UCEs Analysis !!

~~18 June 2015~~ (3)
18 June 2015

Data arrived 8 June 2015.

Download *.fastq.bz2 from Rapid Genomics secure server:

FTP: https://bio.rc.ufl.edu/rgenomics/LSU_4664/
User: [redacted]
Password: [redacted]

→ Installed (wget) from tannerwilliamson.com (Snow Leopard):
~~dl/unzip zip file, cd dir~~
~~sudo mkdir~~

(1) → * wget -r -np -K --user USER --password PASSWORD --no-check-certificate FTP

(2) → Then also download to LSU HPC SuperMike2 (same command)
* /work/cenewman/rawdata/compressed_files_bz2

(3) HPC: unzipped *.bz2 → ~~(necessary? not yet)~~ → cen_bunzip2.qsub
↳ /work/cenewman/rawdata/fastq_files

(4) QC & adapter/barcode trimming – Illumiprocessor

* input files need to be *.fastq.gz → rezip: cen_gzip.qsub
* input files renamed: SLxxxxx_L006_R1_###.fastq.gz
" " " R2_###.fastq.gz

↳ RG sample# — lane — run 1 or 2 — my sample #

→ SEE FILE: LSU_466401_Sample_Data.xlsx
*gz_file_rename_for_Illumiprocessor.qsub
↳ /work/cenewman/rawdata/raw_reads_for_illumiprocessor

* Illumiprocessor config file: Pserratus_illumiprocessor.conf
* HPC job submission file: illumiprocessor.qsub
* Output: /work/cenewman/rawdata/cleaned_reads_from_illumiprocessor

(5) Assembly – Phyluce with Trinity

22 June 2015

(* see: https//github.com/faircloth-lab/phyluce/blob/master/docs/assembly.rst)

* Phyluce Trinity config file: phyluce_trinity.conf
* HPC job submission file: phyluce_trinity.qsub

★ 1st run: error immediately – requires Java 1.7 (had 1.6) →
installed v.1.7 (32bit) → ran 2.5hr, error again = ↓

(trinity.log → "Invalid maximum heap size -Xmx4G"
↳ installed Java v.1.7 64-bit ★
running now... ⟶ THIS WORKS! ⇨

④

23 June 2015

Assembly, cont.

* Phyluce / Trinity took 3.5 hrs to complete 1 sample. → on checkpt 32 GB 1 node, 16 ppn
  ↳ (current version in phyluce) = v. 2.0.6
  - Run on bigmem (256 GB) did not complete more quickly.

* Split ~~96 samples~~ 95 samples (minus the 1 already done) into 10 groups → run 10 jobs simultaneously on checkpt.

- 5 jobs w/ 10 samples = 50 = 35 hrs
- 5 jobs w/ 9 samples = 45 = 31.5 hrs
  95

Jobs:

#HOME/phyluce_trinity_checkpt1.qsub
" " " " " 2.qsub
" " " " " 3.qsub
" " " " " 4.qsub
" " " " " 5.qsub
" " " " " 6.qsub
" " " " " 7.qsub
" " " " " 8.qsub
" " " " " 9.qsub
" " " " " 10.qsub

* Run each job in separate working directory:

~~/work/cenewman/~~ /work/cenewman/rawdata/phyluce_trinity1
" " " " " 2
" " " " " 3
" " " " " 4
" " " " " 5
" " " " " 6
" " " " " 7
" " " " " 8
" " " " " 9
" " " " " 10

* In each working directory: → (2, 3, etc.)
  * Config file: phyluce_trinity1.conf
  * /cleaned_reads 1 (folder of output folders from illumiprocessor)
  * /assemblies_trinity → Trinity output (created by Trinity)

STARTED RUNS: ~ 11:00 AM TUES 23 JUNE

- Job 1 ~~failed (BDT 0554)~~ → ~~moved to end of samples & restarted job.~~
- Job 8 ~~failed (BDT 0941)~~ → " " " " " " " "

28 June 2015 ⑤

I don't know why some assemblies failed, but all completed successfully after 2 or 3 attempts, at most. When run failed, I deleted all created files & restarted run.

All assemblies completed.

/work/~~BBBB~~cenewman/rawdata/assemblies_trinity

UCE Processing (post-assembly)

http://github.com/faircloth-lab/phyluce/blob/master/docs/uce-processing.rst

★UCE probe set fasta: LSU_466401_probes.fasta

IMPORTANT → SQLite does not work on HPC server!

↳ These steps must be done on computer (d/l contigs from HPC)

~~① Match~~

① phyluce_assembly_match_contigs_to_probes
(make sure to rename probes: uce-NNNN_pN

② Creating data matrix configuration file
- no complete matrix (all loci missing at least 1 taxon)
- incomplete: phyluce_assembly_get_match_counts
↳ [dataset1] = all 96 samples (datasets.conf)

③ Extracting FASTA data using data matrix config file
- phyluce_assembly_get_fastas_from_match_counts

└ END OF SQLITE PROCESSES → CAN DO REMAINDER ON HPC ↓]

④ ALIGNMENT & TRIM

- phyluce_align_seqcap_align (output = /mafft-nexus/)
- phyluce_align_get_align_summary_data
★ NOTE: % in "Data matrix completeness" based on total taxa in TaxaMax
- phyluce_align_remove_locus_name_from_nexus_lines
(output = /mafft-nexus-clean/)

ON MY COMP ① downloaded cleaned alignments.

② For each sample:
grep -R 'BDT041' alignments/* > BDT041.txt

③ Created spreadsheet locus x sample → locus-by-taxon-spreadsheet.XLSX

⑥ ④ Omit all samples w/ fewer than ~~100~~ 15% loci amplified:

BDT	078	– 44 loci	– 3%
BDT	121	– 49 "	– 3%
BDT	124	– 101 "	– 7%
BDT	125	– 231 "	– 15%
BDT	159	– 28 "	– 2%
BDT	160	– 80 "	– 5%
BDT	167	– 17 "	– 1%
MUZ	145052	– 20 "	– 1%
UAHC	14920	– 19 "	– 1%

✱ Delete these taxa from ~~dataset1.conf~~ = dataset2.conf

⇓

[dataset2]

✱ Re-run all from phyluce_assembly_get_match_counts using dataset2

HPC

⑤ Build datasets w/ varying % completion:

↳ phyluce_align_get_only_loci_with_min_taxa

↳ (phyluce_align_cull.qsub on HPC)

Getting Coverage data

phyluce_assembly_get_trinity_coverage_for_uce_loci

UCEs: Pop. Gen. Processing | 24 Sept. 2015

make sure Anaconda is installed

https://github.com/mgharvey/seqcap_pop

★Starting w/ "full" set of 87 samples★
★Starting @ Step 4, using CEN131.fasta from Brant's pipeline (the most contigs matching to probes).

④ BWA already installed on Supermikell - add to path.
- use output from cleaned-reads_from_illumiprocessor
- I split into 4 simultaneous jobs.
- None failed
- Output: 5-mapping/*.sa.sai and *.sam → BIG ~~HUGE~~ FILES

⑤ Samtools already installed on Supermikell.
- Output: 5-mapping/*.bam → BIG FILES
- After this step, I deleted *.sam files.

⑥ Output: 6_picard/*-aln_CL.bam → BIG FILES

⑦ Output: 6_picard/*-aln_RG.bam → BIG FILES

⑧ Output: 6_picard/*-aln_MD.bam → BIG FILES

⑨ Start here to get new SNP dataset for different sets of samples!
- Can choose which samples (bam files) to include in merged bam file.
- Output: 7_merge-bams/Pserratus.bam → MASSIVE FILE (30 GB)

⑩ (indexing)

⑪ Output: poggen/CEN131.dict

⑫ (indexing)

⑬ Output: 8_gatk/Pserratus.intervals

⑭ Output: 8_gatk/Pserratus_RI.bam

⑮ Output: 8_gatk/Pserratus_raw_SNPs.VCF

⑯ Output: 8_gatk/Pserratus_SNPs_annotated.VCF

⑰ Output: 8_gatk/Pserratus_SNPs_indels.VCF

⑱ ★add to command:
--disable_auto_index_creation_and_locking_when_reading_rods
- Output: 8_gatk/Pserratus_SNPs_no_indels.VCF

• if errors on HPC

⑥ Pop Gen, Cont.

⑲ Output: 8_gatk/Pserratus_SNPs_pass-only.VCF

⑳ *add same command option as step 18* (if error on HPC)
- Output: 8_gatk/Pserratus_SNPs_phased.VCF

↳ Use this File for Conversion to:

- Structure / Structurama
- fastStructure
- Adegenet
- Genepop

~~Population Level Analysis~~ 17 July 2015

~~Brant Faircloth's github~~ ~~github.com/mgharvey/seqcap_pop~~

Making another phylogenetic dataset

30 Sept. 2015

- Trying to balance # of taxa and # of loci.
- Experimenting with varying amounts of missing data.

Dataset600: all samples that amplified ≥ 600 loci (75 total)
(* excluded outgroup LSUMZ 15569, retained 15568)

- Re-run from phyluce_assembly_get_match_counts (p. 5~~3~~, step 2)

(10)

For *BEAST species tree

[From dataset 600, 80%]

20 Oct 2015

most informative 20 loci

models of evolution (BIC, jModelTest)

Locus	Model	# Informative Sites	Length (bp)	# Samples
uce-14553	JC	39	749	63
uce-123865	JC	31	725	64
uce-130414	JC	30	766	60
uce-84233	JC	27	665	64
uce-87295	K80	27	641	63
uce-122432	JC	27	758	66
uce-69626	JC	26	540	64
uce-129912	K80	25	755	67
uce-515078	JC	25	713	61
uce-28629	JC	24	662	66
uce-35877	JC	24	843	65
uce-36937	JC	24	696	63
uce-65890	JC	24	668	65
uce-69811	HKY	24	679	60
uce-510176	F81	24	790	68
uce-69597	JC	23	546	65
uce-131740	K80	23	724	68
uce-35516	JC	22	550	64
uce-214216	F81	22	684	62
uce-311879	K80	22	682	62

↳ (before deleting admixed individuals)

cont.→

For Starbeast Species tree (cont.)

20 oct. 2015 ①

From K=6 (all_taxa → Hpc results → 6/2 (-17884.7)) ~~(6/3, -14652)~~ – delete admixed (≥ 0.10)

- BDT 042
- BDT 101
- BDT 103
- BDT 109
- BDT 111
- BDT 163
- BDT 164
- BDT 169
- BDT 173
- BDT 175
- FHSM 15547
- " 15548
- " 15549
- BDT 044
- BDT 100
- BDT 168
- BDT 174
- BDT 041 ?
- MVZ 145058
- " 145059
- OMNH 41642

USE ↙

From dataset3-1k → 80% → NO HETS : ↘

Locus	Model	# Informative Sites	Length (bp)	# Samples
uce-44009	JC	35	632	36
uce-113558	JC	33	693	39
uce-28703	F81	31	632	36
uce-6062	F81	31	670	38
uce-123865	F81	30	770	40
uce-125851 X	HKY	29	1067	39
uce-130414	JC	29	780	36
uce-14553	JC	29	744	40
uce-87295	K80	29	681	40
uce-122432	JC	26	762	40
uce-224728	JC	26	726	38
uce-33880	HKY	26	661	38
uce-84233 X	JC	26	695	40
uce-122334	K80	25	626	37
uce-28629	JC	25	710	40
uce-69626	JC	25	554	37
uce-84091 X	K80	25	504	36
uce-126661	F81	24	765	40
uce-321510 X	HKY	24	701	37
uce-35877	K80	24	859	43

X = removed because missing "species" (see ↘)

(12)
26 Oct 2015

*BEAST Species Tree

Dataset #3 - 1K → 80% → NO HETS

- Deleted hets (p. 11)
- Deleted loci that don't have at least 1 sample for each "species" (as designated for *Beast)

Top 20 most Informative Loci

Locus	Model	# Informative Sites	Length (bp)	# Samples
uce-44009	JC	35	632	36
uce-113558	JC	33	693	39
uce-28703	F81	31	632	36
uce-6062	F81	31	670	38
uce-123865	F81	30	770	40
uce-130414	JC	29	780	36
uce-14553	JC	29	744	40
uce-87295	K80	29	681	40
uce-122432	JC	26	762	40
uce-224728	JC	26	726	38
uce-33880	HKY	26	661	38
uce-122334	K80	25	626	37
uce-28629	JC	25	710	40
uce-69626	JC	25	554	37
uce-126661	F81	24	765	40
uce-35877	K80	24	859	43
uce-69597	K80	24	609	41
uce-133907	HKY	23	687	39
uce-3378	HKY	23	646	37
uce-69811	HKY	23	711	42

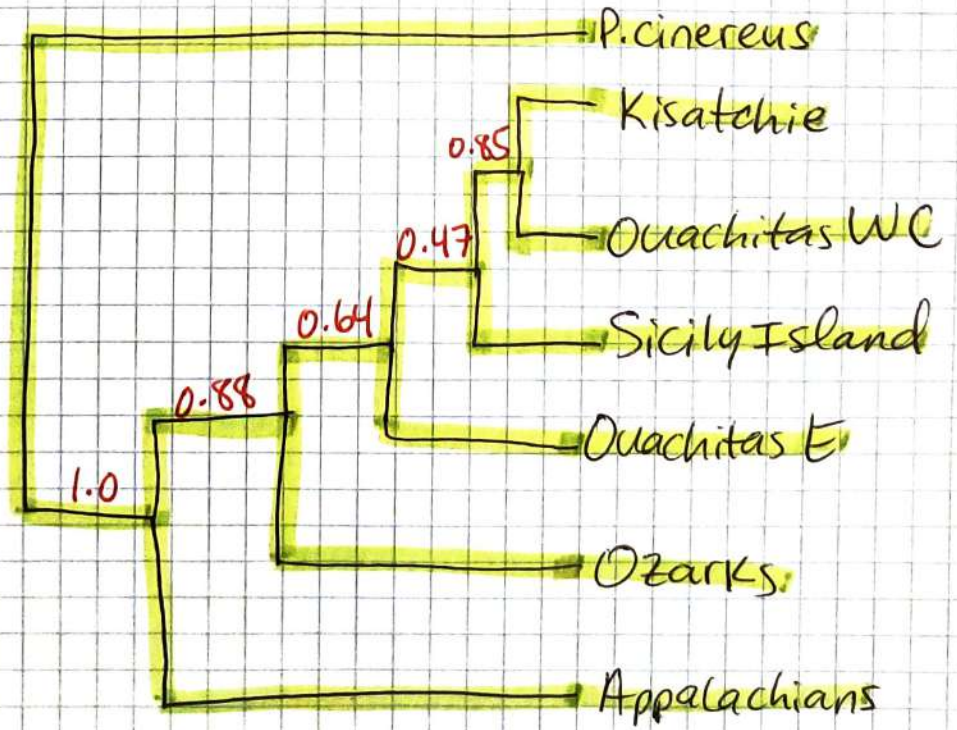
*BEAST Species Tree

26 Oct. 2015

13

Top 20 most informative loci (p.12)

- Model of evolution
  - Either JC or HKY (empirical base freq.)
- All loci : strict clock, fixed rate = 1
- Yule species tree
- Population size: Linear with Constant root
- birthRate.t: Species = exponential
  - ↳ (Smith et al. says $n = 1000$ but I don't know how to set that.)
- pop mean = Lognormal
  - ↳ (Smith et al.: $n = 0.001$, SD = 2; I don't know how to set this.)
- MCMC = 1 billion generations
  Store every 5,000
- Ran on big mac in lab → total time: 48.3 hrs (2m53s/million)
- TreeAnnotator: burn-in = 2,000 trees (10 million generations)
- Tracer: All ESSs > 2,000 !!!



(14) 26 Oct. 2015

* BEAST Species Tree

Top 50 most Informative Loci

Locus	Model	# Informative Sites	Length (bp)	# Samples
uce-44009	JC	35	632	36
uce-113558	JC	33	693	39
uce-28703	F81	31	632	36
uce-6062	F81	31	670	38
uce-123865	F81	30	770	40
uce-130414	JC	29	780	36
uce-14553	JC	29	744	40
uce-87295	K80	29	681	40
uce-122432	JC	26	762	40
uce-224728	JC	26	726	38
uce-33880	HKY	26	661	38
uce-122334	K80	25	626	37
uce-28629	JC	25	710	40
uce-69626	JC	25	554	37
uce-126661	F81	24	765	40
uce-35877	K80	24	859	43
uce-69597	K80	24	609	41
uce-133907	HKY	23	687	39
uce-3378	HKY	23	646	37
uce-69811	HKY	23	711	42
uce-113471	JC	22	722	40
uce-15409	F81	22	911	41
uce-311879	SYM	22	719	40
uce-36937	JC	22	731	39
uce-65890	JC	22	698	42
uce-129912	K80	21	775	43
uce-131753	JC	21	664	40
uce-14545	F81	21	675	36
uce-213502	HKY	21	808	40
uce-35605	JC	21	518	36
uce-35857	F81	21	716	39
uce-113603	JC	20	624	40
uce-123129	JC	20	618	37
uce-223133	F81	20	755	41
uce-24100	JC	20	585	40
uce-35516	K80	20	610	39
uce-69742	F81	20	683	43
uce-113350	HKY	19	635	45
uce-116409	K80	19	761	41
uce-119202	F81	19	653	41
uce-21916	JC	19	714	43
uce-27190	JC	19	698	39
uce-113605	F81	18	686	41
uce-131740	K80	18	739	42
uce-14591	HKY	18	718	41
uce-22324	HKY	18	669	39
uce-225794	F81	18	703	38
uce-317628	K80	18	635	38
uce-3388	K80	18	732	42
uce-418072	K80	18	696	39

* BEAST Species Tree
Top 50 most Informative Loci (p. 14)

26 Oct. 2015

- Same parameters as top 20 (p. 13)
- mcmc = 500 million generations
  Save every 5,000
- Running on big mac in lab