Computational Approaches to Detecting the Signatures of Local Adaptation

Roman Arguello

Academic Year 2019-2020

Contents

1 Intro: What is local adaption? (a circumscribed overview)

- The apparent "fit" of an organism to its environment has long been a topic of interest: phenotypic correlation with environmental variables. What are some examples that you find most compelling?
- This "fit" has been taken as evidence of selection, but is it really? How do we claim that the association between an apparent phenotype an an apparent environmental variable(s) are not by chance?
  - How could this be *tested*?
    - experimental: i.e. transplant experiments (not covered here)
    - statistically: i.e. methods to test for changes in phenotypes or genetic variation
- We are going to focus on (1) genetic/genomic data and (2) recent evidence for adaptation.
- General types of approaches
  - pop. gen
  - GWAS
  - QTL

2 population genomic approaches

- many of the current usages of F_{ST} trace back to the mid 20th century and work of Sewall Wright [1]
- F_{ST} = "fixation indices" = statistical framework for studying the expected level of heterozygosity in populations

- the ability to get measurements required the technology to start accessing heterozygosity (protein variation and DNA variation)
- most common modern usage: allele frequency differences between populations:

$$F_{\mathrm{ST}} = \frac{\sigma^2_{sub}}{\bar{p}(1-\bar{p})}$$

$$F_{\mathrm{ST}} = \frac{\pi_{A\text{-}B}-\pi_A}{\pi_{A\text{-}B}}$$

$$F_{\mathrm{ST}} = \frac{MSP-MSG}{MSP+(n_c-1)MSG}$$

$$\mathrm{MSG} = \frac{1}{\sum_{i=1}^{s} n_i-1}\sum_i^s n_i p_{Ai}(1-p_{Ai})$$

$$\mathrm{MSP} = \frac{1}{s-1}\sum_i^s n_i(p_{Ai}-\bar{p}_{Ai})^2$$

where n_i = sample size in sub population i, $\bar{p} = \frac{n_i p_{Ai}}{\sum_i n_i}$, and where

$$n_c = \frac{1}{s-1}\sum_i^s n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$$

which is the average sample size across the the samples, that incorporates/corrects for the variance in sample size over the subpopulation

- there are multiple modification on the Weir & Cocker formulations, and related
  - blog discussion

2.1 “more standard” F_{ST} approahces

- **raw F_{ST} scans:** determine the empirical/null distribution of F_{ST} on your data the tail(s) should be enriched with candidates for population differentiation
  - simple, no additional info needed
  - sometimes all you really want to know are differentiated positions (especially if they fall w/in “good” candidate loci)
    - examples: [2]
  - lots of false positives, you will always have a tail
  - usually pair-wise, which is not always ideal
  - **software:**
    - custom scripts,
    - arlequin [3],
    - DnaSP [4]
    - fsthet (R package)
- **raw F_{ST} scans + better info on a null distribution:** determine the empirical distribution + provide a threshold that is informed by demographic modeling and simulations (i.e. [5]), or additional inspection of F_{ST} distribution
  - potentially still simple if a demographic model exists for you samples (but how good is the demographic model ?)

    - potentially provides extra protection agains false positives (depends on the demographic model)
        - **simulation software:** coalescent simulator (i.e. ms [6], fastsimcoal (XX), primems (XX), cosi (XXX)

- null distribution via other means:
    - matched data to form a null
        - **software:**
            - SmileFinder [7]
            - posterior predictive simulation for 2 populations: GppFst (R package)
              - "GppFst will compute the probability of observing an empirical proportion of loci within a given F_{ST} range conditioned on the particular coalescent model of population divergence"
    - matched F_{ST} distribution
        - **software:**
            - OutFLANK [8] (R package)
            - attempts to fit χ^2 distribution to central part of a F_{ST} distribution from a reference set of variants
- **Hierarchical approaches:** one may want more flexibility in deciding what are the groupings
    - **software:** HierFstat (R package)
        - while some software allows a limited number of hierarchical levels, HierFstat (which implements the methods of Yang [9]) allows an arbitrary number of levels [10]
- **PCA-based approaches**
    - **software:** pcadapt [11] (R package)
        - very fast and naturally accounts for population structure

2.2 more complicated model-based approaches

- Bayesian "F-models"
    - **software:** BayeScan [12, 13]
        - model choice approach in Bayesian framework
    - **software:** BlockFeST [14]
        - builds on Bayescan but groups variants into predefined blocks

3 haplotype-based approaches

- **software:** hapFLK [15, 16]
    - builds upon a parametric test that is tree-based (includes branching order and pop. size variation) and extends it to haplotypes
- **extended haplotype tests:** iHS, EHH, XP-EHH
    - **software:** rehh [17], fastPHASE [18], hapbin [19], selscan [20]
    - picks up on sweep signals that extend the run of homozygosity

4 combining genetics with environmental traits more explicitly

- **software:** BayeScEnv [21]
    - builds upon BayeScan but with the ability to incorporate environmental variables
- **software:** BayEnv [22, 23]
    - builds on Bayescan but groups variants into predefined blocks

5 QTL

- **software:** R\QTL [24]
  - implements a large set of methods and plotting functions and many tutorials

6 GWAS

- **software:** GWASTools [25], Plink [26], rrBLUP [27], GWASpoly [27] [28], Hail, BGENIE [29]
  - like much of above, this is a huge area of research so this is only a subset of the tools

Bibliography

References

[1] S. Wright, "The genetical structure of populations," *Annals of Eugenics*, vol. 15, pp. 323–354, 2019/10/21 1949.

[2] M. Roesti, S. Gavrilets, A. P. Hendry, W. Salzburger, and D. Berner, "The genomic signature of parallel adaptation from shared genetic variation," *Molecular Ecology*, vol. 23, pp. 3944–3956, Apr 2014.

[3] L. Excoffier, G. Laval, and S. Schneider, "Arlequin (version 3.0): an integrated software package for population genetics data analysis," *Evol Bioinform Online*, vol. 1, pp. 47–50, Feb 2007.

[4] P. Librado and J. Rozas, "Dnasp v5: a software for comprehensive analysis of dna polymorphism data," *Bioinformatics*, vol. 25, pp. 1451–2, Jun 2009.

[5] J. M. Akey, G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, "Interrogating a high-density snp map for signatures of natural selection," *Genome Res*, vol. 12, pp. 1805–14, Dec 2002.

[6] R. Hudson, "Generating samples under a wright-fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.

[7] W. M. Guiblet, K. Zhao, S. J. O'Brien, S. E. Massey, A. L. Roca, and T. K. Oleksyk, "Smilefinder: a resampling-based approach to evaluate signatures of selection from genome-wide sets of matching allele frequency data in two or more diploid populations," *Gigascience*, vol. 4, p. 1, 2015.

[8] M. C. Whitlock and K. E. Lotterhos, "Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of f(st)," *Am Nat*, vol. 186 Suppl 1, pp. S24–36, Oct 2015.

[9] R.-C. Yang, "Estimating hierarchical f-statistics," *Evolution*, vol. 52, pp. 950–956, 2019/10/31 1998.

[10] J. GOUDET, "hierfstat, a package for r to compute and test hierarchical f-statistics," *Molecular Ecology Notes*, vol. 5, pp. 184–186, 2019/10/31 2005.

[11] K. Luu, E. Bazin, and M. G. B. Blum, "pcadapt: anrpackage to perform genome scans for selection based on principal component analysis," *Molecular Ecology Resources*, vol. 17, pp. 67–77, Sep 2016.

[12] M. Foll and O. Gaggiotti, "A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective," *Genetics*, vol. 180, pp. 977–93, Oct 2008.

[13] M. C. FISCHER, M. FOLL, L. EXCOFFIER, and G. HECKEL, "Enhanced aflp genome scans detect local adaptation in high-altitude populations of a small rodent (microtus arvalis)," *Molecular Ecology*, vol. 20, pp. 1450–1462, 2019/10/29 2011.

[14] B. Pfeifer and M. J. Lercher, "Blockfest: Bayesian calculation of region-specific fst to detect local adaptation," *Bioinformatics*, vol. 34, pp. 3205–3207, 09 2018.

[15] M. I. Fariello, S. Boitard, H. Naya, M. SanCristobal, and B. Servin, "Detecting signatures of selection through haplotype differentiation among hierarchically structured populations," *Genetics*, vol. 193, pp. 929–41, Mar 2013.

[16] M. Bonhomme, C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. SanCristobal, "Detecting selection in population trees: The lewontin and krakauer test extended," *Genetics*, vol. 186, pp. 241–262, Jun 2010.

[17] M. Gautier and R. Vitalis, "rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure," *Bioinformatics*, vol. 28, pp. 1176–1177, 03 2012.

[18] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *Am J Hum Genet*, vol. 78, pp. 629–44, Apr 2006.

[19] C. A. Maclean, N. P. Chue Hong, and J. G. D. Prendergast, "hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets," *Mol Biol Evol*, vol. 32, pp. 3027–9, Nov 2015.

[20] Z. A. Szpiech and R. D. Hernandez, "selscan: an efficient multithreaded program to perform ehh-based scans for positive selection," *Mol Biol Evol*, vol. 31, pp. 2824–7, Oct 2014.

[21] O. E. GAGGIOTTI and M. FOLL, "Quantifying population structure using the f-model," *Molecular Ecology Resources*, vol. 10, pp. 821–830, Aug 2010.

[22] G. Coop, D. Witonsky, A. Di Rienzo, and J. K. Pritchard, "Using environmental correlations to identify loci underlying local adaptation," *Genetics*, vol. 185, pp. 1411–23, Aug 2010.

[23] T. Günther and G. Coop, "Robust identification of local adaptation from allele frequencies," *Genetics*, vol. 195, pp. 205–20, Sep 2013.

[24] K. W. Broman, H. Wu, Ś. Sen, and G. A. Churchill, "R/qtl: QTL mapping in experimental crosses," *Bioinformatics*, vol. 19, pp. 889–890, 05 2003.

[25] S. M. Gogarten, T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, D. Levine, T. Lumley, S. C. Nelson, K. Rice, J. Shen, R. Swarnkar, B. S. Weir, and C. C. Laurie, "GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies," *Bioinformatics*, vol. 28, pp. 3329–3331, 10 2012.

[26] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "Plink: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, pp. 559–575, 2019/11/07 2007.

[27] J. B. Endelman, "Ridge regression and other kernels for genomic selection with r package rrblup," *The Plant Genome*, vol. 4, pp. 250–255, 2011.

[28] U. R. Rosyara, W. S. De Jong, D. S. Douches, and J. B. Endelman, "Software for genome-wide association studies in autopolyploids and its application to potato," *The Plant Genome*, vol. 9, 2016.

[29] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, G. McVean, S. Leslie, P. Donnelly, and J. Marchini, "Genome-wide genetic data on 500,000 uk biobank participants," *bioRxiv*, 2017.