遗传 Hereditas (Beijing) 2018 年 11 月, 40(11): 1015－1023
www.chinagene.cn

研究报告

BIG-Annotator：基因组测序数据高效功能注释及其在遗传诊断中的应用

黄莹[1,2]，刘琪[1]，池连江[1]，石承民[1]，吴祯[1]，胡敏[3]，石宏[4]，陈华[1,2]

1. 中国科学院北京基因组研究所，中国科学院精准基因组医学重点实验室，北京 100101
2. 中国科学院大学未来技术学院，北京 100094
3. 云南省骨代谢性疾病机理及药物干预重点实验室，昆明 650214
4. 昆明理工大学灵长类转化医学研究中心，昆明 650500

摘要：近年来二代测序技术发展迅速，在精准医疗和遗传诊断上得到日益广泛的应用。对二代测序数据进行分析的一个核心环节是对遗传变异位点的识别和注释。基于此，本课题组开发了一个能高效对全基因组单核苷酸多态位点进行功能注释的软件——BIG-Annotator。该软件以 JAVA 语言编写，且提供多线程运行模式，运行更为高效，比现有的同类软件或流程提速 10 多倍，适用于人群队列研究、大样本全基因组关联分析等数据量庞大、时效性要求高的分析需求。同时，该软件还集成了目前常用的二代测序遗传变异注释数据库，以及临床数据解读与报告的标准指南(2015 ACMG-AMP《解读报告标准指南》)，并且增加了针对肿瘤组织遗传变异注释的信息。最后，通过两个研究实例具体说明该软件在遗传诊断中的应用。

关键词：二代测序；遗传变异注释；遗传诊断；精准医学

Application of BIG-Annotator in the genome sequencing data functional annotation and genetic diagnosis

Ying Huang[1,2], Qi Liu[1], Lianjiang Chi[1], Chengmin Shi[1], Zhen Wu[1], Min Hu[3], Hong Shi[4], Hua Chen[1,2]

1. CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
2. School of Future Technology, University of Chinese Academy of Sciences, Beijing 100094, China
3. Yunnan Key Laboratory of Basic Research on Bone and Joint Diseases, Kunming 650214, China
4. Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China

Abstract: The next generation sequencing (NGS) technology has been playing important roles in genetic diagnosis of diseases in recent years, and serving as a technological basis of precision medicine. In analyzing NGS data, the variant

收稿日期: 2018−09−29; **修回日期:** 2018−11−02
基金项目: 国家自然科学基金项目(编号：31571370, 91631106)和中国科学院“百人计划”项目资助[Supported by the National Natural Science Foundation of China (Nos. 31571370, 91631106) and the “Hundred Talents Program” of Chinese Academy of Sciences]
作者简介: 黄莹，硕士研究生，专业方向：群体与计算遗传学。E-mail: huangying@big.ac.cn
通讯作者: 陈华，博士，研究员，研究方向：群体与计算遗传学。E-mail: chenh@big.ac.cn
DOI: 10.16288/j.yczz.18-274
网络出版时间: 2018/11/6 17:30:36
URI: http://kns.cnki.net/kcms/detail/11.1913.R.20181106.1730.006.html

annotation is an important step. In this study, we developed a computationally efficient software (BIG-Annotator) to perform functional annotation for whole-genome single nucleotide polymorphisms. BIG-Annotator integrates the widely used databases and pipelines for variant annotation of genetic diseases and tumors, and follows the 2015 ACMG-AMP Standard Guide for Interpretation and Reporting of Clinical Variants. BIG-Annotator is ten times faster than the existing software, and suitable for annotating genomic sequencing data from large samples. Here we present two analysis cases of genetic diagnosis using BIG-Annotator to show its applications.

Keywords: NGS; variant annotation; genetic diagnosis; precision medicine

精准医疗是随着基因组测序技术的进步和生物信息与大数据科学的交叉应用而发展起来的、以个体化医疗为目的的新型医学理念与医疗模式[1]。进行精准医疗的诊疗依据来自于高通量二代测序(next generation sequencing, NGS)数据。随着测序成本的大幅度降低以及测序通量的不断提升，全基因组测序已经成为研究人类疾病最为快速有效的方法之一。然而，二代测序技术产生的原始数据并不能直接应用于遗传病诊断和临床治疗，还需要一系列后续的生物信息学分析，包括 NGS 数据基本分析，如碱基和序列质控、序列比对和变异识别，以及基于数据库或算法的变异注释和对注释结果的验证与解读等[2]。对于精准医疗和遗传诊断来说，在遗传变异注释的基础上还需要进一步给出临床诊断指导和可能的用药指南。

目前常用的遗传变异注释的软件有两类：一类基于功能数据库完成各种变异注释, 如 ANNOVAR[3]；另一类更侧重于对遗传变异的临床意义给出判断，如 InterVar[4]。ANNOVAR 软件基于 Perl 语言开发，能够确定 SNP 或 CNV 是否导致蛋白质编码区变化以及受影响的氨基酸，并且能够由用户自主选择不同类型的基因定义系统(如 RefSeq、UCSC、Ensemble 等)。ANNOVAR 能独立识别非编码区的特定功能区域的变异, 如保守性区域、转录因子结合位点、DNase I 高度敏感的活性染色质区域、组蛋白结合修饰区域；或关联其他数据库，如 dbSNP (https://www.ncbi.nlm.nih.gov/SNP)和 1000 Genome Project[5]等，对遗传变异进行注释。此外，ANNOVAR 还能检索用户自定义的基因组位置，识别与孟德尔遗传性疾病相关的可能的基因列表。InterVar 是一个基于 Python 语言开发的生物信息分析软件，应用 ACMG-AMP《解读报告标准指南》，实现对 ACMG 28 条判读标准中的 18 条进行自动化评分(其余 10 条由于需要后续证据输入或者参数调整, 如 Sanger 测序验证结果), 将变异分类为“良性”，“可能良性”，“不确定重要性”，“可能致病性”和“致病性”。

虽然这些遗传变异注释软件在 NGS 数据分析和遗传诊断中成为极重要的分析工具，但是在性能上仍存在不足。一方面，ANNOVAR 和 InterVar 在使用上不够便捷，计算速度慢，对单个个体全基因组突变的分析动辄耗时 10 多个小时。另一方面，各类对遗传变异的功能注释和临床作用的注释分散在不同软件中，而且缺少复杂疾病，如肿瘤的信息。鉴于此, 本研究基于 Java 语言编写了分析软件 BIG-Annotator。该软件运行更高效，集成了对二代测序数据的变异注释的方法与流程，整合了包括肿瘤在内的多种遗传疾病变异数据库, 以及 2015 年美国分子遗传协会(ACMG)协同全美分子病理协会(AMP)共同出台的临床数据解读与报告的标准指南[6]和 2017 年中国遗传协会遗传咨询分会拟定的《ACMG 遗传变异分类标准中文版专家共识指南》, 临床可靠性更强。

本文首先介绍了 BIG-Annotator 的分析流程, 并将 BIG-Annotator 和常用的注释软件进行了性能上的比较，进而以一个成骨发育不全症的家系数据和一例癌症组织全基因组测序的遗传变异数据为例，应用于实际数据的注释分析，鉴定出可信度较高的可疑致病突变，为临床确诊和治疗提供依据。

1 软件设计

1.1 软件分析流程

BIG-Annotator 软件的流程和运行过程如图 1 所

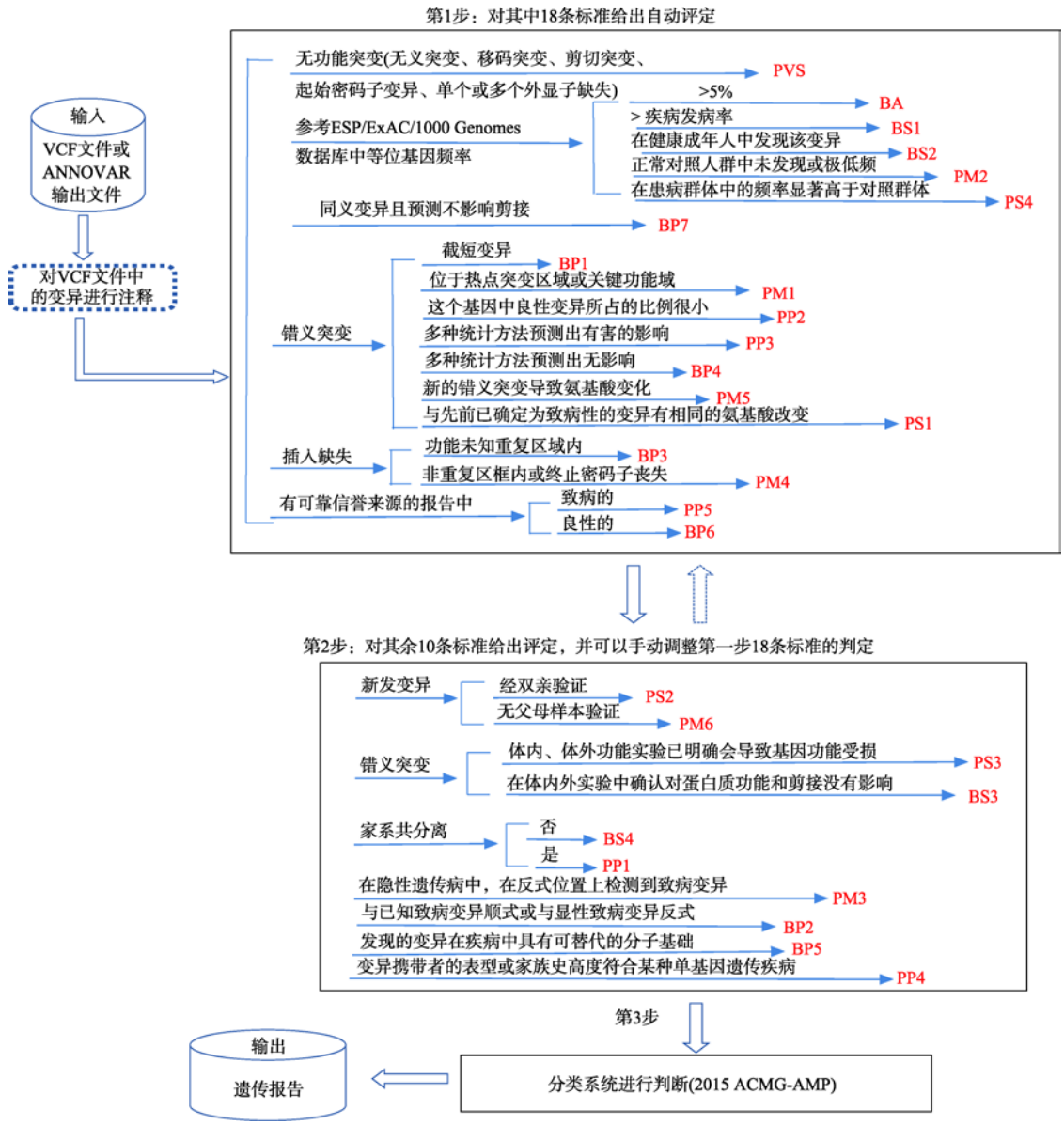

图 1　BIG-Annotator 分析流程

Fig. 1　Pipeline of BIG-Annotator

PVS：当一个疾病的致病机制为功能丧失(LOF)时的无功能变异(无义突变、移码突变、经典±1 或 2 的剪接突变、起始密码子变异、单个或多个外显子缺失)。PM1：位于热点突变区域，和/或位于已知无良性变异的关键功能域(如酶的活性位点)。PM4：非重复区框内插入/缺失或终止密码子丧失导致的蛋白质长度变化。PM5：新的错义突变到氨基酸变化，此变异之前未曾报道，但是在同一位点，导致另外一种氨基酸的变异已经确认是致病性的。PP2：对某个基因来说，如果这个基因的错义变异是造成某种疾病的原因，并且这个基因中良性变异所占的比例很小，在这样的基因中所发现的新的错义变异。PP3：多种统计方法预测出该变异会对基因或基因产物造成有害的影响，包括保守性预测、进化预测、剪接位点影响等。PP5：有可靠信誉来源的报告认为该变异为致病的，但证据尚不足以支持进行实验室独立评估。BS2：对于早期完全外显的疾病，在健康成年人中发现该变异(隐性遗传病发现纯合、显性遗传病发现杂合，或者 X 连锁半合子)。BP1：已知一个疾病的致病原因是由于某基因的截短变异，在此基因中所发现的错义变异。BP2：在显性遗传病中又发现了另一条染色体上同一基因的一个已知致病变异，或者是任意遗传模式遗传病中又发现了同一条染色体上同一基因的一个已知致病变异。BP3：功能未知重复区域内的缺失/插入，同时没有导致基因编码框改变。BP4：多种统计方法预测出该变异会对基因或基因产物无影响，包括保守性预测、进化预测、剪接位点影响等。注意事项：由于做预测时许多生物信息算法使用相同或非常相似的输入，每个算法不应该算作一个独立的标准。BP4 在一个任何变异的评估中只能使用一次。BP6：有可靠信誉来源的报告认为该变异为良性的，但证据尚不足以支持进行实验室独立评估。

示。首先，软件对变异遗传位点进行功能注释，该步骤和 ANNOVAR 功能一致。软件允许输入未注释的 VCF 格式文件，或者经过预注释的 ANNOVAR 输出文件。在功能注释部分，BIG-Annotator 采用 3 种不同的注释方式：基于基因的注释、基于区域的注释和基于过滤的注释。软件提供 6 种主要的功能：数据库下载、蛋白质序列推断、多种文件格式转换、物种转录本自建、不同类型的注释以及灵活注释流程定制。在实际使用中，BIG-Annotator 软件已包含人类基因组已建好的转录本和各种人类疾病数据库。对于其他物种的变异注释，使用者可以根据基因组定义文件或者 GFF3 文件、GTF 文件及序列的 FASTA 文件构建数据库。

在功能注释的基础上，BIG-Annotator 执行以下步骤 :第一步, 依据 ACMG 给出的 28 条标准对 SNP 做判定, 软件根据功能注释对其中 18 条标准给出自动评定；第二步，根据用户提供的资料和证据，对其余 10 条标准给出评定，并可以手动调整第一步 18 条标准的判定；第三步，根据指南给出的分类联合标准规则，对 SNP 生成最终分类结果，最后输出报告文件。

根据 ACMG 2015 年发布的变异位点临床意义判读指南中的 28 条评估标准，可以将变异位点分为 5 级：致病变异、疑似致病变异、临床意义不明确变异、疑似良性变异和良性变异。在指南中这 28 条评估标准具体可以分为 7 类证据：(1) 1 条极强的致病证据 PVS；(2) 4 条强的致病证据 PS1~4；(3) 6 条中等的致病证据 PM1~6；(4) 5 条支持的致病证据 PP1~5；(5) 1 条极强的良性证据 BA，(6) 4 条强烈的良性证据 BS1~4；(7) 7 条支持的良性证据 BP1~7。对每个变异根据遗传变异分类联合标准规则进而从 5 级系统中选择一个分类(例如满足 2 个以上的 PS 证据，或者满足 1 个 PS 证据和 3 个以上 PM 证据，则会判定为致病变异)。

此外，重要的遗传突变往往影响核苷酸的变化和蛋白质的结构功能，因此软件集合了一些常用的疾病遗传变异数据库信息和一些突变功能预测工具。

1.2 数据库和分析工具集成

BIG-Annotator 集成了 ClinVAR、HGMD 和 OMIM 3 个常用的疾病遗传变异数据库相关信息。ClinVAR (https://www.ncbi.nlm.nih.gov/clinvar/)是一个将遗传变异、临床表型、实证数据以及功能注解与分析 4 个方面的信息通过专家评审，逐步形成一个遗传变异–临床表型关联的数据库[7]。该数据库整合了 NCBI 下的 dbSNP、dbVar、gene、GTR、MedGen 和 ACMG 等 10 多个数据库的信息，截止目前收集的条目超过 67 万，分布在 30 181 个基因中。数据库每一个条目的信息包括基因、变异、发生频率、表型、临床意义、评审状态及染色体上的位置等。ClinVAR 的注释在 PVS、PS1、PM5、PM1、PP5 和 BP6 的判定中提供证据。HGMD (http://www.hgmd.cf.ac.uk/ac/index.php)数据库是由英国卡尔地夫医学遗传研究所构建的专门用于收录整理已发表文献中与人类遗传病密切相关的致病位点数据库，每个位点都有参考文献的 PMID[8]。HGMD 从大约 250 种期刊中收集突变信息, 截至目前收录了 9073 个基因中的 127 200 个突变。查询条目详细列出了变异的染色体定位，突变类型列表和相关的表型列表，并将基因内的所有突变定位到 HGMD 的参考序列上。HGMD 对变异的类型做了 10 多种划分，包括编码区、调控区和剪切区域的点突变，大片段或小片段的插入和缺失，基因重组、序列重复，致病性点突变及移码突变，影响可变剪切和疾病相关的多态性位点等。HGMD 根据与疾病的相关程度将变异位点分为 6 个等级：致病突变、疑似致病突变、有功能证据支持的疾病相关多态性变异、实验证明有功能的多态性变异、疾病相关的多态性变异移码或截短突变。HGMD 的注释在 PP5 和 BP6 的判定中提供证据。OMIM (http://www.omim.org/)数据库是一个保持更新的免费的人类孟德尔遗传性疾病数据库。该数据库的信息由世界各地的研究者上传，并提供相关文献，由约翰霍普斯金大学医学院的研究人员进行维护。截止 2018 年 6 月，数据库共收集了 24 576 个条目，记录了 15 909 个基因位点和 8973 种表型信息，包含有 3917 个致病性基因变异。现在 OMIM 关注的表型主要分为单个基因的孟德尔遗传性疾病，具有显著性的单基因致病的复杂疾病，以及染色体缺失和重复综合症。OMIM 的每个条目还包括相关染色体基因位置、临床简介、参考文献的摘要、表

型描述、相似表型的遗传异质性和疾病的分类等。OMIM 的注释在 BA、BS1、BS2、PS4 和 PM2 的判定中提供证据。

BIG-Annotator 中同时还集成了 SIFT 数据分析工具的功能。SIFT (http://sift.jcvi.org/)是一个突变功能预测工具，基于同源蛋白每一个位点上的氨基酸保守性的评估，给出所提交蛋白质每个氨基酸位点发生突变后的评估分数，分数小于 0.05 被认为可能影响到蛋白质功能。

BIG-Annotator 还集成了本课题组开发的一个肿瘤组织遗传变异的数据库。在对肿瘤组织的遗传变异报告中，除了上述的注释，还增加了响应药物的突变、耐药突变、驱动型突变和继发性突变来描述突变的意义。

2　软件应用实例

2.1　成骨不全症致病性突变研究

运用 BIG-Annotator 对来自云南的成骨不全症(osteogenesis imperfect, OI)家系进行全基因组遗传变异的注释。该家系样本包括 5 名 OI 患者和 2 名正常个体。对原始数据的测序质量评估后，以 GRCh37 (hg19)为参考基因组，使用 Samtools v1.0 识别所有的单核苷酸多态性位点(SNPs)，获得 7 例样本的原始 SNP 变异信息对应的 VCF 文件，每例样包含约 309 万个 SNP 位点。以此 VCF 文件作为输入文件，运用 BIG-Annotator，并结合全基因组多态性分析，进一步确定致病变异位点。首先，从 5 名患者的 SNP 数据中提取出共有的等位基因位点(shared allele)，约 79 万个候选 SNP 变异，所有患者在这些位点携带至少一个相同的等位基因。其次，对这 79 万个 SNPs 需要进一步过滤和筛选，下载了 103 个中国汉族(CHB, 1000 genome project phase III)样本约 8000 万 SNP 的 VCF 数据。从 79 万个候选 SNPs 中选取那些在这 8000 万位点中不存在的，或者最小等位基因频率小于 0.01 的等位基因位点，作为候选的疾病突变位点。然后，将候选突变位点集合与测序的 2 个非患者亲属个体的 SNPs 进行比对，滤除掉在非患者中具有相同等位基因的位点，进一步缩小候选致病位点的范围。最后，通过分析剩下的 SNP 的注释信息，找到 OI 的致病突变位点。

2.2　癌症组织变异注释

运用 BIG-Annotator 软件对一例肺鳞癌症患者组织的全基因测序数据进行遗传变异的注释。首先，对样本的全基因组的变异位点进行了注释；然后，对引起蛋白变化的变异进行筛选，包括非同义突变、移码突变等；最后，结合肿瘤注释数据库，挑选出可能致病的位点。

3　实例分析结果

3.1　BIG-Annotator 运行性能

BIG-Annotator 集成了目前常用的二代测序遗传变异注释数据库，以及临床数据解读与报告的标准指南(2015 ACMG-AMP《解读报告标准指南》)，并且增加了针对肿瘤组织遗传变异注释的信息。因此，对于遗传疾病的致病性变异分析，BIG-Annotator 软件最后会生成各个变异位点的详细风险报告，报告中包含的具体信息如表 1 所示。

BIG-Annotator 提供多线程运行模式，运行高效，适用于人群队列研究和大样本全基因组关联分析等数据量庞大、时效性要求高的分析需求。在计算性能的比较中，用 InterVar 软件和 BIG-Annotator 软件分别对一个 30X 的全基因组测序数据进行了遗传变异注释。两种软件都是基于以下数据库做出注释：refGene、refGeneMrna、esp6500siv2_all、ALL.sites.2015_08、avsnp147、dbnsfp33a、clinvar_20170905、exac03、dbscsnv11、dbnsfp31a_interpro、rmsk、ensGene、ensGeneMrna、kgXref、knownGene 和 knownGeneMrna。两个软件的运行都在一个主频为 1009.734 MHz，64 核的服务器上进行。二者都注释了全基因组共 4 451 891 个变异位点，InterVar 耗时 930 min，而 BIG-Annotator 运行耗时 56 min (表 2)。InterVar 每秒钟可注释近 80 个位点，而 BIG-Annotator 每秒钟可注释 1325 个位点(表 2)。BIG-Annotator 的运行时间为 InterVar 的 6%，而注释速度超过后者的 16 倍，更为高效，能更好的满足大样本数据分析的需求。

表 1 BIG-Annotator 软件生成报告解析
Table 1 Items reported by BIG-Annotator

属性	描述
Chr.	染色体号
Position	变异位点在染色体上的绝对位置
Ref	参考基因组碱基型
Alt	样本基因组碱基型
Gene (refGene)	基于 refGene 注释的基因名称，列出该变异所在的基因
Type	软件关于 ACMG 的分类判定
Clinvar	ClinVAR 注释
ExonicFunc (refGene)	外显子区的 SNV 或 InDel 变异类型
Gene (ensembl)	基于 Ensembl 注释的基因号
SNP	dbSNP 数据库关于该位点的描述
Transcripts (ensembl)	基于 Ensembl 的转录本和变异注释
MAF in ExAC_ALL	ExAC_ALL 数据库中的次等位基因频率
MAF in 1000g2014oct	千人数据库中的次等位基因频率
SIFT_score	SIFT 分值，表示该变异对蛋白序列的影响
GERP++_RS	注释变异位点的保守性的 GERP++_RS 分值
dbscSNV	关于 splicing 区的变异注释数据库，基于不同算法给出 Ada 和 RF 分值
OMIM	OMIM 数据库注释的疾病号
Interpro_domain	蛋白序列和蛋白分类数据库 InterPro 中关于结构域的注释

表 2 BIG-Annotator 和 InterVar 运行性能比较
Table 2 Comparison of computational performance between BIG-Annotator and InterVar

软件	运行耗时(min)	注释速率(variants/s)
InterVar	930	79.78
BIG-Annotator	56	1324.97

3.2 成骨不全症致病性突变的鉴定

成骨不全症是一组罕见的伴有全身性结缔组织异常的遗传性疾病[10]，在临床上 OI 亚型多达 15 种，不同亚型患者可能是由不同的基因突变导致，但在表型上是呈连续型变化，典型表现为身材矮小，骨骼发育异常，关节松弛，有多发性骨折[11]，因此仅从临床症状和体征上很难对患者进行疾病亚型的精确诊断与治疗。

本研究对来自云南地区的一个成骨不全症家系的全基因组测序数据进行了分析和注释。经过过滤和筛选，找到了 33 个可能的候选致病位点。这些候选位点都满足以下特点：在患者样本中具有一致性，在患者和非患者样本的对比中具有特异性，在正常人群中最小等位基因频率具有极小性。通过分析候选致病位点 SNP 注释，挑选出所有的位于基因外显子区段上的非同义突变，最终找到 23 个致病变异的候选 SNP 位点(表 3)。运用 BIG-Annotator 对这 23 个 SNP 位点的注释和判定以及用 DAVID 工具对这些候选基因功能注释的结果表明致病变异基因功能和 OI 相关。在该家系测序的 16 名成员中，最终确定 SNP 位点 rs66612022 (*COL1A2*: p.Gly328Ser)上 AG 的基因型在患者中是稳定发生的致病性突变(表 4)，该位点位于常染色体上且表现显性遗传模式(图 2)。BIG-Annotator 对该 SNP 位点生成的报告如表 5 所示。这一发现为该家系中 OI 患者的确诊和个性化治疗具有指导意义。

3.3 癌症组织变异注释分析结果

在对一例肺鳞癌症患者组织的全基因测序数据遗传变异的注释中，BIG-Annotator 分析确定了 2 个可能致病的位点(表 6)。这 2 个位点分别位于 2 号染色体 79523268 碱基位点和 X 染色体 123885987 碱基位点上，在分析个体中均以杂合状态存在，都是由

表 3　成骨不全症相关的 23 个候选 SNP 位点及其对应基因

Table 3　List of 23 candidate SNPs and genes for OI

基因	SNP		
	错义突变	终止密码突变	未知
AGAP3	CHR7_150783917_T2G_L2R	–	–
CD22	rs182604067	–	–
COL1A2	rs66612022	–	–
DNAJC21	rs77129269	–	–
DNASE1L3	rs12491947	–	–
DNTT	CHR10_98079087_G2T_K2N	–	–
HMMR	CHR5_162898459_G2C_E2Q	–	–
HSPA6	rs41297718	–	–
IL17RD	rs140930246	–	–
KCNJ12	rs4985866	–	–
KCNJ18	rs4985866	–	–
MCM7	CHR7_99693641_C2T_A2T	–	–
PDPR	rs10852462	–	–
PRIM2	–	–	rs927192
PTPRG	rs372086949	–	–
RNF19B	rs113840389	–	–
SORBS1	–	CHR10_97096994_G2A_R2X	–
TACC2	rs140280635	–	–
TRPS1	CHR8_116427025_T2G_Q2H	–	–
XKR9	–	rs115507207	–
ZAN	–	–	rs191137
	–	–	rs80067406
ZHX2	CHR8_123965019_C2G_I2M	–	–

对于在 dbSNP 数据库中没有标识的 SNP, 本研究编码了对应的编号, 如 CHR10_98079087_G2T_K2N, 表示位于 10 号染色体的 98079087 位置，G2T 表示 DNA 突变，K2N 表示氨基酸变化，–表示不属于该种变异。

表 4　*COL1A2* 基因外显子片段上发现的变异列表

Table 4　List of mutations occurring in the exon region of *COL1A2*

ID	SNP 编号及其最小等位基因频率					
	rs1801182 (C/0.3398)	rs1800222 (C/0.3835)	rs66612022 (A/na)	rs42524 (G/0.0680)	rs1800238 (T/0.3738)	rs1800248 (T/0.1165)
Ind19	CT	–	–	GG	GT	–
Ind20	CC	–	GA	GG	TT	–
Ind21	CC	CC	GA	GG	TT	–
Ind22	–	–	–	CG	GT	–
Ind23	–	CT	GA	GG	GT	–
Ind24	CT	CT	GA	GG	GT	CT
Ind25	CT	CT	GA	CG	GT	–

“ – ”表示对应个体在该位点上不存在变异。

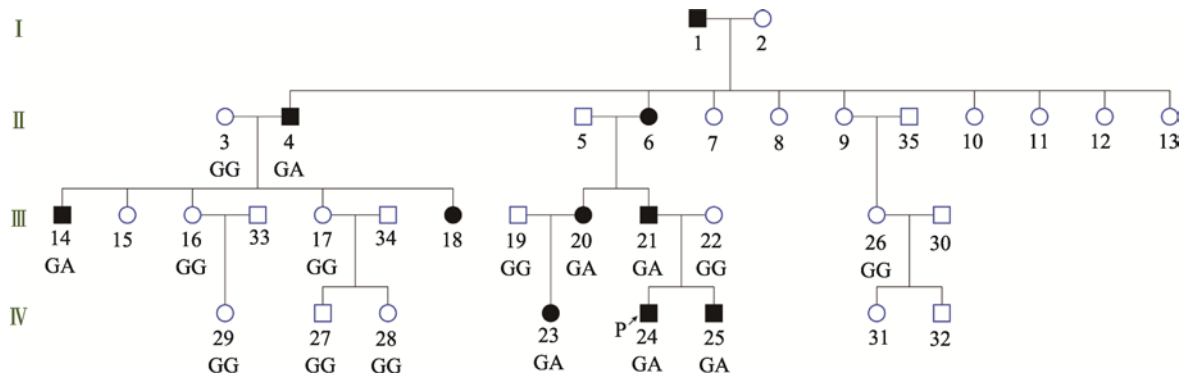

图 2　*COL1A2* 基因 rs66612022 位点基因型在家系上的分布

Fig. 2　Genotypes of rs66612022 of *COL1A2* in the studied pedigree

I~IV 表示家系中的代数，圆形表示女性，正方形表示男性，加黑个体表示为患者，P 表示该个体为家系中的先证者。

表 5　BIG-Annotator 生成的有关 rs66612022 的报告

Table 5　The report on rs66612022 output from BIG-Annotator

属性	描述	属性	描述
Chr.	7	SNP	rs66612022
Position	94039080	Transcripts (ensembl)	ENST00000297268 c.982G>A p.G328S
Ref	G	MAF in ExAC_ALL	. (. means absent)
Alt	A	MAF in 1000g2014oct	.(. means absent)
Gene (refGene)	COL1A2	SIFT_score	0
Type	Pathogenic	GERP++_RS	5.58
Clinvar	UNK	dbscSNV	. (. means absent)
ExonicFunc (refGene)	Nonsynonymous SNV	OMIM	120160
Gene (ensembl)	ENSG00000164692	Interpro_domain	. (. means absent)

表 6　肿瘤组织全基因组测序变异注释报告

Table 6　Annotation report of whole-genome-sequencing variants of tumor cells

染色体	碱基位置	突变	突变种类	影响基因	变异类型	蛋白变化	临床意义	相关疾病	影响方式	靶向药物
Chr.2	79523268	T→TAG TTAAAA ATAATA TACAAA TTTATA	杂合	*CTNNA2*	Frame shift Insertion	CTNNA2:NM_001282598:exon1:c.56_57insAGTTAAAAATAATATACAAATTTATA:p.F19fs	可能致病	胃癌 子宫内膜癌	癌症上的通路	无
Chr.X	123885987	G→GAT	杂合	*XIAP*	Frame shift Insertion	XIAP:NM_001167:exon2:c.325_326insAT:p.G109fs,XIAP:NM_001204401:exon2:c.325_326insAT:p.G109fs	可能致病	小细胞肺癌	癌症上的通路	无

于碱基插入导致移码突变，并分别引起了 *CTNNA2* 基因和 *XIAP* 基因的编码蛋白质的变化，从而影响代谢通路。在数据库的注释中，CTNNA2 基因上的该位点突变与胃癌，子宫内膜癌相关，*XIAP* 基因上位点的突变与小细胞肺癌相关。目前，尚无对这 2 个位点及其相关基因的靶向药物(表 6)。

4　讨　论

本研究基于 JAVA 语言开发编写了软件 BIG-Annotator，且提供了多线程的选项。与现有的同类软件工具相比，BIG-Annotator 在运行性能上有很大提高，时间缩短了 10 多倍，因此能满足大样本二代

测序数据注释分析的需求。BIG-Annotator 既可作为一个二代测序变异注释工具，也可作为一个临床诊断辅助工具，不仅可以对遗传变异给出注释，也提供遗传变异的临床意义和解释，以及对肿瘤的遗传变异给出初步的功能性判断。目前，还鲜有专门针对肿瘤组织的遗传变异的注释解析工具。由于肿瘤细胞突变多为体细胞突变，其意义与遗传性突变的意义(如致病性突变)不同[9]，因此在常见遗传病领域的规范化指导和报告难以直接使用。BIG-Annotator 集成了课题组归纳收集的肿瘤变异数据库，可以对肿瘤组织变异给出初步的遗传注释。

当然，BIG-Annotator 输出结果的可靠性依赖于现有的多个公开的变异位点数据库，在临床实践应用上也受其局限。目前还缺少非常全面的信息库资源，能够保证对每个遗传变异的 28 条标准评估实现自动化判定。因此，在 BIG-Anonotaor 分析流程中允许用户根据自己的主观判断修改某些标准的评估结果。此外，一个可以改进的地方在于，现在只是按照 2015 ACMG-AMP 的 28 条指南来给出临床意义，没有借助一些统计学方法对这 28 条标准做任何的综合性分析，这可能还需要未来能收集到大量实例的软件分析结果和相应的临床验证的比较。最后，对于肿瘤组织的遗传变异的临床解析，目前各种肿瘤组织的变异的临床意义的信息很少，尤其是对耐药性和驱动性突变的鉴定，这方面的数据库信息还有待补充。

参考文献(References)：

[1] Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *N Engl J Med*, 2012, 366(6): 489–491. [DOI]

[2] Middha S. Bioinformatics solution for clinical utilization of next generation DNA sequencing[Dissertation]. 2014, http://hdl.handle.net/11299/168275. [DOI]

[3] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 2010, 38(16): e164. [DOI]

[4] Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet*, 2017, 100(2): 267–280. [DOI]

[5] Altshuler D. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467(7319): 1061–1073. [DOI]

[6] ACMG Board of Directors. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet Med*, 2015, 17(1): 68–69. [DOI]

[7] Landrum MJ, Lee JM, Riley GR, Jang E, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, 2014, 42(Database issue): D980–985. [DOI]

[8] Krawczak M , Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN. Human gene mutation database-a biomedical information and research resource. *Hum Mutat*, 2015, 15(1): 45–51. [DOI]

[9] Slavin TP, Van Tongeren LR, Behrendt CE, Solomon I, Ryback C, Nehoray B, Kuzmich L, Niell-Swiller M, Blazer KR, Tao S, Yang K, Culver JO, Sand S, Castillo D, Herzog J, Gray SW, Weitzel JN. Prospective study of cancer genetic variants: variation in rate of reclassification by ancestry. *J Natl Cancer Inst*, 2018, 110(10): 1059–1069. [DOI]

[10] Monti E, Mottes M, Fraschini P, Brunelli P, Forlino A, Venturi G, Doro F, Perlin S, Cavarzere P, Antoniazzi F. Current and emerging treatments for the management of osteogenesis imperfecta. *Ther Clin Risk Manag*, 2010, 6(2): 367–381. [DOI]

[11] Kataoka K, Ogura E, Hasegawa K, Inoue M, Seino Y, Morishima T, Tanaka H. Mutations in type I collagen genes in Japanese osteogenesis imperfecta patients. *Pediatr Int*, 2010, 49(5): 564–569. [DOI]

(责任编委：杨昭庆)