DOI: 10.13376/j.cbls/2018140
文章编号：1004-0374(2018)11-1157-08

吴家睿，中国科学院上海生命科学研究院生物化学与细胞生物学研究所研究员。现任中科院系统生物学重点实验室主任，上海科技大学生命科学与技术学院执行院长，*Journal of Molecular Cell Biology* 主编，*BMC Systems Biology* 和 *Frontiers in Physiology*, *section: Systems Biology* 副主编，卫生部中国老年保健医学研究会副会长，中国生物化学与分子生物学会分子系统生物学专业委员会主任委员。实验室主要采用系统生物学方法研究糖尿病和肿瘤等重大慢性病发生与发展的分子机制。

人类细胞图谱计划面临的挑战

吴家睿
(中国科学院上海生命科学研究院生物化学与细胞生物学研究所，上海 200031)

摘　要：“人类细胞图谱”(Human Cell Atlas, HCA) 是生命科学领域最近兴起的国际大科学计划，其目标是要采用特定的分子标志物来确定人体的所有细胞类型。该计划的实施面临着巨大的挑战：有些挑战是技术性的，也许在未来能够解决；但是，有些挑战是理论性的，反映出研究者的想法与其研究对象的本性之间有着根本的冲突。

关键词：人类细胞图谱；细胞谱系；单细胞转录组测序
中图分类号：Q2　　**文献标志码：**A

Challenges for Human Cell Atlas

WU Jia-Rui
(Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200031, China)

Abstract: “Human Cell Atlas” (HCA) has been recently initiated as an international grand research project in the area of life science, with the goal to construct a comprehensive catalog of all human cells based on their molecular biomarkers. The HCA project will face grand challenges, among which some are technical and some are theoretical.
Key words: Human Cell Atlas; cell lineage; single-cell RNA-seq

动植物等多细胞生物，通常都是由一个细胞 —— 受精卵，通过连续不断的细胞分裂和细胞分化方式发育而成。在这种个体发育过程中，亲代与子代细胞之间以及同代细胞之间具有人类家族谱系那样的“亲缘关系”，故称为细胞谱系 (cell lineage)。对细胞谱系的研究不仅是发育生物学的主要任务，而且在抗击疾病和衰老等方面也有着重要的应用前景。20 世纪 70 年代，英国科学家 Sulston 通过微分

收稿日期：2018-07-27
通信作者：E-mail: wujr@sibs.ac.cn

干涉显微镜对秀丽隐杆线虫 (*C. elegans*) 的整个发育过程进行了持续的观察，描绘出了至今为止多细胞生物种类中一个最为完整的细胞谱系图 [1-2]。这种线虫是一个长约 1 mm，非常简单的无脊椎动物，从受精卵到成虫过程中总共分裂产生 1 090 个“体细胞”(somatic cell)；这张线虫细胞谱系图清楚地揭示了这 1 090 个体细胞中每一个细胞的身世和命运，例如其中的 131 个细胞会在发育过程的特定时期死去 [1-2]。

随着当前生命科学研究技术的发展，尤其是单细胞测序技术的成熟，研究者于 2016 年 10 月在英国伦敦召开了一个会议，建议实施一个远比线虫细胞谱系研究宏大的“人类细胞图谱”(Human Cell Atlas, HCA) 的国际合作研究计划。该计划堪比“人类基因组计划”，其组织委员会由来自 10 个国家不同研究机构的 27 名科学家组成。2017 年 5 月，HCA 的主要领导人 Regev 和 Teichmann 及其团队发表了关于这次会议的总结报告，系统地论述了 HCA 的意义、目标、任务和实施路径 [3]。在他们看来，HCA 的基本目标是：“采用特定的分子表达谱来确定人体的所有细胞类型，并将此类信息与经典的细胞空间位置和形态的描述连接起来”[3]。但是，与线虫相比，人体大约有 40 万亿到 60 万亿个细胞，HCA 面临的挑战显然要远远超过“人类基因组计划”。

1 要获得完整的细胞图谱面临的挑战

传统生物学对细胞类型 (cell type) 的确定，主要是根据其稳定不变的表型特征，如细胞形态、空间位置和生理性质等，如通过形态差异来区分神经细胞和神经胶质细胞。据此人们推测，人体的细胞类型可能从 200 种到 300 种不等。而 HCA 则试图从分子水平来对细胞进行分类。中国科学家最近发展了一种高通量的单细胞转录组测序技术，并用该技术对小鼠不同生命阶段的近 50 种器官组织的 40 余万个细胞进行了单细胞转录组分析，构建了首个建立在基因表达谱基础之上的哺乳动物细胞图谱 [4]。研究者根据这些转录组数据进行了相应的细胞分类，例如把怀孕小鼠和未受孕 (virgin) 小鼠乳腺组织上形态相同的细胞依据其基因表达差异分出了不同的类型 [4]。

这种按照分子信息来进行细胞分类的策略显然比传统的细胞分类方法更为精确，但同时也使得细胞类型的标准变得模糊，甚至有时变得比较随意。例如，哺乳动物机体内的分泌器官“胰岛”(pancreatic islet) 是由 α 细胞、β 细胞、γ 细胞及 PP 细胞等 4 种分泌型细胞组成。过去人们认为，这些 β 细胞是一种高度均一的胰岛素分泌细胞。但是，德国科学家不久前根据一种称为 Flattop 蛋白的表达与否把成年小鼠的胰岛 β 细胞分为两个亚群，不表达该蛋白的属于未成熟的 β 细胞，而表达该蛋白的则属于成熟的 β 细胞 [5]。而另外一项利用专一结合人体 β 细胞不同膜蛋白的抗体技术的研究揭示，成人的胰岛中存在 4 种亚型的 β 细胞 [6]。HCA 的组织者已经意识到了这样一个问题：他们首先面临的一个挑战是，缺乏对细胞类型和细胞状态 (cell state) 的严格定义。细胞类型意味着稳态特征，而细胞状态则是指瞬态特性，由于所有细胞都是处于变化之中，所以这两个概念间的边界难以区分 [3]。而在分子水平的细胞分类显然让这个问题更为复杂化。

HCA 的组织者提出其理想的终极目标：(1) 确定人体中的每一个细胞；(2) 确定每个细胞的空间位置；(3) 通过细胞谱系确定在人的一生中每一时刻出现过的每个细胞；(4) 根据健康状态、基因型、生活方式和外界环境的不同，对每一个个体的细胞图谱进行注释 [3]。可以这样说，这一终极目标的实现意味着生命科学的终结。面对这样的终极目标，HCA 的组织者不得不承认：“当然，要构建这样的终极图谱是完全不可能的”[3]。

退而求其次，HCA 的组织者把实际目标定为“一个基于所有人类细胞的稳定属性和瞬态特征的综合性参考目录，并包括各种细胞的位置和丰度”[3]。值得注意的是，HCA 的组织者还试图通过两个研究策略把其实际目标分解得更容易实现。首先是借鉴人类基因组计划分阶段完成的方式，“明智的策略是制定一系列构建细胞图谱‘草图’(draft) 的阶段性目标，这些‘草图’将逐渐地增加 (人类细胞图谱) 分辨的精度、覆盖的广度和解释的深度”[3]。

按照这一策略，HCA 目前已经发表了许多人类细胞图谱的阶段性工作，例如，2018 年 3 月 8 日，英国 Sanger 研究所在其网站上宣布，25 万个发育细胞作为人类细胞图谱计划的第一步，已经完成单细胞转录组测序工作；美国 Broad 研究所紧随其后也发布了 50 多万个人体免疫细胞的单细胞转录组测序数据 [7]。此外，中国科学家对人类胚胎 8 到 26 周发育过程中前额叶皮层的 2 300 多个细胞进行了单细胞转录组测序，发现了神经干细胞、兴奋性神经元、抑制性神经元、星型胶质细胞、少突胶质细胞、

小胶质细胞等六大类细胞，并进一步把这六大类细胞精确划分为 35 个独立的细胞亚型 [8]。

虽然这种分阶段研究的策略能够推进 HCA 的工作，不停地产生人类细胞图谱的各种“草图”，但要想获得人体数十万亿个细胞的“全图”显然并非易事。就以人类基因组计划的“小目标”——弄清人类基因组究竟有多少个基因——为例，从 2001 年人类基因组“草图”的发表到今天将近 20 年的时间，研究者尚未得到一个确定性结论。2001 年发表的人类基因组“草图”提出了“高可信度”基因数为 26 588；2004 年的人类基因组“全图”预测出的基因数目在 2 万到 2 万 5 千个之间；美国国立卫生研究院专门负责收集基因的项目“Mammalian Gene Collection”在 2009 年宣称人类基因组只有 18 877 个基因；美国科学家 2018 年在预印本网站“bioRxiv”发表的研究论文提出，人类基因组用来编码蛋白质的基因数目是 21 306 个。

HCA 的组织者拟采用的第二个研究策略是“局部采样”。在他们看来，“为了得到人体细胞的精确图谱去对人体中所有细胞进行研究是不可能的，也没有这个必要”[3]。HCA 的组织者提出，在细胞图谱的研究过程中，可以先进行少量稀疏的细胞采样，经过分析后再决定更深度的采样方式 [3]。此外，作为分子层面的细胞类型分析，也同样可以进行“局部分析”，即首先对采集到的细胞进行低覆盖度 (low coverage) 的转录组测序，以便能找到尽可能多的细胞类型；然后，从中再选择少量的细胞进行深度测序，从而能够帮助解释来自低覆盖度的测序数据并增加检测的精度 [3]。

如果细胞分类仍然采用传统的表型特征，如形态辨识方法，那么这种“局部采样”的研究策略是有一定的合理性：一方面可以按照明确的可操作的标准从一群具有稳定表型特征的细胞中采集样本；另一方面由样本分析得到结果也是具有相应的代表性。但是，HCA 的分类方法是建立在分子水平之上，根据什么标准来采集样本就成了一个问题。例如，胰岛 β 细胞可以按照 Flattop 蛋白的表达与否进行分类 [5]，也可以按照结合 β 细胞不同膜蛋白的抗体特征进行分类 [6]，但这些细胞的分类标准在采样之前是无从知道的，只能是按照的传统取样方法来提取胰岛 β 细胞样品。此外，这些细胞分类结果也很难说具有什么样的代表性，胰岛 β 细胞按照 Flattop 蛋白的表达情况分出的亚型与按照膜蛋白的抗体结合情况分出的亚型之间是什么关系？这两种生物标志物中谁更具有代表性？

目前人类细胞图谱分析的主要技术是单细胞转录组测序。不久前，英美科学家利用该项技术，改写了成年人血液中特定的免疫细胞图谱，其中 DC 细胞 (dendritic cell) 由 4 种类型扩大为 6 种，单核细胞 (monocyte) 由原来的 2 种类型变成了 4 种，还发现了一种新的 DC 祖细胞 (DC progenitor)[9]。在这项工作中，研究者首先采用了一系列特定的抗体来提取和富集血液中的 DC 细胞和单核细胞，然后再对这些细胞进行单细胞转录组测序 [9]。显然，采用的抗体种类对提取细胞样本具有决定性的作用；如果采用了不同的抗体就可能得到不同的细胞，从而出现不同的测序结果。此外，根据单细胞转录组测序数据，研究者对每个细胞选择了平均 5 326 个特定的基因进行分析，寻找出关键的分子标志物，然后才确定出新的细胞类型 [9]。因此，这些新的细胞类型是建立在个别的分子标志物之上，例如，新的 DC 细胞亚类是由细胞表达的 3 个抗原 ——AXL、SIGLEC1 和 SIGLEC6 来决定的，研究者据此把这些细胞命名为“AS DCs”[9]。显然，如果采用不同的分析技术，很有可能从这些 DC 细胞的转录组数据中找到不同的生物标志物，导致不同的细胞分类。

“局部采样”策略建立在抽样的合理性以及其分析结果的代表性之上；临床研究通常采用的“随机对照试验”(randomized controlled trial, RCT) 可以视为这个策略的代表。而通过以上分析可以看到，在分子水平进行“局部采样”策略确定的细胞谱系其实并没有真正的“代表性”，一方面研究者很难建立一个符合科学抽样标准的操作方法，只能是采到什么细胞就分析什么细胞；另一方面研究者只能就分析中获得的特定分子标志物进行细胞分类，看到什么分子标志物就说是什么类型的细胞。

2 来自个体发育和生长的挑战

HCA 面临的一个更大的挑战来自时间尺度：多细胞生物从受精卵分裂到个体发育再到衰老的生长过程中，不停地产生着各种新的细胞类型，同时又有许多细胞类型消失。Sulston 正是通过对线虫整个生长过程的持续观察，才描绘出了一个完整的细胞谱系图。当然，HCA 的组织者并没有忽视这个挑战，并提出了一些应对的措施，“对模式生物来说，可以通过识别共同的祖细胞类型来构建真正的谱系树”；而对人类来说，“更普遍的可行性做法是，通过检测每次细胞分裂时 DNA 变异的稳定积累 (例

如体细胞点突变或者微卫星位点的重复序列扩增) 来追踪谱系”[3]。

多细胞生物，尤其是哺乳动物等高等生物的发育过程非常复杂，人们现在利用新的技术常常获得与传统发育生物学观点不一致的结果。例如，研究者从受精后 5~24 h 期间的非洲爪蟾胚胎上提取了近 14 万个细胞，并进行了单细胞转录组测序；根据 259 个基因表达簇 (gene expression clusters)，研究者分出了 69 个胚胎细胞类型。此外，他们还发现许多胚胎细胞的状态比以往认为的要早得多[10]。这意味着细胞分化在胚胎发育时期的进程要重新定义。另外两项对斑马鱼胚胎发育时期的单细胞转录组测序研究发现，其早期细胞分化的图谱并不是经典的“树状”，在有些“分枝”点的细胞从其转录图谱来看是处于“多种命运”(multiple fates) 的状态下，即处于某个分化路径的细胞可以发生转换而进入到其他的分化路径[11-12]。这种“多种命运”特征有可能使得 HCA 的组织者提出的策略 ——“通过识别共同的祖细胞类型来构建真正的谱系树”——难以实现。

更值得注意的是对发育涉及到的时间尺度的理解。传统的发育生物学认为，对哺乳动物而言，个体的绝大部分细胞类型在胚胎发育过程中都已经完成，出生之后其机体主要是一个非发育性的“成长”过程。因此，细胞谱系的研究主要关注胚胎发育时期而非生长期。例如，在构建首个哺乳动物细胞图谱的工作中，研究者特别强调了选择乳腺组织的优点：“哺乳动物乳腺提供了一个研究组织器官分化独特的模型，因为它是唯一在出生后已经完全发育好的腺体器官”[4]。然而，个体发育的过程并非局限在出生之前。这一观点随着研究工作的推进已经有了更深入的认识。例如，过去认为，小鼠全部冠状动脉都是由心脏外表面血管在胚胎发育过程中“自外向内”扩增而来；但中国科学家不久前利用遗传谱系示踪技术 (genetic lineage tracing) 对小鼠心脏冠状动脉的细胞谱系研究发现，一部分冠状动脉居然是出生后才“自内向外”开始生长。具体来说，冠状动脉源于两种不同的“祖细胞”—— 心脏壁外层的冠状动脉由心外膜下血管祖细胞发育而成，而心脏壁内层的冠状动脉则来源于心内膜祖细胞；而且这两类“祖细胞”的发育时间也有很大差异，心脏壁外层的冠状动脉在胚胎时期就已经开始发育，而心脏壁内层的冠状动脉则在出生后 1 周至 2 周才开始生长[13]。

由此可以看到，如果把细胞谱系研究简单地局限在动物胚胎发育时期，将难以获得完整的个体发育的细胞谱系图。但是，如果不把细胞谱系研究涉及的时间段确定在胚胎阶段，则意味着研究者应该关注从胚胎期到成年期的生长全过程。这对线虫的细胞谱系分析来说容易做到，因为线虫生活周期短，从受精卵发育到可以产卵的成虫不超过三天。但是，对小鼠等哺乳动物就比较麻烦了，其生长期长达数月。而对人来说，要研究个体在数十年生长期里可能出现的细胞谱系显然不是一件易事。

如果把衰老过程视为“逆发育”过程，HCA 的研究者不仅要考虑个体的成长期，而且需要考虑个体在衰老阶段的细胞图谱。已经有多项研究表明，衰老期的细胞与成年期的细胞有很大的差别。早在 20 世纪末，研究者已经知道衰老的哺乳动物细胞具有很特殊的表型，包括停止细胞分裂、抗细胞凋亡、分泌大量的炎性细胞因子和其他种类的蛋白质。一项转录组研究工作指出，年轻人的成纤维细胞与老年人的有 600 多个基因表达的差别，研究者据此基因表达差异确定出衰老细胞特有的转录组“指纹图”(fingerprint)[14]。

更为复杂的是，对免疫细胞的单细胞转录组测序结果表明，在年轻小鼠同类型的免疫细胞中，各个细胞之间的基因表达谱基本一致，没有明显的差异；但在老年鼠体内的同类型免疫细胞中，各个细胞之间的基因表达差异则明显增加[15]。不久前，一项单细胞染色质修饰谱分析工作也发现了同样的现象：老年人体内不同免疫细胞之间的染色质上的组蛋白修饰差异要远大于年轻人的[16]。这种个体内免疫细胞的异质性意味着，如果依照 HCA 对细胞图谱的分子分类标准，那么年轻人体内同一种免疫细胞很有可能在老年人体内就会变成很多种类的免疫细胞，具体的细胞类型将取决于这些衰老细胞的各种生物大分子间的差异情况。

3 基因组随机性突变带来的挑战

HCA 目前重点关注的是细胞在转录组层面的基因表达情况。但是，真正的细胞谱系图不可能忽略了基因组层面的变化，因为从根本上来说，“谱系”的核心是“血缘关系”。传统的生物学观点认为，在实现体细胞一代又一代增殖的有丝分裂过程中，子代细胞的基因组拷贝是由亲代细胞基因组完整而准确地复制和分配而来；因此，用来构成不同组织器官的各种类型体细胞都拥有同样的基因组。也就

是说，在个体发育过程中产生的各种体细胞之间的“血缘关系”都是一样的，所有体细胞的基因组都来自受精卵的基因组；因此，细胞之间的差别与基因组序列关系不大，不需要关注基因组。但是，在人类基因组计划完成的“后基因组时代”，研究者已经证明这种观点是错的；事实上，在每一个个体内，不同的体细胞基因组存在着广泛的差异。

这种传统的“基因组同一性”观点在20世纪英国科学家Sulston研究线虫细胞谱系图的时代不构成大问题，当时他仅仅是依据细胞形态的信息进行细胞谱系的构建。但是，按照HCA的组织者自己确定的目标——在分子水平上进行细胞类型的分析，细胞谱系图研究不考虑基因组层面的信息显然就说不过去。笔者注意到，在讨论如何获得人类发育过程中的细胞谱系图的技术策略时，HCA的组织者已经考虑到要研究基因组层面的变化，“通过检测每次细胞分裂时DNA变异的稳定积累(例如体细胞点突变或者微卫星位点的重复序列扩增)来追踪谱系”[3]。

美国科学家在2018年初发表的一项研究看起来很符合HCA的组织者的思路。在这项工作中，研究者对3个人体胚胎的前脑组织的细胞进行了单细胞全基因组序列分析，发现这些细胞中广泛存在着单核苷酸变异(single-nucleotide variations, SNVs)，平均每个细胞有200个到400个SNV；研究者重构了受精卵最早5次分裂过程中的细胞谱系图，并计算出突变率是每个细胞在每次分裂过程中产生大约1.3个SNV[18]。研究者还指出，在胚胎发育的后期，包括神经发生时期(neurogenesis)，由于氧化损伤作用将导致突变率进一步增加，而且胚胎发育期间的突变会明显多于生长期[17]。

体细胞的突变通常源于DNA复制过程中随机产生的复制错误。有研究文章指出，基因组复制过程中随机产生的突变可以传递到下一代；细胞分裂的次数越多，细胞内积累的复制突变就越多[18]。此外，另一项单细胞全基因组测序研究也发现，在人体早期胚胎发育过程中，细胞每分裂一次平均产生3个体细胞突变(somatic mutation)；更重要的是，这些体细胞突变将以不对称的方式传递给子代细胞[19]。因此，这些随机突变的产生和不对称传递必然导致各个细胞之间的基因组序列有所差别，需要对每个细胞都加以分析才能知道其基因组序列的变异情况。

基因组不仅存在点突变等微小的序列差异，而且还广泛存在着较大的染色体结构差异，如基因拷贝数变异(copy number variant, CNV)。一项研究工作报道，有丝分裂过程通常会导致小鼠胚胎干细胞基因组广泛形成CNV[20]。此外，染色体结构差异在人类胚胎早期发育过程中也是很常见的事件，不仅在大多数卵裂期胚胎的细胞里发现具有非整倍体的基因组，而且在随后的分裂球的细胞内也可以看到各种大片段基因组DNA缺失或者扩增，意味着早期胚胎细胞拥有的是高度不均一的“镶嵌型”(mosaicism)基因组[20]。

值得注意的是，“镶嵌型”基因组在机体的各种体细胞内广泛分布。通过单细胞测序技术对人脑部额皮质的神经细胞基因组分析发现，新产生的CNV(*de novo* CNV)存在于在13%~41%的神经细胞内[21]；而对人体皮肤细胞的基因组分析则指出，大约30%的人体成纤维细胞的基因组内具有源于体细胞的CNV(somatic CNV)[22]。研究者还发现，随机产生变异的“镶嵌型”基因组在人体的整个生长期一直在不停地形成，例如在大脑海马的齿状回区域，每个神经细胞每年大约出现40个体细胞突变，其突变率是胚胎发育期前额皮质神经细胞的2倍[23]。换句话说，人体从胚胎发育到身体衰老，可能就不会产生基因组序列完全一致的两个细胞。

还有一个更大挑战HCA需要面对：环境导致的体细胞基因组序列的随机变异。人类生活的各种环境因素时时刻刻在影响着体内的细胞。众所周知，抽烟会引发基因组变异。不久前一项研究系统地分析了抽烟与突变的关系，发现抽烟的程度与突变程度高度相关，并涉及到碱基置换、插入缺失突变(indels)和CNV等多种突变[24]。更有甚者，即使只是晒晒太阳也有可能引发突变。一项研究指出，长期紫外线照射能够引起正常人体皮肤的上皮细胞基因组发生突变，每个细胞基因组中大约每1百万碱基平均出现2~6个突变，而且在正常皮肤细胞发现的突变中有许多是已知癌基因突变[25]。显然，环境与基因组突变之间的关系是复杂的、随机的，并且是不可忽略的。

由此可见，由于细胞增殖过程中的随机复制错误，以及由于生长环境所导致的随机突变，使得机体内的体细胞广泛地拥有具有不同变异序列的“镶嵌型”基因组。这种“镶嵌型”基因组明显增加了体细胞的遗传复杂性和细胞类型的多样化。更麻烦的是，HCA的组织者如果不是采集“所有”细胞而是用“局部采样”的方式进行细胞图谱研究，势必会遗漏并无从得知没有采集到的细胞里随机出现

的基因组变异信息。

4 不同种类的生物分子标志物之间非线性关系的挑战

HCA 最主要的策略是从分子水平来对细胞进行分类，“这个实施方案显然就是通过确定一系列分子标志物来描述每个人类细胞。例如，可以通过描述编码人类蛋白质的大约 2 万个基因的表达水平来描述每个人类细胞，……当然，这种分子标志物的集合还将包括非编码基因的表达水平、转录本可变剪接的水平、每个启动子和增强子的染色质状态，以及每个蛋白质表达水平和它们的每一种翻译后修饰状态等”[3]。换句话说，细胞内的各种生物大分子的表达水平和修饰状态等信息都将被 HCA 提取出来并用于细胞的分类。

HCA 的组织者提出的这种实施方案看上去很全面，几乎把人们能够想到的生物大分子的信息都包括了。但是，一个由此而来的问题并没有得到很好的思考和解答：这些不同类型的生物大分子之间是什么关系？这其中最主要的是要回答基因表达水平与蛋白质表达水平的关系。按照分子生物学的“中心法则”，基因组上的基因被转录为 mRNA，然后根据 mRNA 合成蛋白质。过去人们认为，基因转录水平与蛋白质合成水平是线性关系，如果基因表达水平高，即作为模板的 mRNA 的拷贝数多，则蛋白质表达水平高，即合成出来的蛋白质就应该多。

随着研究工作的深入，这种基因转录水平与蛋白质表达水平呈简单线性关系的观点已被多项实验所否定。21 世纪初的一篇系统生物学经典论文指出，在酵母细胞中，所检测到的 289 种蛋白质的丰度与其对应的 mRNA 表达水平的相关性并不高，相关系数只达到 0.61；此外，有 30 个蛋白质的丰度在野生型和突变型细胞之间有着明显的差异，但编码这些蛋白质的基因中有一半的 mRNA 表达水平却没有出现相应的显著性变化[26]。随后的另一项研究也同样观察到了酵母细胞的许多蛋白质表达水平和其相应的 mRNA 表达水平之间有较大的差异[27]。对人体肝细胞的研究也同样表明，蛋白质表达水平和其相应的 mRNA 表达水平之间的相关系数只有 0.48，并且这种非线性关系不是随机分布的，例如，在丰度最高的 50 种 mRNA 中有 29 种是用来编码分泌蛋白的；可在丰度最高的 50 种蛋白质中却没有一种是分泌蛋白[28]。对大鼠肝脏转录组和蛋白质组分析也得到了相同的结果：蛋白质表达水平和其相应的 mRNA 表达水平之间相关性不高[29]。

以上这几项研究都只是在测定了大量细胞的基础上得到的统计性结论。今天 HCA 提出的目标是要在单细胞水平上进行分析，显然这种基因表达水平和蛋白质表达水平间的关系就变得更为复杂。一项利用单分子研究技术对单个大肠杆菌细胞分析的结果表明，基因表达水平和蛋白质表达水平不仅受到各个细胞间差异之“外部噪音”(extrinsic noise) 的影响，而且还受到细胞内的“内部噪音”(intrinsic noise) 的影响，两者的丰度关系出现了明显的细胞个体差异[30]。研究者由此得出这样一个结论：“对任何一个给定的基因而言，在单个细胞内的蛋白质拷贝数和 mRNA 拷贝数之间没有相关性”[30]。换句话说，由于基因表达水平与蛋白质表达水平不存在相关性，因此，在同时测量一个细胞里两者的表达水平时就很有可能得到不同的，甚至是相反的结果。

2016 年，一篇综述文章系统地分析了蛋白质表达水平与 mRNA 表达水平的关系，指出这种关系受到细胞内外环境变化、细胞稳态和状态变化以及 mRNA 的时空分布等各种影响。作者总结到：“转录水平本身在许多情况下不足以用来预测蛋白质表达水平以及解释基因型与表型的关系。因此，获取在不同层次的基因表达水平相关的高质量数据是完全理解生物学过程所必不可少的”[31]。

HCA 面临的问题是，如果研究者在一个细胞里同时获得了可用于细胞分类的基因表达水平和蛋白质表达水平的分子标志物，但这两类分子标志物的丰度不一致，例如 A 基因是高表达而其蛋白是低丰度，应该根据哪一类分子标志物来进行细胞的分类？换句话说，这两类分子标志物在用做细胞分类的标准时谁更重要？瑞典科学家在 2017 年发表的一项研究已经涉及到了这个问题。研究者首先根据 56 株人细胞系的基因表达谱构建了一个细胞图谱，在此基础上选择了 22 株细胞系，然后用近 13 993 种抗体检测了这些细胞株上的 12 003 种蛋白质的亚细胞器的空间分布；研究者还通过单细胞分析技术发现了 1 855 种蛋白质存在着表达水平或者空间分布的差异[32]。显然，在这项工作中，研究者是把基因表达谱而非蛋白质表达谱作为细胞分类的标准。不过，德国科学家通过小鼠大脑皮层神经组织发育的研究工作对该问题给出了另外一种答案：他们发现少突胶质细胞 (oligodendrocytes)、星形胶质细胞 (astrocytes)、小胶质细胞 (microglia) 和神经细胞 (neurons) 的转录组与蛋白质组的相关性系数分别只有 0.4~

0.45；在此情况下，研究者提出，蛋白质组的数据能够更好地反映大脑皮层的细胞类型和差异[33]。

从生物学现有的知识来说，基因组的表观遗传修饰、非编码 RNA 的转录和蛋白质翻译后修饰等显然与基因表达或者蛋白质表达之间不存在线性关系。因此，如果 HCA 把这些类型的信息全都视为可用于细胞分类的分子标志物，这将进一步导致细胞分类的复杂性和不确定性。这还只是在分子层面上反映出来的问题，如果再到更高的层次，如细胞形态或生理功能等，这些涉及到细胞分类的各种标准之间的关系将变得更为模糊。HCA 的组织者显然也意识到了这一问题，“关键在于现在始终不清楚基于形态的、分子的和生理性质三者各自得到的分类特征是否相互兼容”[3]。

5 结语

19 世纪法国数学家拉普拉斯 (Laplace) 是科学史上倡导决定论的最著名人物，他于 1814 年提出了“拉普拉斯妖“的假设：如果一个智者知道宇宙中每个原子确切的位置和动量，并能够对这些数据进行分析，就能够用物理定律来展现宇宙中所有事件的全过程，从过去到未来。从 HCA 计划设定的宏伟目标而言 —— 提取所有的生命信息并描绘出人体每个细胞的类型、状态和运行规律，可以把该计划视为“拉普拉斯妖”在生命科学领域的翻版。从以上对该计划面对的各种挑战之讨论中可以看到，最大的挑战正是来自研究者的决定论思路与生命复杂系统的不确定性之间的冲突。

[参 考 文 献]

[1] Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. Dev Biol, 1977, 56: 110-56

[2] Sulston JE, Schierenberg E, White JG, et al. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol, 1983, 100: 64-119

[3] Regev A, Teichmann SA, Lande ES, et al. The human cell atlas. bioRxiv, 2017, doi: http://dx.doi.org/10.1101/121202

[4] Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by microwell-seq. Cell, 2018, 172: 1091-107

[5] Bader E, Migliorini A, Gegg M, et al. Identification of proliferative and mature β-cells in the islets of Langerhans. Nature, 2016, 535: 430-4

[6] Dorrell C, Schug J, Canaday PS, et al. Human islets contain four distinct subtypes of β cells. Nat Commun, 2016, 7: 11756

[7] https://preview.data.humancellatlas.org/

[8] Zhong S, Zhang S, Fan X, et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. Nature, 2018, 555: 524-8

[9] Villani A, Satija R, Reynolds G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science, 2017, 356: eaah4573

[10] Briggs JA, Weinreb C, Wagner DE, et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science, 2018, 360: eaar5780

[11] Farrell JA, Wang Y, Riesenfeld SJ, et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science, 2018, 360: eaar3131

[12] Wagner DE, Weinreb C, Collins ZM, et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science, 2018, 360: 981-7

[13] Tian X, Hu T, Zhang H, et al. Vessel formation. *De novo* formation of a distinct coronary vascular population in neonatal heart. Science, 2014, 345: 90-4

[14] Zhang H, Pan KH, Cohen SN. Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci. Proc Natl Acad Sci USA, 2003, 100: 3251-6

[15] Martinez-Jimenez CP, Eling N, Chen HC, et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science, 2017, 355: 1433-6

[16] Cheung P, Vallania F, Warsinske HC, et al. Single-cell chromatin modification profiling reveals increased epigenetic variations with aging. Cell, 2018, 173: 1385-97

[17] Bae T, Tomasini L, Mariani J, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science, 2018, 359: 550-5

[18] Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science, 2017, 355: 1330-4

[19] Ju YS, Martincorena I, Gerstung M, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature, 2017, 543: 714-8

[20] Vanneste E, Voet T, Caignec C, et al. Chromosome instability is common in human cleavage-stage embryos. Nat Med, 2009, 15: 577-83

[21] McConnell MJ, Lindberg MR, Brennand KJ, et al. Mosaic copy number variation in human neurons. Science, 2013, 342: 632-7

[22] Abyzov A, Mariani J, Palejev D, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature, 2012, 492: 438-42

[23] Lodato MA, Rodin RE, Bohrson CL, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science, 2018, 359: 555-9

[24] Alexandrov LB, Ju YS, Haase K, et al. Mutational signatures associated with tobacco smoking in human cancer. Science, 2016, 354: 618-22

[25] Martincorena I, Roshan A, Gerstung M, et al. High burden and pervasive positive selection of somatic mutations innormal human skin. Science, 2015, 348: 880-6

[26] Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic

and proteomic analyses of a systematically perturbed metabolic network. Science, 2001, 292: 929-34

[27] Griffin TJ, Gygi SP, Ideker T, et al. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. Mol Cell, Proteomics, 2002, 1: 323-33

[28] Anderson L, Seilhame J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis, 1997, 18: 533-7

[29] Low TY, van Heesch S, van den Toorn H, et al. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep, 2013, 5: 1469-78

[30] Taniguchi Y, Choi PJ, Li GW, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science, 2010, 329: 533-8

[31] Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell, 2016, 165: 535-50

[32] Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. Science, 2017, 356: eaal3321

[33] Sharma K, Schmitt S, Bergner CG, et al. Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci, 2015, 18: 1819-31