PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2014, 41(11): 1182~1190
www.pibb.ac.cn

基于 Web 的基因组浏览器研究现状 *

张海川 ** 李 杰 ** 王亚东 ***
(哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001)

摘要 测序技术的发展促使人类基因组测序成本急剧降低，测序速度迅速增加，对这些数据的分析和可视化已成为生命科学领域最重要的课题之一．基因组浏览器技术在基因序列分析，遗传密码解读，复杂疾病研究等方面具有重要意义．本文综述了 9 种主要的基因组浏览器技术，并从可视化内容、可视化形式、软件系统架构等角度分析了它们的特点．最后，探讨了基因组浏览器发展所面临的挑战．

关键词 基因组浏览器，可视化，基因组测序
学科分类号 Q3，Q75，Q78
DOI: 10.3724/SP.J.1206.2014.00055

随着生命科学技术的快速发展，尤其是新一代大规模测序技术的不断进步，使得我们对基因的功能有了更加深入的认识．通过深度测序技术，研究人员对复杂疾病的发病机制，相关致病的发病原理和遗传变化有了崭新的认识．随着海量基因序列的爆炸式增长，发展基因组序列的有效可视化方法，支持生物大数据的深度分析、集成、研究和服务，已经成为基因组学研究领域面临的一项重要课题．目前国内外研究机构和公司开发了多个基于 web 技术的基因组浏览器，以满足基因组可视化、大规模基因组数据分析和应用的需要．根据基因组浏览器的使用形式，可以分为基于桌面的浏览器和 web 浏览器两种．基于桌面的浏览器虽然方便加载本地数据，但是由于基因组数据十分庞大，所以对 PC 的要求比较高，而且需要在本地安装客户端程序．基于 web 的基因组浏览器不用安装，无需进行繁琐的软件配置，用户只需要通过网页浏览器连接到 Internet 就可以使用，对用户个人 PC 的要求不高，而且由于数据存储在服务器端，所以用户本地电脑不用消耗太多存储空间来贮放数据．相比较而言，基于 web 的基因组浏览器更有发展前景．本文重点介绍目前主要的基于 web 的基因组浏览器，并对这些浏览器可展示的内容、可视化方式和软件架构进行详细的分析和比较．

1 基因组浏览器概览

目前功能比较强大，应用比较广泛的基因组浏览器有 UCSC Genome Browser、Ensembl、JBrowse 等，它们在可视化数据类型、可视化方式等方面各有特色，功能都比较强大．下面从可视化内容、可视化方式、实现架构和优缺点 4 个方面对每一种浏览器进行详细介绍．

1.1 UCSC Genome Browser

随着越来越多的物种基因组序列的测定，遗传密码得到解密，对这些序列数据的可视化展示就变得越来越重要．UCSC Genome Browser[1]是由加州大学圣克鲁兹分校(UCSC)开发创立的、功能强大的基因组浏览器，主要用于浏览基因组、查看基因组注释信息．它本身并不下任何结论，只是给用户提供参考信息．UCSC Genome Browser 目前在全世界应用非常广泛，其他很多浏览器网站，如

* 国家高技术研究发展计划(863)(2012AA02A604)，国家自然科学基金(61471147, 61173085)和黑龙江省博士后基金(LBH-Q13084)资助项目.
** 共同第一作者.
*** 通讯联系人.
Tel: 0451-86413316, E-mail: ydwang@hit.edu.cn
收稿日期：2014-03-01，接受日期：2014-06-04

Ensembl 都用到他的基因组序列数据.

1.1.1　可视化内容

UCSC Genome Browser 目前提供人类、黑猩猩等 88 个物种的基因组可视化信息，UCSC 中约有一半的注释信息是 UCSC 通过公共的序列数据计算得到，其余部分则来自世界各地的研究结果. UCSC Genome Browser 可视化内容包括组装序列间隙 / 重叠，mRNA 和表达序列标签队列，多基因预测，跨物种同源性，SNP，序列标记位点，辐射杂交数据，转座子重复等. 进入系统主页后，可以根据基因名、关键词等来检索一个基因，也可以直接根据染色体或者核苷酸碱基范围进行查询. 通过缩放，用户既能从宏观上查看整个基因组各区域的基因密度，也能从微观上查看一个序列区域内的基因信息. 研究者可以从科研或者教育的角度加入自己的注释信息.

1.1.2　可视化方式

UCSC Genome Browser 以 track(轨道)的方式来展示相关信息. Track 表现为横向或纵向的条带，条带上不同区域分别用不同的颜色、线、方块等表示不同的生物含义，因形状类似于赛场的跑道，故翻译为轨道，如图 1 中红色矩形条状框内即为一条轨道. 系统主界面从上到下可分为三块：检索查询、可视化(图 1)和轨道管理. a. 检索查询，包括直接通过染色体区域范围查询、根据基因名称进行查询、对页面中现有范围的左右平移和倍数缩放等；b. 可视化展示，每条样本数据都用一条轨道来表示，并且目前提供 5 种展示示模式(hide、dense、squish、pack、full)；c. 轨道数据管理，主要包括 track 的分组管理、展示模式的管理以及用户自己上传数据的可视化等. 研究者可以通过控制图片下方的下拉菜单来选择展示哪些 track. 每个 track 名字都链接到一个特定的网页，从中可以查看该 track 数据的计算方式，还可能有相应的文献或者序列信息等. 页面中不同的元素代表不同的含义，比如在基因结构视图中，盒形代表外显子，线形代表内含子. 研究者可以点击 Genome Browser 上相应按钮获得更多注释信息，像正向选择(positive selection)基因等.

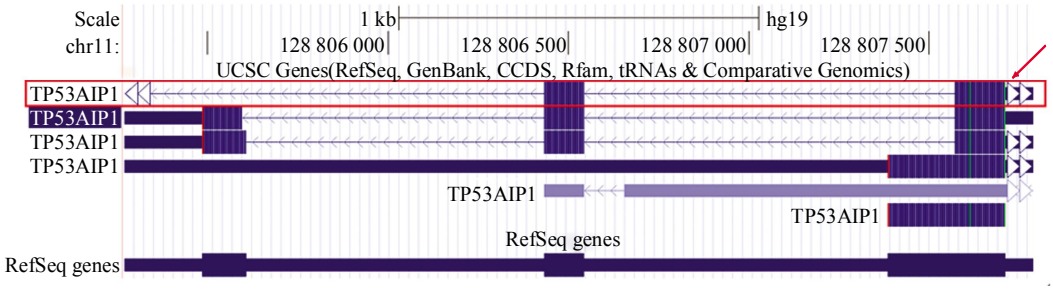

Fig. 1　The visualization form of UCSC Genome Browser: track

图 1　UCSC Genome Browser 的可视化形式——track

较早开发的经典基因组浏览器基本都采用 track 的可视化方式，新一代浏览器也有很多采用这种形式. Track 有多种表现形式，图中从上到下有 6 条横向 track.

1.1.3　实现架构

UCSC Genome Browser 的开发，起源于一小段应用于 *C. elegans* 基因预测拼接图谱的 C 语言脚本[2]，后期通过不断扩充，才变成现在这样强大的一个分析工具. 现在 UCSC 的主要开发语言是 Java/Python，后台数据库依赖于 mysql，而且提供 mysql 的公共接口，只要用户本地电脑装有 mysql 客户端，就可以通过 UCSC 提供的接口访问网站后台的数据库；对于前台要求，UCSC 可以较好地兼容 IE、Chrome、Firefox 等主流网络浏览器. UCSC 是完全开源的，用户可以下载到完整源码.

1.1.4　优缺点

从 web 技术的发展历程，可以将其分为几个阶段：web1.0 时代用户被动接收网站单向发布的信息，以个人网站为代表；web2.0 时代的特征则是开放、去中心化，用户参与信息发布和分享，以博客为代表；新一代 web 技术是对 HTML5 及其周边相关应用开发的统称，它可以支持更多的人机互动，并能实现很大程度的智能化. 局部刷新技术是指，当网页中的某一个元素或标签需要改变时，不

重新加载整个页面，而只对需要改变的部分进行刷新的技术．局部刷新技术最负盛名的应用当属谷歌地图．UCSC Genome Browser 是一个非常综合的基因组浏览器，包含物种数比较多，可视化内容也比较齐全．然而遗憾的是，UCSC Genome Browser 的开发年代尚处在 web1.0 时代，没有充分应用新一代 web 技术和局部刷新技术，所以 UCSC 中部分程序驱动的脚本对用户访问频率有一定限制：以 BLAT 脚本为例，它最多支持每 15 s 一次用户点击加载，一天最多只能支持 5000 次后台加载．

1.2 NCBI Map Viewer

NCBI[3]由美国国家生物技术信息中心开发，建立的初衷是提供一个信息存储处理系统，现在除了建有 GenBank 等核酸序列数据库外，还提供强大的检索和分析功能．NCBI 提供的资源有 Entrez、BLAST 等几十种．Map Viewer 是 NCBI 中一个非常有用的可视化分析工具，通过 Map Viewer 可以了解感兴趣的基因在基因组中的位置、基因序列、内含子 / 外显子的排列等很多有用的信息．Map Viewer 提供了基因组集合、遗传图谱、物理图谱及相关注释信息和对比信息等，可视化工作比较出色．

1.2.1 可视化内容

NCBI Map Viewer 的数据库涵盖了脊椎动物、无脊椎动物、原生动物、植物、菌类等，可视化内容主要包括：染色体图谱、contigs、基因、SNPs、G3/GB4 辐射杂交图、mRNA alignments 等．

1.2.2 可视化方式

与 UCSC 类似，NCBI Map Viewer 也采用 track 的方式进行可视化，不过它采用的是纵向 track．研究者可以通过表格形式查看注释数据，可以下载数据或者将需要获取的基因序列存储到本地硬盘．NCBI Map Viewer 提供多种可视化方法，如 Genetic Linkage Maps 主要是为了研究遗传连锁的重组频率，其原理是 2 个基因位点之间进行重组的频率与它们之间的距离相关联；在 Cytogenetic maps 中，基因 / 拷贝数或者其他生物信息通过荧光或者放射性元素标记在染色体上，探针位置标记在特定的染色体带(band)上，如 TP53 基因的位置是 17p13.1(染色体 17、短臂、band 位置 13.1)；而 Sequence-Based Maps 主要精确地展示由人类基因组计划采集的人类基因组核苷酸层面的序列，视图中各元素的位置通过 BLAST 比对到基因组序列上．和 UCSC 类似，Map Viewer 在全基因组范围内能通过基因名称、关键词或基因位点进行查询，可以单击图形进行放大观看．

1.2.3 实现架构

Map Viewer 的大部分核心是用 C 语言编写的，跨平台性比较好．后台采取多种数据库的组合，有 sql、custom 和 Oracle. Map Viewer 前台可很好地兼容各浏览器，尤其在 IE 下性能最佳．

1.2.4 优缺点

NCBI Map Viewer 提供多种视图来分析不同的功能．但是 NCBI Map Viewer 的可视化界面不是十分美观，在用户操作上也有一些不足，如不支持鼠标拖拽选择可视化区域，也不支持拖动平移等．而且由于 Map Viewer 大都是由 C 语言编写的，虽然方便简洁，效率较高，但 C 语言是面向过程的语言，因而具有面向过程技术的通病：由于方法和数据的分离，封装性不好，耦合和内聚性能也不好；而且面向过程技术编写的各模块之间无法做到完全独立，它们之间往往有关联，模块化不强．NCBI 研究团队对于 Map Viewer 已经不再进行深入开发，并预计在未来几年用其他工具进行替换．

1.3 Ensembl Genome Browser

Ensembl Genome Browser[4]由 Sanger 研究所 Wellcome 基金会、欧洲生物信息研究所共同运营，旨在实现对真核生物基因组的自动注释．全世界很多公司和研究机构都使用 Ensembl 软件系统提供的服务．

1.3.1 可视化内容

Ensembl Genome Browser 目前提供人类、小鼠等 66 个物种的可视化信息，数据主要来源于 InterPro、OMIM、SAGE、HUGO 等．可视化内容包括：基因注释，变异信息，蛋白质结构域和家族，基因转录子，内含子 / 外显子，基因在基因组上的前后序列，不同物种之间基因组的比较等．

1.3.2 可视化方式

Ensembl Genome Browser 采用的可视化方案与 UCSC 有相同之处，也是以横向轨道的方式进行可视化，主要用于对真核生物的基因组进行自动注释．通过检索物种和对应的基因，实现三级分辨率展示(染色体预览、基因区域预览和基因区域细节展示)，分别对应 Multi Mbs、1 Mb 和 100 b 左右．最高精度级别的基因区域细节展示情况详见附件图 S1．点击 Ensembl Genome Browser 的基因名

会链接到特定的网页，从中可以查看该基因区域的详细信息．在 MultiContigView 中，研究者可以同时查看两个基因组的某染色体区域．

1.3.3 实现架构

Ensembl Genome Browser 是基于 BioPerl 框架用 perl 语言进行编写的，部分扩展和接口也用到了 C 和 Java．后台数据库采用开源的 mysql，用户也可以采用 FTP 协议下载数据；客户端可兼容 Firefox3.5+、IE8+、Safari、Chrome 等主流浏览器．整个 Ensembl Genome Browser 都是开源的，其他可视化工具也可以使用 Ensembl 对外提供的 perl 应用接口．总体来说，Ensembl Genome Browser 在开源上做了很多工作，同时数据也很齐全，功能是非常强大的．

1.3.4 优缺点

Ensembl Genome Browser 在对真核生物的注释上功能非常强大，可以做到准确、全自动．然而，由于 Ensembl 的前后台通信采用的是 CGI(通用网关接口)协议，浏览数据时应用的是基于页面的模型，用户的每一个交互动作都需要重新加载整个页面，在局部刷新技术的应用上还存在一些不足．不支持鼠标拖动平移，在鼠标拖拽选择可视化区域上，也是先弹出一个提示框，需点击其中的两条链接才会跳转到选定区域，加载时间也比较长．

1.4 JBrowse

JBrowse[5-6]是一个开源的、应用新一代 web 技术的基因组浏览器．JBrowse 可以看作 GBrowse[7] 的后继，它的主要目的是将谷歌地图等工具中先进的局部刷新技术应用到基因组浏览器中，以达到更加流畅的可视化效果．

1.4.1 可视化内容

JBrowse 可以展示基因组整体视图，也可以细化展示基因跨度、tRNA、转座子、寡核苷酸、蛋白质结合位点、增强子、基因调控区域、非编码 RNA、点突变、序列变异信息等．用户可以自己上传需要可视化的内容，支持 GFF、GFF3、WIG、BED、FASTA、Wiggle、BigWig、BAM 等多种格式的文件．

1.4.2 可视化方式

JBrowse 也用 track 的方式进行可视化，提供平滑的动态移动和缩放功能，也有导航和通道的选择．JBrowse 可以展示多种 track 视图，除基本视图外，还可以显示非翻译区、外显子、内含子结构等．

1.4.3 实现架构

JBrowse 是开源的，主要开发语言为 Javascript 和 HTML5，对外提供 javascript API，数据通过 perl 进行格式规整．前端对浏览器的要求是需要支持 HTML5 中的新标签(如 canvas)，最新发布的版本对浏览器的要求是 Firefox10+、Safari5+、Chrome17+和 IE9+等；后台服务器端主要采用 json 文件进行存储．由于 json 本身就是基于 javascript 的一个子集，可以看作 javascript 中的对象和数组，因此可以实现较快的前后台通信．

1.4.4 优缺点

与上面介绍的三大基因组浏览器相比，JBrowse 属于新一代基因组浏览器，是基于最新的网络技术开发的．在 JBrowse 中，服务器端的负荷得到了极大的释放，后台服务器只需要向浏览器客户端发送静态文件，从繁杂的计算工作脱离出来，大量计算工作被分配到了前端．同时，HTTP cookies 技术也得到了很好的应用，可以有效记录用户的喜好．然而，尽管 JBrowse 把可视化工作放在了浏览器端，但它的可视化方法仅是一些普通的 HTML 标记，前端在绘制图像时需要运行大量的脚本代码，而且要考虑浏览器对 HTML5 中新标签的兼容性问题．

1.5 ABrowse

ABrowse[8]是由北京大学应用新一代 web 技术开发的、完全开源的基因组浏览器．ABrowse 也很好地应用了局部刷新技术，可以通过类似谷歌地图的交互界面来访问全基因组的可视化数据．目前 ABrowse 已经被应用到多个国家机构和国内国际项目中，同时 ABrowse 研究团队还应用其项目成果开发了新一代稻米基因组浏览器[9]等．

1.5.1 可视化内容

ABrowse 可以展示基因组整体视图等．在 track 描述栏中可以查看详细信息，如块长度、序列信息、各种属性信息等．除了提供丰富的接口外，ABrowse 还内置了一个非常强大的检索查询系统，用户可以通过文本或序列的形式查询网站预处理过的及用户操作生成的注释信息．

1.5.2 可视化方式

ABrowse 采用 track 的方式进行可视化，提供连续的移动、缩放展示功能，而且在侧边栏展示详细的 track 信息，可以精确到 base 级别．

1.5.3 实现架构

ABrowse 是完全开源的，主要开发语言为 jsp、java 和 js，对外提供多种网络接口 API，研究者甚至可以使用开源代码中的数据库语句在特定模式的注释数据上配置 ABrowse. 从构架上可以将 ABrowse 分为三部分：用户交互层、需求 / 数据处理层和注释数据库层.

1.5.4 优缺点

与 JBrowse 类似，ABrowse 也属于新一代基因组浏览器. ABrowse 最大的特色是开源性做得非常出色，支持用户通过自己的文件定制 Genome Browser，所有的 ABrowse 注册用户都可以在网站上保存及分享他们的注释信息，通过设置注释信息的隐私级别，用户可以选择将自己的注释信息与大家分享或者仅自己可见. 并且支持在同一个站点内整合多个物种的数据.

1.6 UCSC Cancer Genomics Browser

UCSC Cancer Browser[10]由加州大学圣克鲁兹分校 (UCSC) 开发，用来完善 UCSC Genome Browser，但是可视化内容和方法却与之迥然不同. UCSC Cancer Browser 展示实验样本在全基因组上的基因热图，临床特征也被数字编码并分别展示，以帮助预测病理学特征、对症治疗.

1.6.1 可视化内容

UCSC Cancer Browser 是面向人类癌症基因分析的，现在可以分析肺癌、乳腺癌等 44 种癌症的样本基因组信息. 网站上的数据主要来自 TCGA、Childhood Cancer、SU2C 和全球发表的文献中的数据集. 可视化的生物内容包括拷贝数(变异)、DNA 甲基化、外显子表达、基因表达、mRNA、RNAseq、蛋白质结构和临床信息数据.

1.6.2 可视化方式

UCSC Cancer Browser 采取的可视化方法比较丰富，主要是热图、箱线图和比例图，详见附件图 S2.

热图[11]是以特殊高亮或者不同颜色显示特定区域的技术，通过对页面或图形中不同区域加以不同的标注使枯燥的数据变得更加清晰明了，标注手段一般采用颜色的深浅、点的疏密等. 应用在基因组可视化领域时，热图中不同的颜色代表相应的不同生物含义值. UCSC Cancer Browser 的数据热图中，横向为从 1 号染色体到 Y 染色体的所有核苷酸区段，纵向为不同的病例样本，最右侧为临床数据的热图. 热图中采用不同的颜色来表示不同的数据，以 gene expression 数据可视化后生成的热图为例，红色代表数据值大于 0，绿色代表小于 0，灰色则表示无数据. 用户把鼠标在热图某个位置悬停时，会显示该位置的样本编号、染色体号和核苷酸位置区间，以及数据值.

箱线图[12]，也称盒式图，是统计图中一种常见的形式. 它可以体现样本总体分布情况和数据分散情况. 与热图中每一行对应一个样本基因组不同，在箱线图中，UCSC Cancer Browser 所有的样本都根据右侧的临床数据在列方向上进行重排序，意在展示数据的整体分析情况.

比例图包括最常见的圆饼图，还有一种形式是，先把所有的样本值排序，值的大小分别赋以不同的颜色，达到一种富集的效果，进而查看总体的分布情况，UCSC Cancer Browser 采用的就是后一种方法.

UCSC 还提供生存曲线等其他可视化形式. 临床特征与基因数据库相关联，显示在同一条 track 中，用户可以对右侧的临床特征进行优先级排序并查看排序后的热图. 研究者可以分析一个临床特征与遗传变化之间的关系，也可以对不同临床特征的两种病例组进行对比，统计不同组之间的遗传差异等.

1.6.3 实现架构

UCSC Cancer Browser 的前端可视化代码是用 javascript 脚本编写的，前端的热图则是后台服务器的一段 C 程序画出的. 临床数据热图用 D3(一个非常流行的用来可视化的 javascript 库)进行渲染. 网站总体框架是用 Django 搭建的，后台数据库用的是 mysql.

1.6.4 优缺点

UCSC Cancer Browser 的优点是其癌症数据库比较齐全，应用了较多的可视化形式. 缺点是可视化界面还存在一些不足：通过鼠标操作只能放大查看，不能缩小查看，也不能平移拖动；对某一项疾病的数据，只能切换显示热图、箱线图、比例图中的一项，无法在一个页面中同时显示；用户虽然可以向热图中添加自己的临床数据，但只能是 ID 号和简单的数字值，类目型值(如“positive”，“group 1”)则无法识别.

1.7 IntOGen

随着癌症致病基因数据的不断产生，对数据的整合处理、找到与肿瘤发生相关的基因改变就成为一个重要的课题，IntOGen[13]的主要目的是可视化

那些对肿瘤发生起重要作用的基因.

1.7.1 可视化内容

IntOGen 对 TCGA 中 12 种肿瘤疾病的 3 205 个患者样本进行外显子组测序，主要显示的生物内容为基因扩增、纯合子缺失、突变频率、evidence 等.

1.7.2 可视化方式

IntOGen 的主要可视化手段是表格和矩阵热图，用户选择一种肿瘤类型后，IntOGen 会显示所有与该肿瘤有关的基因样本突变频率，而且频率值按照大小用不同深浅的颜色来进行标记，方便查看该基因与该疾病的关联度大小．用户可以在 IntOGen 的 Search 对话框中检索特定的基因，获得该基因在 IntOGen 中的 ID、染色体号、染色体上的核苷酸位置、长短臂位置、refseq、同源基因、简单描述等．IntOGen 通过研究某一种癌症类型中，差异比较明显(通常采用 *P* 值小于某个阈值的方法)的某个基因或生物模块，来研究该基因或生物模块和特定癌症的关系．研究者可以方便地观察每个个体的实验结果，也可以组合查看有相同临床注释的样本和 pathway.

1.7.3 实现架构

IntOGen 开发主要用到的是 Java 和 Wicket(一种 java web 框架)，后台数据库是 mysql．到目前为止，IntOGen 后台大约有 7 000 万行数据，所以只要有一个性能优良的服务器，mysql 完全可以满足要求．为了应对更大的需求，该研发团队正在开发 Onexus 框架来优化网站设计，并且计划将来在后台应用 NoSQL 技术(Elastic Search 或 Cassandra).

1.7.4 优缺点

与其他浏览器相比，IntOGen 的独特之处在于，其所有的样本都是根据肿瘤疾病国际分类组织的词库人工进行标定的，从而将特定的变异与临床注释关联起来，并把类型和子类型按分层的方式展现出来．由于个人基因的分析并不能很好地解释完整的生物复杂性，因此 IntOGen 是分析生物模块(如 pathway)对癌症的影响．缺点是表格罗列数据信息的方式不如图表直观，而且 matrix 视图加载时间过长.

1.8 cBioPortal

Cerami 等[14]开发的 cBioPortal 是为了进行多维癌症基因组数据集的可视化展示，降低查看复杂基因数据的门槛，让研究者快速、直观地获取大规模癌症基因数据在分子层面的详细信息.

1.8.1 可视化内容

cBioPortal 也是一个癌症基因组网站，截止到目前网站上有 56 种癌症，超过 15 000 个肿瘤样本的数据．可视化内容为序列信息、mRNA、突变频率、拷贝数变异、miRNA、甲基化、RPPA 等．cBioPortal 支持 Pfam 等数据库中蛋白质域的可视化，也可以进行蛋白质和磷蛋白数据的集成分析.

1.8.2 可视化方式

cBioPortal 的可视化方式除表格形式外，主要有三种视图：热图、统计图和网络．热图也是采用不同颜色来区分不同的数据值，但形式比 UCSC Cancer Browser 简单，每次只对一个用户选中的样例进行热图展示，而不是一次性绘制所有样本的热图，通过“oncoprint”可以查看肿瘤样本基因组突变频率和特定的突变区域数据．统计图的形式有直方图、饼状图、散点图．热图和统计图方法中，无论 track 横向或纵向放置，采用的都是线性布局，局限性是显然的：在表示多个对象之间的联系时显得很乏力．而网络可以很好地解决这个问题，网络中对象体现为结点，而它们之间的联系则用边来表示，cBioPortal 即采用了网络可视化方式，详见附件图 S3．另外，cBioPortal 还有一项非常有趣的功能：可以像谷歌地图一样缩放查看详细的组织细胞图.

1.8.3 实现架构

cBioPortal 对外提供数据下载，提供完整的 web 接口和 matlab/R 统计分析包等分析工具．前台采用 java、jsp 和 javascript 进行开发，后台则采用 mysql．用户如果想可视化自己的数据，可以搭载 tomcat 和 mysql 后将软件下载到本地，并加载自己的数据，也可以直接联系网站开发人员.

1.8.4 优缺点

cBioPortal 最大的优点就是简单易用，通过选取癌症类型、生物遗传特性、患者集合和感兴趣的基因，就可以体验到 cBioPortal 的所有功能特性．缺点是不够直观，在数据和视图之间的跳转比较频繁，缺少一个包含所有样本的、友好的可视化界面.

1.9 CRC Aggressiveness Explorer

TCGA 正在进行的一项计划是描述 20 多种癌症中的基因组变化情况，迄今已经发表了两种癌症类型的研究结果[15-16]，人类结肠癌是其中一种．CRC Aggressiveness[17]是可视化结肠癌基因组改变情况的浏览器.

1.9.1　可视化内容

CRC Aggressiveness 是用来分析与直肠癌相关的特定分子标签的. 可视化内容包括外显子序列、编码基因转录水平的改变、基因突变、DNA 拷贝数、启动子甲基化、信使 RNA 和 microRNA 表达、易位、CRC 中改变的 pathway 等.

1.9.2　可视化方式

CRC Aggressiveness 采用热图的方式进行数据展示，但与前面 UCSC Cancer Browser 等稍有不同：CRC Aggressiveness 采用环形热图方式展示整个基因组，详见附件图 S4. 与以往线性布局的视图相比，环形视图优势明显，可以更方便地表示各个特征之间的联系，而且更加美观[18].

1.9.3　实现架构

CRC Aggressiveness 的开发语言主要为 Javascript 和 CSS，后台数据库采用 mysql 进行存储. 数据加载的脚本语言用到了 python. 整个系统都是开源的. 网站对 IE 的兼容性欠佳，对 Chrome 和 Firefox 则可以很好地兼容.

1.9.4　优缺点

CRC Aggressiveness 的可视化方法非常新颖，圆环热图的形式表达信息充分，同时也十分美观. 但目前只分析与直肠癌相关的基因组特征，方向比较单一.

除了上面介绍的 9 种基于 web 的基因组浏览器外，很多需要配置在本地计算机的浏览器功能也是非常强大的, 像 IGV[19-20]、CisGenome Browser[21]、Bluejay[22]、Gitools[23]等. 这些可视化工具的使用需要用户下载应用程序到本地，并安装到自己的 PC 机上，因为它们不是基于 web 的可视化工具，所以在本文中不做详细介绍.

2　各基因组浏览器比较

上面介绍了 9 种基因组可视化工具，它们包含的可视化方法有：track、热图、箱线图、比例图、网络、环形热图等(图 2). 这些工具的可视化方法、可视化内容、前后台架构、开源性及优缺点等比较详见附件表 S1.

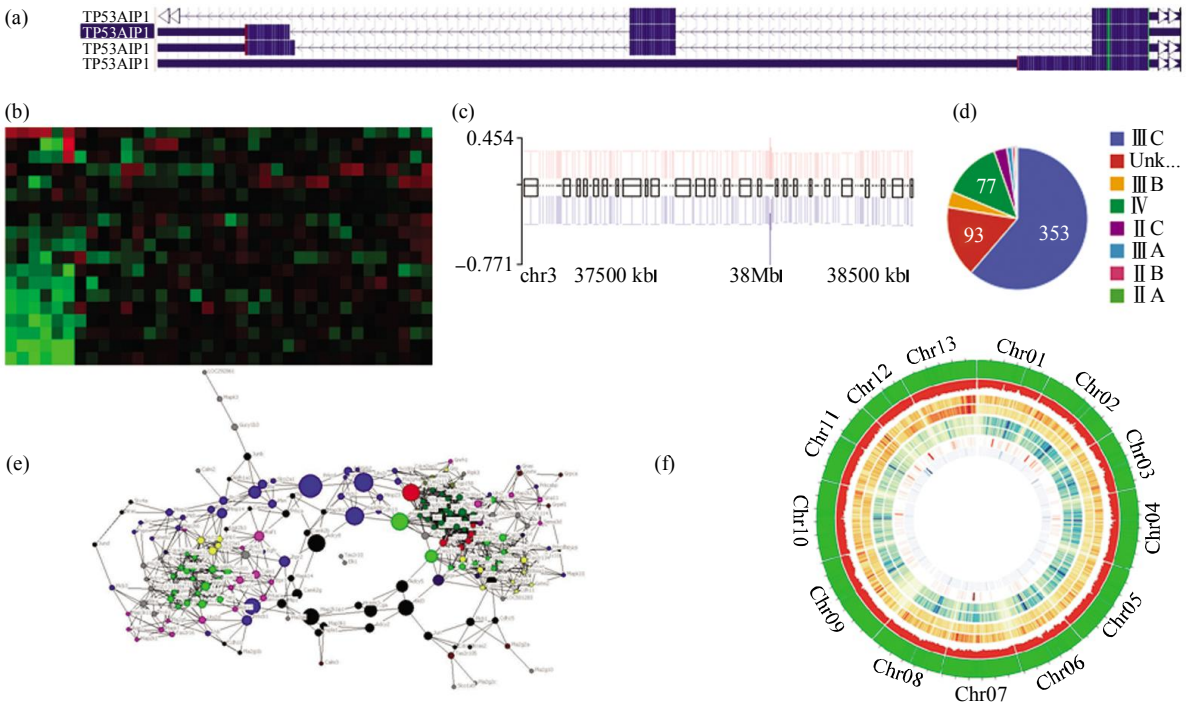

Fig. 2　A review of visualization method in this paper

图 2　本文主要可视化方法汇总

(a) Track. (b) 热图. (c) 箱线图. (d) 比例图. (e) 网络. (f) 环形热图. 侧重序列分析和基因注释的浏览器(如 UCSC Genome Browser，Ensembl 等)，多采用 track 的可视化方式，这种每条数据表现为一个 track 的方式，有利于查看数据的详细信息、做深入研究. 侧重视图浏览的基因组浏览器可视化方法则较多，如图中(b)～(f)所示. 热图的最大优点就是利用人眼对颜色、明暗的敏感性，使不同样本、不同染色体区域的差别淋漓尽致地展现了出来；箱线图和比例图都是展现总体的统计特性；网络的形式有助于查看那些相互关联的基因之间的联系；环形热图的可视化方式非常新颖，不仅有热图的展现，还可以融合 track 和网络的优点，也是一种很好的表现方式.

3 问题和展望

首先，是数据格式不统一的问题．每种不同的生物芯片都有自己的数据生成方法，无论是值的产生方式还是特殊值的处理都有所不同．不同的基因组浏览器对基因的注释方法也是不同的[24]，以UCSC、NCBI Map Viewer 和 Ensembl 为例，有时它们对同一个基因使用不同的符号，所以从它们当中获取的信息并不能直接比较．如果将来能有一个统一的业内标准，可视化工具在选取数据时，就不需要如此繁杂的预处理阶段，可以节省相当可观的工作量．

其次，一般来说，一种基因组浏览器主要面对某一种有特定需求的用户，要想满足所有人的要求是不可能的[25]．侧重序列分析的基因组浏览器，在查看基因序列、加入基因注释方面功能会比较强大，而且基本可以精确到核苷酸碱基位点．而偏重视图型的基因组浏览器，侧重的是“浏览”，是总体情况的概览．研究者可以从大整体中分析出某个局部的信息，或者不同基因组的宏观对比区别，并进一步细化研究．

随着测序技术的进一步发展，要分析的基因组数据可能越来越精细，也越来越大，而目前的网络带宽未必能够跟得上测序的发展速度．即使网络带宽可以发展得足够迅速，服务器向客户端发送未经压缩的数据也是对带宽的极大浪费．后台服务器如果先将数据进行编码压缩存储，在网络上进行传输时就会占用相对较小的带宽，客户端接收到数据后再进行解码，就可以有效地提高前后台通信效率，节省带宽．现在的数据压缩技术已经发展得相对成熟了，像 Google 提出的 protocol buffer[26]协议就是一种轻便高效的结构化数据存储格式，与 xml 相比，用 protocol buffer 处理后的数据要小 3 到 10 倍，占用更少的空间，编解码则要快 20 倍以上，而且向后兼容性好，非常适用于网络传输．当基因组浏览器后台的数据量相当可观时，要做到快速的前后台通信，应用数据压缩技术会是一个不错的选择．

另外，大部分基因组浏览器都是基于上一代 web1.0 的技术开发的，在网站架构、前后台负载均衡上存在不合理之处：大部分数据处理和可视化工作都交给了后台，前台只是被动地接收后台传输的数据或图片，造成了服务器端极大的压力；未能合理地利用局部刷新技术，一个小小的鼠标动作(如拖拽)就重新加载整个页面无疑是极其耗时且浪费带宽的．随着网络技术的不断进步，出现了各种新标签和新功能，像 HTML5 协议中新增加的 canvas 标签，就可以很方便地帮助研究者进行可视化工作．以前发展起来的 SVG 矢量图技术等也可以与新的网络技术相融合，发挥更大的作用．减少页面重量、尽量应用多线程机制，对前后台更有效率的通信都是有益的．

由于基因组浏览器会涉及到很多个人隐私数据，因此数据安全问题是不容忽视的[27]．网站服务器要防止个人隐私基因数据的泄露，保证后台数据库的安全，有效地隔绝跨站脚本攻击和伪造请求．

附件 图 S1～S4，表 S1 见本文网络版附录(http://www.pibb.ac.cn)

参 考 文 献

[1] Kent W J, Sugnet C W, Furey T S, *et al*. The human genome browser at UCSC. Genome Research, 2002, **12**(6): 996−1006

[2] Kent W J, Zahler A M. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. Genome Research. 2000, **10**(8): 1115−1125

[3] Wheeler D L, Church D M, Federhen S, *et al*. Database resources of the national center for biotechnology. Nucleic Acids Research. 2003, **31**(1): 28−33

[4] Hubbard T J P, Aken B L, Ayling S, *et al*. Ensembl 2009. Nucleic Acids Research, 2009, **37**(suppl 1): D690−D697

[5] Skinner M E, Uzilov A V, Stein L D, *et al*. JBrowse: a next-generation genome browser. Genome Research, 2009, **19**(9): 1630−1638

[6] Westesson O, Skinner M, Holmes I. Visualizing next-generation sequencing data with JBrowse. Briefings in Bioinformatics, 2013, **14**(2): 172−177

[7] Stein L D, Mungall C, Shu S Q, *et al*. The generic genome browser: a building block for a model organism system database. Genome Research, 2002, **12**(10): 1599−1610

[8] Kong L, Wang J, Zhao S, *et al*. ABrowse-a customizable next-generation genome browser framework. BMC Bioinformatics, 2012, **13**(1): 2−9

[9] Wang J, Kong L, Zhao S, *et al*. Rice-Map: a new-generation rice genome browser. BMC Genomics, 2011, **12**(1): 165−172

[10] Zun J C, Sanborn J Z, Benz S, *et al*. The UCSC cancer genomics browser. Nature Methods, 2009, **6**(4): 239−240

[11] Wilkinson L, Friendly M. The history of the cluster heat map. The American Statistician, 2009, **63**(2): 179−184

[12] Willianson D F, Parker R A, Kendrick J S. The box plot: a simple visual method to interpret data. Ann Intern Med, 1989, **110**(11): 916−921

[13] Gundem G, Perez-Llamas C, Jene Sanz A, *et al*. IntOGen:

integration and data mining of multidimensional oncogenomic data. Nat Methods, 2010, **7**(2): 92−93

[14] Cerami E, Gao J, Dogrusoz U, *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery, 2012, **2**(5): 401−404

[15] McLendon R, Friedman A, Bigner D, *et al*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008, **455**(7216): 1061−1068

[16] Muzny D M, Bainbridge M N, Chang K, *et al*. Integrated genomic analyses of ovarian carcinoma. Nature, 2011, **474**(7353): 609−615

[17] Sawyer E, Roylance R, Petridis C, *et al*. Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012, **487**(7407): 330−337

[18] Krzywinski M, Schein J, Birol I, *et al*. Circos: an information aesthetic for comparative genomics. Genome Research, 2009, **19**(9): 1639−1645

[19] Thorvaldsdottir H, Robinson J T, Mesirov J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 2013, **14**(2): 178−192

[20] Schroeder M P, Perez A G, Bigas N L. Visualizing multidimensional cancer genomics data. Genome Medicine, 2013, **5**(9): 1−13

[21] Jiang H, Wang F, Dyer N P, *et al*. CisGenome Browser: a flexible tool for genomic data visualization. Bioinformatics, 2010, **26**(14): 1781−1782

[22] Soh J, Gordon P MK, Taschuk M L, *et al*. Bluejay 1.0: genome browsing and comparison with rich customization provision and dynamic resource linking. BMC Bioinformatics, 2008, **9**(1): 450−460

[23] Perez-Llamas, Nuria Lopez-Bigas. Gitools: analysis and visualisation of genomic data using interactive heat-maps. PLoS One, 2011, **6**(5): e19541

[24] Cline M S, Kent W J. Understanding genome browsing. Nature Biotechnology, 2009, **27**(2): 153−155

[25] Nielsen C B, Cantor M, Dubchak I, *et al*. Visualizing genomes: techniques and challenges. Nature Methods, 2010, **7**(3): S5−S15

[26] Muller J, Lorenz M, Geller F, *et al*. Assessment of Communication Protocols in the EPC Network-Replacing Textual SOAP and XML with Binary Google Protocol Buffers Encoding. Industrial Engineering and Engineering Management (IE&EM). 2010, IEEE 17Th International Conference on, 404−409

[27] McIntosh D, Drabic R, Huber K, *et al*. The ethical imperative in the context of evolving technologies. J Psychol, 1996, **31**(3): 3881−3895

Progress on Web-based Genome Browser Technology*

ZHANG Hai-Chuan**, LI Jie**, WANG Ya-Dong***

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, China)

Abstract Advances in sequencing technology have led to a sharp decrease in the cost and rapid increase in the speed of sequencing an entire human genome. It has become one of the most important issues to analyze and visualize genome sequence in life science field. Genome browser technology plays important roles in analyzing genome sequence, interpreting genetic codes, studying complex diseases and so on. In this paper, we review the nine major genome browser technologies and analyze their characteristics in visualized content, visible form, software system architecture and so on. Finally, we discuss the challenges that genome browser faced.

Key words genome browser, visualization, genome sequencing

DOI: 10.3724/SP.J.1206.2014.00055

*This work was supported by grants from Hi-Tech Research and Development Program of China (2012AA02A604), The National Natural Science Foundation of China(61471147, 61173085) and Postdoctoral Science Research Developmental Foundation of Heilongjiang Province (LBH-Q13084).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-451-86413316, E-mail: ydwang@hit.edu.cn

Received: March 1, 2014 Accepted: June 4, 2014

附　　录

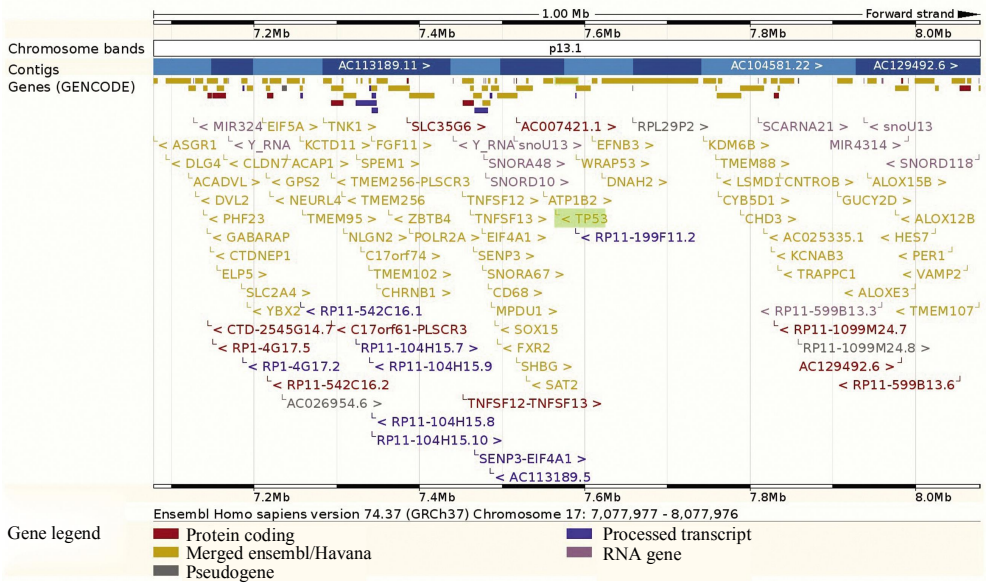

Fig. S1　The "Region in detail" view of Ensembl

图 S1　Ensembl 的基因区域细节展示视图

Ensemble 提供三级分辨率展示(Chromosome summary, region overview, region in detail)，分别对应 Multi Mbs, 1Mb 和 100b 三种尺度. 本图中为精度最高的基因区域细节展示 --region in detail 视图.

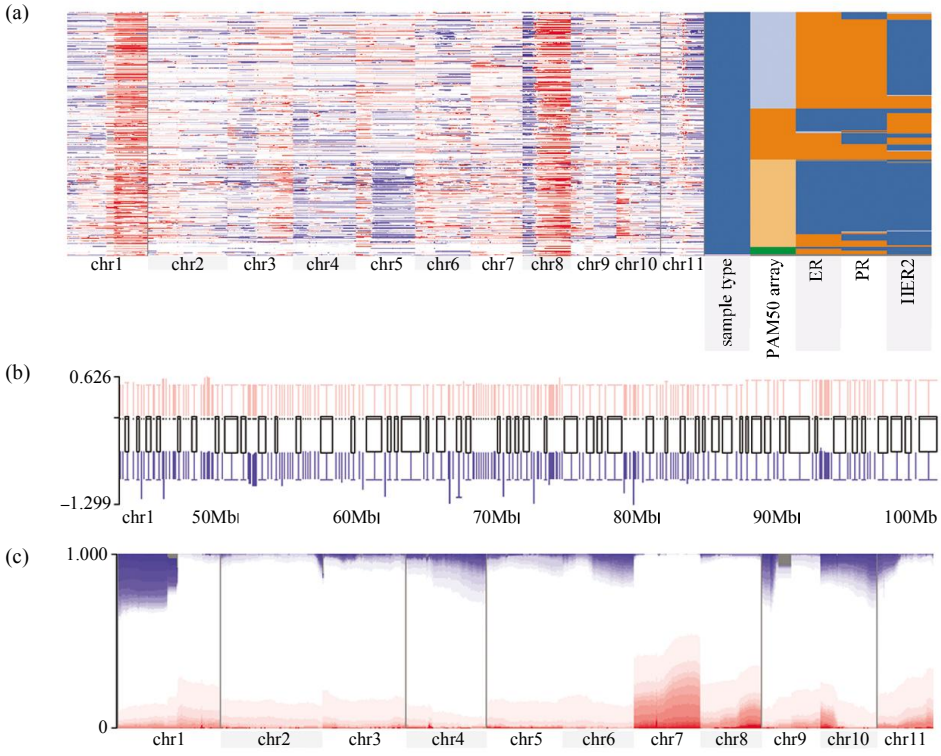

Fig. S2　Heatmap and box plot of UCSC Cancer Browser

图 S2　UCSC Cancer Browser 的数据热图和箱线图

UCSC Cancer Browser 主要的可视化方法是：热图、箱线图和比例图. 热图(a)是将特定的生物值展示为特定颜色的像素点，箱线图(b)和比例图(c)则体现所有样本整体的统计特性.

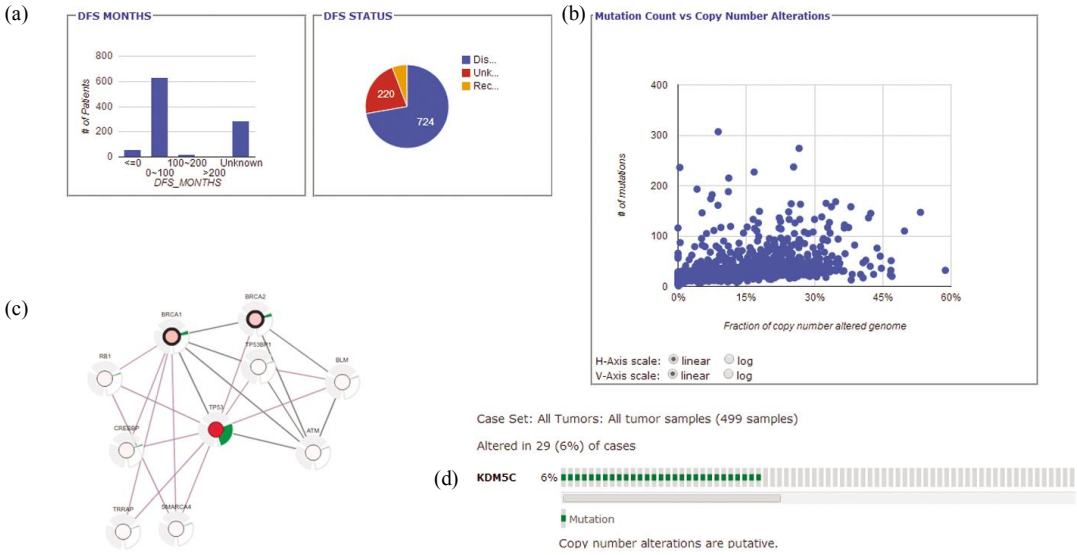

Fig. S3　The visual interface of cBioPortal

图 S3　cBioPortal 的可视化视图

cBioPortal 采用的可视化方案比较丰富，有直方图和饼状图(a)、散点图(b)、网络(c)、热图(d)等. cBioPortal 的热图只针对用户选定的某一个样例，所以形式简单一些，称为 OncoPrint，如 D 所示. cBioPortal 提供多种统计视图：直方图，饼状图，散点图等.

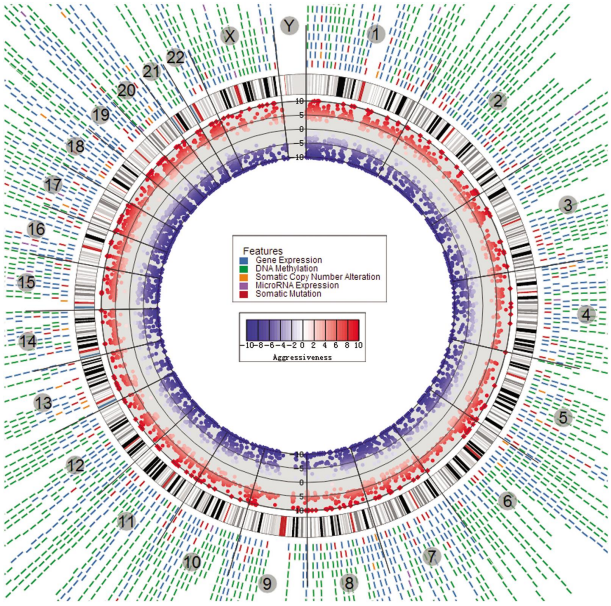

Fig. S4　The visual interface of CRC Aggressiveness Explorer

图 S4　CRC Aggressiveness 浏览器界面

CRC Aggressiveness 采用环形热图的方式，首尾的概念被淡化，圆形的展示方式更美观. 每一条圆环为一条 track，不同颜色的点、线标记分别代表不同的生物含义. 鼠标在每个色块悬停时都会有弹窗链接到其他基因组浏览器.

Table S1 Comparison of genome browsers

表 S1 各基因组浏览器比较

浏览器名称及参考文献	网址	开发语言	存储数据库	数据来源	数据下载	提供 API	开源	针对的物种 / 疾病	处理用户数据	可视化方法	可视化内容	优缺点
UCSC Genome Browser[1]	http://genome.ucsc.edu/	Java/python/C	mysql	UCSC Genome Bioinformatics Group 等	√	√	√	人类, 黑猩猩等 88 种	√	track	组装序列间隙 / 重叠、mRNA 和表达序列标签队列、多基因预测、跨物种同源性、SNP、序列标记位点、辐射杂交数据、转座子重复等	比较综合，功能比较齐全，提供接口比较多. 但是没有充分利用局部刷新技术，服务器端的部分脚本对访问频率有限制
NCBI Map Viewer[3]	http://www.ncbi.nlm.nih.gov/mapview/	C/C++	sql/custom/Oracle	GenBank	√			人类, 大鼠等 242 种		track	染色体图谱、contigs、基因、SNPs、G3/GB4 辐射杂交图、mRNA alignments 等	有几乎最全面的数据库，提供多种视图来分析不同的功能. 可视化操作上仍有一些不足
Ensembl Genome Browser[4]	http:// ensembl genomes.org/	Perl/C	mysql	InterPro, OMIM, SAGE, HUGO 等	√	√	√	人类, 小鼠等 66 种	√	track	基因登录号、基因注释、变异信息、蛋白质结构域和家族、基因转录子信息、内含子 / 外显子、基因在基因组上的前后序列、不同物种之间基因组的比较等	在对真核生物的基因注释上功能非常强大，可以做到准确、全自动. 然而，由于前后台通信采用的是 CGI 协议，在局部刷新技术的应用上存在一些不足
JBrowse[5, 6]	http://jbrowse.org/	Javascript/perl	json 文件	文件	√	√	√	人类	√	track	基因跨度、tRNA、转座子、寡核苷酸、蛋白质结合位点、增强子、基因调控区域、非编码 RNA、点突变、序列变异信息等	充分利用了 ajax 局部刷新和 cookies 技术，缓解了服务器端压力. 但可视化方法仅仅是一些 html 标记，前端绘图时需要运行大量脚本，要考虑浏览器兼容性
ABrowser[8]	http://www.abrowse.org/	Java/jsp/ javascript	mysql	TAIR 9, Vista, GEO, 用户自定义文件等	√	√	√	用户自定义数据	√	track	基因组视图、track 详细属性信息等	开源性好，支持用户通过自己的文件定制 Genome Browser
UCSC Cancer Browser[10]	https:// genome-cancer.ucsc.edu/	Javascript/C	mysql	TCGA, SU2C 等	√			人类 46 种癌症类型	√	热图，箱线图，比例图等	拷贝数 / 拷贝数变异、DNA 甲基化、外显子表达、基因表达、mRNA、RNAseq、蛋白质结构、临床数据等	整合了临床数据，可以根据临床特殊对热图进行排序，可提供统计特征. 用户可视化界面交互上有一些不足
IntOGen[13]	http://beta.intogen.org/	Java	mysql	TCGA 等				人类 12 种肿瘤		热图	基因扩增、纯合子缺失、突变 / 突变频率、evidence 等	所有样本都是根据肿瘤疾病国际分类组织的词库人工进行标定的，而且是分析生物模块(eg, pathway)对癌症的影响，而非分析个人基因
cBioPortal[14]	http://www.cbioportal.org/public-portal/	Java/jsp/ javascript	mysql	TCGA, ICGC 等	√	√	√	人类 56 种癌症	√	热图, 多种统计图, 网络	序列信息、mRNA、突变频率、拷贝数变异、miRNA、甲基化、RPPA 等	简单易用，但缺少一个包含所有样本的、友好的可视化界面
CRC Aggressiveness Explorer[17]	http://explorer.cancerregulome.org/crc_agg/	Javascript/css	mysql	TCGA 等	√	√	√	人类直肠癌	√	环形热图, track	外显子序列、基因突变、DNA 拷贝数、启动子甲基化、信使 RNA 和 microRNA 表达、易位、CRC 中改变的 pathway 等	可视化方法新颖，分析方向较单一