上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

课程论文

COURSE PAPER

论文题目：非洲人泛基因组数据库的构建

课程名称：数据库原理

指导教师：王靖方

学院(系)：生命科学技术学院

组员分工：

朱　宸：ER图构建 结果可视化 论文

张昱朦：数据收集 后端搭建 论文

曾贝琴：前端美化 前后端连接 使用说明

李唯一：前端网页搭建 使用说明 论文

非洲人泛基因组数据库的构建

摘　要

关于人类泛基因组的研究始于 2009 年，测序发现两个不同人之间的基因组就有 4 千万对碱基的差异，存在 73-87 个不同的基因。人类基因组中除原先公认的单核甘酸多态性，插入删除多态性和结构性变异以外，还存在着种群特异甚至个体独有的 DNA 序列和功能基因。构建非洲人泛基因组数据库并对其进行探究可以帮助丰富人类参考基因组并对两个族类的基因差异进行分析，为泛基因组序列的相关研究提供更多帮助。

非洲人泛基因组数据库的数据来源基本来自于 Assembly of a pan-genome from deep sequencing of 910 humans of African descent 公布在 Nature 与 NCBI 上的数据，总共有 910 名非洲人个体，125715 条 contig。在非洲人泛基因组数据库后端的搭建过程中，我们选择用 Django 框架建立关系表格模型，并完成数据表的迁移，即在远端 MySQL 数据库中完成数据库的初步创建。非洲人泛基因组数据库网页使用 HTML，CSS 和 Javascript 进行搭建。我们搭建的非洲人泛基因组数据库中的首要功能是对 contig 的筛选和查找。数据库另一大特色在于基因浏览器可视化核苷酸序列。目前非洲人泛基因组数据库的构建还在起始阶段，尽管还需要更多非洲人泛基因组数据的补充以及开发更多有价值的功能，数据库的主体架构已经构建完成。我们整合的非洲人基因组序列数据库可以用于之后的基础研究之中，对数据进行差异分析并搭建数据库可以为后续工作研究提供帮助。

关键词：非洲人泛基因组，数据库，Django，基因组浏览器

DESIGN OF THE CONSTRUCTION OF AFRICAN PAN-GENOME DATABASE

ABSTRACT

Research on the human pan-genome began in 2009. Sequencing revealed that the genomes of two different people had a difference of 40 million base pairs, with 73-87 different genes. In addition to the previously recognized single nucleotide polymorphisms, insertion and deletion polymorphisms and structural variations in the human genome, there are population-specific and even individual unique DNA sequences and functional genes. The construction and exploration of African pan-genome database can help enrich the human reference genome and analyze the genetic differences between the two races, and provide more help for related research on pan-genomic sequences.

The data source of this database is basically from the Assembly of a pan-genome from deep sequencing of 910 humans of African descent. The data published on Nature and NCBI has a total of 910 African individuals and 125715 contigs. During the construction of the African pan-genomic database backend, we chose to use the Django framework to build the relational table model and complete the data table migration, that is, to complete the initial creation of the database in the remote MySQL database. The African Pan-Genome Database webpage is built using HTML, CSS and Javascript. The primary function of the African pan-genomic database we set up is to screen and find contigs. Another big feature of the database is that the gene browser visualizes nucleotide sequences. At present, the construction of the African pan-genomic database is still in its infancy. Although more African pan-genomic data is needed to supplement and develop more valuable functions, the main structure of the database has been completed. Our integrated African genomic sequence database can be used in basic research in the future. Differential analysis of the data and establishment of a database can help future work.

Key words: African pan-genome, database, Django, Genome Browser

目 录

第一章 非洲人泛基因组数据库的研究背景 1
1.1 泛基因组学研究现状 1
1.2 构建非洲人泛基因组数据库的意义 1

第二章 构建非洲人泛基因组数据库的准备工作 3
2.1 非洲人泛基因组原始数据来源 3
2.2 非洲人泛基因组数据库的设计构想 4

第三章 非洲人泛基因组数据库的开发过程 6
3.1 数据库后端的搭建 6
3.2 数据库前端的网页设计 7
3.3 数据库前后端的连接 7

第四章 非洲人泛基因组数据库的使用说明 9
4.1 筛选查询 contig 序列信息 9
4.2 基因组浏览器可视化功能 9

第五章 非洲人泛基因组数据库的讨论与展望 11

参考文献 13

致 谢 14

第一章　非洲人泛基因组数据库的研究背景

1.1　泛基因组学研究现状

泛基因组 (Pan-genome) 是某一物种全部基因的总称, 其中包括核心基因组 (core genome) 和非必需基因组 (dispensable genome)。其中核心基因组是指该物种所有个体中都存在的基因，一般与物种生物学功能和主要表型特征相关，反映了物种的稳定性；非必需基因组是指只在单个个体或部分个体中存在的基因，一般与物种对特定环境的适应性或特有的生物学特征相关，反映了物种的特性[1]。

鉴定非必需基因组是所有泛基因组研究的重点，因为这类片段代表了物种内基因资源的多样性，也很有可能是使个体产生不同性状（抗病性，抗寒性等）的原因。非必需基因组是基因组结构变异的一部分，但又不同于一般的结构变异，因为它必需在群体水平上展示极端的有和无的分布差异。泛基因组一般是通过对不同品种材料进行基因组测序，组装，将组装好的序列进行整合，从而获得这个物种全部的遗传信息，并对每个个体进行变异检测。

人类泛基因组研究于 2010 年首次进行，分析了非洲和亚洲的两个代表性基因组。在这项研究中，每个个体中检测到大约 5 Mb 的不存在于参考基因组（hg19）的新序列，并且参考基因组中的总缺失序列估计为 19-40 Mb。在随后的研究中，对中国人的 5 Mb 独特序列的重新分析表明 3.7 Mb 序列可以与 GRCh38 5 人参考基因组比对。但是在另一个中国基因组 HX1 中，共有 12.8 Mb 序列在 GRCh38 中未检测到，同时却在亚洲人群中发现了的这些新序列的 68%。

1.2　构建非洲人泛基因组数据库的意义

随着对人类基因组测序研究的广泛开展，测序个体数量的不断增加，科学家们发现，现有的人类基因组参考序列尚不够完整，特别是在一些特定的人群或个体基因组中被测序到现有人类基因组参考序列中缺失的片段。为了填补空白和纠正错误，人类基因组序列自从发表以来就进行了不断的改进。最新版本 GRCh38 包含 3.1 千兆（Gb），仅剩余 875 个 Gaps。

尽管如此，当前的人类参考基因组仍主要来自单个个体，因此限制了其在遗传研究中的作用，尤其是在混合种群（例如代表非洲散居人口的种群）中。近年来，越来越多的研究人员强调了捕获和代表来自不同种群的测序数据并将这些数据纳入参考基因组和基因组学研究的重要性。

因此，构建一个非洲人泛基因组数据库能够捕获非洲特有种群的遗传学信息，为某些疾病（尤其是某些特定种族特有的疾病）驱动基因突变以及非洲人对特殊环境适应性等研究建立基础。而且，非洲人泛基因组相较于以欧美白种人为

主要测序数据来源的人类参考基因组一定会存在很多遗传上的差异，完善的非洲人泛基因组数据库将会是对人类参考基因组的重要补充。

此外，非洲人泛基因组基因组数据库提供了基因组官网的功能，关心非洲人泛基因组的科学家可以通过访问网站了解相关信息、科研进展、做 BLAST 等简单分析、下载基因组数据，数据库中的基因组浏览器可以可视化查看基因结构、变异位点、基因表达等信息。

第二章　构建非洲人泛基因组数据库的准备工作

2.1　非洲人泛基因组原始数据来源

在最新的泛基因组研究中，Sherman 等人报道了一个非洲人的泛基因组[2]。它包含人类参考基因组中缺失的约 300 Mb 独特序列。值得注意的是，这些新序列中的大多数是具有个体特异性的，并且在两个或更多个个体中仅显示了 81 Mb 的共有独特序列。这些研究表明特定人群基因组多样性的重要性。这些非参考基因组区域可能是某些疾病（尤其是某些特定种族特有的疾病）的驱动基因突变，值得我们进行研究。

Sherman 等人使用了由 910 个非洲人后裔组成的深度测序数据集，构建了一组存在于这些个体中但在参考人类基因组中缺失的 DNA 序列。将 910 个个体中的 1.19 万亿个读段与参考基因组（GRCh38）进行了比对，收集了所有未能比对的读段，并将这些读段组装成连续的序列（contigs）。然后将所有重叠群相互比较，发现了目前参考基因组未包含的 302 个独特的 contig，这些序列代表表了参考基因组中缺少的非洲泛基因组区域。此外，这些 contigs 中还存在 1246 个一端能够比对上参考基因组，同样具有进一步研究的潜力，contigs 的分类信息可以参考图2–1。

Table 1 | Novel sequences in the African pan-genome

	Number of sequence contigs	Total length (bp)	Bases with no alignment to GRCh38 (<80% identity)	Longest contig (bp)
Two ends placed	302	667,668	431,656	20,732
One end placed	1,246	3,687,028	1,866,699	79,938
Unplaced	124,167	292,130,588	202,629,979	152,806
Total	125,715	296,485,284	204,928,334	152,806
Non-private only	33,599	80,098,092	50,044,650	152,806

Number and length of novel sequences in the African pan-genome. Bases with no alignment

图 2–1 非洲人泛基因组中的独特 contig 信息

本数据库的数据来源基本来自于Assembly of a pan-genome from deep sequencing of 910 humans of African descent公布在 Nature 与 NCBI 上的数据总共有 910 名非洲个体，125715 余条 contig，其中：contig 与人类个体之间的关系表格为：Supplementary Data 1，contig 位置和基因关系的表格为：Supplementary Tables，contig 序列的 fasta 文件来源于：PDBU01，基因功能则使用 NCBI 的查询功能对基因进行注释。

我们初步根据双端和单端比对上参考基因组上染色体的 1548 个 contig 的信息，包括 contig 的碱基序列，比对上的染色体位置，可以组装成的 scaffold 以及可

能对应的基因名称等，建立关系型数据库，可以即时查找符合搜索条件的 contig 信息。

2.2 非洲人泛基因组数据库的设计构想

我们已知组装得到的 302 个 contig 的两端都能比对上参考基因组，也就能够直接补充人类参考基因组中缺失的部分非洲人泛基因组，但是剩下的 1246 个单端比对上参考基因组的 contig 可能与非洲人族群的进化关系和独特的环境适应性更加密切相关。因此，研究者可能更希望分别研究分析这两类 contig 携带的遗传变异信息和功能。我们首先设计了一张数据表来存放这 1548 条 contig 的信息，将这些记录分为双端比对上和单端比对上两类，此外还在数据表中整合 lNCBI 上这些 contig 的序列长度、完整碱基序列等。在先前 Sherman 对非洲人泛基因组的研究中，特别记录了每个 contig 在 910 个非洲人个体中的出现次数，出现频次可能决定了一些 contig 上包含的变异信息能代表整个非洲人族群的遗传学特征，还是仅仅几个个体基因变异的结果，所以我们也将 910 个非洲人的基因组中包含每个 contig 的个体数目加入到数据表中。

我们进一步考察这 1548 条 contig 比对上人类参考基因组的染色体位置，包括起始和终止位点等。在基因组测序过程中，往往会通过 pair-end 的方法将 contig 按顺序排列组装为不同的 scaffold。为了研究这些 contigs 的相对位置关系，我们可以将它们和已知的 scaffold 进行序列比对，根据比对的 coverage 和 identity 筛选比对结果最好的 contig 对应到 scaffold 上的位置。此外，还有一部分的 contig 可能对应着已知基因，因此我们将这些基因的完整名称、概述和在 NCBI 上的链接补充进数据表中。

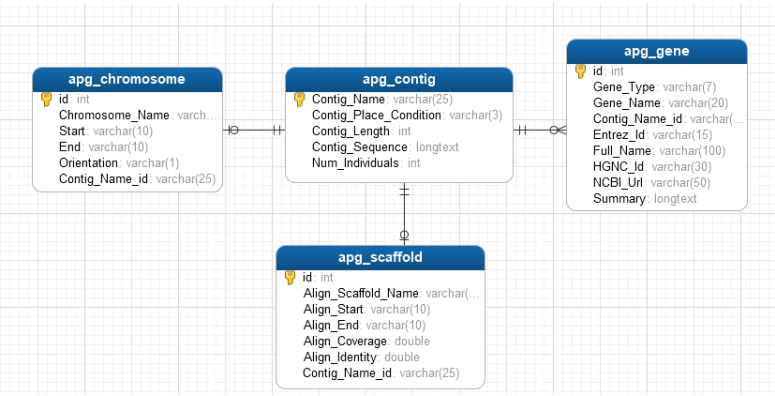

图 2–2 非洲人泛基因组数据库的 ER 模型

根据我们对非洲人泛基因组数据库的需求分析，设计了以下几个实体及其属性，通过 contig 的名称连接相关染色体、scaffold 和基因。实体型与其包含的属性有：

1. contig：contig 名称、配对情况、contig 长度、fasta 序列、在 910 个个体中的出现次数
2. 染色体：染色体名称、起始位点、终止位点
3. Scaffold：Scaffold ID、起始位点、终止位点、比对 coverage、比对 identity
4. 基因：基因名称、基因功能、基因类型、NCBI 链接

我们根据实体间的关系完成了 ER 模型的设计，在图2–2中展示了非洲人泛基因组数据库的 ER 图。

第三章　非洲人泛基因组数据库的开发过程

3.1　数据库后端的搭建

在非洲人泛基因组数据库后端的搭建过程中，我们选择用 Django 框架建立关系表格的模型，通过继承 Django 中的 Model 类定义每一种属性的类型、名称、是否为主键等以及每一类之间的映射关系。接着，通过 python managy.py migrate 命令完成数据表的迁移，即在远端 mysql 数据库中完成数据库的初步创建。数据表模型定义的示例如图3–1所示：

```
class Contig(models.Model):
    Contig_Name = models.CharField(max_length=25,primary_key=True,default=0)
    Place_Condition = (
        ('Two','TwoEndPlaced'),
        ('One','OneEndPlaced'),
    )
    Contig_Place_Condition = models.CharField(max_length=3,choices=Place_Condition)
    Contig_Length = models.IntegerField(default=0)
    Contig_Sequence = models.TextField()
    Num_Individuals = models.IntegerField(default=0)
    def __str__(self):
        return self.Contig_Name
```

图 3–1 非洲人泛基因组数据库模型定义的示例

我们选择 Navicat for MySQL 软件连接远端 mysql 数据库，在可视化界面中将整理完毕的数据表格以.csv 文件的形式导入记录到数据库中。后端搭建好的数据表格如图3–2,3–3,3–4,3–5所示：

Contig_Name	Contig_Place_Condition	Contig_Length	Contig_Sequence	Num_Individuals
CAAPA_1	Two	2495	TTCACAGCTAAGCTCC	450
CAAPA_10	Two	2893	GTACTTCTATCCATAGT	764
CAAPA_100	Two	2429	TATGCAAGGGGGATTC	782
CAAPA_1000	One	2888	GAAGACTGGGACCCGC	85
CAAPA_1001	One	2175	TATATATAAAATATACC	295
CAAPA_1002	One	1420	AGTGTATATTTTGTGCA	271
CAAPA_1003	One	1076	TTTTATCCAGAAAGAAA	258
CAAPA_1004	One	7025	CTCTGGTAAATAATTTA	362
CAAPA_1005	One	5897	GGACCTATGCTCTTTCA	121
CAAPA_1006	One	2334	AAAAAAAAAAAAAAAA	267
CAAPA_1007	One	935	AAATACAAAAAATTAGC	2

图 3–2 Contig 信息数据

id	Chromosome_Name	Start	End	Orientation	Contig_Name_id
1	chr7	141502243	141502243	+	CAAPA_1
2	chr11	42721628	42721637	-	CAAPA_2
3	chr6	166285511	166285511	+	CAAPA_3
4	chr3	95825553	95825557	-	CAAPA_4
5	chr4	30540783	30540794	-	CAAPA_5
6	chr4	6034169	6034169	-	CAAPA_6
7	chr2	26714469	26714469	+	CAAPA_7
8	chr13	89467693	89467766	+	CAAPA_8
9	chr5	32974398	32974467	-	CAAPA_9
10	chr6	79901555	79901555	+	CAAPA_10

图 3–3 染色体上位信息

id	Align_Scaffold_Name	Align_Start	Align_End	Align_Coverage	Align_Identity	Contig_Name_id
1	NC_000001.11	40332916	40333551	59.27	86.64	CAAPA_548
2	NC_000002.12	94734852	94737984	100	82.38	CAAPA_1014
3	NC_000002.12	94734833	94737660	100	82.18	CAAPA_1128
4	NC_000002.12	89799524	89800712	96.89	83.01	CAAPA_1151
5	NC_000002.12	94734974	94735997	100	84.38	CAAPA_1280
6	NC_000002.12	94734852	94738566	99.73	82.31	CAAPA_473
7	NC_000002.12	94734782	94737663	99.79	83.07	CAAPA_558
8	NC_000002.12	94734833	94737658	100	82.17	CAAPA_566
9	NC_000002.12	233787540	233788130	57.57	82.4	CAAPA_652
10	NC_000002.12	194463753	194464308	55.07	80.04	CAAPA_655

图 3–4 Contig 对应的的 Scaffold

id	Gene_Type	Gene_Name	Contig_Name_id	Entrez_Id	Full_Name	HGNC_Id	NCBI_Url	Summary
1	exon	MOGAT2	CAAPA_12	80168	monoacylglyce	23248	https://www.n	"The protein e
2	mRNA	MOGAT2	CAAPA_12	80168	monoacylglyce	23248	https://www.n	"The protein e
3	exon	LOC105372529	CAAPA_60	105372529	uncharacterize	(Null)	https://www.n	(Null)
4	lnc_RNA	LOC105372529	CAAPA_60	105372529	uncharacterize	(Null)	https://www.n	(Null)
5	exon	LOC107986294	CAAPA_89	107986294	uncharacterize	(Null)	https://www.n	(Null)
6	lnc_RNA	LOC107986294	CAAPA_89	107986294	uncharacterize	(Null)	https://www.n	(Null)
7	exon	LOC107985852	CAAPA_120	107985852	uncharacterize	(Null)	https://www.n	(Null)
8	lnc_RNA	LOC107985852	CAAPA_120	107985852	uncharacterize	(Null)	https://www.n	(Null)
9	exon	UBE2QL1	CAAPA_172	134111	ubiquitin conju	37269	https://www.n	
10	mRNA	UBE2QL1	CAAPA_172	134111	ubiquitin conju	37269	https://www.n	

图 3–5 Contig 对应的的基因信息

因为我们选用了 Django 的框架，为了让 Django 能够顺利操作数据库，使用 python manage.py inspectdb 的命令将已经建立完成的数据库迁移到 mysql 的模型中，完成了 Django 后端和数据库的连接。

3.2 数据库前端的网页设计

非洲人泛基因组数据库网页使用 HTML，CSS，Javascript，Jquery 进行搭建。其中 HTML，CSS 用于制作静态的页面，而 Javascript 和 Jquery 则用于页面的交互效果。另外，我们使用了简洁美观的 Bootstrap 作为我们的前端框架。

整体网页设计采用极简风格，前端主要显示界面整体为黑白色调，同时为了增添网页活力，我们在网页左下角额外设置 live2D 模型，增加用户访问过程中的趣味性。图展示了非洲人泛基因组数据网站首页。

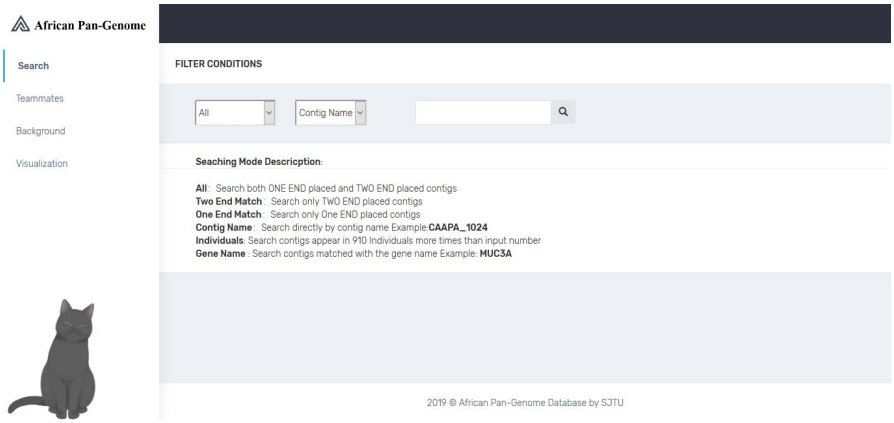

图 3–6 非洲人泛基因组数据网站首页

在 Search 界面，我们设置了选项框、输入框，以及使用方法介绍。用户按照指定形式输入后，点击查询按钮，即可得到搜索结果。在 Teammates 界面，放有小组成员的头像和邮箱，如果用户在使用过程中有任何问题或建议可以与我们联系。为了让用户更好地了解我们的项目背景和意义，我们设置了 Background 背景介绍页面，简要介绍千人基因组计划和 contig ID 的含义，帮助之前不了解非洲人泛基因组项目的用户建立对项目的印象与理解，实现用户友好。 除了基本的查询功能外，我们还开发了可视化功能，参考 IGV 基因组浏览器模板[3]，使用基因组浏览器来演示非洲人泛基因组序列。用户可以选择不同的染色体，搜索框中可输入自己设定的染色体起始位点和终止位点。对应的基因图谱会在下方显示，用户还可以通过不同按键选择是否显示游标导航、中间线、追踪标签等，同时也可以调节显示基因段的大小，并以.svg 格式保存自己设定得到的序列可视化结果。

3.3 数据库前后端的连接

非洲人泛基因组数据库的构建基于 Django 的框架。Django 是一个全栈框架，可以包揽从前端到后端的全部流程。在搭建好前端页面和后端的数据库之后，我们需要把前后端连接起来，也就是说，当前端发送请求时，将请求发送到后端的数据库进行处理，再将处理后的结果返还到前端，呈现给用户。

Django 的核心——MTV 模型为前后端的顺畅连接提供了方便高效的实现方法。MTV 模型将 web 服务的流程划分为三个部分：M 即 Model，数据存取层，处理与数据相关的所有事务;T 即 Template，即表现层，决定如何在页面或其他类型文档中进行显示;V 即 View，视图层，存取模型及调取恰当模板的相关逻辑，是模型与模板的桥梁。我们的前后端连接，主要就是在处理视图层的问题。

在实际使用中，我们在 views.py 文件中建立了视图层的函数。这些函数会在用户输入相应的 url 时，将所需要的网页文件发还给用户。其中最复杂的是处理数据搜索的函数，这个函数会读取通过 get 方式发送的表单信息，包括搜索的类型，双端匹配或者单端匹配等，通过这些信息构建相应的 mysql 搜索命令，获得符合条件的数据记录，向前端发还这些记录。而在 Template 层，也有相应的模板引擎来处理视图层发来的数据。模板引擎是将 View 视图中需要在前端 HTML 页面中展示的数据，通过模板引擎的语法规则（类似于 python）进行业务逻辑的处理。

第四章　非洲人泛基因组数据库的使用说明

4.1　筛选查询 contig 序列信息

用户可以输入 http://106.15.45.44:8000，进入非洲人泛基因组数据库的网站 African Pan Genome。

我们搭建的非洲人泛基因组数据库中的首要功能是对 contig 的筛选和查找功能，查找分为以下三种模式，对应下拉列表中的 Contig Name, Individuals, Gene Name 栏目。在 Contig Name 模式下可以根据 contig 的名称（例如 CAAPA_1024）来搜索对应的序列信息; 在 Individuals 模式下，可以搜索在 910 个个体中出现次数超过输入数字的 contig 信息；在 Gene Name 模式下，可以查找输入基因名称（例如 MUC3A）对应的 contig 信息。此外，在我们的筛选中还可以区分 contig 是双端还是一端比对上参考基因组，由另一个下拉列表中的 Two 和 One 选项实现。查找结果的界面如图所示：

Contig Name	Contig Place Condition	Contig Sequence
CAAPA_316	One	ATTGGGGTTGGTAAGCTACTGATATCAGT...

图 4–1 根据基因名称进行查找的结果界面

查找完毕后，用户可以看到满足筛选条件的 contig 的名称、比对情况（双端或一端），点击想要查看的 contig 名称后得到的界面中包括该条序列的详细信息，包括长度，完整序列，在 910 个样本中的出现次数以及可能组装上的 scaffold 信息、比对上的参考基因组染色体和对应的基因信息（包括 NCBI 上的基因链接）。

4.2　基因组浏览器可视化功能

非洲人泛基因组数据库中另一大重要功能是基因浏览器可视化核苷酸序列。本数据库以 hg38 为参考基因组构建可视化页面。用户可以在下拉列表中勾选不同的染色体，在搜索框中可输入自己设定的染色体起始位点和终止位点。用户还可以通过不同按键选择是否显示游标导航、中间线、追踪标签等，对应按键为 Cursor Guide，Center Line，Track Labels。用户可以调节滑动控制器改变显示基因段的大小，并通过 Save SVG 按键以.svg 格式保存自己设定得到的序列可视化结果，可视化结果见图4–2。

基因组浏览器中若将显示模式调至最大，可以看到两条 Center Line，其中包含的碱基即为当前阅览的中心。Cursor Guide 随光标移动，可以辅助对齐。左上角的字符是样本编号，下方 Refseq Genes 是对参考基因组的注释。用户点击右方

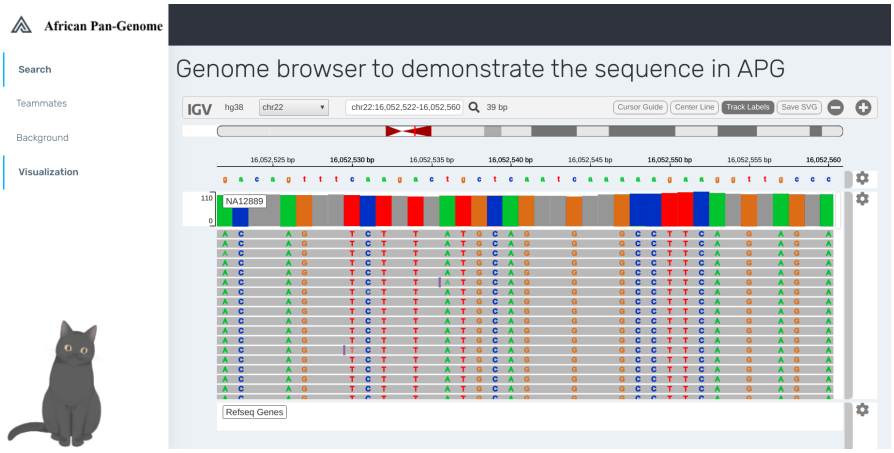

图 4–2 非洲人泛基因组可视化浏览器界面

设置选项，可以有不同的功能，以 Three-frame Translate 为例，它会显示碱基翻译出的氨基酸序列。

第五章　非洲人泛基因组数据库的讨论与展望

目前非洲人泛基因组数据库的构建还在起始阶段，尽管还需要更多非洲人泛基因组数据的补充以及开发更多有价值的功能，数据库的主体架构已经构建完成。非洲人泛基因组数据库的功能基于对满足不同条件的 contig 的筛选，将这些 contig 包含的重要遗传变异信息展示给用户，包括 contig 的序列信息，与从参考基因组的比对结果（染色体上的位置，比对上的 scaffold 以及对应的基因信息）。除了完善的 contig 相关数据的展示结果以外，非洲人泛基因组数据库通过基因组浏览器的功能能够对基因组碱基组成、基因结构、SNP 等变异信息、基因表达量等进行可视化的生动展示。

我们整合的非洲人基因组序列数据库可以用于之后的基础研究之中，对数据进行差异分析并搭建数据库可以为后续工作研究提供帮助。研究人员在获得无法在参考基因组上定位的序列时可以在我们的泛基因组数据库中得到较为准确的定位，开展相关的富集分析与功能分析。用户对数据库中的独特基因进行功能分析，可以帮助实验人员更有针对性地进行调查采样，进一步研究得到非洲人适应性表现和地域、生活习惯上的关系。

当然，目前非洲人泛基因组数据库中具有重大研究价值的 contig 数目偏少，需要收集更多的非洲人泛基因组数据完善数据库的内容。初步收集的泛基因组数据中心存在 12 万多个两端未匹配到参考基因组上的 contig，它们往往只比对到染色体上很短的片段或者两端比对到了不同的两条染色体上，对于这部分的 contig，未来非洲人泛基因组数据库中将会整合初步分析处理后的这部分泛基因数据，展示给研究者进行更进一步的功能分析。

我们目前的数据库还有待改进，最终希望能实现对数据格式的自动检测，实现快速扩充。我们希望能够分析这些变异位点的次要等位基因的数量与频率的关系及在编码区和非编码区的情况。并对所研究的人类个体构建了系统进化树了解了个体的分类情况。并基于对未必对上的基因进行同源分析，预测他们的功能。未来我们计划引入更多的功能提高非洲人泛基因组数据库的实用性，提供用户上传新的非洲人泛基因组 contig 的接口口，调用 BLAST 等生物信息学分析工具的 API，对 contig 进行与参考基因组和已知 scaffold 的比对，并实现可能对应基因的注释。未来非洲人泛基因组数据库的构建设想中还包含对已有的对应不同功能的 contig 进行多序列比对，可视化系统发生树，直观地表现这些遗传变异特征的进化关系。图5–1简要描述了今后非洲人泛基因组分析流程的设想。

对人类泛基因组的研究有着相当广阔的前景，不同族群体现出不同性状的遗传进化分析依赖于对泛基因组的深入研究。我们相信非洲人泛基因组数据库能够为广大研究者提供研究分析的平台，整合常用的分析工具，提供下载上传渠

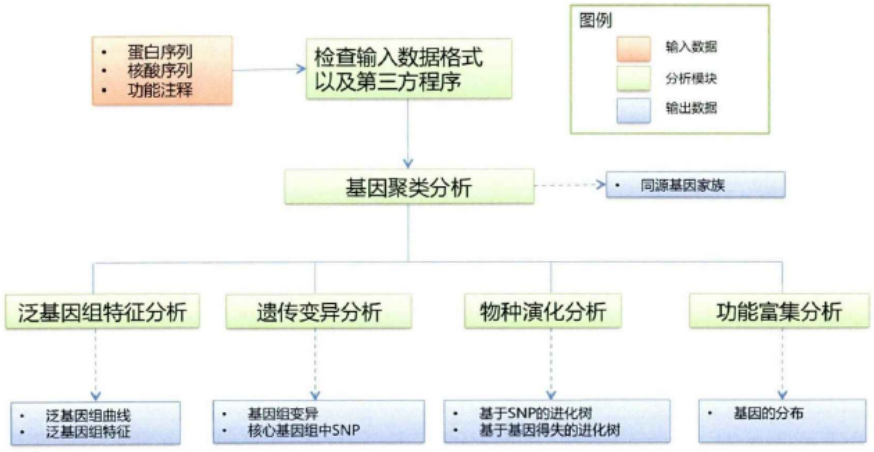

图 5–1 非洲人泛基因组分析流程示意图

道，助力非洲人泛基因组研究的不断发展。用户如果有关于数据库数据内容和功能设计方面的建议，非常欢迎通过邮件联系我们，一同完善非洲人泛基因组数据库的建设。

参考文献

[1] Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial p̈an-genomë[J]. Proc Natl Acad Sci U S A, 102(39): 13950-13955.

[2] Assembly of a pan-genome from deep sequencing of 910 humans of African descent[J]. Nature genetics, 2019, 51(1): 30.

[3] THORVALDSDOTTIR H, ROBINSON J T, MESIROV J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration[J]. Briefings in Bioinformatics, 14(2): 178-192.

致　谢

综上，我们初步了解了数据库的实现方式和常用数据库操作语言原理及使用特点，意识到了服务器处理大量复杂信息时数据库的强大之处。在一学期的数据库原理学习过程中，我们深深感受到了 ER 图的提纲挈领，mysql 特有的方便与快捷，被他们的魅力所折服。团队成员在完成大作业的过程中用到了平时所学，同时自学新知识，丰富充实自己，很期待在未来的研究和学习生活中大展身手。学习永无止境，在接下来的学习和生活中我们将会谨记老师的教诲。

感谢王靖方老师一个学期以来的悉心指导！

感谢组员同学们共同的努力与付出！

感谢大家对我们初步搭建的非洲人泛基因组数据库提供的所有有价值的意见和建议！