Molecular Omics Resources Should Require Sex Annotation: A Call for Action

Kamila M. Bond[1,2], **Margaret M. McCarthy, PhD**[3,4], **Joshua B. Rubin, MD, PhD**[5,6], **Kristin R. Swanson, PhD**[1]

[1]Mathematical Neuro-Oncology Lab, Department of Neurological Surgery, Mayo Clinic, Phoenix, AZ, USA

[2]Mayo Clinic School of Medicine, Rochester, MN, USA

[3]Department of Pharmacology, University of Maryland School of Medicine, Baltimore, MD, USA

[4]Program in Neuroscience, University of Maryland School of Medicine, Baltimore, MD, USA

[5]Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA

[6]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA

Abstract

The most commonly-used omics databases are a compilation of results from primarily male-only and sex-agnostic studies. The pervasive use of these databases critically hinders progress towards fully accounting for the biology of sex differences.

Manuscript

Omics databases are widely used in life sciences research. Scientific investigators, some with limited bioinformatics experience, perform analyses with omics databases under the assumption that they are reliable, although that may not always be the case. For example, two COVID-19 research articles were recently retracted because analyses were based upon an unreliable data registry.[1,2] Concerningly, omics resources rarely provide sex annotation or allow for sex-specific analysis. This diminishes the value of these resources as we increasingly strive to incorporate sex as a biological variable in research. Here we aim to bring attention to the innate bias of omics resources and provide recommendations for addressing this limitation.

The problem

Sex differences in molecular, cellular, and organismal biology accrue from the time of fertilization and impact broadly on normal development.[3] Studying merged male and

female datasets can mask differences that are only revealed when each sex is considered individually.[4] Historically, male subjects have been overrepresented in animal and human research due to concerns that the hormonal variability of females confounds results[5], and the chromosomal sex of cell lines has largely been ignored.[6] Without justification, results from these male-dominant or sex-agnostic studies are assumed to apply equally to both sexes. When comparing female or mixed-sex data to a male standard, false negatives can arise or results may be misinterpreted (Figure 1). Conversely, there are instances when female subjects are overrepresented (e.g. breast cancer, autoimmune diseases), which results in bias against males. This inattention to sex in basic science studies has, in some cases, harmed patients[7,8] and may unintentionally be slowing scientific progress.

Some organizations have raised awareness of the importance of considering sex in research. The National Institutes of Health (NIH) now requires the incorporation of sex as a biological variable in the design of all funded studies[9], and the Horizon Europe program intends to do the same.[10] Some journals follow ARRIVE guidelines[11] and mandate disclosure of the sex of subjects used in the study. While these initiatives are important steps towards ensuring sex equity in research, they are not universally adopted and do not rectify the decades of biased work upon which current omics resources are built.

Current state of sex annotation in omics resources

Omics resources compile the results of thousands of studies to summarize biological relationships. While some investigators regularly consider sex as a biological variable, the NIH has determined that basic and preclinical research continues to suffer from the overrepresentation of males.[9] This in turn gives rise to bias in primary data repositories (e.g. GEO[12]) unless the resource requires sex annotation upon submission (e.g. TCGA[13], GTEx[14]).

There are currently 702 cataloged resources that collectively document all known biological pathways and molecular interactions across 24 organisms.[15] Of these, 370 (53%) provide references to the primary publications that originally described the knowledge. Amongst five of the most-cited resources from which several third-party analysis tools are built, all provide citations but none annotate the sex of the subjects that generated the results (Table 1).

While some resources with niche interests (e.g. DICE[16]) acknowledge the biological importance of sex and have incorporated it into their querying tools, most have yet to adopt this practice. These resources are often used for functional genomic analyses, so research that employs them -- even if sex is considered in the experimental design -- discounts the many molecular mechanisms by which male and female fundamentally differ. It is important to recognize that using these databases as a standard to evaluate both sexes may give rise to misleading results.

Mechanisms by which sex differences arise

At the most fundamental level, X-inactivation and the presence or absence of a Y chromosome drive sex determination. However, sex chromosomes alone cannot explain the innumerable differences between males and females. A striking example of this is androgen

insensitivity syndrome, a condition in which individuals have an XY karyotype but female characteristics due to a nonfunctional androgen receptor.

Across the genome, there are no sex differences in the frequency of single nucleotide polymorphisms[17], and only a few sex differences in rare copy number variations have been described.[18] There are conflicting reports of sex differences in telomere length, telomere attrition rate, and the relationships between telomeres and aging. Males and females accumulate nuclear and mitochondrial DNA mutations at different rates and loci which may contribute to differences in aging and oncogenesis.[19] While there is some sex-based variance in DNA, differences are largely thought to arise at the level of gene expression.[20,21]

When males and females have different fitness optima for the same trait, divergent evolutionary selection can cause sexual dimorphism in a characteristic that was once shared. These selective pressures may act on regulatory factors that can profoundly influence phenotype. Divergent evolution of regulatory factors is increasingly recognized as a contributor to sex differences[22], but their variability and poor characterization make them challenging to identify. Still, sex differences in both coding and regulatory regions have been identified across 29 normal human tissues.[21]

Similar gene expression does not prove the absence of sex differences since the same gene can give rise to two distinct phenotypes in males and females. For example, the male and female glioblastoma transcriptomes are similar, yet cell cycle and integrin-related genes are associated with survival in a sex-specific manner.[4] Similarly, modeling approaches have revealed that chronic obstructive pulmonary disease in males and females is driven by distinct metabolism and mitochondrial networks in the absence of differential expression.[23] Conversely, the same phenotype can be driven by distinct genetic pathways. In a study of over 100,000 humans, thirteen complex phenotypes showed genetic heterogeneity between males and females, and genomic prediction using sex-specific models outperformed a sex-agnostic model.[24]

Additional complexity arises from the effects of environmental exposures and hormonal interactions on molecular phenotypes.[3,17,25] In response to endogenous and exogenous factors, epigenetic modifications regulate the accessibility of DNA to transcriptional machinery.[26,27] This sex-influenced chromatin remodeling can cause differential gene expression in response to the same stimulus.[28,29] Sex hormones can directly modulate the function of transcription factors and other proteins, thereby giving rise to sex-specific regulatory networks.[23,30,31] In this way, identical phenotypes could be generated by two distinct networks in males and females, and diverse transcriptional responses could be generated by the same signal. Network modeling and systems-based approaches have an increased sensitivity to sex differences[21,23,31], so the consequences of neglecting sex in these analyses can be more profound than when considering genes individually.

The importance of incorporating demographic information into primary databases is clearly illustrated by considering immunology research. Women exhibit increased immune responsiveness to acute infection and vaccines compared to men, even when matched for pathogen load.[32] This heightened antigen-specific immune response contributes to the

female bias in autoimmune diseases[32] and may protect young women from cancer.[33] Sex differences in the immune response are not evident in infants and children, suggesting that immunity is modified over the lifespan as a function of age, gonadal and adrenal steroid hormones, and environmental exposures.[32] Consequently, analytical tools that are based upon pooled gene-expression data, without regard to the sex or age of the donor, are not necessarily sensitive nor specific when applied to smaller datasets like those queried by most investigators. Furthermore, they undermine our ability to understand complex biological processes and regulatory mechanisms in their totality.[32]

Conclusions, recommendations, and challenges

Sex differences are a cumulative effect of genetics, epigenetics, transcriptomics, proteomics, environment, social factors, hormonal influences, and network-level modulation. Our understanding of the underlying bases of biological systems requires us to acknowledge and disentangle these complex interactions. Several foundational questions will remain unanswered until omics resources with sex annotation are developed. While sex-unique pathways and networks likely exist across nearly all tissues and species, it is impossible to quantify the error associated with current, sex-agnostic methods. We suspect that databases rooted in gene and protein interactions may suffer disproportionately from this inattention compared to DNA-centric resources as sex differences seem to be most profound at the network level.[21] Despite the uncertainty regarding the degree to which current practices have impacted the quality of past results, it is clear that sex is a critical factor to be considered in omics analyses moving forward. As starting points, we recommend the following:

For scientists

- Perform omics analyses in combined-sex and separated male and female cohorts. Simply adding sex as a covariate to combined-sex investigations is insufficient, but these analyses remain valuable from the perspective of contextualizing sex-specific results in light of previous literature (e.g. results of previous studies were driven by an overrepresentation of one sex).

- Design studies to represent males and females equally and in sufficient numbers to detect sex differences, or provide a justification as to why this is not possible. Although the sex of cell lines is often not available, efforts should be made to conduct experiments on those derived from both sexes. When cell lines are passaged within animals, attention should be given to the evolution of those cells in the sex-matched vs. -mismatched settings.

- Follow the ARRIVE[11] and MIAME[34] guidelines when describing omics studies or depositing data into a public database. When comparing self-generated and public data, report the sex composition of both.

- When using a database that references primary studies, the work that gave rise to any statistically significant pathways/terms should be evaluated for sex, compared to the composition of the experimental cohort, and reported as a part of the results.

- If sex is missing from a tool or database, suggest that curators require subject sex reporting from contributors going forward to facilitate prospective annotation.

- If the terms in an omics database were generated by studies that are sex-incongruent with the experimental design, evaluate the literature for alternative signatures that are sex-specific and may not have been incorporated into the database yet.

For databases

- Provide references to the primary literature from which the information was originally derived.

- Note entries with the sex that the data originated from, and allow users to filter results by the sex that matches their experimental design.

- Actively caution users about the risks of applying female or mixed-sex data to historically male-biased standards.

- Prospectively curate new databases to bring attention to known sex differences and explicitly reference the data that support these conclusions.

For funding agencies

- Provide opportunities for individuals to determine the problem's scope, annotate resources, and use illustrative cases to quantify the impact of sex annotation (or lack thereof) on results.

- Support the generation of data and tools to directly characterize sex differences, or novel statistical or computational approaches to retrospectively address sex differences in data that are not currently amenable to such comparisons.

Challenges—We recognize the hurdles to implementing these recommendations, including:

- Financial burden of running both male and female experiments with the statistical power to detect differences.

- Time to explore the primary publications that contributed to databases and tools to determine the sex composition of these sources.

- Effort to annotate existing and future databases with sample donor sex, race, and age.

- Flexibility to continually expand the numbers of features accounted for in our primary datasets as we learn more about systems-level influences on molecular phenotypes.

Cognizance of sex bias in omics resources and the bioinformatics tools built upon these databases will enhance scientific rigor and improve the quality of work across all biological disciplines. By embracing these recommendations, attention will finally be given to a fundamental variable that has been long overlooked.

Acknowledgments

References

1. Mehra MR, Desai SS, Ruschitzka F & Patel AN RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet (2020) doi:10.1016/s0140-6736(20)31180-6.

2. Mehra MR, Desai SS, Kuy S, Henry TD & Patel AN Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. N. Engl. J. Med (2020) doi:10.1056/NEJMc2021225.

3. Federman DD The biology of human sex differences. N. Engl. J. Med. 354, 1507–1514 (2006). [PubMed: 16598047]

4. Yang W et al. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. Sci. Transl. Med. 11, (2019).

5. Zucker I & Beery AK Males still dominate animal studies. Nature 465, 690 (2010). [PubMed: 20535186]

6. Shah K, McCormack CE & Bradbury NA Do you know the sex of your cells? American Journal of Physiology-Cell Physiology vol. 306 C3–C18 (2014). [PubMed: 24196532]

7. Kosmidou I et al. Long-Term Outcomes in Women and Men Following Percutaneous Coronary Intervention. J. Am. Coll. Cardiol. 75, 1631–1640 (2020). [PubMed: 32273029]

8. Farkas RH, Unger EF & Temple R Zolpidem and driving impairment--identifying persons at risk. N. Engl. J. Med. 369, 689–691 (2013). [PubMed: 23923991]

9. Clayton JA & Collins FS Policy: NIH to balance sex in cell and animal studies. Nature 509, 282–283 (2014). [PubMed: 24834516]

10. Accounting for sex and gender makes for better science. Nature 588, 196 (2020). [PubMed: 33299202]

11. Kilkenny C, Browne WJ, Cuthill IC, Emerson M & Altman DG Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 8, e1000412 (2010). [PubMed: 20613859]

12. Edgar R, Domrachev M & Lash AE Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210 (2002). [PubMed: 11752295]

13. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013). [PubMed: 24071849]

14. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013). [PubMed: 23715323]

15. Bader GD, Cary MP & Sander C Pathguide: a pathway resource list. Nucleic Acids Res. 34, D504–6 (2006). [PubMed: 16381921]

16. Schmiedel BJ et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell 175, 1701–1715.e16 (2018). [PubMed: 30449622]

17. Traglia M et al. Genetic Mechanisms Leading to Sex Differences Across Common Diseases and Anthropometric Traits. Genetics 205, 979–992 (2017). [PubMed: 27974502]

18. Desachy G et al. Increased female autosomal burden of rare copy number variants in human populations and in autism families. Mol. Psychiatry 20, 170–175 (2015). [PubMed: 25582617]

19. Li CH, Haider S, Shiah Y-J, Thai K & Boutros PC Sex Differences in Cancer Driver Genes and Biomarkers. Cancer Res. 78, 5527–5537 (2018). [PubMed: 30275052]

20. Gershoni M & Pietrokovski S The landscape of sex-differential transcriptome and its consequent selection in human adults. BMC Biol. 15, 7 (2017). [PubMed: 28173793]

21. Lopes-Ramos CM et al. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. Cell Rep. 31, 107795 (2020). [PubMed: 32579922]

22. Issler O et al. Sex-Specific Role for the Long Non-coding RNA LINC00473 in Depression. Neuron vol. 106 912–926.e5 (2020). [PubMed: 32304628]

23. Glass K et al. Sexually-dimorphic targeting of functionally-related genes in COPD. BMC Syst. Biol. 8, 118 (2014). [PubMed: 25431000]

24. Rawlik K, Canela-Xandri O & Tenesa A Evidence for sex-specific genetic architectures across a spectrum of human complex traits. Genome Biol. 17, 166 (2016). [PubMed: 27473438]

25. Khramtsova EA, Davis LK & Stranger BE The role of sex in the genomics of human complex traits. Nat. Rev. Genet. 20, 173–190 (2019). [PubMed: 30581192]

26. Liu J, Morgan M, Hutchison K & Calhoun VD A Study of the Influence of Sex on Genome Wide Methylation. PLoS ONE vol. 5 e10028 (2010). [PubMed: 20386599]

27. van Dongen J et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat. Commun. 7, 11115 (2016). [PubMed: 27051996]

28. McCarthy MM & Nugent BM Epigenetic contributions to hormonally-mediated sexual differentiation of the brain. J. Neuroendocrinol. 25, 1133–1140 (2013). [PubMed: 23919286]

29. McCarthy MM et al. The epigenetics of sex differences in the brain. J. Neurosci. 29, 12815–12823 (2009). [PubMed: 19828794]

30. van Nas A et al. Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. Endocrinology 150, 1235–1249 (2009). [PubMed: 18974276]

31. Lopes-Ramos CM et al. Correction: Gene Regulatory Network Analysis Identifies Sex-Linked Differences in Colon Cancer Drug Metabolism. Cancer Res. 79, 2084 (2019). [PubMed: 30987980]

32. Klein SL & Flanagan KL Sex differences in immune responses. Nat. Rev. Immunol. 16, 626–638 (2016). [PubMed: 27546235]

33. Castro A et al. Strength of immune selection in tumors varies with sex and age. Nat. Commun. 11, 4128 (2020). [PubMed: 32807809]

34. Brazma A et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. 29, 365–371 (2001). [PubMed: 11726920]

35. Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000). [PubMed: 10802651]

36. Kanehisa M & Goto S KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000). [PubMed: 10592173]

37. Fabregat A et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–D655 (2018). [PubMed: 29145629]

38. Slenter DN et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 46, D661–D667 (2018). [PubMed: 29136241]

39. Thomas PD PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Research vol. 13 2129–2141 (2003). [PubMed: 12952881]
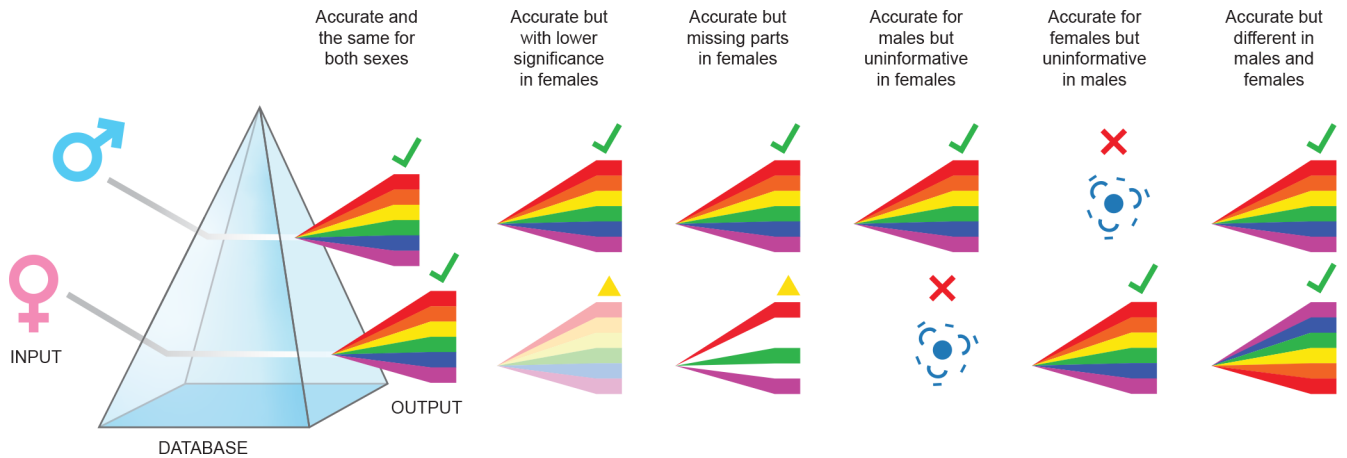
Figure 1.
Analyzing disaggregated male (M) and female (F) data through the perspective of databases that were built upon sex-biased studies (prism) could give rise to misleading results. (A) If the database is male-biased but there are truly no sex differences in the system, the output will be accurate for both M and F. (B) If sex differences exist and the database is M-biased, results could be accurate for Ms but have lower significance in Fs, (C) incomplete in Fs, or (D) uninformative in Fs. (E) In the case that sex differences exist and the database is F-biased, results may be uninformative for M data. (F) If the database annotates for sex, thereby allowing for truly sex-specific analyses, M and F outputs can be both different and accurate.

Table 1.

Five of the most highly-cited, public omics resources do not annotate terms by sex.

Omics Resource	Number of Terms	Primary Sources	Sex Annotation	Popular Dependent Tools
Gene Ontology[35]	44,945	Yes	No	DAVID, Panther, WebGestalt, ClueGO, g:Profiler
KEGG[36]	23,433	Yes	No	DAVID, WebGestalt, ClueGO, g:Profiler
Reactome[37]	21,077	Yes	No	Reactome, Panther, WebGestalt, ClueGO, g:Profiler
WikiPathways[38]	2874	Yes	No	WebGestalt
PANTHER[39]	177	Yes	No	Panther, WebGestalt