

四 人类和其它脊椎动物基因组数据库

Ensembl数据库 (2009更新)

Ensembl (<http://www.ensembl.org>) 是European Bioinformatics Institute (EBI) 与Wellcome Trust Sanger Institute (WTSI) 共同合作开发的数据库项目。Ensembl近年来发展非常迅速, 网站使用频率直追GenBank。

Ensembl为人们提供了一个全面的基因组信息库, 包括基因数据存储、信息整合、数据分析以及生物信息可视化处理等功能。

现阶段, 为了方便用户使用网站, Ensembl作出了许多改进, 包括重新设计网页以及使用Enredo-Pecan-Ortheus系统和新软件设备 (尤其是新开发的支持Ensembl基因组计划的软件<http://www.ensemblgenomes.org/>) 获得新一代多重基因组比对系统 (multiple genome alignment) 和祖先序列 (ancestral sequence)。

1 Ensembl新网页

由于大多数用户都通过网站使用Ensembl, 因此对网站进行改版就显得尤为重要。影响网站使用性最主要的因素就是网站速度和可发现性 (discoverability)。随着网站的发展, 数据变得越来越多, 越来越复杂, 网站也就会变得越来越庞大, 因此速度会变慢, 使用者也越来越难发现网站的新功能, 越来越“不会”使用网站。为了解决这些问题, Ensembl投入了大量人力物力。最终, Ensembl于2008年11月推出了全新改版的Ensembl 51。

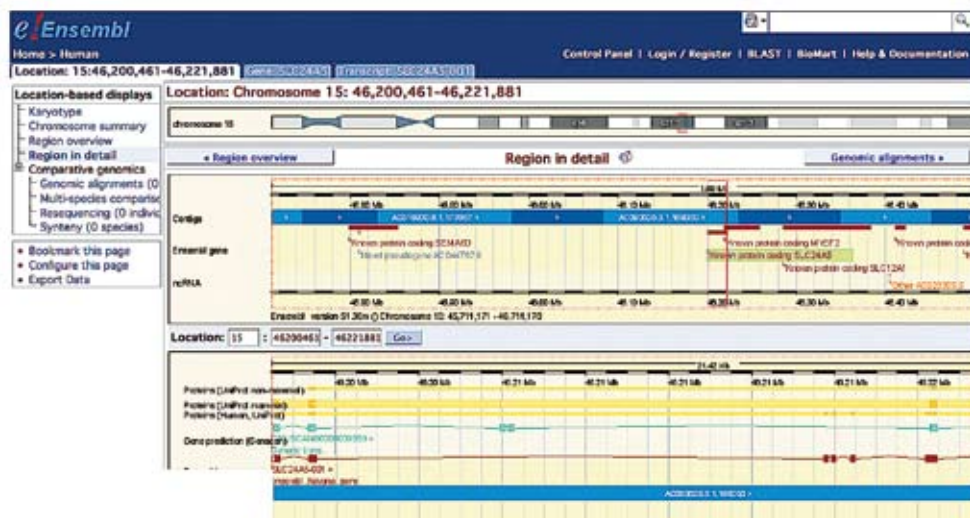
在全新改版的Ensembl 51中, 打开一个网页需要提交给网络服务器的链接数比过去要小得多, 从而大大方便了那些离服务器较远的用户。为了改进Ensembl的服务并测试速度对网站的影响, 2008年初, Ensembl还在世界各地的十几个地方设置了自动持续监测系统来监测Ensembl网站的速度。

修改网页编码以提高网站速度的同时, Ensembl也对网站进行了重新设计, 以增强网站的易操作性 (navigability) 和可发现性 (discoverability) (图1)。

新网站分为四个部分——定位 (Location)、基因 (Gene)、转录体 (Transcript) 和变异体 (Variation)。用户可以轻易地使用每一个网页顶部的标签在这几个部分之间进行切换。

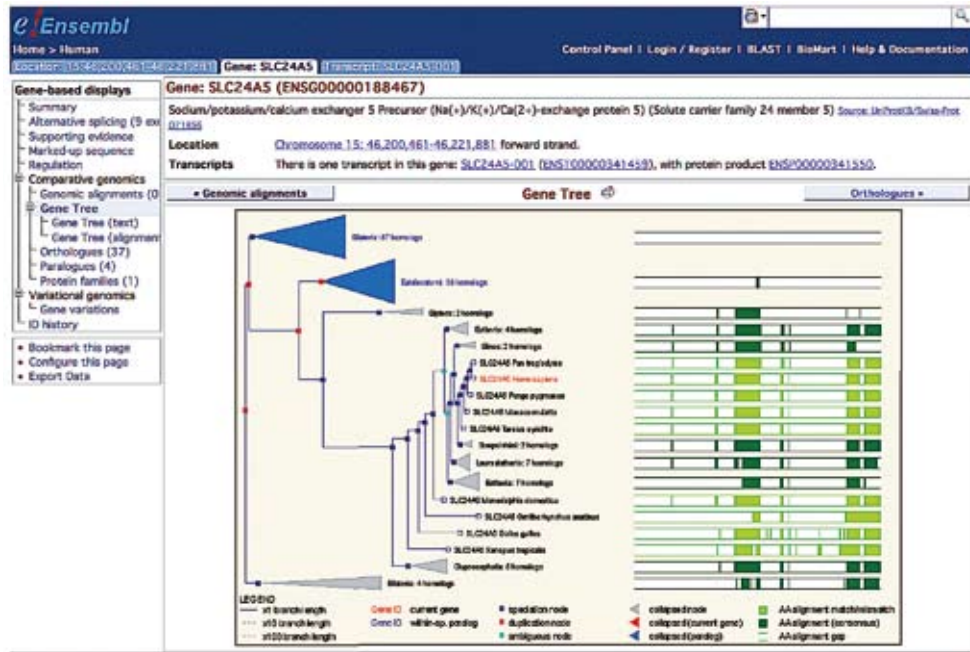
“定位”部分包括不同分辨率的基因组序列图 and 比较形式 (comparative view) 的基因组序列图 (图

A



1A: 定位信息, 显示该基因在基因组中的位置)。“基因”部分包括基因文本信息、该基因在基因组中的定位信息、Ensembl数据库中该基因的直系同源基因和旁系同源基因以及该基因的变异情况等信息(图1B: 基因信息, 显示基因树)。“转录体”部分与“基因”部分类似, 不过着重展示的是转录体结构等相关细节(图1C: 转录信息, 显示转录相关信息)。“变异体”部分主要包括SNP等信息。

B



C

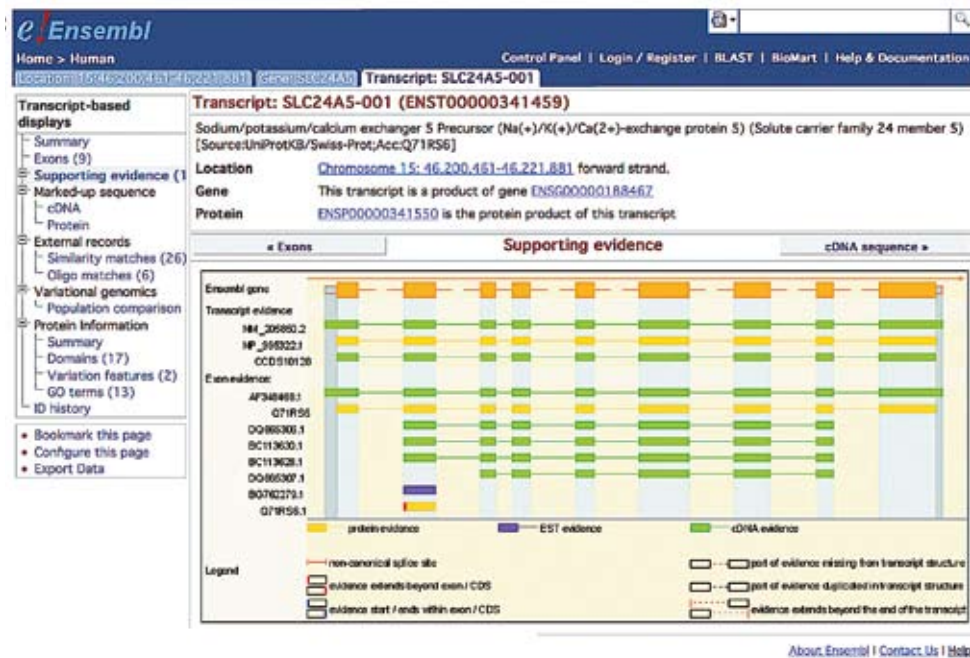


图1 Ensembl 51版网页的截图, 显示该新版本的特征。

A、B、C三幅图使用人类基因SLC24A5举例, 代表了新版本显示内容4个大类中的3大类。

点击网页顶部的3个标签(选项卡)可以快速地在三个显示类别中切换。图1没有显示第4个变异信息标签, 该标签在有SNP时会出现。每一个网页左边的菜单都以列表的形式给出了不同的浏览模式。

在A中包括有基因组形式浏览和比较基因组形式浏览。在B中, 包括基因的文本形式信息, 该基因在人类基因组中的信息, 该基因在Ensembl数据库中的直系同源基因和旁系同源基因, 及这些同源基因中的变异情况等。在C中和B中展示的情况差不多, 不过更加着重于表现转录体的结构细节等信息。除了这些网页导航上的新设计之外还有许多使用AJAX对网页编码进行的改进。例如B中的基因树, 允许进行交互式节点扩展或减少, 这种新功能对于大的基因家族来说非常实用。在C中和A中也会发现许多以前没有的扩展显示选项。

2 新收录的新物种数据和修订的基因注释信息

现在Ensembl中共收录有48个物种的数据，其中41种是脊椎动物。Ensembl数据库的主要任务之一就是提供尽可能准确、全面的基因集合（gene set），在分析新脊椎动物基因组数据时也可以参考这些现存的基因集合数据。最近发表了借助Ensembl数据所获得的新物种基因组数据，包括鸭嘴兽（*Ornithorhynchus anatinus*）、灰色短尾负鼠（*Monodelphis domestica*）以及恒河猴（*Macaca mulatta*）的基因组数据。研究人员是使用蛋白质序列和cDNA序列比对而获得的，不过还需要对这些数据进行进一步处理，使其更准确、覆盖率更高。

Ensembl今年的工作重点是使用同源比较系统（compara homology pipeline）开发一套系统性的后基因结构比较分析系统（systematic post gene build comparative analysis process），以了解进化上表现不一致的基因最初的结构。然后，将这些数据交给更加依赖计算机的、使用更敏感参数的局部基因结构处理系统（localized gene build pipeline）进行第二步处理。其中最主要的问题是发现并确认断裂基因（split gene）、丢失的直系同源基因（missing orthologous gene）、进行了部分预测的基因（partially predicted gene）和假外显子（false exon）。

在用共有20,322个基因的马基因组对该技术进行测试的时候，这种后处理系统（post-processing pipeline）发现了236个断裂基因，发现了除直系同源基因之外的1013个以前没有发现的新基因，将部分预测基因数目增加到了1330个，还剔除了840个假外显子。现在，这种技术已经被系统的用于其它高覆盖率的哺乳动物基因组分析中了。这些新的基因集合数据将在Ensembl的下一版本中公布。

3 比较基因组学多重比对

2008年以来，Ensembl就对全基因组范围比较基因组系统（genome-wide Ensembl comparative genomics pipeline）进行了重大调整，现在称作EPO系统（Enredo-Pecan-Ortheus pipeline）。EPO系统由三个互相合作的程序组成。Enredo负责批量处理基因组数据，生成分段

Ensembl的另一项主要工作，就是和其它组织合作对人类基因组数据进行进一步修正。这些合作组织包括Sanger研究院的HAVANA小组、NCBI的参考序列（RefSeq）项目小组等。CCDS（<http://www.ncbi.nlm.nih.gov/CCDS/>）是公认的高质量蛋白质编码区域及注释信息数据库。现在，CCDS收录的编码序列（CDS）条目数已经从18,290条上升到了20,159条，也就是说基因数从16,003条增加到了17,052条。CCDS中还收录有小鼠数据，小鼠的CDS条目数也从13,374条增加到了17,707条，基因数从13,014个增加到了16,889个。GENCODE以CCDS和ENCODE计划为基础，收录了包括UTR数据在内的转录体数据，还加入了计算机和试验输入及验证数据。一种新的基于GENCODE的，在多种哺乳动物基因组间进行序列比对的计算机方法可以被用来预测基因编码区域。使用Ensembl基因架构系统（gene build pipeline）预测出的好几百个转录体在用上述新方法进行分析后发现得分很低，最终被证实为假基因而从数据库中剔除了出去。Ensembl/HAVANA合作项目包括进一步提高不同数据库间的基因相符率，例如与UniProt进行更紧密的联系，与基因组参考协会（Genome Reference Consortium，<http://www.sanger.ac.uk/sequencing/grc/>）合作找出人类基因组序列和转录体之间不相符的地方等等。

Ensembl中公布的Ensembl/HAVANA人类基因数据是上述所有项目共同努力的成果，其中包括了CCDS中的资料，HAVANA的全长转录体注释信息等。

另一项进展是为所有的基因和物种定义了标准的转录体（canonical transcript）。这标准的转录体可能是最长的CDS（如果该基因有可被翻译的转录体），也可能是最长的cDNA。

图像（segmentation graph），找出那些同血统的同源片段（colinear homologous segment）。和Ensembl以前使用的算法不同，Enredo处理的是特异性复制谱系（lineage specific duplications），例如灵长类谱系复制生成了携带灵长类特异性基因的两个拷贝。这些相同来源的片段再交给Pecan处理，

Pecan是一个多重比对工具，能通过序列比对找出高度同源的区域。使用祖先重复片段（ancestral repeats）评估，Enredo+Pecan就能比其它哺乳动物比对工具加起来做得更好。最后，Ortheus——遗传序列重建程序就能得到准确的祖先序列（ancestral sequences）。Ortheus使用的是分支转换模型（branch transducer model，该模型属于HMM的一种，能发现序列缺失和插入）来推测祖先序列。

图2展示的是在Ensembl中限制各种不同条件后进行基因组进化速率评测（Genomic Evolutionary Rate Profiling, GERP）分析得到的结果。结果发现编码外显子和祖先重复片段间存在非常明显的差别，调节区域位于这两者之间。2008年3月公布的Ensembl 49第一次展示了7种哺乳动物进行EPO比对的结果。2008年7月公布的Ensembl 50中进行这种比对物种范围扩大到23种哺乳动物，包括低覆盖率的基因组数据，还有包括人、黑猩猩、猩猩和猕猴这4种灵长类动物的比对结果。Ensembl计划将来在硬骨鱼类（teleost）动物中也进行一次EPO比对。

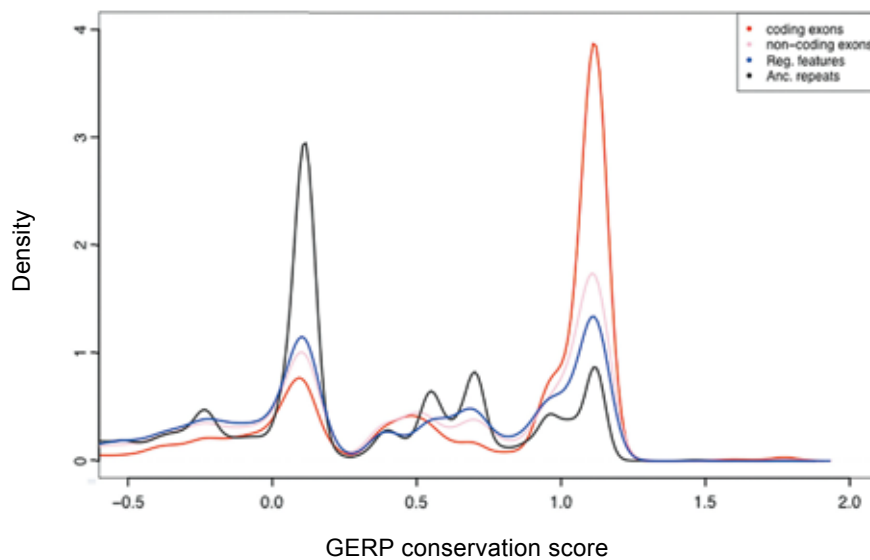


图2 哺乳动物基因组和人类X染色体比对后计算出的GERP保守评分（GERP conservation scores）。红色代表编码外显子（coding exon），粉红色代表非编码外显子（non-coding exon），蓝色代表调节区域（Reg. feature），黑色代表祖先重复片段（ancestral repeat）。GERP 0分表示没有选择压力，而高GERP评分表示有选择压力存在。非编码外显子包括蛋白质编码基因中所有的非编码区。调节区域包括所有收录在Ensembl调节区域数据库中的信息。祖先重复片段只包括由RepeatMasker定义的MER II型转座子。使用EPO比对可以发现外显子和调节区域与重复片段有非常明显的差异。

为了得到23种哺乳动物的EPO比对结果，就要能对低覆盖率的基因组数据进行分析。低覆盖率基因组大部分数据都是片段数据，导致Enredo分析图中会显示出许多“断点（breakpoint）”，Enredo只能用于分析高覆盖率的基因组数据。而使用 and 人类基因组配对比对（pairwise alignment）的方法来获得低覆盖率基因组的共线性区域（colinear region）。对于每一个低覆盖率基因组，通过配对比对方法发现的片段都会链上一串“N”，这有利于最后的多重序列比对。比对结束之后，这些N就会被去掉。

除了提供基因组序列比对之外，Ensembl比较基因组学工具还能构建基因树以及进行直系同源/旁系同源预测。Ensembl 50中最明显的改变是使用SLR程序计算基因树中dN/dS值的方法。这些值允许人们通过比对的方式找出那些承受进化压力不同的位点。

4 功能基因组学和变异体资源

过去几年，基因组学研究领域最大的改变就是功能基因组学研究。由于参与了ENCODE计划和EU FP6资助的HEROIC等国际性研究计划，Ensembl也开展了一些基础工作来处理功能基因组学数据。此外，Ensembl最近又参与了基因组DNA甲基化资源的创建工作，这些资源将来也会收录在Ensembl数据库中。由

于有了新一代的测序技术，有了ChIP-chip技术，功能基因组测序也将变得非常容易。今年主要的工作就是发展ChIP-seq分析系统，包括为ChIP-seq数据分析制定出算法。

Ensembl计划有一个特点就是不仅提供基因数据及其比对服务，还能根据这些数据借助自动分析系统提供高质量的生物学预测服务。例如，Ensembl基因结构系统就能预测蛋白编码区域，Ensembl比较分析系统能“画出”基因树，分析出直系同源基因和旁系同源基因。Ensembl调节系统是最新开发的系统，它能自动对人类基因组内的潜在调节区域进行注释。原始输入信号是使用DNA酶I超敏位点作图法得到的开放染色质图和用ChIP检测到的组蛋白修饰信息。自去年公布了第一个结果之后，Ensembl已经对该方法升级了3次，每一次升级都加入了新的信息，加入了更加精确的染色体构象和修饰信息。该数据库现在收录了来自包括CD4细胞在内的好几种细胞系的大约175,000个基因组区域。还收录了大约40种不同的组蛋白修饰情况和超过2700种这些因素的组合模式，这些组合模式与23,000种蛋白质编码基因或它们的启动子调节机制有关，可以据此对基因相关调节模式或启动子相关模式进行分类。

新一代测序技术也给Ensembl中的数据变化带来了很大影响。然而，Ensembl仍然是从dbSNP数据库载入数据，dbSNP数据库是通过重新测序

由计算机发现的变异数据组成的数据库。该数据库也因为新一代测序技术的飞速发展而数据量猛增。今年，Ensembl继续从三个连续的dbSNP数据库（dbSNP数据库127、128和129）载入数据。此外，它还从鸭嘴兽和猩猩重新测序的SNP数据库（resequencing-based SNP）载入数据，以及从Watson和Venter的人类基因组重测序结果中载入数据。鸭嘴兽的SNP数据已经提交给了dbSNP数据库，它是目前已知SNP数量最大的物种。猩猩的SNP数据将与猩猩的基因组数据一起公布。

在变异体数据库中，Ensembl对拷贝数变异数据的收录增多了，也增加了SNP的注释信息，例如，与疾病有关的SNP和表达数量性状位点（expression QTL）等。Ensembl变异体数据库是与欧洲基因型档案（European Genotype Archive, EGA, <http://www.ebi.ac.uk/ega/>）和欧洲生物信息学研究所（European Bioinformatics Institute, EBI）下属的千人基因组计划（1000 Genomes Project, <http://www.1000genomes.org/>）合作的。EGA开始于2008年春，目前管理着好几个项目汇集来的数据，这些项目包括威康信托基金会病例对照协会（Wellcome Trust Case Control Consortium）的项目和其它几个处于数据公布前夕的项目。这些组织间的协作将有力推动Ensembl变异体数据库的发展，同时有利于即将开展的功能注释工作。

5 未来的发展方向

新一代基因组测序技术的影响力已经开始显现，Ensembl也将做出相应的改变来适应这种新局面。至今，数据规模急剧扩大将是大部分生物信息资源数据库面临的巨大挑战。对于变异数据库来说，最急迫的挑战则是如何展现正在进行的千人基因组计划即将公布的结果。Ensembl已经开始开发出对新一代测序技术得到的转录组数据进行分析的系统。可以想象，这么多不同细胞系来源、不同发育阶段来源的数据汇集到一起，可以给人们带来多少信息。有了这些信息，人们对于可变剪接以及它们表达的时空定位等信息就会有一个全面的了解了。

目前，基因组测序的重点是发现变异，不过由于试验技术和计算机技术的进步，人们可以从头

开始对大基因组进行测序和组装了。这时候再对哺乳动物基因组进行测序就会便宜很多。不过，使用Ensembl技术来处理这么多的基因组数据还在进行之中，Ensembl为此专门设立了一个Ensembl基因组项目（<http://www.ensemblgenomes.org/>），该项目将用Ensembl技术建立一个检索不同生物基因组的入口。对此项目具有重要意义的应用程序接口（API）和构架规划，包括如何在一个核心数据库中存储几个物种的信息等技术正在开发之中。

最后，从人们对网站使用情况的监测中可以很清楚地发现，尽管Ensembl改变了网页编码，提高了网站的速度，但是网络延迟效应（network latency effects）还是会对离服务器较远的用户造成影响。因此，Ensembl会架设平行镜像站点，来

改善网站的速度。由于中国和英国之间的网络链接速度较慢，Ensembl最近与中国深圳的华大基因研究中心（BGI-SZ）合作在中国设立了镜像站点。该站点的服务对象是中国和中国周边用户。稍后，Ensembl也会在美国西海岸架设镜像站点。还将尝试使用“云计算设备（cloud compute facilities）”架设服务器。

原文检索：Nucleic Acids Research, 2009, Vol. 37, Database issue D690–D697



UCSC Genome Browser数据库（2009年更新）

UCSC（University of California Santa Cruz，美国加州大学圣塔克鲁兹分校）建立的基因组数据库（Genome Browser Database, GBD, <http://genome.ucsc.edu>）是一个开放的、收录有众多物种基因组序列信息的数据库，库中还集合了大量的注释信息和比较基因组学信息，它提供人类和众多其它物种的DNA序列查询、搜索服务。此外，该数据库还提供根据参考基因组比对得出的基因注释信息。

用户在使用数据库及其工具（Genome Browser、Table Browser、Gene Sorter、Proteome Browser、VisiGene、Genome Graphs、BLAT等）时可以从以下站点获得大量的适时帮助，包括<http://genome.ucsc.edu/goldenPath/help>、<http://genome.ucsc.edu/FAQ>、<http://openhelix.com>等。还可以写邮件到genome@soe.ucsc.edu获得帮助。

1 新物种信息

目前，GBD新增了13个新物种的基因组序列信息，包括猩猩、绒猴、豚鼠、斑胸草雀、八目鳗、文昌鱼和三种线虫品种——*brenneri*、*remanei*、*japonica*在内的9个以前没有收录的物种信息，以及牛、斑马鱼、海胆、秀丽隐杆线虫（*C.elegans*）这4个已收录物种的更新信息。GBD为每一个新信息都提供了注释，也将这些信息和GenBank中的其它物种序列进行了比对。此外，他们还对上述9种新物种信息中的7个物种进行了多重比对注释，还将6种蠕虫的序列和最新的秀丽隐杆线虫序列进行了比对。

2 UCSC基因组数据库的新注释信息

除了收录新物种序列之外，GBD还在去年新增了200多条注释信息。可以点击Genome Browser上的相应按钮获得更多新注释信息。

对人类基因组集合（数据库）hg18和基因及基因预测组（Genes and Gene Prediction Track Group）中的Pos Sel基因进行新的注释后发现了承受正向选择压力（positive selection）的基因。网站上显示了通过对人类、黑猩猩、猕猴、小鼠、大鼠和狗基因组进行多基因组比对后筛出的全基因组范围内承受正向选择压力的基因。同时，还使用了9种基于Yang和Nielsen发明的branch-site framework模型的似然比检验法（likelihood ratio test, LRT）对一些直系同源基因进行了检测来验证上述结论的正确性。

开放的调控元件注释项目（Open Regulatory Annotation, OregAnno）的研究已经取得了一定成果，获得了人类、小鼠、黑腹果蝇和酵母这四种模式生物调控元件的注释信息。每一条OregAnno的注释信息包括经过试验验证后公开的基因调控序列（如启动子、增强子等）、转录因子结合位点以及调控区域的多态性（regulatory polymorphism）等信息，同时每一条OregAnno的注释信息也都会链接到OregAnno数据库。

数据库中，人类目录下现在还收录了Kidd等人对国际人类基因组单体型图计划（HapMap Project）里8个人的序列同参考序列比对后获得的注释信息（HGSV Discordant Track）。这些注释信息显示了那些人类序列中与参考序列的不符之处，表明该处可能存在着大段的缺失或插入序列，这对寻找人类基因组中的变异

具有非常重大的意义。

小鼠mm9集合（数据库）现在提供了小鼠与其它30种脊椎动物的比对信息。这些比对信息是通过多重比对和phastCons计算（phastCons computation）之后得出的，它们有助于了解不同物种之间在进化上的相关性。GBD还在小鼠mm9集合（数据库）中新增了一个子数据库用来收录从维尔康姆基金会桑格研究所（Wellcome Trust Sanger Institute）miRBase数据库中获取的microRNA信息。

在大鼠rn4集合（数据库）中GBD还提供了从RGD中获取的数量性状基因座（QTL）信息。这些QTL信息与大鼠基因组中1000多个与血压、血糖等处于持续动态波动之中的表型特征相关基因座有关。

在dm3基因组数据库中收录有黑腹果蝇基因组中超过7500个基因插入突变的注释信息。通过网站上提供的这些注释信息加上与位于美国布鲁明顿的信息储存中心之间的链接，可以发现果蝇基因组中由P元件和Minos元件导致的插入。

3 UCSC基因组数据库新收录的基因

2008年9月，最新版的UCSC基因数据——hg18人类基因数据集发布。UCSC基因注释信息包括参考各种数据库（RefSeq、UniProt、GenBank）后使用比较基因组学方法得出的已知编码基因和非编码基因的多种异构体信息。

在CCDS信息和RefSeq信息不一致时，最新的UCSC基因注释信息就会使用CCDS蛋白质信息来选择最合适的比对结果。GBD做出这个选择是因为他们相信国际公认的蛋白质信息比随便比较一下基因串联重复序列和转录体RNA 5'端所获得的差别更有意义。例如在对人类基因IFI35（位于hg18 chr17:38,418,889-38,419,044上，<http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&position=chr17:38418889-38419044&knownGene=pack&refGene=pack>）第四号外显子和第五号外显子之间的剪接情况进行注释时，他们选择的就CCDS信息。

新的UCSC基因库中包括66,803个基因（包括异构体）信息，其中13,767个基因是非编码基因，这些基因共组成26,570个基因簇（表1）。

表1 UCSC中新基因情况简要一览表

UCSC基因种类	基因数	基因簇数	以前基因簇数	改变数目
编码基因	53,036	20,409	20,433	-24
非编码基因	13,767	6161	5871	+290
基因总数	66,803	26,570	26,304	+266

这次更新还在Genome Browser中新增了与外部数据库中几种模式生物内直系同源基因之间的链接。这些模式生物数据库包括小鼠基因组数据库（MGD）、大鼠基因组数据库（RGD）、斑马鱼信息库（ZFIN）、线虫信息库（WormBase）、果蝇信息库（FlyBase）和酵母基因组数据库（Saccharomyces Genome Database）。他们还计划继续定期更新，以保证人类基因组数据和小鼠基因组数据都是最新最全面的。

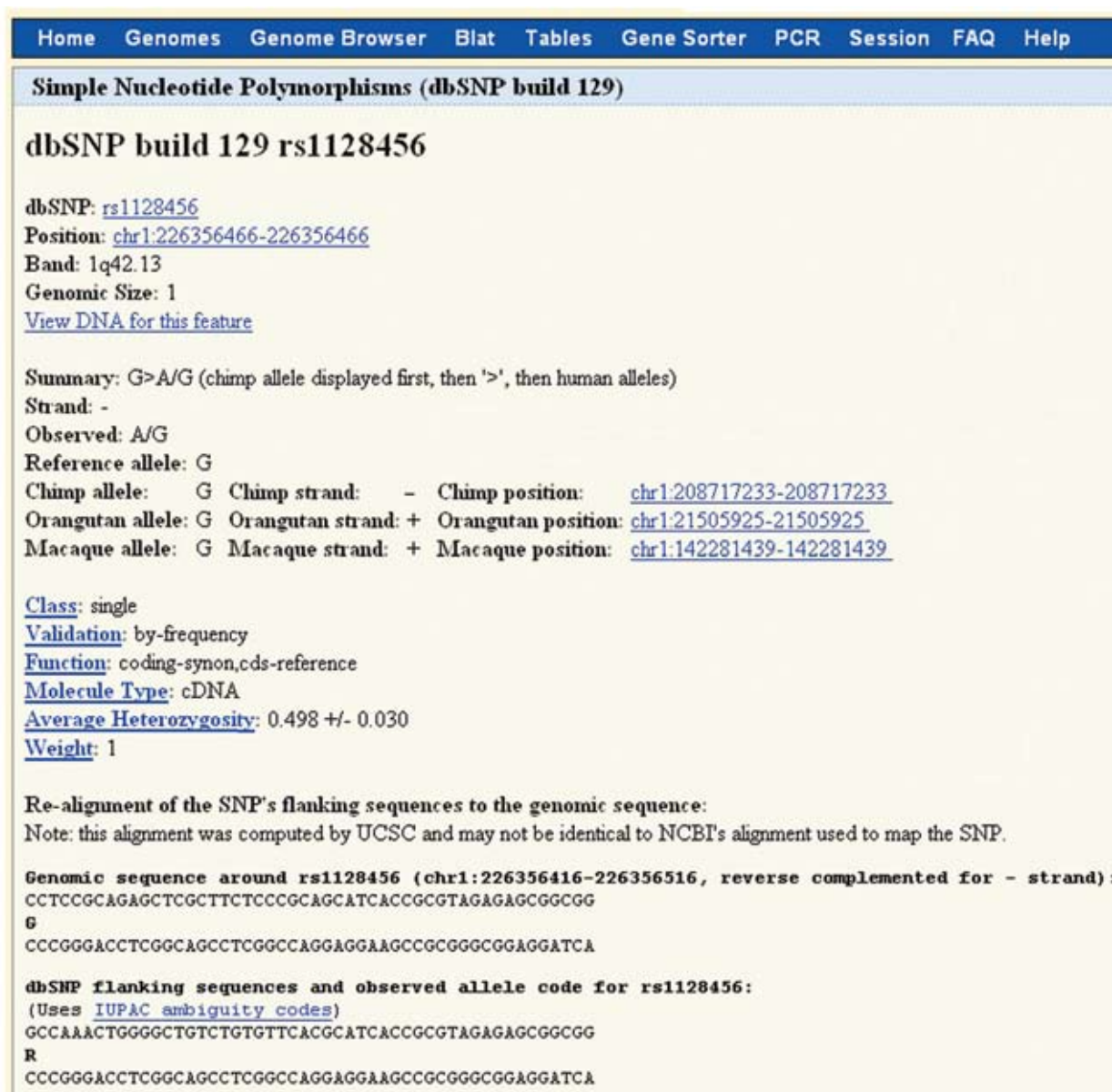
使用Genome Browser浏览UCSC数据库中某一个基因的注释信息，只需要在Genome Browser网页的搜索框中输入该基因的名称等关键词就可以了。同时，可以使用各种基因标识符来进行搜索，例如HGNC名称、UniProt ID、即使是GenBank或UniProt中对该基因的一些描述性关键词都行。通过搜索还可以发现与目的基因产物间有相互作用的基因，对这些基因的注释以RefSeq形式给出。

UCSC Gene track页面还包括与网站内资源之间的链接，例如与GeneSorter、Proteome Browser、VisiGene、*in situ* hybridization image archive等页面的链接。今年还新增了与艾伦脑科学研究所（Allen Brain Institute）的Human Cortex Gene Expression data数据库、Human Genome Epidemiology（HuGE）data数据库、Comparative Toxicogenomics Database（CTD）等外部数据库之间的链接。

4 基因变异信息

hg18人类基因库提供了大量人类基因变异信息，其中有一些信息是去年才新增的。尤其值得一提的是他们从dbSNP 129中补充了一些以前dbSNP 128和dbSNP 126中没有的SNP信息。

Genome Browser中可以查询到这些SNP信息的详细情况，包括SNP的类型（例如编码区、非编码区、同义突变等等）。GBD现在还能将包含这些SNP位点的序列与参考序列进行比对。此外，为了用户比较的方便，他们还给出了几种灵长类动物（黑猩猩、猩猩、猕猴）的直系同源基因供大家使用。图1就显示了SNP 129网页查询SNP rs 1128456的部分结果。



Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Simple Nucleotide Polymorphisms (dbSNP build 129)

dbSNP build 129 rs1128456

dbSNP: [rs1128456](#)
 Position: [chr1:226356466-226356466](#)
 Band: 1q42.13
 Genomic Size: 1
[View DNA for this feature](#)

Summary: G>A/G (chimp allele displayed first, then '>', then human alleles)
 Strand: -
 Observed: A/G
 Reference allele: G
 Chimp allele: G Chimp strand: - Chimp position: [chr1:208717233-208717233](#)
 Orangutan allele: G Orangutan strand: + Orangutan position: [chr1:21505925-21505925](#)
 Macaque allele: G Macaque strand: + Macaque position: [chr1:142281439-142281439](#)

[Class](#): single
[Validation](#): by-frequency
[Function](#): coding-synon, cds-reference
[Molecule Type](#): cDNA
[Average Heterozygosity](#): 0.498 +/- 0.030
[Weight](#): 1

Re-alignment of the SNP's flanking sequences to the genomic sequence:
 Note: this alignment was computed by UCSC and may not be identical to NCBI's alignment used to map the SNP.

Genomic sequence around rs1128456 (chr1:226356416-226356516, reverse complemented for - strand):
 CCTCCGCAGAGCTCGCTTCTCCCGCAGCATCACCGCGTAGAGAGCGGCGG
 G
 CCCGGGACCTCGGCAGCCTCGGCCAGGAGGAAGCCGCGGGCGGAGGATCA

dbSNP flanking sequences and observed allele code for rs1128456:
 (Uses IUPAC ambiguity codes)
 GCCAAACTGGGGCTGTCTGTGTTACGCATCACCGCGTAGAGAGCGGCGG
 R
 CCCGGGACCTCGGCAGCCTCGGCCAGGAGGAAGCCGCGGGCGGAGGATCA

图1 部分显示SNP 129搜索页面。搜索人类1号染色体上的SNP rs 1128456。

此外，他们还从dbSNP 128中获取信息更新了mm9 SNP注释信息，从dbSNP 127中获取信息更新了bosTau3牛数据库信息。

GBD使用比较基因组学对人hg18数据库的注释进行了更新，比较了人、小鼠和狗的基因组进化上保守的部分，找出了人类参考基因组序列中的小插入或缺失片段。其它新增的hg18变异信息还包括DGV结构变异信息（DGV Structural variants）、Segmental Dups、Exapted Repeats、Interrupted Repeats等。

根据NIH的决定，UCSC移除了威廉信托基金会病例控制协会（Wellcome Trust Case Control Consortium）的研究数据以及NIMH对躁狂抑郁双向障碍的研究数据，因为这些数据有可能会在某些特性情况下泄露患者的隐私，这与研究的本意是相悖的。他们还将继续与其它国际研究机构合作，研究如何才能更好的保护那些参与人类基因组相关研究（genome-wide association studies, GWAS）的志愿人员的隐私。他们将来会在保护个人隐私的前提下提供更多的GWAS数据，也会提供更新的图像化工具来浏览、分析临床试验数据。

5 Transmap

有一组新的数据GBD称之为Transmap，并把Transmap放在了脊椎动物基因组集合下。Transmap是在不同物种间使用同线性过滤配对基因组比对的方法（synteny-filtered pairwise genome alignment）来发现这些物种间最有可能的直系同源基因及其相应的注释信息。TransMap中每一条信息都包括mRNA、RefSeq、UCSC Genes等各方面的证据。例如，就人类mRNA证据来说，TransMap使用了23种脊椎动物的超过400,000条mRNA与库中的人类mRNA进行了高严谨性的BLAT比对。然后又使用Blastz这种在区分不同物种的能力上比BLAT更强的方法再次进行了比对。与translated BLAT法（Non-Human RefSeq Gene,图2）相比较，TransMap发现的旁系同源基因更少，而比较的UTR碱基更多（图2）。

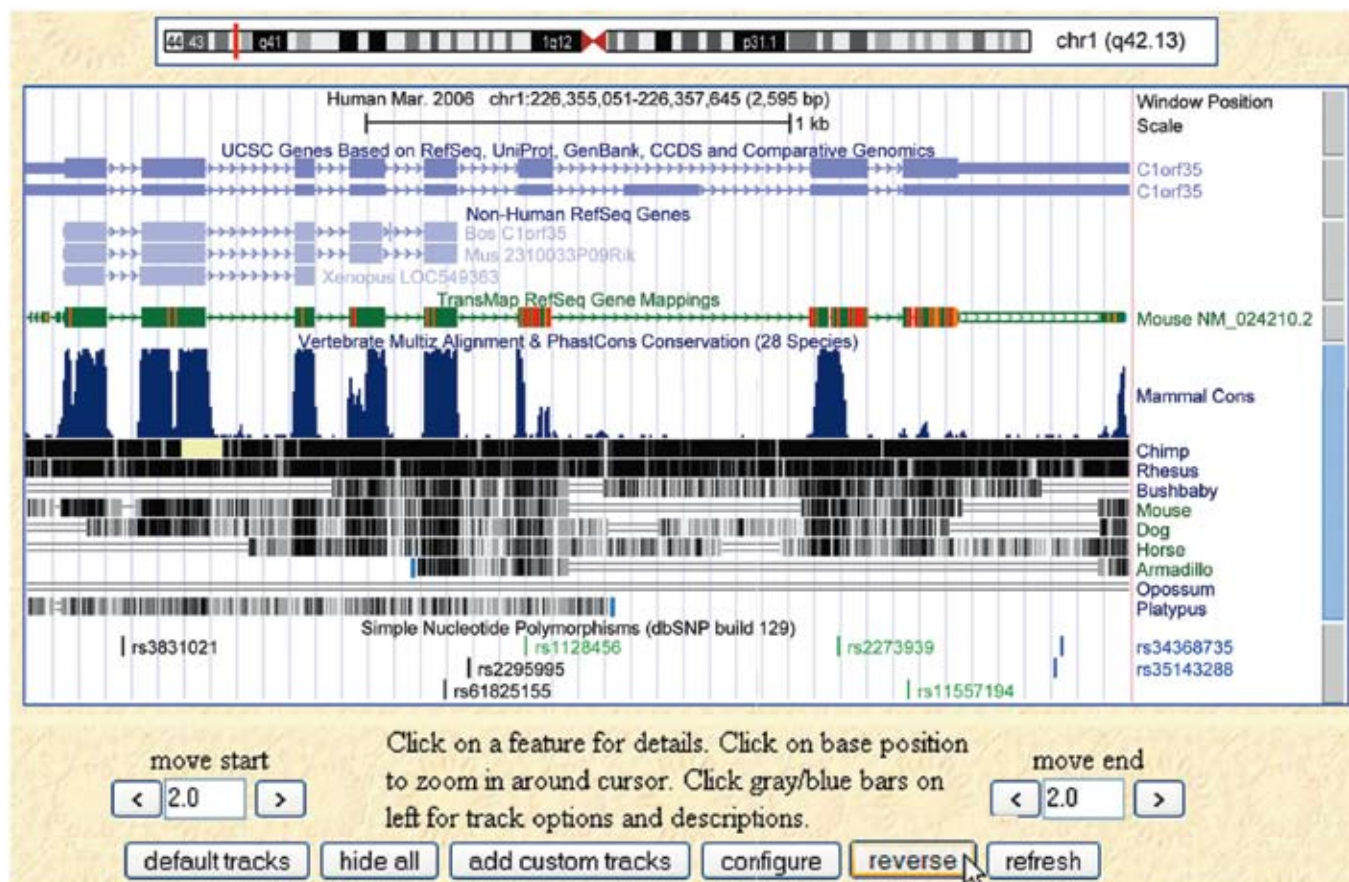


图2 搜索hg18人类基因数据库1号染色体的结果。显示一些最新的改进。从网页顶部到底部分别是带状示意图（Scale bar）、UCSC基因记录、Non-human RefSeq Genes、TransMap RefSeq Genes（图中红色显示的是小鼠RefSeq 3'端与Non-human RefSeq不相符的碱基）、Conservation track和SNP 129 track。整幅图都是使用网页底部箭头所指的反转按钮反转过的，这样就能以5'-3'方向浏览Transmap结果了。

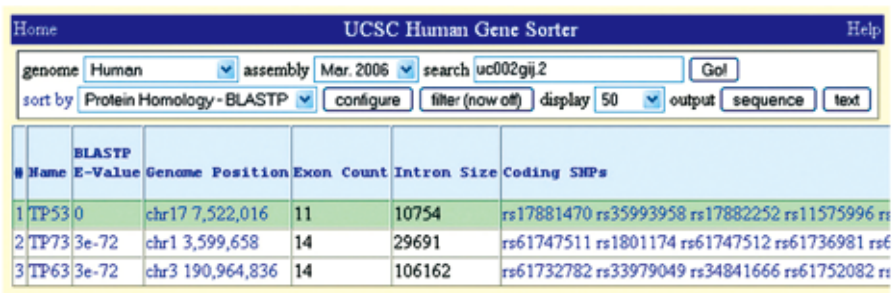
为了研究密切相关的低覆盖率基因组数据，使用的比对方法采用了相互最佳关系，以此来提高线性预测的准确性。就像GBD中所有的注释信息一样，数据库的结构信息也能从相应的Genome Browser页面中了解到。

6 新的基因分类 (Gene Sorter)

用户可以使用Gene Sorter借助不同的标准，例如基因表达谱或蛋白同源性以及其它的一些用户自行设定的条件对基因进行分类。通过网页上的链接可以很方便地切换到Genome Browser页面和其它的UCSC网站页面，了解更多的表达谱信息、蛋白间相互作用信息和其它的相关信息。去年，GBD还在Gene Sorter中添加了几个新的内容，现在Gene Sorter可以对人类、小鼠、大鼠、秀丽隐杆线虫、黑腹果蝇和酵母等六

种模式生物的基因进行分类。

内含子大小栏目 (Intron Size column) 能显示每一个基因内最大的或最小的内含子；编码SNP栏目 (Coding SNP column) 能快速地了解外显子多态性相关信息；CDS值 (CDS Score) 表示该基因是编码基因的可能性大小；基因分类 (Gene Category) 能对基因按编码、非编码、反义等情况进行分类；外显子计数 (Exon Count) 记录外显子数目 (见图3)。



The screenshot shows the UCSC Human Gene Sorter interface. At the top, there are dropdown menus for 'genome' (set to Human) and 'assembly' (set to Mar. 2006), a search bar with 'uc002gij.2', and buttons for 'Go!', 'sort by' (set to Protein Homology - BLASTP), 'configure', 'filter (now off)', 'display' (set to 50), and 'output' (set to sequence and text). Below the search bar is a table with the following columns: #, Name, BLASTP E-Value, Genome Position, Exon Count, Intron Size, and Coding SNPs. The first row is highlighted in green and shows results for TP53.

#	Name	BLASTP E-Value	Genome Position	Exon Count	Intron Size	Coding SNPs
1	TP53	0	chr17 7,522,016	11	10754	rs17881470 rs35993958 rs17882252 rs11575996 rs
2	TP73	3e-72	chr1 3,599,658	14	29691	rs61747511 rs1801174 rs61747512 rs61736981 rs6
3	TP63	3e-72	chr3 190,964,836	14	106162	rs61732782 rs33979049 rs34841666 rs61752082 rs

图3 Gene Sorter网页显示结果中新增TP53基因 (表中第一行以绿色显示的部分)，和其它所有符合标准的基因。图中使用的是Protein Homology-BLAST比对方，表中从左至右分别是BLASTP E值 (BLASTP E-value)、基因组位置 (Genome Position)、外显子个数 (Exon Count)、内含子大小 (Intron Size, 最大内含子大小) 和编码SNP信息 (Coding SNPs)。

7 新的显示特点

GBD在去年为Genome Browser新增了几项显示特点，其中很多都是按照用户的反馈在实用性方面做出的一些改进。碱基位置跟踪 (Base Position track) 现在可以提供一个选项，即在网页上自动显示缩放条配置。在网页下方的反转按钮 (Reverse button) 能让用户方便的将整条序列反向显示，这样就能从5'-3'方向来显示用户的目的基因了 (图2)。现在，直接在定位框 (Position box) 中输入“坐标值 (coordinate)”来定位到一个核苷酸也是允许的了，例如输入chr1:226356466就能定位到SNP rs 1128456位点 (也可以在框中直接输入SNP的rs号)。

此外，在网页中还做了几项改进：例如，Track group增加了下拉菜单，这样用户就可以隐藏不需要或不感兴趣的选项了。Tracks也可以轻易的从一个组 (group) 移动到另一个组，包括

移动到网页顶部的Custom Tracks group，这样用户就能在一个地方选择所有感兴趣的选项了，这项功能是大多数网站都不能提供的。现在每一个选项的标签上都有一个刷新按钮，这样用户就不需要频繁的滚动页面修改选项了。网站上还有许多“隐藏的”改进之处。为了减少网页中用于控制图像的“路径 (track)”数，加快网页的刷新速度，网站将相关的“路径”都组合到一起形成“超级路径 (super-track)”来共享配置选项。

用户进入多重比对网页可以获得genome browser网页中没有的低覆盖率物种的基因组DNA序列。使用UCSC在线验证PCR (UCSC In-silico PCR) 功能还能让用户在Genome Browser网页上直观的看到PCR扩增片段与其它序列 (例如最有用的外显子或内含子) 的关系。如果PCR引物与参考序列不相符，网页上会以红色高亮的方式标示出它们之间的差别。

8 用户自定义选项功能的增强

UCSC Genome Browser提供的用户自定义选项功能允许用户按照自己设定的条件只浏览感兴趣的数据结果。

新增了两种用户自定义数据时提交的数据格式。

用户自定义数据的内在表示法现在是基于专用机器的数据库表，而不再是以前所使用的基于文件系统的执行命令。

最近几年，Genome Browser中最受广大用户欢迎的工具就是“作业保存（session-saving）”功能，用户使用该功能能够保存并共享多个浏览器配置文件以供日后再次使用。与“作业保存”功能相关的用户选项功能也得到了增强。

因为有许多网页设置，还有几乎无数的设置与显示选项组合配置，所以Genome Browser使用了cookies来保存这些设置。

9 未来的发展方向

UCSC将持续收录新的脊椎动物和非脊椎动物模式生物的基因组数据。他们会与NCBI和Ensembl保持密切合作，为获得及处理新数据制定出标准化的流程，以保证所有机构的数据都是一致的。

他们希望能为2×覆盖率的基因组计划（2×coverage species project）提供包含44种物种基因组的多重比对，还希望能为最新的人类基因组数据提供扩展的多重比对。千人基因组计划（1000 Genomes project）获得的数据也将收录进变异信息库，还将收录重组热点的高分辨率图谱（high-resolution maps of recombination hotspot）。

此外，GBD还计划对网站进行进一步的升级。例如扩展网站工具的实用性和可配置性，升级isPCR功能，能让用户查询“RNA空隙（RNA space）”来比对被内含子隔开的序列，还将继续升级用户自定义系统使其支持混合格式数据。

在接下来的几年中，我们计划能让用户在Genome Browser上直接输入信息，就像wiki做的那样，这样各种“基因专家”们就能在网站上发表评论、数据和其它各种信息。最后要说的是，UCSC一直在与全球传染性疾病预防组织（Global Solutions for Infectious Diseases）合作开发限制访问的医药数据系统，例如HIV基因组数据和临床数据。我们还与几个研究机构合作建立了癌症基因组网站（cancer genomics browser）。如果网站的访问和保密问题解决了，我们将很快推出该网站。

原文检索：Nucleic Acids Research, 2009, Vol. 37, Database issue D755–D761

 YORK/编译

五 人类代谢和信号转导途径数据库

人类生物学反应及信号通路数据库Reactome

反应组学（Reactome, <http://www.reactome.org>）是一个汇集了由专家撰写，经同行评阅的有关人体内各项反应及生物学路径的文章的数据库，该数据库相当于一个有效的数据资源以及电子图书。该库目前发布了共计2975个人类蛋白、2907项生物学反应以及4455个引用文献。该数据库为人们提供了一个全新的从整体