 DOI: 10.1086/688018

Appendix from S. Hoban et al., "Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions" (Am. Nat., vol. 188, no. 4, p. 379)

Supplementary Material

Appendix Tables

Table A1: A list of approaches to control for population structure in differentiation outlier tests, genetic-environment associations, and genome-wide associations with phenotype data

Method of correcting for or estimating neutral structure	Approach	Differentiation outlier	Genetic-environment association	Genome-wide association with phenotype	Pros	Cons
No correction for neutral structure	NA	NA	SAM/MatSAM (Joost et al. 2007, 2008); CART (e.g., Jones et al. 2013)	Lasso-penalized logistic regression (multivariate; Wu et al. 2009)	Nonlinear (i.e., logistic) relationships can be modeled; CART makes no assumptions about distribution of the data	High false positive rates because no correction for structure
Simulate null distribution	Demographic null model	fdist2/LOSITAN (assumes island model; Beaumont and Nichols 1996; Antao et al. 2008); hierarchical island model (Excoffier et al. 2009*b*); specific scenario (e.g., range expansion; Eckert et al. 2010)	NA	NA	Can take into account sampling design and missing data	Demographic history must be accurately captured by simulation
Bayesian hierarchical model, uses multinomial Dirichlet (or beta-binomial) distribution	Demographic null model	BayeScan (Foll and Gaggiotti 2008); SelEstim (extension of BayeScan; Vitalis et al. 2014); specific scenario (e.g., paired sampling designs; Foll et al. 2014)	Bayescenv (de Villemereuil and Gaggiotti 2015)	NA	Dirichlet distribution is able to describe the distribution of allele frequencies over a wide range of demographic models, including unequal migration; can be a powerful approach if the underlying model is accurate	Assumes that the samples have not recently exchanged migrants; Bayesian models can be sensitive to priors; long run time
Matrix of covariance in allele frequencies among populations	Sample covariance/ relatedness	Bayenv2 (X^TX; Günther and Coop 2013); BayPass (X^TX; Gautier 2015); Berg and Coop 2014 (multivariate)	Bayenv2; BayPass; Berg and Coop 2014 (multivariate)	BayPass	Explicitly accounts for evolutionary history among samples; can account for pool-seq ability to calculate covariance matrix from subset of SNPs; BayPass can account for LD using a genetic map	Bayesian models can be sensitive to priors; MCMC requires long run times; Bayenv2 assumes loci are not linked

Tree of population histories	Population tree	FLK (Bonhomme et al. 2010); SOM/HMM (Jones et al. 2012)	NA	NA	Explicitly accounts for evolutionary history among samples	Closely related samples may be hard to resolve; user must supply or choose a population to root the tree; for SOM/HMM, approach has not been evaluated by simulation
LMM with random effects for relatedness, covariance, and/or structure	Mixed model	NA	Methods on right have been used with environment instead of phenotype as response variable (e.g., Yoder et al. 2014)	(Phenotype ~ genotype); EMMAX (Kang et al. 2010); TASSEL (Bradbury et al. 2007); GCTA (Yang et al. 2011); LMM-Lasso (multivariate; Rakitsch et al. 2013)	GWAS can be powerful for detecting polygenetic (i.e., small) effects on phenotypes	Using GWAS methods for environmental data (environment ~ genotype) instead of phenotype has not been well evaluated
Latent factor models (random effects of population structure jointly estimated with main effects)	Mixed model	PCAdapt (Duforet-Frebourg et al. 2014; latent factors closely related to principal components)	LFMM (latent factors closely related to principal components; Frichot et al. 2013; Frichot and François 2015); gINLAnd (latent factors modeled as Gaussian; Guillot et al. 2014)	NA	Ability to detect outliers along different axes of structure; computationally fast; gINLAnd allows spatial covariance to be calculated from subset of SNPs	For PCAdapt and LFMM, results may be sensitive to number of latent factors chosen by user; no missing data (missing genotypes must be imputed); assumes loci are not linked
Linear model with principal components as covariate	Mixed model	NA	NA	EIGENSTRAT (Price et al. 2006)	Computationally fast	Results may be sensitive to number of principle components; no missing data
Core distribution used to estimate parameters for neutral distribution	Other	OutFLANK (Whitlock and Lotterhos 2015)	NA	NA	Computationally fast; low false-positive rates	Assumes F_{ST} follows χ^2 distribution
Geographic distance matrix or geographic location used as predictor(s) or to calculate random effects	Mixed model	NA	GDM (Ferrier and Guisan 2006; e.g., Fitzpatrick and Keller 2014); MAGICS (Raj et al. 2013); GEE (e.g., Poncet et al. 2010); GLMM and GAMM (Jones et al. 2013); SAMbADA (Stucki et al. 2014)	NA	Can model nonlinearities; can generate maps; in some cases, can use information criterion to rank competing models; GDM is computationally fast	Geographic distance or location must accurately capture neutral structure; most of these approaches for detecting selection have not been well evaluated by simulation

Table A1 (*Continued*)

Method of correcting for or estimating neutral structure	Approach	Differentiation outlier	Genetic-environment association	Genome-wide association with phenotype	Pros	Cons
Moran's eigenvector map variables/ PCNM (Borcard and Legendre 2002; Dray et al. 2006)	Mixed model	NA	E.g., Fitzpatrick and Keller (2014) use to control structure in gradient forests (Ellis et al. 2012)	NA	Can model nonlinearities; computationally fast; can generate maps	Approach for detecting selection has not been well evaluated by simulation
Individual ancestry coefficients and geographic location	Sample covariance/ relatedness	TESS3 (Caye et al. 2016)	NA	NA	Computationally fast; allows plotting of maps; deals with departures from HWE created by inbreeding or geographically restricted mating	Results may be sensitive to choice of number of ancestral populations
Permutations	Other	E.g., CSS (Jones et al. 2012)	NA	NA	Makes no assumptions about underlying distribution	Computationally slow and intensive

Note: We focus on approaches that have been used to detect selection from genome data (note that other landscape-genetic approaches for detecting gene flow or performing an isolation-by-environment analysis have been excluded). We summarize the pros and cons of each approach. CART = classification and regression trees; SNP = single nucleotide polymorphism; LD = linkage disequilibrium; MCMC = Markov chain Monte Carlo; SOM/ HMM = self-organizing map-based iterative hidden Markov model; LMM = linear mixed model; GCTA = genome-wide complex trait analysis; GWAS = genome-wide association studies; LFMM = latent factor mixed models; GDM = generalized dissimilarity modeling; MAGICS = Mantel-generalized linear models to infer clinal selection; GEE = generalized estimating equations; GLMM = generalized linear mixed models; GAMM = generalized additive mixed models; PCNM = principal coordinate analysis of neighbor matrices; HWE = Hardy-Weinberg equilibrium; CSS = cluster separation score; NA = not applicable.

Table A2: List of reduced representation methods and corresponding references

Reduced representation method	References
Exome-capture arrays	Bamshad et al. 2011
Custom sequence capture probes	Jones and Good 2016
RNA sequencing	De Wit et al. 2012
Single nucleotide polymorphism chips	Wang et al. 1998
Restriction site–associated DNA sequencing	Baird et al. 2008; Andolfatto et al. 2011; Elshire et al. 2011; Peterson et al. 2012; Puritz et al. 2014*b*

Table A3: Data on sampling designs from published and simulation studies used to create figure A1

EmpOrSim, abbreviation	Citation	Organism	No. individuals	No. populations	No. markers	Spatial scale	DOI
Emp:							
Zu2014	Zueva et al. 2014	*Salmo salar L.* (Atlantic salmon)	472	12	4,631	1,000	10.1371/journal.pone.0091672
DA2014	Dell'Acqua et al. 2014	*Brachypodium distachyon* (purple false brome grass)	96	9	16,697	800	10.1186/1471-2164-15-801
Ge2014	Geraldes et al. 2014	Admixture between *Populus trichocarpa* and *Populus balsamifera*	498	30	33,070	3,000	10.1111/evo.12497
Be2015	Berg et al. 2015	*Gadus morhua* (Atlantic cod)	194	7	8,809	700	10.1093/gbe/evv093
Ec2015	Eckert et al. 2015	*Pinus lambertiana* (sugar pine)	241	10	475	40	10.1007/s11295-015-0863-0
Fr2015	Fraser et al. 2015	*Poecilia reticulata* (guppies)	153	8	7,757	20	10.1111/mec.13022
Gr2014	Gray et al. 2014	*Andropogon gerardii* (big bluestem grass)	378	11	384	900	10.1111/mec.12993
Ha2014	Hamilton et al. 2015	*Picea sitchensis* × *Picea glauca* and *Picea glauca* × *Picea engelmannii*	1,492	40	71	3,000	10.1007/s11295-014-0817-y
DK2014	De Kort et al. 2014	*Alnus glutinosa* (black alder)	356	24	1,990	1,200	10.1111/1365-2664.12305
Zh2015	Zhou et al. 2014	*Pinus massoniana* and *Pinus hwangshanensis*	104	26	884	NA	10.1111/mec.12830
Sim:							
LW2015	Lotterhos and Whitlock 2014, 2015	Simulation	300	30	10,000	NA	10.1111/mec.13100; 10.1111/mec.12725
LW2015	Lotterhos and Whitlock 2014, 2015	Simulation	900	90	10,000	NA	10.1111/mec.13100; 10.1111/mec.12725
DM2013	De Mita et al. 2013	Simulation	100	100	1,100	NA	10.1111/mec.12182
DM2013	De Mita et al. 2013	Simulation	384	8	1,100	NA	10.1111/mec.12182
dV2014	de Villemereuil et al. 2014	Simulation	8,000	16	5,050	NA	10.1111/mec.12705
Jo2013	Jones et al. 2013	Simulation	300	100	100	NA	10.1111/evo.12237

Note: The empirical studies used in the table were made from a Google Scholar search on June 4, 2015, to obtain all citations of latent factor mixed models and Bayenv2 (both methods were published in 2013). This is not a comprehensive list but considered to be representative of the current literature.

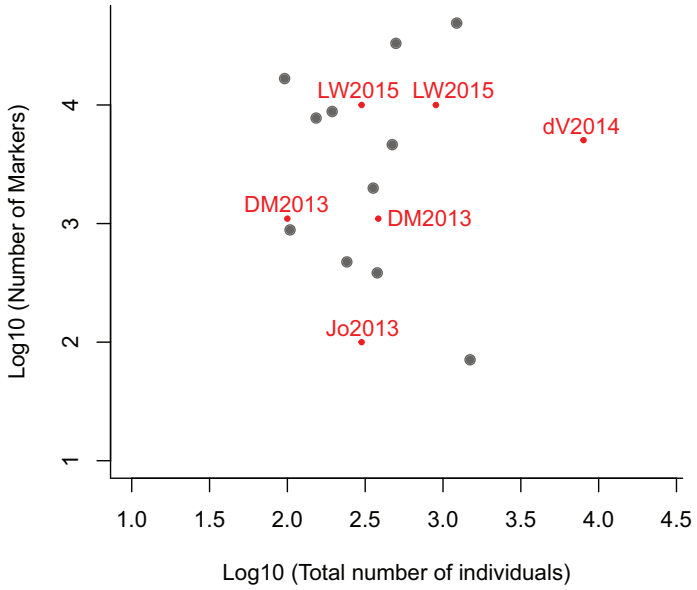

Figure A1: Examples of sampling designs from published studies (black dots) and from simulation studies (labeled red points; data available in table A3). LW2015 = Lotterhos and Whitlock 2015; DM2013 = De Mita et al. 2013; dV2014 = de Villemereuil et al. 2014; Jo2013 = Jones et al. 2013.