Data simulation

MetaDMG

#Use metaDMG-cpp to get the damage and read length info from a preexisting bam file.

metaDMG-cpp getdamage Example.bam -o output

Nucleotide mismatch

#Extract the .bdamage output of metaDMG-cpp, and then print it as a nucleotide mismatch table

metaDMG-cpp print output.bdamage/output.bdamage > output_nucleotide_mismatch_table.txt

Simulation

#Using the mismatch and length file, run NGSNGS to create a simulation. The -i is the reference genome you want to simulate

ngsngs -i Dryas_octopetala.fasta -c 20 -lf output_leng_dist.txt -f fq -qs 40 -seq PE -circ -mf output_nucleotide_mismatch_table.txt -o output_name

-c is depth of coverage

-f is output file format

-qs sets a quality score for reads (for fq)

-seq determines whether the output is single ended or pair ended reads

-circ should be used when visualizing circular genomes

-o is the final output name.

#As an output, if using PE, you should get R1 and R2 of simulated pair ended reads.

#Concatenate the R1 and R2 files separately. Example:

cat Dryas_mt_50_depth_R2.fq Betula_nana_depth_40_R2.fq Crataegus_laevigata_depth_2_R2.fq Populus_alba_depth_30_R2.fq Vaccinium_vitis-idaea_depth_10_R2.fq Carex_divulsa_depth_10_R2.fq Carex_pendula_depth_3_R2.fq Salix_dunnii_depth_20_R2.fq > Plantmix_R2.fq

Assembly

Megahit

megahit -1 reads_R1.fq -2 reads_R2.fq -o Megahit_output_name

#Gives a folder with contigs named final.contigs.fa

SPAdes/metaSPAdes

SPAdes -1 reads_R1.fq -2 reads_R2.fq --phred-offset 33 -o SPAdes_output_name

#Add --meta if doing metaSPAdes

#Gives a folder with contigs named contigs.fasta

CarpeDeam

#CarpeDeam cannot use raw pair ended reads, it needs them merged. Merge using FLASH

flash reads_R1.fq reads_R2.fq -o reads_merged

carpedeam ancient_assemble reads_merged.extendedFrags.fastq CarpeDeam_output.fasta tmp --ancient-damage d_high

#Creates a contig file named CarpeDeam_output.fasta

Quast

#Quast only needs contigs and a reference.

#Example metaquast:

metaquast.py Plantmix_megahit/final.contigs.fa Plantmix_metaspades/contigs.fasta -r Dryas_octopetala.fasta,Betula_nana_plstd.fasta,Populus_alba_plstd.fasta,Crataegus_laevigata_plstd.fasta,Vaccinium_vitis-idaea_plstd.fasta,Carex_divulsa_plstd.fasta,Carex_pendula_plstd.fasta,Salix_dunnii_plstd.fasta -o Plantmix_metaquast

#Metaquast with all references in a directory as references. Very slow.

metaquast.py ngsngs_test/megahit_test1_all/megahit_1mil_short.fa ngsngs_test/spades_1mil_short/spades_1mil_short.fasta ngsngs10000000reads/megahit_10_09test_all/megahit_10mil_short ngsngs10000000reads/spades_10mil_short/spades_10mil_short.fasta 16_09_longreads/megahit_longreads_1mil/megahit_1mil_long.fa

16_09_longreads/spades_1mil_long/spades_1mil_long.fasta 16_09_longreads10mil/megahit_10mil_long/megahit_10mil_long.fa 16_09_longreads10mil/spades_10mil_long/spades_10mil_long.fasta -r $(ls references_119_B3_116_L1_KapK-12-1-41/* | paste -sd,) -o quast

Chloroplast whole genome phylogeny

Get aligned contig:

#In QUAST folder for a specific species reference, extract the names of contigs aligned to that reference

awk -F'\t' '{print $6}' all_alignments.tsv > aligned_contignames.txt

#using seqtk, extract all aligned contigs fasta sequence from the assembled contig file using the contig names file

seqtk subseq contigs.fa aligned_contignames.txt > aligned_contigs.fasta

Align contigs to reference and call consensus:

#index reference for use in bowtie2

bowtie2-build reference.fasta.gz reference.fasta.gz

#Script to align all contigs in current folder to reference and call consensus using angsd

../bam_consensus_script.sh ../reference.fasta.gz

#concatenate consensus sequences to fasta file with references

cat *_consensus.fa alignment_references.fasta > references_and_consensus.fasta

#input references_and_consensus.fasta in online MAFFT tool

Raxml

#construct maximum likelihood phylogeny from MAFFT output

raxmlHPC-PTHREADS -s mafft.fasta -m GTRCATX -n salix3 -p 666 -T 20 -d