

实验项目 2-1：序列组装【应用】

一、分析平台：

- 1.1. 硬件平台：（硬件配置）win10: CPU 1.70GHz 2.40GHz
Ubuntu: CPU 3.3GHz*2
- 1.2. 系统平台：（操作系统及其版本号）WIN10 专业版、Ubuntu【机房s06@HP06】、IBM_S812L 小型机
- 1.3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL 地址）R3.4.4、SUBLIME TEXT3
- 1.4. 数据库资源：Thermoanaerobacter pseudethanolicus 物种的全基因组序列下载地址（fasta 格式）https://www.ncbi.nlm.nih.gov/nucleotide/NC_010321.1 ；
- 1.5. 序列组装软件下载地址：
velvet_1.2.10 (<https://www.ebi.ac.uk/~zerbino/velvet/>)、
ALLPATHS (http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12)
CAP3(<http://seq.cs.iastate.edu/cap3.html>)

二、实验步骤：

2.1. 序列组装软件的概览：

2.1.1、软件搜索练习：

在 PubMed 数据库中，搜索“sequence assembly”，阅读搜索结果。列举至少五个基因组序列组装的常用程序，并简单介绍其功能和使用方法。

2.1.2、软件安装练习：

从网上下载三款与基因组序列拼接有关的软件进行安装，记录安装过程及软件运行测试情况。

[使用 linux 系统进行安装。]

2.1.2.1.cap3 下载安装

```
tar -xvf cap3.linux.x86_64.tar
cd CAP3
ls
./cap3
./cap3 ./example/xyz
```

2.1.2.2.SOAPdenovo2-src-r240 下载安装

```
tar -zxvf SOAPdenovo2-src-r240.tgz
cd SOAPdenovo2-src-r240
ls
##更改以上文件 multi_threads.cpp、pregraph_sparse.cpp、build_preArc.cpp,
##手动添加 #include "unistd.h"
make
./SOAPdenovo-63mer
./SOAPdenovo-127mer
```

||

2.1.2.3.velvet 下载安装

```
wget http://www.ebi.ac.uk/~zerbino/velvet/velvet\_1.2.10.tgz
tar zxvf velvet_1.2.10.tgz
cd velvet_1.2.10
make color 'CATEGORIES=57' 'MAXKMERLENGTH=127' 'BIGASSEMBLY=1' \
'LONGSEQUENCES=1' 'OPENMP=1'
||
##结果文件显示可执行的是 velveth_de 和 velvetg_de, 不是 velveth 和 velvetg
./velveth_de
./velvetg_de
```

2.2、待拼接序列数据源的准备:

2.2.1.各个拼接软件配套的测试数据:

2.2.1.1.Allpaths 数据: genome.fasta

【/home/student/s21/test.genome/ALLPATHS-LG.test_genome/test.genome】

2.2.1.2.SOAP 数据: short_pair_ill.A.fq、short_pair_ill.B.fq、long_pair_ill.A.fq、long_pair_ill.B.fq

(由 2.2.2.2 步获得)

【/home/student/s21】

2.2.1.3.velvet 数据: test_reference.fa

【/home/student/s21/velvet_1.2.10/data】

2.2.2.利用实验一所学习的全基因组测序模拟工具,生成一组单末端和双末端测序结果;同时保留原始基因组模板序列,用来评估序列组装结果;

2.2.2.1.下载数据:

https://www.ncbi.nlm.nih.gov/nuccore/NC_010321.1

2.2.2.2.模拟测序:

【为了避免 illumina 平台会出现 read len 不匹配的问题,所以此次实验使用 linux 系统进行模拟测序】

#短插入片段模拟

```
./art_illumina -ss HS25 -sam -i ./mydata3/NC_010321.1.fasta -p -l 150 -f 20 -m 200 -s 10 -o ./short_pair_ill
```

#长插入片段模拟

```
./art_illumina -ss HS25 -sam -i ./mydata3/NC_010321.1.fasta -mp -l 150 -f 20 -m 2500 -s 10 -o ./long_pair_ill
```

art 参数会自动选自平均片段大小>=2000 的-p 改为-mp

2.3、数据上传: 远程登录 IBM_S812L 小型机,上传 2.2.2.2 步获得的数据。

2.4、全基因组序列拼接软件的使用练习:

2.4.1、使用各个拼接软件配套的测试数据,进行组装实践;

2.4.1.1.allpaths-lg

#代码测试:

```
cd test.genome
cd ALLPATHS-LG.test_genome
prepare.sh
assemble.sh
```

2.4.1.2.SOAP

Step-1 数据准备:

测试使用 illumina 平台生成的短插入序列模拟片段(short_pair_ill.A.fq|short_pair_ill.B.fq) 和 长 插 入 序 列 模 拟 片 段 (long_pair_ill.A.fq|long_pair_ill.B.fq)

Step-2 更改配置文件"third_test.cfg":

```
max_rd_len=90
[LIB]
avg_ins=200
reverse_seq=0
asm_flags=3
rd_len_cutoff=90
rank=1
pair_num_cutoff=3
map_len=32
q1=short_pair_ill.A.fq
q2=short_pair_ill.B.fq
[LIB]
avg_ins=2500
reverse_seq=1
asm_flags=3
rank=2
pair_num_cutoff=5
map_len=35
q1=long_pair_ill.A.fq
q2=long_pair_ill.B.fq
```

Step-3 上传配置文件" third_test.cfg", 代码测试:

```
nohup SOAPdenovo-63mer all -s third_test.cfg -K 30 -o
third_test -p 20 &
```

Step-4查看"nohup.log"文档, 见结果3.5

2.4.1.3.velvet

```
cd velvet_1.2.10
velveth sillyDirectory 21 -shortPaired data/test_reads.fa
#生成 log、Roadmaps 和 Sequences 文件
velvetg sillyDirectory
#生成 contigs.fa 和其他文件
```

2.4.2、利用 blastn 将组装结果与改配套数据中基因组进行对比, 分析其中是否存在差异, 阐述原因

2.4.2.1.Allpaths:

Step-1 数据准备:

使用 genome.fasta 和 final.assembly.fasta 进行比对。

/home/student/s21/test.genome/ALLPATHS-LG.test_genome/test.genom
e 中下载 genome.fasta;

/home/student/s21/test.genome/ALLPATHS-LG.test_genome/test.genom
e/data/run/ASSEMBLIES/test 中下载 final.assembly.fasta。

Step-2 NCBI-blastn 比对

2.4.2.2.SOAPdenovo2-src-r240:

Step-1 数据准备:

使用生成的 SOAPdenovo_out.scafSeq 和原始数据 NC_010321.1.fasta

Step-2 NCBI-blastn 比对

Step-3 创建参考基因的 blast 数据库=》用于分析组装结果的 coverage

#构建数据库

makeblastdb -in NC_010321.1.fasta -input_type fasta -dbtype nucl -title
NC -out NC

#contigs 比对:

blastn -db NC -query SOAPdenovo_out.contig -out contig_blast_outfmt7
-evaluate 1e5 -outfmt 7 -max_target_seqs 1 -num_threads 20

blastn -db NC -query SOAPdenovo_out.contig -out contig_blast_outfmt6
-evaluate 1e5 -outfmt 6 -max_target_seqs 1 -num_threads 20

#scaffold 比对:

blastn -db NC -query SOAPdenovo_out.scafSeq -out
scaffold_blast_outfmt6 -evaluate 1e5 -outfmt 6 -max_target_seqs 1
-num_threads 20

2.4.2.3.velvet

使用 contigs.fa(组装结果)和 test_reference.fa(测试数据)进行比对

/home/student/s21/velvet_1.2.10/data 中下载 test_reference.fa;

/home/student/s21/velvet_1.2.10/sillyDirectory 中下载 contigs.fa。

2.5、全基因组序列拼接软件的使用练习 2:

2.5.1、利用实验一所掌握的全基因组测序模拟工具所生产的测序结果，利用
AllPathsLG 进行拼装：

2.5.1.1.根据 illumina 的命令，更改 in_groups.csv 和 in_libs.csv 中数据

表格 1.根据 illumina 模拟数据更改 in_groups.csv

file_name	library_name	group_name
seq/short_pair_ill.?.fastq	Solexa-25396	frugment
seq/long_pair_ill.?.fastq	Solexa-11542	jumping

表格 2. 根据 illumina 模拟数据更改 in_libs.csv

library_name	project_name	organism_name	type	paired	frag_size	frag_stddev
Solexa-25396	test	test.genome	fragment	1	200	10

Solexa-11542	test	test.genome	jumping	1		
insert_size	insert_stddev	read_orientation	genomic_start	genomic_end		
		inward	0	0		
2500	10	outward	0	0		

2.5.1.2.执行命令:

```
cd my_allpaths
prepare.sh
assemble.sh
```

2.5.1.3.下载数据:

/home/student/s21/my_allpaths/test.genome/data/run/ASSEMBLIES/test
中下载 final.assembly.fasta

2.5.2、利用 blastn 将组装结果与原始基因组进行对比，分析其中是否存在差异，并阐述原因。

将 NC_010321.1.fasta 和 final.assembly.fasta 导入 blastn 中进行比对

三、实验结果：

3.1. 五种基因组组装软件

3.1.1.Modular Open-Source Assembler (AMOS): AMOS 有两个汇编程序:

①Minimus 专门为小数据集设计，如覆盖单个基因或病毒的读取集。Minimus 可应用于较大规模的装配（例如，组装大型细菌基因组和宏基因组数据）。然而，由于其严格性，最终的装配往往会高度分散。

②Minimo 中可以改变装配严格性，从而为基因组装配提供非常细致的方法。然而，对于大型或复杂的基因组，Minimus 和 Minimo 的执行应该附加一些额外的处理步骤。Minimus 和 Minimo 都遵循 Overlap-Layout-Consensus 范例，它由三个主要模块组成，它们通过中央文件库共享信息：（1）散列重叠模块计算重叠使用 Smith-Waterman 局部比对算法的修改版本进行成对读取之间；（2）然后使用 Myers 描述的算法，Tigger 模块使用这些重叠来找出单个重叠群中阅读的排列。（3）构建共识模块通过对每个重叠群中的读段进行多重比对以产生准确的共有序列来改进布局。

3.1.2. CISA: 包括四步，①鉴定代表性重叠群和可能的扩展；②可能错误组合的重叠群的去除和分裂以及位于重叠群末端的不确定区域的剪切；③重叠合并最小 30% 的重叠群并估计重复区域的最大尺寸；④基于重复区域的大小合并重叠群。

使用方法：1）创建 merge.config 配置文件。2）根据上面 merge.config 配置文件信息，对 3 个 Assemblies 的 contigs 序列进行提取，并整合到 merge_contigs.fasta 文件中，同时输出各个 Assemblies 的 contigs size。3）创建 cisa.config 配置文件 4）启动主程序进行运算。基因组越大，Assemblies 数目越多，耗时越多（呈几何级数增加）。

3.1.3.Newbler: 是针对 454 测序数据进行组装的软件，454 下的数据是 sff 格式，无法查看，可以用 sffinfo 进行转换。Newbler 输入文件可以是来自 Illumina 的 fastq 文件，或是 454 的 sff 文件，newbler 可以对 454 测序数据进行组装，也可以对 Illumina 进行组装，以及对 454 和 Illumina 数据进行混合组装。

3.1.4.ABYSS: 引进并行计算的思想，搭建了一个 linux 集群，在集群上建立了一个分布式的 De Bruijn 图结构，将数据分布式存储于每个节点上。其采用 MPI 通信机制完成节点之间的通信，它在运行时间和内存消耗方面占有很大的优势，并且其错误率极低。

3.1.5.SOAPdenovo: 是一个新颖的适用于组装短 reads 的方法，该软件特地设计用来组装 Illumina GA short reads，新的版本减少了在图创建时的内存消耗，解决了 contig 组装时的重复区域的问题，增加了 scaffold 组装时的覆盖度和长度，改进了 gap closing，更加适用于大型基因组组装。

3.2. cap3 下载安装【step-2.1.2.1 结果】

```
s06@HP06:~$ cd CAP3
s06@HP06:~/CAP3$ ls
*****partial screen echoes*****
aceform cap3 doc doc.txt example formcon README
s06@HP06:~/CAP3$ ./cap3
VersionDate: 02/10/15 Size of long: 8
Usage: ./cap3 File_of_reads [options]
```

File_of_reads is a file of DNA reads in FASTA format

If the file of reads is named 'xyz', then
the file of quality values must be named 'xyz.qual',
and the file of constraints named 'xyz.con'.

Options (default values):

- a N specify band expansion size N > 10 (20)
- b N specify base quality cutoff for differences N > 15 (20)
- c N specify base quality cutoff for clipping N > 5 (12)
- d N specify max qscore sum at differences N > 20 (200)
- e N specify clearance between no. of diff N > 10 (30)
- f N specify max gap length in any overlap N > 1 (20)
- g N specify gap penalty factor N > 0 (6)
- h N specify max overhang percent length N > 2 (20)
- i N specify segment pair score cutoff N > 20 (40)
- j N specify chain score cutoff N > 30 (80)
- k N specify end clipping flag N >= 0 (1)
- m N specify match score factor N > 0 (2)

- n N specify mismatch score factor $N < 0$ (-5)
- o N specify overlap length cutoff > 15 (40)
- p N specify overlap percent identity cutoff $N > 65$ (90)
- r N specify reverse orientation value $N \geq 0$ (1)
- s N specify overlap similarity score cutoff $N > 250$ (900)
- t N specify max number of word matches $N > 30$ (300)
- u N specify min number of constraints for correction $N > 0$ (3)
- v N specify min number of constraints for linking $N > 0$ (2)
- w N specify file name for clipping information (none)
- x N specify prefix string for output file names (cap)
- y N specify clipping range $N > 5$ (100)
- z N specify min no. of good reads at clip pos $N > 0$ (3)

s06@HP06:~/CAP3\$./cap3 ./example/xyz

*****partial screen echoes*****

ENV1719.y2- AAACCTT-TTGTGCCTCAGTTCTCTCATCTATGAAAT-TGGTATGCAGATGAGAGA
ACCTC

.....

consensus AAACCTT-TTGTGCCTCAGTTCTCTCATCTATGAAAT-TGGTATGCAGATGAGAGAA
CCTC

	. : . : . : . : . :
ENV1644.y2-	TCTTATAAGCTTGTGTAATGATTAAAG-ACATAATCATGG
ENV1704.y2-	TCTTATAAGCTTGTGTAATGATTAAAG-ACTTAATCATGG
ENV524.x1+	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTAG
AATAG	
ENV1585.x1-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTA
GAATAG	
ENV553.x1-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTAG
AATAG	
ENV1622.y2-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTA
GAATAG	
ENV87.y1-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTAG
AATAG	
ENV1585.x3-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTA
GAATAG	
ENV1095.y1-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTA
GAATAG	
ENV562.y1+	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTAG
AATAG	
ENV1719.y2-	TCTTATAAGCTTGTGTAATGATTAAAGGACTTAATCATGGGTGTAAAGGTCTTA
GAATAG	
ENV63.y1+	GTCTTAGAATAG

.....

安装成功可以使用。

3.3. SOAPdenovo 下载安装【step-2.1.2.2 结果】

```
s06@HP06:~$ tar -zxvf SOAPdenovo2-src-r240.tgz
```

```
s06@HP06:~$ cd SOAPdenovo2-src-r240
```

```
s06@HP06:~/SOAPdenovo2-src-r240$ ls
```

```
s06@HP06:~/SOAPdenovo2-src-r240$ make
```

```
*****partial screen echoes*****
```

```
INSTALL LICENSE Makefile MANUAL sparsePregraph standardPregraph update.log VERSION
```

.....

```
Error in command: g++ -c -O4 -fomit-frame-pointer -D_63MER_ -m64 -I./inc multi_threads.cpp
```

```
Error in command: g++ -c -O4 -fomit-frame-pointer -D_63MER_ -m64 -I./inc pregraph_sparse.cpp
```

```
Error in command: g++ -c -O4 -fomit-frame-pointer -D_63MER_ -m64 -I./inc build_preArc.cpp
```

.....

```
main.c:(.text.startup+0xff): undefined reference to `call_pregraph_sparse'
```

```
collect2: error: ld returned 1 exit status
```

```
make: *** [SOAPdenovo-63mer] Error 1
```

```
##手动更改以上文件，添加#include "unistd.h"
```

```
s06@HP06:~/SOAPdenovo2-src-r240$ make
```

```
s06@HP06:~/SOAPdenovo2-src-r240$ ./SOAPdenovo-63mer
```

```
*****partial screen echoes*****
```

```
Version 2.04: released on July 13th, 2012
```

```
Compile Apr 7 2018 19:25:21
```

```
Usage: SOAPdenovo <command> [option]
```

```
pregraph      construct kmer-graph
```

```
sparse_pregraph construct sparse kmer-graph
```

```
contig        eliminate errors and output contigs
```

```
map           map reads to contigs
```

```
scaff         construct scaffolds
```

```
all           do pregraph-contig-map-scaff in turn
```

```
s06@HP06:~/SOAPdenovo2-src-r240$ ./SOAPdenovo-127mer
```

```
*****partial screen echoes*****
```

```
Version 2.04: released on July 13th, 2012
```

```
Compile Apr 7 2018 19:25:34
```


Usage: SOAPdenovo <command> [option]

pregraph construct kmer-graph
sparse_pregraph construct sparse kmer-graph
contig eliminate errors and output contigs
map map reads to contigs
scaff construct scaffolds
all do pregraph-contig-map-scaff in turn

s06@HP06:~/SOAPdenovo2-src-r240\$ ll

*****partial screen echoes*****

total 1968

drwxr-xr-x 4 s06 s06 4096 4月 7 19:25 ./
drwxr-xr-x 24 s06 s06 4096 4月 7 19:16 ../
-rw-r--r-- 1 s06 s06 448 7月 9 2013 INSTALL
-rw-r--r-- 1 s06 s06 35147 7月 9 2013 LICENSE
-rw-r--r-- 1 s06 s06 2147 7月 9 2013 Makefile
-rw-r--r-- 1 s06 s06 16707 7月 9 2013 MANUAL
-rwxrwxr-x 1 s06 s06 1057876 4月 7 19:25 SOAPdenovo-127mer*
-rwxrwxr-x 1 s06 s06 861318 4月 7 19:25 SOAPdenovo-63mer*
drwxr-xr-x 3 s06 s06 4096 4月 7 19:25 sparsePregraph/
drwxr-xr-x 3 s06 s06 4096 4月 7 19:25 standardPregraph/
-rw-r--r-- 1 s06 s06 3768 7月 9 2013 update.log
-rw-r--r-- 1 s06 s06 10 7月 9 2013 VERSION

安装成功可以使用。

3.4. Velvet 下载安装【step-2.1.2.3 结果】

s06@HP06:~\$ wget http://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz

s06@HP06:~\$ tar zxvf velvet_1.2.10.tgz

s06@HP06:~\$ cd velvet_1.2.10

s06@HP06:~/velvet_1.2.10\$ make color 'CATEGORIES=57' 'MAXKMERLENGTH=127'

'BIGASSEMBLY=1' 'LONGSEQUENCES=1' 'OPENMP=1'

s06@HP06:~/velvet_1.2.10\$ ll

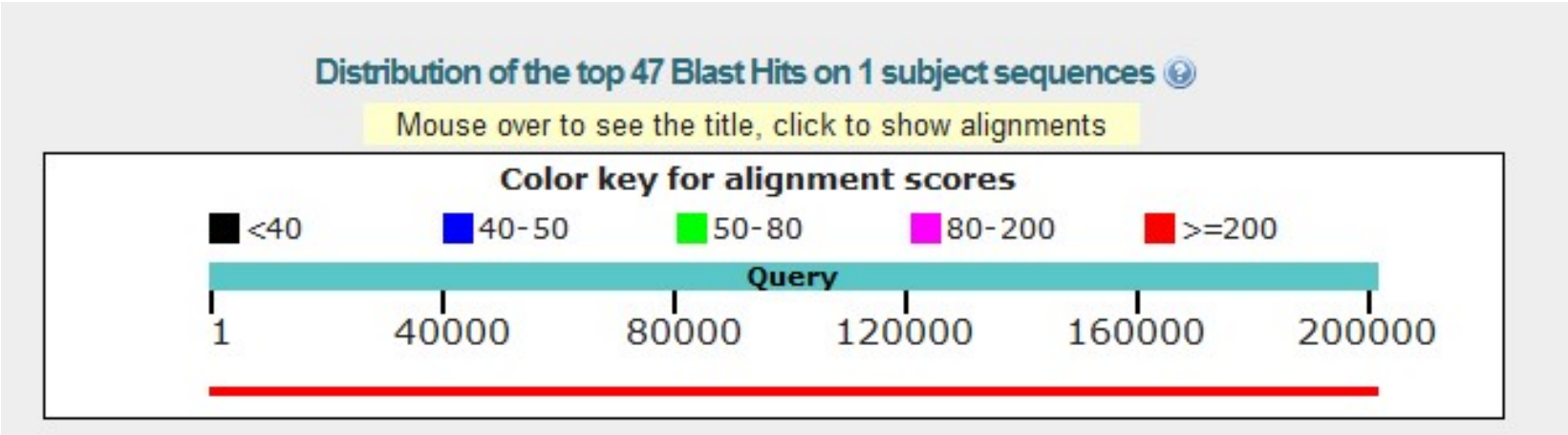
*****partial screen echoes*****

total 1004

drwxr-xr-x 10 s06 s06 4096 4月 7 20:03 ./
drwxr-xr-x 25 s06 s06 4096 4月 7 20:02 ../
-rw-r--r-- 1 s06 s06 12690 10月 18 2013 ChangeLog
-rw-r--r-- 1 s06 s06 90537 10月 18 2013 Columbus_manual.pdf
drwxr-xr-x 17 s06 s06 4096 10月 18 2013 contrib/
-rw-r--r-- 1 s06 s06 1956 10月 18 2013 CREDITS.txt
drwxr-xr-x 2 s06 s06 4096 10月 18 2013 data/
drwxr-xr-x 2 s06 s06 4096 10月 18 2013 debian/
drwxr-xr-x 3 s06 s06 4096 10月 18 2013 doc/
-rw-r--r-- 1 s06 s06 456 10月 18 2013 For_MAC_or_SPARC_users.txt

安装成功可以使用。

3.5. Allpaths 拼接软件配套的测试数据比对结果



图表 1.ALLPATHS 软件配套数据测试比对结果 1

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

Alignments Download Graphics

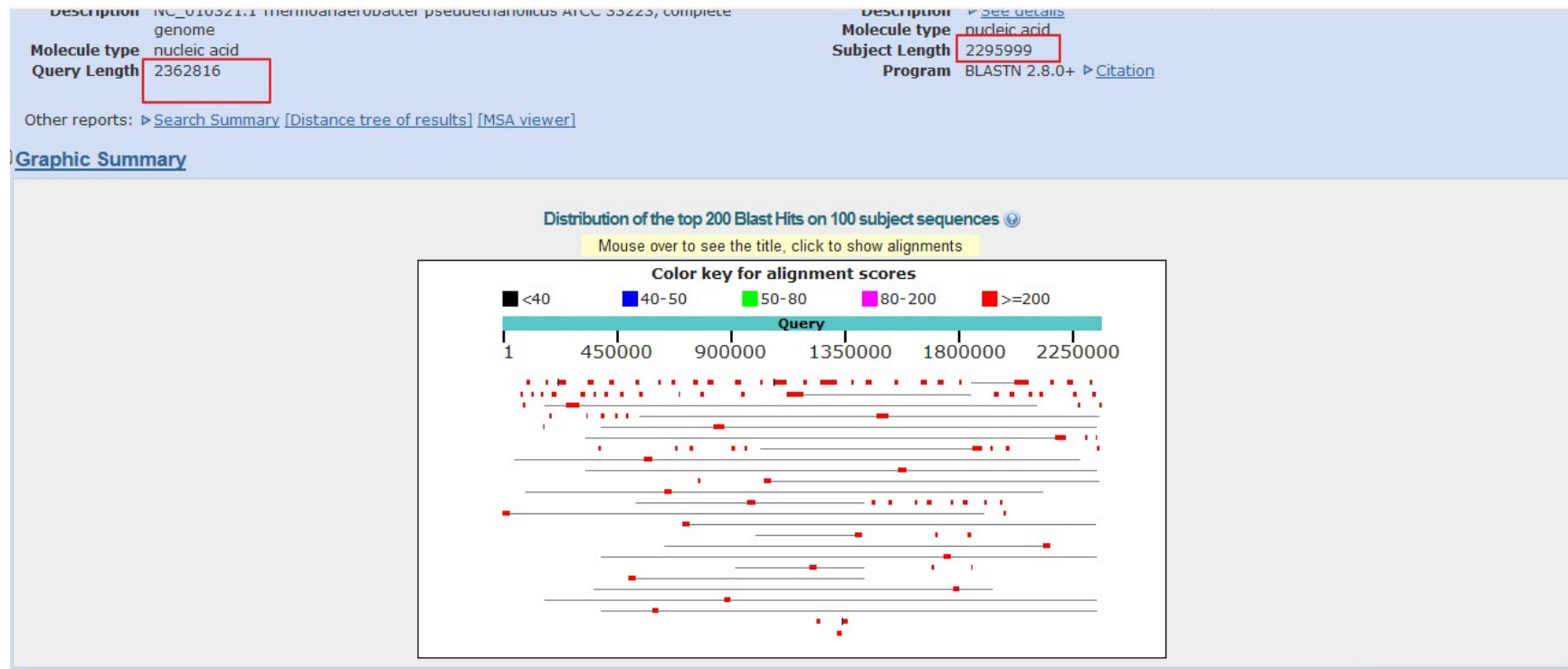
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> scaffold_0		3.697e+05	3.756e+05	99%	0.0	100%	Query_40339

图表 2.ALLPATHS 软件配套数据测试比对结果 2

3.6. SOAP 拼接软件配套的测试数据比对结果

3.5.1 组装结果 nohup.out 文件见附录四中 4.1

3.5.2 contigs 比对



图表 3.SOAPdenovo 组装 illumina 生成片段的 contig 结果 1

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	3988 length 67618 cvq 13.9 tip 1	1.249e+05	1.249e+05	2%	0.0	100%	Query_165475
<input type="checkbox"/>	3986 length 64355 cvq 13.9 tip 0	1.188e+05	1.190e+05	2%	0.0	100%	Query_165474
<input type="checkbox"/>	3984 length 56099 cvq 13.1 tip 0	1.036e+05	1.037e+05	2%	0.0	100%	Query_165473
<input type="checkbox"/>	3982 length 54476 cvq 13.7 tip 0	1.006e+05	1.007e+05	2%	0.0	100%	Query_165472
<input type="checkbox"/>	3980 length 53227 cvq 13.7 tip 0	98292	98949	2%	0.0	100%	Query_165471
<input type="checkbox"/>	3978 length 47783 cvq 13.5 tip 0	88239	88652	2%	0.0	100%	Query_165470
<input type="checkbox"/>	3976 length 41687 cvq 13.8 tip 0	76982	77212	1%	0.0	100%	Query_165469

图表 4.SOAPdenovo 组装 illumina 生成片段的 contig 结果 2

3.5.3 scaffold 比对

Job title: NC_010321.1 Thermoanaerobacter pseudethanolicus

RID [CM4EY1KB114](#) (Expires on 04-09 20:58 pm)

Query ID [Id|Query_171645](#)

Description NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome

Molecule type nucleic acid

Query Length 2362816

Subject ID 113 subjects

Description [▶ See details](#)

Molecule type nucleic acid

Subject Length 2322499

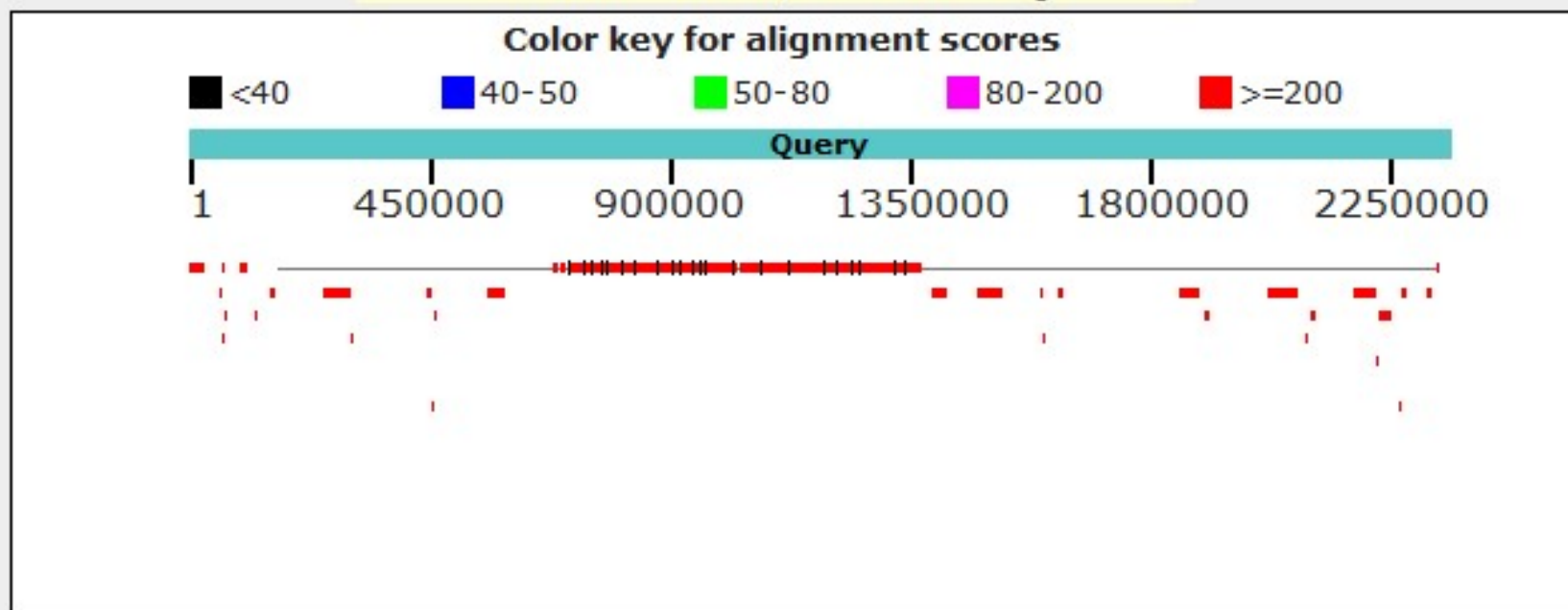
Program BLASTN 2.8.0+ [▶ Citation](#)

Other reports: [▶ Search Summary](#) [\[Distance tree of results\]](#) [\[MSA viewer\]](#)

图表 5. SOAPdenovo 组装 illumina 生成片段的 scaffold 结果 1

Distribution of the top 200 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



图表 6. SOAPdenovo 组装 illumina 生成片段的 scaffold 结果 2

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	scaffold8 13.5	1.248e+05	1.364e+06	29%	0.0	100%	Query_171654
<input type="checkbox"/>	scaffold1 13.4	1.035e+05	3.144e+05	7%	0.0	100%	Query_171647
<input type="checkbox"/>	scaffold15 13.4	98394	4.908e+05	11%	0.0	99%	Query_171661
<input type="checkbox"/>	scaffold14 13.2	88182	2.465e+05	5%	0.0	100%	Query_171660
<input type="checkbox"/>	scaffold20 13.6	76040	1.990e+05	4%	0.0	100%	Query_171666
<input type="checkbox"/>	scaffold9 13.3	63485	4.476e+05	10%	0.0	100%	Query_171655
<input type="checkbox"/>	scaffold4 13.0	62061	3.969e+05	9%	0.0	100%	Query_171650
<input type="checkbox"/>	scaffold7 14.2	53029	96073	2%	0.0	100%	Query_171653
<input type="checkbox"/>	scaffold3 13.5	50630	1.417e+05	3%	0.0	100%	Query_171649
<input type="checkbox"/>	scaffold10 13.0	43748	69116	1%	0.0	100%	Query_171656
<input type="checkbox"/>	scaffold21 13.7	24733	3.811e+05	2%	0.0	100%	Query_171667
<input type="checkbox"/>	scaffold5 13.2	21160	1.192e+05	2%	0.0	100%	Query_171651

图表 7.SOAPdenovo 组装 illumina 生成片段的 scaffold 结果 3

3.5.4 创建参考基因的 blast 数据库，查看结果，详细结果见附录

```
$ makeblastdb -in NC_010321.1.fasta -input_type fasta -dbtype nucl -title NC -out NC
```

```
*****screen echo*****
```

```
Building a new DB, current time: 04/08/2018 22:01:06
```

```
New DB name: /home/student/s21/NC
```

```
New DB title: NC
```

```
Sequence type: Nucleotide
```

```
Keep Linkouts: T
```

```
Keep MBits: T
```

```
Maximum file size: 1000000000B
```

```
Adding sequences from FASTA; added 1 sequences in 0.046113 seconds.
```

```
*****
```

```
$ blastn -db NC -query SOAPdenovo_out.contig -out contig_blast_outfmt7  
-evaluate 1e5 -outfmt 7 -max_target_seqs 1 -num_threads 20
```

```
##生成文件见附录 5.2
```

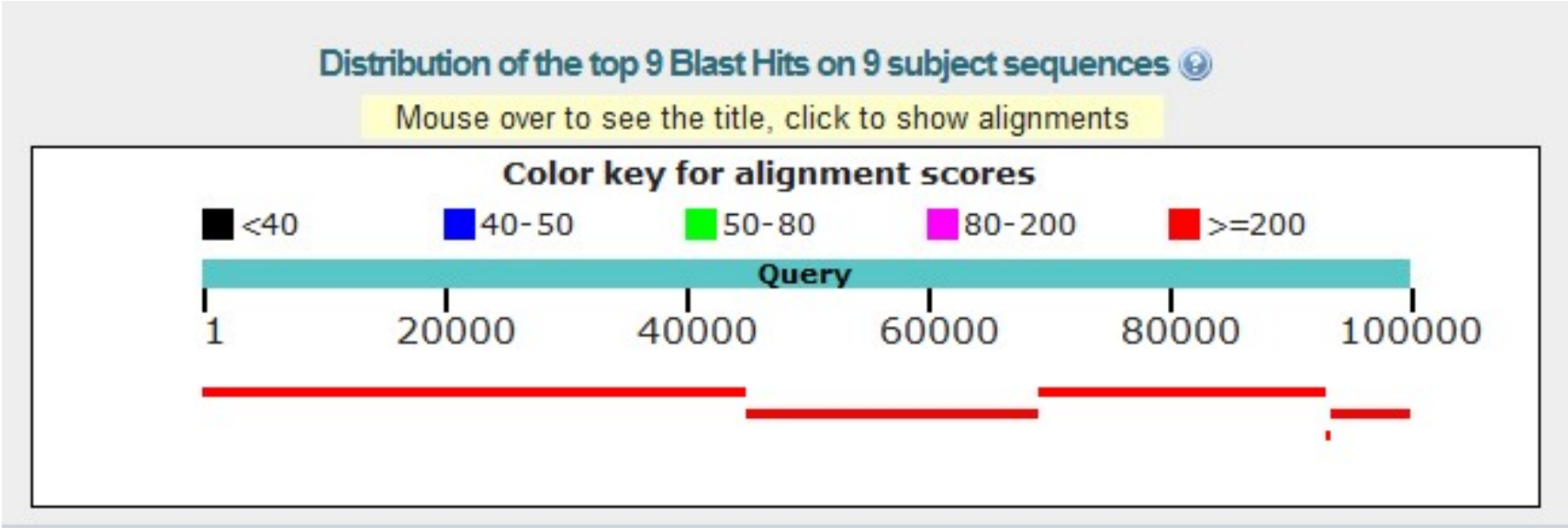
```
$ blastn -db NC -query SOAPdenovo_out.contig -out contig_blast_outfmt6  
-evaluate 1e5 -outfmt 6 -max_target_seqs 1 -num_threads 20
```

```
##生成文件见附录 5.3
```

```
$blastn -db NC -query SOAPdenovo_out.scafSeq -out  
scaffold_blast_outfmt6 -evaluate 1e5 -outfmt 6 -max_target_seqs 1 -num_threads  
20
```

```
##生成文件见附录 5.4
```


3.7. velvet 拼接软件配套的测试数据比对结果



图表 8.velvet 软件配套数据测试比对结果 1

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

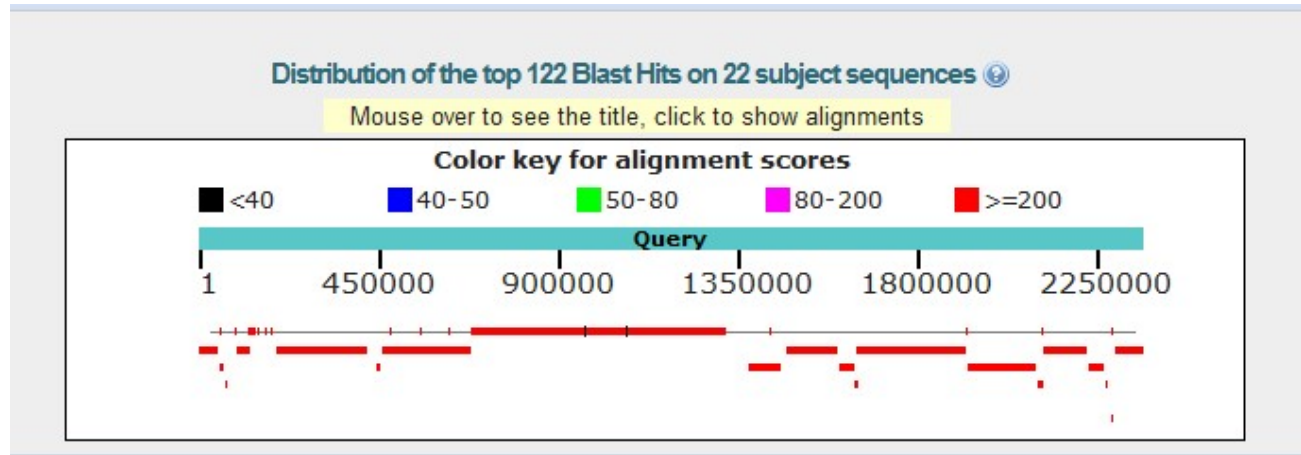
Alignments Download Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	NODE 3 length 44966 cov 18.710159	83074	83074	44%	0.0	100%	Query_152651
<input type="checkbox"/>	NODE 2 length 24184 cov 18.632774	44697	44697	24%	0.0	100%	Query_152650
<input type="checkbox"/>	NODE 4 length 23736 cov 18.616152	43870	43870	23%	0.0	100%	Query_152652
<input type="checkbox"/>	NODE 1 length 6591 cov 18.814899	12209	12209	6%	0.0	100%	Query_152649
<input type="checkbox"/>	NODE 5 length 414 cov 18.190821	802	802	0%	0.0	100%	Query_152653
<input type="checkbox"/>	NODE 12 length 21 cov 15.428572	76.8	76.8	0%	7e-14	100%	Query_152660
<input type="checkbox"/>	NODE 8 length 21 cov 18.904762	76.8	76.8	0%	7e-14	100%	Query_152656
<input type="checkbox"/>	NODE 7 length 21 cov 22.047619	76.8	76.8	0%	7e-14	100%	Query_152655
<input type="checkbox"/>	NODE 6 length 21 cov 14.761905	76.8	76.8	0%	7e-14	100%	Query_152654





图表 9.velvet 软件配套数据测试比对结果 2

3.8. Allpaths 模拟拼装 *Thermoanaerobacter pseudethanolicus* 物种的全基因组序列结果

3.7.1. 用组装结果比对原始序列（NC_010321.1）

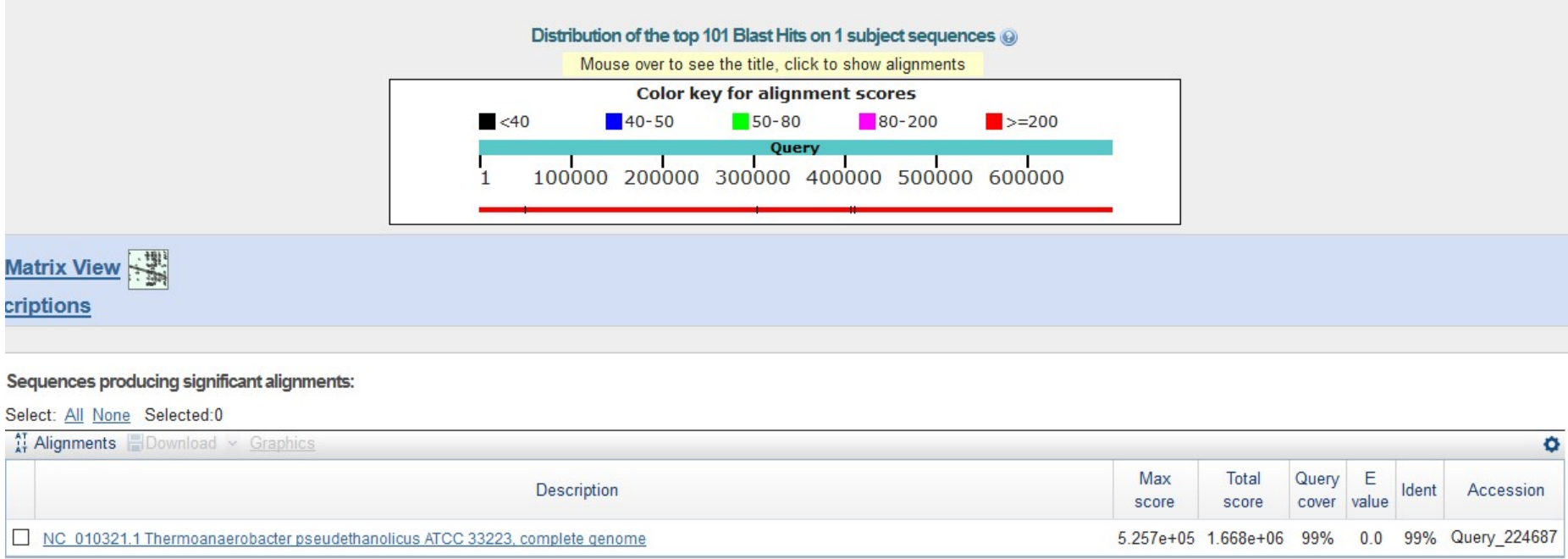


图表 10.Allpaths 组装 illumina 生成片段的结果 1

Alignments  Download  Graphics  Distance tree of results 							
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	scaffold_0	5.257e+05	1.575e+06	30%	0.0	99%	Query_153432
<input type="checkbox"/>	scaffold_2	5.066e+05	6.211e+05	13%	0.0	99%	Query_153434
<input type="checkbox"/>	scaffold_1	4.131e+05	1.090e+06	15%	0.0	99%	Query_153433
<input type="checkbox"/>	scaffold_3	4.073e+05	5.466e+05	11%	0.0	99%	Query_153435
<input type="checkbox"/>	scaffold_4	3.118e+05	3.326e+05	7%	0.0	99%	Query_153436
<input type="checkbox"/>	scaffold_5	2.351e+05	3.489e+05	7%	0.0	100%	Query_153437
<input type="checkbox"/>	scaffold_6	2.026e+05	2.439e+05	5%	0.0	99%	Query_153438
<input type="checkbox"/>	scaffold_7	1.457e+05	1.907e+05	4%	0.0	100%	Query_153439
<input type="checkbox"/>	scaffold_8	1.271e+05	1.517e+05	3%	0.0	100%	Query_153440
<input type="checkbox"/>	scaffold_10	84546	1.394e+05	3%	0.0	99%	Query_153442

图表 11.Allpaths 组装 illumina 生成片段的结果 2

3.7.2. 用原始序列（NC_010321.1）比对组装结果



图表 12.Allpaths 组装 illumina 生成片段的结果 3

四、分析讨论：

1.对于拼接软件配套的测试数据和其组装的结果:本次实验分别使用 allpaths-lg 和 velvet 进行测试,使用 allpaths-lg 的组装结果是一条完整的序列,但是使用 velvet 的组装结果是断的,原因是 velvet 的组装结果生成的是片段文件。

2. allpath 中 in_groups.csv 和 in_libs.csv 文件的说明

①in_groups.csv 文件的解释:

group_name:数据独特的代号,每一份数据有一个代号; library_name:数据所属文库的名字。filename:数据文件所存放位置。可以为相对位置,文件名可以包含'*'和'?'(但是扩展名 中不能有该符号,因为要根据扩展名识别文件类型),从而代表 paired 数据。支持的文件类型有 '.bam','fasta','fa','fastq','fq','fastq.gz'和'fq.gz'。

②in_libs.csv 文件的解释:

library_name:和 in_groups.csv 中的相匹配; project_name:project 的名字; organism_name:测序物种的名字; type:仅仅只是一个信息; paired:0:Unpaired reads;1:paired reads; frag_size:小片段文库插入片段长度的均值; frag_stddev:小片段文库的插入片段长度估算的标准偏差; insert_size:大片段文库插入片段长度的均值; insert_stddev:大片段文库插入片段长度估算的标准偏差; read_orientation:reads 的方向,小片段文库为 inward,大片段文库为 outward; genomic_start:reads 从该位置开始,读入数据,如果不为 0,之前的碱基都被剪掉; genomic_end:reads 从该位置开始,停止读入数据,如果不为 0,之后的碱基都被剪掉。

3.Allpaths-lg 模拟拼装 Thermoanaerobacter pseudethanolicus 物种的全基因组序列结果显示两组序列的匹配程度较好,用组装结果比对原始序列(NC_010321.1)显示位置上都可以匹配,但是分成了多个独立对齐区域;用原始序列(NC_010321.1)比对组装结果可以反向验证,覆盖程度较好。

4.对于 illumina 模拟生成的双末端长插入片段和短插入片段的组装,allpaths-lg 的结果比 SOAPdenovo 好,相比二者结果,序列在 600kb-1350kb 间的组装结果较好,soap 组装碎片多,而且选择不同 kmer,组装的结果差异太大。SOAPdenovo 表现出高比率的组装碎片,contig 重复,重复区压缩,SNPs。

5.SOAPdenovo 组装结果 fasta 格式的 scaffold 序列文件中 contig 之间的 gap 用 N 填充;对于得到的*.scafSeq 文件还需要用 GapCloser 去合并其中的 gap,最后的 contig 文件则是对补洞之后的 scaffold 文件通过打断 N 区的方法得到。

五 . 附录：

5.1.使用 SOAPdenovo 组装 illumina 生成长、短插入片段的 nohup.out 文件:

【部分截取】

```
In third_test.cfg, 2 lib(s), maximum read length 90, maximum name length 256.
```

```
20 thread(s) initialized.
```

```
Import reads from file:
```

```
short_pair_ill.A.fq
```

```
Import reads from file:
```

```
short_pair_ill.B.fq
```

Import reads from file:
long_pair_ill.A.fq
Import reads from file:
long_pair_ill.B.fq
Time spent on hashing reads: 27s, 630080 read(s) processed.
LIB(s) information:
[LIB] 0, avg_ins 200, reverse 0.
[LIB] 1, avg_ins 2500, reverse 1.
4417022 node(s) allocated, 37804800 kmer(s) in reads, 37804800 kmer(s)
processed.
done hashing nodes
4216415 linear node(s) marked.
Time spent on marking linear nodes: 0s.
Time spent on pre-graph construction: 31s.

Start to remove frequency-one-kmer tips shorter than 62.
Total 69499 tip(s) removed.
20 thread(s) initialized.
65992 linear node(s) marked.
Start to remove tips with minority links.
225 tip(s) removed in cycle 1.
0 tip(s) removed in cycle 2.
Total 225 tip(s) removed.
20 thread(s) initialized.
0 linear node(s) marked.
Time spent on removing tips: 1s.

195917 (97959) edge(s) and 2657 extra node(s) constructed.
Time spent on constructing edges: 5s.

In file: third_test.cfg, max seq len 90, max name len 256.
20 thread(s) initialized.
Import reads from file:
short_pair_ill.A.fq
Import reads from file:
short_pair_ill.B.fq
Import reads from file:
long_pair_ill.A.fq
Import reads from file:
long_pair_ill.B.fq
630080 read(s) processed.
Time spent on:
importing reads: 0s,
chopping reads to kmers: 0s,

```

searching kmers: 0s,
aligning reads to edges: 0s,
searching (K+1)mers: 0s,
adding pre-arcs: 0s,
recording read paths: 0s.
0 marker(s) output.
Reads alignment done, 12068 read(s) deleted, 195155 pre-arc(s) added.
LIB(s) information:
[LIB] 0, avg_ins 200, reverse 0.
[LIB] 1, avg_ins 2500, reverse 1.
Time spent on aligning reads: 5s.

```

```

64864 vertex(es) output.
Overall time spent on constructing pre-graph: 0m.

```

```

*****

```

```

Contig

```

```

*****

```

```

Parameters: contig -g third_test -M 1 -s third_test.cfg -p 20

```

```

There are 64864 kmer(s) in vertex file.
There are 195917 edge(s) in edge file.
Kmers sorted.
195917 edge(s) input.
257224 pre-arcs loaded.
48810 none-palindrome edge(s) swapped, 0 palindrome edge(s) processed.
195917 edge(s) sorted.
Arcs sorted.
Start to pinch bubbles, cutoff 0.100000, MAX NODE NUM 3, MAX DIFF NUM
2.
40 start points, 170344 dheap nodes.
39082 pair(s) found, 22224 pair of path(s) compared, 22129 pair(s) merged.
Sequence comparison failed:
Path crossing deleted edge                                0
Length difference of two paths greater than two           7
Mismatch score greater than cutoff (2)                    42
Mismatch score ratio greater than cutoff (0.1)            0
Path length shorter than (Kmer-1)                          46
DFibHeap: 1321 node(s) allocated.
41037 edge(s) concatenated in cycle 1.
4810 edge(s) concatenated in cycle 2.
250 edge(s) concatenated in cycle 3.

```


0 edge(s) concatenated in cycle 4.
Time spent on pinching bubbles: 0s.
Start to destroy weak inner edges.
7602 weak inner edge(s) destroyed in cycle 1.
1 weak inner edge(s) destroyed in cycle 2.
0 weak inner edge(s) destroyed in cycle 3.
15198 dead arc(s) removed.
172 inner edge(s) with coverage lower than or equal to 1 destroyed.
346 dead arc(s) removed.
13418 edge(s) concatenated in cycle 1.
1488 edge(s) concatenated in cycle 2.
58 edge(s) concatenated in cycle 3.
0 edge(s) concatenated in cycle 4.
Before compacting, 195917 edge(s) existed.
After compacting, 13977 edge(s) left.

Strict: 0, cutoff length: 62.
2521 tips cut in cycle 1.
4 tips cut in cycle 2.
2 tips cut in cycle 3.
0 tips cut in cycle 4.
2482 dead arc(s) removed.
2213 edge(s) concatenated in cycle 1.
236 edge(s) concatenated in cycle 2.
18 edge(s) concatenated in cycle 3.
0 edge(s) concatenated in cycle 4.
Before compacting, 13977 edge(s) existed.
After compacting, 3989 edge(s) left.

There are 456 contig(s) longer than 100, sum up 2209634 bp, with average length 4845.

The longest length is 67618 bp, contig N50 is 20802 bp, contig N90 is 4628 bp.

1995 contig(s) longer than 32 output.

Time spent on constructing contig: 0m.

Map

Parameters: map -s third_test.cfg -g third_test -p 20 -K 30

Kmer size: 31.

Contig length cutoff: 33.

1995 contig(s), maximum sequence length 67618, minimum sequence length 32, maximum name length 10.

Time spent on parsing contigs file: 0s.

20 thread(s) initialized.

Time spent on hashing contigs: 0s.

2233397 node(s) allocated, 2235589 kmer(s) in contigs, 2235589 kmer(s) processed.

Time spent on graph construction: 0s.

Time spent on aligning long reads: 0s.

In file: third_test.cfg, max seq len 90, max name len 256

20 thread(s) initialized.

3989 edge(s) in the graph.

Import reads from file:

short_pair_ill.A.fq

Import reads from file:

short_pair_ill.B.fq

Current insert size is 200, map_len is 32.

Import reads from file:

long_pair_ill.A.fq

Import reads from file:

long_pair_ill.B.fq

Current insert size is 2500, map_len is 35.

Total reads 630080

Reads in gaps 34874

Ratio 5.5%

Reads on contigs 626507

Ratio 99.4%

2 pe insert size, the largest boundary is 630080.

LIB(s) information:

[LIB] 0, avg_ins 200, reverse 0.

[LIB] 1, avg_ins 2500, reverse 1.

Time spent on aligning reads: 6s.

Overall time spent on alignment: 0m.

Scaff

Parameters: scaff -g third_test -p 20

Files for scaffold construction are OK.

There are 2 grad(s), 630080 read(s), max read len 90.

Kmer size: 31

There are 3989 edge(s) in edge file.

Mask contigs with coverage lower than 1.4 or higher than 28.0, and strict length 100.

Average contig coverage is 14, 2055 contig(s) masked.

Mask contigs shorter than 33, 532 contig(s) masked.

5536 arc(s) loaded, average weight is 13.

1995 contig(s) loaded.

Done loading updated edges.

Time spent on loading updated edges: 0s.

Start to load paired-end reads information.

For insert size: 200

Total PE links	155709
Normal PE links on same contig	145458
Incorrect oriented PE links	0
PE links of too small insert size	48
PE links of too large insert size	0
Correct PE links	10203
Accumulated connections	4816

Use contigs longer than 200 to estimate insert size:

PE links	145007
Average insert size	199
SD	10

2408 new connections.

For insert size: 2500

Total PE links	155786
Normal PE links on same contig	116434
Incorrect oriented PE links	0
PE links of too small insert size	0
PE links of too large insert size	0
Correct PE links	39289
Accumulated connections	8960

Use contigs longer than 2500 to estimate insert size:

PE links	116405
Average insert size	2379
SD	9

4480 new connections.

All paired-end reads information loaded.

Time spent on loading paired-end reads information: 0s.

Start to construct scaffolds.

For insert size: 200

Total PE links	2407
PE links to masked contigs	1675
On same scaffold PE links	0

Cutoff of PE links to make a reliable connection: 3

Active connections	1464
Weak connections	600
Weak ratio	41.0%

0 circles removed.

Start to remove transitive connection.

Total contigs	3989
Masked contigs	2587
Remained contigs	1402
None-outgoing-connection contigs	597 (42.582027%)
Single-outgoing-connection contigs	747
Multi-outgoing-connection contigs	0

Cycle 1

Two-outgoing-connection contigs	58
Potential transitive connections	27
Transitive connections	13
Transitive ratio	22.4%

Cycle 2

Two-outgoing-connection contigs	45
Potential transitive connections	14
Transitive connections	0
Transitive ratio	0.0%

Start to linearize sub-graph.

Picked sub-graphs	34
Connection-conflict	0
Significant overlapping	12
Eligible	5

```

Bubble structures          0
Mask repeats:
Puzzles                   8
Masked contigs            8
Start to remove transitive connection.
Total contigs              3989
Masked contigs             2603
Remained contigs           1386
None-outgoing-connection contigs  597 (43.073593%)
Single-outgoing-connection contigs  789
Multi-outgoing-connection contigs   0
Cycle 1
Two-outgoing-connection contigs    0
Potential transitive connections    0
Transitive connections              0
Start to linearize sub-graph.
Picked sub-graphs                 0
Connection-conflict               0
Significant overlapping            0
Eligible                          0
Bubble structures                  0
Start to mask puzzles.
Masked contigs                    0
Remained puzzles                  0
Freezing done.

```

```

Rank 1
Scaffold number                101
In-scaffold contig number      456
Total scaffold length          1793791
Average scaffold length        17760
Filled gap number              217
Longest scaffold               130776
Scaffold and singleton number   256
Scaffold and singleton length   2206320
Average length                  8618
N50                            38565
N90                            8371

```

Report from smallScaf: 140 scaffolds by smallPE.

For insert size: 2500

```

Total PE links                4480
PE links to masked contigs    3549

```

On same scaffold PE links 199

Cutoff of PE links to make a reliable connection: 5

Report from checkScaf: 0 scaffold segments broken.

Add large insert size PE links: 0 orientation-conflict links, 61 contigs
acrossed by normal links.

Active connections	678
Weak connections	292
Weak ratio	43.1%

0 circles removed.

Start to remove transitive connection.

Total contigs	3989
Masked contigs	2603
Remained contigs	1386
None-outgoing-connection contigs	265 (19.119770%)
Single-outgoing-connection contigs	1036
Multi-outgoing-connection contigs	38

Cycle 1

Two-outgoing-connection contigs	47
Potential transitive connections	32
Transitive connections	23
Transitive ratio	48.9%

Cycle 2

Two-outgoing-connection contigs	24
Potential transitive connections	9
Transitive connections	0
Transitive ratio	0.0%

Start to linearize sub-graph.

Picked sub-graphs	39
Connection-conflict	0
Significant overlapping	4
Eligible	9
Bubble structures	0

Mask repeats:

Puzzles	11
Masked contigs	5

Start to remove transitive connection.

Total contigs	3989
Masked contigs	2613
Remained contigs	1376
None-outgoing-connection contigs	213 (15.479651%)
Single-outgoing-connection contigs	1158
Multi-outgoing-connection contigs	2

Cycle 1

Two-outgoing-connection contigs	3
---------------------------------	---

Potential transitive connections	1
Transitive connections	0
Transitive ratio	0.0%

Start to linearize sub-graph.

Picked sub-graphs	3
Connection-conflict	0
Significant overlapping	0
Eligible	2
Bubble structures	0

Non-strict linearization.

Start to linearize sub-graph.

Picked sub-graphs	2
Connection-conflict	0
Significant overlapping	0
Eligible	2
Bubble structures	0

Start to mask puzzles.

Masked contigs	2
Remained puzzles	0

Freezing done.

Recover contigs.

Total recovered contigs	46
Single-route cases	30
Multi-route cases	18

All links loaded.

Time spent on constructing scaffolds: 0s.

The final rank

Scaffold number	24
In-scaffold contig number	456
Total scaffold length	2285720
Average scaffold length	95238
Filled gap number	195
Longest scaffold	692194
Scaffold and singleton number	113
Scaffold and singleton length	2319578
Average length	20527
N50	246275
N90	52579
Weak points	0

5.2.contigs 比对结果 1 输出格式 7

```
# BLASTN 2.2.31+
# Query: 7 length 32 cvg_0.0_tip_0
# Database: NC
# Fields: query id, subject id, % identity, alignment length, mismatches, gap
opens, q. start, q. end, s. start, s. end, evalue, bit score
# 6 hits found
7 NC_010321.1 100.00 32 0 0 1 32 388602 388571 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 872036 872067 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 1429811 1429842 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 1558691 1558660 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 2346250 2346219 3e-11 60.2
7 NC_010321.1 100.00 31 0 0 2 32 1766394 1766424 1e-10 58.4
```

.....

5.3.contigs 比对结果 2 输出格式 6

```
7 NC_010321.1 100.00 32 0 0 1 32 388602 388571 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 872036 872067 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 1429811 1429842 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 1558691 1558660 3e-11 60.2
7 NC_010321.1 100.00 32 0 0 1 32 2346250 2346219 3e-11 60.2
7 NC_010321.1 100.00 31 0 0 2 32 1766394 1766424 1e-10 58.4
```

.....

5.4.contigs 比对结果 3 输出格式 6

```
scaffold1 NC_010321.1 99.95 16336 2 1 7195 23524 216417 200082
0.0 30121
scaffold1 NC_010321.1 99.95 8200 0 2 63928 72127 158932 150737
0.0 15117
scaffold1 NC_010321.1 99.92 7064 0 2 32171 39234 190409 183352
0.0 13007
scaffold1 NC_010321.1 100.00 6128 0 0 73515 79642 148848 142721
0.0 11317
scaffold1 NC_010321.1 99.83 5359 0 1 55794 61152 166267 160918
0.0 9838
scaffold1 NC_010321.1 100.00 5271 0 0 40624 45894 181463 176193
0.0 9734
scaffold1 NC_010321.1 99.89 3746 2 2 28431 32176 194178 190435
0.0 6896
scaffold1 NC_010321.1 100.00 3193 0 0 24224 27416 198880 195688
0.0 5897
scaffold1 NC_010321.1 100.00 2599 0 0 217 2815 223402 220804
0.0 4800
scaffold1 NC_010321.1 100.00 2562 0 0 4569 7130 219042 216481
```