中国科学: 生命科学 2020 年 第 50 卷 第 5 期: 538 ~ 548
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 溯本求源大“烤”问: 基因组多维结构和动植物基因组结构差异专题

多维度基因组信息: 起源、内涵与技术瓶颈

于军[1,2*]

1. 中国科学院北京基因组研究所, 基因组科学与信息重点实验室, 北京 100101;
2. 中国科学院大学, 北京 100049
* 联系人, E-mail: junyu@big.ac.cn

收稿日期: 2019-11-21; 接受日期: 2020-01-10; 网络版发表日期: 2020-04-15
国家自然科学基金(批准号: 31671350)资助

摘要 近年来发展起来的染色体多维结构研究很快就遇到了概念盲区和技术瓶颈, 有必要在较高级的思维层次加以澄清和深入探讨. 虽然生命可能起源于操作(基因组DNA以外的大小生物分子)而非信息(基因组DNA), 但是当DNA被选择来承载信息后(或同时), 其操作和平衡功能又分别在不同层面上(不同谱系、物种、细胞、染色质、基因)以不同程度体现出来, 比如参与平衡的物质不仅包括DNA和RNA大分子也包括脱氧核苷酸、核苷酸、修饰核苷衍生物等小分子以及可识别的结构单元包括染色质区域、转座子、染色体结构元件等. 正是这些物质和结构的动态变化和交叉关联构成了多维度的生命活动. 这些活动的机械化学、时间空间界定及对其规律的探索和研究, 构成了复杂的基因调控体系, 从中衍生出由简到繁、由表及里、纵横交错的核心生物学元素和这些元素所构成的科学问题.

关键词 基因组, 染色体, 基因表达调控, 多维基因组信息

我们无法在提出问题的那个层次去解决问题. ——阿尔伯特·爱因斯坦(We cannot solve our problems with the same thinking we used when we created them. Albert Einstein). 生命的发生——尤其是复杂生命, 比如脊椎动物——就形态而言是从单细胞或一个受精卵开始的, 即受精卵所含有的遗传信息和所携带的物质在胞外基本滋养物充足和物理环境允许时就可以形成个体. 如果细胞的生命活动是长期变演(evolution; 变是基因型的属性, 演是表现型或称表型的属性)产生的自组装(self-assemblage)活动, 而细胞间的生命活动相应来说是某种形式的自组织(self-organization), 那么受精卵一定携带着早期生命发生的全部信息与物质. 然而这种信息显然超出了DNA分子的线性编码内容, 这个物种(甚至这个个体)特异的遗传信息一定是多维度的(或称全息的), 因此衍生出了三维基因组结构与信息的概念. 本文主要讨论这些信息的维度概念, 阐述其层次、内涵、关联等, 将结构分析、实验数据和算法等学科细节留给专刊的其他综述文章.

要认识这些全息的、可以用时间与空间(时空)坐标和机械化学(机化)动态公式来表述的三维结构, 显然需要新的思考逻辑和初级科学概念来推演出具有收敛性质的思维框架, 而不仅仅是实验数据. 细胞内的自

引用格式: 于军. 多维度基因组信息: 起源、内涵与技术瓶颈. 中国科学: 生命科学, 2020, 50: 538–548
Yu J. Multi-dimensional genomics information: origin, context, and technical bottlenecks (in Chinese). Sci Sin Vitae, 2020, 50: 538–548, doi: 10.1360/SSV-2019-0276

组装过程就是其生化组分在细胞周期中的活动，而这些活动就复杂生命而言大都不同步，只是忠诚或非忠诚式地自我复制着；细胞间的自组织也是如此，各自分化后形成组织和器官，展示变化着的复杂时空关系，具有自主性、稳健性和可塑性. 因此, 解析这些复杂的关系和相关要素首先需要一系列极简的界定和准确的概念阐述, 将各个维度的基本内涵加以分割和区别, 从而实现关联和贯通. 在线性维度, 基因组信息是构成染色体的DNA分子及其线性序列，其次生信息包括基因组多态性和所有可变化结构单元序列(如端粒). 二维基因组信息主要是以DNA序列为坐标的附加信息，除基因座(gene locus)外，还包括共价修饰核苷酸位点及变化，例如甲基化组(methylome)和CpG岛组(CpG-islandome)研究. 三维基因组信息最为复杂，包括DNA自身复制(如复制起点及变化)、所有转录本的表达调控、与DNA相互作用的所有蛋白质(机器)、染色质与细胞核结构的动态关系等信息. 实现三维基因组形态要素界定最基本的技术应该是：在单细胞水平实现单分子分辨率的基因产物定性(功能表征)、定位(机化表征)、定时(时空表征)和定量(物质表征). 三维基因组研究对科学家的挑战是启动一系列大型科学项目，择选阶段性目标、准备相应技术、组织攻关团队. 问题的复杂性导致每一个环节和各个局部的不充分研究都无法产生有生命力的实验数据. 三维基因组信息的汇聚与研究将系统性解决细胞的自组装规律、过程和原理. 四维基因组信息就是将自组织过程加上时间轴考量，包括细胞周期和细胞分化等细胞本身的时效性变化. 在实际生命体的生命周期上，染色体或基因组还将与细胞内外各种物质(直接交流)和信号(间接交流)形成交流互动, 与其他层面的信息形成第五个维度, 即细胞水平的自组织过程，从而形成组织、器官、生理要素等，总体来讲就是细胞间的互动，其包括整体发育——从一个受精卵到成熟个体. 一维到五维的基因组信息研究构成通生学(holovivology)研究内容[1]. 生物医学研究应归属于第六个维度的基因组信息——整个生命周期的整体化(integration)和阶段性(包括生命周期、生殖周期、生理周期等，比如更年期(menopause))等现象，包括生理与病理状态的比较、生老病死的过程. 当然还有第七、八、九三个维度，分别是种群、生态群和生物圈. 这九个维度分解了“全息”(holo-; 来自希腊语 *holos*, 全、整体之意)的生命信息并表征了“全活”(*vivo*-; 来自拉丁语, 意为活)的生命运动，最后汇聚成通生学的“五流(five tracks)”思维框架，分别为信息流、操作流、平衡流、分室流、可塑流(表1). “五流”思维框架(frameworks)的真谛在于初步将生命体系的复杂性作为多元线性方程体系来思考，通过界定其各流层面特有的分子机制(molecular mechanisms)和细胞过程(cellular processes)，以及这些机制和过程之间的内在关系，系统界定生命现象的收敛性主体或归宿(相对于“组学”定义下的对象无限可分性), 并透过现象揭示其真实的本质和分子机理.

1　染色体多维结构的起源

生命起源于操作、平衡、分室而不是信息——DNA. 人类对生命起源的认识到近20年来才趋于成熟, 摒弃了各种历史性、片面性的单纯性思考. 生命来自于何物的探索一直延续至今，从寻找生命大分子(DNA、RNA、蛋白质、淀粉、纤维素、半纤维素和木质素等)化学构件的起源(核苷酸、氨基酸、戊糖等)，到这些大分子的键合(磷酯键、肽键、酰键等)基础. 目前的认知承认生命起源于生物大分子，而这些大分子的聚合反应应该就是生命起源的第一步——“多聚体世界”(the polymers world). 这些作为生物

表 1　五流说与生物学科领域的关系

Table 1　The five-track frameworks and related biological disciplines

五流(the five tracks)	对应的学科领域	生物学命题举例
信息流(informational track)	遗传学、基因组学、基因组信息学	动植物基因组结构差异
操作流(operational track)	分子生物学、生物物理学、大分子结构	基因表达调控、基因组三维结构
平衡流(homeostatic track)	生物化学、生理学、药理学	细胞能量产生与控制、信号传导
分室流(compartmental track)	细胞生物学、生命起源、发育生物学	干细胞分化、生殖系细胞形成
可塑流(plastic track)	系统生物学、生态学、免疫学、神经生物学	滞育、成瘾、休眠、迁徙

大分子自组装基础的聚合反应包括同质(homopolymer; 相同化合物的聚合, 比如核苷酸的产生[2]和核苷酸聚合成RNA)和异质(heteropolymer; 比如tRNA与氨基酰tRNA)[3]. 在生物大分子的优先选择上, 也从蛋白质和DNA, 转到了RNA, 亦即“RNA世界”(the RNA world)假说. 显然假设生命仅仅是起源于“RNA世界”与第一步的逻辑不符, 因此生命起源的第二步应该是遗传密码在“多聚体世界”的升级版——“[RNA+蛋白质]世界”的逐步确立. 遗传密码的诞生, 或者说标准(通用)遗传密码的诞生过程止于“DNA世界”. 此刻, “RNA世界”遗传编码的多样性被“冻固了”[4]或平衡了[5]或渐进了[6~9]: 所有构成RNA分子的核苷酸都被迫简约为四个构成DNA聚合体的脱氧核糖核酸. 这便是生命起源的“三部曲”. 在细胞的世界里, DNA虽然不是核苷酸总分子当量最多的基本分子, 但无疑是最大的生物大分子. 细胞内一切活动都是为了让这个生物大分子可以高度完整地传给下一代. 这个生物大分子的序列、组分、结构和周期性变化等构成了基因组的物质形态和功能表征——染色体多维结构. 这个以DNA分子为主体的复杂多维结构体在物种的变演中不断丰富变化, 其有序性构成多细胞物种的生殖细胞并以异性各自单一细胞的不对称融合形成配子而实现“传宗接代”.

2 染色体多维结构与通生学“五流”思维框架的关联

染色体和基因组与其多维结构是在变化中诞生的, 也会在变化中生存和变演. 染色体的多维结构都与DNA分子的序列和序列的变化规律有关. 人们在讨论一维或线性序列的时候必须将其抽象出来, 确定一个极简范围, 这就是信息流. 其他维度则与另外“四流”相关.

一维基因组(物种染色体的总和)也有诸多层面的命题. 首先是何为基因的问题, 人们习惯仅将编码蛋白质的DNA序列(染色体上的基因座, locus)称为基因(编码蛋白质的基因). 那么仅编码行使操作(或催化)功能RNA的序列(RNA基因)将如何界定呢? 应该将转录本和转录组统一起来, 让RNA和转录本(transcript)通用化, 不能让转录组特指mRNA. 比如mRNA可称为messenger transcript, 并依此推出信使转录组(messenger-transcriptome或mRNAome)和转运转录组(transfer-transcriptome或tRNAome). 这样所有从DNA模版转录出来的RNA就都是转录组的一部分. 其次是基因结构的问题, 植物基因组与动物基因组在基因结构上有本质的差别[10~14], 这个差别决定了重复序列放在基因内(成为基因的一部分)还是放在基因间(成为基因空间外的染色体部分)的问题; 同时, 也决定了生物学界定的重复序列(biologically-defined repeats, BDR; 主要指不同类型的转座子和其残留序列)在谱系内和物种内的特异性问题[15].

二维基因组主要涉及DNA分子作为信息载体的基因编码、基因附加结构、与功能相关的序列元素, 及其表征或性质, 是操作流在信息流层面的体现. 这个层面的研究对象包括DNA共价修饰位点(如甲基化、羟甲基化)、CpG岛[16~18]、核小体占位[19~22]、着丝粒和端粒等. 可转录的所有DNA序列、基因簇、基因簇的反向调控转录本等也属于这个层面, 还包括染色体结构域、染色体边界等. 此外, 还有一些染色体结构和序列中的顺式(*cis*)元件等也都属于这个层面. 换言之, 二维基因组信息包括两个层面, 其一是染色体DNA分子自身结构序列, 其二是细胞内分子和分子机器在其上的坐标序列.

三维基因组涉及所有与染色体DNA相互作用的生化分子和与这些分子构成的直接或间接的作用机制. 如果人们界定一维和二维基因组信息是信息流的话, 那么其三维信息就是在操作流和平衡流所界定的各种生物学因素. 就目前的数据看, 三维基因组只能在细胞和亚细胞水平界定, 即它是某种细胞的三维基因组, 也是细胞自身的自组装形式和过程. 三维基因组包括DNA自身复制(比如复制起点及变化)、所有转录本的表达与调控、与DNA相互作用的蛋白质(机器)、染色质与细胞核结构关系等. 这个维度其实还没有加入细胞周期的概念, 没有时间这个关键的维度轴. 过去的很多研究, 如复制组和转录组, 其实都还停留在三维. 现在比较热门的单细胞转录组也还没有完全脱离这个维度: 从细胞50万到100万个编码蛋白质的转录本或信使转录组(mTranscriptome)中测定几千到上万个转录本来探索转录调控的机制其实还未从20年前的EST(expressed sequence tag)研究(取样1万个转录本)模式中解放出来[23]. 其实质还是从差异表达基因中找故事, 而不是从分子机制和细胞过程中找故事.

四维基因组就是三维基因组在细胞周期中的变化. 这个维度的认知将会把生命科学基础研究带到一个新的境界. 它要求在单细胞水平、单分子分辨率下表征基因和基因产物的功能、作用对象、运动规律、半衰期、数量变化、动态范围等. 尽管细胞周期研究已经相对成熟, 但是每种细胞乃至每个细胞的周期状态并没有准确界定的技术方案, 尤其是不同细胞的周期(如间期)及变化区间、控制因素等. 而且在生命体内不同细胞的增殖、分化、凋亡、连结、通讯等大都不同步, 这些结构与功能异化和可塑性正是构成细胞自组织能力和过程的分子基础. 这个层面的生命活动构成了通生学框架下的分室流. 多细胞真核生物起源于单细胞真核生物, 而单细胞真核生物物种多而复杂且未知者居多, 因此分室流研究具有非常明显的谱系特异性.

五维基因组信息承载的是细胞自组织能力. 以人类为例, 就是从受精卵到各个胚层的有序分化, 从组织和器官的形成到完整的胚胎和胎儿降生, 再从出生到发育成成熟个体全过程中细胞水平的自组织形式和过程, 也是生理系统之间复杂时空关系的终极展示. 这个层面的生命活动主要对应于可塑流. 可塑流在细胞层面研究的对象和内容主要是细胞周期、节律、应激反应、细胞自噬、凋亡等, 而在生理水平则是免疫、肌肉运动、语言学习、成瘾、滞育、休眠、迁徙等复杂生命现象.

六维基因组主要是强调个体完整生命周期过程中的变化, 包括生理(如生殖周期、更年期、衰老期等)和病理(如肿瘤发生)变化. 本文假设生命周期各个阶段内的生理变化都是同质化的, 而其阶段间是高度异质化的, 通过比较建立研究命题. 这个维度的研究应该是生物医学范畴的大多数科学问题.

同理, 这些维度并不是人为分出来的, 而大都是客观存在和科学历史上形成的研究归纳和轨迹. 比如, 基因组学和生物信息学是研究一维和二维基因组信息的学问, 分子生物学和生物化学是研究三维基因组信息的基本手段和方法, 而细胞生物学和发育生物学分别是研究四维和五维基因组信息的基本学科. 可见, 一维到五维的基因组信息主要体现在分子与细胞水平, 有较多的共性和抽象问题, 需要的是技术和方法学; 六维到九维的基因组信息主要体现在生物更高层次的命题: 整体、群体、生态和生物圈, 但是基本上可分解为具体问题和学科(主要是应用学科). 表2列出了多维和五流基因组信息的分层、比较和举例.

3　界定多维基因组信息与通生学“五流”的必要性

首先要解释的是多维染色体与通生学“五流”的关系. 这两个分层的概念并没有矛盾, 只是收敛在不同层面. 前者强调的是可分性, 后者强调的是可收敛性; 前者关注的是信息的相关性, 后者强调的是物质和信息的统一性; 前者可以作为信息和数据的存储、交流结构, 后者则可用于定义概念和贯通知识. 其次, 本文可以通过三个实例讲解和分析分流和分维的必要性和实效性.

第一个例子是基因组学最基本的概念——植物与

表 2　基因组信息的维度与五流的关系

Table 2　The dimensions of genome information and their corresponding tracks

维度	五流权重	维度元素	元素举例
1	信息流	DNA序列与序列的变化	基因组序列
2	信息流与操作两流	DNA序列坐标下的功能性内容	基因与基因组结构元素
3	操平(衡)信三流流	染色体结构与辅件的关联与分子机制	染色体三维构象与基因表达
4	平操分(室)信四流	细胞周期里和细胞过程中的四流关系	细胞周期的能量平衡
5	分平(可)塑操信五流	细胞、器官与生理系统间的五流关系	细胞分化、发育、生殖
6	五流	生命周期的生理与病理状态比较	心脑血管病、恶性肿瘤、糖尿病
7	五流	物种群体与谱系	脊椎动物从简到繁的变演
8	五流	生态群与生态圈	微生物宏基因组、生物多样性
9	五流	生物圈	磷圈、CO_2循环

动物基因组结构的差异. 人们在大量地组装和注释基因组序列时(一维基因组信息)发现, 几乎所有动物相应谱系(如脊椎动物)的基因都远远大于植物(如有花植物)的基因[19]. 在二维注释中发现可能是BDR序列在作祟: 动物基因组插入的BDR序列会落到内含子中, 成为基因的一部分, 而植物的BDR序列则聚集于基因间区. 推而广之的结论是: 动物(如脊椎动物)的基因常随时间的变迁而不断增大, 而植物基因组增大的则多是基因间区(图1). 比如水稻基因组拥有50%的基因空间(gene space, 含有基因的DNA序列), 大麦基因组仅拥有5%的基因空间. 另外, 动物基因组的编码基因累加起来基本上等于基因组大小的90%[13,14].

尽管BDR具有谱系特异性, 但是为什么在动植物基因组间会有这样的差别呢? 答案显然不会是在信息流层面上, 而是在操作流层面上: 是动物基因组中基因间区在抵抗BDR轰炸性插入呢, 还是植物基因组基因空间在执行某种抵抗性"指令"? 答案是后者, 是植物细胞的剪接机器不能够容忍内含子变大[24]. 动植物基因剪接机制的异化可以追溯到它们的单细胞真核生物的各自祖先. 为什么是各自祖先呢? 因为BDR插入何处的决定是基因组变演的最早期机制之一——类动物(animal-like)和类植物(plant-like)基因组. 这两类基因组可以通过基因组序列分析来确定, 比如前者的代表是面包酵母, 后者的代表是卵菌[25]. 同样地, 绝大部分的基因簇(基因共线性的结构基础)非常保守, 其表达常被共同协调[26~28], 也不能随便被BDR的插入所破坏,

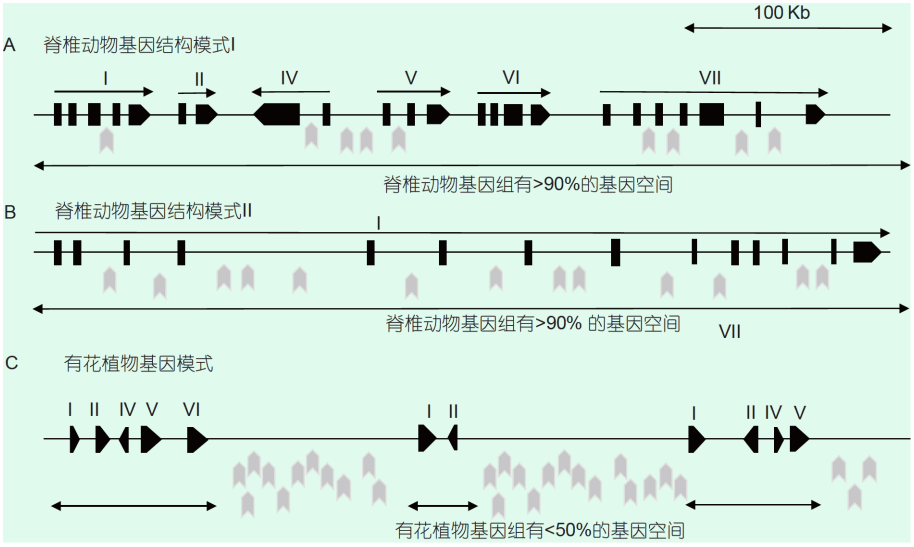

图 1 植物(如有花植物)和动物(如脊椎动物)基因组结构与基因结构均有诸多不同. 脊椎动物基因组有两类结构不同的基因: 基因簇(几个基因形成共调控单元, 占基因总数的70%以上)(A)和超大型基因(>500 kb, 占基因总数的5%左右)(B). 人类基因的平均大小(mean)是100 kb(双尖头实线线标注). 由于重复序列(主要是BDR一类; 灰色箭头所示)不断插入到基因内部(内含子中), 脊椎动物的基因在谱系变演过程中呈不断增大的趋势. C: 有花植物基因组的结构主要是基因簇, 而重复序列主要插入到基因间区(灰色箭头, 距离基因的远近表示插入时间上的差异; 外显子数量和大小没有按对应大小画, 仅为示意图). 因此, 脊椎动物基因组中主要部分为基因空间(虚线双箭头)几乎没有基因间区, 而有花植物的基因间区(主要是LTR类重复序列)一般在50%(拟南芥和水稻)~95%(小麦和大麦)左右. 基因簇和基因(大写罗马数字表示, 横箭头代表基因长度和转录方向)外显子用竖直柱表示(在植物基因组中仅用粗箭头表示), 内含子和基因空间为中央实线

Figure 1 Plants (e.g., Angiosperms) and animals (e.g., Vertebrates) have different genome and gene structures. Vertebrate genes are structured into two different groups: clustered (genes that tend to share their regulatory elements, >70% in number) (A) and super-sized genes (>500 kb, ~5% in number) (B). The mean size of the human genes is about 100 kb (double-headed arrows). As repeats being inserted into the genome continuously (mainly BDR; grey arrowheads), mostly intronic spaces, vertebrate genes appear to be increasing in size. C: Similarly, angiosperm genomes are also composed of clustered genes but repeats in this case are inserted into intergenic spaces (grey arrowheads), and the distance toward the horizontal line (genome sequence) indicates timing of repeat insertions; the size of exons is not drawn in real proportion. Consequently, most vertebrate genomic spaces are genic (dashed lines with double-headed arrows) with little intergenic spaces; in sharp contrast, there are large intergenic spaces in angiosperm genomes, where repeats are mainly LTR related. For instance, such repetitive sequences make up ~50% (in rice and *Arabidopsis*) to ~95% (wheat and barley) of the entire genome lengths. Gene clusters and genes (in Roman numbers) are indicated with horizontal arrows as transcription directions are pointed out. Exons are indicated with vertical solid bars, whereas in the plant genome they are depicted with thick arrows; introns and intergenic spaces are depicted with horizontal lines

但是发生全基因组复制后的多倍化基因组除外. 可见以脊椎动物和有花植物为代表的动植物基因组染色体结构对BDR的插入均有分子水平的限制(图1) .

然而, 在信息流和操作流层面上还不能解释为什么动植物的祖先做了目标迥异的决定性选择: 动物基因组选择将BDR纳入到基因的内含子部分, 而植物选择了将其插入到基因之外的基因间区. 这个决策导致复制需求(主要是dNTP池——胞内全部游离脱氧核苷三磷酸单体)与转录需求(主要是NTP池——胞内全部游离核糖苷三磷酸单体)在两个主要多细胞生物谱系产生了最大需求上的差异(图2). 如果是这样, 在平衡流(能量、物质和信号传导构成的稳态)层面上就有了合理的解释. (ⅰ) 对于细胞整体而言, 其S期(DNA合成期)所使用的能量主要是用于DNA合成, 即细胞分裂前染色质和后续染色体形成. 以复制哺乳动物细胞的两倍体基因组为例, 至少需要两倍于二倍体DNA等量的脱氧核苷三磷酸, 约为10^{10}个dNTP分子, 而转录时则需要至少5×10^{9}个NTP分子(实际最高可能是这个数的数倍左右), 而细胞游离的核苷三磷酸也大致在这个数量级上. 因此, 在植物中, 相对于复制负担而言, 其转录能量负担则会至少减少50%(拟南芥和水稻)~95%(大麦和小麦), 这些多出来的能量会用于其他生物大分子的合成, 如淀粉、纤维素、半纤维素、木质素和其他诸多植物特异的次生代谢产物. (ⅱ) 这个“复制-转录能量负担”的分配显然与剪接机制一样在各自的单细胞祖先时代就已经决定了. 此外, 在更高的维度即分室流和可塑流层面, 人们可以推测动植物基因调控会不断异化, 植物可以果实和种子的多样性与动物的运动性和行为可塑性(如滞育、迁徙和休眠)竞争.

再举一个例子来强调界定“五流”主体的意义, 那就是生物医学最前沿的恶性肿瘤研究. 尽管就基因组信息维度而言, 肿瘤学属于正常生命周期与相应病理现象比较的第五个维度, 但是就“五流”思维框架而言, 肿瘤生物学的本质或主体仍是分室流. 虽然有“癌症是遗传病(强调原癌基因变异)”“癌症是环境病(强调致癌物因素)”“癌症是代谢病(强调代谢失调)”等说法, 但几十年来癌症研究权威性总结指出的十种特征(hallmarks)都是细胞层面的现象[29,30], 是细胞在“五流”框架下的不同分子机制(如DNA损伤和转录调控)和细胞过程(如能量代谢途径)的种种异化[31], 并伴随原发后的器官转移和免疫逃逸等更高维度的变化来实现无限

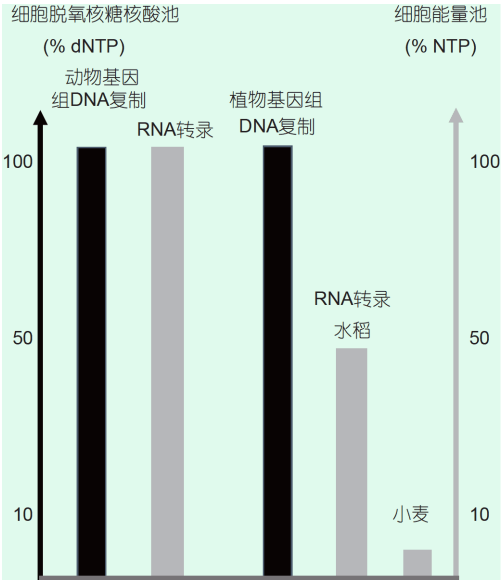

图 2　细胞内平衡流物质和能量分配平衡假说. 假设基因组DNA复制(黑色柱)需要100%的dNTP存储, 那么动物基因组转录(灰色柱)时所需要的能量也应该在同样的数量级. 然而在植物基因组转录时, 所需要的能量则有很大的变化, 从水稻和拟南芥的50%到小麦和大麦的5%不等. 所节省的能量因该被用到其他植物特有的生物大分子(纤维素、半纤维素、淀粉、木质素等)合成和抗逆次生代谢产物等方面

Figure 2　Homeostasis of cellular matters and the energy equal-sharing hypothesis. If we assume that genome replication consumes 100% of the dNTP reserve (vertical solid bars), transcription in animals may consume a similar amount of the NTP reserve (vertical grey bars). However, the transcribed fraction among plant genomes is not approximately equal to the full genome length but ~50% in rice and *Arabidopsis* and ~5% in wheat and barley. The rest of the NTP reserves are assumed to be allocated to other synthetic activities of cellulose, hemicellulose, starch, and lignin, as well as some intermediate metabolites

制增殖. 因此, 恶性肿瘤的分子生物学研究手段也多以细胞模型和动物模型为基础.

本文还要举另外一个例子来强调“五流”思维和多维基因组信息的分层并不是某种拼凑和牵强附会. 比如宏基因组(metgenomics)研究的实质是以第八维基因组信息为基线, 以平衡流为主体的“五流”研究, 是生物圈体系下的超种群(或称跨种群)微生物群体的生态研究. 宏基因组首先是在生物圈历史性地相对固化[32], 然后在地表圈和地球外圈中的大气圈、水圈和生物圈物中持续变演, 再与物中个体共生形成特有的微环境宏基因组. 由于在一个相对固化的环境里, 构成宏基因组的微小生物(包括微小原生动物、单细胞真核生物、古细菌、真细菌和各类寄生性病毒)之间有频繁的基因、代谢产物和生化信号分子交流, 其“五流”主

体之首是平衡流. 然而, 宏基因组的成员毕竟形成一个有机的生态环境, 并具有一定的稳定性和可变性且与生态环境息息相关, 因此其信息具有较高的维度. 例如生活在自然环境中的蜱虫宏基因组, 经过数代无菌条件下的人工繁殖, 从拥有来自29个门1153个属4625个物种减少到来自8个门69个属的70物种[33]. 这个结果展示出了在这个微小的生态圈里, 微生物物种间的“和谐”和物种内的“排斥”——70个物种来自69个属. 不仅如此, 寄生在同一天然物种(如田鼠)的蜱虫具有比宿主更多样宏基因组[33,34]. 因此, 不能说这样的变化不是一个独特而又有意义的“生态圈”, 它带有环境和物种本身的诸多信息, 对疾病的防控具有不可忽视的科学价值.

4 基因组多维信息、染色体多维结构研究的技术瓶颈与前景展望

多维基因组信息的概念很难独立存在, “九维”与“五流”的并存是必然的. 首先, 维度的划分总会有一定的模糊边界, 因为生命体毕竟是一个有机的整体, 细分和收敛会并存, 而“五流”的划分是基于生命的物质基础, 尤其是生命体物质的机化和时空性质, 以及相应的自组装和自组织特征. 其次, 就理论而言, 高纬度信息其实是一定要涵盖低纬度信息的, 而“五流”却不尽是如此. 比如平衡流描述构成生命体物质的动态特征, 分室流表述细胞的自组织特征, 可塑流描述的是生命体与环境的应答特征, 更多地强调有机体组分和结构之间的关系 .

多维基因组信息有组织地获取, 就技术而言面临来自三个层面的挑战: 系统性、规模化、精准性. 系统性是指每个层面上的信息必须要相对应, 从染色体DNA序列到基因在染色体上的坐标, 从基因转录本到功能产物和细胞机器, 从分子机制到细胞过程等, 都是细胞的有序活动, 这些活动必须在一个相顺相通、相辅相成的平台上来解释. 这不仅需要有规划的数据获取, 还要准备强大的算力算法支持. 目前不同维度的组学数据可以罗列的有很多(表3), 并且可以无限细分下去, 但是除了基因组序列——一维基因组信息外, 其他层面都还缺乏计划性和完整性的思考. 人类基因组序列测定从30年前一个人的“人类基因组计划”到如今的“十万基因组”(英国)、“百万基因组”(英国), 甚至“人人基因组”(美国)计划, 成功在于“五位一体”地思考: 建立机构、孵育学科、制定规划、突破技术瓶颈、引导新业态产生. 其中技术瓶颈的突破最为重要, 试想如果没有第二代测序仪(亦称“下一代测序仪”或next-generation-sequencers, NGS)的支持, 这些基因测序计划不都是“天方夜谭”吗? 就目前技术汇聚的现状来看, 规模化与微型化是必然的. 所谓微型化, 就是实现以微流控为核心技术的生化检测(assays)在“片上”(亦可称为芯片上系统或片上系统, 即on-chip; 本文简称“片上”)的量产. 片上操作的必要性是很显然的, 就是要保证低成本、高通量和高速度. 第三代和第四代测序仪就是以片上技术——包括“湿法”(微流控)和“干法”(微电子)——为基本平台. 精准性要通过技术和方法学, 以及重复性、高覆盖数据来实现. 与技术相关的考量之一就是准确性, 就DNA测序而言, 有两个方面: 单核苷酸的准确性(也与序列的覆盖度有关)和序列的连续性(与测序读长有关); 而另一个重要的考量就是成本. 在降低成本的压力下, 很多技术、方法和项目都打了折扣, 其中相当一部分数据难以重复验证, 生命力因此缩水, 比如早期转录组研究的方法, 如SAGE(serial analysis of gene expression)、微阵列(microarrays)、EST等[35]. 数据的准确性是质量保证, 也是数据的生命力所在. 原则上讲, 材料和数据的处理流程越简便, 其数据的真实度和可信度就越高. 细胞RNA组研究(通常指信使转录组)的起点就应该是单细胞转录组本身及其异质化特征(同相同克隆细胞作为一个有物质和信号交流的群体), 然而人们却没有理想的定量技术和设备表征异质化. 目前的单细胞信使转录组研究正在从以界定细胞种类和表达谱差异为主的研究, 转变到转录组本身的研究[36,37]并形成具有系统性、初具规模的一系列大科学项目[38,39], 但是数据的生命力取决于质量, 取决于RNA直测技术和微流控单细胞建库技术的进展.

作为终极目标并考虑RNA中修饰核苷酸的定位、定量, 多维基因组信息获取需要满足“双单四定”(即在单细胞水平实现单分子分辨率的基因产物和细胞组分的定性、定位、定时和定量)的技术指标. 尽管目前的高通量DNA测序和超分辨光学观测技术可用于推动RNA组和三维染色体构象研究[40], 但是还没有能提供四维空间中这些属性较完整的动态数据. 实现“双单四定”的基本前提至少有如下5个方面的工作:

表 3　不同维度基因组信息举例

Table 3　Examples of genome information at various dimensions

细分图谱	信息维度	备注
STR位点	1	不同重复单元的简单重复序列在群体水平的变化
BDR位点	1,2	生物学界定的重复序列位点和群体中动态变化图谱; 可部分拓展为表达谱
假基因座位	1,2	各类假基因座位与谱系发生的起源与变演史
染色体复制起点	1,2	DNA复制的起点具有不确定性, 需在细胞水平确定与邻近基因和染色体结构的相对定位
染色体甲基化	1,2	单细胞分辨率下染色体甲基化位点与动态; 物种内和跨不同物种的谱系内比较
基因调控顺式元件	1,2	全基因组位点谱与细胞特异性元件组合使用; 可拓展为细胞、发育等特异图谱
基因簇	1,2	两个基因以上、基因相对位置特异、有无反向调控等全涵盖性参数; 可拓展为转录组
CpG岛	1~3	全基因组CpG位点以及附近基因座的关系; 可拓展为细胞、发育等特异图谱
核小体占位	1~3	基因座位特异和细胞特异的核小体分布; 可拓展为细胞、发育等特异图谱
标准外显子	1~3	标准外显子在基因和染色体上的分布; 可拓展为转录组
线粒体复制	1~4	细胞周期下线粒体动态复制窗口; 可拓展为细胞特异图谱
rDNA转录	1~5	rDNA转录组与细胞周期的关系, 以及细胞特异、组织特异性表达
rRNA修饰	1~5	rRNA修饰组与细胞周期的关系, 以及细胞特异、组织特异性表达
tDNA转录	1~5	tDNA转录组与细胞周期的关系, 以及细胞特异、组织特异性表达
tRNA修饰	1~5	tRNA修饰组与细胞周期的关系, 以及细胞特异、组织特异性表达
组蛋白修饰	1~5	组蛋白修饰组与细胞周期的关系, 以及细胞特异、组织特异性表达
管家基因谱	1~5	所有细胞都表达的基因和普适性验证
细胞特异基因谱	1~5	细胞特异表达基因
节律调控基因谱	1~5	节律调控基因, 以及细胞和器官表达的特异性

STR, 简单串联重复序列; BDR, 生物学界定的重复序列, 如各类转座子

(ⅰ) 将前两个维度的信息流数据有效地整合在染色体序列(具有种群特异性, 可以通过同源序列界定一个虚拟个体序列[41,42])的坐标上. (ⅱ) 实现RNA直测和单细胞蛋白质组定量技术体系. (ⅲ) 建立三维基因组信息网络, 形成有序、详尽、可视的静态虚拟细胞, 动态界定组织特异性vs.看家基因[43]和干细胞vs.分化细胞等. (ⅳ) 将虚拟细胞激活, 包括平衡流和分室流信息, 并建立细胞特异的动态信息库. (ⅴ) 确立五维和六维基因组信息的科学目标, 探讨实现与这些目标相对应的系列大科学计划.

多维基因组研究对科学家们的挑战是建立协同合作机制和组织实施规划, 尤其是要识别和准备相应技术和组织各个分项的人才团队. 人类基因组计划和精准医学计划的启动和实施以及“五位一体”的内涵, 为未来的大科学计划树立了不朽的典范. “五位一体”的核心是人才培养和技术突破; 人才培养需要时间和延续性, 技术突破不仅需要人才, 还需要有特定目标和有组织攻关. 那些常规的惰性思维方式、固有的技术和方法都无助于未来科学发展的需求, 也跟不上时代的步伐. 事实是目前还没有突破基于单细胞转录组研究的数据瓶颈和单细胞蛋白质组技术的瓶颈. 比如系统建立表皮(epidermis或keratinocytes)细胞从干细胞到分化细胞的基因表达图谱(例如[36]).

5　结束语

多维基因组信息的分析与获取可分为两个部分: 一至五维的“五流”分开和“五流”通悟的六至九维. 前者强调通用性, 其性质是分子(机化原理)和细胞水平(时空原理)的基础研究和技术的分层突破; 后者强调思维框架的整体性和实用性, 是相应领域的应用型研究, 用于解释和揭示生命现象(科学命题)的分子细胞机制. 在未来的5~10年里, 生命科学期待一系列新的、可实现的、能与人类基因组计划相媲美的大科学

表 4 待完成的重大科学研究项目与技术举例

Table 4 Examples of major scientific projects and supporting technologies

五流	项目举例	技术和试验模型举例	终极目标
操作流	人类RNA组计划	单细胞组分分流, 单分子RNA测序, RNA修饰定位	细胞特异的RNA组、信使RNA组、lncRNA组等
操作流	人类小RNA组计划	同上	细胞特异转移RNA组、微小RNA组等
平衡流	人类血与尿代谢组	蛋白质、核苷酸、高能键、生化小分子等组分定量	人类体液组分变化与疾病、生理周期的关联
分室流	人类细胞组计划	细胞微量组分分流, 片上细胞, 片上实验室技术	人类细胞的实验和研究模型
可塑流	人类器官组计划	片上细胞、片上器官技术	器官实验和研究模型
可塑流	人类病理组计划	规模化细胞模型和动物模型建立	按器官、生理体系等以疾病分类的病理学研究
信息流	人类关联组计划	高通量基因分型	基于队列数据的罕见病与复杂疾病的关联研究
可塑流	人类寿命组计划	脊椎动物谱系动物模型体系	基于脊椎动物生命周期变化特征的多维信息整合

计划. 这些大科学计划包括(但不限于)谱系、物种群体、生理病理、发育与器官、细胞与分化、细胞组学等各个水平与层面(表4). 尽管有些项目似乎已经启动(如细胞RNA组), 但是目前诸多技术瓶颈其实还未真正突破, 尤其是实现单细胞、单分子(或分子高分辨率)水平分析的核心技术(比如微纳流控和片上实验室等), 尚需要长期的人才培养和数年、十数年甚至数十年坚持不懈的努力.

致谢 感谢中国科学院北京基因组研究所楚亚男博士帮助整理参考文献和校对本文. 也感谢三位匿名评审专家的建设性意见.

参考文献

1 Yu J. Genomics and Precision Medicine (in Chinese). Shanghai: Shanghai Jiao Tong University Press, 2017 [于军. 基因组学与精准医学. 上海: 上海交通大学出版社, 2017]

2 Zhao Y F, Lu K. Basic chemical rule of molecular evolution (in Chiese). J Xiamen Univ (Nat Sci), 2001, 40: 360–365 [赵玉芬, 卢奎. 分子进化的基本化学规律. 厦门大学学报(自然科学版), 2001, 40: 360–365]

3 Becker S, Feldmann J, Wiedemann S, et al. Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. Science, 2019, 366: 76–82

4 Crick F H. The origin of the genetic code. J Mol Biol, 1968, 38: 367–379

5 Wong J T. A co-evolution theory of the genetic code. Proc Natl Acad Sci USA, 1975, 72: 1909–1912

6 Yu J. A content-centric organization of the genetic code. Genom Proteom Bioinf, 2007, 5: 1–6

7 Xiao J F, Yu J. A scenario on the stepwise evolution of the genetic code. Genom Proteom Bioinf, 2007, 5: 143–151

8 Zhang Z, Yu J. On the organizational dynamics of the genetic code. Genom Proteom Bioinf, 2011, 9: 21–29

9 Zhang Z, Yu J. Does the genetic code have a eukaryotic origin? Genom Proteom Bioinf, 2013, 11: 41–55

10 Yu J, Wong G K S, Wang J, et al. Shotgun Sequencing (SGS). 2nd ed. Hoboken: Wiley-VCH, 2006

11 Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science, 2002, 296: 79–92

12 Yu J, Wang J, Lin W, et al. The genomes of *Oryza sativa*: A history of duplications. PLoS Biol, 2005, 3: e38

13 Wong G K, Passey D A, Huang Y, et al. Is “Junk” DNA mostly intron DNA? Genome Res, 2000, 10: 1672–1678

14 Wong G K, Passey D A, Yu J. Most of the human genome is transcribed. Genome Res, 2001, 11: 1975–1977

15 Wang D, Su Y, Wang X, et al. Transposon-derived and satellite-derived repetitive sequences play distinct functional roles in mammalian intron size expansion. Evol Bioinform Online, 2012, 8: 301–319

16 Zhang L, Xiao M, Zhou J, et al. Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a

jellyfish-based LAUPs analysis application (JBLA). Bioinformatics, 2018, 34: 3624–3630
17 Xiao M, Yang X, Yu J, et al. CGIDLA: Developing the web server for CpG island related density and LAUPs (lineage-associated underrepresented permutations) study. IEEE/ACM Trans Comput Biol Bioinf, 2019, doi: 10.1109/TCBB.2019.2935971
18 Zhang L, Dai Z, Yu J, et al. CpG-island-based annotation and analysis of human housekeeping genes. Briefings BioInf, 2020, doi: 10.1093/bib/bbz134
19 Chen K, Wang L, Yang M, et al. Sequence signatures of nucleosome positioning in *Caenorhabditis elegans*. Genom Proteom Bioinf, 2010, 8: 92–102
20 Chen K, Meng Q, Ma L, et al. A novel DNA sequence periodicity decodes nucleosome positioning. Nucleic Acids Res, 2008, 36: 6228–6236
21 Cui P, Zhang L, Lin Q, et al. A novel mechanism of epigenetic regulation: Nucleosome-space occupancy. Biochem Biophys Res Commun, 2010, 391: 884–889
22 Cui P, Lin Q, Zhang L, et al. The disequilibrium of nucleosomes distribution along chromosomes plays a functional and evolutionarily role in regulating gene expression. PLoS ONE, 2011, 6: e23219
23 Wu J, Xiao J, Zhang Z, et al. Ribogenomics: The science and knowledge of RNA. Genom Proteom Bioinf, 2014, 12: 57–63
24 WU J, XIAO J, WANG L, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. Sci China Life Sci, 2013, 56: 968–974
25 Haas B J, Kamoun S, Zody M C, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature, 2009, 461: 393–398
26 Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) among *Drosophila* and vertebrate genomes. BMC Evol Biol, 2009, 9: 55
27 Cui P, Liu W, Zhao Y, et al. Comparative analyses of H3K4 and H3K27 trimethylations between the mouse cerebrum and testis. Genom Proteom Bioinf, 2012, 10: 82–93
28 Cui P, Liu W, Zhao Y, et al. The association between H3K4me3 and antisense transcription. Genom Proteom Bioinf, 2012, 10: 74–81
29 Hanahan D, Weinberg R A. Hallmarks of cancer: The next generation. Cell, 2011, 144: 646–674
30 Fouad Y A, Aanei C. Revisiting the hallmarks of cancer. Am J Cancer Res, 2017, 7: 1016–1036
31 Liberti M V, Locasale J W. The Warburg effect: How does it benefit cancer cells? Trends Biochem Sci, 2016, 41: 211–218
32 Wu H, Fang Y, Yu J, et al. The quest for a unified view of bacterial land colonization. ISME J, 2014, 8: 1358–1369
33 Sui S, Yang Y, Sun Y, et al. On the core bacterial flora of *Ixodes persulcatus* (Taiga tick). PLoS ONE, 2017, 12: e0180150
34 Estrada-Peña A, Cabezas-Cruz A, Pollet T, et al. High throughput sequencing and network analysis disentangle the microbial communities of ticks and hosts within and between ecosystems. Front Cell Infect Microbiol, 2018, 8: 236
35 Lowe R, Shirley N, Bleackley M, et al. Transcriptomics technologies. PLoS Comput Biol, 2017, 13: e1005457
36 Finnegan A, Cho R J, Luu A, et al. Single-cell transcriptomics reveals spatial and temporal turnover of keratinocyte differentiation regulators. Front Genet, 2019, 10: 775
37 Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity. Genome Biol, 2019, 20: 110
38 Regev A, Teichmann S A, Lander E S, et al. The human cell atlas. eLife, 2017, 6: e27041
39 Ponting C P. The human cell atlas: Making ‘cell space’ for disease. Dis Model Mech, 2019, 12: dmm037622
40 Lee D S, Luo C, Zhou J, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat Methods, 2019, 16: 999–1006
41 Ling Y, Jin Z, Su M, et al. VCGDB: A dynamic genome database of the Chinese population. BMC Genomics, 2014, 15: 265
42 Du Z, Ma L, Qu H, et al. Whole genome analyses of Chinese population and *de novo* assembly of a Northern Han genome. Genom Proteom Bioinf, 2019, 17: 229–247
43 Zhu J, He F, Hu S, et al. On the nature of human housekeeping genes. Trends Genets, 2008, 24: 481–484

Multi-dimensional genomics information: origin, context, and technical bottlenecks

YU Jun[1,2]

1 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; 2 University of Chinese Academy of Sciences, Beijing 100049, China

Multi-dimensional structural studies of chromosomes have met some conceptual hurdles and technical bottlenecks, albeit making significant headways in recent years, so that it is necessary to discuss and clarify several key relevant concepts in a top-down fashion. Life might have started with molecular operations (RNAs, proteins and other macromolecules) rather than information and its inheritability (DNA). Therefore, when DNA became the Chosen One to shoulder genetic information, its multi-potential nature as a set of intracellular macromolecules has been exploited in time and space since, whose operational and homeostatic roles at all levels form a substantial network of harmonious molecular mechanisms and cellular processes. Such a network, defined in molecular details at mechanochemical and spatiotemporal dimensions, along with its rules and complications, has made up the core themes and countless scenarios of modern biology.

genome, chromosome, gene expression regulation, multi-dimensional genome information

doi: 10.1360/SSV-2019-0276