Stacks 2 Hands-On Lecture

Documentation and examples for the *Stacks* 2 hands-on exercise at ConGen 2023, on Aug 28, 2023.

(C) Angel G. Rivera-Colón <arcolon14@gmail.com>

Associated readings

The RADseq data used for this exercise comes from the following preprint:

> Long, KM, Rivera-Colón, AG, Bennett, KFP, *et al.* (2023) **Ongoing introgression of a secondary sexual plumage trait in a stable avian hybrid zone**. *bioRxiv*. DOI: 10.1101/2023.03.30.535000

The data is used here with the authorization of the authors. The authors ask not to distribute this data without prior authorization.

The analysis described follow the general guidelines described in the *Stacks* 2 protocol manuscript:

> Rivera-Colón, AG, Catchen, JM (2022). **Population Genomics Analysis with RAD, Reprised: *Stacks* 2**. In: Verde, C., Giordano, D. (eds) *Marine Genomics. Methods in Molecular Biology*, vol 2498. Humana, New York, NY. DOI: 10.1007/978-1-0716-2313-8_7

For more information regarding PCR duplicates, please check the 2023 *Mol Ecol Resour* publication:

> Rochette, NC, Rivera-Colón, AG, Walsh, J, *et al.* (2023) **On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity.** *Molecular Ecology Resources*, 23, 1299–1318. DOI: 10.1111/1755-0998.13800

For an algorithmic description of the *Stacks* 2 software, please check the 2019 *Mol Ecol* manuscript:

> Rochette, NC, Rivera-Colón, AG, Catchen, JM. (2019) ***Stacks* 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics**. *Molecular Ecology*, 28, 4737–4754. DOI: 10.1111/mec.15253

For information on the download, installation, and documentation of the software, please visit the *Stacks* website.

Repository for the exercise

A copy of this document and the associated data can be found in https://github.com/arcolon14/congen-23.

Preparing the environment

NOTE: the directory hierarchy in the commands below referes to the ConGen2023 server.

Make a directory for the Stacks assignment.

```
$ mkdir stacks-radseq
$ cd stacks-radseq/
```

Copy the raw data from the instructors directory

```
$ cp /data/instructor_materials/Angel_Rivera-Colon/2023/arc-radseq-data.congen23.tar.gz .
```

Uncompress this directory

```
$ tar xvf arc-radseq-data.congen23.tar.gz
```

Check the contents of the directory

```
$ ls *
  alignments:
```

```
  CG_10_LIB1_104.bam  QP_06_LIB1_025.bam  RU_08_LIB1_002.bam
  CG_10_LIB1_107.bam  QP_06_LIB1_028.bam  RU_08_LIB1_004.bam
  CG_10_LIB1_110.bam  QP_06_LIB1_029.bam  RU_08_LIB1_012.bam
  CG_10_LIB1_133.bam  QP_06_LIB1_059.bam  RU_08_LIB1_015.bam
  CG_10_LIB1_135.bam  QP_06_LIB1_061.bam  RU_08_LIB1_019.bam
  CG_10_LIB2_067.bam  QP_06_LIB2_190.bam  RU_08_LIB2_203.bam
  CG_10_LIB2_069.bam  QP_06_LIB2_191.bam  RU_08_LIB2_204.bam
  CG_10_LIB2_070.bam  QP_06_LIB2_195.bam  RU_08_LIB2_205.bam
  CG_10_LIB2_093.bam  QP_06_LIB2_226.bam  RU_08_LIB2_245.bam
  CG_10_LIB2_094.bam  QP_06_LIB2_228.bam  RU_08_LIB2_249.bam
  PR_09_LIB1_098.bam  RO_05_LIB1_039.bam  SS_02_LIB1_076.bam
  PR_09_LIB1_099.bam  RO_05_LIB1_041.bam  SS_02_LIB1_086.bam
  PR_09_LIB1_100.bam  RO_05_LIB1_044.bam  SS_02_LIB1_124.bam
  PR_09_LIB1_114.bam  RO_05_LIB1_049.bam  SS_02_LIB1_153.bam
  PR_09_LIB1_122.bam  RO_05_LIB1_051.bam  SS_02_LIB1_179.bam
  PR_09_LIB2_096.bam  RO_05_LIB2_206.bam  SS_02_LIB2_071.bam
  PR_09_LIB2_102.bam  RO_05_LIB2_221.bam  SS_02_LIB2_081.bam
  PR_09_LIB2_111.bam  RO_05_LIB2_222.bam  SS_02_LIB2_082.bam
  PR_09_LIB2_112.bam  RO_05_LIB2_223.bam  SS_02_LIB2_085.bam
  PR_09_LIB2_115.bam  RO_05_LIB2_254.bam  SS_02_LIB2_090.bam

  info:
  popmap.tsv  popmap.LIB1.tsv  popmap.LIB2.tsv

  stacks_data:
  gstacks  populations
```

Let's move into the `stacks_data` directory and look at the contents.

```
$ cd stacks_data
$ ls
  gstacks  populations
```

The `gstacks` directory contains a pre-existing catalog generated from the alignment data of the 60 *Manacus* samples.

```
$ ls gstacks/
  catalog.calls     catalog.fa.gz  gstacks.log.distribs
  catalog.chrs.tsv  gstacks.log
```

While the `populations` directory contains exports of this same catalog after applying some basic filters (as described in the directory name).

```
$ ls populations/populations.p3.r80.mac3/
  populations.haplotypes.tsv  populations.log.distribs
  populations.haps.vcf        populations.snps.vcf
  populations.hapstats.tsv    populations.sumstats.tsv
  populations.log             populations.sumstats_summary.tsv
```

Inspect input data

Verify the alignments

The `alignments` directory contains the aligned reads (in `bam` format) from 60 *Manacus* individuals (as described by Long et al. 2023). The data for these 60 samples was run through `process_radtags` and aligned to the *Manacus vitellinus* RefSeq assembly (NCBI accession GCF_001715985.3).

For example:

```
$ bwa mem manVit.db CG_10_LIB1_104.1.fq.gz CG_10_LIB1_104.2.fq.gz | \
    samtools view -b -h | \
```

```
    samtools sort -o CG_10_LIB1_104.bam
```

NOTE: For the sake of time, the `bam` files provided here have been filtered to include data from only one chromosome-level scaffold.

Remember, `bam`s store the alignments in binary format (see `bam` documentation). In order to view the alignments as text, we have to run the `samtools view` command:

```
$ samtools view CG_10_LIB1_104.bam | less
```

For more information on how to process and view alignments in the `bam` format, see the `samtools view` documentation.

The population map files

The population map (i.e., `popmap`) files stored in the `info` directory describe the population assignments of each *Manacus* individual.

```
$ cat popmap.ALL.tsv
  SS_02_LIB1_076  020SS
  SS_02_LIB1_086  020SS
  SS_02_LIB2_081  020SS
  SS_02_LIB2_082  020SS
  RO_05_LIB1_039  050RO
  RO_05_LIB1_041  050RO
  RO_05_LIB2_222  050RO
  RO_05_LIB2_223  050RO
  CG_10_LIB1_110  100CG
  CG_10_LIB1_133  100CG
  CG_10_LIB2_093  100CG
  CG_10_LIB2_094  100CG
  ...
```

For example, the individual `SS_02_LIB1_076` belongs to the `020SS` population (shortened ID for transect population #2, San San Drury in Long et al. 2023). Indidivual `RO_05_LIB2_222` belongs to `050RO` (population #5, Rio Oeste). Notice that in addition to an unique numerical identifier (e.g., `076` or `222`), the full name of the samples also describes their population of origin, as well as their library, library 1 (`LIB1`) or 2 (`LIB2`) (more on that later).

Creating a catalog with `gstacks`

In the main `stacks-radseq`, let's create a directory to store a new *Stacks* catalog. This will be the output directory for `gstacks`.

```
$ mkdir stacks-catalog
```

Let's then move into this directory

```
$ cd stacks-catalog
```

Once there, we want to run the *Stacks* `gstacks` program to create a new catalog of RADseq loci and variant sites generated from the aligned reads of our 60 *Manacus* samples, as specified with the popmap file. Since our data is aligned to a genome, we will be running the software in reference mode (by providing the path to the `bam` files). Since these samples were prepared in a single-digest RADseq library and sequenced used paired-end reads, we all able to also remove PCR duplicates when processing our new RAD loci.

Here's an example of the `gstacks` command:

```
$ gstacks \
    -I ~/stacks-radseq/arc-radseq-data.congen23/alignments/ \
    -O . \
    -M ~/stacks-radseq/arc-radseq-data.congen23/info/popmap.tsv \
```

```
    --threads 4 \
    --rm-pcr-duplicates
```

Inspecting the catalog

Once `gstacks` finishes running, we can inspect the coverage and PCR duplicate summary statistics from the `gstacks.log` file.

NOTE: Due to limited time, we are running `gstacks` on a set of subsampled alignments. The examples below show results for a run containing data for the whole genome. A copy of this larger catalog can be found in `~/stacks-radseq/arc-radseq-data.congen23/stacks_data/gstacks`.

```
$ cat gstacks.log | grep -B 3 -A 5 '^Genotyped'
  Removed 9864375 unpaired (forward) reads (6.3%); kept 145744842 read pairs in 130231 loci.
  Removed 102123643 read pairs whose insert length had already been seen in the
    same sample as putative PCR duplicates (70.1%); kept 43621199 read pairs.

  Genotyped 130231 loci:
    effective per-sample coverage: mean=9.0x, stdev=6.9x, min=1.4x, max=24.3x
    mean number of sites per locus: 655.6
    a consistent phasing was found for 1163253 of out 1204439 (96.6%) diploid loci needing phasing

  gstacks is done.
```

This catalog contains 130 thousand assembled loci (average length 656 bp). The average non-redundant coverage of 9x after removing 70% PCR duplicates. 97% of all loci were phased into haplotypes.

Per-individual catalog statistics

The values above are a summary of the whole catalog. Looking at diagnostic distributions at an individual level might provide additional information regarding the properties of the catalog.

Alignment statistics

```
$ stacks-dist-extract gstacks.log.distribs bam_stats_per_sample
  sample          records   primary_kept  kept_frac  primary_kept_read2  primary_disc_mapq  primary_disc_sclip
unmapped  secondary  supplementary
  CG_10_LIB1_104  3233278   2849075       0.881      1416707             203100             74916              93885
0          12302
  CG_10_LIB1_107  2124239   1887469       0.889      937391              117824             48689              60770
0          9487
  CG_10_LIB1_110  3338987   2937839       0.880      1460636             200089             85624              99854
0          15581
  CG_10_LIB1_133  17473133  15522677      0.888      7719372             1006107            410941
463345    0          70063
  CG_10_LIB1_135  7550756   6723388       0.890      3342453             427202             174647
196429    0          29090
  CG_10_LIB2_067  6104763   5200722       0.852      2528449             435756             253311
173531    0          41443
  CG_10_LIB2_069  7023636   6030505       0.859      2928792             479803             277672
195674    0          39982
  CG_10_LIB2_070  5346987   4613356       0.863      2241003             352197             208934
142061    0          30439
  CG_10_LIB2_093  6015970   5125528       0.852      2489385             422268             252187
176579    0          39408
  ...
```

Reformatted here for readability:

sample	records	primary_kept	kept_frac	primary_kept_read2	primary_disc_mapq	primary_disc_sclip	unmapped
CG_10_LIB1_104	3,233,278	2,849,075	0.881	1,416,707	203,100	74,916	93,885

sample	records	primary_kept	kept_frac	primary_kept_read2	primary_disc_mapq	primary_disc_sclip	unmapped
CG_10_LIB1_107	2,124,239	1,887,469	0.889	937,391	117,824	48,689	60,770
CG_10_LIB1_110	3,338,987	2,937,839	0.880	1,460,636	200,089	85,624	99,854
CG_10_LIB1_133	17,473,133	15,522,677	0.888	7,719,372	1,006,107	410,941	463,345
CG_10_LIB1_135	7,550,756	6,723,388	0.890	3,342,453	427,202	174,647	196,429
CG_10_LIB2_067	6,104,763	5,200,722	0.852	2,528,449	435,756	253,311	173,531
CG_10_LIB2_069	7,023,636	6,030,505	0.859	2,928,792	479,803	277,672	195,674
CG_10_LIB2_070	5,346,987	4,613,356	0.863	2,241,003	352,197	208,934	142,061
CG_10_LIB2_093	6,015,970	5,125,528	0.852	2,489,385	422,268	252,187	176,579

Per-sample non-redundant coverage and PCR duplicates

```
$ stacks-dist-extract gstacks.log.distribs effective_coverages_per_sample | grep -v '^#'
  sample          n_loci  n_used_fw_reads  mean_cov  mean_cov_ns  n_unpaired_reads  n_pcr_dupl_pairs  pcr_dupl_rate
  CG_10_LIB1_104  40247   75435            1.874     1.924        53289             1303644           0.945
  CG_10_LIB1_107  31601   49348            1.562     1.591        37140             863590            0.946
  CG_10_LIB1_110  41963   77687            1.851     1.902        57391             1342125           0.945
  CG_10_LIB1_133  78905   401437           5.088     5.465        293328            7108540           0.947
  CG_10_LIB1_135  59591   176098           2.955     3.103        127093            3077744           0.946
  CG_10_LIB2_067  94340   1385738          14.689    16.182       238750            1047785           0.431
  CG_10_LIB2_069  93509   1597878          17.088    18.665       275124            1228711           0.435
  CG_10_LIB2_070  94046   1245928          13.248    14.561       208150            918275            0.424
  CG_10_LIB2_093  94236   1365631          14.492    15.916       238788            1031724           0.430
  ...
```

Reformatted here for readability:

sample	n_loci	n_used_fw_reads	mean_cov	mean_cov_ns	n_unpaired_reads	n_pcr_dupl_pairs	pcr_dupl_rate
CG_10_LIB1_104	40,247	75,435	1.874	1.924	53,289	1,303,644	0.945
CG_10_LIB1_107	31,601	49,348	1.562	1.591	37,140	863,590	0.946
CG_10_LIB1_110	41,963	77,687	1.851	1.902	57,391	1,342,125	0.945
CG_10_LIB1_133	78,905	401,437	5.088	5.465	293,328	7,108,540	0.947
CG_10_LIB1_135	59,591	176,098	2.955	3.103	127,093	3,077,744	0.946
CG_10_LIB2_067	94,340	1,385,738	14.689	16.182	238,750	1,047,785	0.431
CG_10_LIB2_069	93,509	1,597,878	17.088	18.665	275,124	1,228,711	0.435
CG_10_LIB2_070	94,046	1,245,928	13.248	14.561	208,150	918,275	0.424
CG_10_LIB2_093	94,236	1,365,631	14.492	15.916	238,788	1,031,724	0.430

Phasing

```
$ stacks-dist-extract gstacks.log.distribs phasing_rates_per_sample
  sample          n_gts  n_multisnp_hets  n_phased  misphasing_rate  n_phased_2ndpass
  CG_10_LIB1_104  39781  4025             3962      0.016            32
  CG_10_LIB1_107  31283  2270             2245      0.011            21
  CG_10_LIB1_110  41468  4166             4111      0.013            19
  CG_10_LIB1_133  77663  20838            20200     0.031            240
  CG_10_LIB1_135  58785  10114            9901      0.021            78
  CG_10_LIB2_067  91696  38802            37449     0.035            138
  CG_10_LIB2_069  91468  39171            37704     0.037            170
```

```
CG_10_LIB2_070  91450  37588           36469     0.030           143
CG_10_LIB2_093  91671  38126           36867     0.033           129
...
```

Reformatted here for readability:

sample	n_gts	n_multisnp_hets	n_phased	misphasing_rate	n_phased_2ndpass
CG_10_LIB1_104	39,781	4,025	3,962	0.016	32
CG_10_LIB1_107	31,283	2,270	2,245	0.011	21
CG_10_LIB1_110	41,468	4,166	4,111	0.013	19
CG_10_LIB1_133	77,663	20,838	20,200	0.031	240
CG_10_LIB1_135	58,785	10,114	9,901	0.021	78
CG_10_LIB2_067	91,696	38,802	37,449	0.035	138
CG_10_LIB2_069	91,468	39,171	37,704	0.037	170
CG_10_LIB2_070	91,450	37,588	36,469	0.030	143
CG_10_LIB2_093	91,671	38,126	36,867	0.033	129

Filtering the catalog and exporting genotypes

Create general directory

```
$ mkdir ~/stacks-radseq/filter-catalog
$ cd ~/stacks-radseq/filter-catalog
```

General `populations` run

Get r80 loci present in the at least three of the six *Manacus* populations and observe alleles present at least three times (minimum count of 3, i.e., present in at least 2 samples). Run with the popmap containing all 60 samples.

Create the directory of this run. The name of the directory (`populations.p3.r80.mac3`) describes the specific filters applied to the data.

```
$ mkdir populations.p3.r80.mac3
$ cd populations.p3.r80.mac3
```

Run the *Stacks* `populations` module

```
$ populations \
    --in-path ~/stacks-radseq/stacks-catalog \
    --out-path . \
    --popmap ~/stacks-radseq/arc-radseq-data.congen23/info/popmap.tsv \
    --threads 4 \
    --min-populations 3 \
    --min-samples-per-pop 0.80 \
    --min-mac 3
```

Check the outputs of `populations`

Note: for the sake of time, we ran `populations` on the reduced catalog we prepared earlier. We will use a larger run (available in `~/stacks-radseq/arc-radseq-data.congen23/stacks_data/populations/populations.p3.r80.mac3`) to explore the filtering of the catalog.

Go to the large `populations` run:

```
$ cd ~/stacks-radseq/arc-radseq-data.congen23/stacks_data/populations/populations.p3.r80.mac3
```

List the contents of this directory:

```
$ ls populations/populations.p3.r80.mac3/
  populations.haplotypes.tsv  populations.log.distribs
  populations.haps.vcf        populations.snps.vcf
  populations.hapstats.tsv    populations.sumstats.tsv
  populations.log             populations.sumstats_summary.tsv
```

The `sumstats` and `hapstats` files contain the summary statistics assigned per-population for each SNPs and haplotype, respectively. The SNPs and haplotypes are also exported in VCF format.

Inspect the `populations.log` file to obtain the number of loci and variant sites retained after filtering:

```
$ cat populations.log | grep 'Kept'
  Kept 72066 loci, composed of 56612370 sites; 40540446 of those sites were filtered, 155674 variant sites remained.
```

After applying filters, this run kept 72 thousand loci, containing 156 thousand variant sites.

Inspect the `populations.log.distribs` to obtain additional diagnostic distributions and per-sample missing data statistics.

Samples per-locus

```
$ stacks-dist-extract populations.log.distribs samples_per_loc_postfilters | grep -v '^#'
  n_samples   n_loci
  24          1528
  25          1250
  26          1176
  27          631
  28          408
  ...
  56          5404
  57          6720
  58          7612
  59          8920
  60          8308
```

Missing loci per-sample

```
$ stacks-dist-extract populations.log.distribs loci_per_sample | grep -v '^#'
  sample          n_loci  present_loci  missing_loci  frequency_missing
  SS_02_LIB1_076  72066   43799         28267         0.3922
  SS_02_LIB1_086  72066   55188         16878         0.2342
  SS_02_LIB1_124  72066   56081         15985         0.2218
  SS_02_LIB1_153  72066   63975         8091          0.1123
  SS_02_LIB1_179  72066   57717         14349         0.1991
  SS_02_LIB2_071  72066   64933         7133          0.0990
  SS_02_LIB2_081  72066   64924         7142          0.0991
  SS_02_LIB2_082  72066   64904         7162          0.0994
  SS_02_LIB2_085  72066   64935         7131          0.0990
  ...
```

Reformatted here for readability:

sample	n_loci	present_loci	missing_loci	frequency_missing
SS_02_LIB1_076	72,066	43,799	28,267	0.3922
SS_02_LIB1_086	72,066	55,188	16,878	0.2342
SS_02_LIB1_124	72,066	56,081	15,985	0.2218
SS_02_LIB1_153	72,066	63,975	8,091	0.1123
SS_02_LIB1_179	72,066	57,717	14,349	0.1991
SS_02_LIB2_071	72,066	64,933	7,133	0.0990
SS_02_LIB2_081	72,066	64,924	7,142	0.0991
SS_02_LIB2_082	72,066	64,904	7,162	0.0994
SS_02_LIB2_085	72,066	64,935	7,131	0.0990

Missing variant sites per-sample

```
$ stacks-dist-extract populations.log.distribs variant_sites_per_sample | grep -v '^#'
  sample          n_sites  present_sites  missing_sites  frequency_missing
  SS_02_LIB1_076  102800   30421          72379          0.7041
  SS_02_LIB1_086  102800   82220          20580          0.2002
  SS_02_LIB1_124  102800   84021          18779          0.1827
  SS_02_LIB1_153  102800   79498          23302          0.2267
  SS_02_LIB1_179  102800   67139          35661          0.3469
  SS_02_LIB2_071  102800   102249         551            0.0054
  SS_02_LIB2_081  102800   101728         1072           0.0104
  SS_02_LIB2_082  102800   102231         569            0.0055
  SS_02_LIB2_085  102800   102380         420            0.0041
  ...
```

Reformatted here for readability

sample	n_sites	present_sites	missing_sites	frequency_missing
SS_02_LIB1_076	102,800	30,421	72,379	0.7041
SS_02_LIB1_086	102,800	82,220	20,580	0.2002
SS_02_LIB1_124	102,800	84,021	18,779	0.1827
SS_02_LIB1_153	102,800	79,498	23,302	0.2267
SS_02_LIB1_179	102,800	67,139	35,661	0.3469
SS_02_LIB2_071	102,800	102,249	551	0.0054
SS_02_LIB2_081	102,800	101,728	1,072	0.0104
SS_02_LIB2_082	102,800	102,231	569	0.0055
SS_02_LIB2_085	102,800	102,380	420	0.0041

Authors

Angel G. Rivera-Colon[1,2]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA

[2]Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, IL, USA
arcolon14@gmail.com | ariverac@uoregon.edu

Kira M. Long
Program in Ecology, Evolution, and Conservation Biology,

University of Illinois at Urbana-Champaign, Urbana, IL, USA
kiralong778@gmail.com

Julian M. Catchen
Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, IL, USA
jcatchen@illinois.edu