Isabelle R. Onley, Jeremy J. Austin, Kieren J. Mitchell

Sex assignment in a non-model organism in the absence of field records using Diversity Arrays Technology (DArT) data

Conservation Genetics Resources, 2021; 13(3):255-260

Title: Sex assignment in a non-model organism in the absence of field records using Diversity Arrays Technology (DArT) data

Isabelle R Onley[1]*, Jeremy J Austin[1], Kieren J Mitchell[1,2]

[1] Australian Centre for Ancient DNA (ACAD), School of Biological Sciences, University of Adelaide, South Australia, SA

[2] ARC Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Biological Sciences, University of Adelaide, South Australia, SA

* Corresponding author (ORCID 0000-0003-2053-4002)

Word Count: 3448 words

Acknowledgements

The authors wish to acknowledge and thank Dr Katherine Moseby, Shaun Barclay, Bill Sherwin, and the staff of Arid Recovery Reserve for supplying the field data and samples used in this study.

Abstract

Conservation genomics research often relies on accurate sex information to make inferences about species demography, dispersal, and population structure. However, field determined sex data are not always available and can be subject to human error, while laboratory sex assignment methods such as PCR assays can often be costly and challenging for non-model species. Conservation genomics programs increasingly use reduced-representation genome sequencing to assess neutral and functional genetic diversity, population structure, gene flow and pedigrees in threatened species. Here we demonstrate that sex can be determined from reduced-representation sequencing data produced by the increasingly popular Diversity Arrays Technology sequencing workflow (DArT-seq) using a program originally designed for application to shotgun data. This program – *sexassign* – compares the "dosage" of sequencing reads mapping to autosomes versus the X chromosome. In the present study, *sexassign* was used to identify the sex of 60 field-collected Greater Stick-Nest Rat (*Leporillus conditor*) samples, despite the absence of an annotated reference genome for the species. This "read-dosage" approach is not only more accurate and affordable than traditional sex assignment methods, but can be applied to any diploid organism with a heterogametic sex determination system – including non-model and understudied species of conservation importance – by using FASTQs generated by DArT.

Keywords

Conservation genomics, sex assignment, bioinformatics, DArT-seq

Declarations

Funding

This research was supported by the University of Adelaide and funded by the following organisations and awards; Australian Government Research Training Program Scholarship, Nature Foundation South Australia Grand Start Grant (Grant No. 2019-07), Biological Society South Australia/Nature Conservation Society of South Australia Conservation Biology Grant, Field Naturalists Society of South Australia Lirabenda Endowment Fund Research Grant.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability

The reads generated for this study have been deposited at the Sequence Read Archive (NCBI) with BioProject ID PRJNA702840 (http://www.ncbi.nlm.nih.gov/bioproject/702840).

Code Availability

The original code can be found on Dr Graham Gower's GitHub repository < https://github.com/grahamgower/sexassign>.

Ethics Approval

Live animal trapping and sampling at Arid Recovery was conducted under South Australian Wildlife Ethics Committee permit numbers 27/98, 4/99, 22/99, 2/2000, 19/2000, and 18/2000.

Authors' Contributions

IRO and JJA coordinated submission of samples to DArT. IRO and KJM analysed the data. IRO drafted the abstract, introduction, results, and discussion. KJM drafted the materials and methods and figures. All authors contributed to the interpretation of results and provided feedback on the final manuscript.

Consent to Publish

The authors give consent for the publication of this manuscript.

Plant Reproducibility

N/A

Clinical Trials Registration

N/A

91

92

Introduction

Accurate sex assignment is an integral aspect of conservation genomics research, particularly when studying parameters such as relatedness, dispersal, and philopatry. Sexing of individuals used in conservation genomics studies typically takes place in the field at the time of collection. However, sex assignments recorded in the field are not always reliable and there is a wide margin for human error, particularly for species that do not demonstrate sexual dimorphism or when researchers are working in difficult conditions. Further, field records can easily be lost or incorrectly transcribed during trapping and monitoring. Genetic sex assignment is a favourable alternative or complement to field identification, as it is an objective, highly standardised, and accurate approach that eliminates the possibility of upstream sex misidentification confounding genomic studies (Hrovatin & Kunej, 2017).

While PCR-based sex identification methods have been used for several decades to identify and amplify sex chromosomes in individual samples (Akane et al., 1992; Clapcote & Roder, 2005; McFarlane et al., 2013), such processes can be time consuming and expensive. In addition, they require taxon-specific primers that are not always available or applicable to the target species. With the advent of high-throughput sequencing (HTS) technology it is now possible to produce high-resolution genomic data that may allow researchers to determine the sex of sequenced individuals bioinformatically. For example, single nucleotide polymorphisms (SNPs) in the genome can often be linked to the sex chromosomes in model organisms, allowing sex to be determined on chromosomal presence-absence basis (Fowler & Buonaccorsi, 2016; Lambert et al., 2016). For non-model organisms where a well-assembled and well-annotated reference genome is unavailable, the overall “dosage” of sequencing reads

mapping to the sex chromosomes can be assessed to determine whether the individual is heterogametic or homogametic and thus to identify the sex (Bover et al., 2018; Gamble, 2016; Gower et al., 2019; Pečnerová et al., 2017).

Read-dosage-based approaches to sex assignment have only been applied using shotgun sequencing data, where molecules are randomly sampled and sequenced (Flamingh et al., 2020; Motahari et al., 2013; Skoglund et al., 2013). However, many conservation programs employ reduced-representation sequencing approaches (e.g. RADseq), where sequenced molecules belong to a subset of genomic loci. One commercial provider of reduced-representation sequencing that is growing in popularity in the conservation genomics field is Diversity Arrays Technology (DArT) (Cummins et al., 2019; Ewart et al., 2019; Pazmiño et al., 2018; Sansaloni et al., 2011; Schultz et al., 2018; van Deventer et al., 2020). The DArT workflow uses restriction enzymes to reduce genomic complexity, allowing identification of informative markers that are subsequently sequenced for all submitted samples (Kilian et al., 2012). However, despite the growing popularity of DArT for conservation genomics projects, no simple and widely applicable sex-assignment framework has emerged that can be applied to DArT data. In the present study we apply a read-dosage sex-determination approach to DArT data from an Australian rodent, the Greater Stick-Nest Rat (*Leporillus conditor*), and demonstrate that - despite being originally designed for application to shotgun data - this method remains robust when applied to FASTQ files generated as part of the DArT workflow.

Materials and Methods

DNA submitted to DArT was extracted from 60 *L. conditor* tissue samples collected by staff during routine trapping events at Arid Recovery Reserve, South Australia, between 1999 and 2003. DNA extraction was completed following the methods described by Barclay et al. (2006) and samples were subsequently stored at -20°C prior to sequencing by DArT. Following library preparation and sequencing by DArT using their proprietary workflow, we obtained the raw Illumina data in FASTQ format. We used the Paleomix v1.2.14 pipeline to process these data: AdapterRemoval2 v2.3.1 was used to trim residual adapter sequences (using default parameters) and filter reads shorter than 30 bp, after which all remaining reads were mapped against the repeat-masked house mouse genome assembly (GRCm38) using BWA v0.7.17 *mem* algorithm. We then used the *idxstats* command in SAMtools v1.10 to extract the number of reads mapping to each scaffold of the reference assembly.

To determine the sex of the Greater Stick-Nest Rat samples we used Gower et al.'s (2019) python script *sexassign* (https://github.com/grahamgower/sexassign), which uses a likelihood ratio test to assign samples to either male or female on the basis of the observed ratio of reads mapping to the X chromosome versus the autosomes. Following Gower et al. (2019), X chromosome read-dosage is used in preference to the Y chromosome because references for the latter are either unavailable or poorly assembled for most species (Janečka et al., 2018). However, *sexassign* assumes that the X chromosome in homogametes (females, in this case) should receive the same read-dosage as an autosome of the same length (i.e. read dosage of ~1X versus ~0.5X in heterogametic males), so we first checked that our data conformed to this assumption by visualising read-dosage (proportion of total reads mapped versus scaffold length) for each sample using RStudio v1.3.1073 (Fig. 1). We observed that the mean proportion of reads mapping to the X chromosome (length = 171,031,299 bp) for the putatively

female samples (0.0308) was substantially lower than the expectation (0.0656) based on the relationship between the proportion of reads mapped and scaffold length inferred from the autosomes, perhaps due to the DArT marker-selection and filtering process or a depletion of the restriction motif on the X chromosome. Consequently, before proceeding with analysis using *sexassign* we first multiplied the number of reads mapping to the X chromosome for all samples (regardless of putative sex) by a factor of 2.12 (the expected read-dosage for the X chromosome in a female, 0.0656, divided by the observed mean read-dosage for the X chromosome in the putatively female samples, 0.0308).

Results

The proportion of reads mapping to each of the autosomes was highly consistent between samples (Fig. 1). Further, autosomal read-dosage appeared to be positively correlated with scaffold length, as expected if restriction motifs are randomly distributed. We tested this correlation by performing a linear regression in RStudio (proportion of reads ~ scaffold length), which resulted in a slope coefficient of $3.833e^{-10}$ (adjusted $R^2 = 0.7$, $p < 2e^{-16}$). Unlike the autosomes, values for the proportion of reads mapping to the X chromosome formed two clusters, putatively representing females (with higher read-dosage values) and males (with lower read-dosage values).

The read-dosage sex-assignment program (*sexassign*) allowed us to successfully assign all individuals in the dataset as either male (heterogametic, XY; X read-dosage = ~0.5X) or female (homogametic, XX; X read-dosage = ~1X, Fig. 2, Table 1). Of the 60 individuals sequenced, 33 were determined to be female and 27 to be male, consistent with the typical sex ratio in

rodent populations under normal conditions (Labov et al., 1986; Rosenfeld et al., 2003). Genetic sex assignment had a ~94% concurrence rate with field determined sex, a typical human error margin considering the lack of obvious sexual dimorphism within the species and the difficulty of accurately sexing rodents in the field, particularly during non-reproductive periods (Hoffmann et al., 2010; Jacques et al., 2015).

Discussion

Our results demonstrate that the FASTQ-formatted data routinely generated by Diversity Arrays Technology (DArT) as an intermediate step in their workflow can reliably be used to determine the sex of samples from non-model organisms, confirming or replacing field-based sex identification and eliminating the need for additional costly laboratory sexing analyses. Importantly, a reference genome from the species of interest does not appear to be necessary, as we obtained robust results by mapping our data to the reference assembly for the house mouse (*Mus musculus*), which shared a common ancestor with *L. conditor* 10 million years ago (Steppan & Schenk, 2017). While the house mouse genome is assembled to the chromosome-level, making identification of reads mapping to the X chromosome straightforward, this approach should also work with scaffold-level reference assemblies.

Gower et al. (2019) identified X-linked scaffolds in the polar bear genome (UrsMar1.0) by first mapping all scaffolds against the chromosome-level dog reference assembly (CanFam3.1), then applied *sexassign* to shotgun sequencing data from a third species – brown bears (*Ursus arctos*) – that they mapped to the putative polar bear X-linked scaffolds. Given that scaffold-level assemblies are increasingly available for a wide range of taxa, our results suggest that

most DArT end-users working on mammals should be able use their FASTQ data to determine the sex of their samples. Indeed, the read-dosage approach to sex assignment should be applicable to any diploid organism with a heterogametic sex-determination system, such as birds, lizards, and many invertebrates, regardless of which sex is homogametic.

References

Akane, A., Seki, S., Shiono, H., Nakamura, H., Hasegawa, M., Kagawa, M., Matsubara, K., Nakahori, Y., Nagafuchi, S., & Nakagome, Y. (1992). Sex determination of forensic samples by dual PCR amplification of an X-Y homologous gene. *Forensic Science International*, *52*(2), 143–148. https://doi.org/10.1016/0379-0738(92)90102-3

Bover, P., Llamas, B., Thomson, V. A., Pons, J., Cooper, A., & Mitchell, K. J. (2018). Molecular resolution to a morphological controversy: The case of North American fossil muskoxen Bootherium and Symbos. *Molecular Phylogenetics and Evolution*, *129*, 70–76. https://doi.org/10.1016/j.ympev.2018.08.008

Clapcote, S. J., & Roder, J. C. (2005). Simplex PCR assay for sex determination in mice. *BioTechniques*, *38*(5), 702–706. https://doi.org/10.2144/05385BM05

Cummins, D., Kennington, W. J., Rudin-Bitterli, T., & Mitchell, N. J. (2019). A genome-wide search for local adaptation in a terrestrial-breeding frog reveals vulnerability to climate change. *Global Change Biology*, *25*(9), 3151–3162. https://doi.org/10.1111/gcb.14703

Ewart, K. M., Johnson, R. N., Ogden, R., Joseph, L., Frankham, G. J., & Lo, N. (2019). Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species. *Molecular Ecology Resources*, *19*(6), 1578–1592. https://doi.org/10.1111/1755-0998.13082

Flamingh, A. de, Coutu, A., Roca, A. L., & Malhi, R. S. (2020). Accurate Sex Identification of Ancient Elephant and Other Animal Remains Using Low-Coverage DNA Shotgun Sequencing Data. *G3: Genes, Genomes, Genetics*, *10*(4), 1427–1432. https://doi.org/10.1534/g3.119.400833

241 Fowler, B. L. S., & Buonaccorsi, V. P. (2016). Genomic characterization of sex-identification
242 markers in Sebastes carnatus and Sebastes chrysomelas rockfishes. *Molecular*
243 *Ecology*, *25*(10), 2165–2175. https://doi.org/10.1111/mec.13594
244 Gamble, T. (2016). Using RAD-seq to recognize sex-specific markers and sex chromosome
245 systems. *Molecular Ecology*, *25*(10), 2114–2116. https://doi.org/10.1111/mec.13648
246 Gower, G. (2019). *Inferring the Charactersitics of Ancient Populations using Bioinformatic*
247 *Analysis of Genome-wide DNA Sequencing Data* [Doctoral Dissertation]. University
248 of Adelaide.
249 Gower, G., Fenderson, L. E., Salis, A. T., Helgen, K. M., van Loenen, A. L., Heiniger, H.,
250 Hofman-Kamińska, E., Kowalczyk, R., Mitchell, K. J., Llamas, B., & Cooper, A.
251 (2019). Widespread male sex bias in mammal fossil and museum collections.
252 *Proceedings of the National Academy of Sciences*, *116*(38), 19019–19024.
253 https://doi.org/10.1073/pnas.1903275116
254 Hoffmann, A., Decher, J., Rovero, F., Schaer, J., Voigt, C., & Wibbelt, G. (2010). Field
255 Methods and Techniques for Monitoring Mammals. *Manual on Field Recording*
256 *Techniques and Protocols for All Taxa Biodiversity Inventories*, *8*, 482–529.
257 Hrovatin, K., & Kunej, T. (2017). Genetic sex determination assays in 53 mammalian
258 species: Literature analysis and guidelines for reporting standardization. *Ecology and*
259 *Evolution*, *8*(2), 1009–1018. https://doi.org/10.1002/ece3.3707
260 Jacques, M.-E., McBee, K., & Elmore, D. (2015). *Determining Sex and Reproductive Status*
261 *of Rodents*. 4.
262 Janečka, J. E., Davis, B. W., Ghosh, S., Paria, N., Das, P. J., Orlando, L., Schubert, M.,
263 Nielsen, M. K., Stout, T. A. E., Brashear, W., Li, G., Johnson, C. D., Metz, R. P.,
264 Zadjali, A. M. A., Love, C. C., Varner, D. D., Bellott, D. W., Murphy, W. J.,

Chowdhary, B. P., & Raudsepp, T. (2018). Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nature Communications*, *9*(1), 2945. https://doi.org/10.1038/s41467-018-05290-6

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K., Jaccoud, D., Hopper, C., Aschenbrenner-Kilian, M., Evers, M., Peng, K., Cayla, C., Hok, P., & Uszynski, G. (2012). Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms. In F. Pompanon & A. Bonin (Eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols* (pp. 67–89). Humana Press. https://doi.org/10.1007/978-1-61779-870-2_5

Labov, J. B., William Huck, U., Vaswani, P., & Lisk, R. D. (1986). Sex ratio manipulation and decreased growth of male offspring of undernourished golden hamsters (Mesocricetus auratus). *Behavioral Ecology and Sociobiology*, *18*(4), 241–249. https://doi.org/10.1007/BF00300000

Lambert, M. R., Skelly, D. K., & Ezaz, T. (2016). Sex-linked markers in the North American green frog (Rana clamitans) developed using DArTseq provide early insight into sex chromosome evolution. *BMC Genomics*, *17*(1), 844. https://doi.org/10.1186/s12864-016-3209-x

McFarlane, L., Truong, V., Palmer, J. S., & Wilhelm, D. (2013). Novel PCR Assay for Determining the Genetic Sex of Mice. *Sexual Development*, *7*(4), 207–211. https://doi.org/10.1159/000348677

Motahari, A. S., Bresler, G., & Tse, D. N. C. (2013). Information Theory of DNA Shotgun Sequencing. *IEEE Transactions on Information Theory*, *59*(10), 6273–6289. https://doi.org/10.1109/TIT.2013.2270273

Pazmiño, D. A., Maes, G. E., Green, M. E., Simpfendorfer, C. A., Hoyos-Padilla, E. M., Duffy, C. J. A., Meyer, C. G., Kerwath, S. E., Salinas-de-León, P., & van Herwerden, L. (2018). Strong trans-Pacific break and local conservation units in the Galapagos shark (Carcharhinus galapagensis) revealed by genome-wide cytonuclear markers. *Heredity*, *120*(5), 407–421. https://doi.org/10.1038/s41437-017-0025-2

Pečnerová, P., Díez-del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., Nikolskiy, P., Tikhonov, A., Vartanyan, S., & Dalén, L. (2017). Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth. *Current Biology*, *27*(22), 3505-3510.e3. https://doi.org/10.1016/j.cub.2017.09.064

Rosenfeld, C. S., Grimm, K. M., Livingston, K. A., Brokman, A. M., Lamberson, W. E., & Roberts, R. M. (2003). Striking variation in the sex ratio of pups born to mice according to whether maternal diet is high in fat or carbohydrate. *Proceedings of the National Academy of Sciences*, *100*(8), 4628–4632. https://doi.org/10.1073/pnas.0330808100

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., & Kilian, A. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: Genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proceedings*, *5*(S7), P54, 1753-6561-5-S7-P54. https://doi.org/10.1186/1753-6561-5-S7-P54

Schultz, A. J., Cristescu, R. H., Littleford-Colquhoun, B. L., Jaccoud, D., & Frère, C. H. (2018). Fresh is best: Accurate SNP genotyping from koala scats. *Ecology and Evolution*, *8*(6), 3139–3151. https://doi.org/10.1002/ece3.3765

Skoglund, P., Storå, J., Götherström, A., & Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science*, *40*(12), 4477–4482. https://doi.org/10.1016/j.jas.2013.07.004

Steppan, S. J., & Schenk, J. J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. *PLOS ONE*, *12*(8), e0183070. https://doi.org/10.1371/journal.pone.0183070

van Deventer, R., Rhode, C., Marx, M., & Roodt-Wilding, R. (2020). The development of genome-wide single nucleotide polymorphisms in blue wildebeest using the DArTseq platform. *Genomics*, *112*(5), 3455–3464. https://doi.org/10.1016/j.ygeno.2020.04.032

Table 1. Results of DNA-based sex assignment using *sexassign* compared to sex determined in the field for 60 greater stick-nest rats. The length of the X chromosome was 171,031,299 bp and the total length of the autosomes was 2,462,745,373 bp (Gower, 2019).

ID[†]	Field Sex	M_X[‡]	Sex	N_X[§]	N_A[¶]
ET002	nd	0.474	M	26392	802947
ET101	nd	0.962	F	23290	349580
ET102	nd	0.480	M	12663	380070
ET103	nd	0.953	F	25446	385924
ET106	M	0.515	M	12856	359385
ET119	F	0.930	F	60581	941980
ET133	F	0.963	F	26462	396880
ET146	M	0.517	M	8416	234263
ET147	F	0.979	F	26407	388782
ET147B	nd	0.976	F	12858	190022
ET148	M	0.507	M	14526	412334
ET149	F	1.020	F	29618	417327
ET151	F	0.975	F	24507	362393
ET152	F	1.002	F	28024	402617
ET153	nd	0.970	F	26335	391810
ET154	nd	0.946	F	25894	395471
ET155	nd	0.482	M	14054	420275
ET157	M	1.026	F	28484	399170
ET158	nd	0.950	F	26525	403485
ET162	nd	0.503	M	13867	397200
ET163	nd	0.946	F	24215	370137
ET163B	nd	0.473	M	14299	436150
ET17	F	0.942	F	58158	892183
ET173	F	0.956	F	27789	419868
ET176	F	0.938	F	22451	346121
ET177	nd	0.952	F	26275	398926
ET18	F	0.905	F	39676	635045
ET183	nd	0.495	M	13220	384279
ET184	nd	0.487	M	32640	966813
ET185	F	1.010	F	26952	384146
ET186	nd	0.473	M	28294	863292
ET187	nd	0.996	F	25964	375434
ET188	nd	0.503	M	12563	359345
ET189	nd	0.972	F	22913	339929
ET192	M	0.500	M	14444	416297
ET193	M	0.489	M	13108	386485

ET195	nd	0.960	F	25194	378761
ET196	F	0.977	F	27030	398915
ET198	M	0.512	M	28970	813496
ET198B	nd	0.484	M	12733	378965
ET203	M	0.475	M	11469	348138
ET209	F	0.971	F	25533	379373
ET217	M	0.480	M	13460	404344
ET231	nd	0.493	M	12745	372720
ET233	nd	0.480	M	12353	370852
ET255	F	0.952	F	24282	368459
ET259	M	0.478	M	11357	342534
ET261	F	0.958	F	26557	400394
ET277	M	0.488	M	29029	857827
ET29	F	0.939	F	30250	465742
ET29B	nd	0.959	F	23511	354200
ET3	F	0.991	F	27316	397077
ET32	M	0.509	M	13059	369309
ET37	M	0.467	M	26816	828029
ET5	M	0.491	M	27564	809456
ET50	F	0.485	M	11802	350729
ET50.2	nd	0.981	F	23708	348582
ET5967	nd	0.958	F	26131	393719
ET61	M	0.491	M	7566	222226
ET62	F	0.987	F	36015	526076

† ID = ear tag number for *L. conditor* individual, nd = not determined, ‡ M_X = read dosage on X chromosome, § N_X = count of reads mapped to the X chromosome (after multiplying by 2.12), ¶ N_A = count of reads mapped to the autosome.

Figure Legends

Fig 1 Proportion of reads mapped to autosomes and the X chromosome in the *L. conditor* DArT dataset. Colour/symbol combinations represent different individuals. Read dosage of autosomes was positively correlated with scaffold length, while reads for the X chromosome form two distinct "dosage" clusters indicative of homogametic (XX) and heterogametic (XY) individuals

Fig 2 Plot of X chromosome read dosages for all sequenced *L. conditor* individuals, with confidence intervals for male heterogametes (red) and female homogametes (blue)