遗传 HEREDITAS (Beijing) 2012 年 10 月, 34(10): 1339－1347
ISSN 0253-9772 www.chinagene.cn

DOI: 10.3724/SP.J.1005.2012.01339

Otterlace软件在猪全基因组序列人工注释分析中的应用

张杰[1], 尚宗民[2], 曹建华[1], 樊斌[1], 赵书红[1]

1. 华中农业大学, 农业动物遗传育种与繁殖教育部重点实验室&农业部猪遗传育种重点开放实验室, 武汉 430070;
2. 湖北省襄阳职业技术学院生物工程学院生物工程系, 襄阳 441050

摘要: 2009 年 11 月, 美、英等国科学家宣布首次绘制出家猪的基因组草图。近两年, 随着全基因组序列陆续释放, 越来越多的测序片段得到正确拼接组装, 从全基因组水平上对猪功能基因进行注释分析显得尤为迫切。文章以丝切蛋白 1(Cofilin 1, CFL1)基因的注释过程为例, 介绍了运用 Sanger 研究所开发的 Otterlace 软件对猪全基因组的免疫基因序列进行人工分析与注释。通过详细说明 Zmap、Blixem 和 Dotter 3 个注释工具的使用方法, 并给出了注释过程的主要步骤, 以期对 Otterlace 的应用起一个抛砖引玉的作用。运用 Otterlace 软件对 243 个免疫相关基因进行分析, 其中 180 个基因得到完整或部分注释, 这为后续深入开展这些基因的功能研究奠定了基础。

关键词: 猪; 全基因组; 人工注释; Otterlace

Manual annotation of the pig whole genomic sequence using Otterlace software

ZHANG Jie[1], SHANG Zong-Min[2], CAO Jian-Hua[1], FAN Bin[1], ZHAO Shu-Hong[1]

1. *Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education & Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture, Huazhong Agricultural University, Wuhan* 430070, *China*;
2. *Department of Biological Engneering, Xiangyang Vocational & Techinical College, Xiangyang* 441050, *China*

Abstract: In November 2009, scientists from the US, UK, and other countries announced the complete genome sequence draft of the domestic pig. With the release of improved versions of the pig genome assembly and the increase of correctly assembled sequenced fragments over the past two years, it is particularly urgent to have the pig genes annotated at whole-genome level. This article is aimed at introducing an excellent manual annotation tool, Otterlace software, developed by Sanger institute. We used *CFL1* (Cofilin 1) gene as an example to expound the usage of the three main components of Otterlace, Zmap, Blixem, and Dotter tools, and developed a practical procedure for manual annotations. We have analyzed 243 immune-related genes, among which 180 genes have been completely or partially annotated, offering novel information to the porcine functional genomics.

Keywords: pig; whole genome; manual annotation; Otterlace

收稿日期: 2012–03–12; 修回日期: 2012–07–10
基金项目: 国家自然科学基金项目(编号：30901021)和教育部新教师基金项目(编号：20090146120032)资助
作者简介: 张杰, 博士研究生, 专业方向：动物遗传育种。E-mail: bailihongchen@163.com
通讯作者: 赵书红, 教授, 研究方向：动物遗传育种。E-mail: shzhao@mail.hzau.edu.cn
网络出版时间: 2012-8-23 10:14:40
URL: http://www.cnki.net/kcms/detail/11.1913.R.20120823.1014.003.html

2005 年, 中国科学院北京基因组研究所与丹麦猪育种与生产委员会(The Danish Committee of Pig Breeding and Production, DCPBP)联合公布了中国和欧洲 5 个不同家猪品种的 384 万个基因组测序片段[1], 详细序列信息可以从NCBI Trace Repository (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=searchCenter name: "SDJVP"; Project name: "Sino-Danish Pig Genome Project") 和GenBank中得到, 此外他们同时还分析了 100 个不同猪组织和发育阶段的表达序列库。同年, 由美、英等多国科学家及政府和企业代表组成的国际猪基因组测序联盟(The Swine Genome Sequencing Consortium, SGSC)公布了猪基因组测序计划(Porcine Genome Sequencing Project)的路线图[2], 详细介绍了如何实施这一计划。2009 年 11 月初在英国Sanger研究所举行的为庆祝猪基因组测序计划完成而召开的Pig Genome III会议上, 他们宣布首次绘制出了家猪基因组草图, 草图包含杜洛克猪 98%的基因组序列, 约 27 亿个碱基[3], 主要在Sanger研究所完成测序。随着猪基因组的测序完成, 下一阶段的任务之一就是从全基因组水平上对猪进行序列分析注释, 此次会议分别成立了猪免疫应答基因、肌肉发育与脂肪沉积基因、非编码RNA基因等几个注释小组, 十余个国家的研究者参与, 其中免疫应答注释小组(Immune Response Annotation Group, IRAG)由美国依阿华州立大学的Christopher K. Tuggle教授和法国农业科学院(INRA)的高级研究员Claire Rogel-Gaillard博士负责, 本课题组作为中方受邀成员参与该小组的相关注释工作。

目前基因组注释方法主要包括两种。一种是通过自动注释系统, 由计算机综合 3 种方法完成从基因组序列到预测新基因：(1) 分析mRNA 和EST数据直接得到结果; (2) 通过相似性比对从已知基因和蛋白质序列得到间接证据; (3) 基于各种基因结构的统计模型和算法从头预测(ab initio prediction)。另一种是基于自动注释系统的人工注释, 它对于自动注释过程中出现的错误有一个很好的纠正和完善作用。本文所采用的基因组注释工具Otterlace是一款成熟的软件, 已经被广泛运用于人、小鼠及斑马鱼等模式生物日常性的注释工作中, 2009 年发表在*Science*杂志上关于牛基因组序列[4]的一篇文章中人工注释的 4 000 多个基因就是通过运用该软件完成的, 猪基因组测序联盟也决定用这个软件对猪基因组进行注释。本文详细介绍了Otterlace软件包的构成及并以丝切蛋白 1 基因为例对使用方法进行了说明。

1 工具与方法

1.1 工具

本研究采用的基因注释工具是Sanger研究所开发的一款基于Mac OS平台的基因组注释软件“Otterlace”(软件下载地址http://www.genome.iastate.edu/tools/share/otterlace/; 软件使用手册地址http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/otterlace_user_manual.pdf), 它是一个交互式的图形化客户端, 使用集成有Zmap和Perl/TK工具的本地AceDB数据库对基因组进行注释, 注释储存在一个扩展的Ensembl schema(“Otter”数据库)中, 它会为研究者呈现一条染色体上的连续区域。Zmap、Blixem[5]和Dotter[6]是基因注释过程中主要使用的 3 个工具, 都采用C语言编写, 其中Zmap是一个为基因组各种特征提供一个可视化工具的软件包; Blixem是一种BLAST多重比对观察器; Dotter是一个图形化的二维散点图程序, 它们为序列分析提供了丰富的图形环境。

1.1.1 Zmap

该软件包使用gnome工具包(GTK2)在“画布”(即Zmap主界面)上描绘各种基因组特征序列, 如ESTs、mRNAs、RefSeq Protein matches、CpG islands、Annotated transcripts等等。通过使用这个界面, Otterlace可以分析注释已储存在“Otter”数据库中的序列, 同时也会及时更新到Vertebrate Genome Annotation (VEGA)数据库[7]中(http://vega.sanger.ac.uk/index.html)。

1.1.2 Blixem

该工具全称为“BLast matches In an X-windows Embedded Multiple alignment”, 即“在嵌入式多重比对 X 窗口中的 BLAST 匹配”。严格意义上讲, Blixem不是一个多重比对工具, 而是一种“多对一”的比对, 它被用来核对多条核苷酸或氨基酸序列与一个参考序列进行的 Blast 比对, 比对由程序自动完成, 但完成后不能手工移动序列。

1.1.3 Dotter

Dotter 是一个图形化的二维散点图程序，用来对两个序列进行详细比较，Dotter 和 Blixem 都可以单独使用。在这里，被比较的两条序列分别沿着 X 轴和 Y 轴展开，其中一条序列的每个残基都会与另一条序列的每一个残基进行比对，在两序列彼此相似的区域中，一系列高分序对(HSPs)将会以斜对角线的方式呈现在点阵图当中。Dotter 也是由程序自动完成，与 Blixem 不同的是，比对后可以手工调整序列进行精细查看核对剪接位点。

1.2 方法

基于 HAVANA group(Human And Vertebrate Analysis and Annotation)的相关注释规则(http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf)，本文主要以蛋白编码基因 *CFL1*(Cofilin 1 (non-muscle)，丝切蛋白 1)为例，介绍 Otterlace 软件在猪全基因组人工注释分析中的应用。

1.2.1 获取注释基因的相关信息

通过网站A Wish List of Genes for Annotation(www.animalgenome.org/cgi-bin/host/ssc/gene2bacs)确定该基因的基本信息，该网站通过猪基因组 Build.9 版本的染色体位置与两个BAC(bacterial artificial chromosome)文库进行匹配，其中一个是“已完成”文库，即已完成测序的克隆是连续的不含任何内部间隙，误差率少于 0.01%，也即每 10 000 个核甘酸中只允许 1 个以下的错误[8] 的序列，另一个是“未完成”克隆文库。基于此网站有两种方法可以确定基因在染色体上的位置：Ensembl预测和Blasts，详细信息例如Ensembl Gene ID/Blast location，Genome Location，Accession number等都可以通过超文本链接到Ensembl数据库。

1.2.2 在 Otterlace 中定位目的基因

获得相关信息后便可进行基因初步定位，首先是进行物种选择，如人、黑猩猩、长臂猿、小鼠、大鼠、斑马鱼、牛、猪等，其次是选择基因所在染色体，大鼠、斑马鱼、牛、猪等，再次选择染色体拼接组装的 Clone number(s)和 Accession number(s)，最后 Run lace 运行软件，通过链接到 Sanger 网络服务器(http://www.sanger.ac.uk/resources/downloads/other-vertebrates/pig.html)下载注释所需数据，共包括 vertebrate_mRNA、EST_Pig、EST_Human、EST_Mouse、EST_Other、Swissprot、TrEMBL 等 17 种数据资源(图 1，更多的描述信息可以从网址 http://scratchy.internal.sanger.ac.uk/wiki/index.php/Otterlace_filter_descriptionsk 中得到)。下载完成后即弹出 Zmap 的主界面，根据其他物种的同源基因结构信息，例如 NCBI 中人或牛的该基因含有的外显子数目，结合 Ensembl 预测或 Blast 的转录本信息(有时 Ensembl 预测的外显子结构并不是很准确)，如 Transcription ID: ENSSSCT00000014181，寻找该基因的位置，找到后“Mark”标记这一染色体区域，依次打开 vertebrate_mRNA 和 EST_Pig 列(图 2)。

1.2.3 建立主要转录本

选择 vertebrate_mRNA matches 列中最具代表性的一条 mRNA(图 2，左侧窗口成簇排列的 mRNA 中呈黄色突显)，它可能是物种特异的，如猪的 cDNA；也有可能是直系同源证据，如人或牛的 cDNA。本例中为猪特异的 mRNA match。然后“Pfetch”(ACEDB 的外部程序，当需要数据库，如 EMBL 或 GenBank 的数据时，它开始检索，它还可以存储 ACEDB 中所有序列的记录)读取该序列的克隆登录号及序列信息，Blast 该序列确认其所在基因座，就是所要注释的基因。在 Ensembl 预测的转录本(图 3)基础上建立一个新的转录本，然后 Blixem 猪的 EST 和 vertebrate_mRNA，逐个核对每一个外显子的剪接位点是否符合通用规则，即 AG-GT 规则(图 4)。

1.2.4 延伸 UTR 及注释 polyA feature

利用物种特异性即猪的 EST 或 mRNA 证据延伸 5′和 3′UTR，若有很充分的证据支持多聚 A(poly A)特征，如通过 Pfetch 查看某一 EST 证据在 NCBI 的 EST 数据库中显示含有已知的 “poly A tail”或特异性的 mRNA 显示含有 poly A(非基因组本身的多聚 A)，则注释 poly A 特征，否则不予注释，等待以后更多的实验数据支持(图 5，图 6)。另外，根据注释规则，一般情况下 poly A site 可以单独注释，但 poly A signal 必须与 poly A site 一起注释。

1.2.5 添加核苷酸和蛋白质序列证据

在核对外显子的过程中，所用到的最佳 mRNA 将作为 cDNA 证据，此外添加的 EST 证据也要尽可

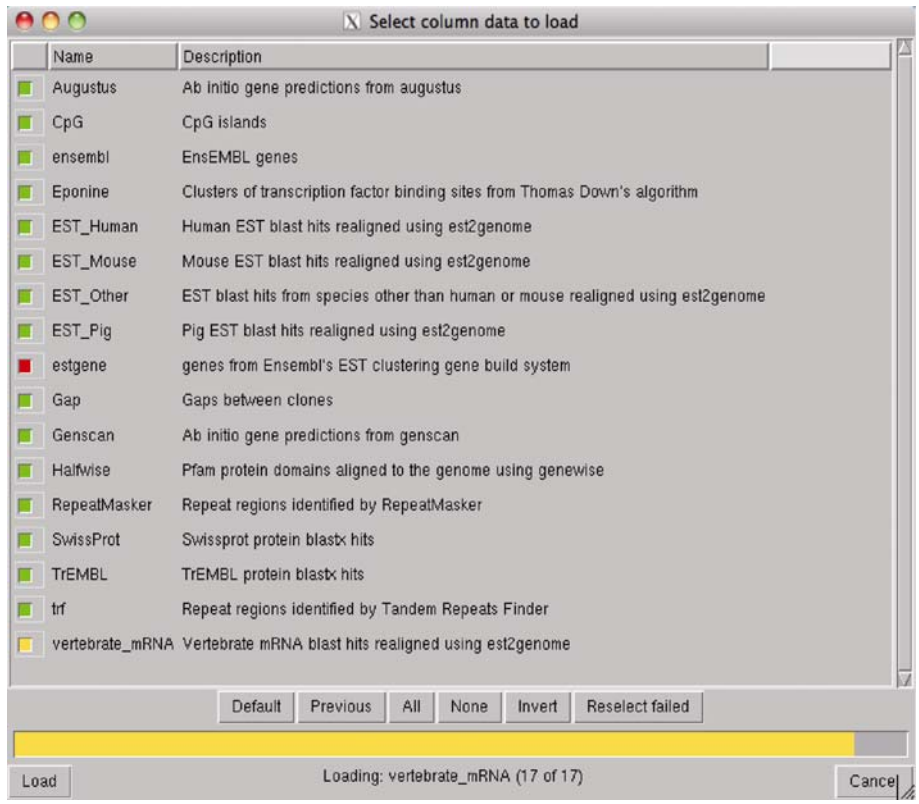

图 **1** 通过 **Sanger** 网络服务器下载注释时所需资源信息
绿色框代表下载成功; 红色框代表下载失败; 黄色框代表未下载。

能地包括所有注释的外显子, 最后注释完主转录本(Main variant)后, 如果有证据支持还需添加蛋白质序列, 一般添加人或牛的蛋白质序列作为证据, 如果有物种特异性(猪)的蛋白质序列(SwissProt蛋白序列证据), 则只需添加该证据即可, 但假如只有猪的TrEMBL 蛋白序列证据, 则还需要添加人或牛的同源蛋白质序列。另外根据注释变异体的相关规则, 确定主转录本的类型, 例如 Known_CDS(猪特异性mRNA 支持)或 Novel_CDS(无特异性 mRNA 支持, 但有其他物种如人或牛的同源基因 mRNA 支持), 并依据该基因是否出现在NCBI的Entrz gene中确定其基因座属于已知(Know locus)还是未知(Unknown locus)。

1.2.6 注释可能存在的选择性剪接体

注释完主要转录本后, 还可以通过挖掘其他一些核苷酸序列信息注释可能存在的选择性剪接体, 在 Zmap 中可以直观地发现外显子是否存在选择性剪接, 如外显子跳跃(Exon skip)、选择性的 5′端或 3′端(Alternative 5′ or 3′ splicing)、内含子保留(Intron retention)等。由于 HAVANA 小组要求所注释的基因结构(转录本)必须要有转录证据, 或者 cDNA、EST, 或者蛋白质序列支持, 所以并不要求所有注释的转录本都必须是完整的, 对于 CDS 可以 5'或 3'端不完整。另外这种支持不需要所用证据都来自于基因位点特异性的证据, 可以是直系同源也可以是旁系同源证据, 因此有时候在核对剪接位点时需借助更加精确的比对工具 Dotter(如图 7), 该工具最大的好处就是可以手动调整序列查看程序比对的结果, 看是否符合标准剪接及剪接后的序列在不同外显子之间是否连续, 再者它还可以找到由于软件本身存在的缺陷导致的某些序列呈现在 Zmap 窗口中未能显示出来的外显子。

2 结果与分析

本研究所有的证据都是基于提交到公共数据库中的序列信息, 如 GenBank、EMBL、DDBJ、UniPROT 等, 通过基因组注释软件 Otterlace 运用生物信息学的方法对猪全基因组免疫相关基因进行注释, 取得

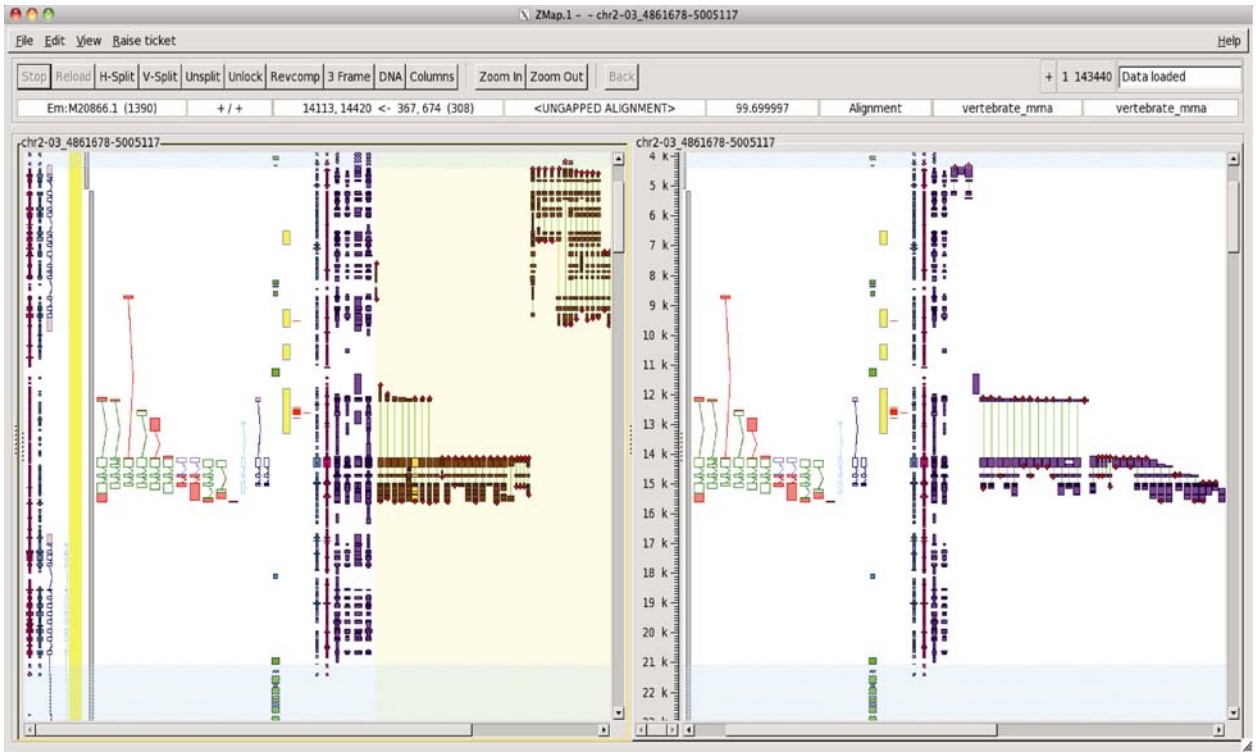

图 2　通过 **Zmap** 主界面结合 **NCBI** 或 **Ensembl** 中的基因相关信息标记所要注释基因的染色体区域

左侧窗口中打开着的成簇排列的棕色线条框代表脊椎动物 mRNA 同源序列(vertebrate_mRNA matches)，紧接其左边的 4 条紫色线条框分别代表：人的同源表达序列标签(EST_Human matches)、鼠的同源表达序列标签(EST_Mouse matches)、猪的表达序列标签(EST_Pig matches)、其他物种同源表达序列标签 EST_Other matches(从左至右)；在 4 条紫色线条框左边的两条线条框分别为灰蓝色的 SwissProt 蛋白证据和粉红色的 TrEMBL 蛋白证据，前者为实验得到数据，后者为计算机预测得到数据。右侧窗口中打开的成簇排列的紫色线条框是 EST_Pig matches。窗口中成簇排列的绿色线条框代表注释的转录本，两端红色部分为 UTR。图中 4 个黄色线框代表 CpG 岛。(下同)

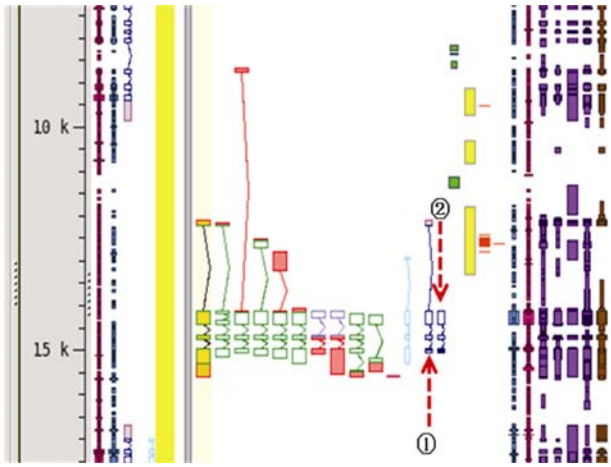

图 3　通过 **Ensembl** 预测的基因转录本建立新的转录本

Ensembl 对 *CFL1* 基因预测的转录本有两个，如图中蓝色线条框所示，这里综合该基因的各种相关信息选择①，在此基础上建立新的转录本。

了较好结果。截至目前，本课题组所承担的分布于不同染色体上(不含 Y 染色体)的 243 个免疫相关基因中，有 180 个基因得到完整或部分注释，其余 63 个基因未能注释。这 63 个基因中的有些基因如 *IGF2*、*CD163* 等可以在 NCBI 的 Entrez Gene 找到部分有关信息，但可能由于克隆片段还未组装添加到 Otteralce 软件中而找不到相关序列；另有一些基因可能还未被完全克隆或其序列比较特异而未能完整测序。另外，在已经完整注释了的基因中发现很多基因含有数目不等的可能的选择性剪接体(Alternative splicing)，有一些是基于物种特异性的证据，但是大部分的证据支持来源于非物种特异性的证据，今后还需对这些剪接体进行实验验证，并进行相关的一些功能研究。本研究的注释结果已经被汇编整合并将呈现在 Ensembl browser 中，这将为后续深入开展基因组功能研究奠定基础。

3　讨 论

本文进行人工注释，主要是相对于计算机的自

图 4　通过 Blixem 比对工具核对每一个外显子的剪接位点是否符合标准剪接

图中红色线条标记的碱基即是内含子剪接的通用规则，亦即 AG-GT 规则。黄色基因组序列下方并行排列的蓝色序列即是公共数据库中的实验数据匹配到该位置上的外显子片段，此种比对属于“多对一”比对，即大量序列与猪的基因组序列进行比对。标记为黄色的基因组序列有两条，它们为互补序列，图中有些序列证据比对到正义链，有些则比对到负义链，这与测序时所测的是哪条链有关。

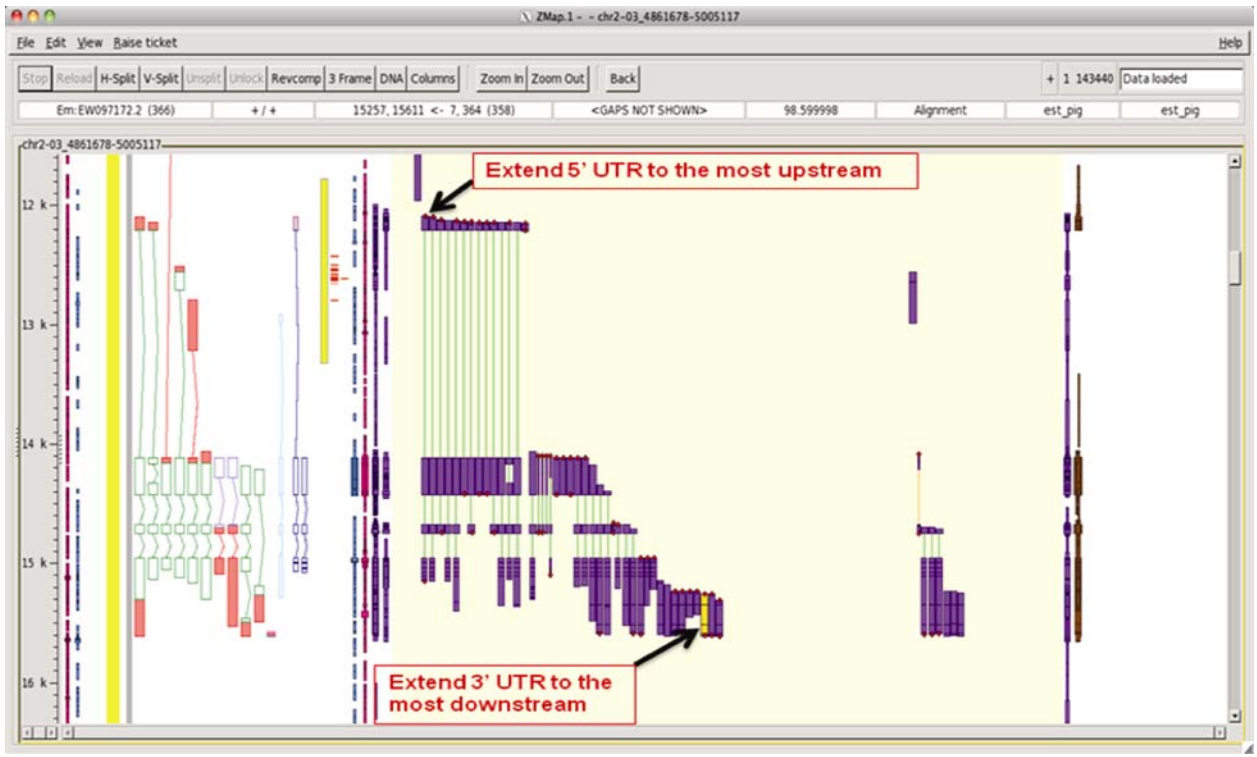

图 5　通过 Zmap 主界面注释基因的 UTR

在 Zmap 主界面中打开特异性物种，即所要注释的物种-猪的 ESTs，找到该基因最上游和最下游的 ESTs 片段序列证据延伸 5'和 3'UTR。

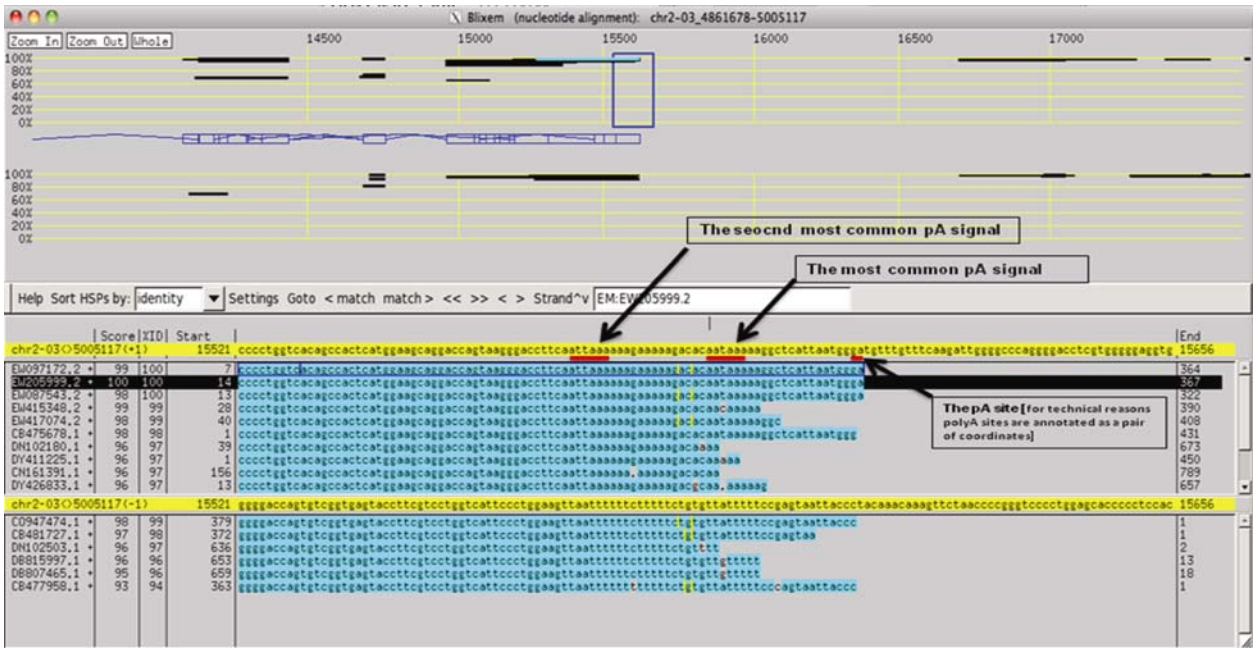

图 6　通过 Blixem 比对工具注释基因的 poly A feature

注释多聚 A 特征(poly A feature)，即多聚 A 保守位点(poly A site) 和多聚 A 信号序列(poly A signal)。本文所举例子 *CFL1* 基因存在两个最常见的加尾信号：aataaa 和 attaaa，通常情况下在加尾信号出现的下游 16 个 bp 左右的地方会进行多聚 A 的添加(ESTs 证据中会显示出一连串的 poly A，而在 Blixem 比对界面结果中则不会显示出来，例如图中标记为黑颜色的猪 EST 序列 EW102900.2 的最后两个碱基为 ga，在其后面是一连串的 poly A，而在比对结果中则未显示出)。

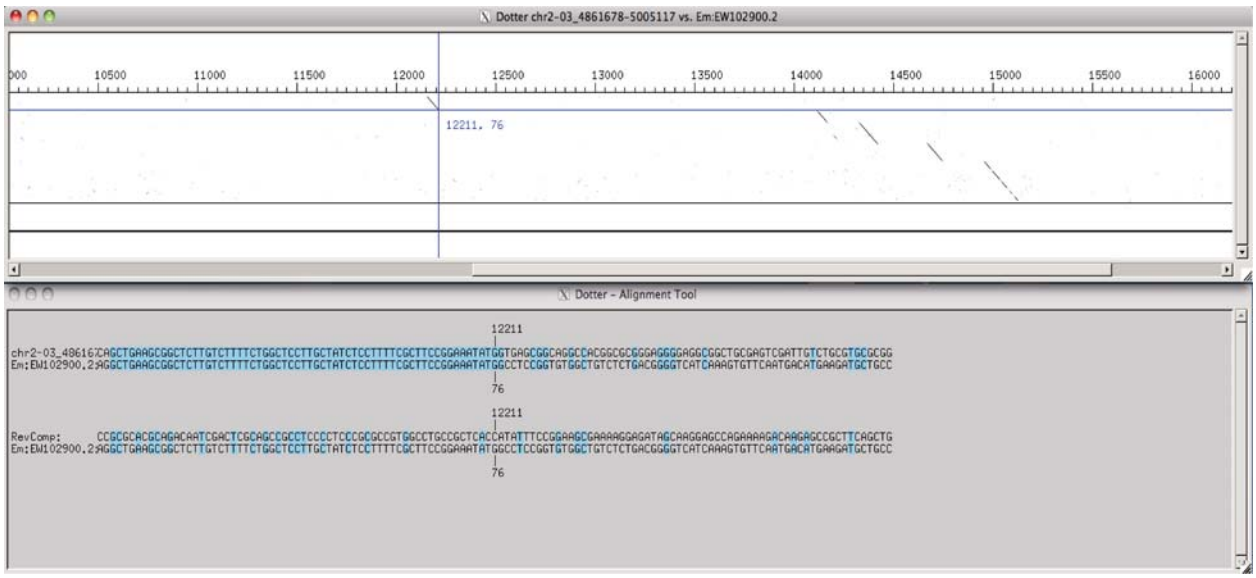

图 7　通过 Dotter 比对工具进行序列的精细比对

图中所示的是猪 mRNA 序列 EW102900.2 与标记的基因组区域进行 Dotter 比对的结果，这种比对属于“一对一”比对，即特定一条序列与特定基因组区域进行比对，故精确性更高。图的上部分中几条黑色斜线代表该 mRNA 序列中的某一段序列匹配上猪的基因组序列，图的下部分中蓝色标记的是与猪基因组序列相匹配的序列片段，此种比对可以进行人工移动碱基位置来比对序列，核查是否符合标准剪接及不同外显子序列间的连续性与否。

动注释系统而言的，当前，不论是 NCBI，还是 Ensembl，都有一套自己的基因组自动注释系统。这类系统通常是基于所有公开数据库中的数据信息，由生物信息学工作者运用生物信息学方法和工具，

编写合适的算法或程序并开发相应的软件系统进行基因组的高通量注释。然而“自动注释”的准确性仍然欠佳，在基因组注释过程中仍然存在各种误差，甚至错误，这时候就需要人工手动去校正这些错误，弥补自动注释的不足，使基因注释工作得以完善。

本研究大部分对已知基因进行注释，在注释基因主要转录本的同时，依据公共数据库中存在的实验数据，mRNA 或 EST 证据，注释可能存在的选择性剪接体，但同时也试图对部分承担的未知基因进行预测和注释。主要分以下几种情况：1)完全未知基因，即没有任何所要注释物种的 EST 或 mRNA 证据支持的基因，但在人或牛中则有该基因的相关信息(即有一种参照作用)，此种基因预测或注释一般是通过调取人的该基因的 mRNA 序列，在 Ensembl 网站中(http://asia.ensembl.org/Multi/blastview)BLAT 到所要注释的物种基因组序列中，然后通过染色体上的相关位置的定位找到目标区域，进而在 Otterlace 软件中进行注释，但是此种完全未知基因包含的信息量太少，而且也与基因组测序列的释放和组装版本存在很大关系，一般此种基因注释的成功率较低且不太可靠，相信随着 Ensembl GeneBuild 版本的更新以后会得到较好的注释；2)部分未知基因，即已经有了一些信息，但是不完全，此种基因注释包括 5′端或 3′端缺少外显子，以及基因中间部分缺少外显子，对于此种情况的处理是基于注释当前所能得到的所有实验证据信息对基因进行“片段注释”，即进行一段一段的注释，缺少的外显子要么需等待组装版本的更新才能进行完整的注释，如某一个克隆还没有组装到或未能正确组装到该基因所在位置；要么需要新的实验证据支持，如有些基因的转录起始位点到翻译起始位点这一段 5′UTR 或翻译终止位点到 PolyA 加尾位点之间的 3′UTR 具有很强的特异性，而注释当前还没有能够得到该基因的 5′RACE 或 3′RACE 实验证据。

本研究进行人工注释的软件是由 Sanger 研究所开发的一款名为 Otterlace 的集成软件，通过本地 ACEDB 数据库最大限度汇集公共数据库中的信息资源，利用可视化的 Zmap 图形界面对基因组数据进行管理和分析，通过这款软件对猪基因组序列进行分析注释将提高基因组草图的精度。本软件的缺点或不便的地方是，如果某一基因位于负链(软件定义位于基因组序列右侧为正链，左侧为负链，如图 3 所示黄色线条代表基因组序列，其右侧灰色线条代表不同的 contigs)，要先通过“RevComp”功能键把正负链进行转换，以便正确地进行注释，当进行 Blixem 或 Dotter 比对时可以更好地符合阅读习惯，即从 5′→3′方向阅读。

本研究基于的猪基因组释放及组装版本是Sus scrofa Build 9.0[9]，前期的自动注释由Ensembl一套标准的哺乳动物注释流水线系统完成，但它也有不能很好地提供注释的地方，如基因的剪接变异体、假基因、重复基因、非编码基因和保守的基因家族，对于这些基因人工注释就显得尤为重要。

本研究承担的是编码基因的相关注释工作，在人工注释过程中发现主要有两方面的问题，一是虽然测序达到 6X基因组覆盖，包含基因组 98%的序列，但这只是一个草图，要想得到更为精细的基因组图谱，测序深度还需提高，并且已测的序列信息完全释放尚需一个过程。就目前这个版本而言，还有一定数量的克隆缺口(Clone gaps)分布于猪全基因组中，导致在注释过程中对某些基因只能注释部分CDS或未能找到相应的染色体物理位置；另一方面在组装克隆片段的过程中，虽然物理图谱能够帮助鸟枪法序列的组装[10]，但对于组装软件如CAP3[11]、EULER[12]、ARACHNE 2[13]、PCAP[14]等来说干扰最大的是基因组中的重复序列，即使是低拷贝的重复序列也会使相对基因组来说很小的DNA片段如BAC克隆的组装出现错误，本研究应用Otterlace通过人工注释手段能够准确发现这些错误，对于主要转录本而言，主要存在以下几方面的错误组装： (1)某一基因的不同外显子位于不同的链上；(2)5′和 3′端外显子顺序发生颠倒；(3)某一克隆片段的重复组装；(4)某一基因的克隆片段组装到另一个基因的序列中从而影响到另一基因CDS的注释(或者导致预测的氨基酸序列提前终止，或者虽然不改变阅读框，但经过蛋白质同源性比对后发现该区域与其他物种或与本物种已知序列存在相当大的差异)，这些错误组装需要在今后更新版本数据时加以修正调整。此外，测序问题如某一基因的相邻两个克隆片段间，由于gap导致片段末端碱基测序不准确及基因组本身存在的测序错误如某一位点碱基的增加或缺失，这两种情况从目前的注释工作来看还算比较少见，

但也会在一定程度上影响CDS的注释，即对于主转录本而言可能会导致预测的氨基酸序列提前终止。今后的任务是准确确定基因的染色体位置、结构及序列信息，深入分析和挖掘更多的选择性剪接体以及注释更多的新基因，这将依赖于更多的提交到公共数据库中的实验数据和基因组测序数据的后续释放以及克隆片段更好的组装。

参考文献**(References)**:

[1] Wernersson R, Schierup MH, Jørgensen FG, Gorodkin J, Panitz F, Staerfeldt HH, Christensen OF, Mailund T, Hornshøj H, Klein A, Wang J, Liu B, Hu SN, Dong W, Li W, Wong GK, Yu J, Wang J, Bendixen C, Fredholm M, Brunak S, Yang HM, Bolund L. Pig in sequence space: A 0.66X coverage pig genome survey based on shotgun sequencing. *BMC Genomics*, 2005, 6: 70. DOI

[2] Schook LB, Beever JE, Rogers J, Humphray S, Archibald A, Chardon P, Milan D, Rohrer G, Eversole K. Swine genome sequencing consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comp Funct Genom*, 2005, 6(4): 251–255. DOI

[3] Rogatcheva MB, He WS, Larkin DM, Marron BM, Ehrhardt MJ, Beever JE, Schook LB. Survey sequencing of the porcine genome using BAC end sequences. In: Plant and Animal Genome XII final Abstract Guide. San Diego, CA, 2004: 59. DOI

[4] Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, *et al*. The genome sequence of Taurine cattle: A window to ruminant biology and evolution. *Science*, 2009, 324(5926): 522–528. DOI

[5] Sonnhammer EL, Durbin R. A workbench for large-scale sequence homology analysis. *Comput Appl Biosci*, 1994, 10(3): 301–307. DOI

[6] Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 1995, 167(1–2): GC1–GC10. DOI

[7] Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, Wilming L, Hubbard T. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*, 2005, 33(S1): D459–D465. DOI

[8] Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC. Genome project standards in a new era of sequencing. *Science*, 2009, 326(5950): 236–237. DOI

[9] Humphray SJ, Scott CE, Clark R, Marron B, Bender C, Camm N, Davis J, Jenks A, Noon A, Patel M, Sehra H, Yang FT, Rogatcheva MB, Milan D, Chardon P, Rohrer G, Nonneman D, de Jong P, Meyers SN, Archibald A, Beever JE, Schook LB, Rogers J. A high utility integrated map of the pig genome. *Genome Biol*, 2007, 8(7): R139. DOI

[10] Warren RL, Varabei D, Platt D, Huang XQ, Messina D, Yang SP, Kronstad JW, Krzywinski M, Warren WC, Wallis JW, Hillier LW, Chinwalla AT, Schein JE, Siddiqui AS, Marra MA, Wilson RK, Jones SJM. Physical map-assisted whole-genome shotgun sequence assemblies. *Genome Res*, 2006, 16(6): 768–775. DOI

[11] Huang XQ, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*, 1999, 9(9): 868–877. DOI

[12] Pevzner PA, Tang HX, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA*, 2001, 98(17): 9748–9753. DOI

[13] Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*, 2003, 13(1): 91–96. DOI

[14] Huang XQ, Wang JM, Aluru S, Yang SP, Hillier L. PCAP: A whole-genome assembly program. *Genome Res*, 2003, 13(9): 2164–2170. DOI