Proc. Assoc. Advmt. Anim. Breed. Genet. 23:23-26

DEEP LEARNING FOR GENOTYPE QUALITY CONTROL

D.P. Garrick[1,2]

[1]Theta Solutions, LLC, Atascadero CA, USA 93422
[2]A.L. Rae Centre for Genetics and Breeding, School of Agriculture and Environment, Massey University, New Zealand

SUMMARY

SNP genotype data are increasingly employed across a range of species for routine use in parentage verification and identification, and single-step evaluations. Robust and automated quality control processes are a critical step in maximizing the value of genotype data in these, and other, applications. Prediction of "genotype sex" is a common quality control metric but can be problematic for example on mammalian chips that do not contain Y chromosome markers because methods based on heterozygosity of X chromosome markers can incorrectly flag inbred females as male. A deep learning model is trained to predict "genotype sex" and validated and tested using real-world data routinely used in the American Hereford Association's single-step evaluation.

INTRODUCTION

A major challenge that comes with the advent of low-cost SNP genotyping is curation and management of the vast quantities of data that are produced. Take the case when the genotype sample for a particular animal fails to verify against its genotyped parents in a SNP based parentage verification. If this was to occur, an ideal system would automatically initiate a search against other relevant genotype samples to try and find the true parent without any extra input from the user. If such functionality is not available, or if such a search fails to find a match, then there is the question of a) is the true parent not genotyped, or b) is one or more of the relevant genotypes involved in the parent verification a bad or mismatched sample. In either case this typically requires the breed society and/or breeder to be contacted in order to generate a list of potential parents or to query any potential issues with the sample. This can be complicated by the use of non-standard or otherwise inconsistent animal, sample, and genotype identifiers. The length of time for this process can be significantly shortened by gleaning various information from the genotype sample(s) in question such as potential relatives or phenotypic characteristics. For example, if a genotype is clearly from a female and the animal in question is definitively male (or animal is black and horned and genotype indicates red and polled), it is reasonable to assume the sample in question is incorrect and the animal should have a new sample taken for regenotyping.

Prediction of "genotype sex" is an important quality control metric for genotype samples and is predicted from the sex chromosomes, i.e. in mammals the X and Y chromosomes for males and two X chromosomes for females. Females inherit one X chromosome from their mother, and one X chromosome from their father. With no inbreeding, the copy of each allele from each chromosome will not always be the same and the resulting SNPs will exhibit heterozygosity. As males only have one copy of the X and Y chromosomes, any alleles called from the unmatched parts of those chromosomes should always be the same, resulting in homozygosity within that region.

Deep learning is a subset of machine learning algorithms that passes an input training dataset through multiple layers of neurons in a neural network to successively transform and extract features from the output of the previous layer (Deng and Yu 2014). Leveraging the unique computational capabilities of Graphics Processing Units (GPUs) developed to render modern video games, deep learning approaches have gained significant media attention recently due to associated large technological advances in applications such as self-driving cars, image recognition and classification,

medical diagnostics, and many others.

Certain X chromosome SNPs, even those outside the pseudo autosomal region (PAR), can be heterozygous in males if they are located in regions exhibiting copy number variation. Further, X chromosome SNPs can be homozygous in females, especially inbred females who may have inherited the same X chromosome from both her sire and dam, e.g. if her sire is also her maternal grandsire. Thus in rules-based approaches selecting an appropriate subset of SNPs and male/female heterozygosity cut-offs can greatly affect the subsequent genotype sex prediction and without Y chromosome SNPs inbred females can be misclassified. On the other hand, given a suitable training dataset with realistic data and known true sex of the associated samples, a deep learning model can in theory account for the nuances and variation of specific SNPs in the given training dataset to generate accurate predictions. This is possible using a table containing the relevant sex SNPs and utilizing approaches for deep learning on tabular data via the fast.ai toolbox (Rachel Thomas 2018). The objective of this study was to determine if a deep learning approach can accurately predict the genotype sex of an animal and to assess the value of such a tool as a routine automated quality control step within a genomic database information system.

MATERIALS AND METHODS

The genotype data employed for the study consisted of a subset of those SNP genotypes from 67,304 animals used in a recent single-step evaluation from the full American Hereford Association genomic database of >110,000 genotyped animals. The samples originate from several platforms, genotyping laboratories, and chips across a number of years but consist predominantly of GeneSeek 50K and 30K genotypes. Of these, a subset of 15,619 "pedigree verified true" male and female genotypes was determined by taking samples from only those animals who were recorded in the current pedigree as a sire or dam and who subsequently passed SNP-based pedigree verification with at least 1 genotyped offspring. For pedigree verification, no samples used in this study had less than 5,000 called SNPs in common. Pedigree verified animals recorded as a sire in the pedigree were then considered a "true" male while those recorded as a dam were considered a "true" female totalling 5,058 and 10,561 for males and females respectively. As the American Hereford Association has utilized the international ICAR ID format for many years, the pedigree recorded sex for each animal is recorded as the 7th character of the ID, e.g., HERUSAM000000000001 is recorded as a male and HERUSAF000000000002 is recorded as a female. Comparing a predicted genotype sex to its pedigree recorded sex is straightforward as a result.

Three approaches for computing "genotype sex" were examined. The first consists of a simple rule-based non PAR (nPAR) X-chromosome heterozygosity check using all available called nPAR X SNPs from a list of 3,035 SNPs which exist across a variety of genotyping chips and platforms. No sample used in this study had less than 700 called nPAR X SNPs. Samples with ≤5% heterozygosity amongst their called nPAR X SNPs were classified as males while samples with >5% were classified as females. The second approach tested is the rule-based protocol developed by ICBF and is as follows using only a specific small subset of 280 nPAR X chromosome SNPs as described by McClure *et al*. 2018: 1) Determine heterozygosity rate (#AB/ (#AA+#AB+#BB)) for nPAR SNP; 2) If ≤5% het rate = male; 3) If ≥15% female; 4) If between 5 and 15%= ambiguous sex. Additionally, ICBF employs a subset of 7 Y chromosome SNPs: 1) Count nPAR chrY genotypes; 2) If 0–1 genotypes = female; 3) If 6–7 =male; 4) If 2–5 = ambiguous sex. Between the X and Y chromosome predictions any non-conflicting unambiguous sex is reported, otherwise an ambiguous or conflicting sex is reported. The Y sex prediction is dependent on samples having been genotyped on a chip where Y SNPs are available and several thousand samples used in this study did not have Y SNPs available. Instead of excluding those samples a two-step ICBF (X+Y) sex prediction was utilised instead of the fully joint

ICBF(X+Y) sex prediction described above and by McClure *et al*. 2018, such that Y chromosome predictions were used only if the X chromosome predictions were ambiguous.

Finally, a deep neural network (DNN) genotype sex predictor was built utilizing the fast.ai deep learning tabular toolbox (Howard & others 2018) in conjunction with a dataset consisting of just the 280 ICBF X chromosome sex SNPs. Some 2,500 male and 5,000 female genotypes chosen at random from the "pedigree verified true" samples were used as the training data for the DNN while the remaining of the 15,619 samples were used as the validation data. The only dependent variable is the sex prediction while each called SNP was treated as an input categorical variable with values -1, 0, 1, or 5 (no call). Prediction accuracy was used as the training metric and neural networks with various numbers of hidden layers and neurons per layer were tested for training over 25 epochs which took ~5-6 minutes each. The sex prediction is output as a probability of being male and a probability of being female. Sex predictions with ≥ 80% probability were taken as the predicted sex with the remaining assumed to be ambiguous.

RESULTS AND DISCUSSION

Table 1 summarises the number of predicted male, female, and ambiguous sex animals from each approach. An ambiguous male or female means the sex prediction was ambiguous and the pedigree recorded sex was male or female respectively. A conflicting male or female refers to the pedigree recorded sex being male or female respectively and the genotype sex predicted as female or male, respectively. The DNN results were reported from a network with 600 hidden layers and 300 neurons per layer which was found to have the most accurate results of those tested. However, other network sizes with neurons on the order of the number of SNPs (280) achieved very similar results. Perhaps unsurprisingly, the DNN achieves the highest accuracy on this "pedigree verified true" dataset as it is the same dataset that was used for training and validation of the neural network.

Table 2 summarises the differences between the sex predictions from each approach compared to the pedigree recorded sex of each animal in the larger genotype database not including samples otherwise used in the training and validation set for the DNN consisting of 67,304-15,619=51,685 samples. In both the "training" and "test" datasets, use of the ICBF Y chromosome data to augment otherwise ambiguous predictions using only the ICBF X chromosome results does appear to improve prediction accuracy. The nPAR X approach with the hard cut-off between male and female means no "ambiguous" sex samples are flagged, however, the overall percentage of animals matching their pedigree recorded sex is roughly the same as the ICBF approach. The DNN achieves a similar percentage of predictions matching the pedigree recorded sex in the test dataset as the rules-based approaches while using only the 280 ICBF X SNPs and after training with a dataset of only 2,500 male and 5,000 female genotypes randomly selected from the 15,619 "true" sexed samples. The remainder of the 15,619 samples were used for cross-validation during training.

Table 1. Results summary against the "pedigree verified true" sex of 15,619 individuals used for training and validation of the DNN

	nPAR (X)	ICBF(X)	ICBF(X+Y)	DNN(X)
% Correctly Predicted	99.76	99.23	99.86	99.88
Total Predicted Female	10,525	10,444	10,542	10,549
Total Predicted Male	5,094	5,074	5,076	5,065
Ambiguous Female	N/A	97	0	5
Ambiguous Male	N/A	4	1	0
Conflicting Female	37	20	20	10
Conflicting Male	1	0	1	3

Table 2. Genotype sex prediction results summary against the pedigree recorded sex of the animals in the 51,685 samples test dataset, which does not include any animals used in the training and validation set

	nPAR (X)	ICBF(X)	ICBF(X+Y)	DNN(X)
% Matching pedigree	99.79	99.40	99.82	99.70
Ambiguous Female	N/A	197	2	10
Ambiguous Male	N/A	37	14	46
Conflicting Female	57	53	53	61
Conflicting Male	51	23	25	82

CONCLUSIONS

This study shows deep learning approaches have potential as an accurate genotype sex prediction tool in routine and automated genotype sample quality control processes. The accuracy of a deep learning tool trained on a random subset of "pedigree verified true" gendered samples is found to be comparable to that of existing rules-based approaches. A purely X chromosome heterozygosity rules-based approach can benefit from using Y chromosome data to improve otherwise ambiguous predictions.

The benefits of a deep learning tool are that it can be integrated and automated with an existing suite of quality control protocols. In a production system the tool could be routinely tuned and further trained against new and verified data as it arrives. This in theory should allow it to better account for the nuances in the specific datasets of interest.

There are a significant number of avenues for further investigation with regards to the deep learning approach. These include greater exploration of the effect of the deep learning parameters on prediction results, e.g. number of hidden layers and neurons per layer, as well as the size of the dataset used for training and validation both in terms of the SNPs included and the particular individuals that comprise the training and validation sets. Other avenues include incorporation of other data features into the deep learning model such as genotyping platform or chip, Y chromosome SNPs, recorded breed, sample call rate or individual SNP GC scores, inbreeding coefficients, and/or other pedigree information. If genotype data on individuals exhibiting sex chromosome defects or being intersex are available, these could also be incorporated. Extension of the model to additional prediction outputs (e.g. breed) would also be valuable.

Some drawbacks of the deep learning approach are that it does require a suitable training dataset, finding the optimal DNN architecture (e.g. number of layers and neurons per layer) and training parameters is unclear, it requires GPU-based hardware and expertise to run. Finally, even though the deep learning model returns the probability a given sample is male or female unlike the rules-based approaches, the abstract nature of the deep learning model can create extra challenges in communicating prediction results back to breeders or other stakeholders.

REFERENCES

Deng, L., & Yu, D. (2014). *Foundations and Trends® in Signal Processing*, ***7*(3–4)**:197. https://doi.org/10.1561/2000000039

Howard, J., & others. (2018). fastai. GitHub.

McClure, M. C., McCarthy, J., Flynn, P., McClure, J. C., Dair, E., O'Connell, D. K., & Kearney, J. F. (2018). *Frontiers in Genetics*, ***9*(84)**. https://doi.org/10.3389/fgene.2018.00084

Thomas, R. (2018). Retrieved from https://www.fast.ai/2018/04/29/categorical-embeddings/