











Genome Assembly with PacBio Reads

J Fass | December 2018

Raw Data

Arabidopsis data release ... [blog post](#).

	Name	Last modified	Size
	Parent Directory		-
	m54113_160913_184949.metadata.xml	2016-09-13 11:49	8.5K
	m54113_160913_184949.scraps.bam	2016-09-22 22:44	11G
	m54113_160913_184949.scraps.bam.pbi	2016-09-22 23:56	14M
	m54113_160913_184949.sts.xml	2016-09-13 21:07	74K
	m54113_160913_184949.subreads.bam 	2016-09-22 22:03	9.3G
	m54113_160913_184949.subreads.bam.pbi	2016-09-22 23:58	5.4M
	m54113_160913_184949.subreadset.xml	2016-09-23 02:08	2.8K
	md5sum.txt	2018-08-03 11:01	477

Raw Data

SAM format

SAMtools

```
module load samtools/1.9  
samtools view file.bam | less
```

Raw Data

```
m54113_160913_184949/4194800/0_1273 4 * 0
255 * * 0 0 TTTCTCCT[...]
===== [...] RG:Z:95dc44e8 cx:i:0
ip:B:C,5,6,31,19,4,8,83,45,19,145, [...]
```

Tab-separated. Field 1 = ZMW and range (id). Field 10 = sequence. Field 11 = base qualities.

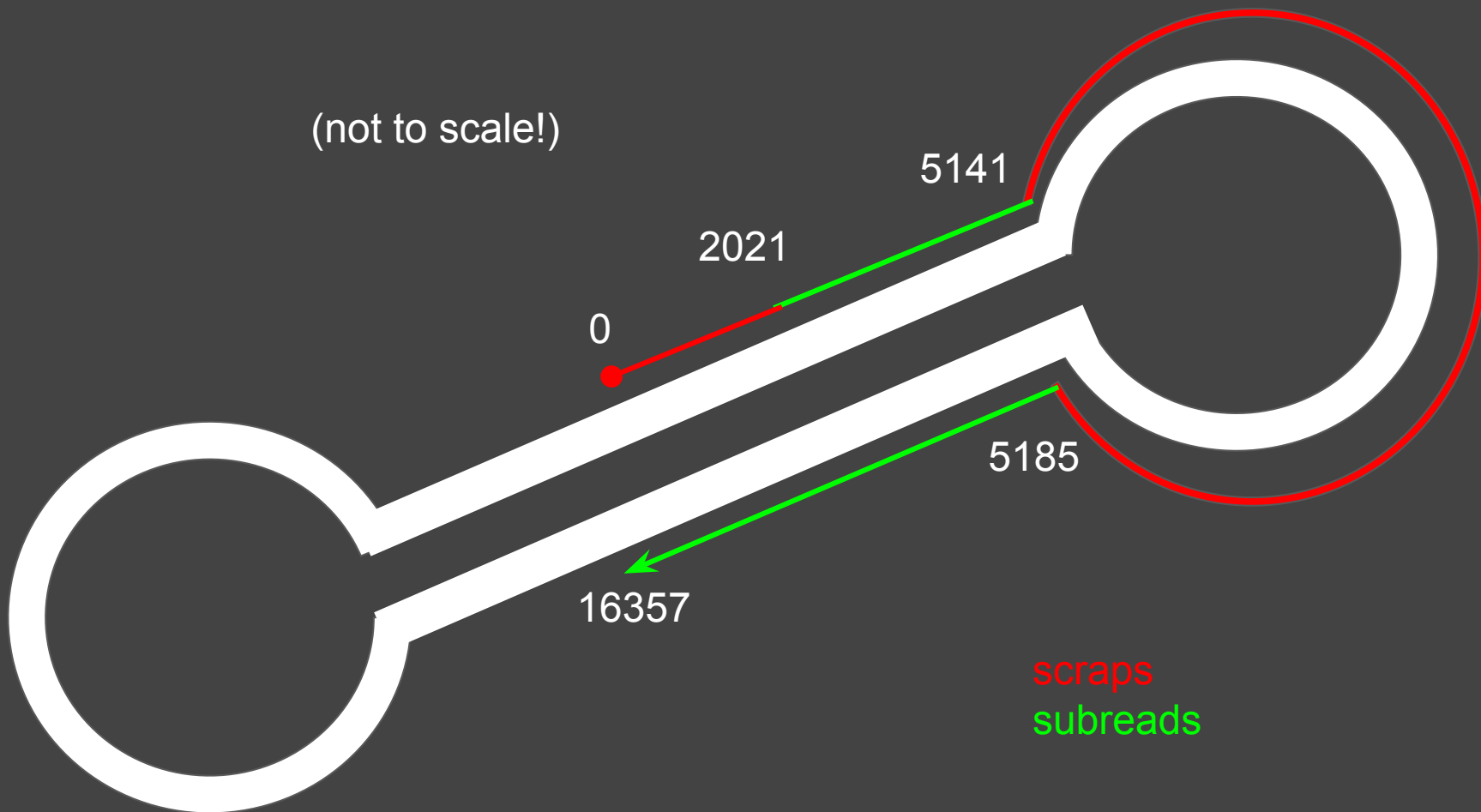
Raw Data

```
$ samtools view m54113_160913_184949.scrap.bam | cut -f1 |  
head -10000 | grep -F "/4326262/"  
m54113_160913_184949/4326262/0_2021  
m54113_160913_184949/4326262/5141_5185
```

```
$ samtools view m54113_160913_184949.subreads.bam | cut -f1  
| head -10000 | grep -F "/4326262/"  
m54113_160913_184949/4326262/2021_5141  
m54113_160913_184949/4326262/5185_16357
```

Raw Data

(not to scale!)



Assembly

CANU (a nü Celera Assembler?; Koren 2017 Genome Research 27:722) is a classic OLC assembler with updates for noisy long reads.

- **Uses MinHash (see “MHAP”) for faster alignment performance.**
- **Can use PacBio or Nanopore reads.**
- **Generates fasta as well as graphical fragment assembly (GFA) output.**

Assembly

Miniasm (& minimap; Li 2016 Bioinformatics 32:2103) is one of the first (if not first) to assemble *uncorrected* reads, rather than trying to correct long reads with shorter reads before assembly. Uses all-versus-all alignment of uncorrected reads with minimap, which aligns quickly because of use of minimizers.

- 1. Minimap all versus all.**
- 2. Miniasm using alignment.**
- 3. Correct assembly.**

The advantage here is that step 1 can be run once, and is the more time-consuming step. Then step 2 can be run multiple times with different parameters. After this, one would run a polishing step, using either high accuracy short reads, or the same long reads used in the assembly, to correct the assembled sequence (quicker than correcting the input reads).

Polishing

Racon (Rapid consensus; Vaser 2017 Genome Research 27:737 doi: 10.1101/gr.214270.116) is a consensus-finding tool intended for assemblers that assemble raw reads into sequences with same accuracy as reads. Can also be used as a read correction tool.

Integration across technologies

QuickMerge (Chakraborty 2016 Nucleic Acids Research 44:e147) uses MUMmer to align scaffolds from different assemblies (say, from different technologies) and construct super-contigs. Can also be used to further scaffold an Illumina assembly using a small (inexpensive) amount of long reads (PacBio, Nanopore).