团　　体　　标　　准

T/SZGIA 2-2018

人类全基因组遗传变异解读的高通量测序数据规范

The Quality Standard of Using High-throughout Sequencing Data to Interpretate Human Whole Genome Variation

2018-01-30　发布　　2018-01-31　实施

深圳基因产学研资联盟　发　布

前　言

本部分编写格式遵循了 GB/T 1.1-2009 给出的规则编写。

本部分由深圳基因产学研资联盟提出并归口。

本部分负责起草单位：深圳华大生命科学研究院、深圳基因产学研资联盟、深圳华大智造科技有限公司、深圳瑞奥康晨生物科技有限公司、北京知因盒子健康科技有限公司、深圳市早知道科技有限公司、深圳中大基因科技有限公司、深圳华大基因科技有限公司。

本部分主要起草人：翦敏、瞿宁、李陶莎、程奇、刘晓、侯勇、蒋慧、杨旭、罗宏敏、郑晨晴、徐驰、张勇、陈钢、马振华、李倩一、杜佳婷、李岱怡、谢强。

本部分为首次发布。

人类全基因组遗传变异解读的高通量测序数据规范

1 范围

本标准明确了高通量测序数据对人类基因组的覆盖度、测序深度指标，以及使用标准品NA12878应达到的检测准确度、灵敏度、特异性的指标。

本标准适用于使用人类全基因组高通量测序数据对个体进行遗传变异解读时，对高通量测序数据的质量和准确度进行评价。本标准为通用标准，临床诊断或科学研究应参考其应用要求制订细化的质量控制参数。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 29859-2013 生物信息学术语

GB/T 30989 高通量基因测序技术规程

GB/T 34798-2017 核酸数据库序列格式规范

SZDB/Z 53-2012 孕妇外周血基因检测胎儿“21-三体综合征”标准

SZTT/SZGIA 1.1-2016 基于高通量测序的环境微生物检测 第1部分:基本规程

SZTT/SZGIA 1.4-2017 基于高通量测序的环境微生物检测 第4部分：临床样本病原微生物检测

3 术语与定义

下列术语与定义适用于本标准

3.1

生物信息学 bioinformatics

应用信息科学及相关学科的方法和技术，研究和分析生物体系和生物过程中信息存储、处理和传递的一门交叉学科。

注1:参见 GB/T 29859-2013。

3.2

序列比对 sequence alignment

比较两个或两个以上核苷酸或氨基酸序列间的相似性的过程。

注1:参见 GB/T 29859-2013。

3.3

测序 sequencing

测定氨基酸或者核苷酸序列的过程。

注1:参见 GB/T 29859-2013。

3.4

核酸数据库 the nucleic acid database

以核酸序列为基本内容，并附有核酸序列注释信息的数据库。

注1:参见GB/T 34798-2017。

3.5

位置 location

一个或一段碱基在另一段较长碱基上的相对坐标位置。

注1:参见GB/T 34798-2017。

3.6

序列编号

序列编号应保证一个序列号码对应一个核酸序列，具有唯一性。序列编号由两个字母加下划线加 6 个数字组成，DNA 序列编号两个字母为 NT(如 NT_123456),RNA 序列字母为 NM(如 NM_123456)，蛋白质序列字母为 NP(如 NP_123456)，整个染色体、质粒等的基因组序列为 NC(如 NC_123456)。提交一个新的序列会系统产生一个新的序列编号，为保证序列的唯一性，当提交的序列在数据库中已经存在，序列将不能被提交。

注1:参见GB/T 34798-2017。

3.7

序列版本号

序列的版本号是由序列编号加一个点号加版本号，如：序列编号.版本号,（NM_123456.1)，当一个序列改变，相应的版本号加1。

注1:参见GB/T 34798-2017。

3.8

读长 read length

测序的下机数据里，每一条序列的平均长度。以碱基（bp）为单位，常见的读长有50bp、90bp、100bp、150bp等。

注1:同GB/T 34798-2017里的“序列长度”。

3.9

测序类型

测序时采用的文库构建类型，和对应的测序方法。一般分为双端测序和单端测序。

3.9.1

单端测序 single end sequencing

将DNA样本处理成片段后，把引物序列连接到片段的一端，然后在引物序列末端加上接头。测序时只测加了引物和接头的这一端。可分为SE50、SE90、SE100等，分别表示读长为50bp、90bp、100bp。

3.9.2

双端测序 pair end sequencing

也叫双向测序。将DNA样本处理成片段后，把引物序列连接到片段的两端，然后在引物序列末端加上接头。对加了引物序列和接头的两端都进行测序。两端序列成对存在，中间的

距离叫插入长度（insert length）。可分为PE50、PE90、PE100等，分别表示每一端的读长均为50bp、90bp、100bp。

3.10

基因检测

指运用测序、荧光定量PCR、基因芯片、液态生物芯片和微流控技术等分子生物学手段，对血液、体液或组织细胞中的DNA进行检测，以期达到预测和诊断疾病的技术。

注1:参见SZDB/Z 53-2012。

3.11

高通量测序

能一次并行对几十万到几百万条DNA分子进行序列测定的测序技术。

注1:参见SZDB/Z 53-2012。

3.12

GC 含量

是在所研究对象的DNA分子中，鸟嘌呤(Guanine)和胞嘧啶(Cytosine)所占的比例。

注1:参见SZDB/Z 53-2012。

3.13

全基因组测序 whole genome sequencing

是指对已知基因组序列的物种的个体核基因组中的全部DNA序列进行测序，并对获得的序列进行差异性分析的方法。使用全基因组测序数据进行遗传解读，需要达到一定的基因覆盖度和深度标准。

3.14

参考序列 reference genome sequence

是指已公开发表的某物种的供参考的全基因组序列,用来供同种物种的不同或相同个体测序后，与之进行比较分析，找出相互间存在的差异。参考序列上存在少部分N区域，即尚不清楚碱基序列的区域。

3.15

参考序列版本

3.15.1

GRCh系列

即Genome Reference Consortium Human Build，由美国国立生物技术信息中心（NCBI）基因组参考序列合作体发布的人类参考基因组，截至2017年12月的最新版为GRCh38版。

3.15.2

hg系列

即human reference genome,由美国加州大学圣克鲁兹分校发布的人类参考基因组序列，截至2017年12月的最新版为hg38版。

3.15.3

GRCh和hg的关系

GRCh38对应着hg38；GRCh37对应着hg19。

3.16

测序重复序列 duplication

上机测序之前对DNA序列进行聚合酶链式反应（PCR）以获得更多的序列，过程中产生的重复分子被测到两次或两次以上时，即为duplication。

3.17

覆盖度 coverage ratio

是指将测序序列比对到参考序列上时，所有被比对到的区域占参考序列总区域的百分比。计算时需要去除参考序列上N碱基区域。

3.17.1

4X覆盖度

指被4条或更多测序序列比对到的位点的总数，占参考序列非N区总数的百分比。

3.18

测序深度 sequencing depth

某一个位置的碱基被n条测序序列覆盖，则被测到了n乘（X），即测序深度是n。一个全基因组测序样本的深度指非N区所有碱基的深度的平均值，计算方法：测序得到的碱基总量（bp）与人类基因组去除N区后的大小（Genome）的比值。

3.18.1

去重测序深度 rmduped sequencing depth

去除duplication后计算的测序深度。

3.19

变异 variation

是指由DNA高通量测序得出的reads经过算法处理后，找出的所有和参考序列不一致的信息，包括单核苷酸多态性、插入删除型变异、结构性变异。

3.20

单核苷酸多态性 single nucleotide polymorphism

简称SNP，是指在DNA中的遗传物质上某一位置的单个核酸被另外一种单个核酸替换的变异。

3.21

插入缺失型变异 insertion and deletion

简称Indel，是指在基因组的某个位置上所发生的小片段序列的插入或者缺失，插入或缺失片段的长度在50bp以下。

3.22

结构性变异 structural variation

包括大片段缺失、大片段重复、倒位、易位。其中大片段缺失和大片段重复又叫拷贝数变异（Copy number variation），即连续较长的序列发生了缺失或者重复，与插入删除型变异的区别在于变异的长度。倒位指染色体上某一段序列发生了180度的颠倒。易位指染色体上的某一片段转移到了其他位置上。

3.23

主要基因型深度比 major allele supporting depth Ratio

是指突变位置上对应的测序下机的读的数目多的一个类型其读的数目占该突变上全部的读数目的比值。

3.24

单核苷酸多态性数据库 the single nucleotide polymorphism database

简称dbSNP，是指一种开放的资源库，包含已知的单核苷酸多态性和插入删除型变异的位置、变异的编号、物种内发生相应变异的频率等信息。

3.25

次等位基因频率 minor allele frequency

简称MAF，指在给定人群中除了参考序列基因型以外，最常见的等位基因出现的频率。

3.26

纯合突变 homogeneous mutation

是指一对等位基因都发生了突变。

3.27

杂合突变 heterogeneous mutation

是指一对等位基因中只有一个发生了突变。

3.28

目标区域 target region

对于提供给用户的结果里的特征（如某基因或某表型）、以及用户诉求且获得承诺的特征（如某基因或某表型），其在基因组上的关联区域，叫作目标区域。即这些区域的变异与上述特征存在已知的相关性。

3.29

NA12878标准品

2014年由美国国家标准与技术研究院(NIST)敲定为检测人类基因测序数据中识别单核苷酸多态性（SNP）与碱基插入缺失（Indel）准确度的标准样品。[1]

3.30

HGVS命名规范

人类基因组变异协会制定的对突变进行命名的规范（sequence variant nomenclature by human genome variation society）

3.31

VCFeval

是RTGtools里的子工具，能准确地对不同类型的高通量测序数据（如VCF文件）进行合并，筛选，比对等操作。当进行比对时，能进行相对于标准集的假阴性、假阳性、灵敏度、

特异性等的计算。RTGtools是由新西兰Real Time Genomics公司发布的高通量测序数据处理工具集[2]。

4 缩略语

下列缩略语适合于本文件。

SNP 单核苷酸多态性

Indel 插入删除型变异

dbSNP 单核苷酸多态性数据库

NCBI 美国国立生物技术信息中心(national center for biotechnology information)

5 高通量测序数据进入变异解读的质量规范

5.1 参考序列

宜使用NCBI发布的最新版本作为参考序列[3][4]。如截至2017年12月的最新版本是GRCh38、hg38。

5.2 测序类型及数据质量

测序类型应选择双端测序。对于从测序仪下机的原始序列，应达到相应高通量测序平台的要求，并由此计算有效的测序数据量。

表1 测序数据的质量要求

类别	要求
测序类型	PE测序
序列读长	≥100bp
错误率大于等于1%的碱基的比例	≤20%（SZTT/SZGIA 1.1-2016）
GC含量	≤44%

5.3 测序数据对基因组的覆盖情况

应计算4X覆盖度和去重测序深度，以此作为指标。

表2 测序数据对基因组的覆盖要求

类别	要求
非N区的4X覆盖度	≥99%
非N区的去重测序深度	≥30X
对解读目标区域的覆盖度(4X)	等于100%，或对所有未覆盖区域采取其他方法进行补测，或向用户明确指出具体未覆盖的区域

5.4 分析时涵盖的变异类型

应至少包含snp和indel的分析。

5.5 突变的命名

对检测出的snp和indel进行命名时，应遵循HGVS命名规范，且为每一个突变名称标记所使用的转录本序列的编号和版本号。对于每个基因，宜使用最长转录本、或最常用的转录本序列、或者二者同时使用；对于每个转录本，宜使用最新版本号。

5.6 snp 和 indel 的准确性评价

对NA12878样本进行测序和分析后，与标准集进行比较，计算出精确度和灵敏度。宜使用VCFeval软件进行计算。

表3 NA12878的**snp和indel的准确性**要求

类别	要求
snp精确度	≥99%
snp灵敏度	≥99%
Indel精确度	≥92%
Indel灵敏度	≥96%

附录A

（资料性附录）

高通量测序数据的分析处理过程

A.1 测序序列生成

步骤包含测序信号处理（signal processing） 和碱基序列转换（ base calling），主要是基于测序的光学或其他信号等，生成含有碱基序列（主要是 A、C、G、T 四种碱基）和每个碱基对应的质量（即碱基读取的可信度）。

本步骤基于测序原始信号，得到原始碱基序列（一般用 ACGT 等核苷酸简称来表示），序列文件一般以 FASTQ 格式存放，供后续分析。FASTQ 是一种存储了核酸序列以及相应的质量评价的文本格式，以 ASCII 编码，是高通量测序的标准格式。

A.2 原始碱基序列过滤

本步骤根据原始 FASTQ 里的碱基质量信息，对于不符合质量要求的序列进行过滤，保留下来的序列才进入后续分析。过滤包含 3 个方面：被测序试剂等污染的序列；平均质量值低的序列；低质量碱基及 N 碱基超过一定比例的序列。

A.3 基因序列比对

序列比对是把样本测序得到的 DNA 序列与参考基因组比对，为后续发现样本的序列突变和差异做准备。这个过程计算密集，为每个短序列读取分配一个 Phred 量表映射质量分数（表明比对过程的可信度）以及读取在参考基因组中的物理位置（计算深度和覆盖度）。序列比对结果通常以 BAM 文件格式存储，是序列比对的 SAM 格式的二进制版本。常用压缩格式为 CRAM 和其加密压缩格式 SECRAM，可以有效节省空间（BAM 文件通常比较大）：CRAM 格式规范（3.0 版本），http://samtools.github.io/hts-specs/CRAMv3.pdf

本步骤基于原始碱基序列，得到比对参考基因组的结果文件（BAM 文件），用于后续分析。

A.4 基因变异分析

这个步骤的输入是 BAM 或其他类似格式文件，基于目前科学界认定的序列变异的类型来进行判断和分析，找出存在的变异，包括单核苷酸突变（SNP）、小的插入和缺失（Indel）、结构性变异（复制，插入，倒位，易位等）。从算法上分析出这些变异的集合。这个步骤的准确性高度依赖于碱基质量（步骤 5.1）和比对质量（步骤 5.2）。对于 SNP 和 Indels，常用的文件格式是 VCF：https://samtools.github.io/hts-specs/VCFv4.3.pdf

A.5 变异过滤

基于数据层面，包括测序数据质量，比对率，深度，覆盖度等。从千或万数量级的序列变异中进行初步过滤。临床、科研用途的数据，依据其具体需求，设计不同的过滤条件。

A.6 变异注释

基于过滤后基因变异，通过已知的和基因功能相关的数据库，例如 dbSNP，OMIM，ClinVar 等，对应基因变异映射到的蛋白等其他信息，从而对变异进行注释。用于对变异的进一步筛选和解读。对于中国人的遗传变异进行解读，应优先使用中国人 / 亚洲人的数据库。

A.7 变异解读

在所有基本数据分析完成后，遗传变异需要结合临床表型、临床知识库、个体化用药指南、科研文献等资料，在遗传学家、遗传咨询师的参与下进一步对得到的若干基因变异进行分析，找出哪些是致病突变，哪些是正常的多态性（简单地理解是无害的），以及判断它们与疾病、药物以及个体特征和表型的关联。

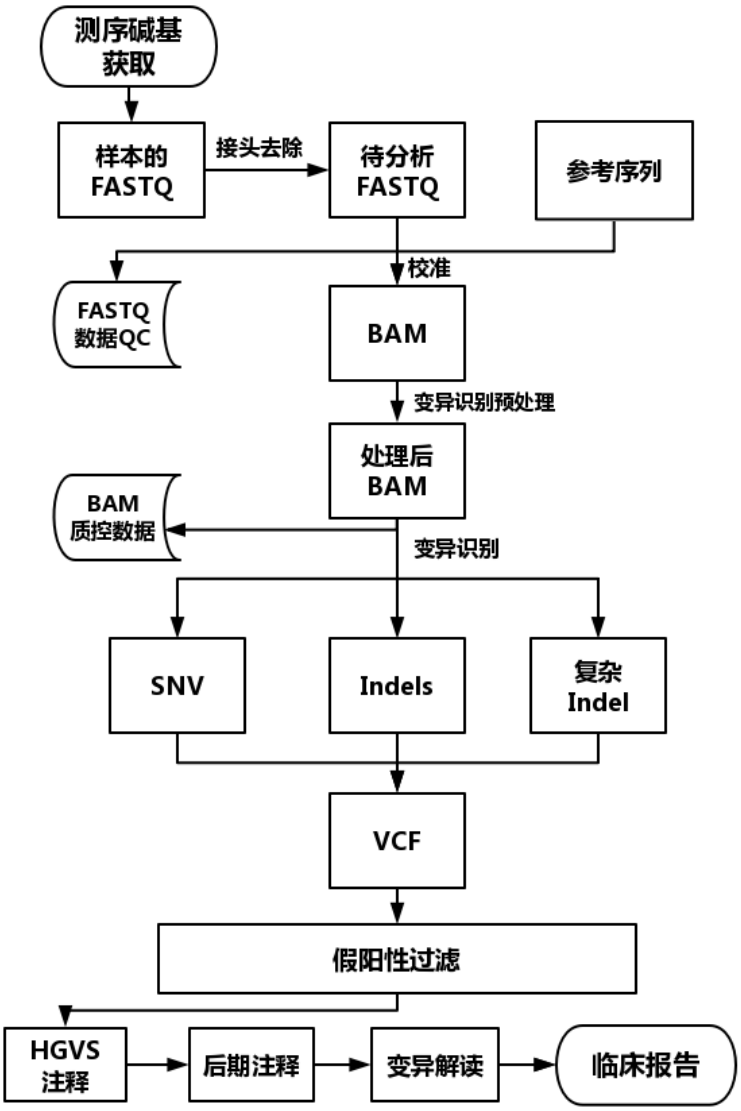


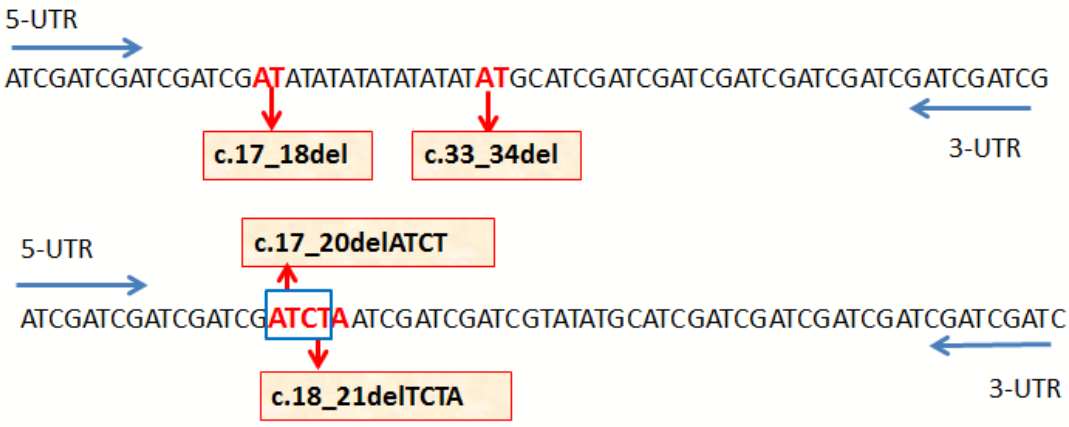
图 A-1 一个典型的生物信息分析流程，图片转化自《The Journal of Molecular Diagnostics》[5]

附录B

（资料性附录）

HGVS命名规范举例

HGVS命名规范，是人类基因组变异协会制定的对突变进行命名的规范（sequence variant nomenclature by human genome variation society），即当序列改变和氨基酸功能一致等价时，突变的命名要以靠近 3-UTR 区来命名。



图B-1 列举了2个HGVS命名矫正的例子。

上例：由于c.17至c.34是一连串AT，当发现有一个AT缺失时，可认为发生在该区段的多个位置，而HGVS矫正后，正确的命名是c.33_34del。

下例：由于蓝色框里TCT的两边都是A，当发生包含TCT的4bp缺失时，没缺的那个A可认为是c.17也可认为是c.21，而HGVS矫正后，正确的命名是c.18_21delTCTA。

官方网站:http://www.hgvs.org/mutnomen/

校正网站:https://mutalyzer.nl/name-checker?description

附录C

（资料性附录）

数据库及软件举例

C．1 频率数据库

C．1.1 dbSNP数据库

https://www.ncbi.nlm.nih.gov/projects/SNP

C．1.2 ExAC数据库

exac.broadinstitute.org

C．1.3 ESP6500数据库

esp.gs.washington.edu

C．1.4 gnomad 数据库

http://gnomad.broadinstitute.org

C．2 疾病及表型数据库

C.2.1 OMIM数据库

https://www.ncbi.nlm.nih.gov/omim

C.2.2 CGD数据库

research.nhgri.nih.gov/CGD

C.2.3 ClinVar数据库

https://www.ncbi.nlm.nih.gov/clinvar

C.2.4 HGMD数据库

www.hgmd.cf.ac.uk

C.2.5 gwasCatalog数据库

http://www.ebi.ac.uk/gwas

C．3 功能预测数据库

dbNSFP 数据库：包含SIFT、polyphen2、LRT、MutationTaster、MutationAssessor、FATHMM、PROVEAN、VEST3、CADD、fathmm-MKL_coding、MetaSVM、MetaLR等对突变可能会造成的影响的预测。

http://varianttools.sourceforge.net/Annotation/DbNSFP

C．4 注释软件

ANNOVAR：可注释的人类基因组版本有 hg18，hg19，hg38。[6]

http://annovar.openbioinformatics.org/en/latest/

参考文献

[1] Justin M Zook, Brad Chapman, etc. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* 32, 246–251 (2014) doi:10.1038/nbt.2835

[2] John G. Cleary[1,§], Ross Braithwaite[1], etc. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* online Aug. 2, 2015; doi: http://dx.doi.org/10.1101/023754.

[3] Yan Guo [a,*], Yulin Dai [a],etc. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 109 (2017) 83–90. doi:10.1016/j.ygeno.2017.01.005

[4] Valerie A. Schneider[1], Tina Graves-Lindsay,etc. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*. doi/10.1101/gr.213611.116.

[5] Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and College of American Pathologists. *The Journal of Molecular Diagnostics* DOI: (10.1016/j.jmoldx.2017.11.003)

[6] Kai Wang[1,*],Mingyao Li[2], and Hakon Hakonarson[1,3].ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data。*Nucleic Acids Res*. 2010 Sep; 38(16): e164.Published online 2010 Jul 3. doi:10.1093/nar/gkq603