人类基因组用户指南

翻译者：吴健、曾爱华、陈昌杰、罗宝正、
张志、陈辉、黄力、王旭

生物软件网（http://www.bio-soft.net）
编辑整理

人类基因组计划将于 2003 年完成，人类基因组数据库成为人类的巨大财富。它对所有公众开放，每个人都有权免费使用这些强大的资源，从而成为生物医学研究者必不可少的工具。但是，面对日益增长的浩瀚的数据海洋，怎样有效地利用它而不至于迷失其中，是一个严峻的问题。据wellcome Trust去年的一项调查，使用序列数据库的研究人员中，只有一半的人能够完全熟悉基因组数据库提供的服务。针对这种情况，2002 年 9 月份，Nature genetics特别出了一本“人类基因组用户指南”，以提问的形式详细讲解了人类基因组数据库的结构和使用方法，带领我们一步步深入其中，获取有用的信息。它是我们开启人类基因组数据宝库的一把金钥匙。读者也可以上Nature杂志网站（http://www.nature.com ）看原文 http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v32/n1s/index.html，这本用户指南的电子版是免费的。

问题 1：如何找到一个感兴趣的基因并确定其结构？一旦基因在图谱上被定位，又如何方便地检测到同一区域的其它基因？

可借此问题介绍 3 个主要的基因组浏览器。将利用所有 3 个站点对基因 ADAM2 进行检测，使读者能对每个站点提供的信息之间的细微的区别有一个正确的认识。

1. 国立生物技术信息中心(NCBI)图谱浏览器(Map Viewer)

可以通过NCBI主页进入NCBI 的人类图谱浏览器，网址为 http://www.ncbi.nlm.nih.gov/ 。点击右栏标有“Human map viewer”的超级链接即可进入图谱浏览器的主页。页面上端的符号标明此为Build 29，或NCBI人类基因组的第 29 次数据装配。Build 29 是以 2002 年 4 月 5 日的序列数据为基础而建立的。在它之前的基因组装配称为Build 28，以 2001 年 12 月 24 日的序列数据为基础而建立。想要寻找图谱上的任何信息，比如基因符号、基因库的登录号、标记物名称或疾病名称，只需在“Search for” 窗口输入相应的术语名，然后点击

“Find”即可。例如，输入“ADAM2”然后点“Find”。而染色体栏“on chromosome(s)” 的窗口会空出以进行基于文本的查找。

结果，浏览器的页面显示了所有人类染色体的示意图，并用指针指出ADAM2 在第 8 号染色体短臂上的位置。搜寻结果表明基因存在于两种 NCBI 图谱上，Genes_cyto 和 Genes_seq。Genes_cyto 指细胞遗传学图谱，而 Genes_seq 指序列图谱，点击任易一种链接将打开相应的图谱。

这方面及其它 NCBI 图谱的详细介绍可通过 http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/humansearch.html .进行查找。若需要了解关于ADAM2 更多的情况包括所有可利用的图谱，点击“Map element” 内相应的选项（本例为ADAM2），将会显示ADAM2 及少数 8p11.2 上的相邻序列。三种图谱都将在本视图显示并将在下面进行详细说明，其它例子所用的图谱可通过Maps & Options 附加到本视图。

最右边的图谱为主要图谱，此图谱提供了最详细的资料。本例中的主要图谱即为 Genes_seq（基因序列）图谱，描述了 ADAM2 的内含子/外显子组成，是通过 ADAM2 mRNA 在基因组上的序列对齐比较（alignment）而建立的。此基因有 14 个外显子。在 ADAM2 基因符号旁的箭头（粉红色区域内）显示了基因转录的方向。基因符号本身与 LocusLink 相链接，这是一类 NCBI 资源，可提供有关此基因的大量信息，包括别名、核苷酸及蛋白质序列，并与其它资源相链接（见问题 10）。基因符号右侧的链接指向了有关此基因的附加信息。

sv,或称序列浏览，表明基因在基因组克隆重叠群（contig）上的位置，包括核酸和编码的蛋白质序列。

ev 给使用者提供证据浏览，显示了支持某特定基因模型的生物学证据。这个视图显示所有的标准序列模型（RefSeq）、基因库 mRNAs（ GenBank mRNAs）、转录子（无论注解的、已知的或潜在的）及与基因组 contig 进行序列对齐比较的

表达序列标签（ESTs）。证据浏览更多的信息可通过点击任意证据浏览页上的 Evidence Viewer Help 链接进入 NCBI 网页查询。

hm 为 NCBI 的人-小鼠同源图谱的链接，显示人类和小鼠之间同源的基因组序列。

seq 允许使用者以文本格式重新获取某一区域的基因组序列,序列显示的区域可很容易地进行替换。

mm 为 Model Maker 的链接，显示当 GenBank mRNAs、ESTs 及基因预测与基因组序列对齐比较时的外显子。随后使用者即可选择特定的外显子创建一个用户化的基因模式。有关 Model Maker 的更多的信息可通过点击任一 mm 页上的“help”栏进入 NCBI 主页获得。

UniG_Hs 图谱显示已经与基因组进行序列对齐比较的人类 UniGene 簇。灰色的柱状图描述了比对的 ESTs 的数目，而蓝色线条显示了 UniGene 簇在基因组中的定位。深蓝色线是进行序列对齐比较的区域（即外显子），浅蓝色划线则表示潜在的内含子。在此例中 UniGene 簇 Hs.177959 在基因组中的定位跟随着 ADAM2 和所有的外显子。

Genes_cyto 图谱显示了基因在细胞遗传学图谱中的位置，橙色条带显示基因位置。尽管 ADAM2 已被很好地定位，并以一条短线表现出来，其它的基因比如它后面一条长线上成组的基因也被按照细胞遗传学定位于第 8 号染色体上较宽的区域。

点击蓝色工具条上的缩放控制区可进行缩小，利于使用者观察第 8 号染色体较大的区域。缩小一个水平可显示 1/100 的染色体区域，在此区域共有 20 条基因，20 条基因均可被显示。ADAM2 基因在所有图谱上的区域均以红色突出。在 Genes_seq 图谱上 ADAM2 定位于 ADAM18 及 LOC206849 之间。

2. UCSC(University of California,Santa Cruz)基因组浏览器

UCSC基因组浏览器的主页为http://genome.ucsc.edu/ 。目前，UCSC不仅提供最新版的小鼠和人类基因组数据，同时也提供许多较早的汇编。使用基因组浏览器时，先在窗口上方蓝色工具条的下拉式菜单中选择相应的生物体（本例为Human），然后点击标有Browser的链接。在结果页，选择相应的人类数据汇编版本进行阅读。2001 年 8 月的基因组浏览器建立于UCSC使用在当时所能获得的序列数据建立的人类基因组汇编。2001 年 12 月的浏览器显示了对NCBI 的人类基因组build 28 的注解。而 2002 年 4 月的浏览器显示了对NCBI的build 29 的注解。因为最近的这个人类资料汇编的注解不及 2001 年 12 月的汇编全面，所以本文所列举的例子来自较早的汇编。在下拉式菜单中选择“Dec. 2001 ”从数据库获得汇编资料。

查询所支持的类型列于文本输入框下面。在标有“position ”处输入“ADAM2”然后点击“Submit”项。查找的结果以两种类别显示，分别为“Known Genes ”和 “mRNA Associated Search Results”。标有“Known Genes ”的部分显示了将 NCBI 的参考 mRNA 序列定位到基因组中。“mRNA Associated Search Results”则代表了 GenBank 的其它 mRNA 序列定位到基因组中。点击“Known Genes ”与 ADAM2 的链接可见 ADAM2 mRNA 参考序列在基因组的状况（NM_001464）。

放大视图显示第 8 号染色体基因组序列从 36234934 到 36280132 碱基的区域，位于 8p12。标记为 Known Genes (来自 RefSeq)的蓝色路径显示已知基因的内含子和外显子结构。垂直框表示外显子而水平线则为内含子。ADAM2 基因似乎具有 14 个外显子，转录的方向由内含子上的箭头示意。标记有 Acembly Gene Predictions, Ensembl Gene Predictions 和 Fgenesh++ Gene Predictions 的路径为基因预测的结果（见问题 7）。其它数据库核酸序列的对齐比较显示在 GenBank 的 Human mRNAs、spliced EST、 UniGene 和来自于 GenBank 路径中的 Nonhuman mRNAs。小鼠和 Tetraodon 基因序列翻译后的序列对齐比较在小鼠和鱼 BLAT

路径内。显示单核苷酸多态性(SNPs)、重复元件及微阵排列数据的路径列于页面底部。关于每个路径附加的细节可通过选择位于底部的 Track Controls 中的路径名获得。

查看 ADAM2 前后基因序列，点击位于右上角的“zoom out”框进行缩小，ADAM2 位于 TEM5 和 ADAM18 之间。

3. Ensembl 网站

Ensembl项目网站（http://www.ensembl.org/ ）为四个物种：人类、小鼠、斑马鱼（zebrafish）和蚊子提供基因组浏览器。点击“Human ”以查看人类基因组的主要条目。目前人类Ensembl的版本为 6.28.1，是以NCBI基因组Build 28 为基础而建立的。欲进行搜索可在文本框中输入“ADAM2”并通过在下拉式菜单中选择“Gene ”以限定搜索范围，点击上方标有“Lookup” 的按钮，点击与ADAM2基因的链接可返回单独的结果。

点击与 ADAM2 的链接可重新回到 GeneView 窗口，此页包含四个部分的数据，第一部份为 ADAM2 的概貌，包括基因登录号，蛋白质结构域和家族的相关链接。链接 Ensembl 查看高度同源的小鼠序列可在“ Homology Matches ”部分获得，以后的例子会在这方面作出更详细的介绍。GeneView 窗的第二部份，提供有关基因转录子的信息，cDNA 序列被列出，其内含子和外显子结构以图表表示，同时在此基因前后位置附近有限数量的基因也以图表形式表示出来。外显子序列在 GeneView 中的第三部份显示，剪接位点显示于第四部份。如果预计基因具有不止一个转录子，则每个转录子拥有各自的转录产物、外显子和剪接位点部分。

ADAM2 完整的前后基因组序列内容可通过返回 GeneView 的第一部份和点击“Genomic Location ”框中的链接来查看。所出现的 ContigView 框的顶端部分描述了染色体，其中最为关键的部分以红色标示。此浏览显示了此基因的基因

组前后序列，包括染色体条带、contigs、标志和在图上靠近 8p12 的基因。点击任意这些项目可显示相关内容，感兴趣的部分在 DNA 图谱上以红色标记。由 Ensembl 注释的 ADAM2 附近的基因为 Q96KB2 和 ADAM18。

ContigView 页的底部即 Detailed View，是一个放大了的区域，标示出已经定位于此区域的人类基因组所有特征。Overview 和 Detailed View 之间的浏览器按钮将视图从左至右移动以及放大和缩小。所显示的内容可通过选择“Features ”的下拉式菜单进行移动以选取需要查看的内容。

所显示的内容为默认值，DNA(contigs)图谱将正链(上方)上的条目从反链(下方)分开，此处反链的唯一特征为 GENSCAN 基因预测程序提出（见问题 7）的单一的 Genscan 转录子。正链表现出了 5 种特征。从底部开始，ADAM2 转录子显示为红色，提示其为一个已知的转录子，对应于接近全长的 cDNA 序列、蛋白质序列或在公共数据库中两者均可得到的转录子。黑色转录子通过 EST 或蛋白质序列的类似性预测。“EST Transcr ”链接于独立的 ESTs 序列对齐比较，而靠近顶端的 UniGene 路径显示了 UniGene 簇。正链上的 Genscan 模式包含了在已知的转录子中发现的外显子。“Proteins and Human proteins ”框指出与本版本的基因组进行序列对齐比较的蛋白质序列。而“NCBI Transcr ”链接于 NCBI Map Viewer。将计算机鼠标放置于任一特征位置则可显示此特征名称，并可链接到更为详细的信息。

NCBI、UCSC 及 Ensembl 有时对同一基因使用不同的符号，所以通过不同的浏览器获得的信息难以进行比较，此外，这 3 个站点保留了独立的注解途径，并且都未尝试将相同的 mRNA 序列排列到基因组中。NCBI 目前显示 build 29, Ensembl 显示 build28，而 UCSC 则提供 build 28（2001.12.）和 build 29（2002.04.）。尽管在本指南中所有 UCSC 的例子都将推荐使用注解较好的 build 28。因为两种汇编数据之间存在的差异，在 NCBI、UCSC 及 Ensembl 中显示的数据就存在极小的差别，但在这 3 个站点中自由地穿梭仍然是很容易的。例如 NCBI 可通过 LocusLink 人类基因入口上方的黑色框链接 UCSC 和 Ensembl，而 Ensembl 指导

NCBI 和 UCSC 使用者通过“Jump to”链接于它的“ContigView”。UCSC 基因组浏览器的一些版本有与 Ensembl 和 NCBI 的 Map Viewer 的链接，链接点位于浏览页顶部的蓝框内。

（吴健 译）

问题 2：如何在 DNA 序列中找到序列标签位点（ESTs）？

NCBI的“electronic PCR（e-PCR）”工具是UniSTS资源库的一部分，可以用来寻找一段目的DNA片段中的STS标记物。UniSTS (http://www.ncbi.nih.gov/genome/sts/) 能提供所有有关STS标记物的资料，包括引物序列、产物大小、作图信息和别名。与之相链接的其他NCBI资源如Entrez、LocusLink 和MapViewer 也同样提供这些信息。e-PCR通过搜寻具有正确的方向和间距的序列且这个序列能代表用于扩增STSs的PCR引物，来寻找一段DNA序列中潜在的STSs。

先在NCBI主页上（http://www.ncbi.nlm.nih.gov/ ）找到e-PCR的主页，然后在右手栏点击“Electronic PCR”链接。再在e-PCR主页的上端大的文本框内粘贴上目的基因序列或键入登陆号（accession number）。例如某个序列的登录号是AF288398，结果显示该序列只包含一个STS： stSG47693（或RH92759），位于此序列的 2102 和 2232 核苷之间。

当点击“Marker”下标记物的名称时，从 UniSTS 中出现 STS 的详细资料。引物的信息、PCR 产物大小以及标记物的替代名称也出现在主页的上端。在不同的图谱中，STSs 常有不同的名称。在“Cross-references”栏目下的 LocusLink、UniGene 和 the Genebridge 4 中，将显示这个 STS 的定位图。在“mapping information”部分包含能链接到 NCBI 的“MapViewer”浏览器。在主页的下端是“Electronic PCR results”，显示了其他序列，包括 contigs（重叠群）、mRNAs 和包含这个 STS 标记物的 ESTs。

为了在所有图谱中看到 STS 标记物及其基因组的状况，则在“Mapping Information”部分的上端点击链接标志 “MapViewer” ，这个图谱浏览器会出现两张图谱。请注意，在这个视窗里，STS stSG47693 被称为 RH92759（用粉红色强调）。99－Genebridge 4 (GM99_GB4，位于左边)基因图谱上有 46000 个 STS 标记被国际放射杂交协会定位到 GB4 杂交面板上。STS 图谱（位于右边）显示了如何使用 e-PCR 将 STSs 序列放置到基因组序列组装。灰色线将两个图谱的标记物连接起来，而红色线条显示 STS RH92759 在两张图谱中的位置。在这个区域，STS 图谱中共有 211 个 STSs，但在这个视窗里只标记了 20 个。在 STS 图谱的右边，点击绿色和黄色圆圈会出现 STS 标记物的图谱。通过左边工具条的缩放工具，可以放大或缩小这个视窗。

（曾爱华 译）

问题 3：定位克隆计划是为了寻找人类疾病基因，已有的连锁分析资料显示目的基因位于两个序列标签位点之间，如何识别该区域已知的或预测的侯选基因？哪些 BAC 克隆含有这些特殊区域？

开始这项研究首先必须浏览 UCSC Genome Browse 网页（http://genome.ucsc.edu/ ）。然后在该网页边缘蓝色下拉菜单从Organism中选择Human这个词。点击Browser，在the Human Genome Browser Gateway网页上，改变assembly成Dec. 2001。要搜寻哪两个序列标签之间的基因，就在search box中输入这两个序列标签，用分号分开。例如，搜寻序列标签D10S1676 和D10S1675之间的基因，在the search box中输入D10S1676；D10S1675，然后点击Submit。因为这些标记定位在基因组中专一的位置，所以这些标记之间的基因很快会出现。

STS Marker 路径（track）上蓝色的道表示遗传图谱标记，黑色的道表示放射杂交图谱标记。点击 STS Markers，就会展开这个路径，列出每一个独立标记。目的标记 D10S1676 和 D10S1675 在这里使用它们的替代名称(分别为

AFMA232YH9 和 AFMA230VA9)，并分别位于这个区间的顶部和底部。

在 Known Genes 路径内显示和列出所有已知的基因名单。这些编码蛋白质的基因来源于 NCBI 汇编的 RefSeq mRNA 序列并使用 BLAT 程序与基因组装配进行系列对齐比较。在该网页搜寻基因名单或其它特征可点击顶端的蓝色条上的 Tables l 链接。关于特殊基因比如（MGMT）的更多的信息，点击这个基因的符号就会得到一系列额外的链接，如在线人类孟德尔遗传规律，PubMed、GeneCards 和小鼠基因组信息 (MGI)。

许多路径包括 Acembly Genes、Ensembl Genes 和 Fgenesh++ Genes 可以显示预测的基因（参见问题 7）。如果想看上述任何种类的全部特征，点击屏幕左边该路径的标题。欲观察这些路径的简要描述以及其它没有提及的特征，点击该路径左边灰色的方框或向下滚动到 Track Controls，再点击你所感兴趣的标题。基因预测程序将在问题 7 中说明。通过点击 reset all 按钮使浏览器默认选择。

想要观察用于测序的 BAC 克隆，回到 Genome browser 页面，点击屏幕左边的 Coverage 展开该路径。在这里分别列出了各个 BAC 克隆，完成的区域用黑色表示，草图区域以不同形状的灰色阴影表示。想要获得更详细的信息如大小和特异克隆覆盖的序列则点击克隆号如 AL355529.21。在这个网页点击该克隆的登录号链接到 NCBI Entrez ，有关于这个克隆的摘要说明。在 Entrez 文挡摘要网页点击 AL355529 可以观察到全部 GenBank 的条目。

根据NCBI的命名协定，该克隆来自RP11 文库，并已经被命名为 85C15。RP11 是NCBI为RPCI-11 指定的名称，由Roswell Park Cancer Institute 制备，是常用的人类BAC文库。有关基因组序列文库命名协定的更多的信息可以在NCBI的Clone Registry查阅http://www.ncbi.nlm.nih.gov/genome/clone/nomenclature.shtml 。还可以在http://www.ncbi.nlm.nih.gov/genome/clone/ordering.html 网页上获得订购克隆的信息。

NCBI 网站

只要两个标记位于主图谱上，就可以在 NCBI MapViewer 上直接观察两个标记之间的区域。例如，主图谱是细胞遗传图，可以搜寻 22 号染色体上 22q12.1 和 22q13.2 之间的区域；如果主图谱是 Gene_Seq，可以找到两个基因之间的区域。

打开http://www.ncbi.nlm.nih.gov/ 网页，点击网页右边的Human map viewer，可以进入the Map Viewer网页。若要观察同一个染色体上多个位点，在search box 中输入的搜寻条件应该用“OR”分开。例如看两个序列标签D10S1676 和 D10S1675 之间的区域，在search box 中输入D10S1676 OR D10S1675 ，然后单击FIND。搜寻结果页面顶端显示染色体图上有两个红色的记号，表明这两个标记在 10 号染色体是紧密靠近的。在搜寻结果网页底部，显示两个标记的别名(AFMA232YH9 和 AFMA230VA9)以及在图谱上的位置。想要同时观察两个标记，在染色体图表中点击chromosome 10 ，显示D10S1676 和 D10S1675 周围区域，用粉红色突出原来的搜寻。红线将两个标记在不同图谱中的位置连接起来。

Maps & Options 链接位于该网页顶端的水平蓝色区，该链接可以让用户按照自己的要求制定显示的图谱和区域。例如，观察该区域已知的和预测的基因，还有作为测序来源的 BAC 克隆。打开 Maps & Options 窗口，首先在 Maps Displayed 框中删除除了 Gene 和 STS 外的其它所有图谱。方法是用鼠标加亮选中的图谱并选择 remove。然后在 Available Maps 框中选择并添加 Transcript (RNA)、GenomeScan、Component 和 Contig 图，再选择“ADD”。

用鼠标加亮 STS 图使它成为支配的图谱，然后选择 Make Master/Move to Bottom。在 Region Shown 框中输入这两个标记名称，就可以使图中只显示 D10S1676 和 D10S1675 之间的 STSs。点击 Apply 可看到排列图，在某种情况下，选择的网页大小比默认值大 20 可以在窗口中浏览到更多的信息。

在 Maps & Options 窗口显示的图谱很详细。STS 右边的绿点显示了遗传标

记在所有图谱的位置。这是 10 号染色体上相当长的区域，并不是每一个 STS 标记都列出来，尽管在该区域有 611 个 STSs，但该页只显示 20 个。对每一个已知基因，基因序列图谱（Genes_Seq map）显示所有已经被绘制到基因组中的外显子。除非基因有不同的剪切形式，对于每个已知 mRNAs 的基因，其外显子也在 RNA 图（转录图）上显示，在 Genes_Seq 和 RNA 图谱上将是一样的。GScan (GenomeScan)图显示 NCBI 的基因预测，所有这些已知或预测的基因都是疾病侯选基因。

NCBI 组装的重叠群（contigs）也叫作NT contigs，可以在Contig图谱中寻找。蓝色的片段来自已完成的序列，橙色来源于草图。这些contigs通过独特的、在构成图[Comp(Component) map]中显示的GenBank序列条目构建而成。草图HTG 记录（1 期和 2 期，见http://www.ncbi.nlm.nih.gov/HTGS/ ）表现橙色而完成的HTG为蓝色。大部分GenBank序列来源于BAC克隆。装配成contigs 的BAC克隆清晰可见。只要点击登录号与Entrez链接，你可以得到该条目更为详细的信息，包括克隆名。如果Comp 图是支配图谱，那么克隆名可以直接在MapViewer看到。点击图谱名称附近的蓝色箭头可很快生成主图谱。

因为是染色体放大图，所以单个基因和 GenBank 条目很难看到。利用蓝色工具条控制可提供某区域更多的细节。另外，点击左边工具条 Data As Table View 可找到全部的资料，包括隐藏在这个窗口中的一个基于文本的表格。

SIDEBAR 网站

你也可以应用Ensembl 的MapView搜寻两个STS标记之间的区域。打开 Ensembl Human Genome Browser （http://www.ensembl.org/Homo_sapiens/ ），点击任一染色体组型进入MapView，在Jump to Contigview中键入遗传标记名称。如想利用Ensembl得到指定的染色体区域的基因目录（或其它注释），在ContigView窗口点击Export GeneList 。

（陈昌杰 译）

问题 4：使用者希望找到两个序列标签位点（STSs）之间所有单核苷酸的多态性。任何单核苷酸多态性都处于基因的编码区域吗？在哪里可以找到有关这些基因的其它功能的信息？

搜寻从 NCBI 单核苷酸多态性数据库（dbSNP）的网址（http://www.ncbi.nlm.nib.gov/SNP ）开始进行。在这一页面上有一系列的连接可供使用，用户可以用数据库自身的信息，也可以使用关于基因或基因座的信息进行搜索。

对于这项搜索，假定所关心的区域是已知的而且限定在两个 STS 标记 RH70674 和 G32133 之间。滚动到页面底部标有“Between Markers”的部分。在两个文本框中键入 STS 标记物的名称“RH70674”和“G32133”，然后点击“Submit STS Markers”。这将会显示所关心区域内总共 81 个 SNP 中的 1～25 个。在页码框中键入“3”然后点击“Display”进入第 3 页。

搜寻结果显示的页面说明了在典型的 dbSNP 页面上所能找到的大多数页面类型。在该表格中，从左边开始，第一栏给出了各个 dbSNP 簇的标识符（全部以“rs”开始）。第二栏，用 Map 标识，显示出某一特定的 SNP 是否已经被定位到基因组中的唯一位点（通过一个绿色箭头显示，就像第一行的例子）还是多位点（这里没有显示）。

之后的几栏， 标识为 Gene，指出这些 SNP 是否与一些详细的特征相关，例如基因、mRNA 或者编码区。这 3 栏（L、T 和 C）中每一行，或者以亮度显示或者以灰色显示，整齐排列。

如果 L（locus）显示蓝色，则标记物的一部分或者全部位置位于基因 5′ 端的 2kb 内或者在基因的 3′ 端 500bp 内。

如果 T（Transcript）显示绿色，部分或者所有标记物的位置与一个已知的 mRNA 重叠。然而这并不意味着 SNP 标记物一定落在编码区内。

如果 C（Codon）显示橙色，部分或者所有的标记物的位置与一个编码区重叠。

下一栏，标识为 Het，显示观察到的标记物的平均杂合度，范围是 0～100%。当读数是 0 时意味着该特异性标记物没有任何信息，然而粉红条带显示标记物的置信区间是 95%。Validation 栏显示该标记是否已经确认（用星号表示）或者尚未确认（用浅蓝色盒表示）。确认的标记已经通过独立的序列再分析来核实。所有尚未确认的标记以 3 个蓝色框来表示，根据顶部栏的刻度，意味着该标记得到确认的几率大于 95%。这个图形指出这个标记物是真的概率（成功率被定义为 1 减去假阳性率）。

在倒数第二栏，符号 TT 表示特定的基因型中存在这个标记。最后，Linkout Avail 栏表示哪一个标记被连接到了其它的数据库。这一栏中 P 表示这种变异已经被定位到一个已知的蛋白质结构。如果要完全描述所有特征，只要点击这一栏之上的标题即可。

回到原来的问题上，如橙色的 C 所显示，在这一页面中显示的其中一个 SNP 确实落在编码区。如果要得到有关任一特定 SNP 的更多信息，只要点击超级链接 SNP 簇的身份标识符即可。例如，点击 rs1059133，产生一个新的页面，显示出该 SNP 的所有信息。在标有“Submitter records for this RefSNP Cluster”的标题下面，是一张一个个 SNP 的列表（在本例中只有一个 SNP），是由单个 SNP 成簇集中在一起形成这种单一的参照 SNP 的。SNP 的序列在下一个标题中出现。在标有“NCBI Resource Links” 的标题下，是与这个 SNP 相关的 GenBank（基因库）和 NCBI RefSeq（参考序列条目）。进一步向下滚动到 SNP 页面的底部，在“LocusLink Analysis”部分显示了这个 SNP 所落在的编码区的基因（ADAM2，disintegrin 和金属蛋白酶结构域 2）。SNP 的等位基因是 G/C，一个导致组氨酸残

基替代天冬氨酸残基的非同义改变。这里也提供了其它的链接，如 NCBI Map Viewer、Ensembl map 和 UCSC 基因组装配（标有 Integrated Maps 的部分）。标有 Variation Summary and Validation Summary 的部分（没有显示）给出了这一特定 SNP 的原始资料。

要回答这一问题的最后部分需要从 dbSNP 转到 LocusLink 主页。要达到这个目的，需要点击该页面 LocusLink 标题下的 ADAM2。这将带领使用者到达 ADAM2 的 LocusLink 页面，并且在页面顶端提供大量到达 NCBI 和相关资源的点击点。通过位于页面左边的位置连接处的 FAQ 连接可以找到更多的信息。通过简单浏览 LocusLink，使用者可以看到 ADAM2 属于一个细胞膜锚锭蛋白质的家族，该家族的蛋白与受精、肌肉发育和神经发生等各种过程有关。

使用者经常忽视的信息来源是 OMIM。这是一个关于人类基因和遗传性疾病目录的电子版，由 Johns Hopkins 大学的 Victor McKusick 制作。OMIM 向使用者提供了来自已发表的大多数人类遗传性疾病文献的简洁原文信息以及遗传基础，并且可以链接到原文献上。OMIM 条目中包含的信息有基因符号、疾病的其它名称、疾病的说明（包括临床、生物化学和细胞遗传学的特征）以及遗传模式的详细资料（包括遗传图谱信息）和临床梗概的细节。这些条目是通过手工整理过的，以保证摘要是最新的并且是准确的。尽管 OMIM 可以直接进行搜索，然而许多 LocusLink 入口也连接到 OMIM 中记录的基因。ADAM2 蛋白的 OMIM 条目页面在图中显示。这个页面可以超级链接到 PubMed、GenBank 和其它的相关数据库。

（罗宝正 译）

问题 5：已知一段 mRNA 序列，怎样在人类基因组图谱中找到对应的 DNA 片段？一旦它的位置确定，如何找到选择性剪接位点位置？

举例说明如下。一个mRNA片段在基因库的登录号为BG334944。首先，登录http://www.ncbi.nlm.nih.gov/Entrez/ ，在NCBI 的 Entrez界面找到这个 EST的

核苷酸序列。在页面上部的对话框中键入登录号BG334944 ，下拉菜单中选择Nucleotide， 点击Go 。结果页面显示有关登录号BG334944 的条目。为了在FASTA 格式（一种生物学信息程序的常用格式）找到这个序列，在这个页面上把下拉菜单变成FASTA 后点击Text，产生一个包含FASTA格式的序列的新页面，然后将序列拷贝下来。

为了确定这段序列在基因组中的位置，使用UCSC的 BLAT 工具。登录 http://genome.ucsc.edu/ ，将你的网页浏览器指到UCSC基因组浏览器的主页开始搜索。在页面一侧的蓝色框里，从Organism 下拉菜单中选择Human ，然后点击Blat 。然后将从上面Entrez 得到的 FASTA格式的序列粘贴到BLAT 搜索页面的大的文本框上 。把Freeze 下拉菜单变成Dec. 2001，将Query Type下拉菜单变成DNA ，然后点击Submit 。服务器将很快找出搜索结果：唯一与之匹配的是一段长为 636bp的片段，位于 9 号染色体上，为正链。

为了得到更加详细的资料，在页面上条目的左边点击 details 链接，得到一个长的页面，界面包含三个部分：mRNA 序列(上部)，基因组序列(中部)以及和基因组序列相对应的 mRNA 序列对齐比较。在序列对齐比较（alignment）图中，和 cDNA 及基因组序列匹配的碱基是用暗绿色的大写字母标记的。缺口用稍低的黑体字标记。淡蓝色稍高的碱基标记的是缺口两边序列对齐比较区域的结合部分，常常是剪接位点。

返回 BLAT 摘要页面搜索，点击 browser。这将产生一个用图解说明特异性的 mRNA 序列在对应的基因组序列上的位置。标记 Chromosome Band（染色体带）的路径提示 mRNA 位于 9q34.11。询问序列本身出现在标记有 Your Sequence from BLAT Search 的直线上。页面上显示的序列是不连续的：相似的区域显示为垂直线，缺口显示为细的水平线，排列的方向由箭头的方向表示。被查询的 EST 的比对排列区域对应于已知基因的外显子立即显示在线条的下面(Known Genes, 在这里是 RAB9P40)。在 UCSC 的搜索框内键入 EST 的名称 BG334944，将会产生一个与上述点击 browser 相似的结果。这个例子的部分目的是阐述 BLAT 的用

途。

大约图谱向下到一半的位置是标记着 Human ESTs That Have Been Spliced 的路径（人类已经剪接的 ESTs）。因为所有的 ESTs 都浓缩在一条线上，这个路径最初显示比较密集，所有的 EST 密集排列在一条直线上。点击该路径标记，可以看到这一区域内与基因组比对排列的所有 EST，这些 EST 可能代表了具有不同剪接位点的转录物（抄本）。这将扩展这个图形的区域，所以每一个 EST 占据一条直线。ESTs 的长度是可变的，但是大部分包含已知基因的相同的外显子并且（大概）以同样的方式剪接。仔细地检查并与已知基因相比较，提示有一些 ESTs 缺失了一个或多个外显子。留心查看标记了 BE798864 和 W52533 的线条，前者缺失第 5 外显子，而后者则缺失第 4、5、6 外显子。

通过点击特定的线条可以考察任何 ESTs 的详细资料。比如，点击 BE798864 所在的线条，可以得到这个 EST 的详细资料页面。这个 EST 与基因组序列有 99.8%的同源性。在标记有 EST/Genomic Alignments 区域点击任何超链接线条都会返回到实际上的一个碱基挨一个碱基的排列。EST 的末端可以不同，但是在推测有外显子缺失附近区域的序列是相同的。

当mRNA改变其编码的野生型蛋白质序列的时候，这个mRNA很可能存在生物学意义上的的选择性剪接。为了确定EST BE798864 是否会编码不同于已知基因(RAB9P40)编码的蛋白质，我们可以用NCBI 的 BLAST 2 Sequences工具直接比较这两个序列。首先，打开一个新的浏览器窗口，因为上面的搜索资料在这儿也需要，当需要使用多个网页工具时，这样将避免过分使用浏览器的前进和后退键。然后从http://www.ncbi.nlm.nih.gov/BLAST 登录BLAST 主页。在Pairwise BLAST标题下选择BLAST 2 Sequences 。在这个页面上，用户可以仅仅输入登录号而不用输入剪切和粘贴的序列进入对话框。

对于 EST 来说，仅在标有 Enter accession or GI for Sequence 1 的对话框中输入 EST 的登录号(BE798864)。获得 RAB9P40 的登录号需要返回前面的图解，然

后点击基因路径。一旦这些都做好了，在标有 Enter accession or GI for Sequence 2 的对话框中输入基因的登录号(NM_005833)。确认 Program 下拉菜单设定在 blastn （比较两个核苷酸序列），然后点击页面底部的 Align 键就会得到所示的比对排列图。序列 1 (the EST)默认为查询序列，而序列 2（已知基因）则被默认为目标序列。 起始于第三行末端排列的已知基因翻译的蛋白序列也显示出来，检查这些排列发现这个 EST 缺失 153 个核苷酸（该 mRNA 第 360－512 核苷酸），对应于 BE798864 缺失的第 5 外显子。这个缺口在开放读码框架内，所以这个 EST 可以编码与已知基因具同源性但稍短的蛋白质。

由于 EST 序列测定的特点决定，ESTs 经常包含测序错配率远远高于已经完成的基因组序列甚而基因组草图序列的错配率。但令人鼓舞的是 EST BE798864 在基因组序列上排列完好，其编码的蛋白质可能与已知基因编码的蛋白质具有相同的结构。另外，从 UCSC 图解来看，这个区域的其他 ESTs 如 BE779110 也会引起 RAB9P40 的第 5 外显子缺失。但是，所有这些预测都必须通过上面讲的 EST－genomic 排列质量来检验。最后的选择性剪接的证据当然还必须在实验室中才能找到。

（张志 译）

问题 6：如何找到一个基因的序列，此序列除了含有所有已注释的外显子和内含子外，还有用于引物设计的一些碱基？

这项搜索从进入 UCSC 基因组浏览器主页开始，网址是 http://genome.ucsc.edu/ 。从标记着Organism的下拉菜单处选择Human, 然后单击 Browser。这样，使用者便进入了人类基因组浏览器通路，可在当前或更早的基因组装配版本中进行许多基于文本或位置的搜索。根据本例的情况，选择Dec. 2001 版本，在position框内键入感兴趣的基因的名称(PTPN1),然后点击Submit（提交）。浏览器将找出以字母‘PTPN1’开头的全部基因。以本例子来说，感兴趣的基因名称为PTPN1，点击PTPN1 的超链接可以观察到这个基因在基因组中的

前后关系。

在页面顶部的文本框内给出了这个基因的碱基对的绝对位置（在 20 号染色体上，位于 48929540－49003636 之间），并说明这个基因长 74 kb 。标记 Chromosome Bands 的路径显示 PTPN1 位于 20q13.13。最后，标记 Known Genes 的路径说明该基因处于正链上，因为路径上的箭头指向右方。这个基因的外显子在 Known Genes 路径中用垂直线表示。

获得一个基因上游序列的方法将在问题 7 中叙及。在这里我们解释一下如何得到一个基因两端的序列。为了得到足够的序列用于设计引物，可以在页面顶部 position 框内改变位置的数字来增加显示区域的长度。例如，为了在 5′ 端增加 1，000 个碱基，并在 3′ 端增加 200 个碱基，将位置（position）框中的内容变为 ‘chr20:4892854-49003836' 然后点击 Jump。这样就会以新的设定刷新屏幕。

要想得到这段区域内的序列，点击该网页顶部的蓝色条带中的 DNA 链接。这样会产生一个新的网页，标题为 Get DNA in Window。点击紧靠 “extended case/color options”的按纽，然后点击提交 Submit。经过这样的选择, 使用者通过改变文本的格式(格子, 下划线, 粗体, 斜体) 和\或颜色(红色, 绿色, 蓝色)，可以强调序列的特征。通过改变标有红、绿和蓝的框中的 0～255 的数字，可以使颜色改变成黑暗或几种颜色的混合色。表格下给出了怎样特异化 RGB（ 红-绿-蓝）3 色的例子。以本例子的情况，在 Known Genes（RefSeq Genes）这行选择 Toggle Case，将红色改成 255 以达到饱和而其他颜色设为零。一旦使用者点击了 Submit，就会产生一个新的网页，包括前面特别设定的序列长度 (chr20:48928540-49003836)，并且这段区域内的外显子用红色的大写字母标记。现在可以保存这个基因组序列，也可以输入引物设计或序列装配程序包，以便做进一步研究。

“extended case/color options”选择页还能用于基因组的路径之间的联合和比较。例如，返回 options 界面，保留前面已选择的 Known Genes 行，但现在也在

标有 Mouse Blat 的那一列选择下划线（Underline）。点击 Submit 产生一个新网页，人外显子仍然是红色大写字母，但和鼠类序列一样的部分现在用下划线标记。在此基因，鼠的保守序列与外显子相重叠。

（陈辉 译）

问题 7：怎样才能使研究者更容易地找到对所感兴趣的基因的结构进行描述的信息汇编？能否获得推定的启动子区的序列？

这项搜寻要从UCSC 基因组浏览器开始，网址为http://genome.ucsc.edu/ 。以编码pendrin (PDS)的基因为例来说明上述问题。PDS 与耳蜗的异常发育、感觉神经性听力下降以及弥散性甲状腺增大（甲状腺肿）有关。

进入 UCSC 的主页后，在 Organism 的下拉菜单中选择 Human，然后点击 Browser。使用者现在到了人类基因组浏览器入口。本例的搜寻很简单：在 assembly 的下拉菜单中选择 Dec. 2001，在 position 框中键入 pendrin，然后点击 Submit 。返回的页面结果显示一个已知的基因和两个 mRNA 序列。继续点击 mRNA 序列的登录号 AF030880 ，出现包含这个 mRNA 区域的图解概要。为了获得这个区域更清晰的图像，点击紧靠 zoom out 的 1.5X 按钮。最后点击页面中部的 reset all 按钮，使各个路径的设置恢复默认状态。

然而，对于本例的搜寻目的来说，默认设置不是理想的设置。按照视图利用页面底部的 Track Controls 按纽，将一些路径设置为 hide 模式（即不显示），其他设置为 dense 模式（所有资料密集在一条直线上）；另一些路径设置为 full 模式（每个特征有一个分开的线条，最多达 300）。在考虑这些路径内究竟存在那些资料之前，对这些路径的内容和表现做一个简要的讨论是必要的，许多这些讨论是由外界提供给 UCSC 的。下面是对基因预测方法的更进一步讨论，这些信息也可以在其他地方找到。

对于 Known Genes （已知基因）和预测的基因路径来说，一般的惯例是以一个高的垂直线或块状表示每个编码外显子，以短的垂直线或块状表示 5′ 端和 3′ 端非翻译区。

起连接作用的内含子以非常细的线条表示。翻译的方向由沿着细线的箭头指示。

Known Genes 来自 LocusLink 内的 mRNA 参照序列，已经利用 BLAT 程序将这些序列与基因组序列进行比对排列。

Acembly Gene Predictions With Alt-splicing 路径是利用Acembly程序将人类mRNA和EST序列数据与人类基因组序列进行比对排列而来的。Acembly程序试图找到mRNA与基因组序列的最好的比对排列以及判断选择性剪接模型。假如有多于 1 个的基因模型具有统计学意义，则它们都全部显示出来。有关Acembly 的更多信息可以在 NCBI 的网站找到（http://www.ncbi.nih.gov/IEB/Research/Acembly/ ）。

Ensembl Gene Predictions 路径由 Ensembl 提供。Ensembl 基因通过许多方法来预测，包括与已知 mRNA 和蛋白质进行同源性比较，ab initio 基因预测使用 GENSCAN 和基因预测 HMMs。

Fgenesh++ Gene Predictions 路径通过寻找基因的结构特征来预测基因内部的外显子，例如剪接位点的给位和受位的结构特征，利用一种动态的程序算法推定编码区域和推定外显子 5′ 端和 3′ 端的内含子区域；这个方法也考虑到蛋白质相似性的资料。

Genscan Gene Predictions 路径由 GENSCAN 方法衍生而来，通过这个方法，可以确定内含子、外显子、启动子区域和 poly(A) 信号。此时，这个方法并不期望查询的序列只出现 1 个基因，因此可以对部分基因或被基因之间的 DNA 分隔

的多个基因进行准确的预测。

Human mRNAs from Genbank 路径显示基因库的人类 mRNAs 与基因组序列的比对排列。

Spliced ESTs 和 Human EST 路径显示来自 GenBank 的 ESTs 序列与基因组的序列对齐比较。由于 ESTs 通常代表了转录基因的片断，一个 EST 很有可能对应于某个外显子区。

最后，Repeating Elements by RepeatMasker 这个路径显示的是重复元件，例如散在的或长或短的核元素(SINEs 和 LINEs)，长末端重复序列(LTRs)和低复杂性区域(http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker)。一般来说，在将基因预测方法应用于核苷酸序列之前，需要去掉或掩饰这些成分。

回到视图显示的例子，可以看到大多数路径返回了几乎同样的基因预测结果。作为一个规则，通过多种方法预测的外显子提高了预测的正确率而不会出现“假阳性”结果。多数方法显示 3′ 端非翻译区，以左侧大而短的块状表示。Acembly 路径显示除了全长序列产物（如这个部分第 3 条线所示）之外还有 3 个可能的选择性剪接，其它大多数路径显示与此预测结果相符。Genscan 路径从左、右方向往远处延伸：GENSCAN 可以被用于预测多个基因。

尽管这些图解概要很有用，然而研究者更需要与这些垂直线或块状相对应的序列。以此为例，用 Fgenesh++ 预测作为获得原始序列数据的基础，但不管选择哪个路径其步骤都是一样的。点击标有 Fgenesh++ Gene Predictions 的路径，出现的是一个描述预测的概要页面。

序列的区域与 pendrin 基因相似（从这个例子一开始就已经知道了）。给出了序列的大小及序列开始和结束的预测，并显示预测是以负链为基础的。想要获得序列，点击 Genomic Sequence。使用者将被带到一个标题为 Get Genomic Sequence

Near Gene 的查询页面，在这个页面上，可以获得转录物、编码区、启动子或转录物加启动子的序列。

点击 Transcript 返回的页面显示完整的转录子，外显子以大写字母表示。

点击 Coding Region Only 得到的是编码区，外显子以大写字母表示。

点击 Transcript + Promoter ，返回的页面显示的是在上述选择 Transcript 所获序列的 5′ 端添加了启动子序列，以大写字母表示外显子。启动子的长度显示在文本框内。

点击 Promoter 返回的页面正好是启动子区。

（黄力 译）

问题 8：如何找到一个基因家族的所有成员

HUGO基因命名委员会（http://www.gene.ucl.ac.uk/nomenclature/ ）一直以来都在努力为人类的每一个基因建立一种独特的符号，和一种更长久更具有描述性的名称。因而很多先前在不同的实验室被克隆出来并且被命名为各种不同术语的基因家族的成员，现在却分享一种共同的基因符号。在任何基因组浏览器进行一项文本搜索时，返回的页面通常都会链接到已经在基因组定位的该基因家族内所有已命名的成员。然而，Ensembl 和UCSC最近列出了所有的基因目录，NCBI不仅提供了基因的目录，还将基因绘制成比较直观的概要图谱。

进入NCBI主页，网址是http://www.ncbi.nlm.nih.gov/ 。点击位于右方的链接Human map viewer进入人类基因图谱浏览器搜索页。在查询框中键入所要查询的词“ADAM* [sym]”。星号或者通配符表示能够搜索到与ADAM有关的所有条目。然而[sym]却对以ADAM为其基因符号的所有搜索结果起到限制作用。可以点击Advanced Search 或者直接阅读在线的文献进行其他高级搜索。这次搜索一共获得了 41 个跟ADAM目标基因相匹配的条目，这些条目包括了ADAM基因家族的所有成员以及以ADAM开头的其它基因家族的一些成员，如ADAMTS和ADAMDEC。为了限定使搜索只指向ADAM基因，以排除掉不需要的基因符号，应该使用逻辑搜寻术语“NOT”。在搜索框中键入ADAM*[sym] NOT ADAMTS*[sym] NOT ADAMDEC1*[sym]， 然后再点击find。返回页面上方的染色体图形上，用红色的线条标明了每个基因的位置。很明显，19 个已定位的ADAM基因分布在 11 条染色体上，有一些如位于 10 号和 14 号长臂顶端上的基因就靠的很近，在染色体图形的下面是ADAM家族的 19 个基因的排列，点击链接到那 19 个基因，便可以查到它们的详细资料。

另外一个在基因组中搜索同源基因的方法是通过在 NCBI 或者 Ensemble 上的基本局部序列对齐比较搜索工具（basic local alignment search tool），简称ＢＬＡＳＴ。ＵＣＳＣ上的ＢＬＡＴ搜索没有ＢＬＡＳＴ敏感，可能没有ＢＬＡＳＴ发现的同源基因多。在这个例子中所有和 ADAM2 蛋白质同源的基因组序列将会在 Ensembl 上通过 BLAST 搜索获得。

从网址http://www.ensembl.org/Homo_sapiens/ 进入Ensembl人类基因组的主页，点击BLAST链接。将ADAM2 的蛋白质序列粘贴到查询框中（GenBank 登录号 NP_001455.2，按照问题 5 的步骤从NCBI的 Entrez数据库中已经获得了蛋白质的序列）。将数据库设置成Homo sapiens, genomic sequence，在Ensembl上查找基因组汇编，选择TBLASTN 运行。其他设置使用默认的参数。当这些全部做完以后，点击查询。得到的页面上将有一个检索符号，当检索完成以后，它将直接连接到检索结果的页面。

在检索结果页面的顶端将出现一个用图形来表示找到的蛋白质的位置。这些搜索结果有的是完整的蛋白质，有的只是一个单一的结构域。它们通过 BLAST 得分被标以不同的颜色。红色表示相似程度最大，蓝色的相似处最少，绿色介于两者之间。有一些区域如位于 10 号和 14 号染色体的长臂上的一对基因，它们的位置跟 NCBI 绘制的 ADAMS 基因的位置有些相似，但是也有一些，如位于第 12 号和 Y 染色体上用 BLAST 搜索的结果就是唯一的。这个唯一的结果可能代表 ADAM 家族中的真正成员，它们有可能还没有被命名，所以还不能在文本中搜索出来。还有可能它们是未被命名的假基因或是 BLAST 搜索不太重要的结果。位于第 1 号染色体上的一个基因也许可以在 NCBI 上以文本为基础的搜索中查到，但不一定在 Ensembl 中用 BLAST 搜索到。这个基因和 ADAM 基因之间的相似之处还没有高到能用 Ensembl 的默认的参数值在 BLAST 中搜索出来。

点击其中一条搜索结果旁边的箭头，激活一个向上的菜单，菜单将显示 BLAST 所搜索结果的细节，并提供一个到 BLAST 序列对齐比较的的链接和 ContigView。在第 12 号染色体上的搜索结果包括了一个终止密码，也许只是一个没有内含子的假基因。在结果页面底部是用 BLAST 搜索的结果摘要。点击一个链接到 BLAST 序列对齐排列的条目，在结果页面中部的链接将会用标准格式表示出 BLAST 报告的所有结果。点击一个 BLAST 搜索出来的条目，将会找到有关条目周围区域的 ContigView。

以下由复旦大学 王旭翻译

问题九. 有没有办法自己订制显示方式和自己指定参数？能不能显示我们自己研究过程中需要的标记和特征？

在这个例子中，我们用UCSC浏览器来查看特定的标记。首先，我们打开UCSC的主页（ http://genome.ucsc.edu ），点击在网页左侧蓝色控制条上的“*Browser*”键，设置你感兴趣的区域到基因组浏览器入口（Genome Browser

Gateway）处。例如：你可以设置为：把genome 设置为 Human，把 assembly 设置为 Dec. 2001, 在 position 下面的框中键入 chr22:38496887-39496866 ，然后点击 Submit 这样就可以显示2001年12月组装的人类基因组的22号染色体上的区段。一些标记已经用密格式（dense format）标了出来（见图9.1）。在所示的图形下面的区域里有下拉菜单，使用者可以在Track Controls的标题下面可以选择是否显示图形（见图9.2）。这些下拉菜单有三个选项：

Hide：在显示中去除所选的特定标记；

Dense：在一行中显示此标记的所有注释和特征；

Full：每一条注释或特征显示为一行，也就是我们在其他几个问题中提到的‘exploded view’。

当我们做好我们想要的选择之后，点击refresh按钮重新作图。若想进一步定制特定的标记，可点击浏览器Track Controls选择中的标记名称。例如。使用者可以把一个库中含有某一关键字的Genebank中的EST标记条目都涂上红色，从而将这样的EST从库中去除。在下面的一系列分析中，浏览器将会保留你所做的设置。你可以通过点击reset all键来恢复默认设置。

UCSC系统一个很吸引人的地方是它可以允许用户在本地显示中添加自己的注释、特征和标记。这些改动不会被读取和写入UCSC的原始数据中。这种显示方式可通过回到Human Genome Browser Gateway的页面，向下拖动滚动条至Add Your Own Tracks section来实现。在这里会出现一个大的文本框，你可以输入或粘贴格式正确的文本。还有一种选择，如果你的文本文档是在本地的网页上，你只要告诉其他同事文档的URL，就可以和他们共享你对自定义标记的注释。他们可以通过在UCSC browser中输入URL到文本框来查阅。

为了举这个例子，我们输入以下的文本（见图9.3），并点击位于页面顶部的submit键。

```
browser position chr22:38496887-39496866
browser hide cytoBand
```

```
browser hide stsMap
browser hide gap
browser hide clonePos
browser full refGene
browser dense mrna
track name="scale" description="our peak"
chr22 38996887 38996888 peak
track name="Microsatellites" description="Microsatellites"
color=0,128,0
chr22 38627059 38627060 D22S276
chr22 39005417 39005418 D22S307
track name="Genotyped SNPs" description="Genotyped SNPs"
color=0,0,255
chr22 38518342 38518343 ss146131
chr22 38705963 38705964 ss2941443
chr22 38884157 38884158 ss141110
chr22 39171390 39171391 ss22916
chr22 39438769 39438770 ss1479794
track name="Upcoming SNPs" description="Upcoming SNPs"
color=0,128,192
chr22 38615712 38615713 ss86855
chr22 38804838 38804839 ss85533
chr22 39077895 39077896 ss141190
chr22 39305065 39305066 ss137027
```

浏览器会忽略position框中的输入，只读取贴在Add Your Own Tracks区域里的文件。显示结果如图9.4所示。

以browser开始的行控制浏览器的全局显示，以track开始的行创建新的标记，track后面的行提供每一项的位置信息，所以：

第一行设置浏览位置为22号染色体的38496887–39496866；

接下来的6个以browser开始的行使浏览器显示*Chromosome Band, STS Markers, Gap, Coverage,Known Genes 和 Human mRNAs*这六项。在这里，格式化的文本必须用每个标记的符号名称（Symbolic names）而不是浏览器中显示的名字。UCSC浏览器使用的符号名称在表9.1中列出。与默认的设置对照可知，the Chromosome Band, STS Markers, Gap and Coverage这几项原来均是hide，Human mRNAs是dense而不是full。（见图9.4）

剩余的行告知浏览器创建scale, Microsatellites, Genotyped SNPs 和 Upcoming SNPs这四个新的标记。浏览器显示时，名称放在左侧。以track开头的行为标记命名，放在最上方，并且设定显示这个标记的描述及颜色（见图9.4）。描述作为浏览器中央的标签出现，颜色由三个RGB值决定。所有Track行下面的行为与每一项相关的标记提供位置信息。例如：peak显示在22号染色体38996887–38996888的位置上。

问题十. 对一个给定的蛋白，怎样知道它是否含有我们感兴趣的功能域？其他何种蛋白有与此蛋白相同的功能域？如何确定它与其他蛋白不光在序列上、而且在结构上有相似性？

为了说明在一个蛋白质中找到功能域，我们用睾丸决定因子（TDF）作为例子。TDF 也称为性别决定蛋白 SRY。

虽然我们可以从NCBI的主页中的Entrez搜索框中开始查找，但更好的方法是从LocusLink中开始查找。使用LocusLink的一个好处是它参考了许多交叉的参考资料来对基因和蛋白质的名称进行了标准化，在更大的程度上保证了一开始就能找到正确的蛋白质。从NCBI的主页中（http://www.ncbi.nlm.nih.gov/ ）通过左上角的下拉菜单选择LocusLink，在查询框键入基因名称“TDF”，并点击“go”。一共返回四个基因座（loci）（见图10.1）第一列给出的是Locus ID，这是此基因座固定的识别标签。点击LocusID可查看LocusLink的报告；关于报告的更多细节信息可以从LocusLink的帮助文档和图形中找到。第二列标记作org，给出物种名称的简写。在这里，一条记录来自果蝇Drosophila(Dm)，一条来自mouse (Mm)，一条来自人类human (Hs) 还有一条来自大鼠rat (Rn)。在每条记录右端的一串字母方框提供的跳到其它数据资源的连接。这里我们感兴趣的基因座是列表中的第

三条，因为它是TDF/SRY在人类中的形式。为了寻找这个蛋白的其他信息，点击此行中第二个P字母（绿色）。这样使用者被带到与此LocusLink条目相关的蛋白质条目处（见图10.2）。这样，使用者就可以通过点击任意一条超链接来查看原始数据库中列出蛋白质数据了。

我们来看列表中的第一条目，是一条accession number为NP_003131的NCBI提供的参考蛋白序列。在accession number的右侧有一系列的超链接。点击Blink标签会把使用者带到所查蛋白的Blink页面（见图10.3）。Blink代表的是BLAST Link，它提供了事先做好的图形化的BLAST搜索结果，此结果不仅是对这条蛋白序列的搜索，而是对Entrez蛋白数据库中的所有蛋白。这个事先做好的TDF/SRY的BLAST结果在标签‘204 aa’下面显示出来。在页面的上方横向并排着一些按钮，它们允许使用者就自己感兴趣的蛋白问一系列的问题。如果我们提问的目标是找出TDF/SRY蛋白中的功能域，可以点击CDD-Search 按钮(Conserved Domain Database Search[18])。这样做我们能够看到此蛋白中存在的功能域的图形显示和所查询序列中功能域的序列比对（见图10.4）。在我们的这个例子中，找到了一个功能域：一个HMG box，是一个在许多细胞核中蛋白质的DNA结合域。此功能域在两个组成CDD的数据库（Pfam和SMART）中均被找到，可通过hit list中的accession numbers进一步查看。

译者注：Pfam：蛋白质家族数据库；
SMART：简单模块搜索工具;
CDD：蛋白质保守结构域数据库。

为了确定哪些其他的蛋白质具有相同的 HMG box 结构域，点击页面顶部图形下方的“show”按钮，调用结构域结构检索工具（domain architecture retrieval tool ，DART）。DART 可以显示某个蛋白的功能域，更重要的是，它还能显示具有相同结构域的其他蛋白质（见图 10.5）。查询条目(the HMG-box)为红色显示在页面顶部。其他 NCBI 非冗余数据库中具有该结构域的每一条蛋白显示在查询条目的下方，它们的 HMG-box 也被涂成了红色。找到的这些蛋白质的其他结构域

也用不同的形状和颜色显示出来，在网页的底部有图例。点击左侧的任何链接可以提供新蛋白的更多信息。

虽然待查蛋白中的蛋白质结构域可以被识别，但还不能提供关于结构域功能的更深层信息。但从DART接下去我们可以通过一个迂回的途径来获得这些信息，一个简单的方法是应用一个叫做InterPro的网页资源。InterPro是一关于蛋白质家族，结构域和功能位点的整合信息资源。它把许多蛋白质功能域相关的资源，如PROSITE, PRINTS, Pfam and ProDom[19]，整合在一起。InterPro简单搜索引擎可以从InterPro的主页进入，网址是http://www.ebi.ac.uk/interpro。点击左侧的Text Search按钮将会把使用者带到搜索页面；对于我们的搜索，在文本框中键入HMG Box并点击Search按钮。返回 3 条记录如图 10.6 所示。为了达到例子中的目的，我们进入第一条记录的链接——高速泳动蛋白家族high mobility group proteins HMG1 and HMG2 (IPR000135)。InterPro的结果摘要页面（图 10.7）提供了功能、细胞内定位、和最重要的在细胞中特定蛋白的代谢功能信息摘要。对于需要更进一步信息的使用者，可以查看网页底部的参考资料。使用者也可以查询包含结构域的全长序列；可以通过阅读InterPro的帮助文档了解更多的细节。

本问题的最后一部分问的是与待查蛋白的相似性是否不仅在序列水平上、而在结构水平上也有相似性。回答这个问题需要在 NCBI Structures 中进行一个新的搜索。在 NCBI 主页上，改变页面顶部的下拉菜单为 Structure，在查询框中键入“SRY”并点击“go”。返回 4 个三维结构，其中一个是 1HRY，核磁共振检测出的人 SRY–DNA 复合物结构。点击 1HRY 的链接可进入 1HRY 的结构概要页面。它可以连接到有关 A 链（由蛋白质组成）和 B 链（由核苷酸组成）的细节信息，以及从 CDD 搜索获得的蛋白质的保守结构域（conserved domain ，CD）。点击 A 链的图形，可以获得用一种叫做 VAST 的方法确认的与原来的 SRY 蛋白在结构上相似的蛋白质的列表；更多有关 VAST 方法和列表中数据的解释可以在其他地方找到 [15]。这里显示 SRY 蛋白与成束蛋白-2-小鼠乙酰胆碱酯酶复合体（fasciculin 2–mouse acetylcholinesterase complex），一种叫做 V-1 Nef 的蛋白，70kD 的热休克蛋白，还有肌球蛋白引擎结构域复合体 myosin motor-domain

complex（见图 10.7）在结构上有一定的相似性。VAST 程序常常可以揭示用简单的 BLAST 或 FASTA 搜索不明显的蛋白质间的差异，所以，推荐读者用它或类似的工具解答有关蛋白家族的问题。

译者注：VAST：vector alignment search tool，矢量连配搜索工具

问题十一：一个研究者鉴别并克隆了一个人类基因，但是在小鼠中的同源基因尚未鉴定。怎样查询小鼠基因组中与人类相似的序列？

为了达到本例子的目的，现假设使用者手上还没有感兴趣的人类基因序列。第一步，在UCSC基因组浏览器中找到感兴趣的人类基因。可以通过指向UCSC基因组浏览器主页开始http://genome.ucsc.edu 。从Organism下拉菜单中选择Human然后点击Browser；这两个按钮都位于页面左侧的导航工具条中。使用者将被带到Human Genome Browser Gateway。选择 2001 年 12 月的UCSC基因组整合版本，在position框中键入AGPS字样然后点击Submit。在返回的结果页面中，进入已知基因部分中AGPS的链接。

关于 AGPS 搜索的结果见图 11.1。在主要的图形上有一系列的分子标记，它们的名字显示在左侧。这些已知的基因标记是关于我们查询的 AGPS 的。点击 AGPS 会返回有关这个基因的信息摘要，包括全名和蛋白质产物（alkylglycerone phosphate synthase precursor），还有通向魏兹曼研究所[20] GeneCards 数据库的链接和通向翻译蛋白、mRNA、基因组序列的链接。我们现在看一下被称作 Mouse Translated Blat Alignments 的标记。这个标记显示的是 2001 年 11 月版的利用 BLAT[8] 程序将小鼠和人类基因组对位排列结果的翻译后蛋白形式。BLAT 算法的更多细节和小鼠 BLAT 标记是如何自动产生的可以点击主要图形显示下方的

Mouse Blat 超链接得到。

在 Mouse Blat 标记中点击任何位置可以扩展单个的 BLAT 标记，从而显示我们感兴趣区域中的每一个小鼠与人类序列的比对（见图 11.2）。特别是在翻译模式下，人类和小鼠的外显子序列比内含子序列更加相似。仔细察看有小鼠序列而来的叫做 chr3 81178k (见图. 11.2 中箭头所示)的两个对位排列。在 Mouse Blat 标记中，棕色的竖线代表对齐，横线代表间隙。这些与蓝色竖线相关的对位排列指示 AGPS 在 Known Genes 标记中的外显子。

要查看翻译的 BLAT 对位排列信息，点击标记为 chr3 81178k 的小鼠基因组序列。结果页面（见图 11.3）提供了反映人类基因组组装途径的对位排列的详细信息。这条长度为 607 核苷酸的小鼠序列分为 8 块与人类序列对位排列。在每一块的内部，小鼠与人类有 78%是一致的。要查看此排列，点击 View details of parts of alignment 链接。在结果页面中（见图 11.4），小鼠序列显示在最上方，对位的区域由蓝色显示。接下来是人类基因组序列，在页面底部是小鼠与人类序列并排的对位排列结果（在图中未显示出）。

NCBI 的 UniGene_Mouse 图谱显示了小鼠和人类基因组 mRNA 和 EST 序列的比对。可使用 Maps & Options 添加这个图谱（见图 3.9）。寻找人类基因在小鼠中的同源基因的最早的方法可能是 Ensembl 的预先计算的同源匹配（precomputed Homology Matches），这些可用的匹配直接从一个人类基因链接到小鼠中推定的同源基因。

问题十二：怎样找到小鼠中与人类基因有关的表型突变？

NCBI提供了一组显示人类和小鼠染色体区域同源性的图谱。这些资源可以直接从网址http://www.ncbi.nlm.nih.gov/Homology/ 进入。在这个例子中，我们用的是一个已知并且已经定位的人类基因，然而，从LocusLink中开始查找酪氨酸酶（tyrosinase）条目更加简单。LocusLink的查询页面可在网址http://www.ncbi.nlm.nih.gov/LocusLink/ 找到。从Organism下拉菜单中选择Human，在查询框中输入tyrosinase然后点击Go。要查看酪氨酸酶（tyrosinase，TYR）条目，点击LocusLink号 7299。

在结果页面中（见图 12.1），在 LocusLink 摘要页面叫做 Relationships 的部分中有通向小鼠同源图谱的链接。在本例中，一共有四个可用的显示小鼠 TYR 对位排列的图谱：

NCBI vs. MGD 比对的是 NCBI 组装的人类基因组和 MGD（小鼠基因组数据库，Mouse Genome Database[21]，在 Jackson 实验室）遗传学图，

UCSC vs. MGD 比对的是 UCSC 组装的人类基因组和 MGD 遗传学图，

NCBI vs. EST-based RH Map 比对的是 NCBI 组装的人类基因组和 Whitehead–MRC RH 图谱，

UCSC vs. Hudson et al. 比对的是 2001 年 10 月 7 号 UCSC 组装的人类基因组和 Whitehead–MRC RH 图谱[22]。

每个图谱旁边的 Hs 和 Mm 链接显示的分别是以人类或小鼠为主的人类-小鼠同源图谱。点击 NCBI vs. MGD 图谱中的 Hs 链接。

结果显示的小鼠-人类图谱表明小鼠与人类 11 号染色体上的基因同源的可能基因（见图 12.2）。根据使用的浏览器的不同，你也许需要点击 View as text 来获得输出结果；这是输出结果将是文字格式，与图 12.2 种显示的稍有不同。小鼠基因的染色体定位被显示出来。绿色的圆圈可链接到每一个位点的 UniSTS 条目；在左侧的链接到人类 UniSTS 条目。细胞遗传学位置（cytogenetic positions）

链接到人类基因图谱浏览器或者小鼠基因图谱浏览器。基因符号链接到LocusLink[10]。酪氨酸酶基因被标记为粉红色高亮，定位到小鼠 7 号染色体 44cM 处，这就是我们下一步需要的信息。

小鼠模式种在Jackson实验室的小鼠基因组信息网站（Mouse Genome Informatics site）上有详细描述。到小鼠基因组信息网站（Mouse Genome Informatics site）主页 http://www.informatics.jax.org 并从Query Forms下拉菜单中选择Linkage Maps选项。在结果页面中，定制搜索区域在小鼠Tyr基因附近。在Chromosome一栏中，设定数字为7；然后设定在40到48cM之间的染色体区域。

许多未克隆的小鼠突变体没有在高分辨率的杂交中定位，许多是在有亲缘关系的数目较少的一些小鼠上进行的，因为这样便于统计定位在同一染色体上的另一突变体的表型。所以，对于可能是未克隆的小鼠突变型，有必要选择一个较宽的范围查找（在所在位置左右±4 cM）。在本例中，NCBI的数据告诉我们此基因在44cM处，所以应当在44到48cM的区域中查找。

向下拖动页面（见图12.3），在Markers下，把Include DNA segments设置为No，这样可以减少显示的标记数。一定要包含syntenic markers，它是尚未被精细定位的与7号染色体连锁的DNA标记和突变等位基因，但是可能与TYR相关的表型连锁。在Comparative Maps， Show homologs from species中，选择human (Homo sapiens)，选择Show all markers。其他的选项均用默认设置，点击Retrieve。

Tyr 基因在输出的第二页上找到，在44cM处（见图12.4）。小鼠染色体以简略图解的方式显示在左侧，在右侧为扩展显示。最右侧的一列是在特定区域中蓝色字体的小鼠标记名称；如果有相关的人类同源基因，便用黑色字体显示其名称。显示出的小鼠标记有的是基因，有的是序列标签位点（STSs），有的是隐性突变（recessive mutants）（全部是小写字母），有的是显性等位基因（dominant alleles）（首字母大写）。在页面底部是是同线性标记（syntenic markers），它们是被定位在7号染色体上但是不知道确切位置的标记。

点击44 cm处蓝色的Tyr链接会打开一个关于此基因的基因、标记和表型的摘要（见图12.5）。在本例中我们特别感兴趣的是等位表型。带有突变型Tyr基因小鼠有99株。

使用者也可以通过使用 Ensembl 的 SyntenyView 来查看小鼠与人类同源的染色体区域，在 ContigView 中点击 Jump to syntenyview（见图 1.14，在中间那个黄色的条中）链接就可以了。

问题十三：一位使用者从模式小鼠中鉴别出一个他感兴趣的表型，但却不能将相关基因的关键区域缩小到 0.5cM 之内。怎样从这个区域中找到小鼠的这个基因？

Ensembl提供了一个小鼠基因组浏览器，与人类组浏览器相似。它与最近组装的小鼠基因组序列是同步更新的，在撰写本文时，显示的是MGSC的组装的第三版的小鼠基因组序列（用的是2002年2月的数据）。据估计，此序列覆盖了96%的小鼠常染色质DNA，Ensembl预测它包含22000多条基因。打开Ensembl小鼠基因组主页，http://www.ensembl.org/Mus_musculus/ 。在下拉菜单中选择Marker，在旁边的框中输入标记名称“RH114718”然后点击Lookup。点击结果中的任意一个链接可以查看这个放射性杂交标记的细节信息。RH114718定位在19号染色体的一个单一的位点上，也叫做MGI:102447, MTH1904 and D19MIT109（如图13.1）。点击chromosomal position可以在基因组背景下察看此标记（如图13.2）。

图 13.2 的 Overview 部分 19 号染色体上以此标记为中心的长度为 1 Mb 的区域，在图中标记为 D19MIT109。在此区域由 30 多个预测的基因，有的是已知的，有的新基因。在页面底部的 Detailed View 是此标记周围区域的放大显示。要得

到这个区域的基因和转录产物的更好的视图，可通过点击 zoom control（离“-”号最近的）中的最长的条状按钮来缩小显示（zoom out）。现在 Detailed View 中显示的仍是这个区域，但是有许多其它的图形（如图 13.3）。基因的剪切模式和基因预测显示了出来，还显示了基因组和其他蛋白质、mRNA 的同源区域。将鼠标指向任何一个图形，使用者可以从打开的小菜单中看到附加的描述的链接。

让我们来看图13.3中红色箭头指示的新基因。要查看关于此基因的基本信息，将鼠标停在这个基因的图形上面并在出现的菜单中选择Transcript Information。打开的GeneView窗口（如图13.4）提供了此基因的描述和一个到推定的人类同源基因GeneView窗口的链接（图13.4中Homology Matches部分）。要查看数据库中可与此小鼠新基因的预测外显子对位排列的序列，将鼠标停在Detailed View中的基因上面并在出现的菜单中选择Supporting evidence。图13.5显示了与此新基因的外显子对位排列的mRNA和蛋白质。点击任意一个绿色方块可以看到这个新的转录物与数据库中序列的对位排列。

Detailed View的缩小显示（zoom out）还提供了计算小鼠和人类基因组区域同源性的链接（图13.3，粉红色工具条）。由于小鼠基因组的组装和注释在人类基因组之后完成，察看在人类基因组同源区域中的人类基因很可能也是有用的。

UCSC还提供了用于最新组装的小鼠基因组序列浏览器和BLAT搜索工具。链接可以在UCSC的主页上找到，http://genome.ucsc.edu/ 。NCBI开发的小鼠基因组分析工具，包括小鼠图谱浏览器和小鼠BLAST页面可以在 http://www.ncbi.nlm.nih.gov/genome/guide/mouse/ 查到。