Tillandsia genome assemblies - Final strategy and data

Status June 2020

Table of Contents

Tillandsia fasciculata

Tillandsia leiboldiana

Tillandsia fasciculata - overview

1. CANU assembly (PacBio data; CANU version 1.8)

Basic assembly of long read data

2. Purge Haplotigs

Since T. fasciculata has relatively high heterozygosity, this step removed redundant contigs from assembled haplotypes

3. Dovetail Chicago + HiC

Scaffolding using first Chicago and then the HiC data

4. Pilon

Final polishing to improve base quality and detect indel errors - note we skip correcting small structural variation since we don't have sufficient Illumina coverage for this

Tillandsia fasciculata - CANU v. 1.8

Because we had coverage at the lower end of the range required (33x), I ran CANU with two rounds of read error correction which resulted in improved contiguity and clearer peaks in the k-mer spectrum of corrected reads. Run with low coverage settings, high heterozygosity setting and to optimise memory for large repeat %

Initial error correction

```
canu -correct -p Tfas -d /scratch2/hess/TfasHiSen genomeSize=800m -pacbio-raw /scratch2/hess/TfasHiSen/DTG-
DNA-541.subreads.fasta.gz maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40' gridOptions='--
constraint=array-8core --nice=10000' minThreads=8 corMhapSensitivity=high corMinCoverage=0 corOutCoverage=200
gridEngineMemoryOption='--mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' stageDirectory=\$TMPDIR
gridOptionsJobName=Tfa correctedErrorRate=0.105 corMhapFilterThreshold=0.0000000002 corMhapOptions=" --repeat-idf-scale
50" mhapMemory=60g mhapBlockSize=500
```

Second error correction

```
canu -correct -p Tfas -d /scratch2/hess/TfasHiSen_rnd2 genomeSize=800m -pacbio-raw /scratch2/hess/TfasHiSen/
Tfas.correctedReads.fasta.gz maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40' gridOptions='--
constraint=array-8core --nice=10000' minThreads=8 corMinCoverage=0 corOutCoverage=200 gridEngineMemoryOption='--
mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' stageDirectory=\$TMPDIR gridOptionsJobName=Tfa
correctedErrorRate=0.105 corMhapFilterThreshold=0.0000000002 corMhapOptions=" --repeat-idf-scale 50" mhapMemory=60g
mhapBlockSize=500
```

Trim and assemble

```
canu -trim-assemble -p Tfas -d /scratch2/hess/TfasHiSen_rnd2 genomeSize=800m -pacbio-corrected /scratch2/hess/
TfasHiSen_rnd2/Tfas.correctedReads.fasta.gz maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40'
gridOptions='--constraint=array-8core' minThreads=8 corMinCoverage=0 corOutCoverage=200 gridEngineMemoryOption='--
mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' stageDirectory=\$TMPDIR gridOptionsJobName=Tfa
correctedErrorRate=0.105
```

Tillandsia fasciculata - Purge Haplotigs (pulled June 2019)

T. fasciculata had a relatively high rate of heterozygosity (see previous reports) so there was a good chance that we have assembled jhfd

Align PacBio data using minimap2

```
minimap2 -t 16 -ax map-pb TfasHiSen_rnd2.contigs.fasta DTG-DNA-541.subreads.fasta.gz --secondary=no \
   | samtools sort -m 1G -o /scratch2/hess/Tfas/align/TfasHiSen_rnd2.contigs.aligned.bam -T $TMPDIR/tmp.ali
```

Run step 1

```
purge_haplotigs  hist  -b /scratch2/hess/Tfas/align/TfasHiSen_rnd2.contigs.aligned.bam  -g 
TfasHiSen_rnd2.contigs.fasta  -t 16
```

Run step 2

# based on coverage histogram set low depth cutoff at 5, high cutoff at 75 and mid cutoff at 22 (no clear trough - best guess)

```
purge_haplotigs  cov  -i TfasHiSen_rnd2.contigs.aligned.bam.gencov -l 5 -h 75 -m 22
```

Run step 3

```
purge_haplotigs purge -g TfasHiSen_rnd2.contigs.fasta  -c coverage_stats.csv -t 16 -d -b /scratch2/hess/Tfas/align/
TfasHiSen_rnd2.contigs.aligned.bam
```

Tillandsia fasciculata - Chicago and HiC

Please refer to Dovetail reports

Tillandsia fasciculata - Pilon v.1.22 / BWA 0.7.16a

Round 1

```
# index genome
bwa index tillandsia_fasciculata_02Oct2019_y1aHk.fasta

# align reads
bwa mem -t 12 tillandsia_fasciculata_02Oct2019_y1aHk.fasta Tfa_0024657_1_trimmed_paired.fq.gz
Tfa_0024657_2_trimmed_paired.fq.gz | samtools view -Sb - | samtools sort -@4 - -o
tillandsia_fasciculata_02Oct2019_y1aHk.aligned.sorted.bam

# index alignment
#samtools index tillandsia_fasciculata_02Oct2019_y1aHk.aligned.sorted.bam

# run Pilon
java -Xmx800G -jar /apps/pilon/1.22/pilon-1.22.jar --threads 8 --fix snps,indels --diploid --genome
tillandsia_fasciculata_02Oct2019_y1aHk.fasta --frags tillandsia_fasciculata_02Oct2019_y1aHk.aligned.sorted.bam
--output tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rnd1.snp.indel
```

Round 2

```
# index genome
bwa index tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.fasta

# align reads
bwa mem -t 12 tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.fasta
Tfa_0024657_1_trimmed_paired.fq.gz Tfa_0024657_2_trimmed_paired.fq.gz | samtools view -Sb - | samtools sort -@4
- -o tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.aligned.sorted.bam

# index alignment
#samtools index tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.aligned.sorted.bam

# run Pilon
java -Xmx800G -jar /apps/pilon/1.22/pilon-1.22.jar --threads 8 --fix snps,indels --diploid --genome
tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.fasta --frags
tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn1.snp.indel.aligned.sorted.bam --output
tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rdn2.snp.indel
```

Tillandsia fasciculata - Read Files

PacBio

DTG-DNA-541.subreads.fasta.gz - PacBio Sequel reads ca. 33x

Illumina (can be combined for higher coverage dataset)

High coverage run to boost coverage for error correction - approx. 35x
HiCov_run.Tfa_0024657_1_trimmed_paired.fq.gz
HiCov_run.Tfa_0024657_2_trimmed_paired.fq.gz

Part of Tillandsia genome survey run dataset (Tillandsia_genomeSurvey_run) - approx 15x*
TillandsiaWGS.Tfa_0024657_1_trimmed_paired.fq.gz
TillandsiaWGS.Tfa_0024657_2_trimmed_paired.fq.gz

Both raw read sets were processed using Trimmomatic v. 0.38 to result in the files above. See full command below for details. The raw read files can be found on the NAS device (need to figure out details together)

```
java -jar ~/Software/Trimmomatic-0.38/trimmomatic-0.38.jar PE -threads 16 TillandsiaWGS.${sam_id}_1.fastq.gz
TillandsiaWGS.${sam_id}_2.fastq.gz TillandsiaWGS.${sam_id}_1_trimmed_paired.fq.gz TillandsiaWGS.${sam_id}
_1_trimmed_unpaired.fq.gz TillandsiaWGS.${sam_id}_2_trimmed_paired.fq.gz TillandsiaWGS.${sam_id}
_2_trimmed_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:75
```

Tillandsia fasciculata - Key Intermediate Files

1. CANU assembly (PacBio data; CANU version 1.8)

TfasHiSen_rnd2.contigs.fasta

2. Purge Haplotigs (sent to Dovetail)

Tfas_HiSenLong_rnd2.curated.fasta

3. Dovetail Chicago + HiC

tillandsia_fasciculata_02Oct2019_y1aHk.fasta (Chicago + HiC)

4. Pilon (Final assembly file)

tillandsia_fasciculata_02Oct2019_y1aHk.pilon.rnd2.snp.indel.fasta

Tillandsia leibolidana - overview

1. CANU assembly (PacBio data; CANU version 1.8)

Basic assembly of long read data

2. Polish using Arrow

Since T. leiboldiana is more homozygous and the assembly size was consistent with a largely homozygous assembly, I did not purge haplotigs. Instead, I used the somewhat higher PacBio coverage for Arrow polishing.

3. Dovetail Chicago + HiC

Scaffolding using first Chicago and then the HiC data

4. Pilon

Final polishing to improve base quality and detect indel errors - note we skip correcting small structural variation since we don't have sufficient Illumina coverage for this

Tillandsia leiboldiana - CANU v. 1.8

Because T. leiboldiana has such a large repeat content I was unable to run two rounds of error correction (the storage required would have been more than 10 TB!). The assembly was still run with high sensitivity settings to accommodate low-ish coverage, and basic settings to mitigate high frequency reapeats.

Error correction

```
canu -correct -p Tlei -d /scratch2/hess/TleiHiSen genomeSize=1.2g -pacbio-raw DTG-DNA-499.subreads.bam.fasta
maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40' gridOptions='--constraint=array-8core'
minThreads=8 ovlMerThreshold=500 corMhapSensitivity=high corMinCoverage=0 corOutCoverage=200 correctedErrorRate=0.105
gridEngineMemoryOption='--mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' gridOptionsJobName=Tle
stageDirectory=\$TMPDIR
```

Trim

```
canu -trim -p Tlei -d /scratch2/hess/TleiHiSen genomeSize=1.2g -pacbio-corrected /scratch2/hess/TleiHiSen/
Tlei.correctedReads.fasta.gz maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40' gridOptions='--
constraint=array-8core' minThreads=8 ovlMerThreshold=500 corMinCoverage=0 corOutCoverage=200 correctedErrorRate=0.105
gridEngineMemoryOption='--mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' gridOptionsJobName=Tle
stageDirectory=\$TMPDIR
```

Assemble

```
canu -assemble -p Tlei -d /scratch2/hess/TleiHiSen genomeSize=1.2g -pacbio-corrected /scratch2/hess/TleiHiSen/
Tlei.trimmedReads.fasta.gz maxMemory=320G maxThreads=20 gridEngineArrayOption='-a ARRAY_JOBS%40' gridOptions='--
constraint=array-8core' minThreads=8 ovlMerThreshold=500 corMinCoverage=0 corOutCoverage=200 correctedErrorRate=0.105
gridEngineMemoryOption='--mem=MEMORY' gridEngineThreadsOption='--cpus-per-task=THREADS' gridOptionsJobName=Tle
stageDirectory=\$TMPDIR
```

Tillandsia leiboldiana - Arrow v.2.3.3 / PBMM2 v.1.0

Round 1

```
# align reads (using raw .bam files)

pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-DNA-499_c1.subreads.bam
scratch2/hess/Tlei/align/Tlei.contigs.c1.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-DNA-499_c2.subreads.bam
scratch2/hess/Tlei/align/Tlei.contigs.c2.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-DNA-499_c3.subreads.bam
scratch2/hess/Tlei/align/Tlei.contigs.c3.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-DNA-499_c4.subreads.bam
scratch2/hess/Tlei/align/Tlei.contigs.c4.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-DNA-499_c5.subreads.bam
scratch2/hess/Tlei/align/Tlei.contigs.c5.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c1.subreads.bam /scratch2/hess/Tlei/align/Tlei.contigs.c6.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c2.subreads.bam /scratch2/hess/Tlei/align/Tlei.contigs.c7.bam
pbmm2 align --sort -j 14 Tlei.contigs.fasta /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c3.subreads.bam /scratch2/hess/Tlei/align/Tlei.contigs.c8.bam

samtools merge /scratch2/hess/Tlei/align/Tlei.contigs.pbmm2.bam /scratch2/hess/Tlei/align/Tlei.contigs.c1.bam /scratch
hess/Tlei/align/Tlei.contigs.c2.bam /scratch2/hess/Tlei/align/Tlei.contigs.c3.bam /scratch2/hess/Tlei/align/
Tlei.contigs.c4.bam /scratch2/hess/Tlei/align/Tlei.contigs.c5.bam /scratch2/hess/Tlei/align/Tlei.contigs.c6.bam /
scratch2/hess/Tlei/align/Tlei.contigs.c7.bam /scratch2/hess/Tlei/align/Tlei.contigs.c8.bam

#index
pbindex /scratch2/hess/Tlei/align/Tlei.contigs.pbmm2.bam
samtools faidx Tlei.contigs.fasta


# run arrow
arrow -j14 /scratch2/hess/Tlei/align/Tlei.contigs.pbmm2.bam -r Tlei.contigs.fasta -o
Tlei.contigs.arrow_rnd1.variants.gff -o Tlei.contigs.arrow_rnd1.consensus.fasta -o
Tlei.contigs.arrow_rnd1.consensus.fastq
```

Tillandsia leiboldiana - Arrow v.2.3.3 / PBMM2 v.1.0

Round 2

```
# align reads
pbmm2 index Tlei.contigs.arrow_rnd1.consensus.fasta Tlei.contigs.arrow_rnd1.consensus.mmi
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-
DNA-499_c1.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c1.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-
DNA-499_c2.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c2.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-
DNA-499_c3.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c3.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-
DNA-499_c4.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c4.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/DTG-
DNA-499_c5.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c5.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c1.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c6.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c2.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c7.bam
pbmm2 align --sort -j 14 Tlei.contigs.arrow_rnd1.consensus.mmi /proj/hess/Tillandsia_genomes/Tlei/DATA/PacBio/SM01-DTG-
DNA-499_c3.subreads.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c8.bam

samtools merge /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.pbmm2.bam /scratch/hess/Tlei/align/
Tlei.contigs.arrow_rnd1.consensus.c1.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c2.bam /scratch/
hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c3.bam /scratch/hess/Tlei/align/
Tlei.contigs.arrow_rnd1.consensus.c4.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c5.bam /scratch/
hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c6.bam /scratch/hess/Tlei/align/
Tlei.contigs.arrow_rnd1.consensus.c7.bam /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.c8.bam

#index
pbindex /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.pbmm2.bam
samtools faidx Tlei.contigs.arrow_rnd1.consensus.fasta


# run arrow
 arrow -j14 /scratch/hess/Tlei/align/Tlei.contigs.arrow_rnd1.consensus.pbmm2.bam -r
 Tlei.contigs.arrow_rnd1.consensus.fasta -o Tlei.contigs.arrow_rnd2.variants.gff -o
 Tlei.contigs.arrow_rnd2.consensus.fasta -o Tlei.contigs.arrow_rnd2.consensus.fastq
```

Tillandsia leiboldiana - Chicago and HiC

Please refer to Dovetail reports

Tillandsia leiboldiana - Pilon v.1.22 / BWA 0.7.16a

Round 1

```
# index genome
bwa index tillandsia_leiboldiana_13Sep2019_jOWHO.fasta

# align reads
bwa mem -t 8 tillandsia_leiboldiana_13Sep2019_jOWHO.fasta HiCov_run.Tle_0024715_1_trimmed_paired.fq.gz
HiCov_run.Tle_0024715_2_trimmed_paired.fq.gz  | samtools view -Sb - | samtools sort -@4 - -o
tillandsia_leiboldiana_13Sep2019_jOWHO.align.sorted.bam

# index alignment
samtools index tillandsia_leiboldiana_13Sep2019_jOWHO.align.sorted.bam

# run Pilon
java -Xmx800G -jar /apps/pilon/1.22/pilon-1.22.jar --threads 8 --fix snps,indels --diploid --genome
tillandsia_leiboldiana_13Sep2019_jOWHO.fasta --frags tillandsia_leiboldiana_13Sep2019_jOWHO.align.sorted.bam --
output tillandsia_leiboldiana_13Sep2019_jOWHO.pilon.rnd1.snp.indel
```

I also ran a second round of Pilon, but it made the results worse than just one round - presumably due to the additional polishing with Arrow.

Tillandsia leiboldiana - Read Files

PacBio

DTG-DNA-499.subreads.bam.fasta.gz - PacBio Sequel reads ca. 40x

Illumina (can be combined for higher coverage dataset)

High coverage run to boost coverage for error correction - approx. 35 x
HiCov_run.Tle_0024715_1_trimmed_paired.fq.gz
HiCov_run.Tle_0024715_2_trimmed_paired.fq.gz

Part of Tillandsia genome survey run dataset (Tillandsia_genomeSurvey_run) - approx 15x*
TillandsiaWGS.Tle_0024715_1_trimmed_paired.fq.gz
TillandsiaWGS.Tle_0024715_1_trimmed_paired.fq.gz

Both raw read sets were processed using Trimmomatic v. 0.38 to result in the files above. See full command below for details.

```
java -jar ~/Software/Trimmomatic-0.38/trimmomatic-0.38.jar PE -threads 16 TillandsiaWGS.${sam_id}_1.fastq.gz
TillandsiaWGS.${sam_id}_2.fastq.gz TillandsiaWGS.${sam_id}_1_trimmed_paired.fq.gz TillandsiaWGS.${sam_id}
_1_trimmed_unpaired.fq.gz TillandsiaWGS.${sam_id}_2_trimmed_paired.fq.gz TillandsiaWGS.${sam_id}
_2_trimmed_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:75
```

Tillandsia fasciculata - Key Intermediate Files

1. CANU assembly (PacBio data; CANU version 1.8)

Tlei.contigs.fasta

2. Arrow polishing (sent to Dovetail)

Tlei.contigs.arrow_rnd2.consensus.fasta

3. Dovetail Chicago + HiC

tillandsia_leiboldiana_13Sep2019_jOWHO.fasta (Chicago + HiC)

4. Pilon (Final assembly file)

tillandsia_leiboldiana_13Sep2019_jOWHO.pilon.rnd1.snp.indel.fasta