ICS 35.240.80
C 07

团　　体　　标　　准

T/CHIA 21.1-2021

组学样本处理与数据分析标准
第 1 部分：全基因组测序数据分析

Specification of omics sample processing and data analysis
Part 1:whole genome sequencing analysis

2021-07-11 发布　　2021-08-01 实施

中国卫生信息与健康医疗大数据学会　发布

目　次

前　言

本文件按照 GB/T 1.1-2020 给出的规则起草。

T/CHIA 21《组学样本处理与数据分析标准》分为以下五部分：

——第1部分：全基因组测序数据分析；

——第2部分：全外显子组测序数据分析；

——第3部分：转录组样本处理；

——第4部分：转录组文库构建；

——第5部分：转录组测序数据分析。

本文件为T/CHIA 21的第1部分。

本文件由中国科学院北京基因组研究所（国家生物信息中心）提出，由中国卫生信息与健康大数据学会归口。

本文件起草单位：中国科学院北京基因组研究所（国家生物信息中心）、中国科学院生物物理研究所、浙江大学、复旦大学、清华大学、中国人民解放军总医院、北京蛋白质组研究中心、中国科学院微生物研究所、北京大学人民医院、中国科学院上海营养与健康研究所、中南大学、空军军医大学（第四军医大学）和华为技术有限公司。

本文件主要起草人：方向东、陈润生、金力、何昆仑、李亦学、张学工、何顺民、段会龙、周水庚、渠鸿竹、龚尚瑾、隋阳、王霞、吕旭东、朱云平、马俊才、杨忠、石乐明、吴松峰、吴林寰、王振、陈先来、贾志龙、张昭军、娄晓敏、阮修艳、单广乐、乔楠、刘登辉、丁子建。

引　言

《组学样本处理与数据分析标准 第1部分：全基因组测序数据分析》为全基因组数据分析提供一套术语规范、定义明确、语义语境无歧义的流程规范，防止流程缺项、术语不规范、配置不合理等问题。

本文件依据目前已有的开源软件、部分常见服务器供货商、以及自产的中国人全基因组数据，搭建全基因组数据分析流程、从准确性、速度、以及使用便利性方面评估不同流程的性能，最终形成全基因组数据分析流程规范。

为了及时反映全基因组数据分析流程的变化情况，本文件将不断更新以符合当前的实际情况。

组学样本处理与数据分析标准

第 1 部分：全基因组数据分析

1 范围

本文件规定了全基因组数据分析流程中涉及的术语和定义。

本文件适用于全基因组数据分析。

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 30989高通量基因测序技术规程

GB/T 35533染色体异常检测基因芯片通用技术要求

GB/T 35890高通量测序数据序列格式规范

3 术语和定义

下列术语和定义适用于本部分。

3.1

全基因组测序 whole genome sequencing

是高通量测序技术下一代测序技术NGS(Next-Generation Sequencing)的一种，利用高通量测序平台对某一个体的全部基因组序列进行测序。

3.2

测序片段 reads

高通量测序平台产生的含有碱基序列和质量值的序列片段。

[GB/T 35890—2018 ， 定义3.2]

3.3

FASTQ格式 fastq format

FASTQ是基于文本的、保存生物序列（通常是核酸序列）和其测序质量信息的、每四行表示一条序列的标准格式。

[GB/T 35890—2018，定义3.9]

3.4

质量控制 quality control

测序数据的质量好坏会影响数据的下游分析，对测序仪下机的原始数据进行质量评估，具体内容包括含N比例、GC含量、duplication情况、序列长度分布情况、碱基平衡情况等。

3.5

参考基因组 reference genome

参考基因组是由科学家组装的一个数字核酸序列数据库，代表一个物种理想条件下基因集合的所有信息。

3.6

GATK the genome analysis toolkit

由布罗德研究所的数据科学平台开发的工具包，内含多种基因组分析工具，可广泛应用于人类和其他物种的数据分析。目前，GATK已经成为了基因组寻找变异的行业标准。

3.7

变异 variation

变异是生物体、病毒或染色体外DNA基因组核苷酸序列的改变。包括单核苷酸变异、核苷酸小片段插入与缺失变异和结构变异。

3.8

基因分型 genotyping

利用生物学检测方法测定个体基因型（Genotype）的技术，主要对变异的纯合和杂合性进行判断。

3.9

注释 annotation

对找到的变异进行注释，确定变异在染色体上的位置，是哪个基因发生突变以及相关蛋白质的变化情况等信息。

3.10

从头组装 de novo assembly

从头组装可以将原始reads拼接成较长的contigs序列，基于contigs和参考基因组的比对结果进行变异识别，可以增加识别的准确度，实现更加全面的短插入缺失变异以及结构变异的检测。

4 全基因组测序数据相关文件格式

4.1 Fastq 格式文件

Fastq文件中储存的是测序原始下机数据，包含测序序列的序列信息及其对应的测序质量，具有独特的格式：每四行是一个单元，对应一条read。其中第一行以‘@’开头，随后是这条read的相关描述，是每条read的唯一标识符；第二行是read的测序信息，由A、T、C、G和N这五种字母构成，其中N代表测序中无法别识别出来的碱基；第三行以‘+’号开头，后面是序列标识符、描述信息或者为空；第四行，与第二行对应，是测序序列的测序质量。

```
@MN00795:3:000H2HKGG:1:11102:18370:1075 1:N:0:9
CGCAGTTCAGAGACNCGCTCCTCTTTCTNGNGGNAGAAGCNCAATTGNGATAGTNGGAGAANGGNAGGTGATGAAGGGGT
NGGGAGGAGGGAAAGAGANGANGGGCACAGGGAAAGNGAGGAAGGGCACAGAAAAATGTAGGGGGAGGACG
+
AFFFFFFFFFFFFF#FFFFFFFFFFFFF#F#FF#FFFFFF#FFFFFF#FFFFFF#FFFFFF#FF#FFFFFFFFFFFFFF
#FFFFFFFFFFFFFFFF#FF#FFFFFFFFFFFFF#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@MN00795:3:000H2HKGG:1:11102:2138:1076 1:N:0:9
CCTTAGCCACCTTCNGCCCAGGAGCCTGNCNGANCATGTGNACCCCCNGCTGGGNTACAACNATNCCCGGGCTTGGTCCT
NACTCAGAGCTGGCACCANCANCAGAGAGAAGCCAGNCAGAGGGAGCAGGCACCCTCATCCCACCGAGGGC
+
AFFFFFFF/FFFFF#FFFFFFFFFFFFF#F#FF#FFFFFF#FFFFFF#FFFFFF#FFFFFF#FF#FFFFFFFFFFFFF
#FFFFFFFFFFFFFFFFF#FF#FFFFFFFFFFFFFF#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@MN00795:3:000H2HKGG:1:11102:16359:1080 1:N:0:9
CTAACTCCAGAGCANGAACAGGATAAATNTNATNTCACAGNCAAAGANACCAGGNAGGAACNTGNGCTTATCAAGAGCAT
NCCCCTGTTGAAATGGTTNGCNAACAACTATAAAAANGTCGGAGCTACATTGGAAATTGTCACATATAAAT
+
AFFFFFFFFFFFFF#FFFFFFFFFFFFF#F#FF#FFFFFF#FFFFFF#FFFFFF#FFFFFF#FF#FFFFFFFFFFFFAF
#FFFFFFFFFFFFFFFFF#FF#FFFFFFFFFFFFFF#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

图 1 Fastq 文件格式

4.2 VCF 格式文件

VCF文件主要由两个主要部分组成：以‘#’为前缀的注释部分；没有‘#’为开头的主体部分。主体部分中的每一行代表一个variant的信息，包括染色体位置、参考基因组的碱基、variant的碱基、质量分数等共10列信息，具体见表1。

表 1 VCF 文件格式说明

列名称	说明
CHROM	染色体名称
POS	变异在染色体上的位置（1-based position）
ID	dbSNP 数据库中的编号
REF	参考基因组上的碱基
ALT	检测到的碱基，若有多个，则使用逗号分隔
QUAL	碱基质量（值越大越可靠）
FILTER	是否通过一系列过滤： PASS 代表通过一系列过滤保留的质量可靠的 SNPs； LowQuality 代表变异的质量分数≥10 和<50； QDFilter 代表 QD≤2.0； MQFilter 代表 MQ < 40.0； FSFilter 代表 FS > 60.0； HaplotypeScoreFilter 代表 HaplotypeScore> 13.0； MQRankSumFilter 代表 MQRankSum< -12.5； ReadPosRankSumFilter 代表 ReadPosRankSum< -8.0 SnpCluster 代表 SNP 邻近聚集
FORMAT	每个 Sample 的格式：GT（基因型）；AD（REF 与 ALT 的测序深度）；DP（测序深度）；GQ（代表基因型准确度的质量分数）；PL（0/0 纯合, 0/1 杂合突变, 1/1 纯合突变这三种基因型的值用来判断 GQ，值越大错误越大）
Samples	FORMAT 对应的数值，不同格式的值用冒号分开，每一个 sample 对应着 1 列；多个 samples 则对应着多列。如 0/1:27,13:43:70:116,0,70：0/1 表示杂合，其中 REF 为 27 个，ALT 为 13 个，测序深度为 43，基因型的准确性 70，（0/0,0/1, and 1/1 的值为 116,0,70 由于 0 为错误最小值所以表型为 0/1）
INFO	过滤指标的数值. ABHom/ABHet 代表变异的纯合和杂合；ABHom: Allele Balance for

	hets (ref/(ref+alt)); ABHom: Allele Balance for homs (A/(A+O)); AC(Allele Count) 表示该 Allele 的数目；AF(Allele Frequency) 表示 Allele 的频率；AN(Allele Number) 表示 Allele 的总数目;BaseQRankSum: Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities；DB: dbSNP Membership; Dels: Fraction of Reads Containing Spanning Deletions; FS: Phred-scaled p-value using Fisher's exact test to detect strand bias; HaplotypeScore: Consistency of the site with at most two segregating haplotypes; MLEAC: Maximum likelihood expectation (MLE) for the allele counts; MLEAF: Maximum likelihood expectation (MLE) for the allele frequency; MQ：RMS Mapping Quality；MQ0：Total Mapping Quality Zero Reads；MQRankSum：Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities；OND: Overall non-diploid ratio (alleles/(alleles+non-alleles)); QD: Quality by Depth；ReadPosRankSum：Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias; VQSLOD: Log odds ratio of being a true variant versus being false under the trained gaussian mixture model; culprit: The annotation which was the worst performing in the Gaussian mixture model, likely the reason why the variant was filtered out

5 全基因组分析流程

5.1 对原始 Fastq 文件进行质控

使用具有统计计数功能的分析工具对原始Fastq进行质量评价，比如FastQC（http://www.bioinformatics.babraham.ac.uk/projects/fastqc/），并去除低质量的reads（常用软件有trim_galore等）。

5.2 将质控后的 Fastq 文件比对到参考基因组

将质控后的数据利用BWA/SOAP等序列比对工具比对到参考基因组，并采用samtools软件对比对后的文件进行sort（按照染色体比对的位置）排序，可输出为sam格式文件，通常直接压缩为二进制格式的bam文件。此外，还可以针对每个样本的bam文件，统计测序数据的比对情况。

5.3 基于 GATK 识别单碱基变异(SNP)和短插入缺失变异（indels）

1）MarkDuplicate

采用Picard 软件去除由于在建库过程中PCR产生的重复序列。如果两条reads具有相同的长度而且比对到了基因组的同一位置，那么就认为这样的reads是由PCR扩增而来，就会被GATK标记。

2）Indel Realignment

根据已知的Indels（如dbSNP142数据库，1KG indels等）进行局部重新比对，去除在比对中出现的错误，提高检测变异的准确性。

3）Recallbrate Base Quality Scores

针对来自测序仪产生的碱基质量分数进行重矫正，使得最后输出的bam文件中reads的碱基质量能够更加接近真实的与参考基因组之间错配的概率。

4）Calling Variant

根据不同的分析要求，采用HaplotypeCaller、MuTect2等工具针对上述处理好的bam文件进行SNP和indel的变异检测。其中HaplotypeCaller适用生殖细胞（Germline）变异检测，MuTect2适用于检测体细胞（Somatic）突变（需要配对样本）。

5）Variant Filtering

采用 Variant Quality Score Recalibration VQSR 策略对过滤后的 call 出来的变异位点进行过滤。

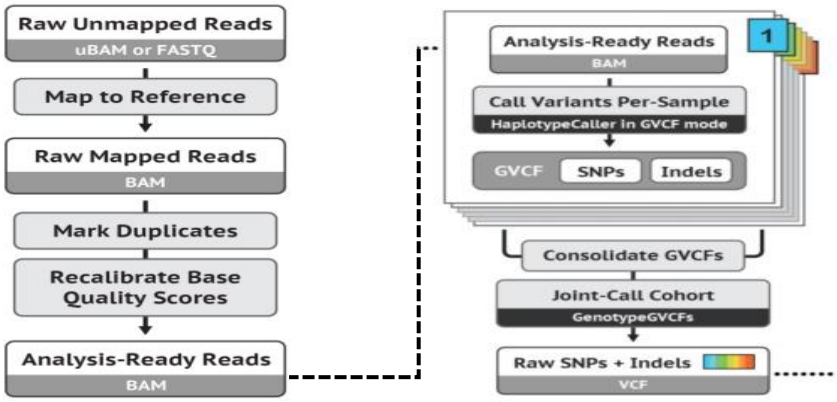

图 2 Germline 分析流程

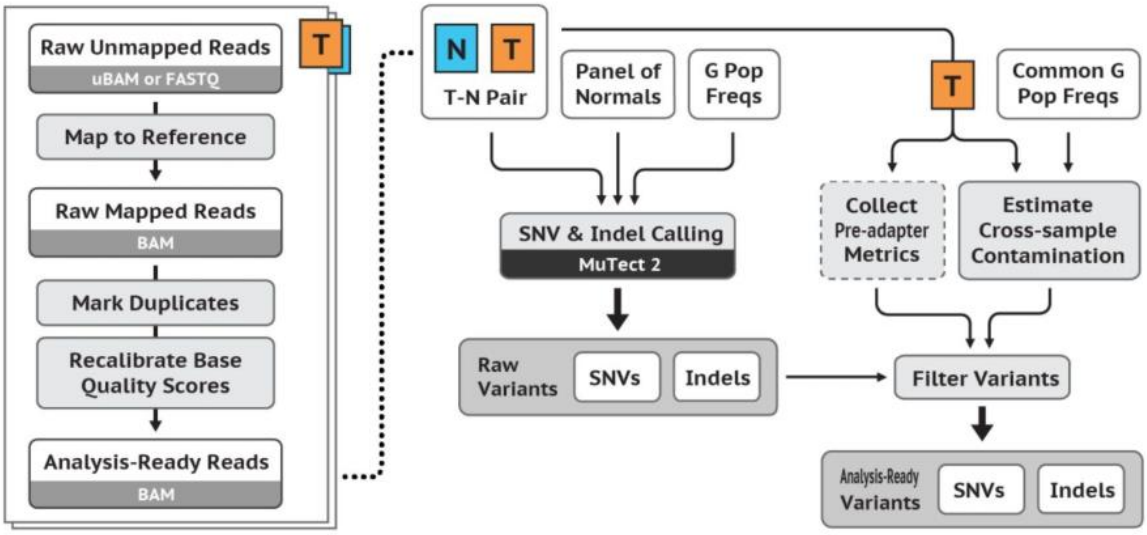

图 3 体细胞突变分析流程

5.4 基于从头组装(de novo assembly)的变异识别

基于从头组装的变异识别可以采用两种策略:

1）采用SOAPdenovo2，ALLPATH-LG等基因组拼接软件，对原始短reads进行拼接，获得高质量的contigs序列，再采用BWA mem比对到基因组后，利用GATK流程对其中的SNP进行识别。或采用MUMmer，AsmVar等工具，比较contigs和参考基因组中的序列差异，并识别其中的SNP和Indel突变以及结构变异。

2）直接采用基于从头组装的变异识别软件，如Fermikit, laSV等，对reads进行拼接和变异识别。

5.5 对变异进行注释

1) 功能注释

突变的位置功能定位，如(在外显子，内含子，剪接位点或者还是基因间区等)；突变所在的基因名称或者邻近的基因；突变如果在编码区域，是否引起氨基酸的改变(同义突变，非同义突变)；如果引起氨基酸的改变，按照 HGVS 命名规则表示--改变的基因 ID，转录本 ID，外显子编号，以及氨基酸改变，如 NOD2:NM_022162:exon8:c.G2722C:p.G908R。默认使用 refSeq 完成基因注释，如果有特殊的要求，可以使用 UCSC known gene，Ensembl，GENCODE，CCDS 等基因注释系统。最常用的突变注释工具有：Annovar, SnpEff 等。

2) 数据库注释

[1] 1000G 注释：检测突变位点是否在 1000 Genomes Projects 数据库中存在，如果存在，显示等位基因频率（allele frequency）。默认是使用所有人种的数据库，如果有特定要求，可以按照要求展示不同人种（比如 AMR, AFR, ASN, EUR）等位基因频率。

[2] dbSNP 注释：检测突变是否在 dbSNP 数据库中存在，如果存在，则显示 rsID。

[3] AVSIFT：检测非同义突变位点重要性，对应非同义突变位点，会给定一个打分，若打分低于 0.05，则表明突变很可能会影响到蛋白质的功能。

[4] UCSC 数据库：UCSC 数据库提供了大量的基因组注释信息，目前关联的数据库如表 2：

表 2 UCSC 数据库中嵌入的数据库

UCSC 数据库	说明
tfbsConsSites	在人/小鼠/大鼠中保守的转录因子结合位点，以 transfac MatrixDatabase (v7.0)为基础
wgRna	snoRNA and miRNA 注释
TargetScanS	TargetScan 预测的 miRNA 靶区域
gwasCatalog	已经发表的各种疾病的 GWAS 结果
genomicSuperDups	基因组中的重复片段
phastConsElements46way	通过 phastCons 对脊椎动物的全基因组比对生成的保守区域，根据用于比对的物种数目，分为 17way, 28way, 30way, 44way 等。默认使用 46way

[5] cosmic63：检测到的突变是否与癌症相关，如果相关，则注释结果包括在 COSMIC 数据库中的 ID、观察到的次数、以及观察到的癌组织。

[6] Oncotator：对检测到的体细胞突变，采用Oncotator软件在线注释SNP和INDEL。注释主要包括以下几个方面：①基因组注释：基因、转录、功能影响（GENCODE）、参考序列、GC含量及人类DNA修复基因注释；②蛋白组注释：位点特异性注释（UniProt）和功能影响预测（dbNSFP）；③癌症变异位点注释：COSMIC突变频率注释，Cancer GenCensus突变基因及变异位点注释，Cancer Cell Line Encyclopedia 重叠变异位点位置，Familial Cancer Database突变基因注释及ClinVar突变位点注释等；④非癌变异位点注释：dbSNP，1000 Genomes和NHLBI GO Exome Sequencing Project (ESP)数据库变异位点注释。

[7] gnomeAD：Genome Aggression Database（简称gnomeAD）是汇集大规模测序及各种疾病研究计划的全外显子组和全基因组测序数据而形成的基因组突变频率数据库。检测突变位点是否在gnomeAD数据库中存在，如果存在，显示等位基因频率（allele frequency）。

6 全基因组数据的个性化分析

6.1 基因融合的预测

基因融合（gene fusion）在基因组中非常普遍，当两个基因分别发生断裂并错误拼接

之后，就有可能形成新的基因片段，这就是融合基因。一般而言，基因融合是指基因组层面的融合，其常见的发生机制有：染色体异位、染色体中间缺失、染色体倒位等。大多数情况下，融合基因可以导致异常序列或功能蛋白质的产生，某些基因的表达失调，进而导致或者促进肿瘤的发生。

6.2 拷贝数变异的预测

拷贝数变异（copy number variation，CNV）是由基因组发生重排而导致的长度为 1 kb 以上的基因组大片段的拷贝数增加或者减少。CNV 是基因组结构变异的重要组成部分，其位点的突变率远高于 SNP，是人类疾病的重要致病因素之一。在获得重矫正质量分数后的 bam 文件后，可以使用 CNVnator 等软件检测拷贝数变。

6.3 结构变异的预测

使用 Breakdancer、CREST、Pindel、Delly、Manta 等软件检测结构变异（Structure Variation，SV），可包括(1) 染色体间易位 (CTX)，(2) 染色体内易位 (ITX)，(3) 倒位 (INV)，(4) 缺失 (DEL)，(5) 插入 (INS)。

6.4 肿瘤纯度检测

肿瘤组织中除了肿瘤细胞之外还有免疫细胞、基质细胞、间质细胞等非肿瘤细胞, 共同影响肿瘤发生发展。肿瘤纯度（tumor purity）是指肿瘤组织中肿瘤细胞所占的比例。肿瘤纯度对肿瘤转录组分析、基因聚类以及分子分类等均有影响。针对成对肿瘤样本，估计肿瘤样本纯度，鉴定校正纯度和克隆组成，为肿瘤样本的微卫星不稳定性检测，体细胞变异做参数设定。

6.5 肿瘤微卫星不稳定性检测

微卫星 (microsatellite) 是遍布于人类基因组中的短串联重复序列, 有单核苷酸、双核苷酸或高位核苷酸的重复，重复次数 10-50 次。与正常细胞相比，肿瘤细胞内的微卫星由于重复单位的插入或缺失而导致微卫星长度的改变，就叫做微卫星不稳定性（MSI, microsatellite instability）。其与肿瘤的发生及免疫疗法有密切的关系。准确检测肿瘤病人MSI对于临床上肿瘤治疗有重要指导意义。

6.6 肿瘤突变负荷检测

肿瘤突变负荷（tumor mutation burden, TMB）被定义为每百万碱基中被检测出的体细胞基因编码错误、碱基替换、基因插入或缺失错误的总数。对体细胞变异检测（如Mutect2），然后对突变进行功能注释，计算肿瘤突变负荷。

6.7 构建肿瘤突变进化树

肿瘤的演变包括所有类型的基因突变的积累，它们共同影响肿瘤细胞的发生发展。对不同时间下，正常组织和肿瘤组织的不同部位进行分析，进而预测肿瘤的进化史，从而更好的了解肿瘤的发生和寻找潜在的生物标志物。

6.8 驱动基因的预测

突变的分类方法有很多种，按照其是否会导致癌症进展，可以分为驱动突变（driver mutation）和乘客突变（passenger mutation）。前者在肿瘤细胞中具有选择性生长优势的突变，后者对肿瘤细胞的选择性生长优势无直接或者间接影响的突变。Driver mutation 倾向

于聚集在基因组某一特定区域，而 passenger mutation 随机分布在整个基因组。