

9. gyakorlat

Genome Assembly

A feladat:

A mai feladat a helicobacter genomjának de-novo összeállítása lesz, amihez a SPADES programot fogjuk használni.

A feladatokat csoportokban kell elvégezni. A szükséges programok és adatok a felhőben lévő szerveren találhatóak meg, aminek a jelenlegi **IP címe** a következő: **23.251.128.6**

A szerverhez a következő parancs kiadásával tudtok csatlakozni:

```
ssh -X <username>@23.251.128.6
```

A felhasználói név a neptun kód nagybetűivel, a jelszó: **Bioinfo2016**

1. Genome assembly

A szükséges adatok a következő útvonalon érhetőek el:

/gfs/data/assembly/

Itt a következő adatok találhatóak:

Group ID	reads	Assembler	Pac bio reads
1	Helico_ins200_p1.fastq	Spades	no
	Helico_ins200_p2.fastq		
2	Helico_ins200_p1.fastq	Spades	yes
	Helico_ins200_p2.fastq		
3	Helico_ins200_p1.fastq	velvet	no
	Helico_ins200_p2.fastq		
4	Helico_ins200_p1.fastq	velvet	yes
	Helico_ins200_p2.fastq		
5	Helico_ins1000_p1.fastq	Spades	no
	Helico_ins1000_p2.fastq		
6	Helico_ins1000_p1.fastq	Spades	yes
	Helico_ins1000_p2.fastq		
7	Helico_ins1000_p1.fastq	velvet	no
	Helico_ins1000_p2.fastq		
8	Helico_ins1000_p1.fastq	velvet	yes
	Helico_ins1000_p2.fastq		
9	Helico_single.fastq	Spades	no
10	Helico_single.fastq	Spades	yes
11	Helico_single.fastq	velvet	no
12	Helico_single.fastq	velvet	yes

SPAded assembly

A genomokat a SPADES program (spades.py) segítségével fogjuk összeállítani. A program használatának a leírását a következő parancs kiadása után láthatjátok.

```
spades.py --help
```

1) Spades esetében (egy sorként, itt csak tördelve van):

```
spades.py -o assembly_nev_ins200_no_pacbio -1  
/gfs/data/assembly/Helico_ins200_p1.fastq -2  
/gfs/data/assembly/Helico_ins200_p2.fastq --careful -t 4 -m 4
```

Részlet

magyarázat

-1 <LEFT_reads.fq>	Paired-end readek egyik párja
-2 <RIGHT_reads.fq>	Paired-end readek másik párja
--careful	Pontosabb összerakás
-t 4	4 szálat használhat
-m 4	4 Gb RAM maximum memória

Opcionális: Összerakás pacbio leolvasásokkal együtt

```
spades.py -o assembly_nev_ins200_with_pacbio -1  
/gfs/data/assembly/Helico_ins200_p1.fastq -2  
/gfs/data/assembly/Helico_ins200_p2.fastq --pacbio  
/gfs/data/assembly/Helico_pac_bio_long_reads.fastq --careful -t 4 -m 4
```

2) Összerakott genom átmásolása a „/tmp” mappába

```
cp the username is your NEPTUN ID (with capital letters), and the password  
is Bioinfo2016.
```

This is a unix server and all the necessary data are available in your home directory on the following path: /gfs/data/xxxxx.

```
assembly_nev_ins200_with_pacbio/scaffolds.fasta /tmp/  
nev_spades_ins200_with_pacbio.fasta
```

Velvet assembly

A genomokat a Velvet program segítségével fogjuk összeállítani. Két lépésből tevődik össze: 1) leolvasások előkészítése és 2) genom összerakás.

1) Readek előkészítése (velveth program):

```
velveth assembly_nev_ins200_kmer31_no_pacbio 31 -fastq -shortPaired1  
/gfs/data/assembly/Helico_ins200_p1.fastq -shortPaired2  
/gfs/data/assembly/Helico_ins200_p2.fastq
```

Részlet

magyarázat

assembly_nev_ins200_no_pacbio

Mappa név

31

kmer méret

-fastq

fastq formátum

-shortPaired1 <LEFT_reads.fq>

Paired-end readek egyik párja

-shortPaired1 <Right_reads.fq>

Paired-end readek egyik párja

Opcionális: Összerakás pacbio leolvasásokkal együtt

```
velveth assembly_nev_ins200_kmer31_with_pacbio 31 -fastq -shortPaired1  
/gfs/data/assembly/Helico_ins200_p1.fastq -shortPaired2  
/gfs/data/assembly/Helico_ins200_p2.fastq -long  
/gfs/data/assembly/Helico_pac_bio_long_reads.fastq
```

2) Genom összeállítása (velvetg program)

```
velvetg assembly_nev_ins200_kmer31_with_pacbio
```

2) Összerakott genom átmásolása a „tmp” mappába

```
cp assembly_nev_ins200_kmer31_no_pacbio/contigs.fa /tmp/  
nev_velvet_ins200_no_pacbio.fasta
```

Genomok összehasonlítása: QUAST

```
quast.py -o comparison nev_spades_ins200_with_pacbio.fasta  
nev_velvet_ins200_no_pacbio.fasta Helico_N1_uniform.fna
```

Dot plot készítés (pairwise alignment): Gepard

Java program, GUI-val