

基因组组装

日期: 2022-11-23

实验者: 生信 2001 张子栋

[MarkdownNotes/软件第4次作业.md at main · Bluuur/MarkdownNotes \(github.com\)](#)

[生物信息学原理/软件第4次作业.md · blur/MarkdownNotes - 码云 - 开源中国 \(gitee.com\)](#)

实验目的

1. 回顾 Linux 系统的常用命令的使用。
2. 掌握常用三代和二代测序组装软件各至少一种的使用，并理解关键参数的含义，熟悉测序数据 fastq 等格式。
3. 会编写程序计算 N50 / L50 等组装连续性指标。
4. 会使用基因组比对工具 MUMmer 进行序列比对，并寻找 SNP 等变异。

实验内容

- 使用组装软件 SOAPdenovo、canu、Hifiasm 分别组装大肠杆菌 *Escherichia coli* K12 基因组的二代和三代测序数据。
- 编写程序计算 SOAPdenovo 组装 contig 和 scaffold 序列的 N50 / L50 等组装质量评估指标。
- 使用基因组比对工具 MUMmer 比较 canu 组装的 contig 序列和 hifiasm 组装的 contig 序列，并寻找两者之间的序列差异。

实验结果

使用 soapdenovo2 组装 *E. coli* 基因组 illumina 二代测序数据

```
[uu01@localhost soap]$ pwd
/home/uu01/01_zzd/soap
[uu01@localhost soap]$ date
Thu Nov 24 14:10:40 CST 2022
[uu01@localhost soap]$ ls -lh
total 250M
-rw-r--r--. 1 uu01 WBJIA0 1.5M Nov 23 21:31 ecoli-soap.Arc
-rw-r--r--. 1 uu01 WBJIA0 0 Nov 23 21:32 ecoli-soap.bubbleInScaff
-rw-r--r--. 1 uu01 WBJIA0 9.3M Nov 23 21:31 ecoli-soap.contig
-rw-r--r--. 1 uu01 WBJIA0 497K Nov 23 21:31 ecoli-soap.ContigIndex
-rw-r--r--. 1 uu01 WBJIA0 204K Nov 23 21:32 ecoli-soap.contigPosInScaff
-rw-r--r--. 1 uu01 WBJIA0 30M Nov 23 21:30 ecoli-soap.edge.gz
-rw-r--r--. 1 uu01 WBJIA0 0 Nov 23 21:32 ecoli-soap.gapSeq
-rw-r--r--. 1 uu01 WBJIA0 1.1K Nov 23 21:30 ecoli-soap.kmerFreq
-rw-r--r--. 1 uu01 WBJIA0 2.9M Nov 23 21:32 ecoli-soap.links
-rw-r--r--. 1 uu01 WBJIA0 3.8M Nov 23 21:31 ecoli-soap.markOnEdge
-rw-r--r--. 1 uu01 WBJIA0 1.2M Nov 23 21:32 ecoli-soap.newContigIndex
-rw-r--r--. 1 uu01 WBJIA0 91M Nov 23 21:31 ecoli-soap.path
-rw-r--r--. 1 uu01 WBJIA0 41 Nov 23 21:32 ecoli-soap.peGrads
-rw-r--r--. 1 uu01 WBJIA0 22M Nov 23 21:31 ecoli-soap.preArc
-rw-r--r--. 1 uu01 WBJIA0 78 Nov 23 21:31 ecoli-soap.preGraphBasic
-rw-r--r--. 1 uu01 WBJIA0 41M Nov 23 21:32 ecoli-soap.readInGap.gz
-rw-r--r--. 1 uu01 WBJIA0 20M Nov 23 21:32 ecoli-soap.readOnContig.gz
-rw-r--r--. 1 uu01 WBJIA0 1.6M Nov 23 21:32 ecoli-soap.scaf
-rw-r--r--. 1 uu01 WBJIA0 270K Nov 23 21:32 ecoli-soap.scaf_gap
-rw-r--r--. 1 uu01 WBJIA0 7.4M Nov 23 21:32 ecoli-soap.scafSeq
-rw-r--r--. 1 uu01 WBJIA0 1.7K Nov 23 21:32 ecoli-soap.scafStatistics
-rw-r--r--. 1 uu01 WBJIA0 6.2M Nov 23 21:31 ecoli-soap.updated.edge
-rw-r--r--. 1 uu01 WBJIA0 14M Nov 23 21:31 ecoli-soap.vertex
-rw-----. 1 uu01 WBJIA0 10K Nov 23 21:32 nohup.out
```

使用 canu 组装 Nanopore 测序读段

```
[uu01@localhost canu-ont]$ pwd
/home/uu01/01_zzd/canu-ont
[uu01@localhost canu-ont]$ date
Thu Nov 24 14:12:27 CST 2022
[uu01@localhost canu-ont]$ ls -lh
total 64M
drwxr-xr-x. 2 uu01 WBJIA0 67 Nov 23 21:36 canu-logs
drwxr-xr-x. 2 uu01 WBJIA0 10 Nov 23 21:36 canu-scripts
drwxr-xr-x. 7 uu01 WBJIA0 263 Nov 23 21:44 correction
-rw-r--r--. 1 uu01 WBJIA0 4.5M Nov 23 21:47 ecoli-ont.contigs.fasta
-rw-r--r--. 1 uu01 WBJIA0 313K Nov 23 21:47 ecoli-ont.contigs.layout.readToTig
-rw-r--r--. 1 uu01 WBJIA0 226 Nov 23 21:47 ecoli-ont.contigs.layout.tigInfo
-rw-r--r--. 1 uu01 WBJIA0 30M Nov 23 21:44 ecoli-ont.correctedReads.fasta.gz
-rw-r--r--. 1 uu01 WBJIA0 41K Nov 23 21:47 ecoli-ont.report
drwxr-xr-x. 5 uu01 WBJIA0 4.0K Nov 23 21:45 ecoli-ont.seqStore
-rw-r--r--. 1 uu01 WBJIA0 738 Nov 23 21:36 ecoli-ont.seqStore.err
-rwxr-xr-x. 1 uu01 WBJIA0 925 Nov 23 21:36 ecoli-ont.seqStore.sh
-rw-r--r--. 1 uu01 WBJIA0 29M Nov 23 21:45 ecoli-ont.trimmedReads.fasta.gz
-rw-r--r--. 1 uu01 WBJIA0 14K Nov 23 21:47 ecoli-ont.unassembled.fasta
-rw-----. 1 uu01 WBJIA0 81K Nov 23 21:47 nohup.out
drwxr-xr-x. 6 uu01 WBJIA0 199 Nov 23 21:45 trimming
drwxr-xr-x. 9 uu01 WBJIA0 4.0K Nov 23 21:46 unitigging
```


使用 hifiasm 组装 PacBio HiFi 测序数据

```
[uu01@localhost hifiasm]$ pwd
/home/uu01/01_zzd/hifiasm
[uu01@localhost hifiasm]$ date
Thu Nov 24 14:13:42 CST 2022
[uu01@localhost hifiasm]$ ls -lh
total 2.6G
-rw-r--r--. 1 uu01 WBJIAO 5.0M Nov 24 07:01 ecoli-hifi.bp.hap1.p_ctg.gfa
-rw-r--r--. 1 uu01 WBJIAO 8.6K Nov 24 07:01 ecoli-hifi.bp.hap1.p_ctg.lowQ.bed
-rw-r--r--. 1 uu01 WBJIAO 508K Nov 24 07:01 ecoli-hifi.bp.hap1.p_ctg.noseq.gfa
-rw-r--r--. 1 uu01 WBJIAO 5.0M Nov 24 07:01 ecoli-hifi.bp.hap2.p_ctg.gfa
-rw-r--r--. 1 uu01 WBJIAO 8.6K Nov 24 07:01 ecoli-hifi.bp.hap2.p_ctg.lowQ.bed
-rw-r--r--. 1 uu01 WBJIAO 504K Nov 24 07:01 ecoli-hifi.bp.hap2.p_ctg.noseq.gfa
-rw-r--r--. 1 uu01 WBJIAO 5.0M Nov 24 07:01 ecoli-hifi.bp.p_ctg.gfa
-rw-r--r--. 1 uu01 WBJIAO 8.6K Nov 24 07:01 ecoli-hifi.bp.p_ctg.lowQ.bed
-rw-r--r--. 1 uu01 WBJIAO 501K Nov 24 07:01 ecoli-hifi.bp.p_ctg.noseq.gfa
-rw-r--r--. 1 uu01 WBJIAO 5.2M Nov 24 07:01 ecoli-hifi.bp.p_ctg.gfa
-rw-r--r--. 1 uu01 WBJIAO 19K Nov 24 07:01 ecoli-hifi.bp.p_ctg.lowQ.bed
-rw-r--r--. 1 uu01 WBJIAO 502K Nov 24 07:01 ecoli-hifi.bp.p_ctg.noseq.gfa
-rw-r--r--. 1 uu01 WBJIAO 5.2M Nov 24 07:01 ecoli-hifi.bp.r_ctg.gfa
-rw-r--r--. 1 uu01 WBJIAO 19K Nov 24 07:01 ecoli-hifi.bp.r_ctg.lowQ.bed
-rw-r--r--. 1 uu01 WBJIAO 502K Nov 24 07:01 ecoli-hifi.bp.r_ctg.noseq.gfa
-rw-r--r--. 1 uu01 WBJIAO 337M Nov 24 06:59 ecoli-hifi.ec.bin
-rw-r--r--. 1 uu01 WBJIAO 4.6M Nov 24 07:00 ecoli-hifi.ovlp.reverse.bin
-rw-r--r--. 1 uu01 WBJIAO 2.2G Nov 24 07:00 ecoli-hifi.ovlp.source.bin
-rw-r--r--. 1 uu01 WBJIAO 4.5M Nov 24 13:06 ecoli-hifi.p_ctg.fa
-rw-r--r--. 1 uu01 WBJIAO 97K Nov 24 07:01 nohup.out
```

比较 canu 的 ONT 组装和 hifiasm 的 hifi 组装

```
[uu01@localhost wga]$ pwd
/home/uu01/01_zzd/wga
[uu01@localhost wga]$ date
Thu Nov 24 14:15:11 CST 2022
[uu01@localhost wga]$ ls -lh
total 568K
-rw-r--r--. 1 uu01 WBJIAO 0 Nov 24 13:09 nohup.out
-rw-r--r--. 1 uu01 WBJIAO 67K Nov 24 13:09 out.delta
-rw-r--r--. 1 uu01 WBJIAO 19K Nov 24 13:15 out.filter
-rw-r--r--. 1 uu01 WBJIAO 513 Nov 24 13:15 out.flt.tab
-rw-r--r--. 1 uu01 WBJIAO 476K Nov 24 13:16 out.snps
```

- SNP 数量: 1018
- indel 数量: 4702

Data	
 snps	5720 obs. of 14 variables
Values	
indel.num	4702L
snps.num	1018L

讨论

在这次上机实验中熟悉掌握了几种基因组组装软件和基因组比对工具。