基于 GBrowse 的多源长非编码 RNA 数据可视化系统①

孙 磊, 陈 璇, 唐 红, 魏李婷, 姬岚洋, 施胜飞, 杨晓华

(扬州大学 信息工程学院, 扬州 225127)

摘 要: 针对长非编码 RNA(long non-coding RNA, lncRNA)数据类型多样带来的有用信息提取困难的问题, 提出基于基因组浏览器 GBrowse(Generic Genome Browser)的多源 lncRNA 数据可视化系统. 该系统主要包括网页服务器和 lncRNA 数据存储. 其中, 网页服务器主要由 HTTP 服务和 GBrowse 网页组件构成, 支持纯文本、MySQL、SQLite 等多种数据存储方式. 系统实现流程包括 GBrowse 安装与配置、多源 lncRNA 数据的收集、数据预处理、数据存储、数据访问及可视化配置. 原型系统收集了六种人类 lncRNA 数据, 包括人类基因注释、基因组序列、组蛋白修饰 H3K4me3 信号及其位点、转录因子 CTCF 绑定位点信号及其位点的数据, 并对数据进行了预处理. 通过 MySQL、SQLite 等建立了 lncRNA 数据库, 对数据的访问方式和可视化参数进行配置. 实验结果表明, 多源 lncRNA 数据在 GBrowse 框架下能够得到整合与可视化, 并在基因组空间同时呈现, 这使得研究者能够以更加直观的方式观测数据, 进而建立新的科学假说.

关键词: 长非编码 RNA; 基因组浏览器; 数据库; 可视化

Visualization System of Multi-Source Long Non-Coding RNA Data Based on GBrowse

SUN Lei, CHEN Xuan, TANG Hong, WEI Li-Ting, JI Lan-Yang, SHI Sheng-Fei, YANG Xiao-Hua

(School of Information Engineering, Yangzhou University, Yangzhou 225127, China)

Abstract: In consideration of the problem that useful information cannot be easily extracted from various types of long noncoding RNA (lncRNA) data, this paper proposes a visualization system of multi-source lncRNA data based on generic genome browser (GBrowse). The system mainly includes a web server including HTTP service and GBrowse components, and lncRNA data storage which supports flat files, MySQL, SQLite and other types of databases. The main steps of constructing the system include GBrowse installation and configuration, multi-source lncRNA data collection, preprocessing, storage, and access and visualization configuration. A demo system is constructed by firstly collecting six sets of human lncRNA data, including human gene annotation, genome sequence, histone modification H3K4me3 signals and their loci predicted, signals of transcription factor CTCF binding sites and their loci predicted. After preprocessing, these data are stored by databases such as MySQL, SQLite and so on, and data access and visualization methods are also configured. The experiment results demonstrate that multi-source lncRNA data can be integrated and visualized within the GBrowse framework, and be showed in the genome spatial space simultaneously, which can make researchers observe the lncRNA data more intuitively, thereby helps to produce novel scientific hypothesis.

Key words: long non-coding RNA; genome browser; database; visualization

长非编码 RNA(long noncoding RNA, lncRNA)是一类具有重要生物学功能的非编码 RNA. 研究表明 lncRNA 参与胚胎干细胞凋零、细胞循环调控等细胞过程[1,2]. 近年来, 随着高通量测序技术的发展和应用(如 RNA-Seq), 成千上万的功能性 lncRNA 被发现, 同时也产生了大量用于分析 lncRNA 功能和机制的生物数

① 基金项目:国家自然科学基金(61301220);扬州大学大学生学术科技创新基金(x2015423, x2015444)

收稿时间:2016-06-23;收到修改稿时间:2016-07-25 [doi:10.15888/j.cnki.csa.005633]

据. lncRNA 数据来源广泛, 主要包括与 lncRNA 直接相关的基因注释、序列、组蛋白修饰、转录因子绑定位点等数据和信息, 以及蛋白质编码 RNA 数据、物种间序列比对、保守性分值等用于与 lncRNA 数据进行比较分析的数据. 如何有效分析这些多源 lncRNA 数据已成为 lncRNA 功能研究的重要挑战.

为了准确推断 lncRNA 的功能和机制, 可首先对多源 lncRNA 数据进行可视化, 后根据数据在基因组空间的关系设立假说并建模, 再通过统计分析对 lncRNA 的功能机制进行推断. 其中, lncRNA 数据可视化是关键步骤. 基于网页技术的基因组浏览器为包括 lncRNA 数据在内的基因数据的可视化和交互操作提供了有效方法. 当前流行的基因组浏览器是加州大学圣克鲁兹分校基因组浏览器(UCSC genome browser)[3], 但由于其服务器远在美国, 因此数据上传和下载可能会受网络连接状况和带宽限制等因素的影响. 另一方面, 类似 UCSC 基因组浏览器的公共浏览器在免费使用情况下并不能提供完善的服务(如数据共享等). 因此, 当研究者的 lncRNA 数据量特别大或需要高级访问服务时, 公共基因组数据浏览器可能无法满足研究需要. 相较而言, 可在本地建立诸如 UCSC 基因组浏览器、GBrowse[4]、JBrowse[5]等浏览器. 在本地私有网络环境下, 数据的传输速率将大大提高. 研究者还可根据需要设置相应的服务选项, 以增加数据整合与可视化的灵活性. GBrowse 是一种开放源代码的通用基因组浏览器(Generic Genome Browser), 它为用户提供了丰富的生物数据存储、交互式管理以及可视化方法. GBrowse 凭借其存储、管理、可视化数据方面的诸多优点, 已广泛应用于如植物 lncRNA 数据库 PLncDB[6]、家禽 lncRNA 数据库 ALDB[7]、深度测序信号可视化 VING[8]、转录起始位点的识别[9]等研究. 针对 lncRNA 功能研究过程中由于 lncRNA 数据量不断增加且类型众多带来的有用信息提取困难的问题, 本文提出了基于 GBrowse 的多源 lncRNA 数据可视化系统. 实验以人类 lncRNA 数据的可视化为例, 详细介绍该系统的实现流程. 在此基础上, 将研究讨论转录因子 CTCF、表观遗传信息与 lncRNA 基因之间的相互关系.

1　系统概述

基于 GBrowse 的多源 lncRNA 数据可视化系统主要由网页服务器和 lncRNA 数据存储构成(如图 1 所示). 其中, lncRNA 数据可根据需要存储于各种类型的数据库, 如 Berkeleydb、SQLite、MySQL、Oracle、PostgreSQL, 以及 GFF 格式文本. 网页服务器除了包括常用的 HTTP 服务进程之外, 最重要的是包含了 GBrowse 网页组件. GBrowse 组件中有丰富的数据访问接口, 提供对以上多种类型数据库的访问.

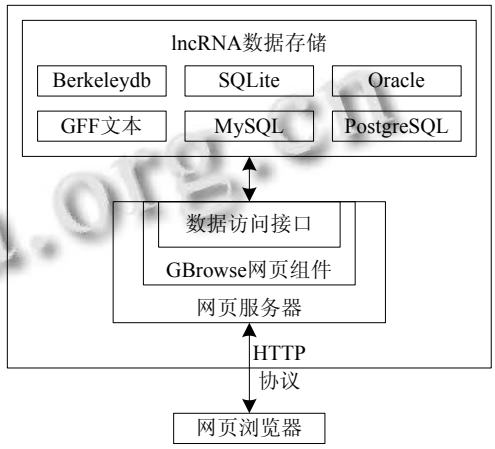

图 1　系统架构图

2　系统实现流程

基于 GBrowse 的多源 lncRNA 数据可视化系统的实现流程主要包括“GBrowse 安装与配置”、“多源 lncRNA 数据的收集”、“数据预处理”、“数据存储”和“数据访问及可视化配置”五个步骤(如图 2 所示). 本节将以人类 lncRNA 数据的可视化为例, 详细介绍系统的实现流程.

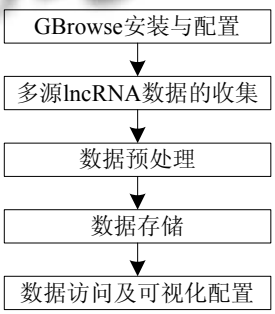

图 2　系统实现流程

2.1 GBrowse 安装与配置

GBrowse 可安装在 Linux 等类 UNIX 操作系统上. 本文采用 Ubuntu 12.04 Linux 操作系统, 通过 Ubuntu 软件中心安装了版本号为 2.42 的 GBrowse 软件. 另外, 对于 GBrowse 及网页服务器运行过程中所依赖的其他

软件(如 Apache2、Perl、MySQL、SQLite 等)的安装, 可参考文档: http://gmod.org/wiki/GBrowse_2.0_Install_HOWTO.

2.2 多源 lncRNA 数据的收集

为了帮助推断人类 lncRNA 的调控机制, 从公共数据库收集了包括人类基因注释、人类基因组序列、转录因子绑定位点和组蛋白修饰在内的多种来源的 lncRNA 数据(如表 1 所示). 其中, 人类基因注释数据(编号: D1) 下载自 GENCODE[10], D1 数据包含了人类基因的位置、结构、ID 号、数据源等信息, 数据格式为 GFF3 (Generic Feature Format Version 3). 为了获取与 lncRNA 相关的基因组序列信息, 从 UCSC 基因组浏览器下载了人类基因组序列数据(编号: D2). 已有研究表明增强子可通过 lncRNA 与基因启动子作用以影响基因转录, 而蛋白质 CTCF 与靶顺序因子的结合可阻断增强子和启动子的相互作用. 为了研究 CTCF 与 lncRNA 之间的关系, 从 ENCODE 项目网站(https://www.encodeproject.org/)下载了利用 ChIP-Seq 技术获得的转录因子 CTCF 的绑定位点信息, 该信息包含了 CTCF 绑定位点的信号(编号: D3)及预测出的最佳信号峰值区域(编号: D4). 另外, 由于三甲基化组蛋白 H3 赖氨酸(H3K4me3)与基因转录起始位点有关, 因此还下载了利用 ChIP-Seq 技术获得的人类骨骼肌细胞基因的 H3K4me3 位置信息(包含了 H3K4me3 的信号 D5 和峰值信号区域 D6). 其中, bigWig 格式数据提供了通过测序方法获得的信号强度信息, narrowPeak (BED6+4)和 broadPeak (BED6+3) 格式数据提供了预测出的最佳目标区域信息.

表 1 多源 lncRNA 数据信息

编号	数据名称	格式	大小	类型说明
D1	gencode.v19.annotation.gff3.gz	GFF3	1.18 GB	人类基因注释
D2	chr1.fa,…,chr22.fa,chrX.fa, chrY.fa, chrM.fa	FASTA	3 GB	人类基因组(版本: hg19/GRCh37)
D3	ENCFF002DC0.bed	narrowPeak (BED6+4)	964 KB	CTCF 绑定位点的最优峰值区域
D4	ENCFF001HJF.bigWig	bigWig	124 MB	CTCF 绑定位点的信号
D5	ENCFF001SXN.bed	broadPeak (BED6+3)	700 KB	H3K4me3 峰值区域
D6	ENCFF000BME.bigWig	bigWig	257 MB	H3K4me3 的信号

2.3 数据预处理

为了达到有效组织和整合 lncRNA 数据的目的, 须要对多源 lncRNA 数据进行预处理, 本实验需要预处理的数据包括 D1、D3、D5. 由于 D1 数据包含了人类编码和非编码基因的注释信息, 因此通过脚本程序提取了其中 lncRNA 基因的注释信息, 并命名为 gencode.v19.lncRNAs.gff3(编号: D1-1, 大小: 44M). 为了便于 GBrowse 处理, narrowPeak 格式的 D3 和 broadPeak 格式的 D5 都转换成了 BED6 格式, 并分别命名为 D3-1 和 D5-1.

2.4 数据存储

对于数据存储, GBrowse 支持多种数据库后端(backend), 如 Berkeleydb、SQLite、MySQL、Chado、BioSQL 等. 为了便于 GBrowse 快速显示数据, 根据已收集数据的类型和大小设计了如下的数据存储方案: 由于 D1 数据(如表 1 所示)包含了 lncRNA 的基因位置、结构、名称、数据源等信息, 内存访问比较缓慢, 因此为其建立了 MySQL 数据库(名称: “hg19”), 以提高 D1-1 的访问速率. 另外, 由于 D2 数据规模较大, 因此也将其导入“hg19”数据库. 其次, 建立了两个 SQLite 数据库, 分别存储 D3-1 和 D5-1 数据. 对于二进制格式的 bigWig 数据, 由于可通过 GBrowse 中的 Perl 模块 bigWig.pm 进行读取, 因此无需对 D4 和 D6 数据建库.

2.5 数据访问与可视化配置

数据存储完成之后, 在 GBrowse 配置文件目录下建立了用于配置数据访问和可视化方法的文件 hg19.conf, 同时在 GBrowse.conf 文件末尾添加关于 hg19.conf 的段落(section). 通过设置 hg19.conf 中的参数对数据访问接口和可视化方法进行配置(如表 2 所示), 以实现对已存储数据的显示, 并优化数据的可视化效果. 表 2 中的访问接口是指与各数据相对应的 Perl 适配模块(adaptor). 不同数据要设置成合适的形状才可以得到正确显示, 而各数据轨道(Track)应设置成容易区分和观察的形状和颜色. 参数说明和配置方法可参考文档: http://cloud.gmod.org/gbrowse2/tutorial/tutorial.html.

表 2　数据访问及可视化的主要配置参数

编号	所在数据库名称	访问接口	形状	颜色
D1-1	hg19(MySQL)	DBI::mysql	gene	蓝色(blue)
D2	hg19(MySQL)	DBI::mysql	dna	红色(Red)
D3-1	ENCFF002DCO.bed6.SQLite	DBI::SQLite	graded_segments	橄榄绿(olive)
D4	ENCFF001HJF.bigWig	DB::bigWig	wiggle_whiskers	默认(Default)
D5-1	ENCFF001SXN.bed6.SQLite	DBI::SQLite	graded_segments	黑色(black)
D6	ENCFF000BME.bigWig	DB::bigWig	wiggle_whiskers	默认(Default)

3　结果与分析

通过以上实现流程, 建立了一个人类 lncRNA 数据可视化的原型系统 (名称: HlncRNAdb-demo, 访问: http://bioinf.yzu.edu.cn:40/cgi-bin/gb2/gbrowse/hg19/), 该系统为研究者提供了直观的人类 lncRNA 数据可视化(如图 3 和图 4 所示).

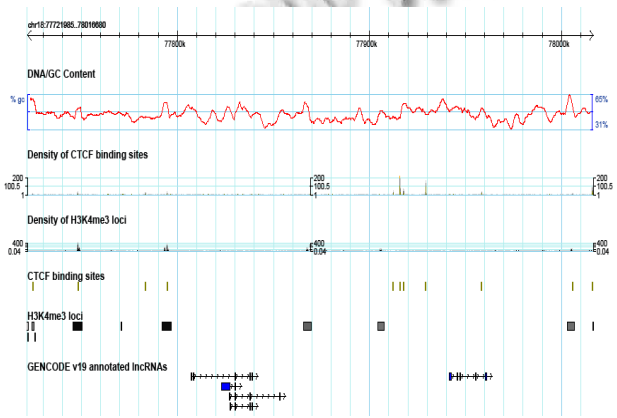

图 3　chr18:77721985-78016680 范围内的 lncRNA 数据显示

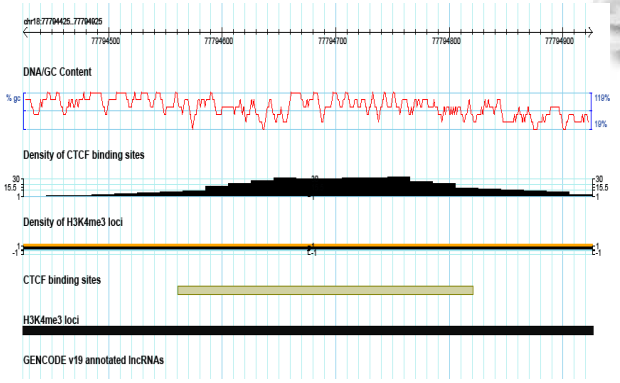

图 4　chr18:777994425-7794925 范围内的 lncRNA 数据显示

3.1 lncRNA 数据的可视化

HlncRNAdb-demo 通过 GBrowse 成功整合了包括人类 lncRNA 基因注释(GENCODE v19 annotated lncRNAs)、人类基因组序列/GC 含量(DNA/GC Content)、组蛋白修饰 H3K4me3 信号(Density of H3K4me3 loci)及其预测位点(H3K4me3 loci)、转录因 CTCF 绑定位点信号(Density of CTCF binding sites)及其预测位点(CTCF binding sites)在内的多源 lncRNA 数据, 数据格式包含 GFF3、FASTA、BED6、bigWig 等.

多源 lncRNA 数据在基因组空间得到整合, 并以各自的形状和颜色加以显示, 从而区别于其他轨迹. 图 3 展示了在比例缩小(zoom out)情况下对基因组 chr18:77721985-78016680(295 kilo base pairs/295kbp) 范围内的整合数据进行可视化的概况, 而图 4 是将比例放大(zoom in)后对 chr18:777994425-7794925(500bp) 范围内的整合数据进行显示的概况. 图中红色波浪状曲线显示的是“DNA/GC Content”轨道, 红色曲线实际上是将区域内计算得到的序列 GC 含量通过图形化显示后的结果. 如果放大倍数足够, 便能够看到基因组序列的碱基构成. H3K4me3 信号的分布及预测的峰值区域分别如图中的“Density of H3K4me3 loci”和“H3K4me3 loci”轨道所示. 类似地, CTCF 绑定位点信号的分布及预测区域分别如图中的“Density of CTCF binding sites”和“CTCF binding sites”所示. 其中, “H3K4me3 loci”和“CTCF binding sites”均采用 graded_segments 形状进行显示, 其中的颜色灰度会根据原 BED 数据文件中的分值进行显示. 图 3 中最下方的轨道“GENCODE v19 annotated lncRNAs”显示的是 GENCODE 发布的 v19 版的 lncRNA 的结构和位置信息. 综上, 研究者能够在同一空间范围内对多源 lncRNA 数据进行观测和比较. 通过鼠标拖放可选取观察范围, 或放大或缩小. 对于每个轨道中的标记对象, 可通过鼠标点击获取结构化的详细数据/信息表, 此表可帮助研究者查看目标图形的数据详情. 另外,

在 GBrowse 界面中, 选定范围内各种数据的特征和相对关系一目了然, 起到了数据显微镜的作用.

3.2 可视化数据的分析

借助基于 GBrowse 的多源 lncRNA 数据可视化系统, 研究者可在基因组空间中清晰地观测多源 lncRNA 数据, 这可以帮助验证已有的假说、推论或建立新的科学假说或模型. 如图 3 所示, CTCF 和 H3K4me3 信号出现在 lncRNA 基因上游启动子附近, 说明CTCF和H3K4me3可能与该基因的表达调控有关联, 此数据显示反映出的特征与当前流行的研究观点保持了一致. 又如图 3, 可以观测到许多 CTCF 信号的出现位点都会伴随有H3K4me3信号的出现, 而其中的本质原因值得进一步探讨, 比如可以建立如下假说: CTCF 能够识别 H3K4me3 位点, 然后绑定到 H3K4me3 区域, 进而对基因产生调控作用. 当然, 假说的验证需要依据后期更多的实验和分析. 由此可见, 基于 GBrowse 的多源 lncRNA 数据可视化系统能够帮助研究者获得更多的关于 lncRNA 的研究信息和思路.

4 结语

本文提出了基于 GBrowse 的多源 lncRNA 数据可视化系统, 并介绍了系统的实现流程. 实验建立了人类 lncRNA 数据的可视化系统原型 HlncRNAdb-demo. 实验结果表明该系统能够实现在同一基因组空间上对多源 lncRNA 数据进行整合与可视化, 便于研究者从中获取信息, 进而助其进行理论验证或建立新的科学假说. 对于本文的后续工作, 有如下计划和建议: ①可根据研究需要收集和整合其它 lncRNA 数据, 以增加 lncRNA 功能研究的信息; ②采用新的方法以提高数据访问的速率, 比如可采用 FastCGI; ③在网页中添加对 lncRNA 二级结构进行可视化的功能, 能在二维或三维空间整合和观测 lncRNA 数据. 综上, 多源 lncRNA 数据在 GBrowse 框架下得到有效整合与可视化, 能够推动 lncRNA 功能研究的发展.

参考文献

1 Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, Manos PD, Datta S, Lander ES, Schlaeger TM, Daley GQ, Rinn JL. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nature Genetics, 2010, 42(12): 1113–1117.

2 Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P, Wang Y, Kong B, Langerod A, Borresen-Dale AL, Kim SK, van de Vijver M, Sukumar S, Whitfield ML, Kellis M, Xiong Y, Wong DJ, Chang HY. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet, 2011, 43(7): 621–629.

3 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Research, 2002, 12(6): 996–1006.

4 Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: A building block for a model organism system database. Genome Research, 2002, 12(10): 1599–1610.

5 Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: A next-generation genome browser. Genome Research, 2009, 19(9): 1630–1638.

6 Jin J, Liu J, Wang H, Wong L, Chua NH. PLncDB: Plant long non-coding RNA database. Bioinformatics, 2013, 29(8): 1068–1071.

7 Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: A domestic-animal long noncoding RNA database. PLoS ONE, 2015, 10(4): e0124003.

8 Descrimes M, Zouari YB, Wery M, Legendre R, Gautheret D, Morillon A. VING: A software for visualization of deep sequencing signals. BMC Research Notes, 2015, 8: 419.

9 Cumbie JS, Ivanchenko MG, Megraw M. NanoCAGE-XL and CapFilter: An approach to genome wide identification of high confidence transcription start sites. BMC Genomics, 2015, 16(1): 597.

10 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S. GENCODE: The reference human genome annotation for the ENCODE project. Genome Research, 2012, 22(9): 1760–1774.