EWASex

An efficient R-package to predict sex in epigenome-wide association studies

Lund, Jesper; Li, Weilong; Mohammadnejad, Afsaneh; Li, Shuxia; Baumbach, Jan; Tan, Qihua

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btaa949

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Lund, J., Li, W., Mohammadnejad, A., Li, S., Baumbach, J., & Tan, Q. (2021). EWASex: An efficient R-package to predict sex in epigenome-wide association studies. *Bioinformatics*, *37*(14), 2075-2076. https://doi.org/10.1093/bioinformatics/btaa949

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

Subject Section

EWASex – An efficient R-package to predict sex in epigenome-wide association studies

Jesper Beltoft Lund[1], Weilong Li[1], Afsaneh Mohammadnejad[1], Shuxia Li[1], Jan Baumbach[2,4], Qihua Tan[1,3,*]

[1]Epidemiology & Biostatistics, Department of Public Health, University of Southern Denmark, [2]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Germany, [3]Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, [4]Computational BioMedicine lab, Department of Mathematics and Computer Science, University of Southern Denmark

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Epigenome-Wide Association Study (EWAS) has become a powerful approach to identify epigenetic variations associated with diseases or health traits. Sex is an important variable to include in EWAS to ensure unbiased data processing and statistical analysis. We introduce the R-package EWASex, which allows for fast and highly accurate sex-estimation using DNA methylation data on a small set of CpG sites located on the X-chromosome under stable X-chromosome inactivation in females. We demonstrate that EWASex outperforms the current state of the art tools by using different EWAS data sets. With EWASex, we offer an efficient way to predict and to verify sex that can be easily implemented in any EWAS using blood samples or even other tissue types. It comes with pre-trained weights to work without prior sex labels and without requiring access to RAW data, which is a necessity for all currently available methods.

Availability: https://github.com/Silver-Hawk/EWASex

Contact: qtan@health.sdu.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

More and more researchers have begun using high-throughput chips or arrays for analyzing biological materials. This has led to an exponential increase in genomic studies (Rossi and Grifantini, 2018) with datasets from large scale cohorts stored in publicly available repositories such as the Gene Expression Omnibus (GEO) (Barrett *et al.*, 2012), which are usually bundled with annotations such as gender.

Epigenome-wide association study (EWAS) analyzing DNA methylation (DNAm) is a well-accepted approach (Rakyan *et al.*, 2011; Lappalainen and Greally, 2017) and popular for studying epigenetics in disease etiology (Greenberg and Bourc'his, 2019). With the rapid growth in EWAS and available data, human and systematic errors are not rare to encounter. As a result, it is not unusual to observe suspectable errors in gender information based on the observed patterns, often leading to a required reduction in sample size due to exclusion or spurious results due to misinformation (Kim et al. 2016). Consequently, a set of methylation-based gender estimators has already been implemented in R-packages such as *minfi* (Aryee *et al.*, 2014) and *sEst* (Jung *et al.*, 2018). However, they rely on raw original information that is not always available in public repositories. We recently found a set of markers that are significantly and stably methylated in females due to X-chromosome inactivated (XCI) and distinguishable between the genders (Li *et al.*, 2020), which sprouted the idea for *EWASex*.

2 Results

We found a set of 49 CpGs that best predict gender with 100% accuracy in the Middle-Aged Danish Twins (MADT) (Skytthe *et al.*, 2013)(McGue and Christensen, 2007), Birthweight Discordant Twins (BWT) (Tan *et al.*, 2014), and Longitudinal Study of Aging Danish Twins 2 (LSADT2) (Christensen *et al.*, 1999) participants (Supplementary Figure 1a). The process of how the 49 CpGs were selected is reported in the supplementary text.

Table 1 displays the accuracy comparisons between the existing *sEst* and *minfi* methods, and the presented *EWASex* method. All three algorithms were executed with default parameters, and the RAW datasets were processed using minfi, without additional normalization. EWASex used

the supplied weights trained for *MADT*, *BWT*, *LSADT2* for all datasets tested. Note that only 43 of the 49 CpGs are available for the Illumina 850K array. Additional information about each dataset can be found in the supplementary text. EWASex performs equally or better on all blood-based datasets tested, with slightly lower accuracy for GSE61107 on brain and GSE151407 on muscle tissues. Additional information about each dataset can be found in the supplementary text.

Table 1: Accuracy of the comparisons between *sEst*, *minfi*, and the presented *EWASex* method.

Dataset	*N*	*Tis.*	*Chip*	*sEst*	*minfi*	*EWASex*
MADT[$]	490	Bl	450K	**100%**	**100%**	**100%**
BWT[$]	310	Bl	450K	99.67%	**100%**	**100%**
LSADT2[$]	224	Bl	450K	**100%**	**100%**	**100%**
GSE51032[*]	845	Bl	450K	98.57%	95.97%	**98.81%**
GSE42861	689	Bl	450K	**100%**	**100%**	**100%**
GSE43976	95	Bl	450K	**100%**	**100%**	**100%**
GSE68777	40	Bl	450K	**100%**	**100%**	**100%**
GSE157341	274	Li	450K	98.90%	98.17%	**99.27%**
GSE61107	48	Br	450K	**95.83%**	**95.83%**	93.75%
GSE131433	440	Bl	850K	98.86%	99.09%	**99.31%**
GSE132181	392	Bl	850K	**100%**	**100%**	**100%**
PCOS	116	Bl	850K	**100%**	**100%**	**100%**
GSE143157	156	Br	850K	**98.71%**	**98.71%**	**98.71%**
GSE151407	78	Mu	850K	**100%**	**100%**	97.43%

N:* Number of samples. *Tis.*: Tissue; Blood (Bl), Brain (Br), Liver (Li), or Muscle (Mu). . [$]:** Used for finding the used 49 CpGs.:** *sEst* and *minfi* accuracy was calculated based on the reported numbers in (Jung *et al.*, 2018), as raw idat files are not available. The highest accuracies are bolded.

3 Discussion

Sex prediction using the XCI CpGs offers a natural way to verify sex through not only statistical modelling but also a biological phenomenon. As a result, our method is characterized by (1) a small number of features to use; (2) simple prediction model building and application; but (3) improved accuracy in blood-based datasets; additionally (4) direct use of raw methylation data (idat files from Illumina methylation chips) before normalization, which avoids introducing wrong sex labelling in the normalization process as information on sex is required by some of the normalization schemes e.g. functional normalization (Fortin et al., 2014). And finally (5) *EWASex* works on the beta-values, either before, during, or after normalization, enabling gender predictions on public datasets where raw methylation channels are not available, which in contrast, is not possible for currently available methods. In conclusion, *EWASex* is to this date, the most lightweight (59KB including data) and accurate tool for whole blood DNAm gender predictions. Its weights are easily transferable between datasets. Iit works with all normalization methods (Supplementary Figure 5) and datasets of all ages (Supplementary Figure 3), while both installation and prediction take literally milliseconds to complete on a standard laptop. The method is also applicable to data on non-blood tissues but prediction accuracy can be slightly lower. . Extended explanations are available in the supplied Supplementary Text.

Software availability

The *EWASex* R-package along with tutorials, documentation, and source code are available at https://github.com/Silver-Hawk/EWASex.

Acknowledgements

This work was supported by the Velux Foundation research grant #000121540. JB is grateful for the financial support of H2020 grant nr. 777111 (RepoTrial) and BMBF grants Sys_Care and SymBoD.

Funding

QT, SL, KC, JB conceived the study and proposed the analysis. JL, WL, AM performed data analysis. KC, QT, SL provided biological inputs and JB provided guidance for the analysis. JL wrote the paper. All authors read and approved the manuscript. *Conflict of Interest:* none declared.

References

Aryee, M. J. *et al.* (2014) 'Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays', *Bioinformatics*, 30(10), pp. 1363–1369. doi: 10.1093/bioinformatics/btu049.

Barrett, T. *et al.* (2012) 'NCBI GEO: archive for functional genomics data sets—update', *Nucleic Acids Research*. Narnia, 41(D1), pp. D991–D995. doi: 10.1093/nar/gks1193.

Christensen, K. *et al.* (1999) 'A danish population-based twin study on general health in the elderly', *Journal of Aging and Health*, 11(1), pp. 49–64. doi: 10.1177/089826439901100103.

Fortin, J.-P. *et al.* (2014) 'Functional normalization of 450k methylation array data improves replication in large cancer studies', *Genome Biology*. BioMed Central Ltd., 15(11), p. 503. doi: 10.1186/s13059-014-0503-2.

Greenberg, M. V. C. and Bourc'his, D. (2019) 'The diverse roles of DNA methylation in mammalian development and disease', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 590–607. doi: 10.1038/s41580-019-0159-6.

Jung, C. H. *et al.* (2018) 'sEst: Accurate sex-estimation and abnormality detection in methylation microarray data', *International Journal of Molecular Sciences*. MDPI AG, 19(10). doi: 10.3390/ijms19103172.

Kim, J. H., Park, J.-L. and Kim, S.-Y. (2016) 'Non-negligible Occurrence of Errors in Gender Description in Public Data Sets', *Genomics & Informatics*. Korea Genome Organization, 14(1), p. 34. doi: 10.5808/GI.2016.14.1.34.

Lappalainen, T. and Greally, J. M. (2017) 'Associating cellular epigenetic models with human phenotypes', *Nature Reviews Genetics*. Nature Publishing Group, pp. 441–451. doi: 10.1038/nrg.2017.32.

Li, S. *et al.* (2020) 'Exploratory analysis of age and sex dependent DNA methylation patterns on the X-chromosome in whole blood samples', *Genome Medicine*. BioMed Central, 12(1), p. 39. doi: 10.1186/s13073-020-00736-3.

McGue, M. and Christensen, K. (2007) 'Social activity and healthy aging: A study of aging Danish twins', *Twin Research and Human Genetics*, 10(2), pp. 255–265. doi: 10.1375/twin.10.2.255.

Rakyan, V. K. *et al.* (2011) 'Epigenome-wide association studies for common human diseases', *Nature Reviews Genetics*. Nature Publishing Group, pp. 529–541. doi: 10.1038/nrg3000.

Rossi, R. L. and Grifantini, R. M. (2018) 'Big Data: Challenge and Opportunity for Translational and Industrial Research in Healthcare', *Frontiers in Digital Humanities*. Frontiers Media SA, 5(13), p. 13. doi: 10.3389/fdigh.2018.00013.

Skytthe, A. *et al.* (2013) 'The Danish Twin Registry: Linking surveys, national registers, and biological information', *Twin Research and Human Genetics*, 16(1), pp. 104–111. doi: 10.1017/thg.2012.77.

Tan, Q. *et al.* (2014) 'Epigenetic signature of birth weight discordance in adult twins', *BMC Genomics*. BioMed Central Ltd., 15(1). doi: 10.1186/1471-2164-15-1062.