MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 1661–1667

NEWS AND VIEWS

MEETING REVIEW

Recent novel approaches for population genomics data analysis

KIMBERLY R. ANDREWS* and GORDON LUIKART†

**School of Biological & Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK; †Flathead Lake Biological Station, Fish and Wildlife Genomics Group, University of Montana, Polson, MT 59860, USA*

Next-generation sequencing (NGS) technology is revolutionizing the fields of population genetics, molecular ecology and conservation biology. But it can be challenging for researchers to learn the new and rapidly evolving techniques required to use NGS data. A recent workshop entitled 'Population Genomic Data Analysis' was held to provide training in conceptual and practical aspects of data production and analysis for population genomics, with an emphasis on NGS data analysis. This workshop brought together 16 instructors who were experts in the field of population genomics and 31 student participants. Instructors provided helpful and often entertaining advice regarding how to choose and use a NGS method for a given research question, and regarding critical aspects of NGS data production and analysis such as library preparation, filtering to remove sequencing errors and outlier loci, and genotype calling. In addition, instructors provided general advice about how to approach population genomics data analysis and how to build a career in science. The overarching messages of the workshop were that NGS data analysis should be approached with a keen understanding of the theoretical models underlying the analyses, and with analyses tailored to each research question and project. When analysed carefully, NGS data provide extremely powerful tools for answering crucial questions in disciplines ranging from evolution and ecology to conservation and agriculture, including questions that could not be answered prior to the development of NGS technology.

Keywords: bioinformatics, computational analyses, genomic course, next-generation sequencing, statistical genetics, workshop

Received 29 October 2013; revised 25 January 2014; accepted 29 January 2014

Correspondence: Kimberly R. Andrews, Fax: +44 191 33 41201;
E-mail: kimandrews@gmail.com

The challenge of learning NGS data analysis skills

It is exciting that population geneticists can now use several orders of magnitude more genetic markers, thanks to next-generation sequencing (NGS) technologies. However, it can be problematic that this technology requires researchers to learn many new and rapidly evolving laboratory and bioinformatic methods; keeping up with these rapidly changing methods can present a major challenge.

To address this challenge, a workshop entitled 'Population Genomic Data Analysis' was held from 2 September to 8 September 2013 at the University of Montana Biological Station on Flathead Lake. This workshop was the most recent of a series of six 'ConGen' (abbreviated from 'Conservation Genetics') workshops held since 2006, and brought together 16 instructors and 31 student participants from eleven countries. Instructors were leading academic and government researchers in the fields of population and conservation genomics. The participants were graduate students, postdocs, academic faculty, government personnel and other researchers. The workshop provided training in conceptual and practical aspects of data production and analysis for population genomics, with an emphasis on NGS data analysis. The format consisted of lectures, discussions and hands-on activities regarding experimental design and statistical methodologies, as well as designated time for students to receive one-on-one advice from instructors regarding specific research projects and analysis of their own data.

Here, we describe the major themes and lessons from the workshop lectures and discussions, with the goal of helping students and researchers improve their careers and their knowledge and skills in NGS data production and analysis. We also hope this article will promote more workshops and courses such as this ConGen course.

The NGS revolution in population genomics

Fred Allendorf (University of Montana) and Gordon Luikart (University of Montana) started off the workshop with overview talks regarding the exciting 'information explosion' that NGS is bringing to the fields of population and conservation genetics. However, Allendorf warned that even though we have masses of data and many available analytical software programs, we cannot conduct empirical population genetics without a strong understanding of the population genetics theory and models underlying the computational analyses conducted by these software packages. He illustrated the three essential components of empirical population genetics using the analogy of a stool supported by three legs: (i) an understanding of population genetics theory, (ii) appropriate data collection to address the research questions and (iii) appropriate implementation of statistical analyses (Allendorf *et al.* 2013).

Allendorf recommended two books that 'everyone conducting population genetic data analysis should own' to ensure their solid understanding of population genetic theory: An Introduction to Population Genetics Theory (Crow & Kimura 1970) (now reprinted and available at Amazon.com for ~$50); and Joe Felsenstein's Theoretical Evolutionary Genetics (unpublished book, available online for free at http://evolution.genetics.washington.edu/pgbook/pgbook.html). Allendorf, and later Jim Seeb (University of Washington), warned against suffering from 'automation addiction,' in which we use available computer programs without really understanding the assumptions underlying the statistical analyses, because this approach can lead to errors in data interpretation. A memorable and helpful analogy was provided in a PowerPoint slide of a plane crash caused by a pilot who did not know how to respond when the plane's autopilot stalled (http://abcnews.go.com/Technology/automation-addiction-pilots-forgetting-fly/story?id=14417730).

Throughout the workshop many of the instructors reiterated the message regarding the importance of understanding the theory underlying population genetics data analysis. For example, Luikart quoted John McCutcheon (University of Montana) by saying 'There's no such thing as a data analysis pipeline,' referring to the idea that researchers should tailor their analyses to their data set and research question, rather than simply using a set of analytical methods or scripts developed for another research project. This quote was subsequently repeated several times and eventually became a mantra for the workshop.

NGS methods

To help researchers understand the utility of NGS for answering population genomics questions, the workshop instructors provided advice about population genomics data analysis for each of the currently most popular NGS methods, as described below.

Restriction-site associated DNA sequencing (RADseq)

Mike Miller (University of California, Davis) and Paul Hohenlohe (University of Idaho) taught about RADseq data production and analysis. The use of RADseq is exploding in the field of population genetics, as evidenced by the fact that over 50% of the workshop participants had their own RADseq data or were planning to soon have such data. RADseq can be used to discover and genotype thousands of SNPs throughout the genome by sequencing short regions adjacent to restriction enzyme cut sites (Miller *et al.* 2007; Baird *et al.* 2008).

RADseq requires no prior genome information and therefore can be used in any species. However, having at least a preliminary draft genome from the study species can improve several aspects of RADseq population genetics data analysis, such as the identification of paralogs (see 'Filtering paralogs' section below). In addition, a reference genome provides access to information regarding the genomic positions of loci, and therefore facilitates the identification of loci that are close together in the genome and likely non-independent due to linkage. Identification of linked loci can also facilitate one of the most exciting capabilities of RADseq population genetic data analysis, which is the identification of linkage groups or genomic regions under divergent or balancing selection (e.g. Hohenlohe *et al.* 2010).

One challenge of RADseq data analysis is the identification of PCR duplicates, which can be more complicated for RADseq than some other NGS methods (see 'Filtering PCR duplicates' section below). Fortunately, identification of PCR duplicates for RADseq can be effectively accomplished when using a library prep method which has a random sheering step, along with generation of pair-end sequence reads (PE-RADseq) (Davey *et al.* 2013; Hohenlohe *et al.* 2013). Alternatively, there is potential to avoid PCR duplicates using PCR-free RADseq, e.g. using PCR-free Illumina library prep kits (Toonen *et al.* 2013).

Exon capture (DNA enrichment)

Jeff Good (University of Montana) and Steve Amish (University of Montana) provided overviews and practical advice for another increasingly popular NGS method called exon capture (Hodges *et al.* 2007). This method involves sequencing protein-coding genomic regions of interest that have been 'captured' by direct hybridization to oligonucleotide probes, or 'baits,' which are typically 60- to 120-bp long (reviewed in Mamanova *et al.* 2010; Good 2011). This method is particularly useful for studies focused on variation in protein-coding genes and the genetic basis of fitness and adaptation, because it provides a gene-targeted approach with SNPs from thousands of genes or the entire exome (Bi *et al.* 2012). Good described an exon capture study testing for the effects of climate warming on genetic variation in chipmunks, including a test for identifying candidate adaptive loci using the site frequency spectrum (SFS) (Bi *et al.* 2013). He also showed that exon capture could be used with low-quality genomic DNA samples or samples with >90% exogenous DNA from other species such as bacteria found in faecal and historical bone samples (Perry *et al.* 2010; Bi *et al.* 2013), whereas RADseq is less feasible or impossible with such DNA samples.

One limitation of exon capture is it requires knowledge of the exon sequences from the study species, or a related species, to design the capture bait oligonucleotides. However, Steve Amish showed that exons can be captured from a nonmodel species that has no published genome using baits designed from a related species that is over 50 million years divergent (or ~8% sequence divergence), although species with >5% sequence divergence may result in lower average sequence coverage (Cosart *et al.* 2011; Vallender 2011; Bi *et al.* 2012; Cosart 2013).

RNAseq

Another popular NGS method is RNAseq, which generates sequences from all expressed genes (the transcriptome) by sequencing cDNA produced by reverse transcription of total RNA, followed by de novo assembly or mapping of NGS reads to a reference genome (Hansen 2010). RNAseq can present several logistical challenges in a population genetics study, such as the challenges of acquiring high-quality RNA samples, and accounting for temporal variability in gene expression in the study organism. However, Good and Tyler Linderoth (University of California, Berkeley) showed how RNAseq de novo assembly for just one specimen of the study species can be used to identify thousands of exon sequences for exon capture array design (see previous section), which allows gene-targeted genotyping in population genetics studies of nonmodel species (Bi *et al.* 2012).

Whole genome sequencing

In whole-genome sequencing (WGS), short sequences (<500 bp) generated through shotgun sequencing are assembled into an entire genome. Of all the NGS methods, WGS generates the most data. However, WGS is also substantially more expensive than other methods. Several instructors suggested that most population genomic research questions do not require WGS, which might contribute to this method being rarely used in population genomics (Allendorf *et al.* 2010). However, WGS can answer some questions that other methods cannot. For example, Tiago Antão (University of Oxford) presented WGS data from 2000 mosquitos that revealed the presence of chromosomal inversions that might not be detected with RADseq or other NGS techniques; detection of inversions is important as they can influence fitness and gene flow, and can cause genotype-calling errors if undetected. In addition, using WGS to generate at least a preliminary draft genome can aid RADseq, RNAseq or Exon capture population genetic analyses, because of the advantages gained in these analyses when a reference genome is available (e.g. increased ability to identify paralogs). Therefore conducting WGS for one or a few specimens can be a relatively inexpensive way to improve other NGS analytical methods.

Filtering

'The main dilemma is filtering, in a word'. This quote, spoken by Luikart and J. Seeb, referred to one of the most important and the most challenging aspects of all NGS population genomics data analysis: filtering of raw sequence data to identify and remove errors. Filtering must be conducted at multiple levels of NGS analysis (e.g. during genotype calling and identification of outlier loci) and can have a dramatic impact on data interpretation. Several of the instructors provided advice regarding filtering of NGS data, as summarized below.

Filtering sequencing errors

Ken Warheit (Washington Department of Fish & Wildlife) provided an overview of our current knowledge of sequencing error on the Illumina platform, as well as methods for filtering to detect these errors. Error rates are much higher for Illumina sequencing (~ 1-1.5%) than traditional Sanger sequencing (~ 0.001%) (Shendure & Ji 2008). Because of these high error rates, all genotype-calling algorithms incorporate methods that account for sequencing errors. Many of these methods rely on estimates of sequencing error rates, and Warheit illustrated how sensitive these genotype-calling methods can be to incorrect estimates of error rates when depth of coverage is low. He advised that researchers use high depth of coverage (≥20×) and empirical estimates of sequencing error rates to avoid errors in genotype calling.

Filtering PCR Duplicates

Some of the sequencing reads obtained with NGS can be PCR duplicates, which are clonal reads arising from the same parent genomic DNA fragment during the PCR replication step in NGS library preparation. Failure to identify and filter out these duplicates can inflate coverage estimates, cause a heterozygote to look like a homozygote during genotype calling, or make an allele containing a PCR error appear to be an actual allele (false allele). Usually PCR duplicates are identified as fragments that are identical in insert length and sequence composition.

Hohenlohe and Miller pointed out that identifying PCR duplicates can be more difficult for RADseq than some other NGS methods, because all reads for a given RADseq allele begin at the same genomic position, that is, the restriction cut site, and therefore will have identical single-end sequences. The RADseq methods that can easily detect PCR duplicates have a random sheering step and also generate paired-end sequence reads (PE-RADseq) (Davey *et al.* 2013; Hohenlohe *et al.* 2013). When using these methods, RADseq PCR duplicates can be identified as reads that start at an identical sequence position in both the first and second read of the paired-end fragment. Notably, this method cannot be used to identify PCR duplicates for methods that do not incorporate a step which sheers DNA at random genomic locations, such as double digest RADseq (ddRADseq) (Peterson *et al.* 2012) and 2b-RADseq (Wang *et al.* 2012), and therefore several instructors discouraged the use of these methods. However, the development of methods to identify PCR duplicates for RADseq protocols that do not have a random sheering step is an area of active research, and therefore this may be possible in the future. Additionally, the development of new PCR-free RADseq methods provides potential for avoiding PCR duplicates (e.g. ezRAD, Toonen *et al.* 2013).

Filtering paralogs

Warheit also described the genotyping problems that can occur in NGS data analysis when paralogous regions, or repetitive regions in the genome, go unidentified and unfiltered. Because NGS reads are usually short, paralogous regions can be difficult to distinguish from one another. Several methods are available for identifying paralogous regions, but the simplest and most effective method is the alignment of reads against a reference genome, because paralogous regions are expected to align to multiple locations in the genome (reviewed in Hohenlohe *et al.* 2012). Therefore, when a reference genome is not available for a NGS population genetics project, it can be beneficial for researchers to generate at least a preliminary draft reference genome for the study species (Haussler *et al.* 2009). Once identified, paralogous regions are usually removed from subsequent population genetics analyses because they violate the assumption of orthology underlying most of these analyses.

Filtering loci with inconsistent depth of coverage

Several factors can cause variability in sequencing depth between and within individuals for some loci; when not accounted for, these factors may lead to biases in genotyping and population genetic inference. For example, the erroneous identification of reads from paralogous loci as being from only one orthologous locus can lead to high depth of coverage for the erroneously identified locus. Alternatively, the erroneous identification of orthologous alleles as separate loci could result in low depth of coverage for each erroneously identified locus. For RADseq, mutations in restriction cut sites can also cause variability in depth of coverage between and within individuals (Arnold *et al.* 2013; Gautier *et al.* 2013). Generally, loci with inconsistent coverage are removed from NGS data analyses, although in some cases, this solution may still be problematic (for further discussion see Arnold *et al.* 2013; Gautier *et al.* 2013).

Filtering outliers

Detection and filtering of outliers were discussed by many of the instructors. An important 'big picture' take-home message regarding filtering of outliers was from J Seeb: Be careful when filtering your NGS data to remove errors, paralogs, or outliers, because you might remove important biological signals. He used an analogy of how a well-funded USA research team studying global atmospheric ozone levels failed to notice the hole in the ozone layer over Antarctica because they had set quality control software to filter out ozone level data outside the expected norms. Instead, this team was scooped by the British Antarctic Survey, who discovered the ozone hole (Farman *et al.* 1985) and in the words of J.Seeb, 'saved the planet and got 1808 citations'.

Outlier detection has become a popular analytical tool among population geneticists for identifying putative adaptive loci in NGS data. Lisa Seeb (University of Washington) provided an overview of methods for NGS outlier detection when looking for selection, including tests for loci that behave as outliers when computing locus-specific statistics such as F_{ST}, heterozygosity, nucleotide diversity or correlations with environmental variables (reviewed in Oleksyk *et al.* 2010). Luikart recommended that researchers understand and apply the different methods for detecting outliers because agreement among methods might improve confidence in outlier detection. He also suggested researchers consider how the removal of outliers influences conclusions, and to report the results with and without outliers removed.

Genotype calling

Genotype calling can be complex for NGS data due to high error rates as well as biological phenomena such as paralogy (see previous section). Several instructors provided descriptions of different genotyping methods for NGS sequence data. Hohenlohe described a maximum-likelihood method implemented in the popular RADseq statistical analysis package STACKS (which also filters for sequence quality and paralogs) (Catchen *et al.* 2011, 2013). When using this method, genotype calling is expected to be highly accurate if sequencing coverage is deep (i.e. ≥20× average depth), and therefore Hohenlohe recommended that researchers 'do not lowball sequencing coverage' because 'reducing sequencing depth is a false economy'.

However, Linderoth described recently developed Bayesian methods that can improve the accuracy of population genomic inferences based on genotypes even when depth of coverage is as low as 2× (Fumagalli *et al.* 2013), as long as data are available for multiple individuals per population (Nielsen *et al.* 2012). The hallmark of these methods is that instead of strictly calling genotypes or SNPs, the posterior probabilities for all possible genotypes or allele frequencies at a given genomic site can be used to estimate F_{ST} or other population genetics parameters and statistics. The posterior probabilities for genotypes can be obtained using priors based on models such as Hardy–Weinberg Equilibrium (HWE), and the posterior probabilities for allele frequencies can be obtained using the sample-wide distribution of sample allele frequencies (i.e. SFS). These methods are implemented in the analysis package Analyses of Next-Generation Sequencing Data (ANGSD, http://www.popgen.dk/angsd/index.php/Main_Page) and its companion ngsTools (http://github.com/mfumagalli/ngsTools).

Other genotype-calling programmes mentioned during the workshop included GATK (http://www.broadinstitute.org/gatk/) and SAMtools (Li *et al.* 2009). Antão provided a sobering cautionary note that the use of SAMtools to call genotypes with default settings can lead to severe genotyping errors when a locus is not in HWE, because the default settings use HWE as a prior (as do some other genotype-calling programmes). However, he pointed out that using the allele log-likelihoods in SAMtools can circumvent this problem.

The importance of using Linux command line and scripting

One take-home message from several instructors was the importance of learning to use the Linux command line and to write scripts, which are generally required to analyse NGS data. For example, Luikart quoted Steve O'Brien (St. Petersburg State University) who recently said that >90% of the cost of WGS is the bioinformatics, emphasizing the need for scripting and programming skills. This estimate will likely remain true for many years if the cost of WGS production continues to decline.

Miller led a session providing basic strategies for learning to use the Linux command line. He emphasized that we should not be intimidated by learning Linux commands or scripting, because there are numerous resources available at our fingertips to make these skills easy to learn, such as Google, which can be used to search for any command. Also helpful are other people's scripts (such as the Perl scripts on his laboratory web page), the Seqanswers website which is dedicated to NGS discussion and education (http://seqanswers.com/), online tutorials such as the one at http://linuxcommand.org/index.php, and cheat sheets such as the one at http://ss64.com/bash/. To help researchers, we provide an audio recording of Miller's introduction to Linux command line, as well as recordings of several other instructors' talks, on the web page http://flbs.umt.edu/consecol/videos.

Application of NGS to population genomics questions

The instructors described how NGS can provide valuable information to aid in many population genomics inferences. For example, Mike Schwartz (US Forest Service, Montana) described a study which used whole mitogenome sequences to identify management units of fishers (*Martes pennanti*) that were not evident when using only one locus of the mitogenome (the D-loop) (Knaus *et al.* 2011). Ben Fitzpatrick (University of Tennessee) showed how both admixture and heterozygosity metrics can be used together to improve characterization of advanced generation hybrid individuals (Fitzpatrick 2012). He also described new analytical methods for detecting admixture outlier loci, that is, nonnative alleles introgressed to high frequency, likely by natural selection (Fitzpatrick 2013).

Brian Hand (University of Montana) explained how thousands of RADseq SNPs can be filtered for quality control and then used to identify corridors and landscape features that facilitate gene flow using novel landscape genomics modelling approaches including both circuitscape and least cost path approaches. He also showed how uncertainty in corridor and resistance model identification can be quantified using leave-one-out and sensitivity analysis methods; he explained how uncertainty assessment is seldom used but is badly needed to make the field of Landscape Genetics more reliable and scientific.

Robin Waples (NOAA Fisheries, Seattle) showed how the ratio of effective population size per generation (N_e) and N_b (effective number of breeders in one reproductive cycle) can be estimated for age-structured populations using two or three simple life-history traits (Waples *et al.* 2013). He described how large numbers of SNPs generated using NGS could be used to estimate N_e or N_b with high precision using the method based on linkage disequilibrium (LD), but he also cautioned that failure to explicitly account for physical linkage in NGS data sets can lead to serious problems. First, estimates of effective size will be biased downwards if LD due to physical linkage is interpreted as a signal of genetic drift. Second, overlapping sets of locus-pair comparisons (1 vs. 2, 1 vs. 3, 2 vs. 3, etc.) and redundant information from comparisons involving physically linked loci can lead to a great deal of pseudo-replication. This means that actual precision is much less than what could be achieved if all pairwise comparisons were independent, and this in turn means that confidence intervals can be substantially wider than those actually calculated under the standard assumption of independence (e.g. using LDNe software, Waples & Do 2008). Developing practical methods for dealing with these complications for LD analyses in NGS data sets is a high-priority research topic.

How can both neutral and adaptive variation revealed by NGS help to identify management units? This question generated much discussion among students and instructors (reviewed in Funk *et al.* 2012). Morten Limborg (University of Washington) showed how adaptive gene markers discovered using NGS might help identify cryptic adaptively differentiated populations (Limborg *et al.* 2012). While there was much excitement and hope to use adaptive genetic markers to identify conservation units, many agreed to first prioritize the conservation of genome-wide neutral variation and only then to also consider whether adaptive genetic markers might also be useful. One serious risk of prioritizing conservation decisions based on adaptive genetic markers is that they might identify loci or alleles adaptive to past environments, rather than current or future environments (Allendorf *et al.* 2013).

Parting words of wisdom

At the end of the workshop, instructors and students grabbed drinks, gathered around picnic tables overlooking the lake and reflected on the take-home messages of the workshop. Instructors were asked to provide overall words of advice, and several gave general career advice such as 'Have fun' (Miller), 'Don't be afraid to ask questions' (Fitzpatrick), 'Do what you love' (Allendorf), 'Like what you do and work damn hard. Try to think about the problem from a different perspective all the time' (Schwartz), and 'If you can't do what you love, love what you do' (Waples). J. Seeb recommended that everyone read E.O. Wilson's book entitled 'Letters to a Young Scientist' (Wilson 2013), which provides advice on how to succeed in science. Miller reiterated that scripting is important and easy to learn, and Warheit advised that researchers 'be prepared to spend a few thousand dollars [e.g. on sequencing lanes] to learn to produce and analyse NGS data'.

Linderoth said the most valuable tool for analysing NGS data was 'the Site Frequency Spectrum'. Miller then asked him what the second most important tool was, and Linderoth replied 'the 2D Site Frequency Spectrum'. Linderoth explained again that the SFS contains all the raw information about allele frequency distributions used in most population genetics approaches for detecting selection and inferring demography, as well as filtering to detect sequencing errors. One student was thankful to have learned it was surprisingly easy to simulate data to evaluate the performance of a statistical method; she was referring to an exercise led by Waples which used Easypop (Balloux 2001) to simulate data to evaluate N_e estimation methods under different demographic scenarios and with different numbers of loci.

Many instructors and students touched on the overarching message of the workshop that the analysis of NGS data requires careful attention to theory and computational assumptions to prevent drawing erroneous conclusions. Waples advised researchers to 'be one of the few to actually read the user's manual' for data analysis programmes. We must also be careful not to filter out and ignore biologically interesting information, such as signatures of selection, from our data. Ultimately, each NGS project will require a unique set of analyses tailored to the data set and research questions. Provided that we pay careful attention to the limitations of our analytical methods, NGS will bring many exciting new discoveries and improved understanding to the fields of population genomics, molecular ecology and conservation biology.

Acknowledgements

We thank Mike Miller, Jim Seeb, Tyler Linderoth, Robin Waples, Nolan Kane and three anonymous reviewers for helpful comments regarding the manuscript. K.R.A was supported by a European Union Marie Curie International Incoming Fellowship. G.L. was supported in part by funding from US National Science Foundation Grants DEB-1258203 and DEB-1067613.

References

See the course web page for the institute of origin of the 16 instructors (http://www.popgen.net/congen2013/). Audio recordings of course talks of Allendorf, J. Seeb, Miller, and Hohenlohe are available at http://flbs.umt.edu/consecol/videos.

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

Allendorf FW, Luikart GH, Aitken SN (2013) *Conservation and the Genetics of Populations*, 2nd edn. Wiley-Blackwell, West Sussex.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 1–14.

Bi K, Linderoth T, Vanderpool D *et al.* (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018–6032.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3-Genes Genomes Genetics*, **1**, 171–182.

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

Cosart T (2013) *Evaluation of new Methods for Large-Scale and Gene-Targeted Next Generation DNA Sequencing in Nonmodel Species*. PhD dissertation, University of Montana, Missoula, MT, 104 pp.

Cosart T, Beja-Pereira A, Chen S *et al.* (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, **12**, 347–355.

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory*. Burgess Pub. Co., Minneapolis, MN.

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

Farman JC, Gardiner BG, Shanklin JD (1985) Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature*, **315**, 207–210.

Fitzpatrick BM (2012) Estimating ancestry and heterozygosity of hybrids using molecular markers. *BMC Evolutionary Biology*, **12**, 1–14.

Fitzpatrick BM (2013) Alternative forms for genomic clines. *Ecology and Evolution*, **3**, 1951–1966.

Fumagalli M, Vieira FG, Korneliussen TS *et al.* (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Good JM (2011) Reduced representation methods for subgenomic enrichment and next-generation sequencing. In: *Molecular Methods for Evolutionary Genetics, Methods in Molecular Biology*, vol. **772** (eds Orgogozo V, Rockman MV), pp. 85–103. Springer Science+Business Media, New York.

Hansen MM (2010) Expression of interest: transcriptomics and the designation of conservation units. *Molecular Ecology*, **19**, 1757–1759.

Haussler D, O'Brien SJ, Ryder OA *et al.* (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, **100**, 659–674.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

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

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. In: *Data Production and Analysis in Population Genomics: Methods and Protocols, Methods in Molecular Biology*, vol. **888** (eds Pompanon F, Bonin A), pp. 235–260. Springer Science+Business Media, New York

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.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK (2011) Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC Ecology*, **11**, 10-Article No: 10.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, e37558.

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 185–205.

Perry GH, Marioni JC, Melsted P, Gilad Y (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, **19**, 5332–5344.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Toonen RJ, Puritz JB, Forsman ZH, et al. (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Vallender EJ (2011) Expanding whole exome resequencing into non-human primates. *Genome Biology*, **12**, 1–10.

Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, **9**, 808–812.

Waples RS, Do C (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, **8**, 753–756.

Waples RA, Luikart G, Faulkner JR, Tallmon DA (2013) Simple life-history traits explain key effective population size ratios across diverse taxa. *Proceedings of the Royal Society. Series B: Biological Sciences*, **280**, 20131339.

Wilson EO (2013) *Letters to a Young Scientist*. Liveright, New York.

G.L. was an organizer of the workshop. K.R.A. and G.L. wrote the paper.

doi: 10.1111/mec.12686