Received: 22 January 2018 | Revised: 22 January 2018 | Accepted: 24 January 2018

DOI: 10.1111/1755-0998.12761

NEWS AND VIEWS

Perspective

WILEY MOLECULAR ECOLOGY RESOURCES

Let's talk about sex: A rigorous statistical framework to assign the sex of individuals from reduced-representation sequencing data

Jean-Sébastien Moore | Laura Benestan

Institut de Biologie Intégrative et des Systèmes, Département de Biologie, Université Laval, Québec, QC, Canada

Correspondence
Jean-Sébastien Moore, Institut de Biologie Intégrative et des Systèmes & Département de Biologie, Université Laval, Québec, QC, Canada.
Email: jean-sebastien.moore@bio.ulaval.ca

Molecular markers have been used to identify the sex of sampled individuals for several decades, but the time-consuming development phase prevented their application in many systems. Recently, a growing number of papers have applied reduced-representation sequencing (RRS) protocols to the identification of sex-specific markers without the use of test crosses or prior genomic information. While such an approach has great advantages in terms of versatility and ease of use, the "shotgun sequencing" nature of RRS data sets leads to a high amount of missing data, which results in statistical challenges to the confident assignment of sex to individuals. In this issue of Molecular Ecology Resources, Stovall et al. (Molecular Ecology Resources, 18, 2018) provide a statistical framework to answer two questions: (1) how many individuals of one sex only must possess a genotype for this locus to be considered significantly sex-specific? and (2) How many sex-specific loci must an individual of unknown sex possess (in a given data set) to be confidently assigned a sex? The statistical pipeline introduced, and applied to samples of New Zealand fur seal (*Arctocephalus forsteri*) to identify 90 sex-specific loci, should be broadly applicable to a large number of species and constitutes a nice addition to the molecular ecology toolkit in the genomics era.

KEYWORDS
conservation genetics, evolution of sex, genomics/proteomics, mammals

Sex is a major preoccupation for many, perhaps especially for ecologists and evolutionary biologists who often need to interpret observations in the light of the sex of the individuals sampled. While physical attributes such as sexual dimorphisms or secondary sexual characteristics are widely spread, the collection of accurate sex information remains challenging in several organisms. Furthermore, in many applications it is impossible or impractical to infer sex from observation of physical attributes (e.g., forensics, fisheries surveys). It is no surprise then that the tools of molecular ecology were applied early on to the problem of sex assignment (e.g., Griffiths & Tiwari, 1993). However, characterization of markers that could reliably identify the sex of organisms (hereafter sex-specific markers) could be time-consuming and technically challenging. Reduced-representation sequencing (RRS) approaches (i.e., RADseq, genotyping-by-sequencing and other similar protocols; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016) are appropriate to address this challenge because they randomly sample the genome, which in the case of species with genotypic sex determination (Bachtrog et al., 2014) may lead to the discovery of some markers located in sex-specific regions. Consequently, there have been numerous recent articles that have applied RRS data to the problem of identifying sex-specific markers. Early examples of the use of RADseq data typically relied on linkage mapping or test crosses (e.g., Anderson et al., 2012; Baxter et al., 2011), but more recent studies have identified sex-specific markers without experimental crosses or prior genomic information (Fowler & Buonaccorsi, 2016; Gamble & Zarkower, 2014). In many cases, these

sex-specific markers have also advanced our understanding of the sex-determination system of nonmodel organisms (Benestan et al., 2017; Brelsford, Lavanchy, Sermier, Rausch, & Perrin, 2017). While the "shotgun sequencing" aspect of RRS approaches makes it a versatile tool to identify sex-specific markers, it also comes with its own set of problems. Indeed, RRS data sets typically contain large chunks of missing data, which makes it possible, or even likely for very small sample sizes, that a marker may be absent from one sex entirely just by chance, leading to the erroneous conclusion that it is sex-specific. Additionally, an individual might be missing genotypes at many of the previously identified sex-specific markers, but how many are required to confidently assign a sex to that individual?

The article by Stovall et al. (2018) in the current issue of *Molecular Ecology Resources* takes a significant step towards providing a statistically rigorous framework for the identification of sex-specific markers from RRS data, which they illustrate with a genotyping-by-sequencing (Elshire et al., 2011) data set from New Zealand fur seal (Figure 1). First, their approach defines markers that are definitely not sex-specific (i.e., markers that are found in individuals of both sexes), and randomly subsamples them in two groups ignoring sexes, to generate a null distribution of the frequency of markers that are found exclusively in one group by chance alone. They then used the 99th percentile of this distribution as a conservative threshold for inferring the number of genotyped individuals needed to statistically validate that a marker is sex-specific (here male-specific) versus non-sex-specific, as these non-sex-specific markers will be outside of the distribution. This number was defined as a sex-specific locus threshold (SSLT). From this threshold, they confidently identify a set of potential sex-specific markers without any prior information on their genomic location. Here, authors only aimed to identify male-specific loci as an XY sex-determination system was previously known (i.e., sex-specific markers are only located on the Y chromosome of males).

FIGURE 1 The article by Stovall et al. (2018) uses the New Zealand fur seal (*Arctocephalus forsteri*) as a model to demonstrate the usefulness of their protocol for the statistically rigorous identification of sex-specific markers using reduced-representation sequencing techniques. The seal is a common bycatch in commercial fisheries, and the authors used their newly proposed protocol to conclude that males were vastly over-represented in bycatch. This observation is consistent with observed behavioural differences between the sexes and will undoubtedly provide useful information for conservation. Photo: Will Stovall

Then, the next statistical issue that needed to be addressed was how many sex-linked markers would be required to confidently assign the sex of an individual. Indeed, sequencing errors and incorrect genotype calls are fairly common in RADseq data sets and may lead to erroneous sex assignation. Furthermore, the large amounts of missing data typical of RADseq data sets mean that even in the presence of many confidently identified sex-specific makers, some individuals of unknown sex may have genotypes calls only at a few of them, which leads to lower confidence in sex assignation. To answer this question, they use a 10-fold cross-validation method, which consists of dividing the data set into nine "training" sets where individuals have available sex information, and one "test" set where sex is unknown for all individuals. Based on this test set, they were able to estimate the number of individuals accurately assigned to its "true" sex and then draw a distribution of the frequency of individuals accurately assigned as a function of the number of sex-specific markers they possess. The authors then revealed that this distribution contains a break between the females, which possess no male-specific markers, and males, which possess a variable but non-zero number of sex-linked markers. The mid-point in the break of that distribution corresponds to the minimal number of sex-specific markers for an individual to be accurately called a male, which the authors named the significant sex-assignment threshold (SSAT). While the values of both the SSLT and the SSAT will vary according to each data set, the framework proposed here also provided straightforward recommendations, broadly applicable to a large number of study systems, for validating sex-specific markers through RRS protocols.

One interesting fact stemming from the work of Stovall et al. (2018) is that all of the loci identified as sex-specific are in fact monomorphic. This makes sense as these loci were almost all male-specific and found on the Y chromosome, which appears to lack diversity in this species, a common pattern in mammals. In a more traditional use of RRS data sets, these loci would have been thrown out at the filtering steps. Because this vast amount of sequencing information is usually wasted, the fact it might contain nuggets of useful information will be welcome news to some researchers.

Some will argue that using RRS to identify sex-linked markers is a brute force approach. There is some truth to that, but it is quickly evaporating with the rapidly decreasing costs of sequencing and the increasing availability of technical know-how relating to RRS techniques in laboratories across the world. Nonetheless, Stovall et al.

(2018) followed up their sex-specific loci discovery with the implementation of PCR-based methods using the information on the flanking regions of the RRS data. In that way, they demonstrated the efficiency of these sex-specific markers not only in the focal species, but also showed that they worked successfully for three other closely related species. This only served to further highlight the versatility of RRS methods for sex-specific marker identification. In cases where sex assignment is the major goal, a RRS library can be constructed from a smaller set of individuals and used to develop sex-specific markers; whereas when sex assignment is a secondary goal, the statistical methods presented by Stovall et al. (2018) can be used to maximize information content available in RRS data sets. This statistical framework therefore certainly will have value for many researchers in the field, and the use of RRS approaches to identify sex-specific markers will undoubtedly become increasingly common.

AUTHOR CONTRIBUTIONS

Both JSM and LB conceived and wrote the manuscript.

ORCID

Jean-Sébastien Moore http://orcid.org/0000-0002-3353-3730

REFERENCES

Anderson, J. L., Rodríguez Marí, A., Braasch, I., Amores, A., Hohenlohe, P., Batzel, P., & Postlethwait, J. H. (2012). Multiple Sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics (L Orban, Ed,). *PLoS ONE*, *7*, e40701.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., . . . The Tree of Sex Consortium. (2014). Sex determination: why so many ways of doing it? *PLoS Biology*, *12*, e1001899.

Baxter, S. W., Davey, J. W., Johnston, J. S., Shelton, A. M., Heckel, D. G., Jiggins, C. D., & Blaxter, M. L. (2011). Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism (PK Ingvarsson, Ed,). *PLoS ONE*, *6*, e19315.

Benestan, L., Moore, J.-S., Sutherland, B. J., Le Luyer, J., Maaroufi, H., Rougeux, C., . . . Bernatchez, L. (2017). Sex matters in Massive Parallel Sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. *Molecular Ecology*, *26*, 6767–6783.

Brelsford, A., Lavanchy, G., Sermier, R., Rausch, A., & Perrin, N. (2017). Identifying homomorphic sex chromosomes from wild-caught adults with limited genomic resources. *Molecular Ecology Resources*, *17*, 752–759.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), e19379.

Fowler, B. L., & Buonaccorsi, V. P. (2016). Genomic characterization of sex-identification markers in Sebastes carnatus and Sebastes chrysomelas rockfishes. *Molecular Ecology*, *25*, 2165–2175.

Gamble, T., & Zarkower, D. (2014). Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources*, *477*, 902–913.

Griffiths, R., & Tiwari, B. (1993). The isolation of molecular genetic markers for the identification of sex. *Proceedings of the National Academy of Sciences*, *90*, 8324–8326.

Stovall, W., Taylor, H., Black, M., Grosser, S., Rutherford, K., & Gemmell, N. (2018). Genetic sex assignment in wild populations using GBS data: a statistical threshold approach. *Molecular Ecology Resources*, *18*, 179–190.

How to cite this article: Moore J-S, Benestan L. Let's talk about sex: A rigorous statistical framework to assign the sex of individuals from reduced-representation sequencing data. *Mol Ecol Resour*. 2018;18:191–193. https://doi.org/10.1111/1755-0998.12761