Complete a three-generation sequencing assembly in three days?

From 徐洲 更生 信媛 2018-06-22

A company sent me a second-generation + three-generation transcript group training course. After these days of training, I found an embarrassing fact. I am afraid that no company training can train me.

So I took the time to write this tutorial to talk about three generations of genome assembly, and I will find an article to write a three-generation transcriptome analysis.

How to assemble three generations of sequencing

The article we used to practice is published in *Nature Communication* , "High contiguity Arabidopsis thaliana genome assembly with a single nanopore

flow cell", very shameless, this article was issued by my teacher's laboratory.

Simply telling the story, they bought a nanopore instrument in their lab, which is the following. At present, the price of the instrument is about 8K in China. Of course, the price of sequencing is also said. Just like buying a PS4 console, you have to buy a game, buy a SLR, you have to buy a lens. The instrument is just the beginning of the defeat!

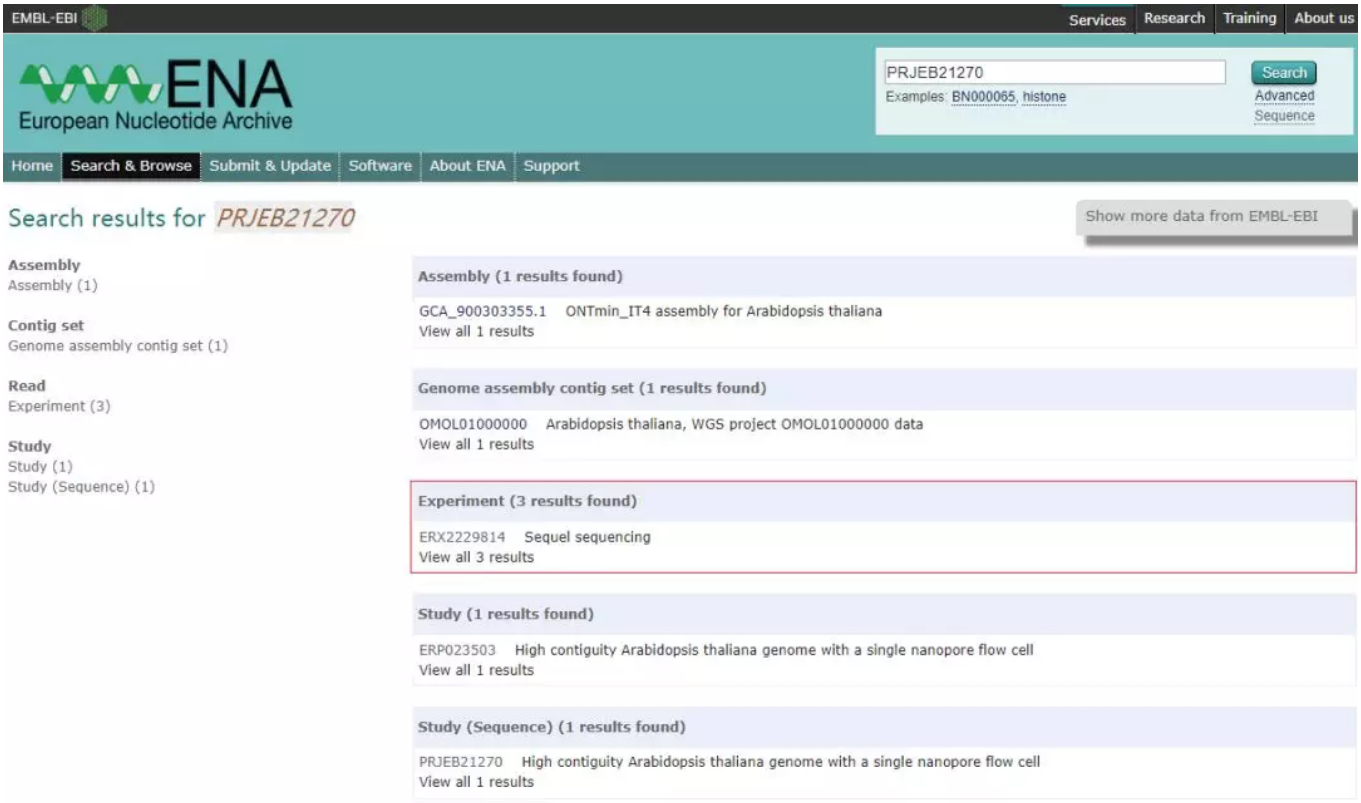
They believe that the three generations of sequencing currently have two major problems, the measurement is not long enough and not accurate. Nanopore solves one of the problems and is not long enough. *Arabidopsis thaliana* was sequenced with a generation. Although it can be considered as the gold standard for assembly, there are still many areas where BAC and BAC libraries are uncertain, so this instrument was used to measure the *Arabidopsis thaliana* . Obviously it is impossible to publish a nanopore, or a species of known sequence, so they measured a wave with Pacbio sequel. Finally, it was verified once with the bionano optical map (please calculate how much it costs).

Light sequencing is not enough, you have to assemble it. The traditional assembly method is to find a way to correct errors with high depth and random errors, then assemble with long sequences after error correction, and finally use the second generation for error correction. It takes about ten days and a half for a good server (20W to start). The author may think that it would be too stupid to buy a 20-w peripheral with less than 1w of the sequencer, so he used a tool developed by Li Heng, Minimap+miniasm for assembly, and then corrected with racon+pillon Wrong, it took 4 days to use

a Macbook Pro 15.6 inch to get it, and compared with the conventional tools, it is still worth it.

The following is the formal analysis:

According to the project number "PRJEB21270" provided in the article, find the download address on the European Nucleotide Archive.



After entering this page, you can download all the data used by the author. We download the data of Sequel and MinIon and Illuminia, and the amount of data adds up to almost 30G.

```
## Sequal
wget -c -q ftp://ftp.sra.ebi.ac.uk/vol1/ERA111/ERA1116568/bam/pb.bam
wget -c -q ftp://ftp.sra.ebi.ac.uk/vol1/ERA111/ERA1116568/bam/pb.bam.bai
## MinION
wget -c -q ftp://ftp.sra.ebi.ac.uk/vol1/ERA111/ERA1116595/fastq/ont.fq.gz
# Illuminia MiSeq
wget -c -q ftp://ftp.sra.ebi.ac.uk/vol1/ERA111/ERA1116569/fastq/il_1.fq.gz
wget -c -q ftp://ftp.sra.ebi.ac.uk/vol1/ERA111/ERA1116569/fastq/il_2.fq.gz
```

Once we have the data, we can repeat it with the analysis process provided by the author. The address is https://github.com/fbemm/onefc-oneasm/wiki/Assembly-Generation

> This is the confidence of the great god, give you the code, you can't understand it anyway. Of course, I use the latest software when I repeat it, so it will be different.

The first step: take the 80% to 90% correct ratio of the original data to each other, find the Overlap between the sequences. This step, I spent 30 minutes

```
time ~/opt/biosoft/minimap2/minimap2 -t 10 -x ava-ont ont.fq ont.fq > gzip -1
```

Step 2: Find the Overlap and you will be able to assemble it. I took 2 minutes in this step.

```
time ~/opt/biosoft/miniasm/miniasm -f ont.fq ont.paf > ONTmin.gfa &
awk '/^S/{print ">"$2"\n"$3}' ONTmin.gfa | seqkit seq > ONTmin_IT0.fasta &
```

Step 3: The original assembly result is full of errors, so you need to correct it. Error correction is divided into two types, one is to use three generations of its own data, and the other is to use second generation data for error correction. Of course, these two steps are all needed.

First of all, using three generations of data for error correction, the old saying has the cloud "things are not three" and iteratively three times. These three steps took almost an hour.

```
# Iteration 1
~/opt/biosoft/minimap2/minimap2 ONTmin_IT0.fasta ont.fq > ONTmin_IT0.paf &
time ~/opt/biosoft/racon/build/bin/racon -t 10 ont.fq ONTmin_IT0.paf ONTmin_IT
# Iteration 2
~/opt/biosoft/minimap2/minimap2 ONTmin_IT1.fasta ont.fq > ONTmin_IT1.paf
time ~/opt/biosoft/racon/build/bin/racon -t 10 ont.fq ONTmin_IT1.paf > ONTmin_
# Iteration 3
~/opt/biosoft/minimap2/minimap2 ONTmin_IT2.fasta ont.fq > ONTmin_IT2.paf
time ~/opt/biosoft/racon/build/bin/racon -t 10 ont.fq ONTmin_IT2.paf > ONTmin_
```

Then use the second generation data for error correction. Although the second-generation data is short, the quality of the sequencing is high, so it is generally used for error correction. It is recommended to use 30X PCR free illuminia sequencing data.

Step 1: Data pre-processing, filtering low-quality short-reading and de-joining. A lot of tools, commonly used is trimmomatic, cutadapter. I am a good tool for domestic Hai Proos.

```
# data clean
fastp -q 30 -5 -l 100 -i il_1.fq.gz -I il_2.fq.gz -o i1_clean_1.fq -O i1_clear
```

The standard here is: the average quality is higher than Q30, low-quality base deletion is performed on the 5' end, and short readings greater than 100 bp are retained.

Step2: Compare, this step basically only uses bwa

```
# align
bwa index ONTmin_IT3.fasta
bwa mem -t 8 ONTmin_IT3.fasta il_clean_1.fastq il_clean_2.fastq | samtools sor
```

Step3: Correction using the aligned BAM file

```
# short read consensus call
java -Xmx16G -jar pilon-1.22.jar --genome ONTmin_IT3.fasta --frags ONTmin_IT3.
```

The time for second-generation error correction is significantly longer than before, and it takes a day.

Finally, everyone took out their own notebook and actually felt it. Nanopore should send me an instrument, feeling like writing a soft text.

> I will save this step of software installation. I should introduce a special video course next week.

references

- Nanopore assembles Arabidopsis thaliana: High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell
- No error correction assembly: Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences
- Three-generation assembly software evaluation: Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data

Article was 2018-06-22 modifications