ICS 07.80
CCS A40

团 体 标 准

T/LTIA 15—2021

人类基因组的低深度全基因组重测序的基因型推断和遗传变异解读

Human genomics' 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

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、深圳市生命科技产学研资联盟、深圳市早知道科技有限公司、深圳华大基因科技服务有限公司、深圳瑞奥康晨生物科技有限公司、深圳华大智造科技股份有限公司、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、北京知因新生活细胞生物科技有限公司。

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

本文件为首次发布。

人类基因组的低深度全基因组重测序的基因型推断和遗传变异解读

1 范围

本文件规定了低深度全基因组重测序人类基因组数据的下机数据质控要求。

本文件适用于进行人类个体遗传变异解读时，对低深度重测序的下机数据质量、基因型推断技术的准确性和灵敏度以及遗传变异解读结果的规范性进行评价。

本文件不适用于临床诊断。

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。

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

3 术语和定义

T/LTIA 14-2021界定的术语和定义适用于本文件。

4 缩略语

下列缩略语适用于本文件。

bp：碱基对（base pair）

dbSNP：单核苷酸多态性数据库(the single nucleotide polymorphism database)

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

OMIM：在线人类孟德尔遗传数据库(online mendelian inheritance in man)

GRCh：基因参考序列联盟发布的人类基因组参考序列(genome reference consortium human build)

hg：美国加州大学圣克鲁兹分校发布的人类参考基因组序列（human reference genome）

HGVS：人类基因组变异协会（human genome variation society）

GP：基因型后验概率(genotype posterior probability)

5 变异分析前的数据分析流程和质量要求

5.1 数据分析流程

人类基因组的低深度全基因组重测序数据分析流程应符合T/LTIA 14-2021中第5章的要求。

5.2 参考序列

宜使用NCBI发布的最新版本人类参考基因组作为参考序列。

注1：NCBI发布的人类参考基因组有GRCh系列和hg系列。GRCh和hg版本具有关联性：GRCh38对应hg38，GRCh37对应hg19。

注2：截至目前最新版本是2017年12月发布的GRCh38[1]和hg38[2]。

注3：该参考序列版本须有较为完善的对应注释相关数据或数据库作为支撑。

5.3 测序类型和下机数据质量

人类基因组的低深度全基因组重测序类型和数据质量要求见表1。

表 1　人类基因组的低深度全基因组重测序类型和数据质量要求

类别	要求
测序类型	双端测序（PE）
读长	≥100bp
Q30比例	≥80%
GC含量	≤45%
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%

6　变异分析

6.1　变异分析的范围

变异分析的范围应符合T/LTIA 14-2021中7.1的要求。

6.2　VCF 过滤

VCF的过滤应符合T/LTIA 14-2021中7.2的要求。

7　基因型推断

7.1　一般要求

基因型推断应符合 T/LTIA 14-2021 中第 8 章的要求，基因型推断结果文件格式示例见附录 A。

7.2　参考基因组

7.2.1　宜使用最新版本国际千人基因组计划项目（1000 Genomes Project）的基因组数据作为标准参考基因组。

注：国际千人基因组计划项目（1000 Genomes Project）于2015年搭建并发布Phase 3版参考基因组，该参考基因组涵盖了非洲、欧洲、东亚、南亚、拉丁美洲五大地域的多种族和多民族数据[3]。

7.2.2　如所有测序样本来自于单一的人种或民族，宜在最新版本国际千人基因组计划参考基因组的基础上，结合大量该人种或民族的其他样本，重新搭建满足特殊需求的参考基因组，提高低深度数据基因型推断结果的准确性和灵敏度。

示例：当研究对象为中国汉族时，在国际千人基因组计划参考基因组（或者只选取其中的东亚 EAS 样本）的基础上结合大量其他的中国汉族样本，搭建出满足需求的中国汉族参考基因组。与单纯使用国际千人基因组计划参考基因组相比，针对中国汉族的参考基因组可以显著提升中国汉族样本的基因型推断结果的准确性和灵敏度。

7.3　基因型推断结果质量

7.3.1　基因型推断前应使用标准品及其标准变异集进行准确性和灵敏度的评估。宜使用国际通用标准品及基于国际千人基因组计划 Phase 3 的参考基因组进行基因型推断。当有特殊需求时，可使用自行搭建的目标人群参考基因组进行评估。

注：目前国际通用标准品为NA12878标准品，即2014年由美国国家标准与技术研究院(NIST)发布的检测人类基因测序数据中识别单核苷酸多态性与插入缺失型变异准确度的标准样本之一[4]。

7.3.2　应对标准品进行低深度全基因组重测序和数据处理，并基于参考基因组进行基因型推断分析，

用基因型推断后的结果与标准变异集[4]进行比较，计算出准确性和灵敏度。

7.3.3　当使用自行搭建的参考基因组评估时，也应使用 NA12878 标准品进行基因型推断并对推断结果的准确性和灵敏度进行评估。

7.3.4　宜使用表 2 中的参数作为参考标准。

表 2　基因型推断后 NA12878 标准品的 SNP 质量要求

类别	测序深度	要求（GP≥0.6）	要求（GP≥0.9）
SNP准确性	1X	≥93%	≥93%
SNP灵敏度	1X	≥80%	≥70%
SNP准确性	4X	≥97%	≥97%
SNP灵敏度	4X	≥95%	≥90%

注1：参考标准的验证和制定过程：综合比较多款主流基因型推断软件（如Beagle[5]、IMPUTE[6]和Minimac[7]等）的表现（准确性和灵敏度），最终得出适中的质量标准(最差表现对应数值)。

注2：通常基因型推断结果会伴随产出GP信息，基因型的GP值越接近1，表示该基因型的准确性越高。因此可基于GP值对结果进行过滤，例如GP≥0.6通常被视为默认过滤标准。

注3：由于常规的基因型推断通常是基于SNP数据，该标准只适用于SNP的基因型推断，并不适用于Indel。

注4：多个样本的联合变异检测（Joint Calling）作为常规全基因组测序分析流程里的常见数据分析方法，适用于本标准。当分析流程涉及多个样本的联合变异检测（Joint Calling）时，对应的基因型推断的灵敏度和准确性应优于表格中提出的质控要求。

8　遗传变异解读

8.1　变异注释结果文件

变异注释结果文件格式示例见附录B。

8.2　突变的命名

8.2.1　对检测出的 SNP 和 Indel 进行命名时，应遵循 HGVS 的命名规范，并为每一个突变名称标记所使用的转录本序列编号和版本号。

8.2.2　对于每个基因，应符合 T/LTIA 14-2021 中 9.2.1 的规定。

8.2.3　对于每个转录本，应符合 T/LTIA 14-2021 中 9.2.2 的规定。

8.3　变异注释相关数据库

变异注释应涵盖基本的基因功能相关的公共核酸数据库，如dbSNP[8]、OMIM[9]、ClinVar[10]、gnomAD[11]等（见附录C）。

附 录 A
（资料性）
基因型推断结果文件

以基于beagle5软件的基因型推断结果为例，常规结果文件应是VCF格式，并包含基本的头文件信息。结果的主体内容，每一行应至少展示染色体信息、基因组位置、参考序列碱基信息、变异碱基信息、变异在参考基因组中的频率、基因型信息以及基因型后验概率信息。

示例：

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA12878
chr1	10519	.	G	A	.	PASS	DR2=0.01;AF=0.0067;IMP	GT:DS:GP	0|0:0.01:0.99,0.01,0
chr1	10563	.	C	A	.	PASS	DR2=0;AF=0;IMP	GT:DS:GP	0|0:0:1,0,0
chr1	10575	.	C	G	.	PASS	DR2=0;AF=0;IMP	GT:DS:GP	0|0:0:1,0,0
chr1	10582	.	T	C	.	PASS	DR2=0;AF=0;IMP	GT:DS:GP	0|0:0:1,0,0
chr1	10583	.	G	A	.	PASS	DR2=0.02;AF=0.0209;IMP	GT:DS:GP	0|0:0.04:0.96,0.04,0

附 录 B
（资料性）
变异注释结果文件

以基于bcfanno1.4软件(https://github.com/shiquan/bcfanno)并结合用户自行下载并定义的数据库(如Clinvar和OMIM)的注释结果为例，常规结果文件应是VCF格式，并包含基本的头文件信息。结果的主体内容，每一行应展示染色体信息、基因组位置、参考序列碱基信息、变异碱基信息、变异所在基因信息、变异所在基因位置信息、变异对应转录本编号信息、变异对应HGVS命名信息、变异类型信息、变异在OMIM数据库中的信息、变异在常见参考基因组中的频率信息、基因型信息等。

示例：

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA12878
chr1    20604981        rs60369023      G       A       951.77  .
DP=2;AF=0.5;AN=2;AC=1;ExonIntron=E2;Gene=CDA;Transcript=NM_001785.2;HGVSnom=NM_001785.2:c.208G>A(p.A
la70Thr/p.A70T);AAlength=147;VarType=Missense;RS=60369023;g1000_AF=0.00139776;g1000_EAS_AF=0.005;gno
mad_exomes_AF=0.000252139;gnomad_exomes_AF_EAS=0.00301554;Gene_MIM_id=123920;Phenotypes=.;Possible_I
nheritance=NA;SIFT_pred=D;SIFT_score=0.002;LRT_pred=D;LRT_score=0.000000;MutationTaster_pred=D;Mutat
ionTaster_score=1;FATHMM_pred=T;FATHMM_score=-0.53  GT  0/1
```

附 录 C
（资料性）
注释数据库

C.1 变异频率数据库

C.1.1 dbSNP数据库，访问地址为：https://www.ncbi.nlm.nih.gov/projects/SNP。

C.1.2 gnomAD数据库，访问地址为：https://gnomad.broadinstitute.org/。

C.2 疾病及表型数据库

C.2.1 OMIM数据库，访问地址为：https://www.ncbi.nlm.nih.gov/omim。

C.2.2 ClinVar数据库,访问地址为：https://www.ncbi.nlm.nih.gov/clinvar。

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] 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68.

[4] 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.

[5] 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.

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

[7] 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.

[8] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001 Jan 1;29(1):308-11.

[9] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005 Jan 1;33(suppl_1):D514-7.

[10] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2018 Jan 4;46(D1):D1062-7

[11] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434-43.