DETECTING SELECTION IN NATURAL
POPULATIONS

INTRODUCTION

# The past, present and future of genomic scans for selection

JEFFREY D. JENSEN,*† MATTHIEU FOLL‡
and LOUIS BERNATCHEZ§

*School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; †Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; ‡Genetic Cancer Susceptibility, International Agency for Research on Cancer, Lyon, France; §IBIS (Institut de Biologie Intégrative et des Systèmes), Université Laval, Québec, Québec, Canada

The notion of utilizing expected patterns of variation produced by genetic hitchhiking to scan genomes for recent adaptive fixations has remained prevalent in the field of population genetics for many years, being first inspired by the seminal study of Lewontin & Krakauer (1973). In this vein, test statistics have been designed to identify regions of reduced variation, skewed site frequency spectra, elevated linkage disequilibrium, and/or elevated rates of interpopulation or interspecies divergence. Along with these dozens of test statistics have come hundreds of empirical applications in a wide array of organisms.

Despite the initial, and in many ways increasing enthusiasm for genomic scans of selection following the development of powerful high-throughput sequencing methods, this literature also gives rise to a number of notable points of caution (see Haasl & Payseur 2015). These include the general observation that other factors may produce genomic regions with patterns of variation strongly resembling that expected under hitchhiking models, including population size change (e.g. Teshima et al. 2006), population structure (e.g. Excoffier et al. 2009) or background selection (e.g. Stephan 2010) – potentially resulting in a high false-positive rate. However, they also include the less cited aspect of potentially low power of detection due to heterogeneous mutation and recombination rates (Roesti et al. 2012; Tine et al. 2014), past demographic events (even when accounted for) (Bierne et al. 2011) or the genetic architecture of the selected trait (Le Corre & Kremer 2012). This problem is exacerbated when ascertaining regions for specific 'hitchhiking-like' patterns of variation without appropriate P-value correction in subsequent testing (Thornton & Jensen 2007).

Correspondence: Jeffrey D. Jensen, E-mail: jeffrey.jensen@epfl.ch

Accompanying these critiques has been the general observation that despite hundreds of genomic scans, the number of identified loci that have been subsequently connected in a meaningful way to phenotypic or fitness-level changes is limited. In a pessimistic sense, these scans may simply produce long lists of candidate genes (where the list appears to be largely dependent on the statistical method chosen; see Biswas & Akey 2006 and Schlamp et al. 2015) for which any follow-up functional study would be extremely difficult and are hence rarely undertaken. However, there are notable exceptions for scans limited to candidate loci defined a priori (e.g. Bernatchez & Landry 2003; Colosimo et al. 2005; Hoekstra et al. 2006).

Thus, this special issue has been organized in order to ask many of the leading empirical researchers and theoreticians in this area to discuss how we can go beyond standard genomic scans in order to overcome some of these difficulties identified over the past two decades. We here briefly note the main themes that have emerged from the resulting articles. These themes can be summarized under two general terms: alternative models and alternative data sources.

## Alternative models

As noted above, the notion that nonequilibrium demographic models may replicate patterns associated with genetic hitchhiking has been widely discussed and explored in the literature, with a major focus on the effects of population size change (reviewed in Crisci et al. 2012). However, the ability to co-estimate demographic and selection parameters remains a major challenge despite this focus, although some notable progress has indeed been made for accurately estimating selection parameters in a way that is robust to certain nonequilibrium perturbations (e.g. Ormond et al. 2015). In this special issue, the authors bring attention to a number of alternative models that must now be explored more fully and tested empirically in order to increase the accuracy and biological reality of the assumptions underlying scans for selection. With regard to demography, there is a strong focus on the need for a more in-depth understanding of hitchhiking effects, and a description of expected neutral patterns of variation, under both models of population structure with migration and for recently introgressed populations – with results here particularly demonstrating the difficulty of identifying selected loci across much of this parameter space (Vatsiou et al. 2015). This special issue also presents a number of empirical studies examining the dynamics of selection in subdivided and introgressed populations, in systems ranging from mussels to lizards (Ferchaud & Hansen 2015; Fraïsse et al. 2015; Laurent et al. 2015; McGee et al. 2015; Ravinet et al. 2015; Schweizer et al. 2015 a, b; Wenzel et al. 2015).

Apart from additional demographic considerations, there is also a strong call for alternative selection models to be explored and incorporated. In particular, the case is made that the genomic effects of background selection can no longer be ignored (see Charlesworth 2012) – and it is

demonstrated here via simulation that neglecting this effect may result in a strong misinference of demographic parameters (Ewing & Jensen 2015). Given that these demographic models often in turn become the nulls in genomic scans for selection, this observation is indeed troubling. Thus, estimating parameters of background selection in addition to demography ought to become the null model of interest. As an alternative approach, supplementing the popular Sweepfinder framework (Nielsen *et al.* 2005), Huber *et al.* (2015) introduce some corrections to account for the effects of background selection in genome scans without directly estimating the underlying parameters. As an illustration, Renzette *et al.* (2015) discuss the ability to distinguish between the genomic effects of genetic hitchhiking and background selection utilizing theoretical results of Innan & Stephan (2003), arguing for a pervasive role for background selection in populations of human cytomegalovirus (HCMV). Relatedly, by collecting data over a broad phylogenetic coverage, James *et al.* (2015) make the case that the evolution of mitochondrial DNA is indeed dominated by the presence of slightly deleterious mutations – though, upon accounting for this effect, still find strong evidence for positively selected fixations as well.

In addition to these developments surrounding background selection, other alternative selection models are also highlighted. Specifically, the ability of genomic scans themselves to identify sites underlying polygenic adaptation is discussed; the case for statistical approaches focused on this model is strongly made, and the requirements underlying them are detailed (Stephan 2015). This point is also highlighted in a RAD sequencing approach to argue for the occurrence of a polygenic adaptation in North Atlantic Eels (Laporte *et al.* 2015), providing additional empirical validity for the relevance of this model. This study also highlights the power of combining genomic and environmental data (in this case pollutants) in a multivariate statistical framework in order to identify potential selective agents shaping observed patterns of allelic variation. Relatedly, Forester *et al.* (2015) examine the effects of such landscape heterogeneity on the spatial genetic signature of positive selection.

### Alternative data sources

On the empirical side, the authors here have also considered what type of data may serve to further improve inference and augment biological interpretation of genomic scans. Three common themes have emerged. First, the advantages of time-sampled polymorphism data – where the additional temporal dimension of allele frequency change is affording tremendous inference power. Considerable methodological progress has been made over the past few years (e.g., Malaspinas *et al.* 2012; Foll *et al.* 2014, 2015), and this continues to be a very fruitful area for research; in fact, work herein extends these approaches to models of fluctuating selection (Gompert 2015). Besides the classical applications in natural or experimental population studies, it is also clear that there are at least two important areas of research that will

benefit from time-sampled data – one being the growing field of ancient DNA (Malaspinas 2015), and the other the clinical application of population genetics (Wilson *et al.* 2015). Indeed, the latter is also a quickly growing area, where statistical inference methods are producing clinically relevant results that will shape patient treatment strategies in the future (see reviews of Pennings 2013; Renzette *et al.* 2014).

The second theme is the ability of experimental evolution approaches to inform natural population inference. This is a topic of growing interest (see review of Bank *et al.* 2014), with some controversy regarding the ultimate ability of experimental populations to provide information about adaptive processes in nature (Bailey & Bataillon 2015). Despite this potential caveat, it is clear that the capability to experimentally control the demographic history of, and selective pressures experienced by, experimental populations is a powerful means to gain novel insights into the process of adaptation – insights that have simply not been possible in natural populations owing largely to the many confounding issues discussed above, as well as the complexity of potential selective pressures at play.

Finally, on a more technical note, authors here discuss the future promise of next generation sequencing to provide better information, from building reference genomes (Manel *et al.* 2015), to RAD mapping (Qiu *et al.* 2015) to novel strategies of high-throughput target capture (Jones & Good 2015) – arguing that improved data generation strategies, in addition to the theoretical and methodological improvements discussed above, will also prove fruitful in better quantifying evolutionary processes.

To conclude, we trust that the readership of *Molecular Ecology* will find this special issue to be both timely and insightful – spanning theoretical, statistical, empirical and experimental advances in this area. In addition, we hope that this collection will convince the reader that there are a number of very promising avenues being actively explored, and others that are in great need of further exploration. If made, these advances will continue to improve our ability to identify and quantify beneficial mutations using genomic data, to understand the relative contribution of positive selection in shaping population level variation, to identify the selective agents driving the mode and tempo of adaptation and to broaden the scope of population genetics to touch other diverse research communities ranging from ecology to virology.

### References

Bailey SF, Bataillon T (2015) Can the experimental evolution program help us elucidate the genetic basis of adaptation in nature? *Molecular Ecology*, **25**, 203–218.

Bank C, Ewing G, Ferrer-Admetlla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, **30**, 540–547.

Bernatchez L, Landry C (2003) MHC studies in non-model vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Biswas S, Akey JM (2006) Genomic insights into positive selection. *Trends in Genetics*, **22**, 437–446.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**, 5–22.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in Sticklebacks by repeated fixation of Ectodysplasin alleles. *Science*, **307**, 1928–1933.

Crisci JL, Poh Y, Bean A, Simkin A, Jensen JD (2012) Recent progress in polymorphism-based population genetic inference. *Journal of Heredity*, **103**, 287–296.

Ewing GB, Jensen JD (2015) The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, **25**, 135–141.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Ferchaud A, Hansen MM (2015) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: threespine sticklebacks in divergent environments. *Molecular Ecology*, **25**, 238–259.

Foll M, Poh Y, Renzette N *et al.* (2014) Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genetics*, **10**, e1004185.

Foll M, Shim H, Jensen JD (2015) WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, **15**, 87–98.

Forester B, Jones M, Joost S, Landguth E, Lasky J (2015) The effects of landscape heterogeneity on the spatial genetic signature of local adaptation. *Molecular Ecology*, **25**, 104–120.

Fraïsse C, Belkhir K, Welch JJ, Bierne N (2015) Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Molecular Ecology*, **25**, 269–286.

Gompert Z (2015) Bayesian inference of selection in a heterogeneous environment from genetic time-series data. *Molecular Ecology*, **25**, 121–134.

Haasl RJ, Payseur BA (2015) Fifteen years of genome-wide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, **25**, 5–23.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Huber CD, Degiorgio M, Hellman I, Nielsen R (2015) Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, **25**, 142–156.

Innan H, Stephan W (2003) Distinguishing the hitchhiking and background selection models. *Genetics*, **165**, 2307–2312.

James J, Piganeau G, Eyre-Walker A (2015) The rate of adaptive evolution in animal mitochondria. *Molecular Ecology*, **25**, 67–78.

Jones MR, Good JM (2015) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Laporte M, Pavey S, Rougeux C *et al.* (2015) RAD-sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, **25**, 219–237.

Laurent S(*), Pfeifer SP(*), Settles M *et al.* (2015) The population genomics of rapid adaptation: disentangling signatures of selection and demography in white sands lizards. *Molecular Ecology*, **25**, 306–323.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics*, **74**, 175–195.

Malaspinas A (2015) Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Molecular Ecology*, **25**, 24–41.

Malaspinas A, Malaspinas O, Evans SN, Slatkin M (2012) Estimating allele age and selection coefficient from time-serial data. *Genetics*, **192**, 599–607.

Manel S, Perrier C, Pratlong M *et al.* (2015) Genomic resources and their influence on the detection of the signal of positive selection on genome scans: application to conservation biology. *Molecular Ecology*, **25**, 170–184.

McGee M, Neches R, Seehausen O (2015) Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. *Molecular Ecology*, **25**, 260–268.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomics scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Ormond L, Foll M, Ewing GB, Pfeifer SP, Jensen JD (2015) Inferring the age of a fixed beneficial allele. *Molecular Ecology*, **25**, 157–169.

Pennings PS (2013) HIV drug resistance: problems and perspectives. *Infections Disease Reports*, **5**, s1e5.

Qiu S, Bergero R, Guirao-Rico S *et al.* (2015) RAD mapping reveals an evolving, polymorphic and fuzzy boundary of a plant pseudoautosomal region. *Molecular Ecology*, **25**, 157–169.

Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M (2015) Shared and non-shared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, **25**, 287–305.

Renzette N, Gibson L, Jensen JD, Kowalik TF (2014) Human cytomegalovirus intrahost evolution – a new avenue for understanding and controlling herpesvirus infections. *Current Opinion in Virology*, **8**, 109–115.

Renzette N, Kowalik TF, Jensen JD (2015) On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. *Molecular Ecology*, **25**, 287–305.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko A, Messer P (2015) Evaluating the performance of selection scans to detect signatures of selective sweeps in dogs. *Molecular Ecology*, **25**, 342–356.

Schweizer RM, vonHoldt BM, Harrigan R *et al.* (2015a) Genetic subdivision and candidate genes under selection in North American gray wolves. *Molecular Ecology*, **25**, 342–356.

Schweizer RM, vonHoldt BM, Harrigan R *et al.* (2015b) Targeted capture and resequencing of 1040 genes reveals environmentally driven functional variation in gray wolves. *Molecular Ecology*, **25**, 357–379.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B*, **365**, 1245–1253.

Stephan W (2015) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, **25**, 79–88.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multi-locus genome scans for selection. *Genetics*, **175**, 737–750.

Tine M, Kuhl H, Gagnaire PA *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.

Vatsiou AI, Bazin E, Gaggiotti OE (2015) Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology*, **25**, 89–103.

Wenzel M, Douglas A, James M, Redpath S, Piertney S (2015) The role of parasite-driven selection in shaping landscape genomic structure in red grouse (Lagopus lagopus scotica). *Molecular Ecology*, **25**, 324–341.

Wilson B, Garud N, Feder A, Assaf Z, Pennings P (2015) The population genetics of drug resistance evolution in microbial pathogens. *Molecular Ecology*, **25**, 42–66.