DOI:10.12113/202106002

面向大规模人群的基因组注释系统

闫贞磊,国宏哲*
(哈尔滨工业大学 计算学部,哈尔滨 150001)

摘 要:基因组注释是识别出基因组序列中功能组件的过程,其可以直接对序列赋予生物学意义,由此方便研究者探究和分析基因组功能。基因组注释可以帮助研究从三个层次上理解基因组,一种是在核苷酸水平的注释,主要确定DNA序列中基因、RNA、重复序列等组件的物理位置,包括转录起始,翻译起始,外显子边界等具体位置信息。同时可以注释得到变异在不同人群中的变异频率差异,这是解读不同人群表型差异的图谱基础。第二种是蛋白水平的注释,主要解读基因或变异的可能功能异常,评估变异所在基因位置、变异类型等对蛋白质改变的影响。第三种是生物学功能/过程注释,主要解读不同基因相互作用的对生物学过程和通路的影响,可以从系统生物学角度解释基因或调控元件对生命生化过程或功能的影响。自从人类基因组计划完成之后,各国陆续启动了基因组测序计划,完成绘制了人类基因详尽的基因多态性谱图,记录了不同表型群体的变异分布和频率差异情况等注释信息。我们结合已有的注释数据库知识,开发了具有高准确性和高效的面向大规模人群的基因组注释系统,实现对大规模的人群变异数据进行全自动化的功能性注释分析计算,进一步助力未来人群遗传变异分布等方面的研究。

关键词:人群基因组;基因组注释;基因组变异
中图分类号:Q343.1 **文献标志码**:A **文章编号**:1672-5565(2022)01-011-09

Workflow of large-scale population genome annotation

YAN Zhenlei, GUO Hongzhe*
(*Faculty of Computing*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: Genome annotation is the process of identifying functional components in a genome sequence, which can directly assign biological significance to the sequence, thus facilitating the exploration and analsis of genome functions. Genome annotation can help researchers understand the genome from three levels. The first is the annotation at the nucleotide level, which mainly determines the physical locations of genes, RNA, repetitive sequences. and other components in the DNA sequence, including specific location information such as transcription initiation, translation initiation, and exon boundaries. At the same time, the difference in the variation frequency of the variation in different populations can be obtained via annotation, which is the basis for the interpretation of the phenotypic differences of different populations. The second is the annotation at the protein level, which mainly interprets the possible functional abnormalities of genes or mutations, and evaluates the impact of the location of the mutations in the genes and the types of mutations on protein changes. The third type is biological function/process annotation, which mainly interprets the impact of different gene interactions on biological processes and pathways, and can explain the impact of genes or regulatory elements on life biochemical processes or functions from the perspective of systems biology. Since the completion of the Human Genome Project, countries across the world have successively launched genome sequencing projects, completed detailed gene polymorphism maps of human genes, and recorded annotation information such as variation distribution and frequency differences of different phenotypic groups. Based on the existing knowledge of annotation databases, a highly accurate and efficient genome annotation system for large-scale populations was developed, realizing fully automated functional annotation analysis and calculation of large-scale population variation data, and further assisting future research on variation distribution of population genetics and other aspects.

Keywords: Population genomes; Genome annotation; Genome variation

收稿日期:2021-06-02; **修回日期**:2021-06-23.
资助项目:国家重点研发项目(No. 2017YFC0907503).
作者简介:闫贞磊,男,研究方向:生物信息学. E-mail:1226685740@qq.com.
*通信作者:国宏哲,男,助理教授,研究方向:生物信息学. E-mail:hzguo@hit.edu.cn.

1 材料与方法

1.1 数据下载

注释数据来源于多个公共数据库,下载的数据经过格式及参考基因组版本转换以备注释使用。

1.1.1 Annovar 整合数据下载

Annovar 数据下载方式为:annotate_variation.pl -buildver hg38 -downdb -webfrom annovar dataname humandb/。从公共数据库 annovar[1] 下载了包含人群变异频率数据、疾病变异数据、祖源数据、SNP 数据、全基因组关联分析数据、保守性数据、变异位置及功能等数据。

1.1.2 CADD 数据下载

CADD 数据需要下载多版本数据,目前使用的是 1.6(hg38/hg19),注意下载的数据是拆分了 snv 和 indel 的,都要下载。

从公共数据库 CADD[2] 下载了包含 bscores、chromhmm、dbscSNV、encode、encodeMotifDB、exons、gerp、grantham、mirSVR、mirTargetScan _ v71、mmsplice、mutation _ density、phastCons、phyloP、reference、regulatoryBuilt、Remap、spliceai、transcripts、vep 等数据。

1.1.3 UCSC 数据库数据下载

UCSC 数据库使用 FTP 下载方式在 HG38 goldpath 路径下载数据,同时也通过 Tablebrowser 下载数据。

从 UCSC[3] 数据库下载的数据分为 phastCons20way placental、phastCons100way vertebrate、phyloP20way placental、phyloP100way vertebrate、Recombinate Rate、repeatmasker、TAD 等。

1.1.4 Fantom 5 数据库

CAGE Promoters 和 CAGE Enhancers 下载于 Fantom 5[4] 数据库,直接选择 hg38 版本的 F5.hg38.enhancers.bed.gz、hg38 _ fair + new _ CAGE _ peaks _ phase1and2.bed.gz 文件。

1.1.5 Ensembl Regulatory 数据

Ensembl Regulatory Build 下载于 Ensembl [5] 数据库,提供的链接可能直接复制下载会失效,可以按照链接提供的目录自己逐级查找,下载的文件为 homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20190329.gff.

1.1.6 ORegAnno3 下载(2016)

ORegAnno3[6] 存储为调控元件信息数据库,数据处理后保留所有位置信息及调控原件信息内容。

1.1.7 LINSIGHT 得分

LINSIGHT[7] 是一个统计模型,用于估计人类基因组中非编码序列的阴性选择。LINSIGHT 得分衡量了非编码位点上的阴性选择概率,可用于对与遗传疾病相关的 SNVs 进行优先排序,或量化调控序列(如增强子或启动子)的进化约束。

1.1.8 miRNA 数据

miRNA 信息来源于 miRBase r21[8] and snoRNABase v3[9]。

1.1.9 GenCode 数据

GenCode[10] 数据包含基因位置、名称、转录本、类型等信息。

1.2 数据格式及参考基因组版本转换

所有已经下载的数据参考基因组版本为 hg38 的数据不需要进行参考基因组版本转换,将其处理成注释可用格式即可。参考基因组为 HG19 的数据需要进行数据版本转换,版本及格式转换完成后的数据处理为 varnote 可用的知识库,以供 varnote 工具对变异数据进行注释。

1.2.1 liftover 等工具进行版本转换

liftover 的获取地址:wget http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver。同时,获取坐标注释文件 hg38Tohg19 和 hg19Tohg38,且获取命令如下:

wget http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz

wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz (下载得到的 * over.chain.gz 不需要解压)其他版本的注释文件可在 UCSC 数据库查找。

Input 文件 BED 格式文件,BED 格式文件只定义前三列:chr start end,无表头(注:start 不等于 end)。坐标转换命令:liftover inputfile,over.chain.gz,outputfile,unmapfile。

1.2.2 其他版本转换工具

部分文件转换版本的时候用的是 CrossMap.py[11],尤其是 hg18 及更老版本转换到 hg38 的版本的时候,可以实现一步转换。CrossMap.py bed * _to_ *.chain.gz input.bed out.bed。

1.2.3 bw 文件处理

bw 文件可以用于 UCSC 及 IGV 等可视化工具展示,但是注释的时候不适合使用,要转换问 bed 文件才能进行注释。bw 文件转换使用的是 UCSC 提供的 bigWigToBedGraph[12] 工具,运行命令为:bigWigToBedGraph file.bw out.bedGraph(如果是 HG19 的数据按照上面版本转换方法进行版本转换)。

1.2.4　VCF 数据预处理

所有.vcf 格式变异均处理为 avinput(参见 annovar avinput 处理)格式,知识库及得到的变异数据文件都要进行相应处理,vcf 格式知识库数据还包括 GWAS、GRASP 和 CADD 等文件。

1.3　数据注释

1.3.1　流程结构(见图 1)

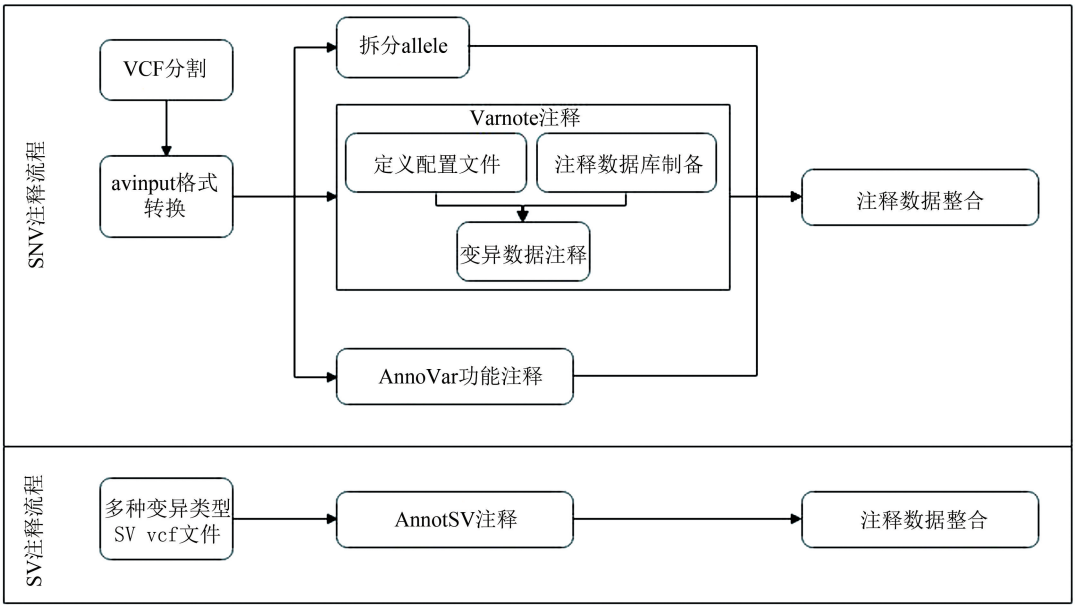

图 1　基因组注释流程图

Fig.1　Flow chart of genome annotation

1.3.2　注释步骤

转换 avinput:avinput 文件是 annovar 的标准输入文件,其自带数据库均依据该文件中 ref/alt 的格式建立,当前注释流程的依赖数据库中包含了很多 annovar 自带的数据库,且无法通过 avinput 的 ref/alt 格式逆推回原本的格式。同时,avinput 文件定义的 ref/alt 格式可以有效地将整合后的 vcf 文件中的 ref/alt 标准化,如原本的 ref/alt 为 C/T,但整合后有可能会变为 AC/AT,显然整合后的 ref/alt 并不能与数据库中的信息匹配,故标准化是保证注释命中的关键步骤。由于 annovar 自带的标准化过程限定条件太多,部分不满足要求的行会被自动过滤掉,且这一过程无法通过参数进行调节,故这一步骤采用了改写 annovar 代码的方式,自行编写出了 python 版本的转换代码。

拆分同位置上多 allele:由于该注释流程获得的上游输入文件是多个样本整合后的 vcf,而整个项目最终会需要将多次收到的输入文件进行整体整合。对于单个收到的 vcf 来说,其中样本的基因型是依照当前的 vcf 给出的,这一基因型会与最终整合的 vcf 不匹配。同时,转换后的 avinput 每一个 allele 都是单独一行,若不进行拆分则不利于后续的整合步骤。

1.3.3　annovar 注释 refGene

refGene 是基于基因的注释,varnote 无法注释该类型的数据,故采用 annovar 对该数据库进行注释。

1.3.4　varnote 注释

生成 varnote 配置文件:基于注释框架中指定的注释库、项目、输出项目名自动化生成 varnote 使用的配置文件。

通过 avinput 建立数据库:varnote 进行注释时,对于完全注释不到的行,不会在结果文件中进行输出,这样的输出结果不符合实际需求以及后续整合的需要,且这一行为无法更改。故将转换后的 avinput 文件建立为一个数据库,通过自己注释自己的方式,使得每一行信息都能得到注释,从而保证结果文件中不会丢行。

注释:通过读取流程框架中配置文件的数据库信息、参数信息拼接出 varnote 的实际执行语句进行调用注释。

1.4　整合注释结果

1.4.1　去除无用数据列

无用数据列包括原始 vcf 文件中包含的部分信息、avinput 数据库产生的额外信息等。在最后的整合过程中予以剔除。

1.4.2 重新对数据列排序

根据配置文件中指定的数据列排列顺序,将注释结果的信息列进行重新排序。

1.4.3 整合 refGene 注释结果

将 annovar 单独注释出的 refGene 信息整合到最终结果文件中。

1.4.4 整合样本基因型

将拆分后的多 allele 样本基因型合并到注释结果中,形成最终结果文件。

1.5 AnnotSV 注释

1.5.1 注释各类型 SV 变异

结构变异注释的输入文件是根据 SV 类型产生的多个 vcf 文件,部分类型的文件可能不存在或内容为空。注释时给定的输入文件参数并非是实际的文件名,而是前缀名,经过拼接后自动查找目录中是否有相应类型的文件,逐个调用 annotSV 进行注释。

1.5.2 整合 SV 注释结果

由于输入文件多个文件,产生的结果文件同样是多个,且无法保证每种 SV 类型的结果文件都存在。对各个类型的结果文件进行检验,查看其是否存在,将存在的文件进行合并拼接,得到最后的完整 SV 注释结果。

2 结果与分析

2.1 注释数据统计

2.1.1 人群变异频率数据

为了进行注释收集整理了多类变异频率数据,其中包含人群变异频率数据,例如:the Exome Aggregation Consortium(外显子组整合数据库)、Genome Aggregation Database[13](简称 gnomAD)是由各国研究者联合发展起来的基因组突变频率数据库,汇集众多大规模测序计划的基因组数据、NHLBI 计划汇集的外显子组突变频率数据、1000G计划汇集的全基因组突变频率数据、KAVIAR[14] 整合了 HGP、1000G、CGI69、UK10K 等共 35 个计划的基因突变频率数据、HRCR 整合了 32 k 个样本整合的基因组突变频率数据、巴西人群外显子组突变频率数据[15]、中东地区人群外显子组变异频率数据、以及收集整理的中国人群变异频率数据。我们对人群变异频率数据的总数据量、SNP/INDEL 数量、稀有突变、低频突变、高频突变的数量进行了统计,各人群数据库内稀有变异比例最高,SNP 数量大于 INDEL 数据量,位点数量最多的数据库为 gnomad,近 3 亿个变异位点,其他全基因组数据库包含变异位点数量大多小于 1 亿个。统计结果(见图 2)。

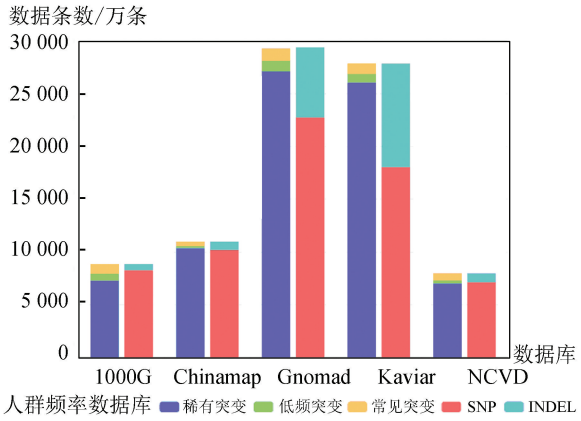

图 2 人群变异频率数据

Fig.2 Illustration of frequency of population genome variation

2.1.2 临床及表型注释数据

为了了解变异是否是已知的致病变异我们整理了多个 DNA 变异和人类疾病关系知识库,包括:60 种人类肿瘤细胞系外显子组变异频率数据、预测内含子单核苷酸变异的致病性影响数据库、NCBI 维护的疾病相关的人类基因组变异数据库、EBI 负责维护的一个收集已发表的 GWAS 研究的数据库[16]、GRASP 2.0[17](SNP 和表型关联数据库)、癌症基因突变数据库[18],此外还有基于美国医学遗传学与基因组学学会(The American College of Medical Genetics and Genomics,ACMG)曾制定过序列变异解读指南而分析变异数据的数据库-InterVar[19]。我们统计了这些疾病及表型相关的数据库内的变异数量及分类情况,大多数变异数据和临床关系为"未明确"或者"可能致病/良性",明确的"良性"及"致病性"位点比例较低,InterVar 和 clinvar 统计结果(见图 3 和图 4)。

2.1.3 序列保守性数据库

收集整理了多个评估序列保守性的数据库以及对相应保守性的得分信息,比如灵长类动物基因组序列 PhyloP[20] 方法保守性打分数据、哺乳动物基因组序列 PhyloP 方法保守性打分数据、脊椎动物基因组序列 PhyloP 方法保守性打分数据、灵长类动物基因组序列 PhastCons 方法保守性打分数据、哺乳动物基因组序列 PhastCons 方法保守性打分数据、脊椎动物基因组序列 PhastCons 方法保守性打分数据、由 GERP++[21] 方法计算得到的 GerpN 值等,同时还下载整理了祖源等位基因情况。以上的保守性打分数据对基因组上所有的位点进行打分,每个位点都有相应的分值。

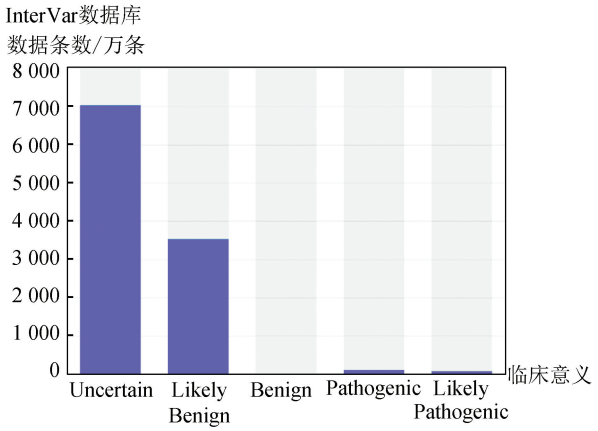

图 3　InterVar 临床变异数据统计

Fig.3　Statistics of clinical variation based on the InterVar database

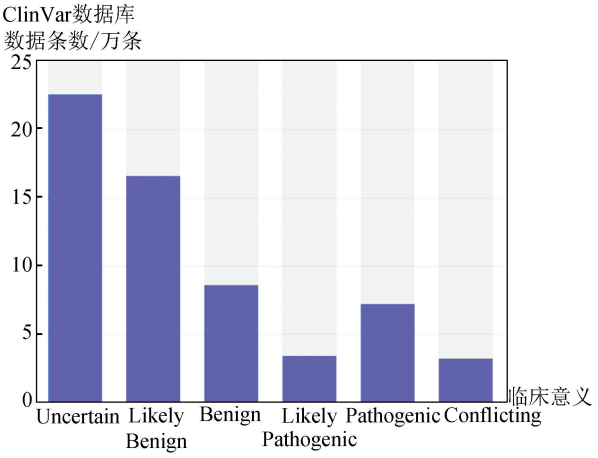

图 4　Clinvar 临床变异数据统计

Fig.4　Statistics of clinical variation based on the clinvar database

2.1.4　表观注释数据

为了对变异区域进行表观注释,我们下载了多个相关的数据库,其中包括识别 promoter 和 enhancer 区域的数据、细胞中染色质标记状态比例数据、预测得到的转录因子 motif 数量的数据、与转录起始和终止位置的距离数据、表观遗传学标记和检测相应转录调控原件的数据、染色质拓扑结构域识别的数据等。

2.1.5　变异综合得分

为确定每个变异的功能和致病性风险程度,多个数据库基于多种类的注释结果对变异数据进行了综合打分,这些数据库包括了 LINSIGHT 得分,LINSIGHT 用于估计人类基因组中非编码序列的阴性选择。LINSIGHT 得分衡量了非编码位点上的阴性选择概率,可用于对与遗传疾病相关的 SNVs 进行优先排序,或量化调控序列(如增强子或启动子)的进化约束。还包含了在编码区和非编码区对遗传变异进行注释的 Eigen[22] 得分、CADD 得分、fitcons[23] 得分等。各数据库包含位点数量(见表 1)。

表 1　变异综合得分总数

Table 1　Total number of variants from multiple databases

数据库	变异综合得分位点总数/万
LINSIGHT	22 102
FATHMM-XF[24]	804 741
Eigen	8 329
Fitcons	8 329
REVEL[25]	8 208
CADD	1 236 254

2.1.6　变异基础信息

变异区域的 GC 比例,以及变异所在的区域所在的功能区,变异在相应的功能区域的相对位置,例如是否位于 promoter 区域,是否在编码区域,在第几个外显子上,在第几个内含子上,这些变异的信息相关的数据都进行了整理,以及用 dbSNP(150 版本)数据对每个位置的变异进行了注释。这些信息覆盖整个基因组。

2.1.7　蛋白功能数据

基因变异会对蛋白功能造成相应改变,基于氨基酸序列的同源性和物理性质来预测氨基酸的替换对蛋白质功能是否造成影响,用来评估基因变异的有害程度的数据有 SIFT[26] 得分及 polyphen[27] 得分,除了这两种打分数据我们还整理了 MA[28]、LRT[29] 等蛋白功能预测数据对基因组变异进行打分。

2.1.8　miRNA 数据

MicroRNA (miRNA) 是一类内生的、长度约为 20-24 个核苷酸的小 RNA,其在细胞内具有多种重要的调节作用。每个 miRNA 可以有多个靶基因,而几个 miRNA 也可以调节同一个基因。miRNA 的基本信息及靶向关系的信息来源于 miRBase 和 TragetScan[30] 数据库等数据库,各库包含的数据量(见表 2)。

表 2　miRNA 数据量统计

Table 2　Statistics of miRNA based on multiple databases

数据库	位点数量
miRBase	4 801
snoRNA	402
TragetScan	122 606

2.1.9　结构变异注释数据

结构变异注释使用的注释工具室 AnnotSV[31],注释数据使用的是软件自带的,注释项目包括:变异的基本信息,比如变异所在的转录本,所在的 CDS

长度,变异长度等信息,还提供变异位置其他人群机构变异频率的注释,比如 DGV、gnomAD、1000G、IMH 数据库。变异致病性注释使用的是 dbVAR、ClinGen、clinvar 等数据。结合以上所有信息对相应的结构变异进行打分,得出致病性结论。

2.2 注释流程涉及软件及工具

2.2.1 LiftOver

将所有的数据转换为同一个参考基因组之后进行注释是一个必经的步骤,LiftOver 是一个高效的位置信息转换工具,在我们使用工具进行转换的时候,成功转换的比例大于 99.97%,2 000 万条变异的数据进行基因组位置转换在五分钟内完成,注释消耗时间与目标文件的大小成正比。

2.2.2 Annovar

Annovar[1]是一个高效的软件工具,可以利用最新的信息对从不同基因组(包括人类基因组 hg18、hg19、hg38,以及小鼠、蠕虫、苍蝇,酵母和许多其他基因组)检测到的遗传变异进行功能注释。给出带有染色体、起始位置、结束位置、参考核苷酸和观察到的核苷酸的变异列表。其功能包含基于基因的注释、基于区域的注释、基于过滤的注释及其它辅助功能。

在工作流中,Annovar 负责对样本 VCF 进行基于基因注释,主要依赖于 refGene 数据库。在测试过程中,针对 500 万行的 VCF 文件,完成注释工作需要耗时约 8 分钟,针对 2 500 万行的 VCF 文件,完成注释工作需要耗时约 40 分钟。可见 ANNOVAR 的工作效率是比较稳定的,注释消耗时间与目标文件的大小成正比。

2.2.3 VARNOTE

VARNOTE[32]是 ANNOVAR 的作者最新开发的注释工具,该工具主要针对基于过滤及基于区域的注释,由于算法上的优化,其注释效率较 ANNOVAR 有极大的提升,且注释过程中不再需要将依赖的数据库完整读入内存中,有效地降低了对机群资源的占用。使用 VARNOTE 进行注释需要自行构建依赖的数据库,这一过程由于需要进行排序、压缩、建立索引等过程,相较于 ANNOVAR 比较繁琐,但 VARNOTE 的效率提升主要便是依托于对依赖数据库的处理,故这一过程必不可少。

在工作流中,VARNOTE 负责对样本 VCF 进行所有的基于过滤以及区域的注释工作,当前包含的数据库共 56 个。在测试过程中,针对 500 万行的无 INDEL 位点 VCF 文件,VARNOTE 对其进行 56 个数据库的注释仅需要 3 分钟。但在实际工作过程中,VARNOTE 的注释效率波动较大。由于其算法原因,VARNOTE 对 INDEL 位点的注释效率较低,当目标 VCF 文件中 INDEL 位点的占比较大时,其注释效率会有比较明显的下降。但整体而言,VARNOTE 的注释效率仍然是比较理想的。

2.2.4 AnnotSV

AnnotSV 是一款使用 Tcl 语言编写的用于注释以及评估结构变异的软件,其目的在于发现结构变异的潜在致病性以及过滤出假阳性结构变异。该软件接受 VCF 格式以及 BED 格式的文件作为输入,以 tab 分割的文本文档作为输出文件。AnnotSV 在注释时需要使用各种不同的参考数据库来进行支持,这部分不需要用户自行处理,在注释过程中若本地不存在相应的数据库,软件会自动进行下载操作。

在工作流中,AnnotSV 负责对结构变异数据进行注释工作。在测试过程中,针对27 000行的结构变异 VCF 文件,完成注释工作需要 13 分钟,针对 49 000行的结构变异 VCF 文件,完成注释工作需要 25 分钟。注释消耗时间与目标文件的大小大体上符合线性规律。

2.3 变异注释结果统计

对注释后的样例数据结果进行统计,发现在示例样本中变异类型中大多数是单碱基替换变异,插入和缺失变异数量相近,并且远少于单碱基替换位点数量,样例数据中位点变异数量超2 000 万个。多数变异位点位于基因间区及内含子区域。因为样例数据中单碱基替换位点比例较高,蛋白结构变异类型中同义突变及非同义突变数量最多,非同义突变数量大于同义突变。样例数据涉及变异类型统计结果(见图 5),样例数据变异位点区域分布(见图 6)。

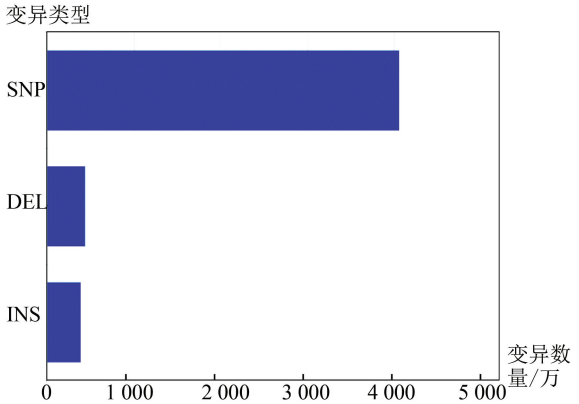

图 5 样例数据变异类型数量

Fig.5 Statistics of multiple types of variation

因为样例数据中单碱基替换位点比例较高,蛋白结构变异类型中同义突变及非同义突变数量最

多,远超移码突变,非同义突变数量大于同义突变,而且占总突变数量的一半以上。移码突变中,不是 3 或 3 的倍数碱基缺失或增加的移码突变(编码氨基酸序列改变)数量少于其他移码突变数量。突变导致终止密码子获得或缺失在所有变异类型内比例最低。编码区蛋白结构变异类型比例(见图 7)。

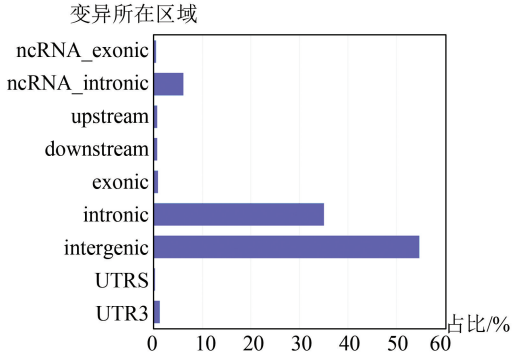

图 6　样例数据变异所在区域分布比例

Fig.6　Statistics of regional distribution of genomic variation

2.3.1　转换和颠换

碱基颠换(transversion)是指在碱基置换中嘌呤与嘧啶之间的替代,而转换(transition)则是一个嘌呤被另一个嘌呤,或者是一个嘧啶被另一个嘧啶替代。在样例数据中,转换变异数量是颠换变异数量两倍有余(约 2.1 倍),其中 G>T 和 C>A 碱基变异比例最高,均超过了全部变异数量的 20%,T>G 和 A>C 次之,均超过全部变异数量的 10%,各种颠换变异的碱基比例均较低。各碱基变异情况统计结果(见图 8)。

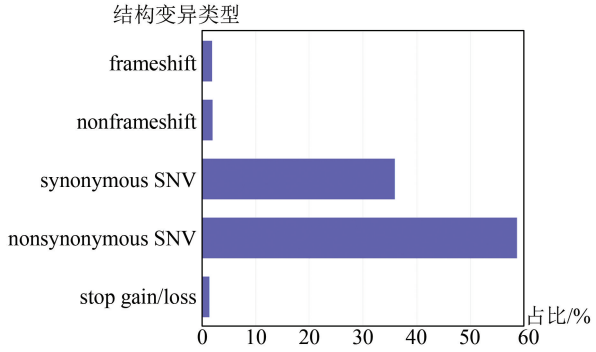

图 7　样例数据编码区蛋白结构变异类型比例

Fig.7　Statistics of ratio of protein structural variation in encoding region

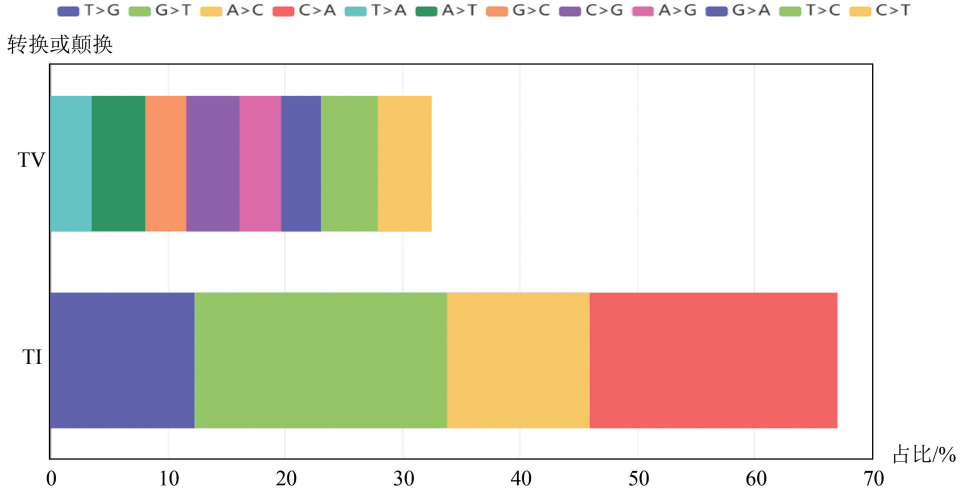

图 8　转换及颠换变异比例

Fig.8　Statistics of ratio of transition and transversion variations

2.3.2　样本变异数量分布

样例数据中每个样本的变异数量均有所差异,经过我们统计单样本变异数量在 470 万到 500 万区间,样本变异数量在 485 万附近的样本最多,样本变异数量分布(见图 9)。

2.3.3　结构变异数量分布

每个样本中有多种类型的结构变异,样例数据的结构变异数量在 7 000 到 1.2 万之间,大部分样本内的结构变异数量在 9 000 个左右,样例数据结构变异数量分布(见图 10)。

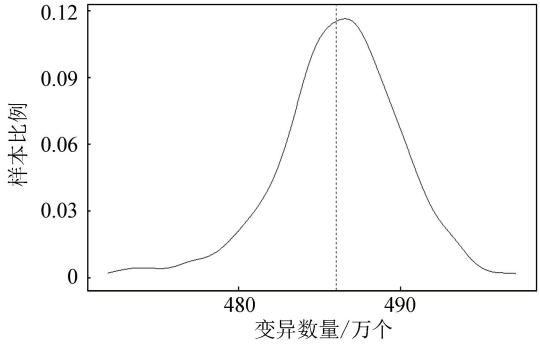

图 9　样本变异数量分布

Fig.9　Distribution of variation quantity

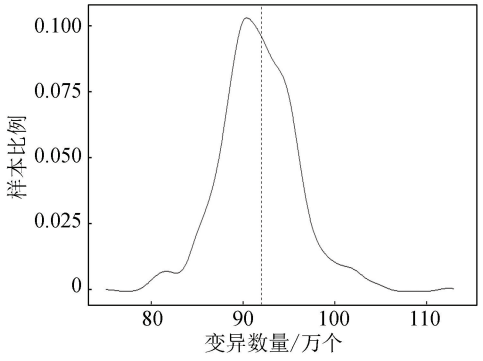

图 10 样本结构变异数量分布

Fig.10 Distribution of the number of structural variations

3 结 论

1)基因组注释是确定基因组序列中基因及编码区位置的过程,其可以直接对序列赋予生物学意义,故基因组注释是解读基因组序列重要途径。基因表达及变异对其生物学功能发挥具有决定性意义,通过基因组注释,可以确定基因调控区域及编码区域是否变异,以及变异的影响,从而帮助研究者从生物功能的角度理解基因组。个体表型的差异究其根本是由于遗传物质的差异造成的,通过基因组注释可以帮助研究者发现个体差异的遗传本质,发现群体的遗传特征。此外,通过对基因组进行注释,结合已有变异与表型的关联知识,可以评估个体的健康情况并指导个性化用药。

2)我们研究开发的面向大规模人群的自动化基因组注释系统利用国际具有高置信度的人类基因组变异数据库,综合运用不同注释方法和高效的索引技术,实现了对个人基因组和群体基因组的高效、快速,准确和全面的基因组信息注释。注释系统对每一个变异位点实现了基因组区域类型、变异类型、基因类型等基础基因组相关信息标注。同时,系统对每一个变异位点进行综合功能性分析,例如:基因组转录、甲基化、功能损失以及各类疾病的风险预测等分析。该注释系统计算并展示个人基因组的差异化特征,助力预测个体基因组变异的相关疾病的发病风险。同时,该系统可应用大规模人群基因组数据分析中,提供全面且准确的基因组变异数据的注释服务,为不同地区和民族的人群遗传学分析提供了坚实基础。

参考文献(References)

[1]WANG K, LI M, HAKONARSON H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data[J]. Nucleic Acids Research,2010,38(16): e164. DOI:10.1093/nar/gkq603.

[2]RENTZSCH P,WITTEN D, COOPER G M, et al. CADD: predicting the deleteriousness of variants throughout the human genome[J]. Nucleic Acids Research, 2019, 47(D1): D886-D894. DOI:10.1093/nar/gky1016.

[3]ZWEIG A S, KAROLCHIK D, KUHN R M,et al. UCSC genome browser tutorial[J]. Genomics, 2008, 92(2):75-84. DOI:10.1016/j.ygeno.2008.02.003.

[4]LIZIO M,HARSHBARGER J, SHIMOJI H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas [J]. Genome Biology, 2015,16(1):1-14. DOI:10.1186/s13059-014-0560-6.

[5]YATES A D,ACHUTHAN P, AKANNI W,et al. Ensembl 2020[J]. Nucleic Acids Research, 2020, 48(D1):D682-D688. DOI:10.1093/nar/gkz966.

[6]LESURF R, COTTO K C, WANG G, et al. ORegAnno 3.0: A community-driven resource for curated regulatory annotation[J]. Nucleic Acids Research, 2016, 44(D1): D126-132. DOI:10.1093/nar/gkv1203.

[7]HUANG Y F, GULKO B, SIEPEL A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data[J]. Nature Genetics, 2017, 49(4): 618-624. DOI:10.1038/ng.3810.

[8]GRIFFITHS-JONES S, GROCOCK R J, VAN DONGEN S, et al. miRBase: microRNA sequences, targets and gene nomenclature[J]. Nucleic Acids Research, 2006, 34(Database issue): D140-144. DOI:10.1093/nar/gkj112.

[9]XIE Jun, ZHANG Ming, ZHOU Tao, et al. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs[J]. Nucleic Acids Research, 2007, 35(Database issue): D183 - 187. DOI: 10.1093/nar/gkl873.

[10]HARROW J, FRANKISH A, GONZALEZ J M, et al. GENCODE: the reference human genome annotation for The ENCODE Project[J]. Genome Research, 2012, 22(9): 1760-1774. DOI:10.1101/gr.135350.111.

[11]ZHAO Hao, SUN Zhifu, WANG Jing, et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies[J]. Bioinformatics, 2014, 30(7):1006-1007. DOI:10.1093/bioinformatics/btt730.

[12]KENT W J, ZWEIG A S, BARBER G, et al. BigWig and BigBed: enabling browsing of large distributed datasets [J]. Bioinformatics, 2010, 26(17): 2204-2207. DOI: 10.1093/bioinformatics/btq351.

[13]KARCZEWSKI K J, FRANCIOLI L C, TIAO G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans[J]. Nature, 2020, 581(7809): 434-443. DOI:10.1038/s41586-020-2308-7.

[14]GLUSMAN G, CABALLERO J, MAULDIN D E, et al. Kaviar: an accessible system for testing SNV novelty[J].

Bioinformatics, 2011, 27(22): 3216-3217. DOI: 10.1093/bioinformatics/btr540.

[15] LOHMUELLER K E, SPARSØT, LI Q, et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes[J]. The American Journal of Human Genetics, 2013, 93(6): 1072-1086. DOI: 10.1016/j.ajhg.2013.11.005.

[16] BUNIELLO A, MACARTHUR J A L, CEREZO M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019[J]. Nucleic Acids Research, 2019, 47(D1): D1005-D1012. DOI: 10.1093/nar/gky1120.

[17] EICHER J D, LANDOWSKI C, STACKHOUSE B, et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes[J]. Nucleic Acids Research, 2015, 43(D1): D799-D804. DOI: 10.1093/nar/gku1202.

[18] TATE J G, BAMFORD S, JUBB H C, et al. COSMIC: the catalogue of somatic mutations in cancer[J]. Nucleic Acids Research, 2019, 47(D1): D941-D947. DOI: 10.1093/nar/gky1015.

[19] LI Q, WANG K. InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines[J]. American Journal of Human Genetics, 2017, 100(2): 267-280. DOI: 10.1016/j.ajhg.2017.01.004.

[20] SIEPEL A, BEJERANO G, PEDERSEN J S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes[J]. Genome Research, 2005, 15(8): 1034-1050. DOI: 10.1101/gr.3715005.

[21] DAVYDOV E V, GOODE D L, SIROTA M, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++[J]. PLoS Computational Biology, 2010, 6(2): e1001025. DOI: 10.1371/journal.pcbi.1001025.

[22] IONITA-LAZA I, MCCALLUM K, XU B, et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants[J]. Nature Genetics, 2016, 48: 214-220. DOI: 10.1038/ng.3477.

[23] LIU X, WU C, LI C, et al. dbNSFP v3.0: A One-Stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs[J]. Human Mutation, 2015, 37(3): 235-241. DOI: 10.1002/humu.22932.

[24] SHIHAB H A, GOUGH J, COOPER D N, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models[J]. Human Mutation, 2013, 34(1): 57-65. DOI: 10.1002/humu.22225.

[25] IOANNIDIS N M, ROTHSTEIN J H, PEJAVER V, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants[J]. The American Journal of Human Genetics, 2016, 99(4): 877-885. DOI: 10.1016/j.ajhg.2016.08.016.

[26] KUMAR P, HENIKOFF S, NG P C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm[J]. Nature Protocols, 2009, 4(8): 1073-1081. DOI: 10.1038/nprot.2009.86.

[27] ADZHUBEI I A, SCHMIDT S, PESHKIN L, et al. A method and server for predicting damaging missense mutations[J]. Nature Methods, 2010, 7(4): 248-249. DOI: 10.1038/nmeth0410-248.

[28] REVA B, ANTIPIN Y, SANDER C. Predicting the functional impact of protein mutations: application to cancer genomics[J]. Nucleic Acids Research, 2011, 39(17): e118. DOI: 10.1093/nar/gkr407.

[29] CHUN S, FAY J C. Identification of deleterious mutations within three human genomes[J]. Genome Research, 2009, 19(9): 1553-1561. DOI: 10.1101/gr.092619.109.

[30] AGARWAL V, BELL G W, NAM J W, et al. Predicting effective microRNA target sites in mammalian mRNAs[J]. Elife, 2015, 4: e05005. DOI: 10.7554/eLife.05005.

[31] GEOFFROY V, HERENGER Y, KRESS A, et al. AnnotSV: an integrated tool for structural variations annotation[J]. Bioinformatics, 2018, 34(20): 3572-3574. DOI: 10.1093/bioinformatics/bty304.

[32] HUANG Dandan, YI Xianfu, ZHOU Yao, et al. Ultrafast and scalable variant annotation and prioritization with big functional genomics data[J]. Genome Research, 2020, 30(12): 1789-1801. DOI: 10.1101/gr.267997.120.