem Cell](#), 2011, 9: 563–574
- 42 Zhong C, Yin Q, Xie Z, et al. CRISPR-Cas9-mediated genetic screening in mice with haploid embryonic stem cells carrying a guide RNA library. [Cell Stem Cell](#), 2015, 17: 221–232
- 43 Li Q, Li Y, Yin Q, et al. Temporal regulation of prenatal embryonic development by paternal imprinted loci. [Sci China Life Sci](#), 2020, 63: 1–17
- 44 Wang L, Li J. ‘Artificial spermatid’-mediated genome editing. [Biol Reprod](#), 2019, 101: 538–548
- 45 Li Y, Li J. Technical advances contribute to the study of genomic imprinting. [PLoS Genet](#), 2019, 15: e1008151
- 46 Bai M, Wu Y, Li J. Generation and application of mammalian haploid embryonic stem cells. [J Intern Med](#), 2016, 280: 236–245
- 47 Shi L, Yang H, Li J. Haploid embryonic stem cells: an ideal tool for mammalian genetic analyses. [Protein Cell](#), 2012, 3: 806–810
- 48 Ding Y F, Li J S, Zhou Q. Generation and applications of mammalian haploid embryonic stem cells (in Chinese). [Sci Sin Vitae](#), 2019, 49: 1635–1651 [丁一夫, 李劲松, 周琪. 哺乳动物单倍体胚胎干细胞的建立与应用. 中国科学: 生命科学, 2019, 49: 1635–1651]
- 49 Wang H, Yang H, Shivalila C S, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. [Cell](#), 2013, 153: 910–918
- 50 Gu B, Posfai E, Rossant J. Efficient generation of targeted large insertions by microinjection into two-cell-stage mouse embryos. [Nat Biotechnol](#), 2018, 36: 632–637
- 51 Wu Y, Zhang J, Peng B, et al. Generating viable mice with heritable embryonically lethal mutations using the CRISPR-Cas9 system in two-cell embryos. [Nat Commun](#), 2019, 10: 2883

- 52 Wang L, Li M Y, Qu C, et al. CRISPR-Cas9-mediated genome editing in one blastomere of two-cell embryos reveals a novel Tet3 function in regulating neocortical development. *Cell Res*, 2017, 27: 815–829
- 53 Niu Y, Shen B, Cui Y, et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, 2014, 156: 836–843
- 54 Wang X, Kang J Y, Wei L, et al. PHF7 is a novel histone H2A E3 ligase prior to histone-to-protamine exchange during spermiogenesis. *Development*, 2019, 146: dev175547
- 55 Wang L, Zhang Y, Fu X, et al. Joint utilization of genetic analysis and semi-cloning technology reveals a digenic etiology of Müllerian anomalies. *Cell Res*, 2020, 30: 91–94
- 56 Zhong C, Li J. Efficient generation of gene-modified mice by haploid embryonic stem cell-mediated semi-cloned technology. In: Reeves A, ed. *In Vitro Mutagenesis. Methods in Molecular Biology*. New York: Humana Press, 2017. 1498: 121–133
- 57 Yin Q, Wang H, Li N, et al. Dosage effect of multiple genes accounts for multisystem disorder of myotonic dystrophy type 1. *Cell Res*, 2020, 30: 133–145
- 58 Jiang J, Yan M, Li D, et al. Genome tagging project: tag every protein in mice through ‘artificial spermatids’. *Natl Sci Rev*, 2019, 6: 394–396
- 59 Meng X, Liu X, Guo X, et al. FBXO38 mediates PD-1 ubiquitination and regulates anti-tumour immunity of T cells. *Nature*, 2018, 564: 130–135
- 60 Qu C, Yan M, Yang S, et al. Haploid embryonic stem cells can be enriched and maintained by simple filtration. *J Biol Chem*, 2018, 293: 5230–5235
- 61 Li Q, Li Y, Yang S, et al. CRISPR-Cas9-mediated base-editing screening in mice identifies DND1 amino acids that are critical for primordial germ cell development. *Nat Cell Biol*, 2018, 20: 1315–1325
- 62 Lin C D, Zhang Y H, Zhang K, et al. Fever promotes T lymphocyte trafficking via a thermal sensory pathway involving heat shock protein 90 and $\alpha 4$ integrins. *Immunity*, 2019, 50: 137–151.e6
- 63 Liu H, Huang T, Li M, et al. SCRE serves as a unique synaptonemal complex fastener and is essential for progression of meiosis prophase I in mice. *Nucleic Acids Res*, 2019, 47: 5670–5683
- 64 Wang X, Liu R, Zhu W, et al. UDP-glucose accelerates SNAIL mRNA decay and impairs lung cancer metastasis. *Nature*, 2019, 571: 127–131
- 65 Bai M, Han Y, Wu Y, et al. Targeted genetic screening in mice through haploid embryonic stem cells identifies critical genes in bone development. *PLoS Biol*, 2019, 17: e3000350
- 66 Peng G, Suo S, Cui G, et al. Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature*, 2019, 572: 528–532
- 67 Yang X, Hu B, Liao J, et al. Distinct enhancer signatures in the mouse gastrula delineate progressive cell fate continuum during embryo development. *Cell Res*, 2019, 29: 911–926

“Artificial spermatid”-based Genome tagging project

CHANG BoRan & LI JinSong

CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China

Protein is the basis of all biological processes. Protein studies, especially the discovery of protein functional regulation network in life activities, have become an important scientific issue that has drawn increasing attention from scientists after the Human Genome Project. However, large-scale protein research needs an efficient and standard platform. Building a resource platform of model organism where every protein is tagged at organismal level is the key to solving this problem. However, although genome-wide protein labelling has been achieved in *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila*, genome-wide protein tagging in mice is still very challenging. Genome tagging project (GTP) in mice has been proposed based on two state-of-art biotechnologies, i.e., androgenetic haploid embryonic stem cell (so-called “artificial spermatid”)-mediated semi-cloning technology and CRISPR-Cas9-based genome-editing technology. GTP involves two steps: first, the “artificial spermatid” lines carrying a tagged protein are constructed *in vitro* by endogenously inserting a tag sequence into a specific protein-coding gene via CRISPR-Cas9; second, the tagged “artificial spermatids” can be injected into oocytes to obtain tagged mice in one step. This strategy by combining “artificial spermatids” and CRISPR-Cas9 is simple, efficient, and low-cost, and thus can be used to generate tagged mice at the genome-wide scale. The goal of GTP is to tag every protein in mice and, we believe, the implementation of GTP will promote to investigate the protein expression, localization and function *in vivo* through standardized tag antibody-based assays, which will greatly enhance the advances of life sciences.

Genome tagging project, protein tag, “artificial spermatids”, semi-cloning technology, CRISPR-Cas9

doi: [10.1360/SSV-2020-0071](https://engine.scichina.com/doi/10.1360/SSV-2020-0071)