4P v 1.00

Parallel Processing of Polymorphism Panels

Benazzo Andrea[1], Panziera Alex[1,2] and Bertorelle Giorgio[1]

[1]Department of Life Sciences and Biotechnology, University of Ferrara, via L. Borsari, 46, 44100 Ferrara, Italy

[2]Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, via E. Mach 1, 38010 S. Michele all'Adige, Italy.

August 2014

Email: andrea.benazzo@unife.it

The program is available at https://github.com/anbena/4p

1. OVERVIEW: WHAT 4P CAN DO

4P (Parallel Processing of Polymorphism Panels) is a software for computing population genetics statistics from large SNPs dataset.

The input data types handled by the program include ped/map, arp and vcf files.

The software allows the user to calculate the allele frequencies, the mean and locus by locus heterozygosities (expected and observed), and several measures of genetic distance between populations (Nei's Gst (1973), Nei's Gst (1983), Hedrick's G'st (2005), Jost's D and Weir&Cockerham's Fst (1984)). It also allows the computation of the single and joint allele frequency spectrum (folded and unfolded, using different dimensions) and the matrix of similarity and dissimilarity indices (based on allele sharing) between individuals.

4P is written in the C programming language, using the OpenMP library to distribute computation among available CPU cores. This feature allows the program to make full use of the computational power of multi-core systems, allowing a significant reduction in computational times. 4P, being written in the fully compiled language C, is able to handle datasets larger than those used in environments like R or with scripting languages (as Perl or Python).

2. INSTALLING 4P

4P is distributed at https//github.com/anbena/4p in the form of pre-compiled binaries, meaning that no installation is required. Windows users, however, needs to download and install the MinGW (http://www.mingw.org/) package before running 4P. This software provides essential windows libraries that are not included in 4P. Linux and Mac users can run immediately the software after downloading it .

The 4P package includes the program files for Windows, MacOS, and Linux, this manual in pdf, and some examples.

3. INPUT FILES

3.1 Genetic Data

Genetic data should be provided using one of the following formats:

ped/map

This is the standard Plink format, where the ped file contains information about the genotype of a sample, while the map file stores information about the locus. Further information about this format could be obtained at http://pngu.mgh.harvard.edu/~purcell/plink/. Note: the actual release of 4P does not support the binary Plink input file format (.bed), but you can easily convert it with the "–recode" flag in Plink.

arp

This is the Arlequin format, but the current implementation of 4P reads only the arp format produced by simulation using fastsimcoal (cmpg.unibe.ch/software/fastsimcoal/). In this version of the arp file, individual genotypes are specified as combinations of ancestral (0) and derived (1) alleles. If real data are available in the standard Arlequin format (with the ATCG notation), we suggest to convert to a ped or vcf file (using for example PGDSpider, cmpg.unibe.ch/software/PGDSpider/) before running 4P .

vcf

This is the Variant Call Format (version 4.1), which is used by different genomic databases like the 1000 Genome Project. Further information about this format can be obtained at 1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41.

3.2 Populations

Only when the vcf format is used, an additional input file should be provided to specify the population of origin of each individual. This information is already included in ped/map and arp files. The format of this additional file is as follows:

<Individual's ID>   <Population>

<u>Example</u>

```
IND1    Pop1
IND2    Pop1
IND3    Pop2
…       …
```

The supported delimiters are tab or space.

3.3 Ancestral/derived alleles

In order to compute the unfolded allele frequency spectrum using the Plink and the VCF formats, the user needs to provide an additional input file to specify the ancestral allele. The format of this additional file is as follows:

<Chromosome>   <SNP position>   <Ancestral allele>

Example

```
1       10000  A
1       10050  T
2       900    T
2       950    C
…       …      …
```

The supported delimiters are tab or space. The chromosomes, or the scaffolds in a de novo assembly, should be coded using increasing integers.

3.4 Parameters

The user needs to specify in the parameters file (mandatory name: *sumstat.par*) which summary statistics 4P should compute. Each line (in any order) of the file specifies the settings for each statistic (or group of statistics), using a keyword followed by 2 or 3 digits separated by #. The sumstat.par file is compulsory and it must be in the same folder of 4P. Lines can be omitted, or commented out (using // at the beginning of the line), but if a line is included it should have values for all the digits required for the corresponding keyword.

Keyword: BASEFREQ
Base frequencies. Two digits should follow this keyword. The first digit is just for confirming that base frequencies should be computed (1=compute; 0=not compute). The second digit specifies if locus by locus base frequencies (1) or mean and standard deviation (0) should be printed in the output file (the latter corresponding to nucleotide proportions data). If the first digit is set to 0, the second digit is of course not relevant.
Example: BASEFREQ#1#1

Keyword : HET

Heterozygosities. Three digits should follow this keyword. The first digit indicates if the OBSERVED heterozygosity should be calculated (1) or not (0). The second digit indicates if the EXPECTED heterozygosity should be calculated (1) or not (0). The third digit specifies if locus by locus base frequencies (1) or mean and standard deviation (0) should be printed in the output file. If the first two digits are set to 0, the third digit is of course not relevant. In the expected heterozygosity computation, we used the sample size correction proposed by Nei (1987).

Example: HET#0#1#1

Keyword: AFS

Allele frequency spectrum. Three digits should follow this keyword. The first digit is just for confirming that the allele frequency spectrum should be computed (1=compute; 0=not compute). The second digit specifies the number of dimensions of the allele frequency spectrum (from 1 to 3). The third digit indicates if the allele frequency spectrum should be unfolded (0) or folded (1). The unfolded spectrum can be computed only if the ancestral allele is known. If the first digit is set to 0, the second and third digits is of course not relevant.

Example: AFS#1#1#1

Keyword: DIST

Pairwise genetic distances among populations. Two digits should follow this keyword. The first digit is just for confirming that genetic distances should be computed (1=compute; 0=not compute). The second digit is used to specify if the locus by locus (1) or the mean (0) genetic distances across loci will be printed in the output file. In this latter case (mean as output), the computed distances are multi-locus estimates of the distance and not simply the mean across loci. The genetic distance indices are Nei's Gst (1973), Nei's Gst (1983), Hedrick's G'st (2005), Jost's D and Weir & Cockerham's Fst (1984). If the first digit is set to 0, the second digit is of course not relevant.

Example: DIST#1#0

Keyword: DALL
Similarity/dissimilarity index among individuals (based on their allele sharing). The index is the proportion of shared/not-shared alleles between pairs of individuals. The proportion is computed using all loci in which there is at least 1 chromosome sequenced for both individuals. Two digits should follow this keyword. The first digit is just for confirming that the matrix should be computed (1=compute; 0=not compute). The second digit indicates that the matrix similarity values (0) or dissimilarity values (1) will be computed. If the first digit is set to 0, the second digit is of course not relevant.
Example: DALL#1#1

4. LAUNCHING 4P

4P has a command line interface, meaning that the user needs to launch the binary file (for example, 4P_64_win) from the console. After opening the console session, launch the program followed by the options (flags) required. Some flags are mandatory and cannot be omitted, others are optional.

Mandatory flags	
-f [filename]	Name of the genotypes input file
-i [number]	Input file type (0: Plink, 1: arp, 2: vcf)
-m [filename]	Map input file name (only for Plink files)
-n [number]	Total number of diploid individuals
-s [number]	Total number of polymorphisms (SNPs)
-p [filename]	Name of the populations input file (only when vcf files are used for the genotypes)

Optional flags	
-a [filename]	Ancestral allele input file name (only for Plink and vcf files)
-t [number]	Number of cores used for computation. If no value is specified, all the cores will be used
-v	Print on the screen the genotypes when they are loaded
-y	Only summary files are produced for each statistic (not locus by locus)
-h	Print the list of available flags

5. OUTPUT FILES

Depending on the options selected by the user in the parameters file, a certain number of output files will be produced. Here we describe them referring to the groups of statistics as they were subdivided in the parameters file. Only a summary file for each selected statistic, which include mean and standard deviation across loci in each population, is produced using the “-y” flag.

5.1 Base frequencies

If base frequencies are computed locus by locus, K+1 files are produced, where K is the number of populations. K files called SNP_BASE_FREQUENCY_[*popname*].txt will contain the results for each population, and SNP_BASE_FREQUENCY_whole.txt will contain the results for the total data set (all populations together). These output files will have this format:

CHR POS RS A C T G MISS MAFALL MAF

where CHR is the chromosome number (or the scaffold number), POS and RS are the position and the ID of the locus, respectively (if available, “-“ if not), A,C,T and G are the nucleotides frequencies at each locus, MISS refers to the missing data frequencies, MAFALL is the allele with the minor frequency and MAF is its frequency.

If only means and standard deviations are computed across loci, or the general “-y” flag is specified, the output is condensed in a single file called SNP_BASE_FREQUENCY_SUMMARY.txt with the following format:

population mean(A) mean(C) mean(T) mean(G) mean(MISS) sd(A)sd(C) sd(T) sd(G) sd(MISS)

where “population” is the population name, A,C,T and G are of course the nucleotides and MISS refers to the missing data.

5.2 Heterozygosities

If herozygosities are computed locus by locus, K+1 files are produced for both the observed and the expected heterozygosity values (total of 2K+2 files if both statistics are computed). The generic name for the single-population statistics is HET_OBS_[*popname*].txt and HET_EXP_[*popname*].txt, and the files HET_OBS_whole.txt and HET_EXP_whole.txt will contain the results for the total data set (all populations together). These output files will have this format:

CHR POS RS H_OBS

or

CHR POS RS H_EXP

where CHR is the chromosome number (or the scaffold number), POS and RS are the position and the ID of the locus, respectively (if available, "-" if not), and H_OBS or H_EXP are the observed and expected heterozygosities at that locus.

If only means and standard deviations are computed across loci, or the general "-y" flag is specified, the output is condensed in two files called HET_OBS_SUMMARY.txt and HET_EXP_SUMMARY.txt. These files will have the following format:

population mean(H_OBS) sd(H_OBS)

or

population mean(H_OBS) sd(H_OBS)

where the population field corresponds to the population name and it is followed by means and standard deviations of observed or expected heterozygosity across loci, respectively.

5.3 Allele frequency spectrum

The allele frequency spectrum, folded or unfolded, is computed in each population (1 dimension) or between each couple or triplets of populations (2 or 3 dimensions). Depending on the number of dimensions selected in the parameters file, i.e., 1, 2, or 3, the number of output files is K, (K^2-K)/2, and $\binom{K}{3}$, respectively.

To ensure compatibility, the output is provided in "dadi" (Gutenkunst et al 2009) format. The generic name of the allele frequency spectrum file computed in a single population is AFS-[typeofafs]_[popname], where "typeofafs" could be F or U in case of folded or unfolded spectrum. When multiple dimension are selected, the popname field is substituted with the names of the populations separated by "_". When the "-y" flag is selected, all the allele frequency spectra are reported in a single file called AFS-U_SUMMARY, or AFS-F_SUMMARY depending on the spectrum type (unfolded or folded). Two output file are produced, one with extension .dadi.txt ("dadi" format) and one with extension .txt (general text file).

The allele frequency spectrum is computed using only loci at which the alleles are present in all individuals (loci with no missing data). This choice is motivated by the fact that is impossible to identify the frequency category in the presence of missing data. Other programs can be used to impute sporadic missing genotypes (e.g., IMPUTE2 (Howie et al 2012); Shapeit (Delaneau et al

2012); Beagle (Browning & Browning 2007)) before running 4P, or individuals with many missing data should be preliminarily excluded from the analysis.

The allele frequency spectrum is reported in a single line in the output file. The within population spectrum (1 dimension) is the set of absolute frequencies of sites where the minor (folded spectrum) or the derived (unfolded spectrum) allele are observed in 0, 1, 2, … chromosomes in the population. The joint spectrum (2 dimension) is the set of absolute frequencies of sites where the minor (folded spectrum) or the derived (unfolded spectrum) allele are observed in 0 chromosomes in the first population and 0 chromosomes in the second population, 1 chromosomes in the first population and 0 chromosomes in the second population, 2 chromosomes in the first population and 0 chromosomes in the second population, etc. , considering all the possible pairwise combinations. The three dimensional spectrum is based on the same kind of counting, but it considers three populations.

For example, the one dimension folded AFS format will look like:

```
#f_Pop10      #f_Pop11      #f_Pop12      #f_Pop13      #f_Pop14      …
X folded
847     31      2       9       6       …
0 0 0 0 0 0 0
```

where the first line includes the frequency categories (#f_Pop13 means, for example, the number of sites where the minor allele (**f**olded spectrum) is observed in the **Pop**ulation **1** in **3** individuals); the second line specifies the number of the frequency categories (X) followed by the type of spectrum; the third line specifies the number of SNPs in each category; the fourth line allows the user to mask (exclude) some categories in the dadi analysis (0 to unmask and 1 to mask).

The first line for the spectrum in 2 or 3 dimensions will have names as #f_[*pop1name*]1_[*pop2name*]0 or #u_[*pop1name*]0_[*pop2name*]0_[*pop3name*]0, respectively.

5.4 Genetic distances

If genetic distances are computed locus by locus, (K^2-K)/2 files will be produced, each corresponding to a pairwise population comparison. The generic name for these files is PAIR_DIST_[*pop1name*]_[*pop2name*].txt, and their format is

```
CHR     POS     RS     GSTNEI73      GSTNEI83      GSTHED05      DJOST   FSTWC84
```

where CHR, POS and RS are the chromosome, the position and the ID of the locus where genetic distances are computed, and the other columns correspond, respectively, to the following distances: Nei's Gst (1973), Nei's Gst (1983), Hedrick's G'st (2005), Jost's D and Weir&Cockerham's Fst (1984)

If only means and standard deviations are computed across loci, or the general “-y” flag is specified, the genetic distances between all populations are summarized in five upper triangular matrix (one for each differentiation statistic) and stored in a single file called PAIR_DIST_SUMMARY.txt.

5.5 Similarity/dissimilarity matrix

One file is produced, IND_SIM_MATRIX.txt or IND_DIS_MATRIX.txt, which includes a triangular matrix of similarities or dissimilarities among all pairs of individuals. The matrix is based on the average fraction of alleles shared (or different) per locus.

This is an example of this individual based comparison (where IDX is the name of X individual):

	ID1	**ID2**	**ID3**	**ID4**	**…**
ID1	**0**				
ID2	**0.5**	**0**			
ID3	**0.22**	**0.29**	**0**		
ID4	**0.14**	**0.8**	**0.74**	**0**	
…					

6. EXAMPLES

6.1 Computing genetic distances

The dataset includes 100 SNPs typed in 1000 individuals from two populations (500 from Pop1 and 500 from Pop2). The data are generated using the *fastsimcoal2* simulator (Excoffier et al., 2013). The input file is example_1000ind_100snp.arp and the analysis will be run using 8 out of 16 available cores. The user is interested in means and standard deviations of different genetic distances.

In this analysis, the sumstat.par file should contain at least one line

```
DIST#1#0
```

the command line to run the analysis is

```
4P –f example_1000ind_100snp.arp –i 1 –n 1000 –s 1000 –t 8
```

and only one output file will be generated with the name DIST_Pop1_Pop2.txt.

6.2 Computing heterozygosities

The dataset includes 15000 SNPs typed in 800 individuals from two populations (350 from Pop1 and 450 from Pop2), converted using Plink in a ped/map format. The input files are 15kSnp.ped (genotype) and 15kSnp.map (information about the loci), and the analysis will be run using all cores available. The user is interested in locus by locus measures of expected and observed heterozygosity.

In this analysis, the sumstat.par file should contain at least one line

```
HET#1#1#1
```

the command line to run the analysis is

```
4P –f 15kSnp.ped –m 15kSnp.map –i 0 –n 800 –s 15000
```

and six output file will be generated with the following names: HET_OBS_Pop1.txt; HET_OBS_Pop2.txt; HET_OBS_whole.txt; HET_EXP_Pop1.txt; HET_EXP_Pop2.txt; HET_OBS_whole.txt

6.3 Computing genetic distances, heterozygosities and the unfolded allele frequency spectrum

The dataset includes 15 millions SNPs typed in 1200 individuals from four populations (Pop1,Pop2,Pop3,Pop4). The data were downloaded from a genomic database and the input file name is 15M_1200ind.vcf. The analysis will be run using 16 out of 24 available cores. The user is interested in genetic distances (locus by locus), heterozygosities (both expected and observed, locus by locus), and the unfolded one-dimensional allele frequency spectra.

Since the vcf file does not contain the information on the population of origin of each individual and on the status (ancestral or derived) of each allele, two additional input files are required in this case (see sections 3.2 and 3.3 of this manual). In the population file (here called pop_info.txt, 1200 lines and 2 columns), each line specifies the individual code and its population of origin. In the ancestral/derive allele file (here called anc_allele.txt, 15 million lines and two columns), each line specifies the SNP code and the ancestral allele.

In this analysis, the sumstat.par file should contain at least three lines

```
DIST#1#1
HET#1#1#1
AFS#1#1#1
```

the command line to run the analysis is

```
4P –f 15M_1200ind.vcf –i 2 –n 1200 –s 15000000 –p pop_info.txt –a anc_allele.txt –t 16
```

and the total number of output files will be 20 (4 for the genetic distances, 10 for the heterozygosities, and 4 for the allelic spectra).

6.4 Integrating 4P in an ABC (Approximate Bayesian Computation) pipeline

4P can be directly used in an ABC pipeline as a mean to compute summary statistics in simulated data sets. For example, 4P can be integrated in scripts developed by the user or already available as tools in the ABCToolbox (Wegmann et al 2010), in order to create the reference table required the by ABC.

Here we provide an example of input file for the ABCsampler (the ABCtoolbox module that creates the reference tables) where the data are generated by the coalescent simulator simcoal2, and the last two lines of the file specify that 4P is used to compute the summary statistics.

```
//a simple inputfile for ABCsampler
samplerType standard
//only uniform priors
estName test_uniform.est
obsName test.obs
outName test_output
nbSims 1000000
writeHeader 1
//settings regarding the simulation program
simulationProgram simcoal2
simInputName test.par
simParam FILENAME#1#1#0 //see simcoal2 manual for details
//program to calculate summary statistics
sumStatProgram 4p_64_linux
sumStatParam -f#SIMINPUTNAME.arp#-i#1#-n#NIND#-s#NSNP#-t#NCORES
```

After this general setting, the user have to specify two additional lines to select which statistics will be stored in the ABC reference table. For example, if we want to use the unfolded allele frequency spectrum as summary statistics, we have to set this line in the sumstat.par file:

```
AFS#1#1#1
```

and add the following line to the ABCsampler input:

```
launchAfterSS script.sh
```

where “script.sh” is a bash script that moves the allele frequency spectrum computed with 4P in the summary_stats_temp.txt file. The latter file is used by ABCsampler to create the reference table, therefore all statistics the user wants to include in the analysis have to moved in this file.
To include only the unfolded allele frequency spectrum, the script.sh file should contain this two lines:

```
#!/bin/bash
mv AFS-U_\”Sample\ 1\”.txt summary_stats_temp.txt
```

7. FURTHER INFORMATION

How to cite 4P

Benazzo A, Panziera A, Bertorelle G () 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *XXX*.

Technical details

The software was compiled using gcc4.3 (Linux), MinGW-w64 (Windows) and X-code (Mac Os X).

The software was tested on Ubuntu 12.04 (32 and 64 bit), Windows 7 (64 bit) and Mac OS X 10.6 Snow Leopard (32 and 64 bit).

References

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *The American Journal of Human Genetics*, 81:1084-1097.

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179-81.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLOS Genetics*, 9(10):e1003905.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genetics, 5(10):e1000695.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, 59:1633-1638.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955-9

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology*, 17:4015-4026.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences U S A*, 70:3321-3323.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, 47:253-259.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, 11:116

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38:1358-1370.