ICS 07.80
CCS A40

团 体 标 准

T/LTIA 14—2021

低深度全基因组重测序的基因型推断和遗传变异解读的通用要求

General requirements of genotype imputation and variation interpretation in lowpass whole genome resequencing

2021－11－23 发布　　2021－11－30 实施

深圳市生命科技产学研资联盟　　发 布

仅供试读

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由深圳市生命科技产学研资联盟提出并归口。

本文件起草单位：深圳华大生命科学研究院、深圳市早知道科技有限公司、深圳华大智造科技股份有限公司、深圳瑞奥康晨生物科技有限公司、深圳市生命科技产学研资联盟、深圳华大基因科技服务有限公司、北京知因新生活细胞生物科技有限公司、青岛华大智造科技有限责任公司、深圳华大基因科技有限公司。

主要起草人：翦敏、陈奕、吴静静、黄飞、张勇、李胜康、黄志博、黎宇翔、陈钢、杨旭、吴莉萍、王理中、陶勇、张艳艳、苟洪兰、唐森威、李延红、李陶莎、罗宏敏、梁鑫明、赵霞、滕飞、徐驰、张和、刘姗姗、武庆超、吴昊、李倩一、骆顺。

本文件为首次发布。

低深度全基因组重测序的基因型推断和遗传变异解读的通用要求

1 范围

本文件规定了低深度全基因组重测序中的下机数据质控、基因型推断技术和遗传变异解读结果的要求。

本文件适用于人类、动物和植物等一系列已建立国际通用标准参考基因组的物种，且全基因组重测序平均深度低于10X的基因组数据分析，对个体进行基因型推断和遗传变异解读。

本文件不适用于临床诊断的基因数据分析。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 34798-2017 核酸数据库序列格式规范

3 术语和定义

GB/T 29859-2013和GB/T 34798-2017界定的以及下列术语和定义适用于本文件。

3.1

测序 sequencing

测定氨基酸或者核苷酸序列的过程。

[来源：GB/T 29859-2013，2.4.13]

3.2

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。

[来源：GB/T 34798-2017，3.1]

3.3

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。

[来源：GB/T 34798-2017，3.6]

3.4

读长 read length

测序的下机数据里，每一条序列的长度，也称为测序长度或序列长度。

注：以碱基对(bp)为单位，常见的读长有50bp、90bp、100bp、150bp等。

3.5

低深度全基因组重测序 lowpass whole genome resequencing

是指基于下一代测序技术，对已知基因组序列的物种个体基因组中的全部DNA序列进行全覆盖测序，测序深度在10X以下（真实应用场景下为0.5X～6X，如1X或4X），并进行个体或群体差异性分析的一种技术。

3.6

参考序列 reference genome sequence

是指已公开发表的某物种可供参考的全基因组序列，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。

注：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。

3.7

测序重复率 duplication rate

指被重复测序到的数据量占总共测序数据量的比例。

3.8

覆盖度 coverage

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。

注：计算时需要去除参考序列上N碱基区域。

示例 1：1X 覆盖度：指被 1 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。

示例 2：4X 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。

3.9

测序深度 sequencing depth

是指测序得到的碱基总量与基因组大小的比值。

注：某一个位置的碱基被n条测序序列覆盖，则被测到了n层，即测序深度是n。一个全基因组测序样本的深度指非N区所有碱基的平均深度。

3.10

去重后测序深度 duplicates removed sequencing depth

是指去除测序重复序列后计算的测序深度。

3.11

比对率 mapping rate

是指将测序序列比对到参考序列上时，比对上的测序序列条数占所有测序序列的百分比，或比对上的碱基数占所有碱基数的百分比。

3.12

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、插入缺失型变异、结构性变异。

3.13

单核苷酸多态性 single nucleotide polymorphism（SNP）

是指在基因组上同一位置由单个核酸的变异所形成的多态性遗传标记。

3.14

插入缺失型变异 insertion and deletion（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。

3.15

结构性变异 structural variation

是指生物染色体上结构的变异，由一个物种基因体中的多种变异所组成，包括大片段缺失、大片段重复、倒位、易位。

注1：大片段缺失和大片段重复又叫拷贝数变异，即连续较长的序列发生了缺失或者重复，与插入缺失型变异的区别在于变异的长度。

注2：倒位指染色体上某一段序列发生了180度的颠倒。

注3：易位指染色体上的某一片段转移到了其他位置上。

3.16

基因型推断 genotype imputation

是指利用公共的或特殊搭建的群体参考基因组，在个体测序数据完成比对后，结合区域的上下游基因型规律，判断缺失的位点或区域的基因型，最终达到补齐缺失基因型信息目的的分析过程。

3.17

参考基因组 reference panel

是指基于群体大量样本的单倍型信息，搭建的基本涵盖全基因组的参考性基因型数据集。

4 缩略语

下列缩略语适用于本文件。

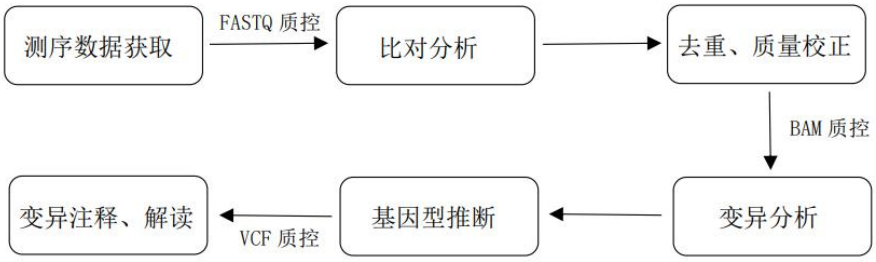
bp：碱基对（base pair）

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

VCF：文件格式（variant call format）

5 低深度全基因组重测序数据分析流程

低深度全基因组重测序数据分析流程应包括测序数据获取、比对分析、数据去重及质量校正、变异分析、基因型推断和变异注释解读，如图1所示。



注：若基因型推断软件以BAM文件作为输入，则该流程中的变异分析包涵在基因型推断中。

图 1 低深度全基因组重测序数据分析流程

6 下机数据质控要求

6.1 一般要求

6.1.1 序列文件应为 FASTQ 格式以供后续分析。

6.1.2 应使用国际通用核酸数据库作为比对的参考序列，例如人类样本宜使用 NCBI 发布的最新版人类基因组参考序列（GRCh38[1]和 hg38[2]）作为参考序列；家牛样本则宜使用美国农业部在 NCBI 上提交的最新版家牛基因组参考序列（ARS-UCD1.2[3]）作为参考序列。

6.2 测序类型和数据质量

6.2.1 低深度全基因组重测序宜使用双端测序，读长应≥100bp。

注：双端测序是指将DNA样本处理成片段后，将引物序列连接到片段的两端，并在引物序列末端加上接头，对加了引物序列和接头的DNA片段两端都进行测序。

6.2.2 下机数据质量控制参数应包含原始碱基数、Q30 比例和 GC 含量。

注1：Q30是指在所研究对象的DNA测序数据中，质量值大于等于30的碱基所占的比例。

注2：GC含量是在所研究对象的DNA分子中，鸟嘌呤和胞嘧啶所占的比例。

6.2.3 序列比对结果应以 BAM 或 CRAM 文件格式存储。BAM 质控应包括：测序重复率、比对率、覆盖度、测序深度、去重后测序深度等。具体要求见表 1。

表 1　低深度全基因组重测序的 BAM 数据质量要求

类别	要求
测序重复率	≤20%
比对率	≥80%
1X测序深度下的1X覆盖度	≥60%
4X测序深度下的1X覆盖度	≥90%
注1：测序深度和去重后测序深度应符合低深度的要求（低于10X），具体可根据真实场景需求来判断。 注2：此质控要求适用于人类、动物、植物等一系列已建立国际通用的标准参考基因组的物种样本分析。	

7　变异分析

7.1　变异分析应包括单核苷酸突变(SNP)和插入缺失型变异(Indel)，可选择性涵盖结构性变异(大片段缺失、大片段重复倒位、易位等)。对于 SNP 和 Indel，宜使用文件格式 VCF，文件格式参考附录 A。

7.2　应结合具体要求，设计不同的 VCF 过滤标准。

示例：可以基于变异位点的测序深度和质量分数，以及变异位点在群体中的突变频率等参数进行过滤。

8　基因型推断和结果评估

8.1　基因型推断

8.1.1　应进行基因型推断以补全 VCF 里缺失的基因型信息。

8.1.2　基因型推断应基于公共或特殊搭建的群体参考基因组。

8.1.3　基因型推断应使用专用软件算法。

注：专用软件有Beagle[4]、IMPUTE[5]、Minimac[6]等

8.2　推断结果的准确性和灵敏性评估

8.2.1　为保证基因型推断结果的质量，应对推断结果的准确性和灵敏度进行评估。

8.2.2　评估内容应涵盖不同测序深度（如 1X 和 4X），基于多款基因型推断软件，并结合一个或多个群体参考基因组，进行综合比较，从而给出最为适当的参考标准。

注：不同物种样本的推断结果的准确性和灵敏度的标准有所不同。

8.2.3　应基于目标物种的标准品数据，用基因型推断后的结果与标准品数据集进行比较。

注：标准品数据集有人类基因组NA12878标准品[7]、玉米基因组B73标准品[8]等。

9　遗传变异解读

9.1　一般要求

应以过滤后的 VCF 作为输入，对应基因变异映射到蛋白等信息，并对变异信息进行注释。

9.2　突变的命名

9.2.1　对于每个基因，宜使用最长转录本，或最常用的转录本序列，或两者同时使用。如有特殊需求，可采用所研究组织中高表达的转录本。

9.2.2　对于每个转录本，宜使用最新版本号。突变对应的核酸或转录本序列的编号标记应符合 GB/T

34798-2017 中 6.2 的要求，版本号标记应符合 GB/T 34798-2017 中 6.3 的要求。

9.2.3 变异注释应涵盖基本的基因功能相关的公共核酸数据库。

附 录 A
（资料性）
VCF 格式文件示例

以GATK软件分析出来的包含SNP和Indel的变异结果VCF文件为例，建议采用VCFv4.2版本的格式规范。

VCF文件应包含基本的以##开头的头文件信息和主体内容变异信息。主体内容每一行应至少展示染色体信息（CHROM）、基因组位置(POS)、变异ID(ID)、参考序列碱基信息(REF)、变异碱基信息(ALT)、变异质量值信息（QUAL）、过滤信息（FILTER）、其他统计信息（INFO）、基因型信息格式（FORMAT)、样本编号及对应基因型信息。所有展示于主体内容的信息，在头文件部分应给出对应的定义和解释。例如，主体内容的其他统计信息（INFO）中展示有AC信息（AC=1），在头文件部分就应对AC给出定义和解解释（##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">)。VCF文件中的头文件所有行都应以两个“#”作为起始，以区分主体内容。

示例：

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr1,length=248956422>
#CHROM POS    ID           REF  ALT    QUAL     FILTER  INFO                      FORMAT  sampleID_demo
chr1  10439 rs112766696   AC   A      377.77   .       AC=1;AF=0.500;AN=2;DP=18  GT:AD:DP:GQ:PL
0/1:4,14:18:74:406,0,74
chr1   13273 .            G    C      218.77   .       AC=1;AF=0.500;AN=2;DP=40  GT:AD:DP:GQ:PL
0/1:28,12:40:99:247,0,749
chr1   13417 .            C    CGAGA  214.77   .       AC=1;AF=0.500;AN=2;DP=18  GT:AD:DP:GQ:PL
0/1:11,7:18:99:243,0,610
chr1   13613 .            T    A      43.77    .       AC=1;AF=0.500;AN=2;DP=39  GT:AD:DP:GQ:PL
0/1:32,7:39:72:72,0,758
```

参 考 文 献

[1] Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics. 2017 Mar 1;109(2):83-90.

[2] Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome research. 2017 May 1;27(5):849-64.

[3] Null D, VanRaden PM, Rosen B, O'Connell J, Bickhart D. Using the ARS-UCD1. 2 reference genome in US evaluations. Interbull Bulletin. 2019 Oct 17(55):30-4.

[4] Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics. 2018 Sep 6;103(3):338-48.

[5] Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the positional burrows wheeler transform. PLoS Genetics. 2020 Nov 16;16(11):e1009049.

[6] Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D. Next-generation genotype imputation service and methods. Nature genetics. 2016 Oct;48(10):1284-7.

[7] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature biotechnology. 2014 Mar;32(3):246-51.

[8] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P. The B73 maize genome: complexity, diversity, and dynamics. science. 2009 Nov 20;326(5956):1112-5.