ICS 07.080
A 40

T/SZGIA

团 体 标 准

T/SZGIA 7—2019

基因组从头组装质量指南

The guide of genome de novo assembly quality

2019－06－25 发布　　2019－06－30 实施

深圳基因产学研资联盟　发布

目　　次

前　　言

本标准编写格式遵循了 GB/T 1.1-2009 给出的规则编写。

本标准由深圳基因产学研资联盟提出并归口。

本标准负责起草单位：深圳华大智造科技有限公司、深圳基因产学研资联盟、深圳华大基因科技有限公司、安诺优达基因科技（北京）有限公司、武汉古奥基因科技有限公司、北京诺禾致源科技股份有限公司、武汉未来组生物科技有限公司、广州基迪奥生物科技有限公司

本标准主要起草人：陈芳、刘心、高强、谢寅龙、唐静波、陈恬、刘欢、李启业、杨林峰、范广益、李陶莎、程奇、李倩一、吴昊、高玉池、杨伟飞、刘涛、肖世俊、朱世林、李东野、李瑞强、江文恺、苏亚南、汪德鹏、刘山林、胡江、周煌凯、张羽、艾鹏

本标准为首次发布。

基因组从头组装质量指南

1 范围

本标准给出了基因组从头组装质量指南，包括用于组装数据的输入输出格式指南和组装结果可构成物种序列的质量标准。

本标准适用于对基因组测序技术产生数据（详见附录A）进行组装的组织、机构和个人，作为对基因序列的分析及新物种序列的构建的交付要求。

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 生物信息学术语

GB/T 35537 高通量基因测序结果评价要求

GB/T 35890 高通量测序数据格式规范

T/SZGIA 2 人类全基因组遗传变异解读的高通量测序数据规范

3 术语与定义

下列术语与定义适用于本文件。

3.1

条形码 barcode

若干碱基组成的寡核苷酸链标记，用于在混合测序时，区分不同样本。

3.2

FASTA 格式 FASTA format

FASTA是基于文本的、保存生物序列（通常是核酸序列）的、每两行表示一条序列的标准格式。

每一条测序序列用以下2行信息表示：

a)首行以字符>开头，后面为测序片段ID，字符>与测序片段ID之间不应有空格；

b)第二行为测序的序列（核苷酸或氨基酸编码符号）信息。

3.3

叠连群 contig

测序得到的DNA片段，根据相互间的重叠性，构成的一个长的无缺失的DNA片段。

3.4

骨架序列 scaffold

骨架序列为叠连群的有序集合，其相对位置通常是根据大片段两端序列和barcode等其它信息来推断。骨架序列中叠连群之间的间隙碱基类型及长度通常是未知的。

3.5

N50

将组装序列按长度从大到小排序，从最长序列开始进行逐条累加，累加总长度达到50%的基因组长度时，当前累加的这条序列的长度即为基因组的N50。

3.6

定相块 N50 phase block N50

定相块长度为基因定相后第一个和最后一个杂合变异位点之间的碱基数量。将定相块序列按长度从大到小排序，从最长定相块序列开始进行逐条累加，累加总长度达到50%的基因组长度时，当前累加的这条定相块序列的长度即为定相块N50。

3.7

Hi-C 挂载率 Hi-C alignment ratio

对于已有高通量测序、单分子测序或光学图谱辅助组装的的基因组序列草图和已知染色体数目基因组，利用Hi-C（或类似Hi-C的技术）测序数据将基因组序列草图进行染色体群组的划分，并确定各序列在染色体上的顺序和方向，被挂载到染色体水平的基因组序列长度总和除以原基因组序列草图长度得到的比例即为Hi-C挂载率。

3.8

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。

4 缩略语

下列缩略语适用于本文件。

bp——碱基对（base pair）

DNA——脱氧核糖核酸（deoxyribonucleic acid）

BUSCO———一种普遍通用的单拷贝直系同源基准评测软件（Benchmarking Universal Single-Copy Orthologs）

NCBI——美国国立生物技术信息中心（National Center for Biotechnology Information）

CNGB——中国国家基因库（China National GeneBank）

5 输入输出数据格式指南

5.1 输入格式

用于基因组从头组装的数据为FASTA格式、FASTQ格式、BAM格式等。

5.2 输出格式

基因组从头组装结果应输出对应的contig序列和scaffold序列，可以为FASTA格式、FASTQ格式等 ，碱基字符统一按大写表示。

5.3 数据提交

基因组从头组装的读长文件和组装结果序列文件建议提交到NCBI、CNGB或国家基因组科学数据中心等公共数据库。

6 基因组从头组装质量分级指南

6.1 适用范围

为区分基因组从头组装结果的质量差异，本指南设立以下构建物种序列分级质量要求，适用于大部分常规基因组组装项目及产品(特殊材料可参考,不建议直接使用,针对特殊材料的评估条件具体如下：基因组大小小于100 Mbp，或者大于4 Gb，或重复率高于70%，或杂合率高于1.5%）：

6.2 高级要求

基因组从头组装质量分级为高级的要求如下：

（1）基因组从头组装结果的叠连群N50大于1 Mbp 。

（2）基因组从头组装结果的骨架序列N50大于10 Mbp。

（3）Hi-C挂载率高于95%（N区长度设定为100 bp）。

（4)BUSCO组装质量评估,完整基因的比例高于90%,缺失基因的比例低于5%,或者与近源物种相比,完整基因的比例不比近源物种低5%，缺失基因的比例不比近源物种高5%。

（5）对于含有高通量测序读长数据的组装，需将读长比对回组装序列，统计比对率。高通量测序读长比回率大于95%（样品有污染或者高杂合的情况除外）。

（6）对于能提供多倍体基因组结果的方法，多倍体基因组定相块N50长度大于1 Mbp。

（7）高通量测序读长在基因组中的覆盖度，对于含有高通量测序数据的基因组，需将高通量测序读长比对回组装序列，统计不同深度下，组装序列被高通量测序数据覆盖的比例，1X以上的覆盖度大于97% 。

（8)对于含有高通量测序读长数据的组装，须将读长比对回组装序列，评估单碱基错误率,纯合SNP及纯合INDEL的占比,两者比例均不超过0.01%。

6.3 中级要求

基因组从头组装质量分级为中级的要求如下：

（1）基因组从头组装结果的叠连群N50大于300 Kbp。

（2）基因组从头组装结果的骨架序列N50大于1 Mbp。

（3）Hi-C挂载率高于90%（N区长度设定为100 bp）。

（4）BUSCO组装质量评估，缺失基因的比例低于10%，或者与近源物种相比， 缺失基因的比例不比近源物种高5%。

（5）高通量测序读长比回率大于93%（样品有污染或者高杂合的情况除外）。

（6）对于能提供多倍体基因组结果的方法，多倍体基因组定相块N50长度大于300 Kbp。

（7）高通量测序读长在基因组中的覆盖度，对于含有高通量测序数据的基因组，需将高通量测序读长比对回组装序列，统计不同深度下，组装序列被高通量测序数据覆盖的比例，1X以上的覆盖度大于90% 。

（8)对于含有高通量测序读长数据的组装，须将读长比对回组装序列，评估单碱基错误率,纯合SNP及纯合INDEL的占比,两者比例均不超过0.05%。

6.4 初级要求

基因组从头组装质量分级为初级的要求如下：

（1）基因组从头组装结果的叠连群N50大于10 Kbp 。

（2）基因组从头组装结果的骨架序列N50大于500 Kbp。

（3）Hi-C挂载率高于80%（N区长度设定为100 bp）。

（4）BUSCO组装质量评估，缺失基因的比例低于15%，或者与近源物种相比，缺失基因的比例不比近源物种高5%。

（5）高通量测序读长比回率大于90%（样品有污染或者高杂合的情况除外）。

（6）对于能提供多倍体基因组结果的方法，多倍体基因组定相块N50长度大于10 Kbp。

（7）高通量测序读长在基因组中的覆盖度，对于含有高通量测序数据的基因组，需将高通量测序读长比对回组装序列，统计不同深度下，组装序列被高通量测序数据覆盖的比例，1X以上的覆盖度大于80% 。

（8）对于含有高通量测序读长数据的组装，须将读长比对回组装序列，评估单碱基错误率,纯合SNP及纯合INDEL的占比,两者比例均不超过0.1%。

附 录 A
（资料性附录）
基因组测序技术列举

本标准适用的基因组建库及测序技术，如Illumina的高通量测序技术、Oxford Nanopore Technologies的单分子纳米孔测序技术、PacBio的单分子实时测序技术（Single Molecule Real-Time Sequencing，SMRTs）、10x GENOMICS的Linked-reads建库技术、华大智造MGI的单管长片段（Single tube Long Fragment Read, stLFR）建库技术等。

参 考 文 献

[1]Teeling E , Vernes S , Davalos L M , et al. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for all Living Bat Species[J]. Annual Review of Animal Biosciences, 2018, 6(1):annurev-animal-022516-022811.

[2]Nagarajan N, Pop M. Sequence assembly demystified[J]. Nature Reviews Genetics, 2013, 14(3):157-67