Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genome Analysis

seXY: A tool for sex inference from genotype arrays

David C. Qian[1], Jonathan A. Busam[2], Xiangjun Xiao[1], Tracy A. O'Mara[3], Rosalind A. Eeles[4], Frederick R. Schumacher[5], Catherine M. Phelan[6], Christopher I. Amos[1,*]

[1]Department of Biomedical Data Science, Dartmouth Geisel School of Medicine, Lebanon, NH 03756, USA, [2]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA, [3]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia, [4]Division of Genetics and Epidemiology, Institute of Cancer Research, London, SW7 3RP, UK, [5]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA, [6]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL 33612, USA,

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Checking concordance between reported sex and genotype-inferred sex is a crucial quality control measure in genome-wide association studies (GWAS). However, limited insights exist regarding the true accuracy of software that infer sex from genotype array data.
Results: We present seXY, a logistic regression model trained on both X chromosome heterozygosity and Y chromosome missingness, that consistently demonstrated >99.5% sex inference accuracy in cross-validation for 889 males and 5,361 females enrolled in prostate cancer and ovarian cancer GWAS. Compared to PLINK, one of the most popular tools for sex inference in GWAS that assesses only X chromosome heterozygosity, seXY achieved marginally better male classification and 3% more accurate female classification.
Availability: https://github.com/Christopher-Amos-Lab/seXY
Contact: Christopher.I.Amos@dartmouth.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) are being conducted at an unprecedented rate due to the precipitous fall in genotyping cost over time (Begum et al., 2012). An essential quality control (QC) step in these studies is verifying concordance between self-reported sex and genotype-inferred sex. Disagreements can prompt researchers to double-check their data, and either fix entry errors or discard samples if unreliable phenotype ascertainment is suspected. Cytogenetic analyses, such as karyotyping, are gold standard methods of inferring sex (Nagy et al., 2015). They allow not only detection of the X and Y sex chromosomes, but also visualization of potential aneuploidy. On the other hand, high-throughput genomic experiments introduce more uncertainty in sex inference when chromosomes are not evaluated in their entirety, but rather as oligonucleotide fragments.

Single-nucleotide polymorphism (SNP) genotype arrays for GWAS are one such example. Since males have an unpaired X chromosome, hybridization of their X chromosome fragments to SNP array probes is ideally expected to produce 0% heterozygous signals. With two X chromosomes, females should display heterozygosity that is much higher than 0%. In reality, males do not display precisely 0% X chromosome heterozygosity (XH) due to infrequent platform errors. The software PLINK infers sex based on the assumption that males should on average have lower XH compared to females (Purcell et al., 2007). PLINK continues to be one of the most popular tools for GWAS QC given its ease of use and legacy status among the earliest genome analysis software. However, improvements in sex inference from genotype arrays are warranted, since valuable information on the Y chromosome has yet to be leveraged. Although other software such as GenomeStudio (Illumina) and SNP & Variation Suite (Golden Helix) do facilitate examination of hybridization fluorescence intensities on the X and Y chromosomes, raw genotype data are required. Allele calls (e.g. PLINK format and GenomeStudio matrix format) tend to be much more accessible to researchers for shared use than the raw data that genotyping sites often harbor privately.

We propose a new sex inference tool, seXY, that accepts called genotype data and jointly considers information on the X and Y chromosomes. Beyond XH, seXY also accounts for Y chromosome missingness (YM). In spite of pseudoautosomal regions from which X chromosome fragments can cross-hybridize with Y chromosome array probes, females should exhibit substantially greater YM than males. To our knowledge, accuracies of existing array-based sex inference software have never been assessed using a reference that is more reliable than self-reported sex. In this Application Note, we compare the performances of PLINK and seXY when applied to X and Y chromosome SNP array data from the prostate and ovarian cancer projects of the OncoArray Consortium (Amos et al., 2016). Individual sex was established based on verified prostate or ovary presence.

2 Methods

SNP array data were downloaded for 910 males (Prostate Cancer Batch 1, Project Code 762, contact Ros.Eeles@icr.ac.uk) and 5,403 females (Ovarian Cancer Batch 3, Project Code 901, contact Catherine.Phelan@moffitt.org) in Illumina *.idat format. Twenty-one male and 42 female samples were removed for originating from the HapMap project and/or for not having consent forms. Genotype calls at the 15,258 X chromosomes markers and 397 Y chromosome markers of both datasets were converted to matrix format using GenomeStudio v2011.1 with the default QC setting GenCall score greater than 0.15. For every individual, seXY computed XH as the fraction of all markers on the X chromosome that have two different allele calls, excluding markers with missing calls. YM was computed as the fraction of all markers on the Y chromosome that have missing calls. Two-fold cross-validation (CV) was then performed using the following logistic regression model:

$$\ln \frac{P(Sex_i = \text{female})}{1 - P(Sex_i = \text{female})} = \alpha + \beta_1 \cdot XH_i + \beta_2 \cdot YM_i \quad \text{(Equation 1)}$$

Individual *i*'s sex was inferred to be female if $P(Sex_i = \text{female}) > 0.5$; otherwise, sex was inferred to be male. Due to the asymmetry of available genotype data, training set females greatly outnumber training set males in 50/50 two-fold CV. It has been shown that highly unbalanced training sets have the potential to impair classification accuracy in test sets, regardless of class proportions in the test sets (Wei and Dunbrack, 2013). In order to achieve more balanced training sets, a modified version of 80/20 CV was also performed. Implemented over 5 rotating rounds as usual, 80% of the males (711 individuals) and 20% of the females (1072 individuals) formed training sets to fit Equation 1 for evaluation on remaining individuals.

3 Results

As expected, XH and YM plot as distinct clusters for the majority of males and females (Supplementary Figure S1). X chromosome markers with high minor allele frequencies demonstrated the largest difference in heterozygous prevalence between males and females (Supplementary Figure S2). Both PLINK and seXY inferred male sex with nearly 100% accuracy (Table 1). PLINK misclassified the two highest-XH males, while seXY did not. By taking into account YM, seXY consistently outperformed PLINK in accurately classifying females. The few females misclassified by seXY have XH and YM values that closely mirror those of males.

The extent of unbalanced sex proportions in training sets did not influence prediction performance. Results were similar across all rounds of CV. We have therefore made seXY available for public use as Equation 1 trained on all 889 males and 5,361 females in this study. Accuracy was ensured to not be dependent on markers that are inherently specific to the OncoArray platform, as seXY was robust to 10%, 25%, and 50% random omission of markers (Supplementary Table S1).

Table 1. Comparison of sex inference accuracy using PLINK versus seXY.

	PLINK		seXY	
	Male	Female	Male	Female
50/50 CV round 1	99.8	96.8	100.0	99.8
50/50 CV round 2	99.8	96.4	100.0	99.9
80/20 CV round 1	99.4	96.6	100.0	99.8
80/20 CV round 2	100.0	96.4	100.0	99.7
80/20 CV round 3	100.0	96.7	100.0	99.9
80/20 CV round 4	99.4	96.8	100.0	99.9
80/20 CV round 5	100.0	96.7	100.0	99.6

Accuracies are displayed as percent of test set individuals whose sexes were correctly predicted. CV, cross-validation.

4 Conclusion

While XH alone appeared sufficient for sensitively identifying males, female classification was improved through simultaneous consideration of YM by seXY. The 3% gain in accuracy among females can be attributed to those who have distinguishing YM despite low XH. For large GWAS consortia such as the OncoArray where hundreds of thousands of samples are interrogated, seXY may salvage up to several thousand samples from removal due to incorrect prior sex inference. Misclassification of the remaining <0.5% females with both low XH and low YM is likely caused by a combination of 46XX/46XY mosaicism, 45XO/46XY mosaicism, large-scale duplications or deletions, loss of DNA, and other errors in laboratory handling (Qu et al., 2011).

Acknowledgments

We would like to thank rotation student Alexander J. Titus for conceiving the name of this sex inference tool. We would also like to thank Katherine L. Nathanson, Simon A. Gayther, and the Endometrial Cancer Association Consortium led by Amanda B. Spurdle for contributing samples to the ovarian cancer OncoArray.

Funding

This work has been supported by National Institutes of Health grant P20GM103534 (CIA) and the Paul K. Richter and Evalyn E. Cook Richter Memorial Fund (JAB).

Conflict of Interest: none declared.

References

Amos,C.I. *et al.* (in press) The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol. Biomarkers Prev.*

Begum,F. *et al.* (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.*, **40**, 3777–3784.

Nagy,B. *et al.* (2015) Detection of sex chromosome aneuploidies using quantitative fluorescent PCR in the Hungarian population. *Clin. Chim. Acta*, **445**, 2–6.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Qu,C. *et al.* (2011) Cost-effective prediction of gender-labeling errors and estimation of gender-labeling error rates in candidate-gene association studies. *Front. Genet.*, **2**, 31.

Wei,Q. and Dunbrack,R.L.,Jr. (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS One*, **8**, e67863.