MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2967–2977

NEWS AND VIEWS

MEETING REVIEW

# Conservation genomics of natural and managed populations: building a conceptual and practical framework

LAURA MARILYN BENESTAN,*[1] ANNE-LAURE FERCHAUD,*[1] PAUL A. HOHENLOHE,† BRITTANY A. GARNER,‡ § GAVIN J. P. NAYLOR,¶ ILIANA BRIGITTA BAUMS,** MICHAEL K. SCHWARTZ,† † JOANNA L. KELLEY‡ ‡ and GORDON LUIKART‡ §

*Departement de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec G1V 0A6, Canada; †Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA; ‡Flathead Lake Biological Station, Fish and Wildlife Genomic Group, Division of Biological Science, University of Montana, Missoula, MT 59812, USA; §Wildlife Program, Fish and Wildlife Genomic Group, College of Forestry and Conservation, University of Montana, Missoula, MT 59812, USA; ¶Hollings Marine Lab, College of Charleston and Medical, University of South Carolina, 331 Fort Johnson Rd., Charleston, SC 29412, USA; **Department of Biology, Pennsylvania State University, 208 Mueller Lab, University Park, PA 1680, USA; ††USDA Forest Service, National Genomics Center for Wildlife and Fish Conservation, 800 E. Beckwith Ave., Missoula, MT 59801, USA; ‡‡School of Biological Sciences, Washington State University, Pullman, WA 99164, USA*

## Abstract

**The boom of massive parallel sequencing (MPS) technology and its applications in conservation of natural and managed populations brings new opportunities and challenges to meet the scientific questions that can be addressed. Genomic conservation offers a wide range of approaches and analytical techniques, with their respective strengths and weaknesses that rely on several implicit assumptions. However, finding the most suitable approaches and analysis regarding our scientific question are often difficult and time-consuming. To address this gap, a recent workshop entitled 'ConGen 2015' was held at Montana University in order to bring together the knowledge accumulated in this field and to provide training in conceptual and practical aspects of data analysis applied to the field of conservation and evolutionary genomics. Here, we summarize the expertise yield by each instructor that has led us to consider the importance of keeping in mind the scientific question from sampling to management practices along with the selection of appropriate genomics tools and bioinformatics challenges.**

*Keywords*: conservation genomics, effective population size, genetic population structure, local adaptation, massively parallel sequencing, SNPs discovery and filtering

*Received 13 November 2015; revision received 12 March 2016; accepted 6 April 2016*

Conservation and evolutionary genetics are rapidly shifting from a genetic to a genomic perspective, where studies assess thousands of DNA markers in hundreds of individuals (Allendorf *et al.* 2010; Ouborg *et al.* 2010; Stapley *et al.* 2010; Narum *et al.* 2013; McMahon *et al.* 2014). The field has benefited from previous developments in population genomic studies of model organisms, especially in human (see examples reviewed in Allendorf *et al.* 2010). For instance, the Human Genome Project has reshaped life sciences and medicine by making life digital as a part of the 'big data' era (Tyler-Smith *et al.* 2015). A practical and conceptual framework for effective study design and analytical approaches is needed to help guide the new generation of population geneticists in using large-scale genomic data set. Indeed, integrating knowledge about many of the new molecular and computational tools available for analysing genomic data sets is crucial to answering questions in evolutionary and conservation biology. With knowledge of the tools available, researchers should use the underlying scientific question to guide all aspects of a conservation or evolutionary genomic study, from experimental design through data analysis (see Fig. 1).

To help educate population genomics researchers, 15 experts in the field of conservation genomics directed a 1-week workshop called 'ConGen 2015' (abbreviated from Conservation Genetics) at the University of Montana Flathead Lake Biological Station. This meeting review was written for everyone interested in population genomics, from graduate students to professors and resource managers. Here, we highlight the key topics and important take home messages discussed during the workshop, with an emphasis on the recent pertinent literature in this field. More particularly, we described (i) how to design a massively parallel sequencing (MPS) study for a model or non-model species, (ii) how to filter DNA sequence data from MPS data (i.e. extracting loci and/or SNPs on the basis of criteria) and (iii) to analyse MPS data using classic (e.g. clustering algorithms) and recent approaches (e.g. likelihood algorithms), within traditional or new pipelines (e.g. GALAXY). This overview will allow researchers to better

Correspondence: Anne-Laure Ferchaud, Fax: (141) 865 6717; E-mail: annelaureferchaud@gmail.com

[1]These authors contributed equally to this review.

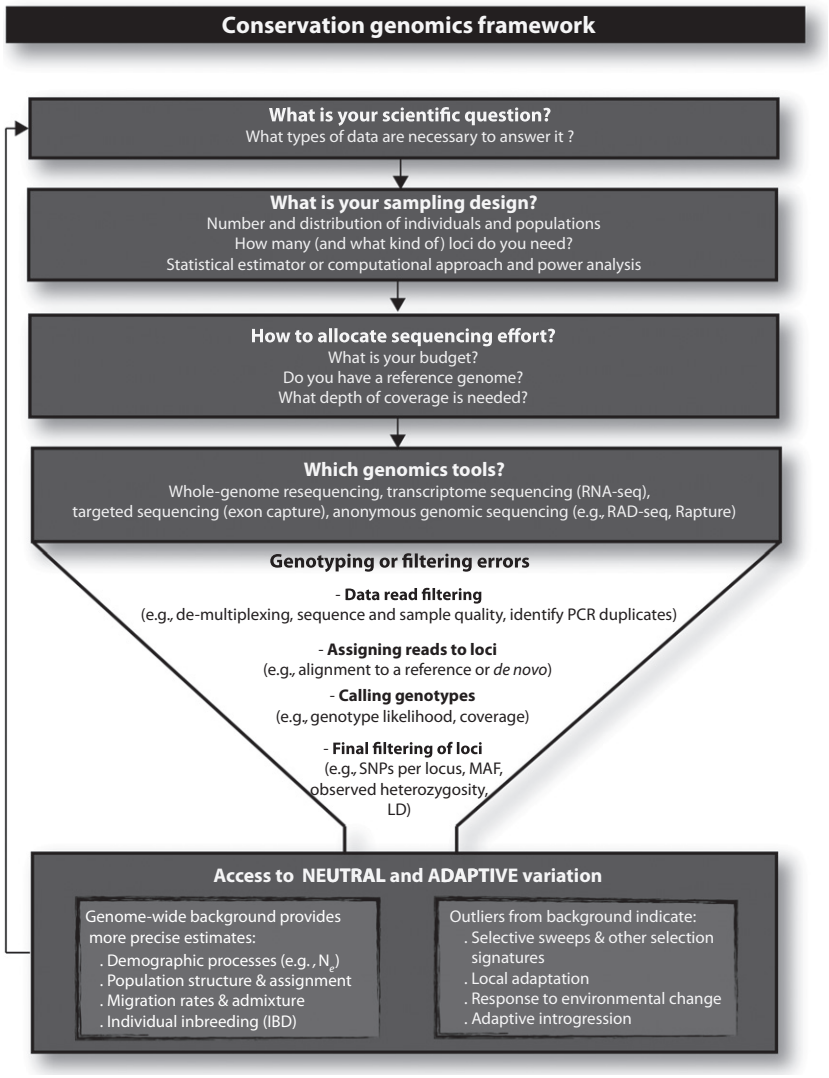



**Fig. 1** Practical framework with steps for designing a MPS study. All along the process researchers involved in MPS projects are faced with logistical trade-offs in order to accurately and efficiently answer their scientific question. The process is not straightforward and unidirectional but feedbacks and/or interactions are possible and common among all steps. 'What kind' of loci refers to characteristics such as loci in genes, linked loci or haplotypes (for genealogical information), mapped loci often required for QTL studies or runs of homozygosity, or long loci (e.g. long RAD contigs from paired-end reads). The 'distribution of populations' refers to the need to sample populations from different landscape locations or across environmental gradients when conducting landscape genetic or genomic studies. 'SNPs per locus' refers to the fact that researchers might use only one SNP per RAD locus (to ensure independent SNPs). Rapture, MAF, LD and IBD are acronyms for RAD-capture (Ali *et al*. 2015), Minor Allele Frequency, Linkage Disequilibrium, and Identity By Descent respectively. Note that SNP chips are an alternative genomic tool (not in this figure), often use MPS for SNP discovery.

understand some of the strengths and limits of recent molecular and computational approaches.

## Designing a MPS study: keeping in mind your biological question

One of the biggest differences in using MPS data vs. classical genetic data (e.g. microsatellites) is the amount of time spent on data analysis, with data production outpacing our ability to analyse it. As stated by ConGen instructor Paul Hohenlohe, it is not just about generating data. Conservation genomics offers an unprecedented genomic perspective by using large numbers of markers to simultaneously genotype putatively neutral and adaptive loci, thus offering glimpses into adaptive potential (Allendorf *et al*. 2010; Harrisson *et al*. 2014). Then, designing a MPS study requires the consideration of a large number of factors represented by Fig. 1 and recently reviewed by Andrews *et al*. (2016).

The most important starting point remains, 'What is your scientific/biological question?' This should determine how a researcher navigates all subsequent questions such as 'What is your sampling design and how should you allocate your budget among samples, populations, individuals, loci and depth of sequence coverage?' Question-driven rather than method-driven research allows researchers to not be limited by the methodological tools available, thus offering the flexibility and openness required to find the appropriate method that answer their question. For instance, recent simulation studies showed that a sampling design with geographically close populations (with recent gene flow and thus low genome-wide $F_{ST}$) across a selection gradient (environmentally distinct locations) had high power to detect local adaptation (Lotterhos & Whitlock 2015).

When do you need to sequence the entire genome vs. only genotype hundreds or thousands of loci to answer your question (Ellegren 2014)? For equivalent budgets, a large number of individuals can be genotyped at lower coverage, if you are interested in accurate estimates of population parameters (e.g. gene flow, $F_{ST}$-outlier loci),

whereas few samples could be genotyped at a higher coverage when you need to genotype individuals accurately (e.g. to assess individual inbreeding level). Do you have an a priori hypothesis about the features of your biological model system or species (e.g. the colonization history, the generation time, small and isolated vs. large and panmictic populations, dispersal capabilities, heterogeneous vs. homogeneous habitats) that could help you to predict the level of genetic diversity, the effect of genetic drift and the extent of the selective pressures?

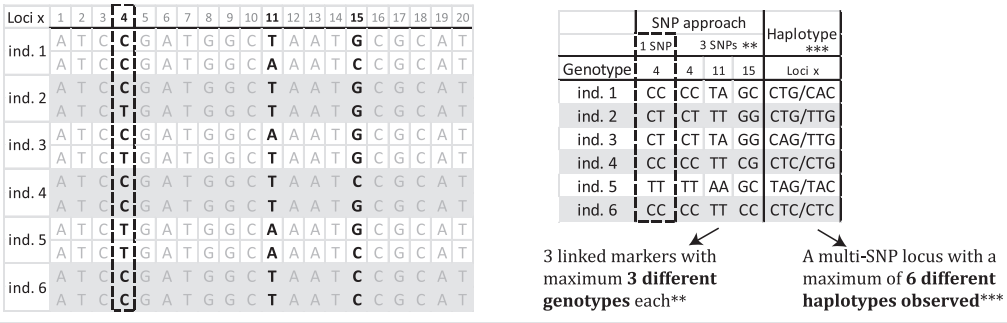
## Existing methods for MPS data analysis

### *Low-coverage genotyping methods and genotype likelihoods (Mike Miller)*

Novel Bayesian methods that aim to analyse efficiently low-coverage genomic data are blooming (Le & Durbin 2011; Yu & Sun 2013; Cantarel *et al.* 2014). Understanding theory behind the application of Bayesian models to low-coverage genomics data is crucial and begins with learning how to calculate genotype likelihoods from DNA sequences. Thus, Mike Miller instructed students how to calculate genotype likelihood based on sequencing errors, coverage and priors probabilities (i.e. uniform or Hardy–Weinberg Equilibrium model). He showed how these key factors could significantly affect genotype likelihood results and then many downstream analyses (Sims *et al.* 2014). These analyses may then suffer from SNP calling and genotype uncertainty, which lead to inaccurate demographic inferences (Nielsen *et al.* 2011). One way to overcome this bias could be to sample larger numbers of individuals at the expense of coverage depth in order to gather more information about population parameters, as suggested by Buerkle & Gompert (2013).

The importance of removing PCR duplicates (reads resulting from PCR clonal amplification of the same original DNA strand) was also underscored because of their potentially distorting influence on the calculation of genotype likelihoods (overconfidence into a genotype called only based on PCR duplicates) as suggested by Puritz *et al.* (2014b). PCR duplicates can easily be removed from paired-end restriction site associated DNA (RAD) sequencing data sets by identifying paired-end reads starting at identical position (Davey *et al.* 2013) and from genotyping-by-sequencing (GBS) data sets by using degenerate-base adaptors (Tin *et al.* 2015). Similarly, paralogs should be excluded from the analysis by detecting reads with high coverage, although genomic data sets often have high variance in coverage across loci (see Fig. 2; Malhis & Jones 2010). Finally, M. Miller also presented new computational approaches to detect genotyping errors, along with a new genotyping approach that combines RAD-seq with DNA

| Primary problem | Possible filtering solution | References |
|---|---|---|
| **Sequencing errors** | Ensuring accurate SNP calling: keeping SNPs with sufficient coverage, quality scores and genotype likelihood * | Davey *et al.* (2011), Kim *et al.* ( 2011), Nielsen *et al.* (2011), Catchen *et al.* (2013) , Marinier *et al.* (2015), Mastretta-Yanes *et al.* (2015), Andrews *et al.* (2016) and Laehnemann *et al.* (2016) |
| | Removing singletons | |
| | Correcting substitution errors to improve the quality of assemblies | |
| **Missing data** | Keeping SNPs genotyped in at least a certain percent of individuals and populations. This threshold will largely be influenced by the number of samples initially genotyped and the quality of data required for the research question | Hohenhole *et al.* (2010) and Benestan *et al.* (2015) |
| **Duplicated loci** | Keeping biallelic SNPs by individual for diploid species | Gayral *et al.* (2013), Pujolar *et al.* (2013)and Mandeville *et al.* (2015) Ferchaud & Hansen (2016) |
| | Removing loci with too high coverage (e.g., the mean plus 2*SD) | Bianco *et al.* (2014) and Ferchaud *et al.* (2014) |
| | Keeping SNPs with heterozygosity inferior to 0.5 | Hohenhole *et al.* (2010) |
| **Linkage desequilibrium (LD)** | Keeping only independent loci (required for many approaches), e.g., keeping only one SNP per loci, or using a cut-off of $r^2$ if a reference genome is available and physical position of loci is known | Larson *et al.* (2014), Baird (2015) and Waples *et al.* (in review) |
| **Hardy–Weinberg** | Keeping SNPs in Hardy–Weinberg proportions (HWP) in most of the populations (some populations could have sampling error that create spurious HWP). Nevertheless, SNPs out of HWP should not be removed if the main goal of the study is to detect outliers potentially under selection | Hess *et al.* (2012), Miller *et al.* (2012), Lexer *et al.* (2014), Lozier (2014), Benestan *et al.*(2015) and Waples (2015) |
| **Polymorphism** | Keeping informative SNPs based on a minor allele frequency (MAF) threshold (e.g., MAF > 0.05 at the population level if only informative SNPs are necessary to revel population structure or MAF > 0.0001 at global level for removing sequencing errors) | Roesti *et al.* (2012) |




Here is an example of 6 diploid individuals (ind.) genotyped at loci x, 20 bp long. Among this subset of individuals, 3 SNPs were discovered and accurately called (*), at nucleotide positions 4, 11 and 15. These 3 SNPs could be treated as three different markers (**). Several classic analysis would treat these 3 markers as independent whereas they are physically linked. To counteract this problem, researchers often retain only one SNP, for example the first one, here SNP 4 (see dashed line). However, in order to make use of all the 3 SNPs, the haplotype approach (combining the 3 SNPs in a single haplotype) could be used (***) when filtering and genotyping.

**Fig. 2** Roadmap for filtering reads from massively parallel sequencing (MPS).

capture arrays for low cost genotyping (Ali *et al.* 2016; Norgaard *et al.* in press). Calling genotypes based on their likelihoods can be easily performed with ANGSD (Korneliussen *et al.* 2014) and GATK (DePristo *et al.* 2011) programs.

*Mapping reads to a reference genome (Paul Hohenlohe)*

Aligning anonymous sequence reads against a reference genome assembly provides many advantages for filtering data (e.g. removing erroneous or clonal PCR duplicate reads) and identifying loci (Hand *et al.* 2015a). If a reference genome is unavailable for the focal species, P. Hohenlohe advised using well-assembled genomes from related taxa. Efforts such as the Genome 10K project (https://genome10k.soe.ucsc.edu/) and the i5k Insect Genome project (https://arthropodgenomes.org/wiki/i5K) are rapidly growing the number of taxa for which this is possible. The issue of how closely related is 'closely related enough' to be useful for alignment depends on details of the data set, such as the sequence read length and whether more conserved regions such as genes are targeted for sequencing. A poorly assembled reference genome can still be useful for assigning reads to loci and finding functional genes linked to candidate markers, even if it does not provide a complete physical map of the genome (e.g. Hand *et al.* 2015a,b).

Techniques like paired-end RAD sequencing (or exon capture) can also be used to build a set of contig sequences for nonmodel species, which then provide a reference for further population-level sequence data (Hohenlohe *et al.* 2013; Jones & Good 2016). When faced with limited resources, P. Hohenlohe cautioned against pool-sequencing (i.e. pooled sequencing of many individuals without barcode) because of the pitfalls associated with estimating allele frequencies (missing rare variants), identifying paralogs, distinguishing true alleles from sequencing error and hidden population structure. Whereas pooling showed promising results for accurate allele frequencies estimates (Futschik & Schlötterer 2010; Ferretti *et al.* 2013; Lynch *et al.* 2014), this approach is often less desirable than individual sequencing for a wide range of applications such as *Structure* analysis, parental assignment and genome scans (review in Cutler & Jensen 2010).

*STACKS workflow tutorial, STACKR package and GALAXY (Laura Benestan and Tiago Antao)*

There is a need for standardization and documentation of the many filtering and processing steps (Fig. 2) required to clean and use MPS data (e.g. by multiple researchers within a research group or the larger scientific community). Laura Benestan also emphasized that standardization helps ensure repeatability. The STACKS workflow tutorial created by Éric Normandeau for Louis Bernatchez's research group at Laval University was designed to facilitate, standardize and document (in a log file) each of many filtering and analysis steps in discovery and genotyping of putative SNP markers from GBS/RAD sequencing data using the STACKS program (Catchen *et al.* 2013). STACKS is a widely used pipeline for analysis RAD-seq data but other pipelines such as PYRAD (Eaton 2014), RADTOOLS (Baxter *et al.* 2011), GATK (McKenna *et al.* 2010), DDOCENT (Puritz *et al.* 2014a) could also be used for calling SNPs. More particularly, PYRAD, DDOCENT and more recently STACKS are promising workflow programs that can handle insertion–deletion polymorphism into the alignment of the reads.

The STACKS workflow uses universal tools, including custom scripts, to standardize and make repeatable all aspects of the pipeline, while also highlighting areas where the researcher should exercise caution in the choice of parameter values. The workflow is freely available on GITHUB (https://github.com/enormandeau/stacks_workflow). The included manual describes each step required for performing MPS analyses in STACKS from downloading and installing STACKS to filtering the results. Raw single-end data produced by Illumina or Ion Proton technology are supported.

Post-STACKS analyses and data filtering (Fig. 2) can be conducted with the R package STACKR (Gosselin & Bernatchez 2016). This package is freely available on GITHUB (https://github.com/thierrygosselin/stackr). STACKR contains several R functions that allow users to: (i) read and modify outputs from STACKS, (ii) filter markers based on coverage, genotype likelihood, number of individuals, number of populations, minor allele frequency, observed heterozygosity and inbreeding coefficient ($F_{IS}$), (iii) explore distributions of summary statistics and create publication-ready ggplot2 figures, (iv) impute missing data using a Random Forest algorithm and (v) export data sets in vcf, genepop, fstat files or as genind objects to be easily integrated into other R packages for population genomics analyses.

Tiago Antao also demonstrated web-based GALAXY software platform (https://galaxyproject.org), which could help with standardization of filtering and genotyping. GALAXY produces flow-chart diagrams (of filtering steps) and log files to help researchers reproducing and sharing complete 'pipeline' analysis with others. GALAXY is an interesting tool for data visualization as it could efficiently draw graphics (i.e. graphics of the distributions of quality scores) that allow users to explore and navigate their data. Running STACKS and related filtering approaches could be also done easily from this web-platform.

*The 'F-word': filtering (Jim Seeb)*

Genomics involves the genotyping of thousands of loci, genome-wide, to bring unprecedented resolution to problems of conservation planning (Allendorf *et al.* 2010; Shafer *et al.* 2015). However, this bright future for genomics hides the numerous *filtering* issues inherent to MPS data sets, which Jim Seeb referred to as the 'F-word'. For instance, merging data sets filtered using different parameters could create spurious results such as strong and significant (but false) $F_{ST}$-outlier values between differently filtered population samples. The lack of necessary details on the filtering steps in many of today's publications using MPS data

would affect the transparency and reproducibility of the results. This would contribute to the trend that most of the MPS studies cannot be accurately verified (Nekrutenko & Taylor 2012). To encourage scientist to publish and understand these important filtering steps, Fig. 2 reports some of the main filtering issues (associated with sequencing and assembling errors) that should be addressed in a MPS project. A complete and exhaustive publication will bring detailed recommendations and pipeline to conduct accurate filtering steps on MPS data in a forthcoming Population Genomics in R—Molecular Ecology Resources special issue. Identifying markers of interest through filtering steps could be done according to single SNP or haplotype approach, the latter being a possible alternative to overcome issues regarding linkage disequilibrium (LD; Fig. 2). Nevertheless, it is important to keep in mind that the appropriate level of filtering will always depend on the scientific question and the available data set (Andrews *et al.* 2016).

In addition, RAD locus discovery and genotyping is often inhibited by the existence of duplicated genes and genomic regions (Allendorf *et al.* 2015; Andrews *et al.* 2016). Gene duplication occurs because of segmental duplication (unequal crossing over) or whole genome duplication (Amores *et al.* 2011). Loci can be assayed in duplicated regions by constructing linkage maps and genotyping with SNP chips or potentially with very deep GBS coverage (Waples *et al.* 2015). Distinguishing and including paralogous loci on the linkage map will allow researchers to circumvent the issues of producing an incomplete picture of the genome (Brieuc *et al.* 2014; Kodama *et al.* 2014) and introducing bias into genetic estimates parameters (Meirmans & Van Tienderen 2013).

### *Structure program insights and tips (Jonathan Pritchard)*

Jonathan Pritchard provided an overview and practical advice about the application of the program *Structure* (Pritchard *et al.* 2000). J. Pritchard explained that it is often unrealistic to expect that there is one 'true' $K$ that is best for modelling a particular data set. Through simulations, Kalinowski (2010) showed that sometimes *Structure* clustered individuals in unpredictable ways, which is because the *Structure* model is a cartoon (simplification) of more complicated natural population. Therefore, viewing and reporting plots for multiple $K$-values is an important step (Gilbert *et al.* 2012) because different values of $K$ can give insights into different levels of structure. Similarly, the selection of the optimal $K$ is not an exhaustive procedure and has to be done with regard to the biology and the history of populations studied (Kalinowski 2010). For instance, the optimal number of clusters ($K$) found by *Structure* or subsequent analysis (e.g. Evanno *et al.* 2005) may have no biological reality and could result from a context of isolation by distance, where *Structure* tends to overestimate genetic structure (Frantz *et al.* 2009). In the same vein, a recent simulation study showed that unbalanced sample size leads to wrong demographic inferences where smaller samples tend to be merged together (Puechmaille 2016). To overcome these issues, alternative methods such as principal component analysis or evolutionary trees could be tested in regard to *Structure* analysis (Jombart *et al.* 2010; Kalinowski 2010; Kanno *et al.* 2011; Benestan *et al.* 2015).

Reviewers often request extremely long *Structure* runs (millions of iterations). J. Pritchard claimed that is generally unnecessary and wasteful of researcher time (and carbon footprint). For most data sets, he would recommend to do about 10 000 steps, but multiple times to assess robustness and convergence of the results. *Structure* tends to converge fairly quickly, but the program does not do a great job of exploring between local peaks in parameter space of the posterior distribution. Therefore, for an exploratory analysis, it would be more efficient to spend the computation time on independent runs (which have a good chance of finding distinct modes) than doing extremely long runs where the algorithm will be simply wandering around within one mode. Nevertheless, a certain minimum burn-in and run length helps overcome the stochasticity of the Monte Carlo approach, as recommended by Gilbert *et al.* (2012).

### *Improving our detection of local adaptation (Lisa Seeb)*

Understanding genetic basis of local adaptation is one of the most exciting potential contributions of genomics to conservation biology (Allendorf *et al.* 2010). The most widely used methods for detecting evidence of selection are genome scan approaches based on differentiation ($F_{ST}$) outlier tests originally developed by Lewontin & Krakauer (1973), then refined more recently by Beaumont & Nichols (1996) and others (Beaumont & Balding (2004), Foll & Gaggiotti 2008). Several programs were designed to perform genome-wide outlier scan analysis such as LOSITAN, which is based on a stochastic $F_{ST}$ null distribution (Antao *et al.* 2008), Arlequin (Excoffier & Lischer 2010) which includes a hierarchical model, and BayeScan (Foll & Gaggiotti 2008) which is based on a Bayesian $F_{ST}$ distribution. Since each method has its drawbacks, requiring outliers (candidate adaptive genes) to be identified in multiple methods can help to reduce the incidence of false positives (Narum & Hess 2011; Villemereuil *et al.* 2014; François *et al.* 2016) because these different methods may tend to agree more on true positives than on false positives. Lisa Seeb described the work of DeMita *et al.* (2013) who investigated the robustness of eight methods to detect loci potentially under selection according to eight demographic scenarios along an environmental gradient. Their work showed that whereas genotype–environment correlation methods have more power to detect signal of selection than genome scans, these methods were more prone to false positives when assessing these associations.

The importance of incorporating neutral genetic structure into genotype–environment correlation methods has led to the emergence of two recent software packages: BAYESCENV (Villemereuil *et al.* 2014) and LFMM (Frichot *et al.* 2013). However, as well as using suitable methods, L. Seeb

emphasized that an appropriate sampling design is crucial to test for evidence of local adaptation. For instance, analysing sets of independent populations ('replicates') across similar environmental gradients helped Larson *et al.* (2014) to find signals of selection in Chinook salmon. In addition, mapping outliers to find chromosomal islands of divergence can help to identify functional genes involved in local adaptation. We advise scientists interested in the utilization of environmental association analysis in genomics to read Hand *et al.* (2015b), Rellstab *et al.* (2015) or Van Heerwaarden *et al.* (2015). Researchers should be aware that new and improved tests as well as evaluations of tests are published frequently (e.g. see Foll *et al.* 2014; Whitlock & Lotterhos 2015; Jensen *et al.* 2016).

## The use of genomics for management decisions

### *Effective population size ($N_e$) estimation (Robin Waples)*

Robin Waples taught concepts of the effective population size by using an analogy of a lottery. Imagine the ability of parents to produce viable offspring for the next generation depends on a lottery system. In a Wright-Fisher (ideal) population, everyone has the same number of tickets, and sampling is with replacement. In real populations, different individuals have different numbers of lottery tickets because some of them will reproduce more than others, and hence they have different probabilities of being parents, thus reducing $N_e$ compared to census size. He enumerated the different methods that can be used to estimate contemporary $N_e$: temporal methods, LD methods, approximate Bayesian computation (ABC) methods, and other single estimators based on heterozygote excess (Pudovkin *et al.* 1996), molecular co-ancestry (Nomura 2008), and sibship analysis (Wang 2009). ConGen participants were also reminded that these methods make several important assumptions: no migration, no selection, mutation is unimportant, discrete generations, random sampling of an entire generation and loci not physically linked.

R. Waples mentioned that genetic estimates of either contemporary or long-term $N_e$ benefit from the proliferation of the number and types of markers, but this also introduces challenges, largely because of a) LD, which is unavoidable when large numbers of markers have to be packaged into a small number of chromosomes, and b) pseudo-replication because of linkage, markers are not independent, so adding more and more loci does not increase precision as fast as it would under complete independence. LD is predicted to be the next big issue in dealing with genomics data since multilocus sampling improves whereas classic analyses such as $N_e$ estimation, genome scan and clustering algorithms treated the loci as independent (Baird 2015). Kemppainen *et al.* (2015) present a useful exploratory tool (named LDna) able to give a global overview of LD associated with diverse evolutionary phenomena and identify potentially related loci. Based on simulations, (R.K. Waples, W.A. Larson, R.S. Waples, in review) showed that more loci do not increase the fraction that is physically linked, since most random pairs of loci are not linked. If linked loci downwardly bias $N_e$ estimates (Larson *et al.* 2014), the bias from ignoring linkage is less severe when the number of chromosomes is large. Finally, strategies that filter out a locus in outlier pairs of loci are only partially effective and a bias correction factor based on the number of chromosomes is probably more effective (R.K. Waples, W.A. Larson, R.S. Waples, in review). Videos recording R. Waples' $N_e$ lecture can be viewed at https://www.youtube.com/watch?v=ErhACWXRLss and https://www.youtube.com/watch?v=N3JbKZbKO5w

### *Defining conservation units: ESUs and MUs (Robin Waples)*

Integrating genomic data into management can be challenging in practice. For instance, there is no single best or correct way to answer the questions 'what is a population' and 'how to identify the suitable conservation unit' (e.g. ESU, MU, etc.) because the definitions of these terms can be vague, not quantitative, and depend on the management objective (Waples & Gagiotti 2006). Since several 'population' concepts can be found in literature (Fraser & Bernatchez 2001), R. Waples suggested choosing the population concept (ESU, MU, etc.) that is appropriate to the objective(s) of each study. One way to detect the number of populations is to test for a statistically significant genetic differentiation. Statistical power is influenced by (i) population differences (effect size) and (ii) data richness (numbers of individuals, number of samples, number of loci and alleles). Then, important biological differences might be missed if data are limited (low power). On the other hand, statistical significance does not guarantee biological significance, especially when large amounts of data are available (i.e. high power detects even trivial differences, see Palsboll *et al.* (2007). This failing should be a major concern in the age of genomics. Also, standard statistical tests usually do not properly answer the question 'Is it different enough?' because they reject only the null hypothesis of no differentiation (panmixia).

Another way to detect population structure and identify population units is to use Bayesian clustering methods such as *Structure* (Pritchard *et al.* 2000), Baps (Corander *et al.* 2004) and Admixture (Alexander *et al.* 2009), but these methods may have reduced power with high gene flow species (Jombart *et al.* 2010; Kanno *et al.* 2011; Benestan *et al.* 2015). Nevertheless, absence of genetic differentiation at neutral markers does not mean absence of adaptive differences (Allendorf *et al.* 2010). Therefore, using markers influenced by selection could be a promising research avenue for delineating important conservation units (see study conducted on herring by Limborg *et al.* (2012a) for an example), particularly in high gene flow species (Gagnaire *et al.* 2015). However, a pattern of adaptive divergence may not necessarily match the neutral pattern (e.g. when one adaptive group overlaps two neutral ones) as the processes affecting adaptive and neutral genetic markers are different. Then, combining neutral and adaptive markers in a hierarchical approach to define conservation units, as suggested by Funk *et al.* (2012) may encounter practical

issues in delineating conservation units. Yet, few studies already used information on adaptive differentiation to improve conservation decisions (Limborg *et al.* 2012a,b; Bourret *et al.* 2013; Larson *et al.* 2014). Nevertheless, given the considerable proportion of false positive in outliers detection (DeMita *et al.* 2013; François *et al.* 2016), it is crucial to complement the pattern of adaptive divergence arising from genomics data with ecological, phenotypic and environmental data. Further research is needed to assess this issue in the future.

*Adaptive genomics as a first step (Michael Schwartz)*

Michael Schwartz, Director of the National Genomics Laboratory for Wildlife and Fish Conservation (in Montana), led a discussion that focused on the extent of direct use of genomic data in conservation and natural resource management (Shafer *et al.* 2015; Garner *et al.* 2016). One side of the debate suggests that genomics has advanced fish and wildlife conservation by increasing the number of markers assayed, but has failed to live up to its promise to elucidate the genetic basis for adaptation in a way that can be used by managers (Shafer *et al.* 2015). The other side notes that genomics is currently being used by management agencies in a variety of taxa, but that the nonacademic nature of some laboratories applying genomics to conservation can lead to a lag in publishing in academic journals. Participants and instructors suggested reasons for a potential gap between genomics and direct management application, most noticeably, that of cost and a lack of familiarity (e.g. some managers are more comfortable with the vocabulary or concepts surrounding microsatellite data (and data analysis) than with novel genomic techniques in decision-making).

The group then discussed how to avoid false positives when identifying outliers by applying statistical correction for multiple testing such as Bonferroni or false discovery rate (FDR) correction (Narum 2006). Power to detect true outliers seems to be highly dependent on sampling and statistic test used, whether it controls or not for population structure (Lotterhos & Whitlock 2015). There was an overall recognition by those using genomic approaches that careful identification of outlier-loci was a first step. Then, additional empirical evidence showing the functional importance of the outlier in a relevant ecological context is a mandatory step to confirming that these genes is a target of selection. For that purpose, common garden and transplant experiments, thought difficult to perform in most of the nonmodel species, would be required (Barrett & Hoekstra 2011). When such experiments are not possible, the observation of the same outlier-loci in multiple independent population sets can help confirm local adaptation signatures (Bradbury *et al.* 2010; Laporte *et al.* 2016).

*RNA-sequencing for management decisions (Joanna Kelley)*

Studying gene expression differences among individuals and populations can provide insight into (i) the molecular basis of phenotypic differentiation, (ii) variation in response to environmental conditions, disease, *etc*., and (iii) management decisions regarding how and where to manage or transplant populations. For example, Barshis *et al.* (2013) compared transcriptome-wide gene expression (*via* RNA-sequencing (RNA-seq) using Illumina sequencing technology) among conspecific thermally resilient corals to identify the molecular pathways contributing to coral resilience. RNA-seq can be also used directly in management decisions. Narum & Campbell (2015) detected differential transcriptomic response to heat stress among ecologically divergent populations of redband trout, which will likely influence future conservation including avoiding translocations between the divergent populations.

The approaches to measuring gene expression including limited gene studies (qPCR and Northern blots) and transcriptome level studies (microarrays and RNA-seq, see Kodama *et al.* 2014). There are two RNA enrichment techniques, polyA+ selection and ribosomal depletion, to remove the highly abundant ribosomal RNAs from the pool of total RNA, prior to library preparation (Cui *et al.* 2010). Both methods are efficient and their use depends largely on financial resources and whether researchers are interested in coding transcripts or transcripts that may be regulatory (for example, long noncoding RNAs). Directional RNA-seq libraries are recommended to find sense and antisense transcripts, which may be relevant for regulatory processes. Additionally, reference bias was briefly discussed. In that context, combining all data sets and generating *de novo* transcriptome assemblies carefully would be very useful in any comparative analysis. She discussed the pipeline and analyses described in Kelley *et al.* (2012). Finally, Joanna Kelley referred to the Simple Fool's Guide from Stephen Palumbi's lab (Wit *et al.* 2012) for calling single nucleotide polymorphisms (SNPs) based on RNA-seq data.

*General advice from instructors*

The common advice given by each instructor was to keep the scientific question of the study in mind at each step from the initial study design to publication. There is no single pipeline for analysing all (or even any two) MPS data sets, and thus the analysis of MPS data requires an investment in scripting and writing computer code (http://korflab.ucdavis.edu/Unix_and_Perl/; Antao 2015). In addition, students and professionals alike can gain a competitive edge in an increasingly competitive job market by understanding new computational methods and being comfortable operating in some kind of programming language. These skills are particularly desirable now as the sheer size of genomic data sets alone demands computational and scripting or coding prowess.

Robin Waples mentioned the importance of understanding all steps in the process from data production to genotype analysis (by filtering data) to avoid conducting analyses that are not adequate and could lead to data misinterpretation. Instructor Tiago Antao disagreed somewhat by suggesting that one single person cannot expertly understand every single step of a genomics project;

however, instructor and ConGen coordinator Gordon Luikart addressed these concerns by recommending close collaboration with people who are experts in some of the different steps of the process.

As a career advice, Jonathan Pritchard recommended early-career researchers to submit manuscripts online at the ArXiv or bioRxiv web page (e.g. Ali *et al*. 2015), so they can show them on their CV when applying for jobs and to perhaps get early feedback (edits) from the scientific community. Submission to bioRxiv could also advance the field of conservation genomics and ecology faster than by waiting until the paper is actually accepted by a traditional journal. Many journals no longer have an embargo and allow early online publication.

In summary, the growing potential for current application of genetic and genomics approaches to conservation is exciting. However, it also requires increasing the development of next-generation approaches and great caution when using massive parallel sequencing. Along with this meeting review, Figs 1 and 2 provide a conservation genomics framework and highlight important issues arising from the massive scale data sets.

## Acknowledgements

LB and ALF received travel grants from RAQ (Réseau Aquaculture du Quebec) to attend the workshop and helpful supported by Louis Bernatchez and the Canadian Research Chair in Genomics and Conservation of Aquatic Resources. We are grateful to Robin Waples and Jonathan Pritchard for helpful comments regarding the manuscript as well as four Anonymous reviewers for their useful comments. We thank the ConGen instructors: Mike Miller, Fred Allendorf, Jim and Lisa Seeb, Tiago Antao, Jeff Good, Brian Hand, Tabitha Graves, Ryan Kovach, Brice Sarver. GL and ConGen were supported in part by grants from the US National Science Foundation (DEB-1258203), NASA (NNX14AB84G), and time and advice from Michelle Quinn at the University of Montana Montana's School of Extended & Lifelong Learning.

## References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Ali OA, O'Rourke SM, Amish SJ *et al*. (2015) RAD capture (Rapture): flexible and efficient sequence-based genotyping. *BioRxiv*, doi:http://dx.doi.org/10.1101/028837. Pre-print.

Ali O, O'Rourke S, Amish S *et al*. (2016) RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics*, **202**, 389–400.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Allendorf FW, Bassham S, Cresko W *et al*. (2015) Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, **106**, 217–227.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

Antao T (2015) *Bioinformatics with Python Cookbook*. Packt Publishing Ltd., Birmingham.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323.

Baird SJE (2015) Exploring linkage disequilibrium. *Molecular Ecology Resources*, **15**, 1017–1019.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barshis D, Ladner J, Oliver T *et al*. (2013) Genomic basis for coral resilience to climate change. *Proceedings of the National Academy of Sciences of the USA*, **110**, 1387–1392.

Baxter S, Davey J, Johnston J *et al*. (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, **6**, e19315.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont M, Nichols R (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **263**, 1619–1626.

Benestan L, Gosselin T, Perrier C *et al*. (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, **24**, 3299–3315.

Bianco L, Cestaro A, Sargent DJ *et al*. (2014) Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (*Malus* × *domestica* Borkh). *PLoS ONE*, **9**, e110377.

Bourret V, Dionne M, Kent M, Lien S, Bernatchez L (2013) Landscape genomics in Atlantic salmon (*Salmo salar*): searching for gene-environment interactions driving local adaptation. *Evolution*, **67**, 3469–3487.

Bradbury I, Hubert S, Higgins B *et al*. (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 3725–3734.

Brieuc M, Waters C, Seeb J, Naish K (2014) A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3: Genes|Genomes|Genetics*, **4**, 447–460.

Buerkle C, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Cantarel BL, Weaver D, McNeill N *et al*. (2014) BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics*, **15**, 1.

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Cui P, Lin Q, Ding F *et al*. (2010) A comparison between ribominus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, **96**, 259–265.

Cutler DJ, Jensen JD (2010) To pool, or not to pool? *Genetics*, **186**, 41–43.

Davey J, Hohenlohe P, Etter P, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davey JW, Cezard T, Fuentes Utrilla P *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

DeMita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer H (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Ferchaud A, Hansen MM (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology*, **25**, 238–259.

Ferchaud A, Pedersen S, Bekkevold D *et al.* (2014) A low-density SNP array for analyzing differential selection in freshwater and marine populations of threespine stickleback (*Gasterosteus aculeatus*). *BMC Genomics*, **15**, 867.

Ferretti L, Ramos Onsins SE, Pérez Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L (2014) Widespread signals of convergent adaptation to high altitude in Asia and America. *American Journal of Human Genetics*, **95**, 394–407.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology*, **10**, 2741–2752.

Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Funk W, Mckay J, Hohenlohe P, Allendorf F (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Gagnaire P-A, Broquet T, Aurelle D *et al.* (2015) Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, **8**, 769–786.

Garner BA, Hand BK, Amish SJ *et al.* (2016) Genomics in conservation: case studies and bridging the gap between data and application. *Trends in Ecology & Evolution*, **31**, 81–83.

Gilbert KJ, Andrew RL, Bock DG *et al.* (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular Ecology*, **21**, 4925–4930.

Gosselin T, Bernatchez L (2016). stackr: GBS/RAD Data Exploration, Manipulation and Visualization using R. R package version 0.2.1. https://github.com/thierrygosselin/stackr. doi: 10.5281/zenodo.19647

Hand BK, Hether TD, Kovach RP, Muhlfeld CC (2015a) Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, **61**, 146–154.

Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G (2015b) Landscape community genomics: understanding eco-evolutionary processes in complex environments. *Trends in Ecology & Evolution*, **30**, 161–168.

Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008–1025.

Hess J, Campbell N, Close D, Docker M, Narum S (2012) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.

Hohenlohe P, Bassham S, Etter P, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Day MD, Amish SJ *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, **22**, 3002–3013.

Jensen JJ, Foll M, Bernatchez L (2016) The past, present and future of genomic scans for selection. *Molecular Ecology*, **25**, 1–4.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kalinowski S (2010) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity*, **106**, 625–632.

Kanno Y, Vokoun JC, Letcher BH (2011) Fine-scale population structure and riverscape genetics of brook trout (*Salvelinus fontinalis*) distributed continuously along headwater channel networks. *Molecular Ecology*, **20**, 3711–3729.

Kelley JL, Passow CN, Plath M *et al.* (2012) Genomic resources for a model in adaptation and speciation research: characterization of the *Poecilia mexicana* transcriptome. *BMC Genomics*, **13**, 652.

Kemppainen P, Knight CG, Sarma DK *et al.* (2015) Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, **15**, 1031–1045. g3.114.012294

Kim SY, Lohmueller KE, Albrechtsen A *et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, **12**, 231.

Kodama M, Brieuc M, Devlin R, Hard J, Naish K (2014) Comparative mapping between Coho salmon (*Oncorhynchus kisutch*) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. *G3: Genes|Genomes|Genetics*, **4**, 1717–1730.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.

Laehnemann D, Borkhardt A, McHardy AC (2016) Denoising DNA deep sequencing data – high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, **17**, 154–179.

Laporte M, Pavey SA, Rougeux C *et al.* (2016) RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, **25**, 219–237.

Larson W, Seeb L, Everett M *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**, 355–369.

Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, **21**, 952–960.

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

Lexer C, Wüest RO, Mangii S *et al.* (2014) Genomics of the divergence continuum in an African plant biodiversity hotspot, I: drivers of population divergence in *Restio capensis* (Restionaceae). *Molecular Ecology*, **23**, 4373–4386.

Limborg MT, Helyar SJ, De Bruyn M *et al.* (2012a) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, **21**, 3686–3703.

Limborg M, Blankenship S, Young S *et al.* (2012b) Signatures of natural selection among lineages and habitats in *Oncorhynchus mykiss*. *Ecology and Evolution*, **2**, 1–18.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Lozier J (2014) Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in North American bumble bees using RAD sequencing. *Molecular Ecology*, **23**, 788–801.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-genetic inference from pooled-sequencing data. *Genome Biology and Evolution*, **6**, 1210–1218.

Malhis N, Jones S (2010) High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, **26**, 1029–1035.

Marinier E, Brown D, McConkey B (2015) Pollux: platform independent error correction of single and mixed genomes. *BMC Bioinformatics*, **16**, 10.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McMahon B, Teeling E, Höglund J (2014) How and why should we implement genomics into conservation? *Evolutionary Applications*, **7**, 999–1007.

Meirmans P, Van Tienderen P (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity*, **110**, 131–137.

Miller M, Brunelli J, Wheeler P *et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, **21**, 237–249.

Narum SR (2006) Beyond Bonferroni: less conservative analyses for conservation genetics. *Conservation Genetics*, **7**, 783–787.

Narum S, Campbell N (2015) Transcriptomic response to heat stress among ecologically divergent populations of redband trout. *BMC Genomics*, **16**, 103.

Narum S, Hess J (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, **13**, 667–672.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications*, **1**, 462–474.

Norgaard ZK, Ali OA, Amish SJ *et al.* (in press) Caution and skepticism are critical when analyzing massively parallel sequencing data.

Ouborg NJ, Angeloni F, Vergeer P (2010) An essay on the necessity and feasibility of conservation genomics. *Conservation Genetics*, **11**, 643–653.

Palsboll P, Bérubé M, Allendorf F (2007) Identification of management units using population genetic data. *Trends in Ecology & Evolution*, **22**, 11–16.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144**, 383–387.

Puechmaille S (2016) The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources*, **16**, 608–627.

Pujolar J, Jacobsen M, Frydenberg J *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, **13**, 706–714.

Puritz JB, Hollenbeck CM, Gold JR (2014a) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431.

Puritz JB, Matz MV, Toonen RJ *et al.* (2014b) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

Roesti M, Salzburger W, Berner D (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, **12**, 94.

Shafer ABA, Wolf JBW, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30**, 78–87.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2015) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, **15**, 329–336.

Tyler-Smith C, Yang H, Landweber L *et al.* (2015) Where next for genetics and genomics? *PLOS Biology*, **13**, e1002216.

Van Heerwaarden J, van Zanten M, Kruijer W (2015) Genome-wide association analysis of adaptation using environmentally predicted traits. *Plos Genetics*, **11**, e1005594.

Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti O (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Wang J (2009) A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology*, **18**, 2148–2164.

Waples RS (2015) Testing for Hardy-Weinberg proportions: have we lost the plot? *Journal of Heredity*, **106**, 1–19.

Waples RS, Gagiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Waples RK, Seeb LW, Seeb JE (2015) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*, **16**, 17–28.

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST*. *American Naturalist*, **186**, S24–S36.

Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Yu X, Sun S (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, **14**, 1.

---

A.L.F., L.B. and G.L. conceived and designed the manuscript. G.L organized the workshop. J.K., M.S. and P.H. were instructors during the workshop, and they wrote the paragraph corresponding to their field of expertise. B.G. realized videos and provided the related links provided in the text. A.L.F. and L.B. elaborated the figures. All co-authors critically revised the manuscript and approved the final version to be published.

---

doi: 10.1111/mec.13647