中国科学: 生命科学
SCIENTIA SINICA Vitae
life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述 中国知名大学及研究院所专栏 中国科学技术大学生命科学学院专辑

生物信息技术在表观遗传调控机制研究中的应用

金晨, 瞿昆*

中国科学技术大学生命科学学院, 合肥 230027

* 联系人, E-mail: qukun@ustc.edu.cn

收稿日期: 2016-11-08; 接受日期: 2016-12-07; 网络版发表日期: 2017-01-19

摘要 表观遗传调控是指在不改变实际DNA序列却控制基因表达的过程, 并在决定细胞功能和发育中起着至关重要的作用. 表观遗传基因组包括组蛋白的位点、排列, 以及化学修饰、DNA 甲基化、非编码 RNA 的结构和表达, 以及转录因子调控网络和染色体三维结构等各种因素共同协调作用来控制细胞表型. 尽管复杂, 各种二代测序, 尤其是生物信息学技术的发展, 极大地促进了人们对于表观遗传机制的理解. 本文简要介绍当下生物信息技术在揭示表观遗传调控机制研究中的应用, 从而为更深入理解整个基因调控机制, 乃至为人类疾病治疗提供新的见解.

关键词 生物信息, 非编码 RNA, 表观遗传调控, 基因调控网络, 云计算

越来越多的证据显示, 基因的表达调控与疾病的发生有着重要的联系, 基因的正确表达在机体功能的实现过程中发挥关键性的作用. 在调控基因表达的诸多因素中, 表观遗传调控是指在不改变实际 DNA 序列却控制基因表达的过程, 它在决定细胞功能和发育中起着至关重要的作用. 整个表观遗传基因组是包括组蛋白的位点、排列, 以及化学修饰, DNA 甲基化, 非编码 RNA 的结构和表达, 以及转录因子调控网络和染色体三维结构等各种因素在内的复杂体系, 它们共同协调作用来控制细胞表型(图 1). 例如, 沉默子(Silencer)、绝缘子(Insulator)、增强子(Enhancer)、启动子(Promoter)、非编码 RNA(lncRNA)等表观遗传调控原件, 都是在没有改变 DNA 序列的情况下调控这些基因表达的过程. 近年来, 各种二代测序和生物信息学技术的发展, 极大地促进了人们对于表观遗传机制的深入理解. 本文将从 RNA, DNA 以及方法论三部分, 简要阐述当下生物信息技术在揭示表观遗传调控机制研究中的应用, 从而为人类提供疾病治疗的新见解.

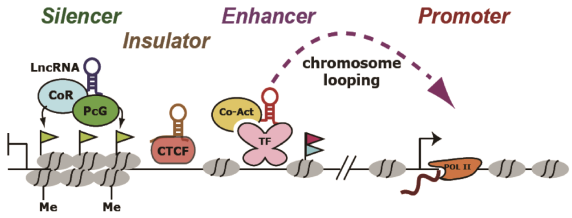

图 1 基因表观遗传调控特征示意图(网络版彩图)

引用格式: 金晨, 瞿昆. 生物信息技术在表观遗传调控机制研究中的应用. 中国科学: 生命科学
Jin C, Qu K. Application of bioinformatics techniques in revealing the mechanisms of epigenetic regulation. Sci Sin Vitae, doi: 10.1360/N052016-00316

1 非编码 RNA 对基因的转录和转录后调控

一直以来认为, 基因的功能主要是通过 DNA 转录成 RNA(主要是编码 RNA)再翻译成蛋白质, 以蛋白质为主体来实现的, 即所谓的中心法则. 随着基因测序和分析技术的发展, 人们发现有 90%以上的人类基因转录产物并没有作为合成蛋白质的模版, 而是以 RNA 的形式存在于细胞中(即非编码 RNA). 最近几年, 越来越多的证据证实非编码 RNA 通过调控基因转录、翻译, 以及翻译后修饰等来调节细胞发育、分化、癌变等重要的生物学进程, 在包括心脏病、自身免疫性疾病以及癌症等一系列重大疾病中都起着非常重要的调控作用[1~8]. 同时期, 各种二代测序技术的发展产生出海量基因组学数据, 如何深度分析这些生物大数据资源, 来系统性研究非编码 RNA 的功能和机理已成为当下研究的热点问题.

1.1 非编码 RNA 的发现和全转录组测序

二代测序技术出现以前, 在基因组中搜索非编码 RNA 并非易事, 因为这些 RNA 并不翻译成蛋白质, 也没有明确的 RNA 序列, 属于转录组当中的"暗物质". 后来, 人们发现, RNA 的转录和组蛋白 H3 的两个氨基酸的三甲基化有着密切的关联, H3 赖氨酸 4 的三甲基化(H3K4me3)信号和基因转录的起始位点(启动子)有很强的相关性, H3 赖氨酸 36 的三甲基化(H3K36me3)则伴随着整条 RNA 转录区域的延伸. 利用 H3K4me3 和 H3K36me3 组合信号, 就可以估计在 DNA 的哪些区域可能发生着基因转录. 第一个反式调控基因表达的 lncRNA *Hotair*, 就是通过 H3K4me3 和 H3K36me3 组合信号 tiling array 芯片技术被发现的[6].

转录组测序(RNA-seq)技术的诞生极大促进了非编码 RNA 研究领域的发展, 种类和数量巨大的非编码 RNA 在从细菌到人的各种生物体系以及各类细胞中被发现[9~11]. 与此同时, 为了分析海量的转录组数据, 从中精选出有用的信息, 一系列生物信息分析软件和算法被陆续发展出来. 其中最为著名的有 Tophat/Cufflinks[12,13], Scripture[10] 和 Trinity[14,15]. Tophat 的核心思想是利用已知的 DNA 序列做参考, 在充分考虑 RNA 剪切配对的情况下, 将由 RNA-seq 产生的短序列(Read)比对到参考序列上面. Cufflinks 和 Scripture 则利用已有的基因标注信息, 将 Tophat 比对的结果, 拼接成相对完整的转录组. 相比之下, Trinity 可以在没有任何先验 DNA 信息的基础上, 完成转录组的 *de novo* 组装, 因此在计算资源上有更高的需求并且在计算时间上也相对要长很多, 所以通常只作为一个辅助手段. 在获得转录组之后, 过滤掉已知的编码基因, 留下潜在的"非编码"基因. 研究人员可以轻松获取这些基因的序列, 以此来计算它们转录成蛋白质的可能性. CPC(coding potential calculator)[16]这样的工具就可以直接利用 RNA 序列计算出该 RNA 是否可能是真正的非编码 RNA. DESeq[17]和 edgeR[18]是两款最常用的 R 软件, 用来计算在不同条件下 RNA 表达是否存在显著性差异, 以此来进一步过滤出具有生物学功能的非编码 RNA.

正是利用这一系列生物信息学方法陆续发现了一大批与细胞分化、炎症、癌症等相关的非编码 RNA. 例如, 本实验室发现, lncRNA *TINCR*[2]促进皮肤干细胞向角质细胞的分化, *ANCR*[19]则会抑制这种分化过程. 还发现 lncRNAs *SeCATs* 调控一种名为 Sézary syndrome 的极富有攻击性的 T 细胞淋巴瘤[20]; lncRNA *BANCR* 调控黑色素瘤癌细胞转移[21]. 此外, 还发现 lncRNA *Lethe* 被经由 NF-kB 的致炎症的细胞因子特异性激活, 并且对 NF-kB 信号通路产生负反馈[22]. 随着更多相关工作在不同物种的不同体系中展开, 这一整套通过对转录组数据的分析, 搜寻新的功能性非编码 RNA 的流程已经逐渐成为领域内的标准.

1.2 非编码 RNA 的功能和调控机制

目前, 对非编码 RNA 的研究热点之一是在探索和发现新的非编码 RNA 在不同体系中的功能和机理[23]. 迄今为止, 已知的非编码 RNA 对基因表达的调控主要通过以下 4 种方式[24]: 信号、诱饵、向导和支架. 鉴于目前研究中还存在各种困难, 一方面, 需要发展新的二代测序技术, 在全基因组层面研究 RNA 和 DNA, RNA 和 RNA, 以及 RNA 和蛋白质之间的互作机制; 另一方面, 需要开发新的生物信息学方法来整合分析、利用这些生物大数据, 通过机器学习, 准确预测非编码 RNA 的功能以及它们对生命活动的调控机理, 并且由此建立一整套系统性研究非编码 RNA 的生物信息学方法、数据库和软件系统. 本实验室在这两方面都做了大量的工作.

由于相当数量的已知非编码 RNA 对基因的调控

作用是通过对组蛋白的表观遗传修饰来实现的, 因此了解某个非编码 RNA 和染色质的精确结合位点, 对于揭示该 RNA 的功能和机理有非常重要的意义. 本实验室发展了 ChIRP(chromatin isolation by RNA purification)测序技术来检测 RNA 在染色质上的结合位点[25]. ChIRP-seq 通过寡脱氧胸苷酸以探针的形式和 RNA 耦合在一起, 然后再通过磁珠将和该 RNA 黏合的蛋白以及 DNA 沉淀下来, 在分离蛋白和 DNA 之后对 DNA 进行测序, 得到 RNA 和染色质的结合位点. 由于探针本身也会对 DNA 形成交联耦合, 因此需要利用奇偶两套探针, 做两次实验来过滤假阳性, 提高信噪比. 为此设计了一种单核苷酸限制分析方法, 通过对由两个正交的探针组产生的 DNA 序列的整合分析, 即在单核苷酸层次比较标准化后的 ChIRP-seq 数据, 来有效地过滤假阳性, 增强信噪比. 并开发了新的生物信息分析方法和工作流程, 能够在全基因组层面上快速定位非编码 RNA 在染色质上的结合位点. 在改进后的实验和计算机分析方法(dChIRP)[26]中, 通过整合果蝇染色体 3D 结构等海量公共高通量数据, 首次揭示了两个重要的长非编码 RNA, *roX1* 和 *roX2* 在果蝇性染色体计量补偿中的不同作用机理. 分离出来的蛋白质通过质谱分析后, 得到和该 lncRNA 互作的蛋白质组(ChIRP-MS 质谱). 通过 ChIRP-MS 精确地发现了 81 个和 lncRNA *Xist* 结合的蛋白, 并且揭示了 *Xist* 在调控 X 染色体失活中的新机制[27].

也有部分非编码 RNA 富集在细胞质中, 它们的功能可以通过与其他 RNA 或蛋白耦合来实现, 调控基因转录后修饰. 本实验室发展了 RIA-seq(RNA interactome analysis)技术, 来检测 RNA 和 RNA 之间的相互作用[2]. RIA-seq 和 ChIRP-seq 的共同点在于利用奇偶两套探针来过滤假阳性, 不同的地方在于前者是对 RNA 测序, 后者是对 DNA, 因此对 RIA-seq 的数据处理, 更多采用改进版的 RNA-seq 的处理技术.

RNA 的二级结构对基因的功能和调控也至关重要, 直接影响着 RNA 的转录、转运、翻译等最基本的生物学过程. 本实验室发展了 PARS-seq(parallel analysis of RNA structure)技术, 首创全新的计算机分析方法, 得到整个转录组中每一个 RNA 上每个碱基单链和双链的概率[28~30]. 该技术的核心是发现 RNA 剪切酶 S1 和 V1 在 RNA 上的剪切位点. 运用该方法, 首次揭示了人类全转录组 RNA 的二级结构, 发现 RNA 结合蛋白 Argonaute 在 mRNA 上的结合位点多属于单链结构. 接着将 PARS-seq 应用到测量一个家族系的全转录组 RNA 二级结构, 通过对比父本和母本等位基因二级结构, 再结合蒙特卡洛模拟, 首次观察到 riboSNitch, 即可以改变 RNA 二级结构的单核苷酸多态性(SNP)[29]. 为了研究 RNA 二级结构随环境的变化, 测量在不同温度下酵母细胞中 RNA 的二级结构[31]. 提出了一种新的阶越函数算法, 通过建立数据模型, 揭示了 mRNA 在 5’端和 3’端的不同稳定性; 计算在全基因组层面每个 RNA 折叠或展开时所需要的能量, 由此预测非编码 RNA 二级结构在总体上比 mRNA 更稳定, 并取得实验验证. 此外, 新的检测 RNA 二级结构的方法也在不断更新中, icSHAPE 方法可以直接在活体细胞中检测全转录组层面的 RNA 二级结构[32], PARIS 则可以检测更远距离的 RNA 二级折叠[33]. 这些实验技术和分析算法对更深入了解非编码 RNA 的调控机制起到了积极的推动作用.

此外利用公共数据对非编码 RNA 功能的理论预测也已经取得一些进展. 例如, 有报道称通过对大量(尽管不是最新)转录组数据的相关性分析, 可以准确预测非编码 RNA 的功能[34]; 有报道利用先前癌症基因芯片数据, 来预测非编码 RNA 在癌细胞中的调控作用[35]; 也有报道开发分析软件和数据库研究非编码 RNA[36~38]等. 然而目前为止, 如何利用最新最全的遗传和表观遗传数据, 应对新的测序技术产生的新的数据类型, 来全方位、系统性地研究非编码 RNA 的功能和作用机理, 也还需要更多的工作. 希望能够发展新的生物信息学方法, 整合包括全基因组、全转录组、转录因子免疫沉淀、RNA 二级结构、RNA 结合蛋白组、DNA 以及组蛋白表观遗传修饰和 DNA 调控组等在内的各种公共组学数据, 来预测非编码 RNA 的结构、功能以及在各类细胞和疾病中对生命过程的调控机理(图 2). 这些生物信息分析技术的发展和广泛应用极大地促进了人们对基因表观遗传调控的认识.

2 表观遗传调控和基因调控网络

2.1 转录因子和基因调控网络的构建

转录因子是基因调控中最重要的部分之一, 由转录因子构成的调控网络也是整个调控机制中最直

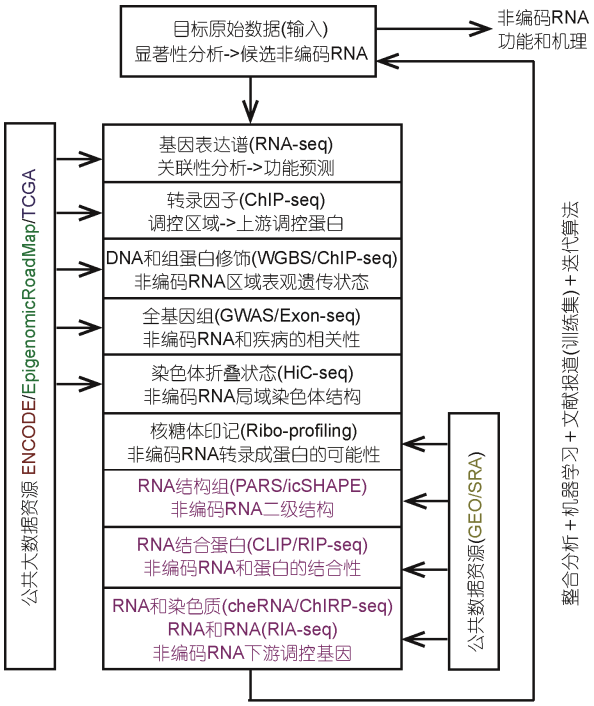

图 2　整合分析组学数据(网络版彩图)

接有效的部分. 在诱导纤维组织母细胞重编程到神经元细胞的过程中, 实验发现, 有 3 个重要的转录因子 Ascl1, Brn2 和 Myt1l 对该重编程进程起着至关重要的作用. 然而, 各个因子之间的调控机理尚不清楚. 通过对该过程中转录因子 ChIP-seq 数据以及转录组数据的整合分析, 发现在细胞重编程过程中转录因子作用的等级结构, Ascl1 作为“先驱”转录因子, 首先结合并打开染色质封闭区域, 再吸引其他转录因子结合到这些位点以调节下游基因的表达. 整合其他公共数据后, 发现一组带有 H3K27ac, H3K4me3 和 H3K9me3 修饰的三价区域, 可以用来区分转录因子是否可以诱导纤维组织母细胞向神经元细胞的重编程[39]. 利用这个发现, 成功预测了可以通过 Ascl1 转化成神经元细胞的细胞类型, 以及它们重编程到神经元细胞的效率. 这项研究成就了两个重要结果, 一是揭示了从纤维细胞向神经元细胞重编程的机理, 更为激动人心的是发现了纤维组织细胞中存在只有在神经元母细胞中才存在的染色质三价区域, 并可以以此为依据指导其他的细胞重编程实验[40].

类似的, 通过研究皮肤干细胞向角化细胞分化过程中基因表达谱, 利用 ModuleMap 算法[41,42], 成功预测了新的转录因子 MAF 和 MAFB 在细胞分化中的调节功能和机制, 通过整合公共数据资源, 重新构建了包括转录因子、DNA 甲基化酶、非编码 RNA 等在内的表皮细胞分化动态基因调控网络[43], 揭示了目前最完整的表皮细胞基因调控网络图景. 还利用相似的方法, 成功构建了胰脏细胞分化动态基因调控网络[44]. 这些工作都显示出生物信息分析方法, 尤其是整合分析多种数据的算法, 对构建基因调控网络有相当积极的作用.

2.2　基于 DNA 开放位点的表观遗传调控

DNA 以特殊的方式缠绕在核小体上面, 造成 DNA 上的空白和封闭位点. 转录因子、绝缘子、RNA 转录酶等调控基因转录的蛋白通常会结合在 DNA 的开放位点(也称活性位点)上, 因此知道 DNA 活性位点的动态分布对研究整个基因调控网络有着至关重要的意义. 传统的检测 DNA 上活性位点的方法, 是利用 DNA 酶敏感位点测序(DHS-seq)[45], 这种方法的弱点是需要的样品量大、步骤繁琐、效率低, 因此很难真正推广到临床应用. 本实验室发展了一种 ATAC-seq (assay for transposase-accessible chromatin using sequencing) 方法, 可以通过少量细胞快速检测 DNA 上的活性位点[46], 将灵敏度提高了近百万倍. ATAC-seq 甚至可以做到单细胞层次[47]. 应用 ATAC-seq 技术, 首次实现对人类活性 T 细胞个性化调控组的研究[48]. 发现了超过 4000 个活性调控位点存在着个性化差异, 通过比对男性和女性常染色体和性染色体上 ATAC-seq 信号的差别, 发现在女性 X 染色体上, 存在 116 个活性位点调控 X 染色体失活逃逸, 并首次发现和验证了 6 个基因逃离 X 染色体失活. 利用对已知转录因子在 DNA 上的特异结合位点分析, 构建正常人 T 细胞基因调控网络, 并发现存在一种针对不同性别的基因调控机制, 使得女性更容易患上自身免疫性疾病. 对于 ATAC-seq 数据的分析, 主要侧重于寻找 Tn5 酶切位点(峰值), 并研究在存在显著性差异的峰值处富集的转录因子 DNA 结合位点, 以此构建基因调控网络. ATAC-seq 技术, 以及相应的生物分析方法, 可以大规模应用到临床样品中, 真正意义上构建癌细胞、免疫细胞等复杂体系的表观遗传调控网络.

2.3 个性化调控机制和精准医疗

尽管越来越多的大规模基因组学项目陆续启动并顺利完成, 距离基于对疾病和预后反应的个性化认识的精准医疗梦想依然遥远. 究其原因, 主要在于: (i) 个性化基因组并不等同于个性化医疗, 对个体患者疾病的基因调控组的个性化认识是提供精准医疗方案的必要条件; (ii) 想要从微量临床样品中获取疾病的调控组信息, 目前还存在着实验技术上的困难; (iii) 整个基因调控是一个遗传和表观遗传修饰的复杂系统, 然而在获取对调控组的系统认识上, 还面临着挑战. 因此, 开发和利用能够从少量活体临床样本中获取调控组的实验和生物信息分析技术, 构建人类疾病个性化基因调控网络, 是对疾病的精准治疗的前提条件. 本实验室发展了一整套以 ATAC-seq 技术为核心, 快速准确检测全基因组, 包括 DNA 活性位点、DNA-蛋白相互作用和核小体分布在内的表观遗传信息新方法, 从获取患者血液样本(5~10 mL), 提取 T 细胞, 建库到测序和数据分析的完成, 只需 10 h. 正在利用它对大量的临床样本进行研究, 并结合计算分析来构建 T 细胞淋巴瘤个性化调控组. 这将是首个通过直接获取 T 细胞淋巴瘤患者活体细胞, 实时构建癌症个性化调控组的工作, 希望它能为研究其他疾病的个性化调控组建立范式, 从而推动精准医疗事业的发展.

3 基于云计算的在线生物信息分析系统

对基因调控机制的研究产生海量的各类组学数据, 由于数据量巨大, 数据类型和分析方法繁多, 以及公共数据库不能有效利用等诸多原因, 众多的实验生物学家很难对自己的数据进行深度分析, 从而影响研究工作的进程. 因此有必要创建一个操作简单、内容丰富, 且计算资源近乎无限的生物信息分析系统, 来推动整个基因组学的发展. 生物信息学家正逐渐将各种分析平台搬到云端以满足日益增长的数据分析需求. 本实验室在同 Broad 研究所的一项合作中, 创建了一个以云计算为基础的, 面向实验生物学研究人员的在线生物信息分析系统-GenomeSpace[49]. 目前该平台涵盖了全世界最常用的 20 个生物信息分析软件, 各软件之间的数据结构已经做到无缝衔接, 能够实现一次性利用多个软件进行一些复杂的整合分析. 通过 GenomeSpace, 可以整合胚胎干细胞、诱导多功能干细胞和诱导癌症干细胞基因表达谱、乳腺癌肿瘤基因表达谱和 DNA 拷贝数, 以及分子签名(MSigDB)等公共数据资源, 利用 SLAMS (stepwise linkage analysis of microarray signatures)的分析方法[50], 重新构建了以 MYC 为核心的乳腺癌基因调控网络, 并将该网络在 GenomeSpace 上实现可视化. GenomeSpace 通过数据和工具的整合可以简化癌症组学数据的分析并产生重要的结果, 使得在诸多常用软件间建立了轻便的桥梁, 让它们有机的统一, 为实验数据的计算和分析带来便利[51]. 本实验室认为, 以 GenomeSpace 为代表的建立在云端的、在线的、开发的研究平台将是未来迅速处理生物大数据的发展方向. 目前, 该平台已经拥有超过 1 万个常用注册用户, 并仍在迅速增长中.

4 展望

基因的表观遗传调控是一个多种因素共同协调作用的复杂系统. 研究人员正从各个层面深入了解表观遗传调控机制中的诸多细节, 在此过程中, 也伴随产生海量的各类组学数据. 因此, 如何整合各种基因组学数据, 构建机器学习和深度分析模型乃至形成新的生物信息学的研究方法, 对于准确预测和构建基因表观遗传网络, 全面和系统地了解表观遗传机制有着至关重要的意义. 本实验室试图通过建立一整套研究基因调控机制的生物信息学方法, 数据库和软件系统, 希望有助于更清晰、更深入、更系统地了解基础生物医学过程, 并拓展其在临床治疗等方面的应用.

随着单位资本可以获得数据量成几何级数增长, 数据的产生变得越来越容易, 相对而言, 数据的存储、分析、共享等变得越来越昂贵, 也使得生物信息技术逐渐成为推动整个研究领域迅速发展的瓶颈. 此前大量的实验和计算分析结果也表明, 生物信息学可以为基因调控机制乃至更多生物医学问题提供更为深入的分析方法, 为进一步实验提供指导方向. 相信随着计算方法和数据库的不断完善, 一种基于云计算的、开放式的、在线生物信息分析平台必将逐渐被广泛使用(图 3). 这将有助于聚集、分析、共享各种生物医学信息, 推动对糖尿病、心血管疾病、癌症等重大疾病的基础和临床研究, 对于各种重大疾病的诊断和治疗意义重大.

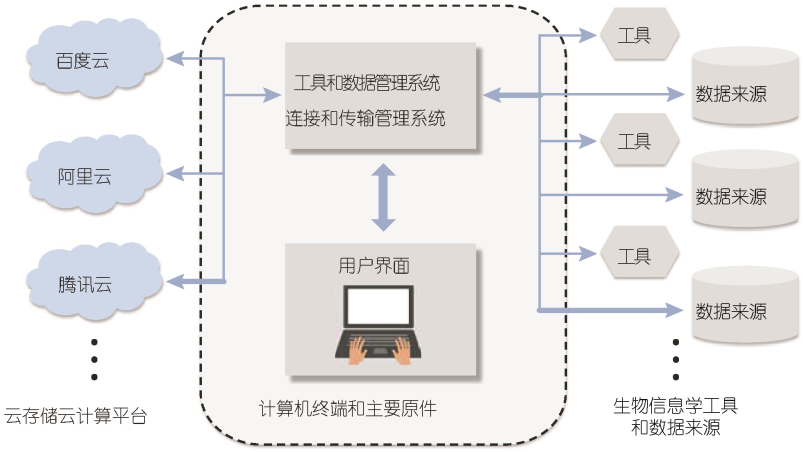

图 3　基于云计算的在线生物信息分析系统(网络版彩图)

参考文献

1 Gupta R A, Shah N, Wang K C, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010, 464: 1071–1076
2 Kretz M, Siprashvili Z, Chu C, et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature, 2013, 493: 231–235
3 Flynn R A, Chang H Y. Long noncoding RNAs in cell-fate programming and reprogramming. Cell Stem Cell, 2014, 14: 752–761
4 Satpathy A T, Chang H Y. Long noncoding RNA in hematopoiesis and immunity. Immunity, 2015, 42: 792–804
5 Batista P J, Chang H Y. Long noncoding RNAs: cellular address codes in development and disease. Cell, 2013, 152: 1298–1307
6 Rinn J L, Kertesz M, Wang J K, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 2007, 129: 1311–1323
7 Rinn J L, Chang H Y. Genome regulation by long noncoding RNAs. Annu Rev Biochem, 2012, 81: 145–166
8 王斐, 宋晓元. 长链非编码 RNA 与脑衰老的研究进展. 生命科学, 2016, 28: 602–614
9 Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, 458: 223–227
10 Guttman M, Garber M, Levin J Z, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol, 2010, 28: 503–510
11 Guttman M, Donaghey J, Carey B W, et al. LincRNAs act in the circuitry controlling pluripotency and differentiation. Nature, 2011, 477: 295–300
12 Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
13 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515
14 Grabherr M G, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods, 2011, 8: 469–477
15 Grabherr M G, Haas B J, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011, 29: 644–652
16 Kong L, Zhang Y, Ye Z Q, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucl Acids Res, 2007, 35: W345–349
17 Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol, 2010, 11: R106
18 Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26: 139–140
19 Kretz M, Webster D E, Flockhart R J, et al. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev, 2012, 26: 338–343

20 Lee C S, Ungewickell A, Bhaduri A, et al. Transcriptome sequencing in Sezary syndrome identifies Sezary cell and mycosis fungoides-associated lncRNAs and novel transcripts. Blood, 2012, 120: 3288–3297
21 Flockhart R J, Webster D E, Qu K, et al. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. Genome Res, 2012, 22: 1006–1014
22 Rapicavoli N A, Qu K, Zhang J, et al. A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. Elife, 2013, 2: e00762
23 Chu C, Spitale R C, Chang H Y. Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol, 2015, 22: 29–35
24 Wang K C, Chang H Y. Molecular mechanisms of long noncoding RNAs. Mol Cell, 2011, 43, 904–914
25 Chu C, Qu K, Zhong F L, et al. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell, 2011, 44: 667–678
26 Quinn J J, Llik I A, Qu K, et al. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. Nat Biotechnol, 2014, 32: 933–940
27 Chu C, Zhang Q, da Rocha S T, et al. Systematic discovery of Xist RNA binding proteins. Cell, 2015, 161: 404–416
28 Kertesz M, Wan Y, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature, 2010, 46: 103–107
29 Wan Y, Qu K, Zhang Q, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature, 2014, 505: 706–709
30 Wan Y, Qu K, Ouyang Z, et al. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. Nat Protoc, 2013, 8: 849–869
31 Wan Y, Qu K, Ouyang Z Q, et al. Genome-wide measurement of RNA folding energies. Mol Cell, 2012, 48: 169–181
32 Spitale R C, Flynn R A, Zhang Q F, et al. Structural imprints *in vivo* decode RNA regulatory mechanisms. Nature, 2015, 519: 486–490
33 Lu Z, Zhang Q, Lee B, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. Cell, 2016, 165: 1267–1279
34 Jiang Q, Ma R, Wang J, et al. LncRNA2 function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. BMC Genomics, 2015, 16 Suppl 3: S2
35 Du Z, Fei T, Verhaak R, et al. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol, 2013, 20: 908–913
36 Wu Y, Shi B, Ding X, et al. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucl Acids Res, 2015, 43: 7247–7259
37 Hu X, Wu Y, Lu Z J, et al. Analysis of sequencing data for probing RNA secondary structures and protein-RNA binding in studying posttranscriptional regulations. Brief Bioinform, 2015, bbv106
38 Yang Y C, Di C, Hu B, et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics, 2015, 16: 51
39 Wapinski O L, Vierbuchen T, Qu K, et al. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell, 2013, 155: 621–635
40 Hysolli E, Park I H. Trivalent chromatin marks the way in. Cell Stem Cell, 2013, 13: 510–512
41 Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer. Nat Genet, 2004, 36: 1090–1098
42 Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003, 34: 166–176
43 Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003, 34: 166–176
44 Benitez C M, Qu K, Sugiyama T, et al. An integrated cell purification and genomics strategy reveals multiple regulators of pancreas development. PLoS Genet, 2014, 10: e1004645
45 Neph S, Vierstra J, Stergachis A B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 2012, 489: 83–90
46 Buenrostro J D, Giresi P G, Zaba L C, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods, 2013, 10: 1213–1218
47 Buenrostro J D, Wu B, Litzenburger U M, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 2015, 523: 486–490
48 Qu K, Zaba L C, Giresi P G, et al. Individuality and variation of personal regulomes in primary human T cells. Cell Syst, 2015, 1: 51–61

49 Qu K, Garamszegi S, Wu F, et al. Integrative genomic analysis by interoperation of bioinformatics tools in Genome Space. Nat Methods, 2016, 13: 245–247
50 Adler A S, Lin M, Horlings H, et al. Genetic regulators of large-scale transcriptional signatures in cancer. Nat Genet, 2006, 38: 421–430
51 Marx V. Cancer: smoother journeys for molecular data. Nat Methods, 2015, 12: 299–302

Application of bioinformatics techniques in revealing the mechanisms of epigenetic regulation

JIN Chen & QU Kun

School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

Epigenetic regulation, the process of controlling gene expression without altering the actual DNA sequence, plays essential roles in determining cellular function and dictating development. The epigenome consists of signals as diverse as the position/arrangement and chemical modifications of histones, DNA methylation, noncoding RNA structure and expression, self-reinforcing transcription factor networks, and chromosome three-dimensional structure, which all function in concert to enforce the cellular phenotype. Although complex, the development of a variety of next-generation sequencing and bioinformatics techniques will greatly facilitate ourunderstanding of the mechanisms of epigenetic regulation. This article provides a brief review of the current applications of bioinformatics techniques in revealing the mechanisms of epigenetic regulation, thereby to providing novel insights into ourunderstanding of the entire gene regulation mechanism and potential therapies for the treatment of human diseases.

bioinformatics, noncoding RNA, epigenetic regulation, gene regulatorynetwork, cloud computing

doi: 10.1360/N052016-00316