RAD Population Genomics Programs

Paul Hohenlohe (hohenlohe@uidaho.edu) 6/2014

I. Overview

These programs are designed to conduct population genomic analysis on RAD sequencing data. They were designed for cases with a reference genome (including a reference set of RAD contigs), although many of the analyses are possible without alignment to a reference.

They will soon be a seamless part of Stacks (http://creskolab.uoregon.edu/stacks). In the meantime, they are presented here as command-line programs under the GNU General Public License (http://www.gnu.org/licenses). They are presented without any warranty; without even the implied warranty of merchantability or fitness for a particular purpose.

All the programs described below are written in C and can be compiled with gcc in a unix terminal. They all require the header file RADfunx.h, and they may need the option -lm in gcc. For example:

```
gcc collate.c -lm -o collate
```

Command line usage for each program can be produced by typing

```
program -h
```

II. File formats

a. Genotype call file

The standard genotype call file format is produced by genotype and taken as input by the other programs. The first line of this file has the total number of nucleotide sites and then the number of individuals in the file, separated by a tab or spaces. Each subsequent line is a single nucleotide position, with the following tab-separated fields: any number of ID fields (strings); chromosome/linkage group name (a string); position (an integer); diploid genotype for each individual (integer from 0-10; see below). genotype prints one ID field by default (position within each RAD tag), and the other programs below expect one ID field by default; Stacks may output one or more ID fields. The positions must be ordered such that linkage groups are together, positions are in increasing order within each linkage group, and the linkage groups are ordered to match the LGnames.txt file (required by the programs genotype or collate). Example:

```
24035   5
tag6    groupI  426     5       5       5       0       6
tag6    groupI  427     8       8       0       0       8
tag6    groupI  428     10      9       10      8       8
…….
```

b. Genotype codes

Unordered diploid genotypes are coded as 0-10. 0 means no genotype was called at that position. There are no hemizygous or dominant marker genotypes (e.g. T_). Genotype codes are:

		Nucleotide 1		
Nucleotide 2	**A**	**C**	**G**	**T**
A	1	2	3	4
C	2	5	6	7
G	3	6	8	9
T	4	7	9	10

c. Chromosome, LG, or contig names

The programs genotype and collate require a text file (default LGnames.txt) giving the names of all chromosomes, linkage groups, or RAD contigs expected in the data. The first line is an integer giving the number of linkage groups, followed by the names. genotype will output genotype calls in this order, and input files for collate must follow this order exactly. An error will result if a chromosome or contig name appears in the data file (IId below) and does not appear in this name file.

d. Sequence data files

genotype requires sequence data files to be in a standard tab-separated format, reflecting a subset of the output from bowtie or other alignment software. These must be one file per individual. Each line contains a sequence read: the first column is the linkage group or contig name (must match one of the names in the LGnames.txt file), the second column is the alignment position of the left end of the read, third column is the sequence read, and fourth is the ascii-coded quality scores. Quality scores are optional (see below). Barcodes should already be trimmed from the reads and quality scores, but restriction enzyme sites may or may not be. These reads do not have to be sorted in any way, although it may reduce computational time if they are (see -sorted option in genotype below).

Either single-end or paired-end RAD sequencing data can be analyzed by genotype. The default expects single-end sequencing; the option -paired must be specified in the genotype program is paired-end data is used, and note that the forward and reverse reads must be of different length (this should occur when barcodes are trimmed off the forward reads).

e. Pathnames to datafiles

genotype requires a text file giving the pathnames to all sequence datafiles and other information. The first line of this file is the number of datafiles to be read (i.e. number of individuals or samples). The second line is the read length of forward reads (or all reads for single-end sequencing). If paired-end data are used, the third line is the read length for the reverse reads; the lengths must be different for forward and reverse reads in order for the program to correctly parse paired-end RAD data. The next line is the quality score threshold for each single nucleotide to be counted (0 is no filter; higher numbers filter out single nucleotide reads). Each subsequent line gives the pathname to

the sequence datafile and the file name for the output of genotype calls. Example for single-end data:

```
48
94
10
datafiles/lane4_100218/lane4_100218_CGATAC.map    calls1.txt
datafiles/lane4_100218/lane4_100218_CGGCGT.map    calls2.txt
…
```

Example for paired-end data:

```
48
94
100
10
datafiles/lane4_100218/lane4_100218_CGATAC.map    calls1.txt
datafiles/lane4_100218/lane4_100218_CGGCGT.map    calls2.txt
…
```

f. Collate pathnames file

collate requires a text file giving the pathnames to all genotype call files to be collated. The first line should give the number of files, followed by the pathnames to each file. Each call file should be in the standard format (IIa), can contain one or more individuals, but must contain linkage groups in the order given by the LGnames.txt file (IIc). Example:

```
3
folder/file1.txt
folder/file2.txt
folder/file3.txt
```

g. Single population ID file

SNPstats0 and statistics0 require a text file identifying the individuals to be included in the population. The first line of this file gives the number of individuals in the population (tab-separated or separate lines), followed by their column number in the call file, starting from 1 for the first individual after the nucleotide position column. Not all individuals in the call file need to be included, and they do not have to be in order. Comments are allowed only at the end of this file. Example:

```
6
1       2       5       9       4       13
```

h. Multiple population ID file

SNPstats1 and statistics1 require a text file identifying the individuals to be included in each population. This file contains (either tab-separated or by line) the number of populations, number of individuals in each population, and column number in the call file for the individuals in each population. Individuals do not have to be in order in the call file, and not all individuals in the call file need to be included. Comments are allowed only at the end of this file. Example:

3
6
8
5
7 8 9 10 21 22
1 2 3 4 5 6 11 12
13 14 15 16 17

III. Programs

genotype

This program reads in a set of single-end RAD sequencing data aligned to a reference genome or reference set of RAD contigs and assigns diploid genotypes to each nucleotide position. Because it was designed for RAD data, it expects forward reads to contain the remaining recognition sequence from a restriction enzyme and tracks some statistics (such as depth of coverage) at the tag level, rather than at the nucleotide level. It outputs a file of genotype calls for each individual in the standard format (IIa above), plus a file with genotype calls by RAD tag for all tags that pass the filters below. Usage is:

genotype filenames.txt [-n LGnamefile.txt] [-stats] [-counts] [-re enzyme {1 for Sbf1, 2 for EcoR1, 3 for PstI}] [-noQ] [-alpha 0.01] [-majorityrule] [-epslower 0.001] [-epsupper 0.1] [-mincalls 0] [-lnL] [-mindepth 0] [-maxdepth 0] [-sorted] [-phred 33] [-paired 0] [-max_tags 100000]

Options

- filenames.txt: See IIe above.
- -n: See IIc file above.
- -stats: Invoking this option will produce a file for each individual giving either AIC weights (default) or lnL values (if -lnL option is used) for each of the ten possible genotypes at each nucleotide position. This option adds computational time and produces **large** files. These data are used only for downstream analyses not included in this group of programs, such as a Hidden Markov Model of genotyping.
- -counts: This option adds a column for read counts per tag to the tag_ files.
- -re: Option for specifying the restriction enzyme for which to filter tags. The default is Sbf1; specifying -re 0 will not filter at all.
- -noQ: Invoke this option if input sequence files do not contain quality scores; see IId above.
- -alpha: Specify the significance level for the likelihood ratio test at each position. Default is 0.01.
- -majorityrule: If there is a single most likely genotype, assign it regardless of the likelihood ratio statistic. May be useful for carrying all available data through an analysis, but not recommended to believe the resulting genotypes.
- -epslower: Lower bound on the uniform prior distribution of epsilon. Must be between 0.0 and 1.0; default is 0.001.

- -epsupper: Upper bound on the uniform prior distribution of epsilon. Must be between 0.0 and 1.0 and greater than epslower; default is 0.1.
- -mincalls: Report only genotype calls for RAD tags with this number or greater of genotype calls within the tag; default is 0 (no filter).
- -lnL: See stats above
- -mindepth: Report only genotypes for RAD tags with this sequence read depth or higher; default is 0 (no filter).
- -maxdepth: Report only genotypes for RAD tags with this sequence read depth or lower; default is no filter (specified as 0).
- -sorted: Invoke this option if reads that aligned to the same genomic location are grouped together in the sequence data input files. This saves computational time without changing any of the analysis.
- -phred: Specify the phred quality score coding with an integer (e.g. 33 or 64). Default is 33.
- -paired: Specify this option if paired-end data are to be used. This will cause the program to expect different read lengths for forward and reverse reads in the header file (see IIe above). The integer specifies the maximum size in bp of DNA fragments used in paired-end sequencing; i.e. the distance from the restriction site to the far end of reverse reads. This is used in memory allocation and determines how many nucleotide positions will be reported in the genotype call file. If paired reads fall outside this range they will be discarded.
- -max_tags: This controls memory allocation for the total number of RAD tags expected. Default is 100,000. Time and memory can be saved by reducing this number, but the program will quit if it's actually exceeded. Best is to count the maximum actual number of positions across a genome at which forward reads have aligned and use this value.

collate

The collate program reads in two or more files of genotype calls and produces a single file with all individuals and nucleotide positions in order. If sites are present in some but not other individuals or input files, 0's are assigned for all individuals missing those data. Usage is:

```
collate datafiles.txt [-o outfile.txt] [-n LGnames.txt] [-min 0] [-IDfields 1]
```

Options

- datafiles.txt: See IIf above.
- -o: Specify the name for the output genotype call file. Default is outfile.txt.
- -n: See IIc above.
- -min: Minimum sample size of individuals for a site to be printed to the output file. Default is 0 (no filter).
- -IDfields: Specify the number of columns expected to the left of the linkage group name on each line of the input files. This number must be the same across all files. Default is 1. The ID fields will be output to the outfile in the same format.

At each nucleotide site, they will match those in the last file in which that site occurred (e.g. in the example under IIf, they will come from file3.txt unless the position did not appear in file3.txt).

SNPstats0

This program returns population genetic statistics for each nucleotide site in a single population or group. Usage is:

SNPstats0 callfile popfile [-o outfile] [-cov coverage.txt] [-all] [-min 0] [-maxpi -1] [-minpi -1] [-maxalleles -1][-minMAF -1] [-maxMAF -1] [-minhet -1] [-maxhet -1] [-minFis -1] [-maxFis -1] [-IDfields 1] [-chrom all] [-filter] [-012] [-PCA]

Options

- callfile: See IIa above. This is required, and must be the first argument after the program name.
- popfile: See IIg above. This is required and must be the second argument after the program name.
- -o: Name of output file; default is "SNPstats.txt".
- -cov: if a collated file of coverage is available (see -cov option in ContigRefGenotype above), path can be specified here and SNPstats0 will report coverage information. This collated coverage file **must** contain the exact same set of individuals and nucleotide positions in the same order as the genotype call file.
- -all: If this option is specified, statistics will be calculated for all sites. Default is only to include SNPs.
- -min: Exclude sites with fewer than this number of individuals genotyped; default is 0.
- -maxpi: Exclude sites with π greater than this value; default is no filter (-1).
- -minpi: Exclude sites with π less than this value; default is no filter (-1).
- -maxalleles: Exclude sites with more than this number of alleles; default is no filter (-1).
- -minMAF: Exclude sites with less than this minor allele frequency; default is no filter (-1).
- -maxMAF: Exclude sites with greater than this minor allele frequency; default is no filter (-1).
- -minhet: Exclude sites with less than this observed heterozygosity; default is no filter (-1).
- -maxhet: Exclude sites with greater than this observed heterozygosity; default is no filter (-1).
- -minFis: Exclude site with less than this F_{IS} value; default is no filter (-1).
- -maxFis: Exclude site with greater than this F_{IS} value; default is no filter (-1).
- -IDfields: Number of ID fields appearing to the left of the linkage group name in the call file. These will appear in the outfile. Default is 1.
- -chrom: report statistics for only one chromosome (name must exactly match that in the callfile and the LGnames file. Default is all (no filter).

- -filter: Invoking this option changes the output to a call file (format as in IIa) for only those sites that pass all filters specified in the above options; values for the statistics are not printed.
- -012: If this option is invoked in addition to –filter, genotypes are output for only biallelic SNPs in the format 0 (homozygous for dominant allele), 1 (heterozygous), 2 (homozygous for minor allele), or N (uncalled).
- -PCA: If this option is invoked in addition to –filter, genotypes are output as normalized values following Price et al 2006 (Nat Genet 38: 904).

SNPstats1

This program produces population genetic statistics per SNP for 1 level of population subdivision. Usage is

SNPstats1 callfile popfile [-o outfile] [-cov coverage.txt] [-all] [-min -1] [-IDfields 1] [-maxalleles -1] [-minMAF -1] [-maxMAF -1] [-minhet -1] [-maxhet -1] [-minpi -1] [-maxpi -1] [-minFst -1] [-maxFst -1] [-minWeirFst -1] [-maxWeirFst -1] [-minGp -1] [-maxGp -1] [-filter] [-PCA]

Options

- callfile: See IIa above. This is required, and must be the first argument after the program name.
- popfile: See IIh above. This is required and must be the second argument after the program name.
- -o: Name of the output file; default is "SNPstats.txt".
- -cov: if a collated file of coverage is available (see -cov option in ContigRefGenotype above), path can be specified here and SNPstats1 will report coverage information. This collated coverage file **must** contain the exact same set of individuals and nucleotide positions in the same order as the genotype call file.
- -all: If this option is specified, statistics will be calculated for all sites. Default is only to include SNPs.
- -min: Exclude sites with fewer than this number of individuals genotyped in any population. If set to 0, requires all individuals to be genotyped in all populations. Default is no filter (-1).
- -IDfields: Number of ID fields appearing to the left of the linkage group name in the call file. These will appear in the outfile. Default is 1.
- -filter: Invoking this option changes the output to a call file (format as in IIa) for only those sites that pass all filters specified in the above options; values for the statistics are not output.
- All further options as described for SNPstats0 above. Thresholds are applied across all populations together.

statistics0

This program conducts a kernel-smoothing sliding window average of population genetic statistics along the genome within a single population. Output files are _allLGs.txt, the

sliding window statistics, and _summary.txt, a summary of genome-wide averages. Usage is

statistics0 callfile popfile [-o outfile_prefix] [-w 100000] [-kernel 150000] [-boot 0] [-min -1] [-gboot 10000] [-het] [-Fis] [-IDfields 1]

Options

- callfile: See IIa above. This is required, and must be the first argument after the program name.
- popfile: See IIg above. This is required and must be the second argument after the program name. The first line of this file gives the number of individuals in the population (tab-separated or separate lines), followed by their column number in the call file, starting from 1. Sample:

```
6
1	2	5	9	10	13
```

- -o: Prefix for multiple output files to be produced.
- -w: Specify the number of nucleotide positions between window averages. Must be an integer. Default is 100,000 (=100 kb).
- -kernel: "Standard deviation" of the Gaussian kernel smoothing function. Must be an integer. Default is 150,000.
- -boot: Number of bootstrap replicates for significance of window averages. Must be an integer. Default is 0. Ramping this up is recommended to gauge computational load.
- -min: Ignore sites with fewer than this number of individuals genotyped. Default is no filter (-1); setting this equal to 0 requires all individuals to be genotyped at each site.
- -gboot: Number of replicates for genome-wide bootstrapping to calculate percentiles on genome-wide average. Must be an integer. Default is 10000.
- -het: Conduct bootstrapping significance on observed heterozygosity. Default is on π.
- -Fis: Conduct bootstrapping significance on F_{IS}. Default is on π.
- -IDfields: Number of ID fields appearing to the left of the linkage group name in the call file. These will be ignored. Default is 1.

statistics1

This program conducts a kernel-smoothing sliding window average of population genetic statistics along the genome with a single level of population subdivision. Usage is

statistics1 callfile popfile [-o outfile_prefix] [-w 10000] [-kernel 150000] [-boot 0] [-min -1] [-gboot 10000] [-Weir] [-IDfields 1]

Options

- callfile: See IIa above. This is required, and must be the first argument after the program name.
- popfile: See IIh above. This is required and must be the second argument after the program name.
- -o: Prefix for multiple output files to be produced.
- -w: Number of nucleotide positions between window averages. Must be an integer. Default is 100,000.
- -kernel: "Standard deviation" of the Gaussian kernel smoothing function. Must be an integer. Default is 150,000.
- -boot: Number of bootstrap replicates for significance of window averages. Must be an integer. Default is 0.
- -min: Ignore sites with fewer than this number of individuals genotyped. Default is no filter (-1); setting this to 0 requires all individuals to be genotyped.
- -gboot: Number of replicates for genome-wide bootstrapping to calculate percentiles on genome-wide average. Must be an integer. Default is 10,000.
- -Weir: Conduct bootstrapping significance on Weir & Cockerham's F_{ST}. Default is on nonparametric F_{ST}.
- -IDfields: number of ID fields appearing to the left of the linkage group name in the call file. These will be ignored. Default is 1.

IV. Output

Many of the fields are common to the output across programs. In the case of multiple populations, many of the statistics below are given for all populations together (shown by _0) and for within each population (shown by _1, _2, etc., in the order of the popfile).

- n (in SNPstats programs): number of individuals genotyped.
- n (in statistics programs): number of nucleotide sites within 3 kernel SD of the window center.
- SNPs: number of SNPs within 3 kernel SD of the window center.
- pi: SNP nucleotide diversity = expected heterozygosity.
- het: observed heterozygosity (proportion of individuals heterozygous).
- Fis: F_{IS} within a single population.
- alleles: number of alleles observed at a site.
- MAF: minor allele frequency, calculated as $1 - p$ where p is the frequency of the most common allele.
- genos: diploid genotype codes at a site.
- PA: private allele = 1 if an allele is found only in this population, 0 otherwise.
- FST: F_{ST}.
- WeirFST: Weir & Cockerham's F_{ST}.
- chi_allele: Chi-square statistic for allele frequencies.
- G_allele: G statistic for allele frequencies.
- chi_geno: Chi-square statistic for genotype frequencies.
- G_geno: G statistic for genotype frequencies.
- p-val: p-value for preceeding Chi-square or G test.
- pi-boot or FST-boot: proportion of bootstrap replicates below the observed value.

- TajD: Tajima's *D* within one kernel SD of the window center.