Workshop on Population and Speciation Genomics, 2020
Practical worksheet: Detecting selection using SNP data

Rachael Dudaniec, Macquarie University
rachael.dudaniec@mq.edu.au

Practical learning goals:

- Learn code for running a variety of popular differentiation-based and environmental association analyses (EAA)
- Modify the code to apply different parameter sets and evaluate how different settings impact outlier detection.
- Plot and interpret results of different tests
- Compare results across tests and determine reasons for differences across methods

What we'll cover (each Part has a separate R script file):

- Part 1.0 Load and examine genetic data, plot and correlate environmental variables
- Part 2.0 Differentiation-based outlier detection with pcadapt and OutFLANK
- Part 3.0 Multivariate EAA with redundancy analysis (RDA)
- Part 4.0 Univariate EAA with latent factor mixed models (LFMM)

Recommendations:

- Work together with your neighbor(s) (i.e. try out different parameter sets)
- Please ask if you have questions!

Data for the practical comes from:
Dudaniec RY, Yong CJ, Lancaster LT, Svensson EI, Hansson B (2018) Signatures of local adaptation along environmental gradients in a range-expanding damselfly (*Ischnura elegans*). Molecular Ecology. 27(11): 2576-2593

Code for this practical have been adapted from:
Dudaniec RY, Yong CJ, Lancaster LT, Svensson EI, Hansson B (2018) Signatures of local adaptation along environmental gradients in a range-expanding damselfly (*Ischnura elegans*). Molecular Ecology. 27(11): 2576-2593; https://doi.org/10.1111/mec.14709

Forester BR, Lasky JR, Wagner HH, Urban DL (2018) Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Molecular Ecology* 27, 2215-2233

Some of the code and annotations for this practical are adapted from:
Forester et al. online tutorial at: https://popgen.nescent.org/2018-03-27_RDA_GEA.html#references and Brenna Forester: Practical 2 of the Workshop on Ecological and Evolutionary Genomics 2019, Sydney, Australia.

Part 1.0: The Data (readme)

The data are from adult damselflies of *Ischnura elegans*, collected in southern Sweden. This damselfly is undergoing a poleward range shift under climate change. Here are interested to know how environmental selection are operating on the species during range expansion.

The genetic data ("SNPdata.txt"):

- Contain 13,612 SNPs, from RAD sequencing, filtered as described in Dudaniec et al. (2018), with a MAF = 0.05 and mean 15x depth of coverage.
- Contain 426 individuals collected from 25 sites (10-20 per site), sampled following a temperature gradient, with latitudinal replicates from south to north.
- Are not at equilibrium (i.e. not at migration-selection balance), genetic structure is evident with 4 main genetic clusters identified in Dudaniec et al. (2018), but with moderate gene flow between clusters that increases towards the range limit.

The environmental data ("envDat.env"):

- Consist of 5 environmental variables important for the ecology and survival of *I. elegans*: Mean Annual Temperature, Mean Annual Precipitation, Maximum Mean Summer Temperature, Wind Speed, and % Tree Cover.
- Are extracted from the WorldClim (BioClim) database and from the Global Land Cover Facility
- Are at 1km resolution (raster cell size)
- The distribution of environmental values is generally correlated with latitude, and the range expansion axis.

********** **Work through the R script "Part1_TheData.R"**********

Questions to consider:

1.1How do you think the demographic history of this dataset (i.e. range expansion) will impact outlier detection rates and, if so, how? A: genetic structure may be an issue = false discoveries. Allele surfing could be incorrectly interpreted as adaptation.

1.2. How could including highly correlated environmental variables in the analysis affect our outlier detection? A: it could lead to inflation/bias in the N outliers detected, lot of common loci too.

Part 2.0: Differentiation-based outlier detection (readme)

Differentiation-based outlier detection is most useful for detecting loci of large effect, or those with large differences in allele frequencies between locations. The approach does not examine environmental selection, and detects outliers using genetic data only.
Here we will run two methods of differentiation-based outlier detection, ***OutFLANK*** (Whitlock and Lotterhos 2015) and ***pcadapt*** (Luu et al. 2017). We will compare the output from each program, with different parameters. Some points about these differing approaches:

- *pcadapt* is based on Principal Component Analysis to detect outliers where each SNP is regressed against each principal component, with outliers extracted using z-scores.
- *OutFlank* calculates a normal distribution of Fst values from all SNPs and detects outliers using left and right-tail trim fractions.
- *pcadapt* is not impacted by admixture as it does not require 'populations' to be defined - *OutFlank* requires individuals to be grouped into genetic clusters or populations.
- The genomic inflation factor (GIF) is used in *pcadapt* to correct for inflation of the test score at each locus, which occurs when population structure or other confounding factors are not appropriately accounted for. See François et al. (2016), Mol Ecol.
- Both methods apply a False Discovery Rate (FDR) that the user decides, which is a cut-off applied to identify outliers with a given number of expected false positives.

********** **Work through the R script "Part2_Outliers.R"** **********

2.1 Running OutFLANK

Table 1. *Fill in your outlier results using different OutFlank parameters. You may wish to divide the tests with your neighbour(s) and fill in the table. K is set to 25, the number of sites. Note that $dfInferred is produced in 'head(OFoutput,5)', and is the inferred degrees of freedom for the chi-square distribution of neutral FST (from which outliers are detected).*

q-threshold (FDR rate)	L + R Trim Fraction?	Number of outliers?	$dfinferred
0.10	0.10	333	13.62
	0.05	290	13.26
	0.01	50	8.92
0.05	0.10	191	13.67
	0.05	174	13.33
	0.01	34	8.92
0.01	0.10	51	13.40
	0.05	51	13.29
	0.01	22	8.91

Questions to consider:

Q 2.1: How do the q-threshold vs L+R Trim Factors appear to affect outlier detection?

A: Increasing the Trim factor reduces the breadth of the null Fst distribution so you get more outliers. A lower trim fraction includes more loci with extreme Fst values in the null distribution so you get fewer outliers. Higher q thresholds raise the acceptable proportion of false discoveries in the data so you get more outliers.

Q 2.2. What effect do the trim factors have on the chi-square Fst distribution? Why would values for the 0.10 and 0.05 Trim Fraction be similar? A: the degrees of freedom are very similar for the two trim factors so it isn't changing the null distribution very much, so the proportion of outliers being included/excluded is very similar. This may vary among datasets/depends on the data.

Q 2.3. How would you describe the relative effects of adjusting the q-threshold versus the Trim Fraction?

A: Adjusting the q threshold has a greater effect on N outliers than the Trim factor, which is more specific to the distribution of the Fst values among the loci in your data-(sets the values from which an outlier can be identified). Q threshold applies a cut-off according to the p-values -whatever that distribution may be.

2.2 Running *pcadapt*

Table 2. *Fill in your results using different pcadapt parameters (FDR, K). For K, compare K = 2 and K = 4. You may wish to divide tests with your neighbour(s) and fill in the table. Note: Keep the modified GIF the same across tests to enable meaningful comparisons.*

K	FDR (qval)	Modified GIF	Number of outliers *pcadapt*?
2	0.10	1.10	1511
	0.05	1.10	1304
	0.01	1.10	1031
4	0.10	1.10	1617
	0.05	1.10	1408
	0.01	1.10	1070

Questions to consider:

2.2.1. How does the number of pcadapt outliers detected differ across parameter values (i.e. K and FDR?) and across the two detection methods (pcadapt vs OutFLANK)? Do numbers of outliers detected vary proportionately across methods? A: Increasing K results in slightly more outliers (greater partitioning of variance?) but not strikingly. Reducing FDR decreases N outliers in similar proportion across K. Number of pcadapt outliers is a lot higher than OUTFLANK> Overlap is low. Why? Different methods. OutFlank better for larger effect loci… Ordination is multivariate and may include more small effect loci in the outliers that covary.


© Pete Hillman 2017

Part 3.0: Multivariate Environmental Association Analysis (readme)

Multivariate EAAs are a powerful complement to univariate detection approaches. RDA is a multivariate ordination technique that analyzes matrices of loci and environmental predictors simultaneously. Some points about RDA:

- It determines how groups of loci covary in response to the multivariate environment, as opposed to univariate approaches that apply multiple tests per locus. This makes it more useful for detecting weaker, polygenic signatures of adaptation.
- RDA performs multivariate linear regression on genetic and environmental data, producing a matrix of fitted values. Then PCA of the fitted values is used to produce canonical axes, which are linear combinations of the environmental predictors.
- RDA doesn't require corrections for multiple tests because it analyzes all genomic and environmental data simultaneously. However, post processing includes checking and modification of the GIF and p-value distribution, with application of the FDR threshold.

Though we do not cover it here, see Partial Redundancy Analysis (pRDA), which is a good complement to RDA that integrates effects of geographic distance on outlier detection (REF).

RDA can be used on both individual and population-based sampling designs. Code for running an RDA with population level data is provided in this practical. See Forester et al. (2018) Molecular Ecology for more details.

********** **Work through the R script "Part3_MultivariateEAA.R"**********

Here we run an RDA with our environmental variables and examine the data in two different ways. Firstly (1) we use an FDR approach based on Mahanolobis distance calculation and apply the modified GIFs to examine outlier numbers. Secondly (2) we use the standard deviation p-value 'cut-off' method of Forester et al. (SD Approach) which does not depend on the assumptions of a 'flat' p-value distribution required for reliable FDR application.

Table 3. *Fill in your results using different RDA parameters (i.e. FDR, GIF) when retaining all 4 PC axes. You may wish to divide tests with your neighbour(s) and fill in the table together. Note: Keep the modified GIF the same (e.g. 1.0) across tests to enable meaningful comparisons.*

PC axes retained	FDR (qvalue) cut-off	Original GIF	Number of outliers	Modified GIF	Number of outliers
4	0.10	1.28	356	1.0	1062
	0.05	1.28	486	1.0	883
	0.01	1.28	611	1.0	561

Questions to consider:

Q 3.1. Looking at the results of Table 1, and the p.value distributions of the original and modified GIFs, do you think these results are reliable?

Q. 3.2. Using the SD approach, which environmental variables appear to explain adaptive genetic variation the most, and which the least?

Part 4.0: Univariate EAA: Latent Factor Mixed Models

LFMM is a regression model that includes unobserved variables (latent factors) that correct the model for confounding effects. The latent factors are estimated simultaneously with the environmental and response variables, which can help improve power when environment and demography are correlated.
The previous version of LFMM (v1.5, implemented in the LEA package) uses an MCMC algorithm to identify GEAs while correcting for confounding factors. MCMC made it (very!) time-intensive for large data sets. LFMM v.2 computes LFMMs for GEA using a least-squares estimation method that is substantially faster than v1.5.

Citation: Caye et al. (2019) LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies.

********** **Work through the R script "Part4_UnivariateEAA.R"**********

Table 4. *Complete the table below for the first variable 'Max Temp' in the LFMM, applying the modified vs. original GIF*

K	FDR (qval)	Adjusted/ Modified GIF	Number of outliers	Unadjusted, original GIF	Number of outliers
4	0.01	1.10	1312	2.62	3437
	0.001	1.10	744	2.62	1913
	0.0001	1.10	453	2.62	1176
**OPTIONAL*					
5	*0.01*	*1.10*	*990*	*2.14*	*2784*
	0.001	*1.10*	*597*	*2.14*	*1669*
	0.0001	*1.10*	*407*	*2.14*	*1070*

******OPTIONAL Table 5**. Modify the R script to test for the different variables (see code annotations). Complete the table below for the rest of the environmental variables in the LFMM, with K =4, and FDR = 0.0001, applying the modified GIF only (you may insert result for Max Temp from Table 4).*

K = 4	FDR (qval)	Adjusted/ Modified GIF	Number of outliers
Max Temp	0.0001	1.10	407
Precipitation	0.0001	1.10	192
Tree Cover	0.0001	1.10	165
Wind Speed	0.0001	1.10	222

Questions to consider:

Q 4.1: What effect does increasing K have on outlier detection? What effect does increasing FDR have on outlier detection?

Q.4.2: How does GIF modification affect outlier detection and why?

****Q 4.3 Which environmental variable had the most SNP associations? Which had the least?*

*****4.6: OPTIONAL TASK: Run PC predictor variables in LFMM**

There is no agreement among statisticians on multiple test corrections for this case: we are testing e.g. 13612 SNPs and 4 predictors = 54,448 tests, with each SNP tested 4 times. But we are not making a correction for that. This may be problematic but there doesn't appear to be a right answer.

One solution is to run a PCA on the predictors and only run tests on the first one (or two, or three) PCs...this will minimize the number of tests, while maintaining the information in our set of predictors. However, PCs should also be biologically meaningful, and using this approach may depend on the objectives of your study.

Things to consider in Environmental Association Analyses:
How do you want to handle multiple environmental predictors (lfmm only)?
How sensitive are the results to your choice of K (lfmm only)?
How much do you want to adjust the p-value distribution (a.k.a adjust the GIF)?
How sensitive are the results to different cutoff thresholds/values of K?
How sensitive are the results to the MAF filter applied to your data?
How sensitive are the results to missing data?

END