Using XIST, X-linked, and Y-linked genes to predict sex phenotype and complements for cancer cell lines

Ali Termos

Abstract

This study utilized the comprehensive dataset from the Cancer Cell Line Encyclopedia (CCLE) to develop a “cell line sex classifier prototype”. The focus was on the differential gene expression patterns linked to the X and Y chromosomes in cancer cell lines. Through rigorous analysis, the study revealed notable variations in classifier performance, particularly influenced by missing data and gender differences. The model demonstrated high precision in identifying male cell lines but encountered challenges in accurately predicting female ones, indicated by a substantial false discovery rate. The development of the sex classifier highlights the necessity of incorporating sex-based analyses in cancer research. The study calls for further refinement of predictive models to enhance their sensitivity and accuracy in recognizing the complexities inherent in sex differences in cancer biology.

Keywords. Cancer Cell Line Encyclopedia (CCLE), Sex Classifier Prototype, Gene Expression Patterns, XIST Gene, Predictive Models, False Discovery Rate.

Introduction

In a recent review study by Lopes-Ramos et al. (2020), sex-based disparities in cancer have been thoroughly examined; highlighting how males often face higher cancer incidence and mortality rates compared to females, even among cancer types that are prevalent in females. What is more alarming in the review is that sex-biases are heavily reflected in treatment responses and drug efficacy (Lopes-Ramos et al., 2020), yet studies in sex differences within cancer is a recent emerging field. In this direction, recognizing and incorporating sex-biases to strategies and interpretations of results in cancer biology is pivotal in advancing cancer research and developing better individualized treatments.

In this study, we analyzed a large data set of cancer cell lines from the Cancer Cell Line Encyclopedia (**CCLE**) and constructed a "cell line sex classifier prototype" that predicts the phenotype and a possible sex compliment for a given cell line in the data. This paper addressed several key questions: What candidate genes have been selected for classification, and why? What criteria did the model used for classification? What was the overall accuracy of the model? And what were the false discovery rates for male and female predictions made by the model?

Methods

We attempt to analyze the RNA-Seq read counts spanning various genes for a selection of cell lines. For this purpose, the 2018 Cancer Cell Line Encyclopedia (CCLE) gene count data is retrieved from Depmap portal - with **raw counts** of 56,202 genes and **annotation data** of 1,461 cancer cell lines. The gene count data was a result of RNA Sequencing performed on Illumina's HiSeq 2000/2500 - gathering at least 100 million sequences, 101 nucleotides each, per sample (Mahmoud et al., 2019). These sequences were mapped to the human GRCh37 genome using STAR 2.4.2a (Mahmoud et al., 2019). The data was imported to R-studio (R version 4.3.1) for pre-processing, data manipulation, exploratory data analysis, model construction, model testing, and generating various results. Refer to **Supplementary Note 1** for details on generated results. For the remaining of this methods section, we will go through the selection of candidate genes and the criteria of classification.

Selection of candidate genes

We examined candidate genes on both sex-chromosomes (X and Y) that can help us differentiate between the two biological sexes. In context of the current literature, this was a step filled with significant variability of choices that depended on a thorough interpretation of the expression levels of different genes in typical versus atypical (cancer state) scenarios. The fact is,

there is no one selection of genes that is deemed perfect, as the list can grow or shrink depending on new discoveries about genes and ones' plausible interpretability of genes and their selection. Primarily, we were interested in *genes that are not tissue-specific, are primarily expressed in adults, and share the least sequence similarity between the X and Y regions*. The reason for establishing such constraints was to be able to differentiate between the two sexes with sufficient certainty.

The sex chromosomes have evolved from a set of autosomal chromosome pairs; except for two small regions (PAR1 and PAR2) found on the tip of both ends of each of the X and Y chromosomes (Tricarico et al., 2020). Genes on the PAR regions behave like autosomal genes (hence the name) - with corresponding genes on those regions sharing extreme sequence similarities. For this reason, selecting genes that lie on the PAR regions is trivial to our study. On the flip side, genes that lie outside the PAR regions are commonly known as X-linked and Y-linked genes – these gene tend to show little to no similarities among their paralogs and are of interest to this study. Following the results of a recent study by Godfrey et al. (2020), quantitative profiles of Y-linked gene expression was examined across 36 human tissues from hundred individuals - showing significant differences in expression levels between paralogs for the selected genes and there X paralogs. In our study, we based our definition of the chosen X and Y linked gens in accordance with the insights draw from Godfrey et al. For simplicity and a consequential limitation, we only considered existing equivalent X-linked paralogs of the following Y-linked genes. We used the sum expression levels for each gene category below (sumXLinked and sumYLinked) to denote the corresponding expression levels for each of the X and Y chromosomes:

- Tissues where Y-linked genes show highest expression in males:
    - Testis: DDX3Y, ZFY, SRY, TGIF2LY
    - Skin: UTY, PRKY
    - Brain: USP9Y, PCDH11Y, NLGN4Y
    - Prostate: RPS4Y2
    - Adrenal Gland: KDM5D
    - Breast: RPS4Y1
    - Thyroid: TBL1Y
    - Heart: EIF1AY
    - Colon: TMSB4Y
- Corresponding X-Linked genes:
    - DDX3X, ZFX, TGIF2LX
    - PRKX
    - USP9X, PCDH11X, NLGN4X
    - RPS4X
    - KDM5C
    - TXLNG

- TBL1X
- EIF1AX
- TMSB4X

In a typical non-cancerous state, females possess two X chromosomes (XX complements), while males have one X and one Y chromosome (XY complements). However, to balance the gene dosage with males, only one X chromosome is active in females, a process known as X chromosome inactivation (XCI). This is where the XIST gene plays a crucial role. Located in the X-inactivation Center (XIC), it produces a long non-coding RNA that inactivates one X chromosome in cells with multiple X chromosomes, thereby preventing the overexpression of X-linked genes. In contrast, cancerous states often deviate from this norm. For instance, X chromosome reactivation is observed in females (Spatz et al., 2004), and the loss of the Y chromosome in males (LOY) is noted in various cancer types (Abdel-Aziz et al., 2023; Dunford et al., 2017). Consequently, we have included the XIST gene in our study due to its dynamic role in interpreting expression levels. This concludes our discussion on gene selection, which encompassed the XIST gene, a set of X-linked genes, and a collection of Y-linked genes.

Criteria for classification

In the previous subsection, we looked at why we chose XIST, a group of X-linked genes, and a group of Y-linked genes to be part of the classification model. In this subsection, we define the criteria of classification used by the model, given the choices of genes. The criteria of classification were accomplished by defining and testing against thresholds for expression levels (read counts) of the selected genes. The minimum and maximum levels of the inter-quartile range (IQR) were used to define the read count thresholds for the sum of X-linked expression (**Figure 2**). Given the bimodal nature of the distribution for both XIST and sum of Y-linked gene expressions, two different sets of thresholds were estimated by observation on the graph to define low and high thresholds for each (**Figure 2**). Accordingly, we defined the criteria of classification as follows:

- Male, XY:
  - **Reasoning**: Any level of X-linked gene expression is consistent with a single X chromosome. High Y-linked gene expression indicates the presence of a Y chromosome. No (NO) XIST expression is expected as there is only one X chromosome which remains active. This is consistent with an XY complement.
  - **Criteria**: [H(X) **OR** M(X) **OR** L(X)] **AND** H(Y) **AND** NO(XIST)
- Female, XX:
  - **Reasoning**: High or Medium X-linked gene expression is typical for females with two X chromosomes. Low or No Y-linked gene expression indicates mismatches in alignment or the absence of a Y chromosome respectively. High XIST

expression is expected due to the inactivation of one X chromosome. This is consistent with a typical XX female.
    - **Criteria**: [H(X) **OR** M(X)] **AND** [L(Y) **OR** NO(Y)] **AND** H(XIST)
- Male, XXrY:
    - **Reasoning**: High (H) or Medium (M) X-linked gene expression indicates the presence of at least one active X chromosome. High Y-linked gene expression confirms the presence of a Y chromosome. Low (L) XIST expression suggests the presence of a reactivated X chromosome. This is consistent with an XXrY complement.
    - **Criteria**: [H(X) **OR** M(X)] **AND** H(Y) **AND** L(XIST)
- Male, XXY:
    - **Reasoning**: High or Medium X-linked gene expression suggests the presence of more than one X chromosome. High Y-linked gene expression confirms a Y chromosome. High XIST expression indicates one of the X chromosomes is inactivated. This is consistent with an XXY complement.
    - **Criteria**: [H(X) **OR** M(X)] **AND** H(Y) **AND** H(XIST)
- Male, LOY:
    - **Reasoning**: Low X-linked gene expression suggests reduced X chromosome dosage, possibly due to loss. Low Y-linked gene expression indicates a loss of Y chromosome (LOY). No XIST expression is consistent with a single X chromosome and a bias towards a male choice. This is potentially an LOY outcome.
    - **Criteria**: L(X) **AND** L(Y) **AND** NO(XIST)
- Female, XXr:
    - **Reasoning**: High or Medium X-linked gene expression indicates two X chromosomes. Low or No Y-linked gene expression confirms the absence of a Y chromosome. Low or No XIST expression suggests X reactivation. This is consistent with an XXr complement.
    - **Criteria**: [H(X) **OR** M(X)] **AND** [L(Y) **OR** NO(Y)] **AND** [L(XIST) **OR** NO(XIST)]
- Female, X0:
    - **Reasoning**: Low X-linked gene expression is consistent with a single X chromosome. No Y-linked gene expression confirms the absence of a Y chromosome. Low or No XIST expression is expected as there is no need for X-chromosome inactivation with only one X chromosome. This is potentially an X0 outcome.
    - **Criteria**: L(X) **AND** NO(Y) **AND** [L(XIST) **OR** NO(XIST)]
- NA, NA:
    - **Reasoning**: all other combinations that fall outside the above are excluded, due to either lack of interpretability.

- **Criteria**: else assign NAs

Results

The analysis of the model's performance reveals a marked disparity in accuracy contingent on the inclusion of missing values (NAs) and the differentiation by gender. When NAs are considered in the model's predictions, accuracy is substantially reduced to 35%, as indicated by the red bar in the graph (**Figure 3**). In contrast, the model's accuracy improves to 69% when NAs are excluded from the analysis, as depicted by the green bar (**Figure 3**). This notable increase underscores the substantial impact that missing values can have on the perceived performance of a predictive model.

Furthermore, the model exhibits a pronounced gender bias in prediction accuracy. The False Discovery Rate (FDR), which measures the proportion of incorrect positive predictions, stands at a substantial 50% for female predictions, represented by the pink bar in the accompanying graph (**Figure 4**). This indicates that half of the instances predicted as female by the model were in fact not female. Conversely, the model demonstrates perfect precision for male predictions, with an FDR of 0, as illustrated by the blue bar (**Figure 4**).

Discussion

This stark contrast in FDR highlights a significant disparity, with the model being highly reliable in predicting males but less so for females. The comparison of these results illuminates the need for model refinement to address both the handling of missing values and the reduction of gender-based predictive inaccuracies. However, given the variability in the interpretability of gene selection and choice of criteria for classification, such refinements are always possible.

Conclusions

In conclusion, our study made use of the Cancer Cell Line Encyclopedia (CCLE) to develop a cell line sex classifier prototype, focusing on XIST gene, along with X and Y chromosome-linked gene expressions. The model's performance, influenced by missing values and gender, revealed a marked disparity in accuracy. Notably, the model demonstrated high precision in predicting male labels but faced challenges with female predictions, as reflected in a significant false discovery rate. This finding underscored the necessity for refined models that more accurately capture sex-based differences in cancer research. Nonetheless, our study added valuable insights to the field of sex-based cancer research. By developing a model to predict sex-linked gene expression patterns, we aimed to contribute to the field of cancer biology by

emphasizing the need for models that are not only accurate but also equally sensitive to the complexities of sex differences.

References

Abdel-Hafiz HA, Schafer JM, Chen X, Xiao T, Gauntner TD, Li Z, et al. Y chromosome loss in cancer drives growth by evasion of adaptive immunity. Nature. 2023;619: 624–631. doi:10.1038/s41586-023-06234-x

Dunford A, Weinstock DM, Savova V, Schumacher SE, Cleary JP, Yoda A, et al. Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nat Genet. 2017;49: 10–16. doi:10.1038/ng.3726

Godfrey AK, Naqvi S, Chmátal L, Chick JM, Mitchell RN, Gygi SP, et al. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. Genome Res. 2020;30: 860–873. doi:10.1101/gr.261248.120

Lopes-Ramos CM, Quackenbush J, DeMeo DL. Genome-Wide Sex and Gender Differences in Cancer. Front Oncol. 2020;10: 597788. doi:10.3389/fonc.2020.597788

Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, ... Todd R. Golub, Levi A. Garraway & William R. Sellers. 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508 (2019).

Spatz A, Borg C, Feunteun J. X-chromosome genetics and human cancer. Nat Rev Cancer. 2004;4: 617–629. doi:10.1038/nrc1413

Tricarico R, Nicolas E, Hall MJ, Golemis EA. X- and Y-Linked Chromatin-Modifying Genes as Regulators of Sex-Specific Cancer Incidence and Prognosis. Clin Cancer Res. 2020;26: 5567–5578. doi:10.1158/1078-0432.CCR-20-1741

Figures

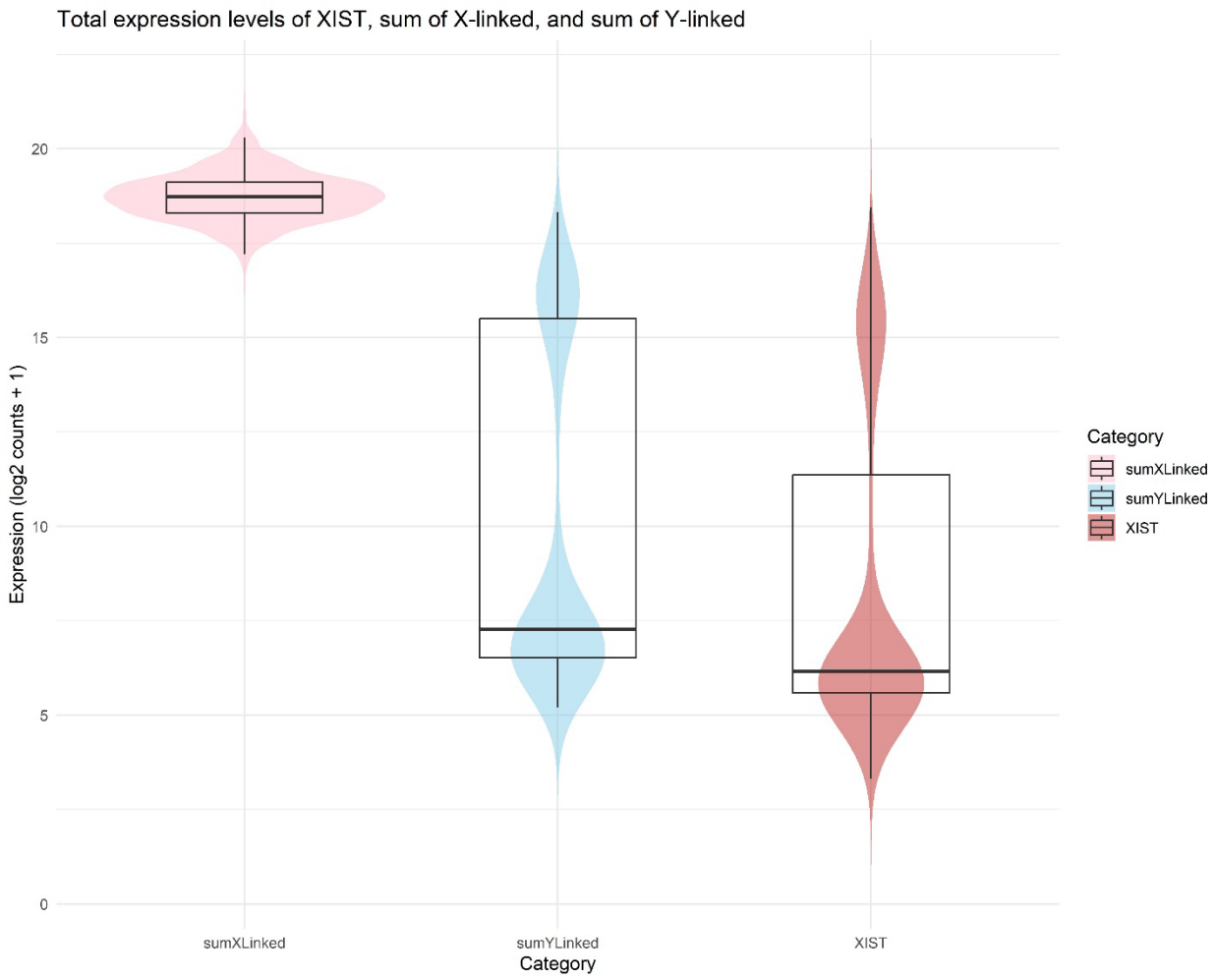

Figure 1: ***Expression levels of XIST, sum of X-linked genes, and sum of Y-linked genes across different categories.*** The violin plot overlays a box plot to provide a distribution overview and median values for each category. Sum of X-linked genes in pink, sum of Y-linked genes in blue, XIST expression is depicted in red. The box plot components within each violin plot illustrate the interquartile range (IQR) where 50% of the cell line read counts for each category exist. The median is indicated by a horizontal line within each box. Expression values are presented on a log2 scale with an added constant of 1 to accommodate zero counts.

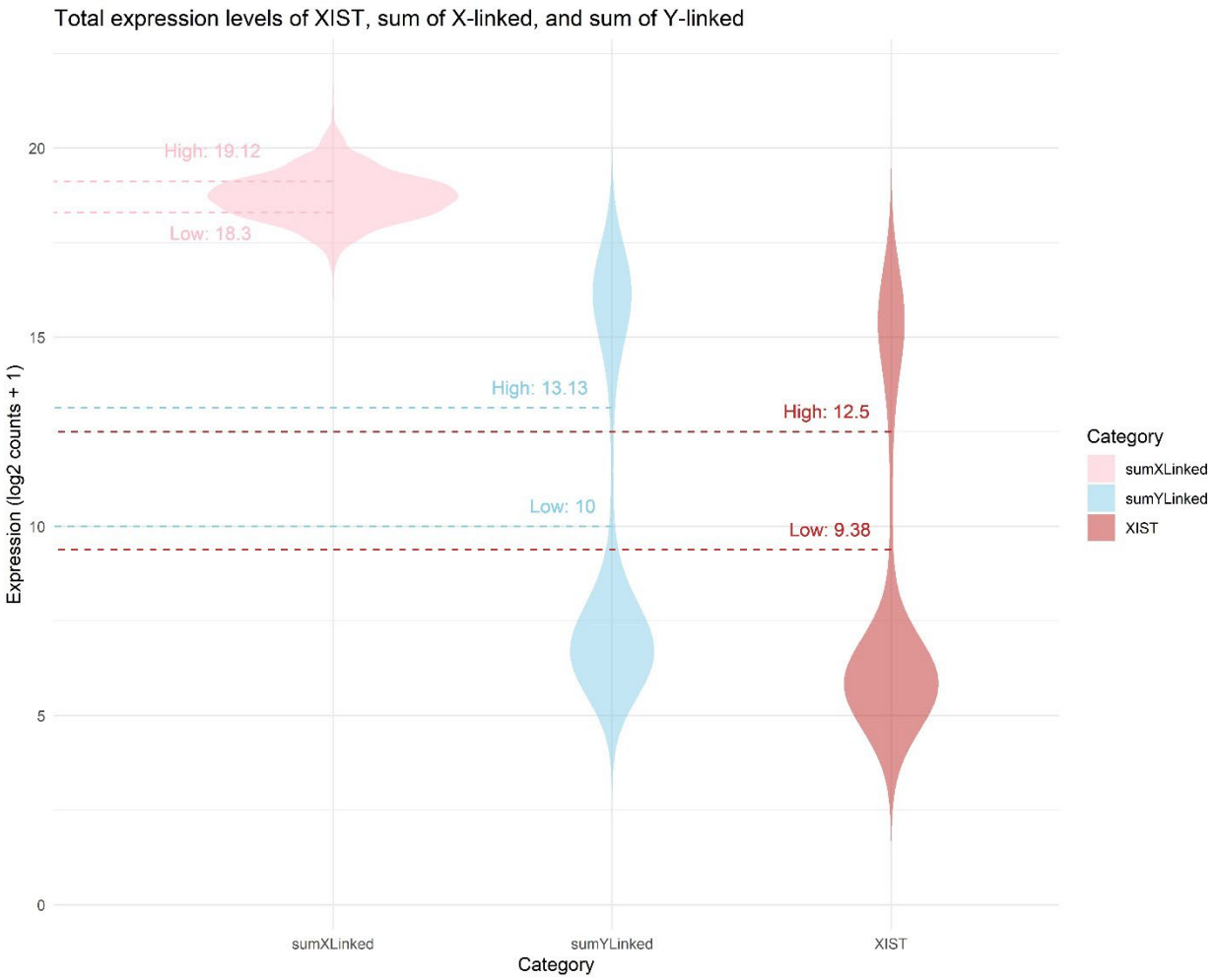

Figure 2: ***Comparative visualization of expression levels for XIST, sum of X-linked, and sum of Y-linked genes with thresholds.*** The plot for X-linked genes is shown in pink, sum of Y-linked genes in blue, and XIST expression in red. The threshold lines for high and low expression levels in each category are depicted as dashed lines. For sum of X-linked genes, the high expression threshold is marked at 19.12 and low at 18.3, for sum of Y-linked genes high at 13.13 and low at 10, and for XIST gene high at 12.5 and low at 9.38. These thresholds aid in distinguishing between different levels of gene expression. The data are transformed using a log2 scale (plus a constant of 1), facilitating the interpretation of expression levels across the categories.

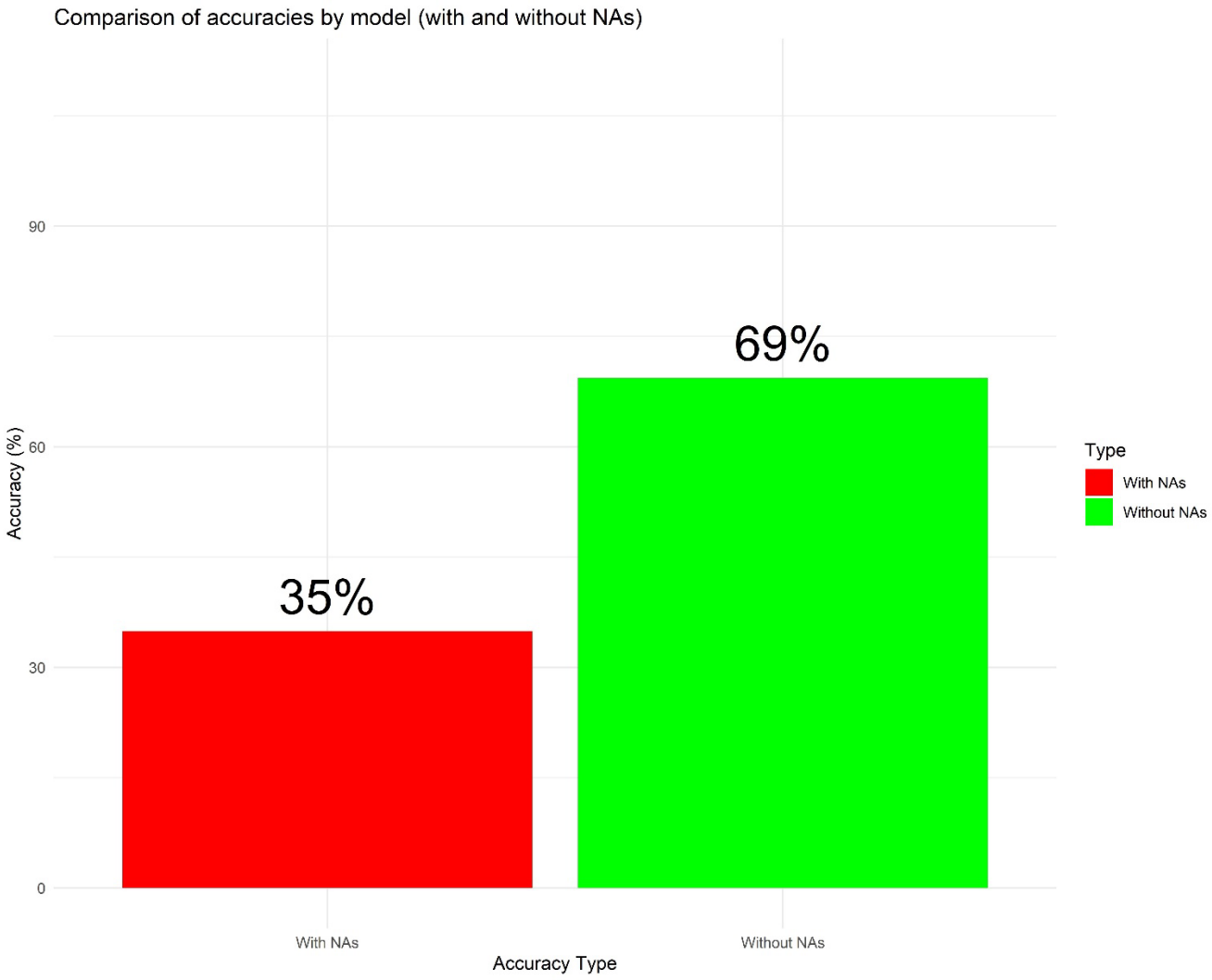

Figure 3: ***Bar graph representing the comparison of model accuracies, considering cases with missing values (NAs) and without.*** The red bar indicates the model accuracy including NAs at 35%, while the green bar represents the model accuracy excluding NAs, which stands at 69%. The percentage values are prominently displayed above each bar for clear and immediate comparison. This figure succinctly illustrates the impact of missing values on the accuracy of the model, highlighting the significant difference in performance metrics depending on whether NAs are accounted for in the accuracy calculation.

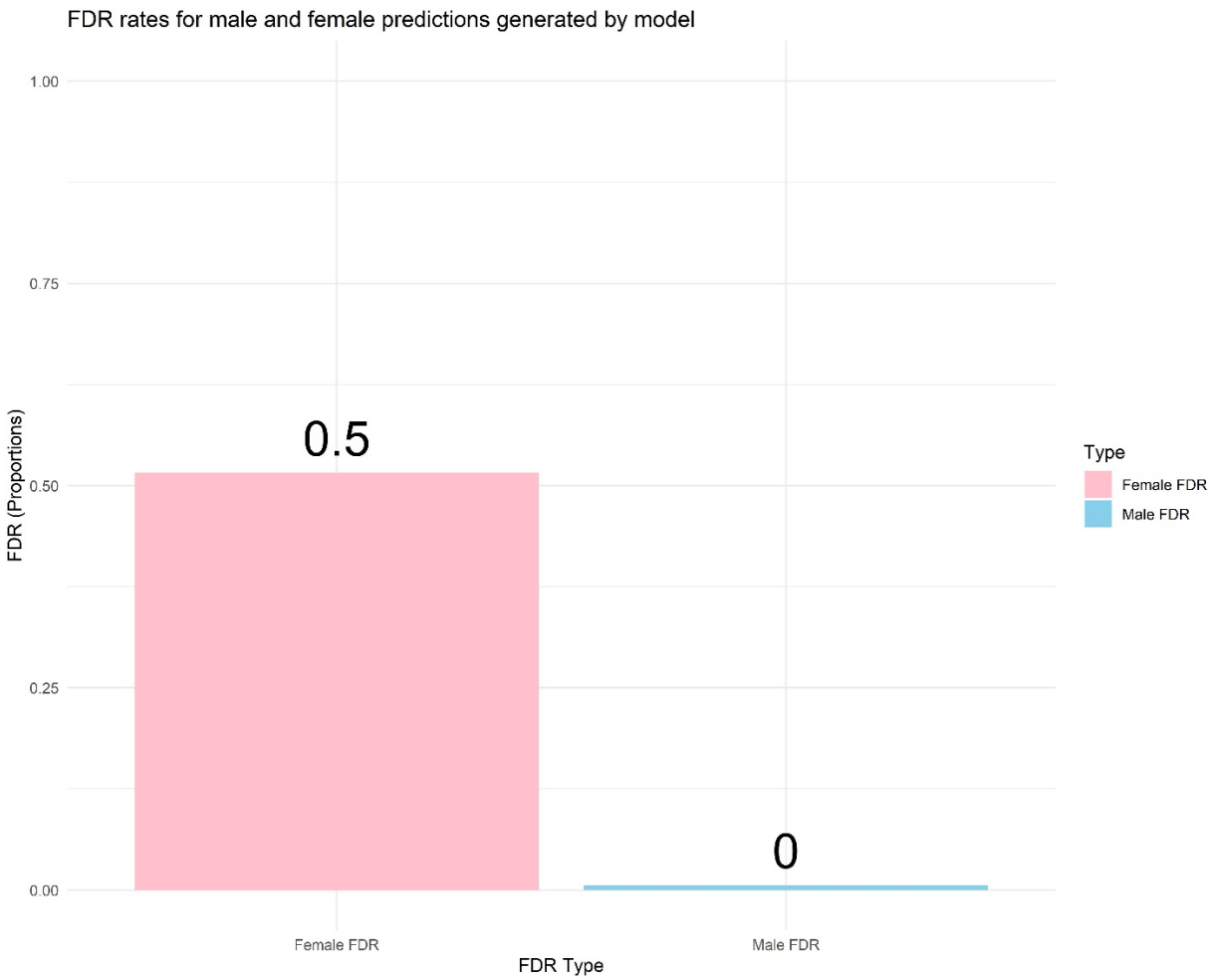

Figure 4: ***Bar graph showcasing the False Discovery Rate (FDR) for gender predictions made by the model.*** The pink bar represents the FDR for female predictions at a rate of 0.5, indicating that half of the positive predictions for females were false discoveries. In contrast, the blue bar indicates an FDR of 0 for male predictions, suggesting no false discoveries were made in predicting male labels. The absence of a visible bar for males emphasizes the model's precision in identifying male instances correctly. This visual comparison underscores the model's disparity in predicting female versus male classifications, with a significantly higher error rate for females.

Supplementary Materials

Supplementary Note 1. For reproducibility and/or validation purposes, the R script and related data files, used to generate the results of this study, can be found on the following GitHub page: **read-count-based_sex-classifier.**