Library preparation and sequencing

ddRAD libraries are produced using an IGATech custom protocol, with minor modifications with respect to Peterson *et al*. 2012 (Peterson et al. 2012). To select the best combination of the two restriction enzymes, an *in silico* analysis on the reference genome of a closely related species (if available) is performed. Selected enzymes are reported in ddRAD_analysis_report.pdf. Genomic DNA is fluorimetrically quantified, normalized to a uniform concentration and double digested. Fragmented DNA is purified with AMPureXP beads (Agencourt) and ligated to barcoded adapters. Samples are pooled on multiplexing batches and bead purified. For each pool, targeted fragments distribution is collected on BluePippin instrument (Sage Science Inc.). Gel eluted fraction is amplified with oligo primers that introduce TruSeq indexes and subsequently bead purified. The resulting libraries are checked with both Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and Bioanalyzer DNA assay (Agilent technologies, Santa Clara, CA). Libraries are processed with Illumina cBot for cluster generation on the flowcell, following the manufacturer's instructions and sequenced with V4 chemistry paired end 125bp mode on HiSeq2500 instrument (Illumina, San Diego, CA).

Double digest restriction-site associated DNA (ddRADseq) standard bioinformatic analysis includes:

- Demultiplexing of raw Illumina reads using the process_radtags utility included in Stacks v2.0 (Catchen et al. 2013).
- Assembly of the short-reads of each sample into exactly matching stacks using the ustacks utility included in Stacks v 2.0 (Catchen et al. 2013).
- Creation of the loci catalog (i.e. a set of consensus loci from all the analyzed samples) using cstacks and matching each sample against the catalog using sstacks and tsv2bam utilities included in Stacks v2.0 (Catchen et al. 2013).
- Using gstacks (Catchen et al. 2013) to pull in paired-end reads (if available), assemble the paired-end contigs and merge it with the single-end locus, align reads to the locus, and call single nucleotide polymorphisms (SNPs).
- Filtering of detected loci using the populations program included in Stacks v2.0 (Catchen et al. 2013). populations is run with option –r=0.75 in order to retain only loci that are represented in at least the 75% of the population.

Delivery Files:

- Summary report (*ddRAD_analysis_report.pdf*).
- Tab-delimited file (*Sequencing_report.txt*) with sequencing statistics.
- Folder *sequences* containing the demultiplexed FASTQ files.
- Folder *stacks* containing the following files:
  - catalog.fa.gz: a FASTA file reporting a representative consensus sequence of all the detected loci.
  - catalog.calls.gz: a VCF file reporting all the sites included in the catalog of loci. For each site, the coverage (i.e. the number of reads covering the position) and the genotype (if the site is polymorphic in the population) are reported for each sample included in the population. NOTE: site coordinates are referred to the catalog loci: CHROM is the locus ID while POS is the position with respect to the locus.
  - populations.loci.fa: a FASTA file reporting a representative consensus sequence of the retained loci, i.e. loci that are represented in at least the 75% of the population.
  - populations.snps.vcf: a VCF file with the population-wise SNP calls.
  - populations.haps.vcf: a VCF file with the population-wise haplotype calls.
  - populations.structure: polymorphic sites in Structure format.
  - populations.snps.genepop: polymorphic sites in GenePop format.
  - populations.haps.genepop: haplotypes in GenePop format.
  - populations.plink.map: data converted in the PLINK map format.
  - populations.plink.ped: data converted in the PLINK ped format.
  - populations.sumstats.tsv: a tab-delimited table reporting a standard set of population genetic statistics calculated for every variant site. File format:

Column Name	Description
Locus ID	Catalog locus identifier
Chr	Chromosome with respect to the reference genome
BP	Position on the reference genome
Col	The nucleotide site within the catalog locus, reported using a zero-based offset (first nucleotide is enumerated as 0)
Pop ID	The ID supplied to the populations program, as written in the population map file
P Nuc	The most frequent allele at this position in this population
Q Nuc	The alternative allele
N	Number of individuals sampled in this population at this site
P	Frequency of most frequent allele
Obs Het	The proportion of individuals that are heterozygotes in this population
Obs Hom	The proportion of individuals that are homozygotes in this population
Exp Het	Heterozygosity expected under Hardy-Weinberg equilibrium
Exp Hom	Homozygosity expected under Hardy-Weinberg equilibrium
Pi	An estimate of nucleotide diversity

Smoothed Pi	A weighted average of π depending on the surrounding 3σ of sequence in both directions
Smoothed Pi P-value	If bootstrap resampling is enabled, a p-value ranking the significance of π within this population
Fis	The inbreeding coefficient of an individual (I) relative to the subpopulation (S)
Smoothed Fis	A weighted average of F_{IS} depending on the surrounding 3σ of sequence in both directions
Smoothed Fis P-value	If bootstrap resampling is enabled, a p-value ranking the significance of F_{IS} within this population
HWE P-value	The probability that this variant site deviates from Hardy-Weinberg equilibrium
Private	True (1) or false (0), depending on if this allele only occurs in this population

- populations.hapstats.tsv: a tab-delimited table reporting a standard set of population genetic statistics calculated for every variant locus, taking the phased SNPs as a set of haplotypes. File format:

Name	Description
Locus ID	Catalog locus identifier
Chr	Chromosome with respect to the reference genome
BP	Position on the reference genome
Pop ID	The ID supplied to the populations program
N	Number of alleles/haplotypes present at this locus
Haplotype Cnt	Haplotype count
Gene Diversity	A measure of locus haplotype richness, similar to nucleotide-level π
Smoothed Gene Diversity	
Smoothed Gene Diversity P-value	
Haplotype Diversity	A measure of locus haplotype richness that takes into account how different haplotypes are from one another in terms of nucleotide distance
Smoothed Haplotype Diversity	
Smoothed Haplotype Diversity P-value	
HWE P-value	The probability that this locus deviates from Hardy-Weinberg equilibrium. Calculated using Guo and Thompson’s MCMC walk.
HWE P-value SE	The standard error for the HWE p-value
Haplotypes	A semicolon-separated list of haplotypes/haplotype counts in the population

References:

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. [Internet] 22:3124–3140. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23701397

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species.Orlando L, editor. PLoS One [Internet] 7:e37135. Available from: http://dx.plos.org/10.1371/journal.pone.0037135