?? ?? ???? (??)

Bacterial Genome data ??

이제 몇 가지 샘플 코드를 실행해 보겠습니다 .

bwa, samtools 도구가 설치되어 있는지 확인해보겠습니다 .

```
which bwa
```

출력에는 bwa가 설치된 위치가 표시됩니다 .

```
which samtools
```

출력에는 samtools가 설치된 위치가 표시됩니다 .

Basic Bacterial Genome Sequence Analysis

1. Get a reference sequence:

```
mkdir -p /tmp/outbreaks/SG-M1
```

```
cd /tmp/outbreaks/SG-M1
```

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/275/545/GCF_001275545.2_ASM127554v2/GCF_001275545.2_ASM127554v2_genomic.fna.gz
```

```
gunzip GCF_001275545.2_ASM127554v2_genomic.fna.gz
```

```
mv GCF_001275545.2_ASM127554v2_genomic.fna SG-M1.fna
```

2. Map and call SNPs:
Note: For an annotation of the programs used below and other bioinformatics tools, check out our course github page.

Reference indexing

```
bwa index SG-M1.fna
```

Mapping

```
bwa mem SG-M1.fna /tmp/fastq/SRR6327950/SRR6327950_1.fastq.gz
/tmp/fastq/SRR6327950/SRR6327950_2.fastq.gz | samtools view -bS - > SRR6327950.bam
```

BAM Sorting

```
samtools sort SRR6327950.bam -o SRR6327950-sort.bam
```

BAM Indexing

```
samtools index SRR6327950-sort.bam
```

Variant calling

```
lofreq faidx SG-M1.fna
```

```
lofreq call -f SG-M1.fna -r NZ_CP012419.2:400000-500000 SRR6327950-sort.bam > SRR6327950-400k.vcf
```

Mapping takes ~5 min on a t2.medium. Sorting takes ~2 min. Running lofreq on this limited section of the genome takes ~1 min.

3. Assembly (runs ~4 min then will run out of RAM if you're on a t2.medium):

```
spades.py -t 2 -1 /tmp/fastq/SRR6327950/SRR6327950_1.fastq.gz -2
/tmp/fastq/SRR6327950/SRR6327950_2.fastq.gz -o SRR6327950_spades
```

NOTE: This assembly above will complete on a t3a.large and takes about 5 hours.

훌륭합니다 ! 이 작업은 AWS EC2 인스턴스에서 쉽게 실행할 수 있는 매우 일상적인 작업입니다 . 3단계에서 어셈블리를 수행할 때 경험했듯이 , 작업에 적합한 머신을 선택하는 것은 매우 중요합니다 . RAM이나 디스크 공간이 부족하면 작업이 중단될 수 있습니다 . 다행히도 인스턴스 유형을 변경하거나 다른 EBS 볼륨을 머신에 연결하면 이러한 문제를 쉽게 해결할 수 있습니다 .

Revision #7
Created 27 September 2023 04:57:08 by Hyunmin Kim
Updated 9 May 2024 03:04:20 by Hyunmin Kim