Zhang-Lab 生信小课堂 第廿五期

Applied Bioinformatics Club（ABC）

和趣求真 乘实生信

（张建伟生物信息学课题组 https://zhang.hzau.edu.cn）

三代测序数据组装参数的选择

2025.12.05 二综C102 15:00 欢迎大家交流学习！

主讲人：王欢

2025/12/05

目录

01 三代HiFi测序数据介绍与质控

02 HiFi数据组装

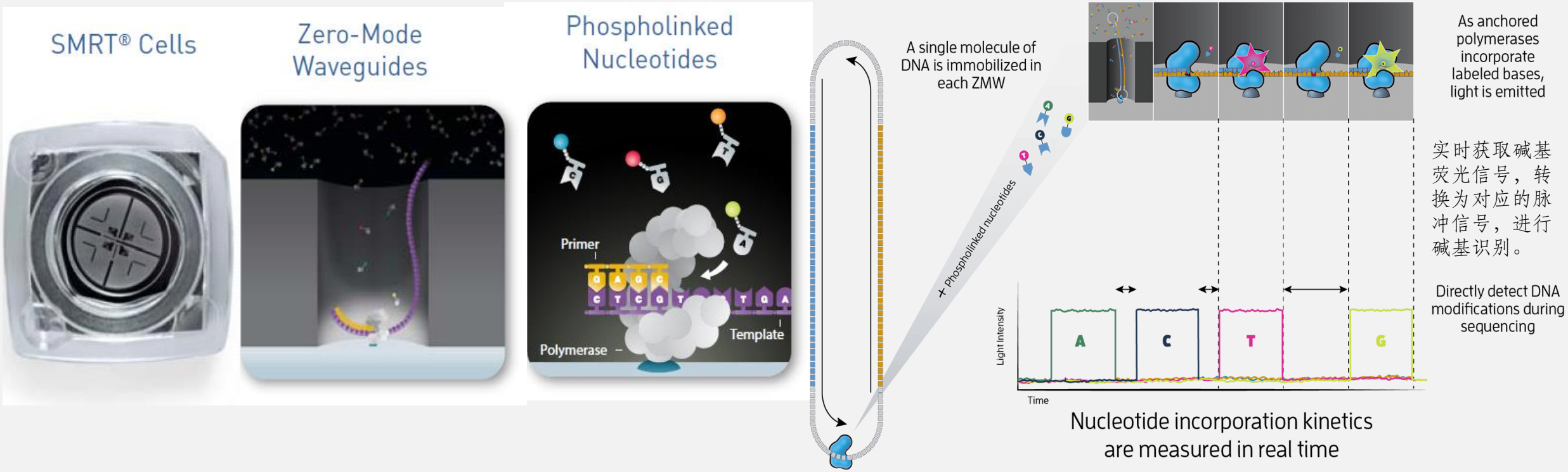
Part 01

三代测序数据介绍与质控

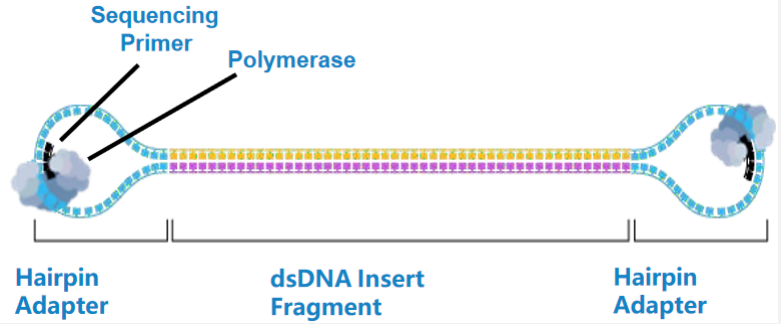
1.1 PacBio HiFi数据介绍

PacBio三代测序技术又称为SMRT测序（Single Molecule Real-Time）：

➢ 零模波导孔技术（Zero-mode Waveguides, ZMWs），对大量单分子DNA测序，减少了背景噪音干扰

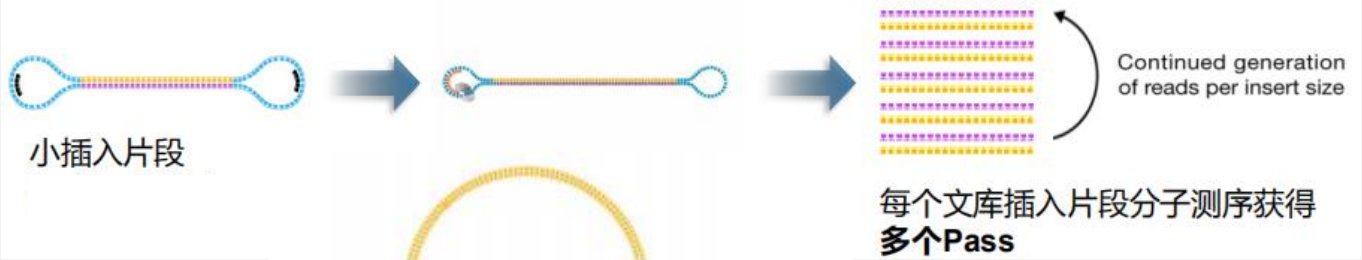


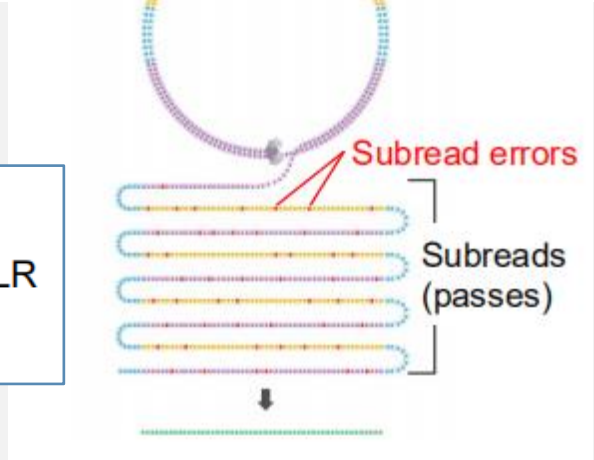
1.1 PacBio HiFi数据介绍



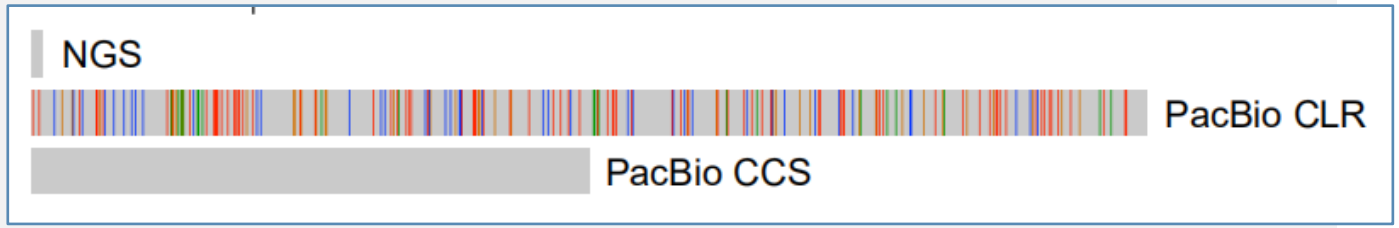
CLR

CCS





HiFi



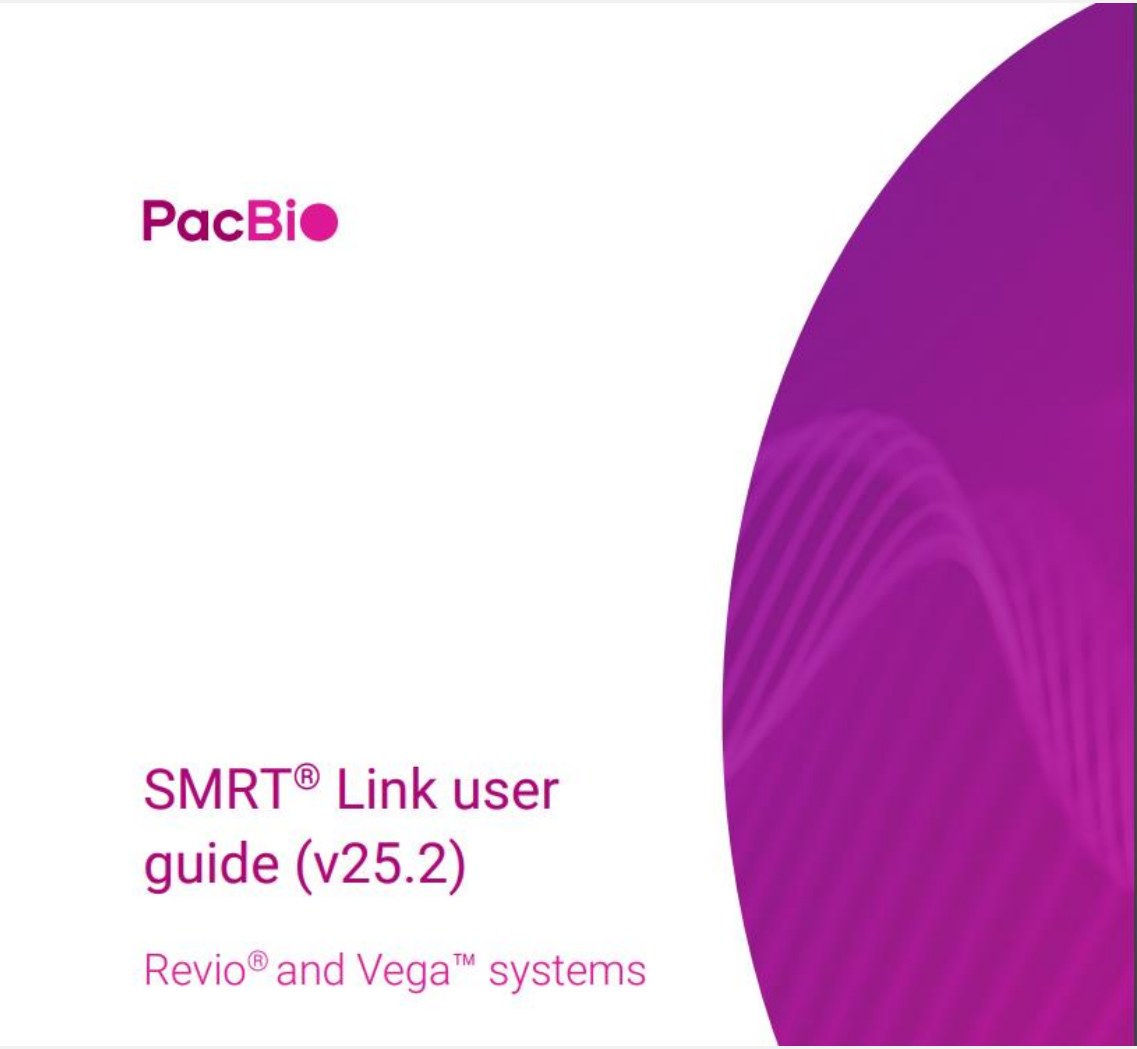
1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

1. Smrtlink 简介：

① PacBio开发的处理HiFi测序数据的open-source工具；

② 基于网页的端对端工作流程管理，也可以用linux command-line分析；

③ 它包括用于构建样品，监控测序过程，分析和测序数据的软件应用。



https://www.pacb.com/wp-content/uploads/SMRT-Link-v25.2-user-guide.pdf

1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

2. smrtlink下载和安装

官网： https://www.pacb.com/support/software-downloads/

下载： smrtlink-release_25.2.0.266456.zip

安装：

```
$ unzip smrtlink-release_25.2.0.266456.zip
$ ./smrtlink-release_25.2.0.266456_linux_x86-64_libc-2.17_anydistro.run
```

- 根据提示一步一步进行配置，基本都是直接使用默认值，按Enter键即可。
- 设置软件临时文件夹的时候，要选择一个较大分区所对应的目录。

1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

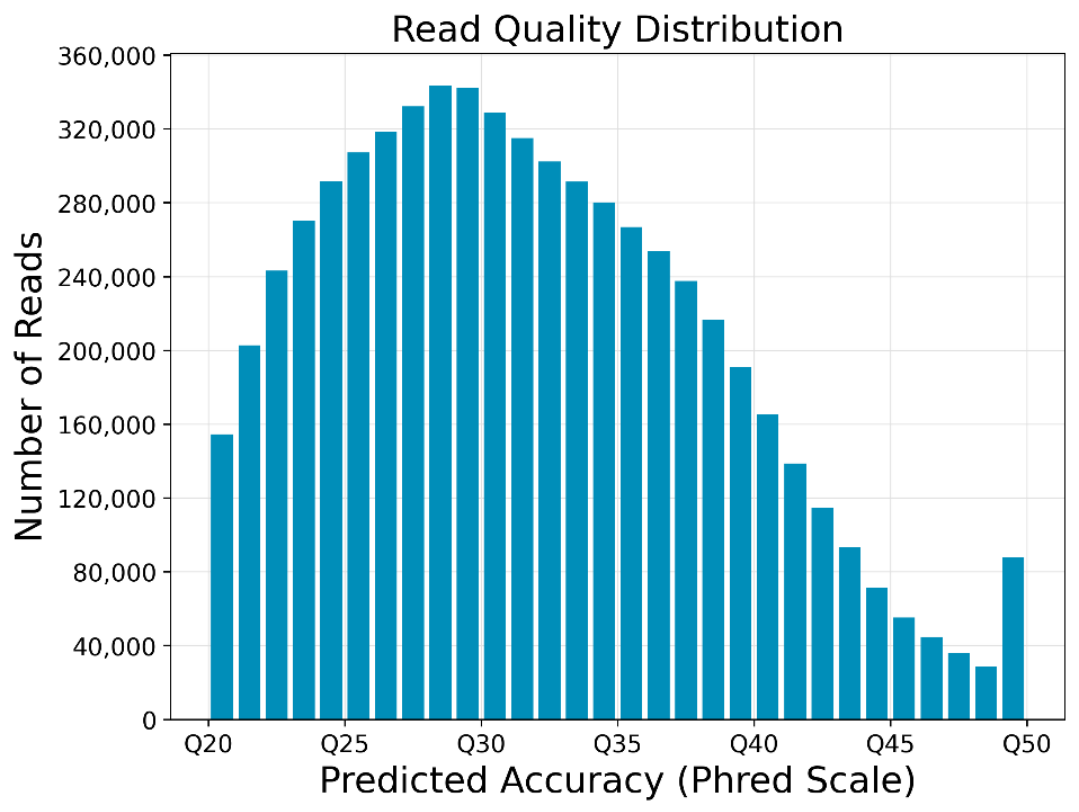
3. Smrtlink运行runqc-reports: HiFi数据统计

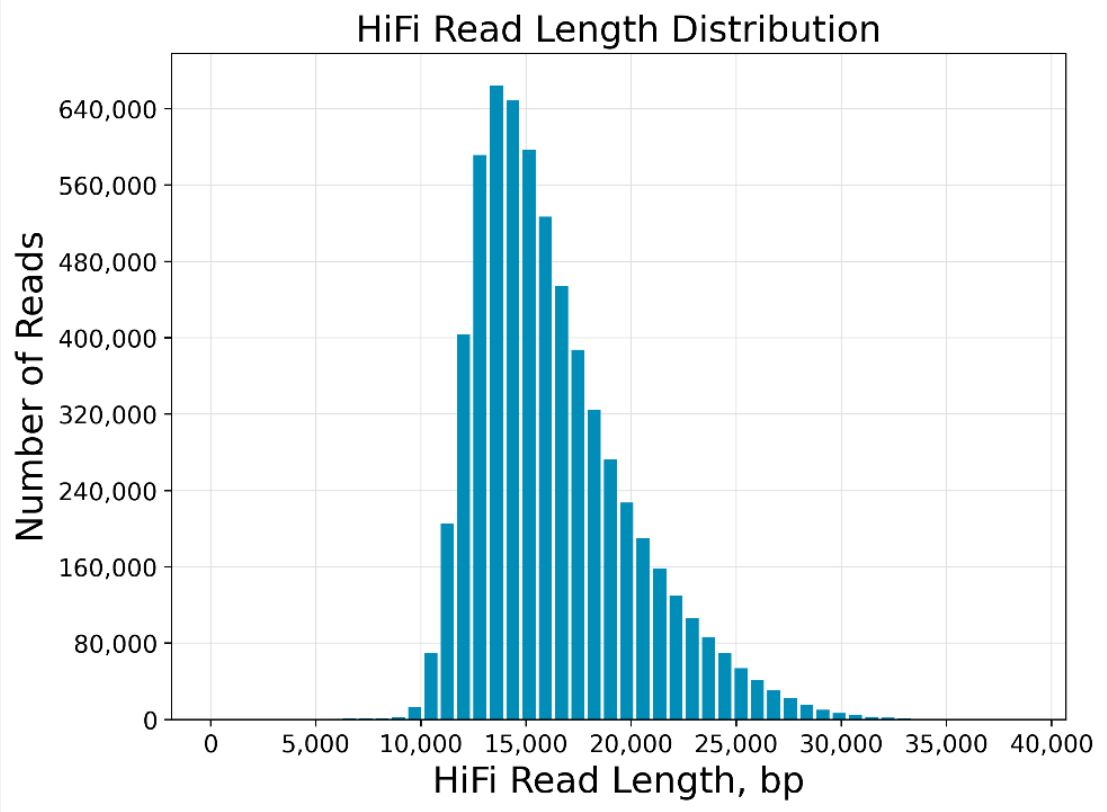
```
$ HiFibam = sample_hifi.bam
$ prefix= sample_hifi
$ export PATH=/work/fscourse/software/smrtlink/smrtlink/smrtcmds/bin:$PATH
$ dataset create --force --type ConsensusReadSet ${prefix}.xml ${HiFibam}
$ runqc-reports ${prefix}.xml
```

- dateset命令： 输入HiFi.bam， 输出HiFi.xml
- runqc-reports命令： 输入HiFi.xml，输出HiFi的数据统计结果

1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

4. Smrtlink运行runqc-reports输出结果: HiFi reads质量分布和长度分布





1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

4. Smrtlink运行runqc-reports输出结果: ccs.report.json


```
"_comment": "Generated with pbcommand version  at 2023-07-19T11:53
"attributes": [
    {
        "id": "ccs2.number_of_ccs_reads",
        "name": "HiFi Reads",
        "value": 6328444
    },
    {
        "id": "ccs2.total_number_of_ccs_bases",
        "name": "HiFi Yield (bp)",
        "value": 102970178540
    },
    {
        "id": "ccs2.mean_ccs_readlength",
        "name": "HiFi Read Length (mean, bp)",
        "value": 16271
    },
    {
        "id": "ccs2.median_ccs_readlength",
        "name": "HiFi Read Length (median, bp)",
        "value": 15478
    },
    {
        "id": "ccs2.ccs_readlength_n50",
        "name": "HiFi Read Length N50 (bp)",
        "value": 16308
    },
    {
        "id": "ccs2.median_accuracy",
        "name": "HiFi Read Quality (median)",
        "value": "Q31"
```


HiFi Yield (bp)	102,970,178,540
HiFi Reads	6,328,444
HiFi Read Length N50 (bp)	16,308
HiFi Read Length (mean, bp)	16,271
HiFi Read Quality	Q31

1.2 PacBio HiFi数据质控 - Smrtlink软件的使用

5. HiFi.bam转HiFi.fasta格式

```
$ HiFibam = sample_hifi.bam
$ prefix= sample_hifi
$ bam2fasta ${HiFibam} -o ${prefix} -u
```

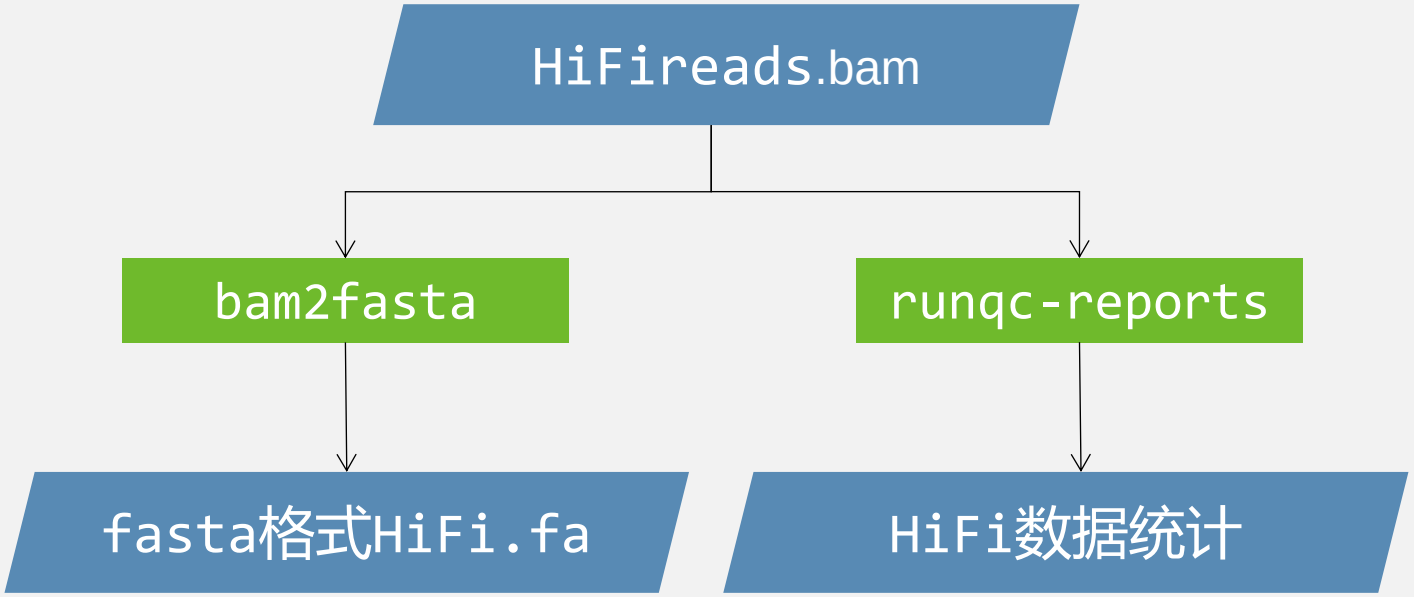
注意事项:

1. 需要Smrtlink25.1或更新版本的bam2fasta

>|> 20230720 06:40:19.529 -|- FATAL -|- Run -|- 0x2baa610d0880|| -|- bam2fasta ERROR: map::at

2.目录下需要有index文件，即bam.pbi文件

1.3 PacBio HiFi数据质控总结



Part 02

HiFi数据组装

2.1 HiFiasm软件介绍和安装

1. HiFiasm简介：

① 基于HiFi reads的快速单倍型从头组装软件；速度快，资源消耗少；

② 给定家系（父母本、子代）数据的情况下，可实现子代来自不同亲本的单倍体的组装；

③ 2022年更新：可使用HiC数据实现单倍型fully phased组装；

④ 2024年更新：自从0.17.3版本，可使用ultra-long ONT整合模式；

⑤ 2025年更新：自从0.21.0版本，支持单独使用ONT数据组装。

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › nature methods › brief communications › article

Brief Communication | Published: 10 May 2024

Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph

Haoyu Cheng, Mobin Asri, Julian Lucas, Sergey Koren & Heng Li

Nature Methods **21**, 967–970 (2024) | Cite this article

BRIEF COMMUNICATION
https://doi.org/10.1038/s41587-022-01261-x
nature biotechnology
Check for updates

Haplotype-resolved assembly of diploid genomes without parental data

Haoyu Cheng[1,2], Erich D. Jarvis[3,4], Olivier Fedrigo[3], Klaus-Peter Koepfli[5,6,7], Lara Urban[8], Neil J. Gemmell[8] and Heng Li[1,2]

Routine haplotype-resolved genome assembly from single samples remains an unresolved problem. Here we describe an algorithm that combines PacBio HiFi reads and Hi-C chromatin interaction data to produce a haplotype-resolved

hifiasm (Hi-C) here, is built on top of phased hifiasm assembly graphs[2] but differs from the published hifiasm (trio) algorithm in sequence partition. In a hifiasm graph, each node is a unitig assembled from HiFi reads with correct phasing, and each edge repre-

ARTICLES
https://doi.org/10.1038/s41592-020-01056-5
nature methods
Check for updates

Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm

Haoyu Cheng[1,2], Gregory T. Concepcion[3], Xiaowen Feng[1,2], Haowen Zhang[4] and Heng Li[1,2]

Haplotype-resolved de novo assembly is the ultimate solution to the study of sequence variations in a genome. However, existing algorithms either collapse heterozygous alleles into one consensus copy or fail to cleanly separate the haplotypes to produce high-quality phased assemblies. Here we describe hifiasm, a de novo assembler that takes advantage of long high-fidelity sequence reads to faithfully represent the haplotype information in a phased assembly graph. Unlike other graph-based assemblers that only aim to maintain the contiguity of one haplotype, hifiasm strives to preserve the contiguity of all haplotypes. This feature enables the development of a graph trio binning algorithm that greatly advances over standard trio binning. On three human and five nonhuman datasets, including California redwood with a ~30-Gb hexaploid genome, we show that hifiasm frequently delivers better assemblies than existing tools and consistently outperforms others on haplotype-resolved assembly.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.14.648685; this version posted April 17, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

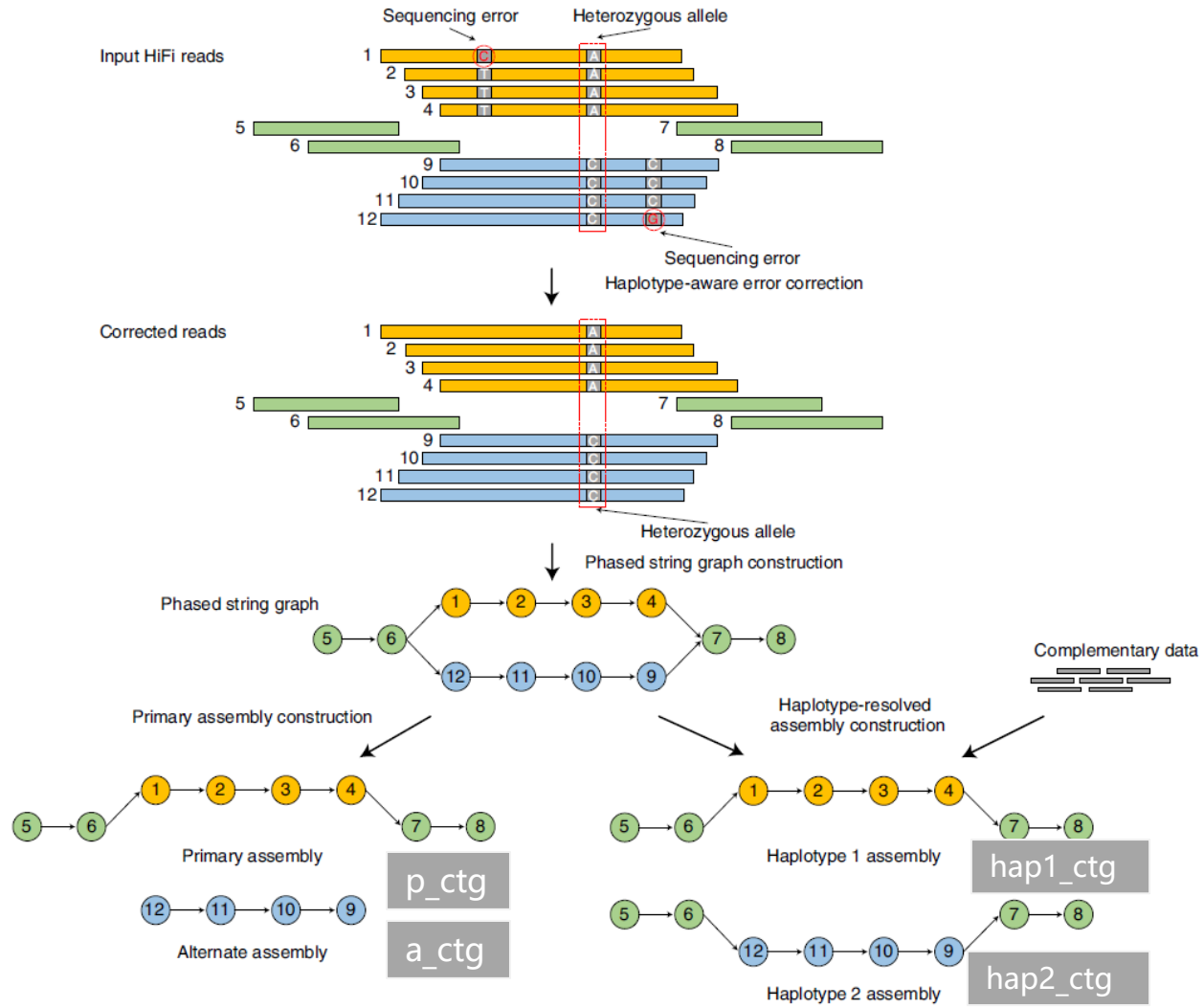
Efficient near telomere-to-telomere assembly of Nanopore Simplex reads

Haoyu Cheng[1,5,*], Han Qu[3,5], Sean McKenzie[4], Katherine R. Lawrence[4], Rhydian Windsor[4], Mike Vella[4], Peter J. Park[3], and Heng Li[2,3,*]

[1]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA
[2]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[4]Oxford Nanopore Technologies, Oxford, United Kingdom
[5]Authors contributed equally
[*]To whom correspondence should be addressed: haoyu.cheng@yale.edu, hli@ds.dfci.harvard.edu

2.1 HiFiasm软件介绍和安装

2. HiFiasm原理：



Step1 Reads互相纠错：将所有的reads进行all-vs-all比对并进行纠错。基于reads间的overlap信息，如果read上有一个碱基与其他碱基不同，并有至少3条reads支持，则认为它是SNP并保留，否则认为是错误并进行纠正。HiFiasm只使用相同单倍型的数据纠错，避免过度校正，保留来自不同单倍型的杂合变异信息，HiFiasm可以对杂合SNP进行定相phasing。

Step2 String-graph构建：HiFiasm构建以reads为顶点、重叠区为边的定相string-graph。区别于一般三代数据组装时构建的string-graph，Hifiasm会保留全部的气泡（bubble），因而可以保留下来基因组上全部的单倍型信息，以便后续对于单倍型的处理。

Step 3 组装构建：如果没有其他辅助分型数据，Hifiasm会任意选择bubble的一侧为主要组装结果（primary contigs），类似Falcon unzip和HiCanu。如果有父母本的测序数据，Hifiasm可以通过亲本特有的kmer在图上识别出来自父母本的序列，从而得到两套单倍体基因组。

2.1 HiFiasm软件介绍和安装

3. HiFiasm安装

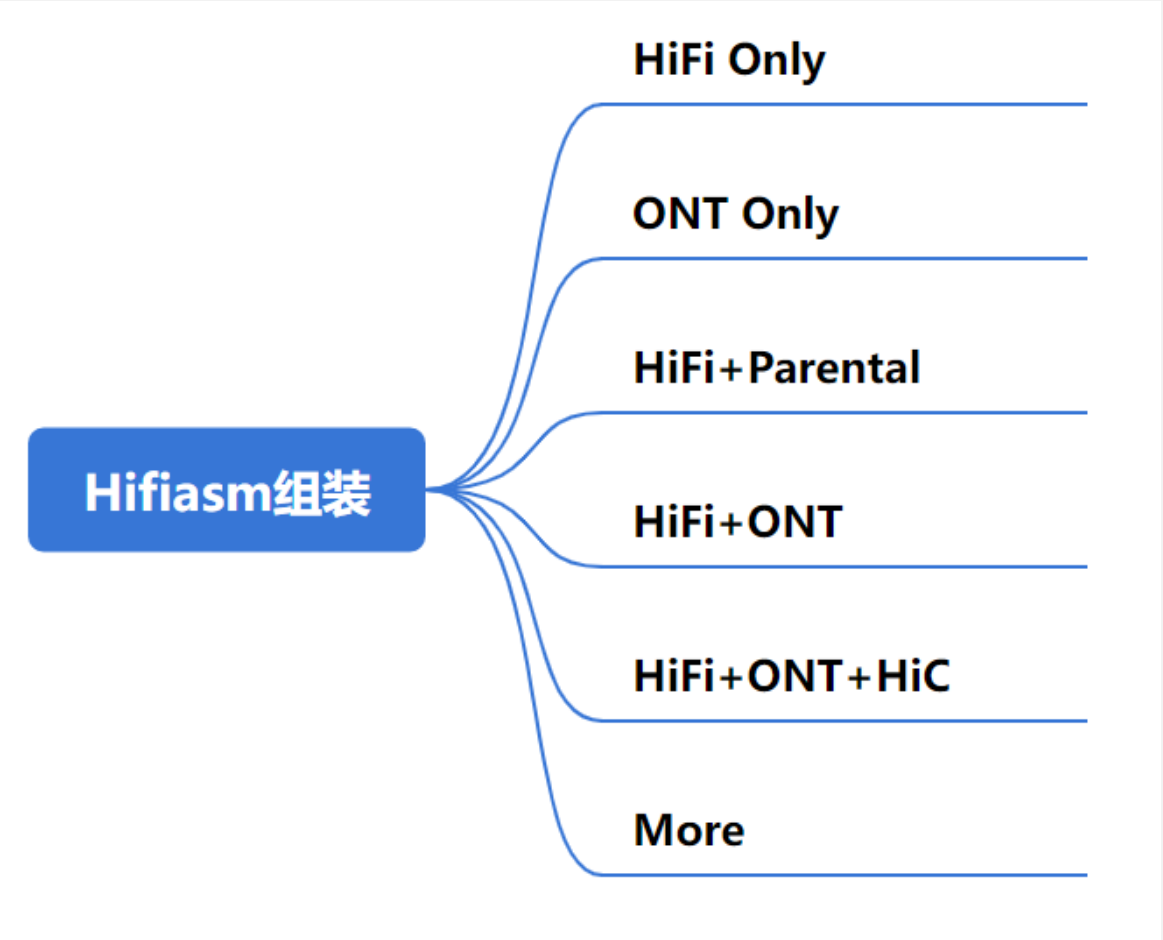
官网：https://github.com/chhylp123/hifiasm

安装:

```
$ git clone https://github.com/chhylp123/hifiasm
$ cd hifiasm
$ make
```

2.2 HiFiasm软件运行

运行类型



2.2 HiFiasm软件运行

1. HiFi only assembly（Partially单倍体组装）

```
$ HiFi= …/deliver/rawdata/HiFi.fastq.gz
$ prefix=Fragaria_vesca

$ hifiasm -o ${prefix} -t 26 -l3  ${HiFi} 2> out.log
$ awk '/^S/{print ">"$2;print $3}' ${prefix}.bp.p_ctg.gfa > ${prefix}.bp.p_ctg.fa
$ /usr/bin/java -jar /path/gnx.jar ${prefix}.bp.p_ctg.fa > N50
```

- -t 线程数
- -o 输出前缀
- -l 去杂合等级参数(l3 for unzip, l0用于纯和/自交基因组)
- -s 去杂合相似性(默认0.55 for -l3)
- --primary 输出主要组装（p_ctg）和替代组装（a_ctg）

HiFi Only输出文件	文件说明
prefix.bp.p_ctg.gfa	assembly graph of primary contigs.
prefix.bp.a_ctg.gfa	assembly graph of alternate contigs.
prefix.bp.hap1.p_ctg.gfa	partially phased contig graph of haplotype1.
prefix.bp.hap2.p_ctg.gfa	partially phased contig graph of haplotype2.

2.2 HiFiasm软件运行

2. ONT only assembly

```
$ ONT= …/deliver/rawdata/ONT.fastq.gz
$ prefix=Rosa_hybrida

$ hifiasm --primary -o ${prefix} -t 30 -l0 --ont $ONT > out.log
$ awk '/^S/{print ">"$2;print $3}' ${prefix}.p_ctg.gfa > ${prefix}.p_ctg.fa
$ /usr/bin/java -jar gnx.jar ${prefix}.p_ctg.fa > N50
```

- --ont 单独使用ONT组装

ONT Only输出文件	文件说明
prefix.p_utg.gfa	Unitig graph without small bubbles
prefix.p_ctg.gfa	Assembly graph of primary contigs
prefix.a_ctg.gfa	Assembly graph of alternate contigs

2.2 HiFiasm软件运行

3. Trio-binning Assembly（有家系数据-Fully单倍体组装分型）

```
$ HiFi= …/deliver/rawdata/HiFi.fastq.gz
$ prefix=Fragaria_vesca

$ yak count –k31 -b37 -t16 -o pat.yak paternal.fq.gz
$ yak count –k31 -b37 -t16 -o mat.yak maternal.fq.gz
$ hifiasm -o ${prefix} -t26 -1 pat.yak -2 mat.yak ${HiFi}
```

```
#   Trio分型模式默认不进行purge haplotig
#   需要提前安装yak：https://github.com/lh3/yak
```

trio-binning输出文件	文件说明
prefix.dip.hap1.p_ctg.gfa	fully phased paternal/haplotype1 contig graph keeping the phased paternal/haplotype1 assembly.
prefix.dip.hap2.p_ctg.gfa	fully phased maternal/haplotype2 contig graph keeping the phased maternal/haplotype2 assembly.

2.2 HiFiasm软件运行

4. Ultra-long ONT integration Assembly （HiFi+ONT：Partially单倍体组装）

```
$ HiFi= …/deliver/rawdata/HiFi.fastq.gz
$ ONT= …/deliver/rawdata/ONT.fastq.gz
$ prefix=Fragaria_vesca

$ hifiasm -o ${prefix} -t 26 -l3 --ul ${ONT} ${HiFi} 2> out.log
$ awk '/^S/{print ">"$2;print $3}' ${prefix}.bp.p_ctg.gfa > ${prefix}.bp.p_ctg.fa
$ /usr/bin/java -jar /path/gnx.jar ${prefix}.bp.p_ctg.fa > N50
```

输出结果与HiFi only assembly类似

加入ONT数据提高组装的连续性，旨在生成telomere-to-telomere assembly

HiFi Only输出文件	文件说明
prefix.bp.p_ctg.gfa	assembly graph of primary contigs.
prefix.bp.hap1.p_ctg.gfa	partially phased contig graph of haplotype1.
prefix.bp.hap2.p_ctg.gfa	partially phased contig graph of haplotype2.

2.2 HiFiasm软件运行

5. Ultra-long ONT + Hi-C Integrated Assembly (HiFi+ONT+HiC：Fully单倍体组装)

```
$ HiFi= …/deliver/rawdata/HiFi.fastq
$ ONT= …/deliver/rawdata/ONT.fastq
$ hic1= …/deliver/rawdata/HiC_1.fastq.gz
$ hic2= …/deliver/rawdata/HiC_2.fastq.gz
$ prefix=Fragaria_vesca

$ hifiasm  -o ${prefix} -t 26 --ul ${ONT} --h1 ${hic1} --h2 ${hic2} ${HiFi} 2> out.log
$ awk '/^S/{print ">"$2;print $3}' ${prefix}.hic.hap1.p_ctg.gfa > ${prefix}.hic.hap1.p_ctg.fa
$ awk '/^S/{print ">"$2;print $3}' ${prefix}.hic.hap2.p_ctg.gfa > ${prefix}.hic.hap2.p_ctg.fa
```

Hi-C partition输出文件	文件说明
prefix.hic.p_ctg.gfa	assembly graph of primary contigs.
prefix.hic.hap1.p_ctg.gfa	fully phased contig graph of haplotype1 where each contig is fully phased.
prefix.hic.hap2.p_ctg.gfa	fully phased contig graph of haplotype2 where each contig is fully phased.

2.2 HiFiasm软件运行

6. 多倍体组装

```
$ HiFi = …/deliver/rawdata/HiFi.fastq
$ prefix=Fragaria_vesca

$ hifiasm --primary --n-hap 4 -o ${prefix} -t 26 ${HiFi} 2> out.log
```

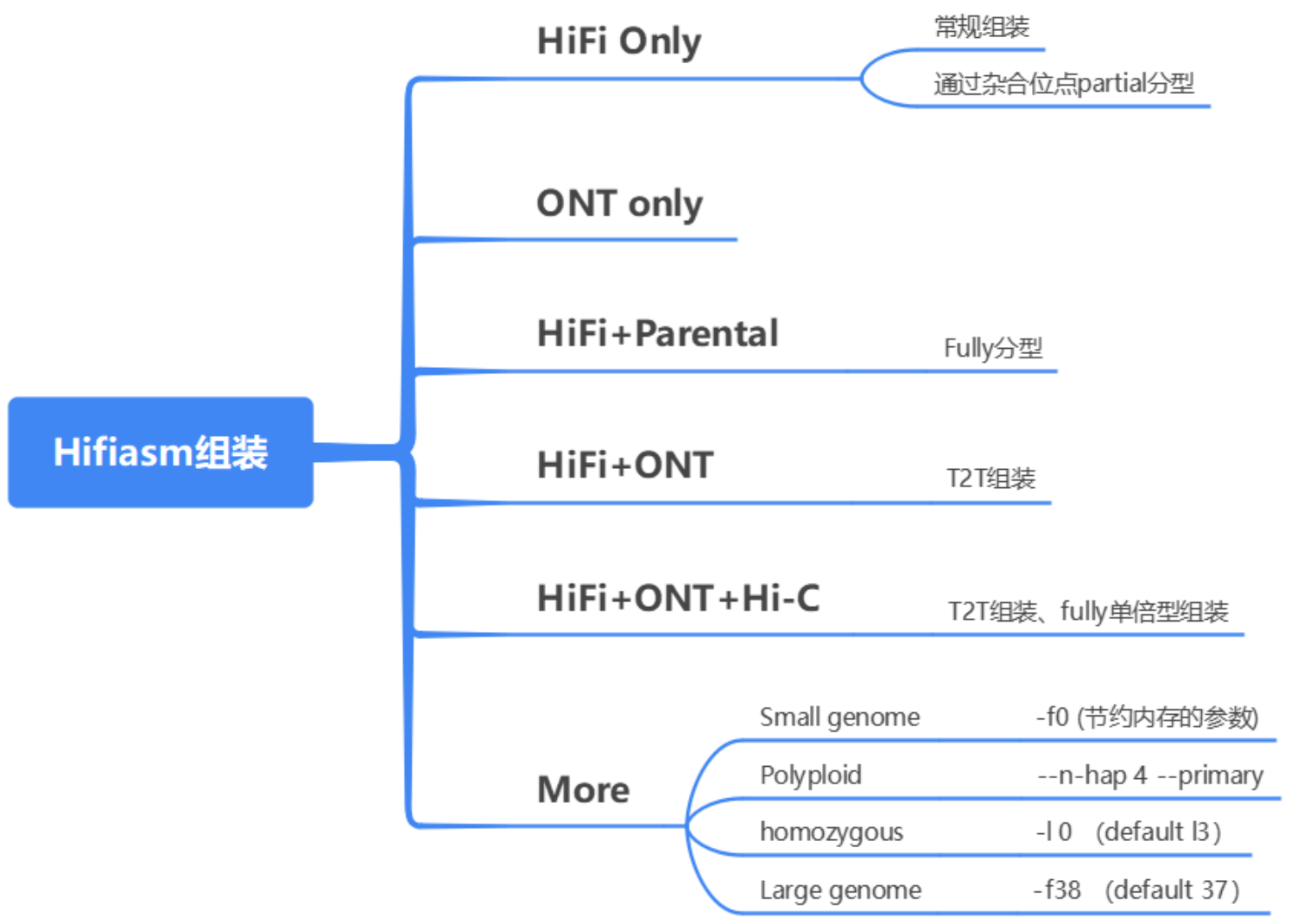
- 默认hifiasm组装模块是为二倍体设计的，不支持多倍体多个Hap的分型；
- 将--n-hap 设置为 > 2，使用--primary 模式，有助于将单倍型信息保留在组装图中。

Are polyploid genomes supported?

The `*r_utg.gfa` and `*p_utg.gfa` are lossless so that they also work for polyploid genomes. However, currently the contig-generation modules of hifiasm are designed for diploid samples, which means both the partially phased assembly and the fully-phased assembly does not directly support polyploid genomes. If it is set to >2, the quality of primary assembly for polyploid genomes might be improved. Please use primary assembly for polyploid samples and run multiple rounds of purging steps using third-party tools such as purge_dups.

2.3 HiFiasm软件小结

运行小结



2.3 HiFiasm软件小结

常见问题

Hifiasm 常见问题

- **1.超高杂合基因组partially phased时hap1和hap2大小差异明显?**
  - 减小-s（default: 0.55）：去掉更多杂合
  - 指定--hom-cov
- **2.Primary的N50连续性为何一般高于hap1和hap2?**
  - Primary有个额外的joining step，链接两个单倍型
  - Alternate assembly格外碎片化
  - 部分纯和区域的cov过低时，hap1/hap2在该处有contig break，但primary会选其中一个hap的higher cov去填补该break
- **3.组装连续性不够时，如何提高?**
  - 提高-D or -N 参数：增长repeat区域分辨率
  - 提高--purge-max：会影响repeats和SD区域组装
- **4.如何避免misassemblies?**
  - 减小--purge-max, -s和-O参数 去掉更多杂合
  - 使用-u：禁止post-join step提高N50，但可能会引入错误

Q&A