Sex Chromosome Determination of Cancer Cells Using CCLE Expression Counts

Authors

Gloria Grama[a], Dr. Seema Plaisier[a], Dr. Melissa A. Wilson[a,b]

Affiliations

[a]*School of Life Sciences, Arizona State University, Tempe AZ*

[b]*Lincoln Center Applied Ethics, Arizona State University, Tempe AZ*

Abstract: Genomic instability is an established hallmark of cancer, however, aberrant sex chromosome complements observed in cancer cells and their subsequent effects are largely under researched. X inactivation escape and loss of Y mutations seen in cancer cells have been observed to alter the prognosis of patient outcomes, yet sex as a biological factor continues to be predominately overlooked. Expression counts for eight sex linked genes: *XIST, DDX3Y, EIF1AY, KDM5D, ZFY, RPS4Y1, USP9Y,* and *UTY* will be isolated in order to determine the sex chromosome complements present in 1,019 cell lines obtained from the Cancer Cell Line Encyclopedia (CCLE).

Keywords: cancer cell lines; sex chromosome determination; sex linked genes

1. Introduction

Genomic instability is an established hallmark of cancer, however, aberrant sex chromosome complements observed in tumor cells are largely ignored despite having significant effects on cancer prognoses[1]. Here we will identify conclusive sex linked genes to be used in predicting sex chromosome complements present in cancer line data provided by the Cancer Cell Line Encyclopedia (CCLE).

Cancer incidence rates between men and women is one of the most apparent sex differences observed, with men showing higher incidence of almost every cancer type with only a small subset occuring more commonly in women than men[2]. Research conducted has concluded that cancer cells evolving genetic mutations resulting in escape from inactivation of the additional X chromosome in women can support the reduced cancer incidence seen in women due to additional tumor suppressor genes located on it. The gene responsible for carrying out X inactivation is the long non-coding region RNA (lncRNA) *XIST*[3]. Studies have shown that there is no correlation between *XIST* expression and sex/gender in cancer cells, which results in this research also corroborated; however, high levels of expression have been linked to poor prognoses[4,5]. Male cancer cells have also been shown to lose their Y chromosome in loss of Y mutations (LOY). Tumors displaying LOY mutations have been found to have an increasingly immunosuppressive tumor microenvironment contributing to a more aggressive growth phenotype, resulting in a worse overall outcome in these individuals. However, LOY individuals were shown to be more responsive to certain immunotherapies[6]. Gene content analysis will be conducted on 1019 cell lines grown from tumor samples across a wide range of cancer types obtained from the CCLE[7]

[1] Rubin et al., "Sex Differences in Cancer Mechanisms."
[2] Rubin et al.
[3] Dunford et al., "Tumor Suppressor Genes That Escape from X-Inactivation Contribute to Cancer Sex Bias."
[4] Fang et al., "Upregulation of Tissue Long Noncoding RNA X Inactive Specific Transcript Predicts Poor Postoperative Survival in Patients with Non-Small Cell Lung Cancer."
[5] Zhu et al., "Prognostic and Clinicopathological Value of Long Noncoding RNA XIST in Cancer."
[6] Abdel-Hafiz et al., "Y Chromosome Loss in Cancer Drives Growth by Evasion of Adaptive Immunity."
[7] Barretina et al., "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity."

in order to determine the sex chromosomes present. The following ten sex linked genes: *XIST, DDX3Y, EIF1AY, KDM5D, NLGN4Y, RPS4Y1, TMSB4Y, USP9Y, UTY, and ZFY* will be isolated for analysis. High expression counts will be correlated with evidence that the corresponding sex chromosome is present, genes with the most distinct bimodal distributions will be ascertained and used for final analysis. Growing understanding of how sex chromosome differences impact and alter the function of cancer cells will aid in further tailoring of prognoses and therapies.

2. Methods

2.1 CCLE data

Gene expression data obtained from the CCLE was derived from RNA-seq reads aligned to the GRCh37 reference human genome using STAR 2.4.2a59. RNA-SeQC v1.1.860 was used to calculate gene level RPKM and read count values[8]. Alignment files were used to calculate gene abundance using the rsem-calculate-expression function and following steps of RSEM v.1.2.22 tool workflow outputting isoform-level expression counts in TPM[9]. All methods were run in accordance with the pipeline developed by the GTEx Consortium[10]. Gene expression values for: *XIST, DDX3Y, EIF1AY, KDM5D, NLGN4Y, RPS4Y1, TMSB4Y, USP9Y, UTY, and ZFY* were reported in RPKM and converted to log values. Reported sex associated with each cancer cell line's CCLE_ID derived from annotation data also found in the CCLE[11] was pulled in order to generate violin plots for each isolated gene using the ggplot function of the ggplot2_3.4.3[12] package of R software version 4.2.2[13] using Arizona State University's high performance cluster Sol.

[8] Ghandi et al., "Next-Generation Characterization of the Cancer Cell Line Encyclopedia."
[9] Li and Dewey, "RSEM."
[10] Aguet et al., "Genetic Effects on Gene Expression across Human Tissues."
[11] "Cancer Cell Line Encyclopedia (CCLE)."
[12] Wickham, "Ggplot2: Elegant Graphics for Data Analysis."
[13] "R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www.R-Project.Org/."

2.2 Thresholds

The violin plots generated (figure 1) displayed the bimodal distribution of each gene allowing for thresholds to be called which were used to assign sex based on the level of gene expression. For cell lines displaying expression counts over the now determined high threshold of the X-linked gene *XIST* were assigned "high_expression" equating to the presence of two X chromosomes, one which is inactivated. Cell lines below the low threshold were assigned "male_range" indicating likely presence of just one X chromosome or X inactivation escape. The nine remaining y-linked genes: *DDX3Y, EIF1AY, KDM5D, NLGN4Y, RPS4Y1, TMSB4Y, USP9Y, UTY, and ZFY* were assigned using slightly different conditions. Cell lines showing gene expression levels above the high threshold were assigned "high_expression" corresponding to the presence of the Y chromosome, those below the low threshold were assigned "female_range" corresponding to the absence of the Y chromosome. For both groups, expression levels falling in between designated thresholds were assigned "intermediate_range".

2.3 Sex/gender assignment

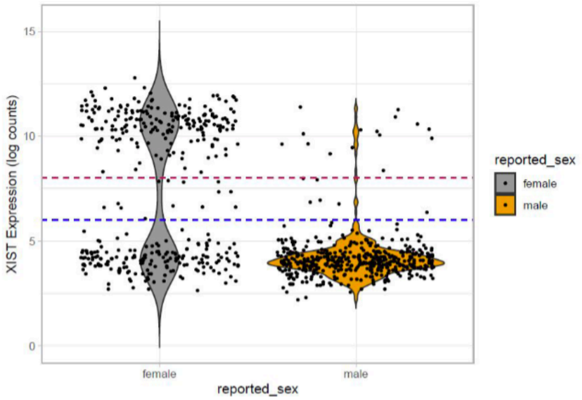
For *XIST* data, cell line expression counts above the high threshold were assigned "female" and those below were assigned "male". For y-linked genes, expression counts above the high threshold were assigned "male" and those below were assigned "female", intermediate values were not changed for both groups. Male and female assignments are meant to correlate with the associated chromosomes for their respective expression levels, ie. reported females expressing low levels of *XIST* fall within what would be expected of an XY individual, therefore determined to be a reported female, assigned male. However, if this experiment were to be conducted again, the exact chromosomes demonstrated to be present would be reported rather than assigning broader gender categories.

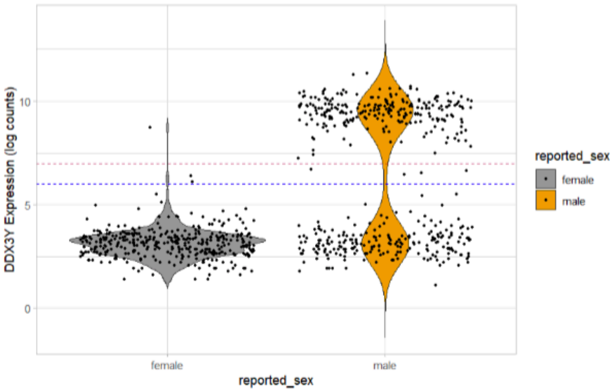
2.3 Gene determination

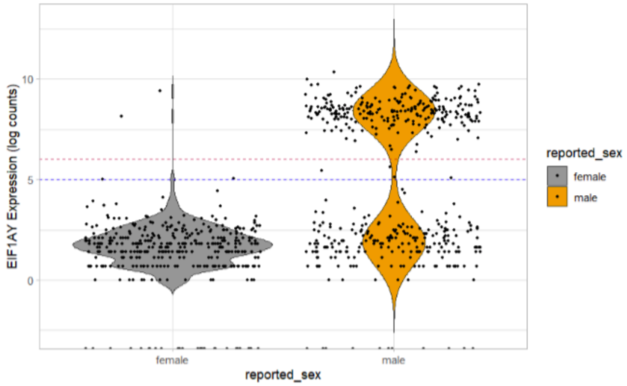
Cancer cell lines' reported sex, gene expression level values, and their corresponding assignments were then compiled for analysis. An upset plot using the package UpSetR_1.4.0 in the R software was produced to inspect commonly occurring trends. R's package dplyr_1.1.3[14] was used to generate data in table 2 assigning TRUE values to cell lines in which all genes analyzed made the same assignment call and FALSE if at least one gene differed. This data was then analyzed in combination with distributions observed in violin plots made earlier in order to ascertain which genes were most distinct and ideal for sex chromosome complement determination. R's package tidyr_1.3.0[15] was used to pull the assignment made most frequently across the eight genes to make the final determination. All codes used are available on GitHub: https://github.com/gloriagrama/compiled_gene_data

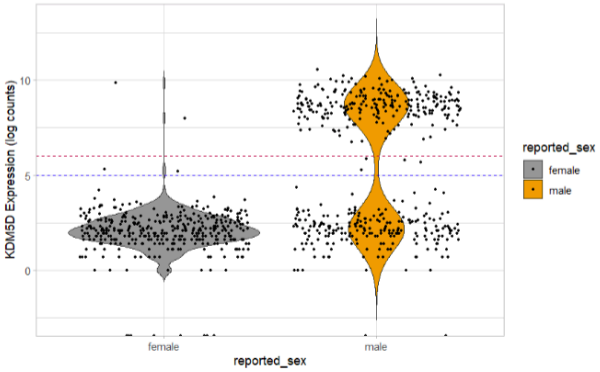
[14] Wickham et al., "Dplyr."
[15] Wickham et al., "Tidyr."

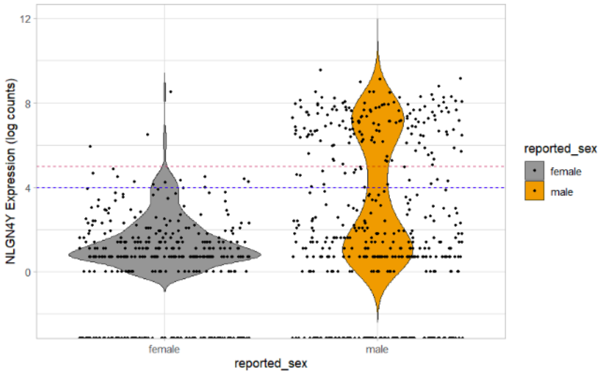
3. Results and Discussion

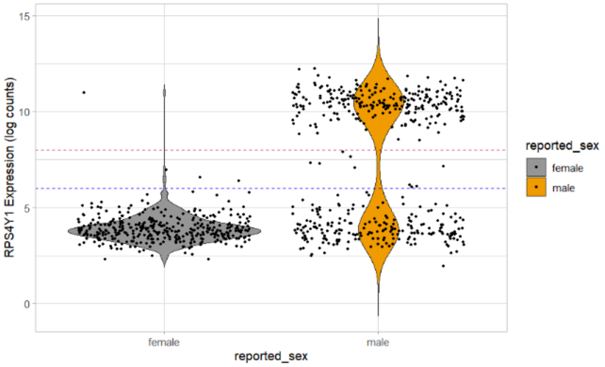


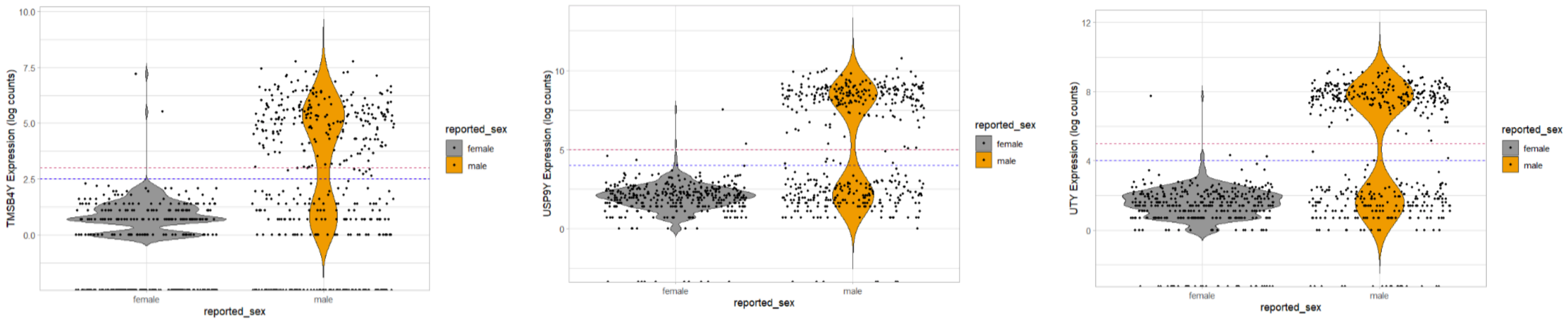












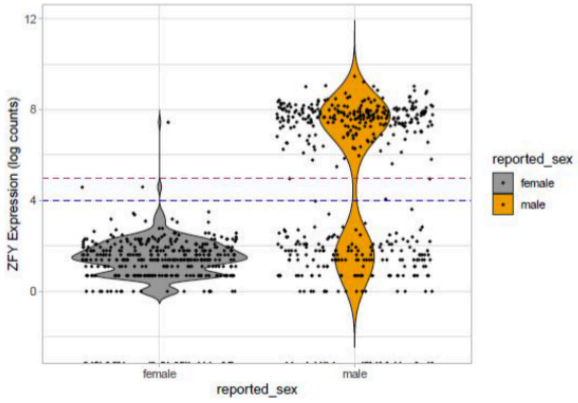


Figure 1. Violin plots for genes: *XIST, DDX3Y, EIF1AY, KDM5D, NLGN4Y, RPS4Y1, TMSB4Y, USP9Y, UTY, and ZFY* generated using ggplot in R plotting reported sex against gene expression levels. Thresholds plotted provide cut offs for high, low, and intermediate expression designations.

Table 1. Intermediate expression counts and threshold width per sex-linked gene

Gene	**Intermediate Expression Counts**	**Threshold Width (log)**
XIST	19	2
DDX3Y	7	1
EIF1AY	5	1
KDM5D	6	1
NLGN4Y	**33**	**1**
RPS4Y1	12	2
TMSB4Y	**12**	**0.5**
USP9Y	10	1
UTY	6	1
ZFY	5	1

Table 2. Number of cell lines in agreement, where all gene expression levels were within the same assignment threshold versus cell lines where at least one gene was in disagreement.

	Cell lines in agreement	**Cell lines in disagreement**
Initial genes	794	225
Final genes (without *NLGN4Y* and *TMSB4Y*)	922	97

The following genes: *XIST, DDX3Y, EIF1AY, KDM5D, ZFY, RPS4Y1, USP9Y,* and *UTY* were determined to be most decisive in determining the presence or absence of each corresponding sex chromosome. Distributions for the genes *NLGN4Y* and *TMSB4Y* (figure 1) were observed to be more scattered in comparison, indicating that determining the presence of each genes' corresponding sex chromosome may be more challenging and less accurate, therefore making them less ideal for these purposes. Something of note to mention, in figure 1, while bimodal expressions of male and female ranges can be observed, lower level expressions should theoretically be zero since reported males would not be expected to express any level of *XIST* or reported females any level of y-linked genes. Cell lines reporting levels of genes that should not be present is likely a result of mismapping to their associated gametolog[16]. For example, DDX3Y in females is likely mapping to the X chromosome gametolog DDX3X[17]; as for *XIST*, exact mechanisms for expression in male cell lines are not known, studies however have conferred similar results[18].

Data shown in Tables 1 and 2 corroborate the exclusion of *NLGN4Y* and *TMSB4Y*. *NLGN4Y* was found to have the highest number of intermediate expression values, indicating that its bimodal distribution is less distinct. *TMSB4Y* and *RPS4Y1* were both observed to have equally high intermediate expression counts, however, *TMSB4Y's* threshold is disproportionately smaller than *RPS4Y1's* when accounting for their different max expression values seen in figure 1 indicating *TMSB4Y's* values are more distributed. Despite *XIST* having more intermediate expression counts than *TMSB4Y* it was still included in the final genes used for analysis since it had one of the wider threshold widths used and the overall distribution seen in figure 1 was distinctly bimodal in comparison. Analysis was performed determining the ratio of cell lines completely in agreement, where every gene made the same sex chromosome assignment call, against cell lines in which at least one call differed. When *TMSB4Y* and *RPS4Y1* were removed from the data set, agreement amongst genes and cell lines grew from 77.92% to 90.48%. More genes could have been excluded from the data to attain an alpha level of 0.05, however, chromosome determination was based on the majority call made by all genes. Omitting too many genes could potentially make chromosome assignments less accurate by making outlier calls disproportionately more prevalent.

[16] Webster et al., “Identifying, Understanding, and Correcting Technical Artifacts on the Sex Chromosomes in next-Generation Sequencing Data.”
[17] Gelfand et al., “Survey of Commercial Antibodies Targeting Y Chromosome-Encoded Genes.”
[18] Sadagopan et al., “Somatic XIST Activation and Features of X Chromosome Inactivation in Male Human Cancers.”

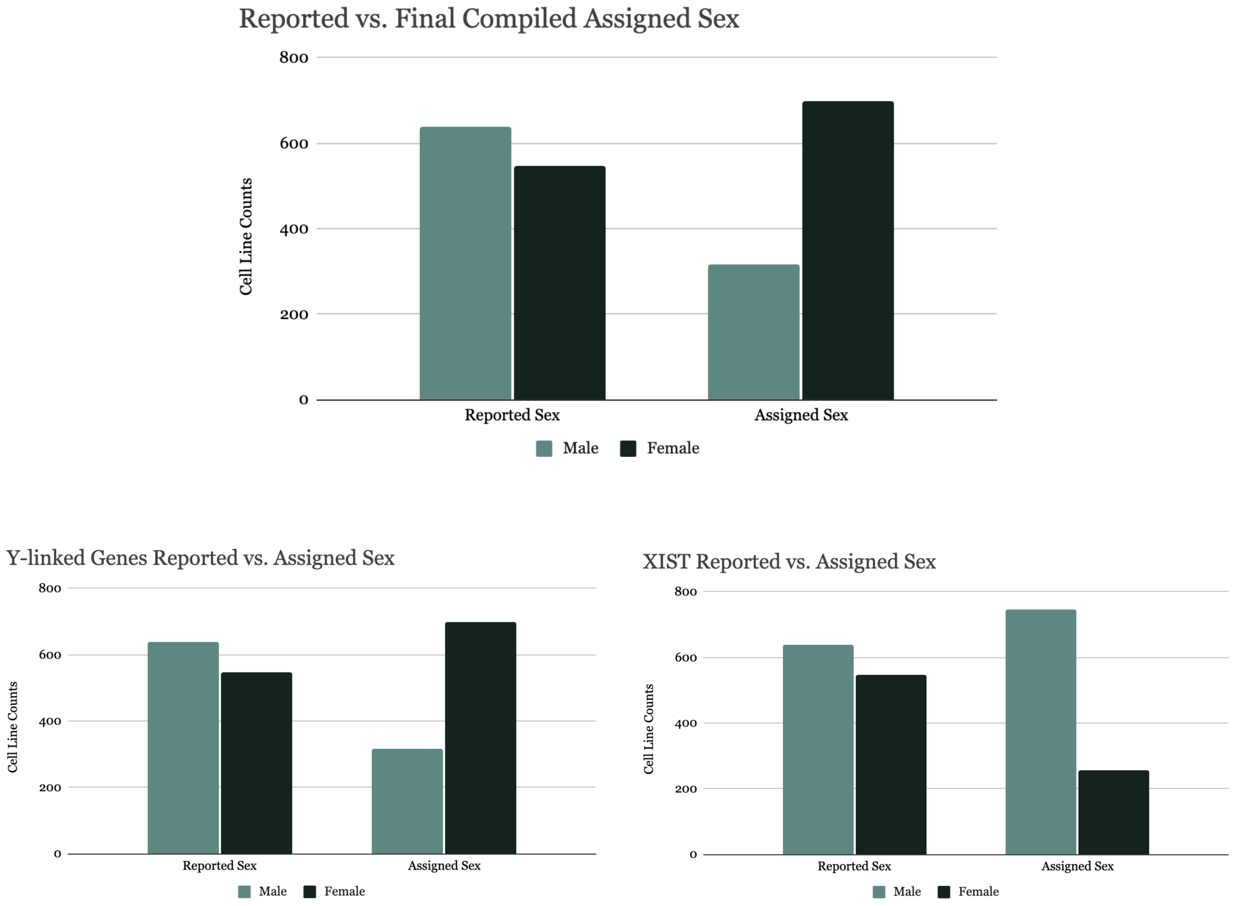


Figure 2. Total counts of reported versus assigned sex for three groups: combined final genes, and two subgroups, Y-linked genes and *XIST* gene data.

After excluding *NLGN4Y* and *TMSB4Y*, just under 150 additional cell lines were assigned female than reported, this pattern is closely mirrored in data displayed for y-linked genes exclusively (removing *XIST* data), likely due to these genes comprising a majority of the genes looked at. Using data displayed in figure 2, it was determined that 325 cancer cell lines originally reported male have lost their Y chromosomes. Interestingly, when solely examining *XIST* determinations, the opposite is true. Of the reported female population, the assigned female group is more than halved and 293 reported female cell lines have likely escaped X inactivation.

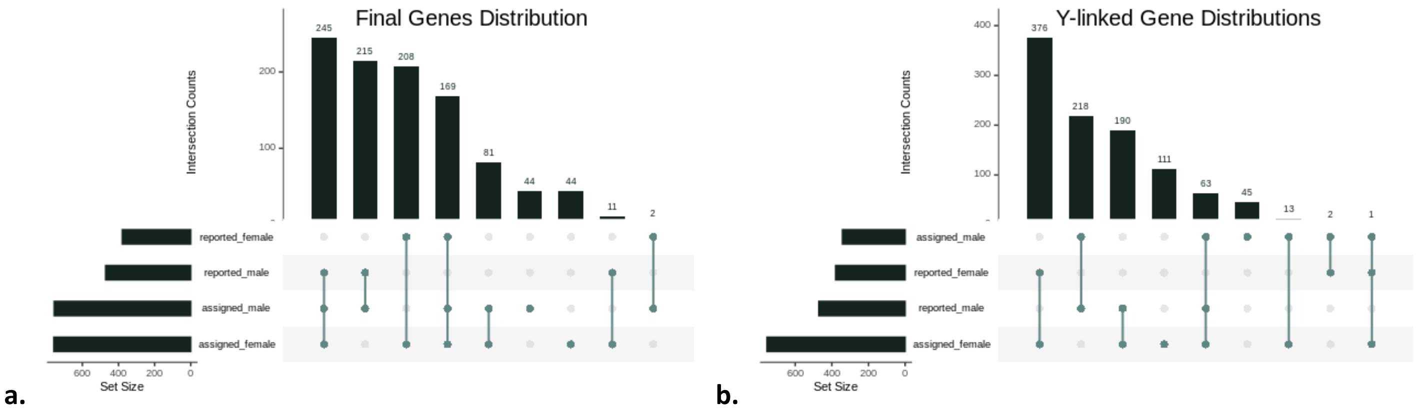


Figure 3. Upset plots displaying intersections for a.) final genes (without genes *NLGN4Y* and *TMSB4Y*) and b.) Y-linked genes (excluding *XIST* data)

The largest group accounted for in figure 3a. is composed of reported males assigned both male and female level gene expressions which is disappointing considering the data used for this plot excluded the genes *NLGN4Y* and *TMSB4Y.* It was suspected that *XIST* data was accountable since cell lines reported male could fall within the male range using *XIST* parameters but fall within female ranges for y-linked data (suggesting loss of Y chromosome and presenting evidence for only one X chromosome: XO individuals). Reported females falling in both assignment categories made up the fourth largest group in this data subset which could also be explained by contrasting calls made by *XIST* and y-linked genes. Combined, these groups make up almost half of cancer lines analyzed. This confounding variable was no longer observed to be as prevalent when *XIST* data was excluded in figure 2b. with 77 cancer lines making opposing assignments in comparison to 495 total in figure 3a. If this experiment were to be conducted again, sex chromosome assignments (XX_i, XX_a, XY, XO) would be determined rather than gender assignments (male/female). Only minor alterations to the original source code[19] would need to be made in order to produce remarkably more descriptive data.

[19] Grama, “Gloriagrama/Compiled_gene_data.”

4. Conclusion

The following sex linked genes: *XIST, DDX3Y, EIF1AY, KDM5D, RPS4Y1, USP9Y, UTY, and ZFY* can be instrumental in cancer cell analysis with the purpose of identifying expressed sex chromosome complements. Expression counts used in this analysis were obtained freely using already published, public data cultivated by the CCLE. Accurate methods of sex determination will provide integral knowledge of how genomic instability in cancer cells operate and the subsequent impacts it can have making treatment more effective and prognoses more informed. Sex as a biological variable is continually overlooked in science, not just in cancer research, growing consideration will make findings more encompassing and provide more nuanced explanations.

Acknowledgments

We thank NSF *Improving Undergraduate STEM Education* (IUSE) for funding this course based undergraduate research, thank you Dr. Wilson and Dr. Plaisier for developing this course and designing this experiment.

Author Contributions

Experimental design and data collected by Dr. Melissa A. Wilson and Dr. Seema Plaisier. Figure 1 generated by Dr. Melissa A. Wilson and Dr. Seema Plaisier. Figures 2-3 and tables 1-2 generated by Gloria Grama. Data analysis in Github repository source code conducted by Gloria Grama. Manuscript written by Gloria Grama.

References

Abdel-Hafiz, Hany A., Johanna M. Schafer, Xingyu Chen, Tong Xiao, Timothy D. Gauntner, Zihai Li, and Dan Theodorescu. “Y Chromosome Loss in Cancer Drives Growth by Evasion of Adaptive Immunity.” *Nature* 619, no. 7970 (July 2023): 624–31. https://doi.org/10.1038/s41586-023-06234-x.

Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550, no. 7675 (October 2017): 204–13. https://doi.org/10.1038/nature24277.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483, no. 7391 (March 2012): 603–7. https://doi.org/10.1038/nature11003.

“Cancer Cell Line Encyclopedia (CCLE).” Accessed November 23, 2023. https://sites.broadinstitute.org/ccle/.

Dunford, Andrew, David M. Weinstock, Virginia Savova, Steven E. Schumacher, John P. Cleary, Akinori Yoda, Timothy J. Sullivan, et al. “Tumor Suppressor Genes That Escape from X-Inactivation Contribute to Cancer Sex Bias.” *Nature Genetics* 49, no. 1 (January 2017): 10. https://doi.org/10.1038/ng.3726.

Fang, Hengxiao, Liushan Yang, Yue Fan, Chunrong Mo, Lei Luo, Daying Liang, and Yi Jiang. “Upregulation of Tissue Long Noncoding RNA X Inactive Specific Transcript Predicts Poor Postoperative Survival in Patients with Non-Small Cell Lung Cancer.” *Medicine* 99, no. 50 (December 11, 2020): e21789. https://doi.org/10.1097/MD.0000000000021789.

Gelfand, Bradley D., Dionne A. Argyle, Joseph J. Olivieri, and Jayakrishna Ambati. “Survey of Commercial Antibodies Targeting Y Chromosome-Encoded Genes.” bioRxiv, July 27, 2023. https://doi.org/10.1101/2023.07.26.550552.

Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, et al. “Next-Generation Characterization of the Cancer Cell Line Encyclopedia.” *Nature* 569, no. 7757 (May 2019): 503–8. https://doi.org/10.1038/s41586-019-1186-3.

Grama, Gloria. “Gloriagrama/Compiled_gene_data,” November 23, 2023. https://github.com/gloriagrama/compiled_gene_data.

Li, Bo, and Colin N. Dewey. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12, no. 1 (August 4, 2011): 323. https://doi.org/10.1186/1471-2105-12-323.

“R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www.R-Project.Org/.,” n.d.

Rubin, Joshua B., Joseph S. Lagas, Lauren Broestl, Jasmin Sponagel, Nathan Rockwell, Gina Rhee, Sarah F. Rosen, et al. “Sex Differences in Cancer Mechanisms.” *Biology of Sex Differences* 11, no. 1 (April 15, 2020): 17. https://doi.org/10.1186/s13293-020-00291-x.

Sadagopan, Ananthan, Imran T. Nasim, Jiao Li, Mingkee Achom, Cheng-Zhong Zhang, and Srinivas R. Viswanathan. “Somatic XIST Activation and Features of X Chromosome Inactivation in Male Human Cancers.” *Cell Systems* 13, no. 11 (November 16, 2022): 932-944.e5. https://doi.org/10.1016/j.cels.2022.10.002.

Webster, Timothy H., Madeline Couse, Bruno M. Grande, Eric Karlins, Tanya N. Phung, Phillip A. Richmond, Whitney Whitford, and Melissa A. Wilson. “Identifying, Understanding, and Correcting Technical Artifacts on the Sex Chromosomes in next-Generation Sequencing Data.” *GigaScience* 8, no. 7 (July 1, 2019): giz074. https://doi.org/10.1093/gigascience/giz074.

Wickham, H. “Ggplot2: Elegant Graphics for Data Analysis.” New York: Springer-Verlag, 2016.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan, Posit Software, and PBC. “Dplyr: A Grammar of Data Manipulation,” November 17, 2023. https://cran.r-project.org/web/packages/dplyr/index.html.

Wickham, Hadley, Davis Vaughan, Maximilian Girlich, Kevin Ushey, Posit, and PBC. “Tidyr: Tidy Messy Data,” January 24, 2023. https://cran.r-project.org/web/packages/tidyr/index.html.

Zhu, Jianwei, Fanyang Kong, Ling Xing, Zhendong Jin, and Zhaoshen Li. “Prognostic and Clinicopathological Value of Long Noncoding RNA XIST in Cancer.” *Clinica Chimica Acta; International Journal of Clinical Chemistry* 479 (April 2018): 43–47. https://doi.org/10.1016/j.cca.2018.01.005.

Supplementary Materials

Cell line description: https://www.researchsquare.com/blog/what-is-a-cell-line

Hallmarks of cancer: https://www.cell.com/action/showPdf?pii=S0092-8674%2800%2981683-9

Information on tumor microenvironments: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8194051/