# Molecular omics resources should require sex annotation: a call for action

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 toward fully accounting for the biology of sex differences.

Kamila M. Bond, Margaret M. McCarthy, Joshua B. Rubin and Kristin R. Swanson

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 retracted because analyses were based on 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 broadly influence 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 over-represented in animal and human research owing 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 (Fig. 1).

Conversely, there are instances when female subjects are over-represented (for example, breast cancer and 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 slow scientific progress.

Some organizations have raised awareness of the importance of considering sex in research. The US 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 toward ensuring sex equity in research, they are not

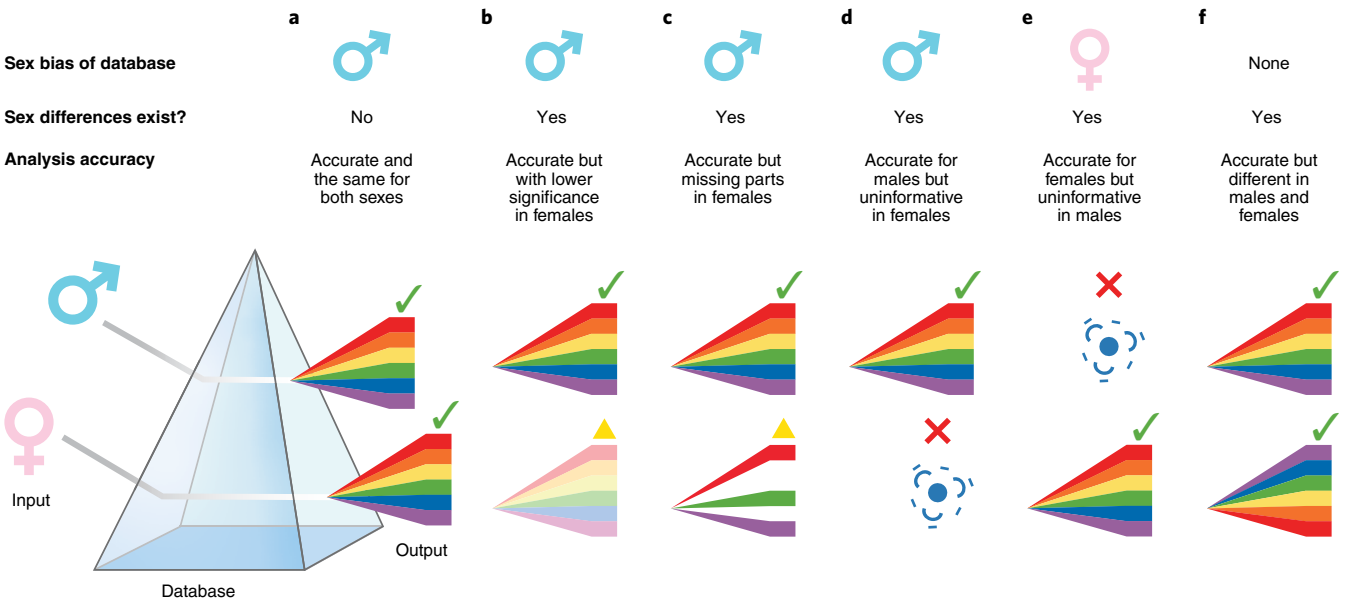


**Fig. 1 | Analyzing disaggregated male and female data through the perspective of databases that were built upon sex-biased studies (prism) could give rise to misleading results. a**, If the database is male-biased but there are truly no sex differences in the system, the output will be accurate for both male and female. **b**–**d**, If sex differences exist and the database is male-biased, results could be accurate for males but have lower significance in females (**b**), be incomplete in females (**c**), or be uninformative in females (**d**). **e**, If sex differences exist and the database is female-biased, results may be uninformative for male data. **f**, If the database annotates for sex, thereby allowing for truly sex-specific analyses, male and female 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] | 2,874 | Yes | No | WebGestalt |
| PANTHER[39] | 177 | Yes | No | PANTHER, WebGestalt |

universally adopted and do not rectify the decades of biased work on which current omics resources are built.

## The 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 over-representation of males[9]. This in turn gives rise to bias in primary data repositories (for example, GEO (Gene Expression Omnibus)[12]) unless the resource requires sex annotation upon submission (for example, TCGA (The Cancer Genome Atlas)[13], GTEx (Genotype–Tissue Expression)[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. Among 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 (for example, DICE[16] (the Database of Immune Cell Expression, Expression Quantitative Trait Loci (eQTLs) and Epigenomics)) 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 as a result of 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, 13 complex phenotypes showed genetic heterogeneity between males and females, and genomic prediction using sex-specific models outperformed a sex-agnostic model[24].

Further 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 elevated 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 illustrated by considering immunology research. Women exhibit greater immune responsiveness to acute infection and vaccines than 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]. Thus, 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 affected 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 (for example, if results of previous studies were driven by an over-representation 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. sex-mismatched settings.
- Follow the ARRIVE[11] and MIAME[34] guidelines when describing omics studies or depositing data in a public database. When comparing self-generated and public data, report the sex composition of both.
- When using a database that references primary studies, evaluate the work that gave rise to any statistically significant pathways or terms for sex, compare it to the composition of the experimental cohort, and report it 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 now 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 on these databases will enhance scientific rigor and improve the quality of work across all biological disciplines. Embracing these recommendations will finally bring attention to a fundamental variable that has been long overlooked. ❐

Kamila M. Bond[1,2], Margaret M. McCarthy[3,4], Joshua B. Rubin[5,6] and Kristin R. Swanson[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.*
✉*e-mail: swanson.kristin@mayo.edu*

Published online: 7 June 2021
https://doi.org/10.1038/s41592-021-01168-6

References

1. Mehra, M.R., Desai, S.S., Ruschitzka, F. & Patel, A.N. *Lancet* https://doi.org/10.1016/s0140-6736(20)31180-6 (2020).
2. Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D. & Patel, A. N. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMc2021225 (2020).
3. Federman, D. D. *N. Engl. J. Med.* **354**, 1507–1514 (2006).
4. Yang, W. et al. *Sci. Transl. Med.* https://doi.org/10.1126/scitranslmed.aao5253 (2019).
5. Zucker, I. & Beery, A. K. *Nature* **465**, 690 (2010).
6. Shah, K., McCormack, C. E. & Bradbury, N. A. *Am. J. Physiol. Cell Physiol.* **306**, C3–C18 (2014).
7. Kosmidou, I. et al. *J. Am. Coll. Cardiol.* **75**, 1631–1640 (2020).
8. Farkas, R. H., Unger, E. F. & Temple, R. *N. Engl. J. Med.* **369**, 689–691 (2013).
9. Clayton, J. A. & Collins, F. S. *Nature* **509**, 282–283 (2014).
10. *Nature* **588**, 196 (2020).
11. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. *PLoS Biol.* **8**, e1000412 (2010).
12. Edgar, R., Domrachev, M. & Lash, A. E. *Nucleic Acids Res.* **30**, 207–210 (2002).
13. Cancer Genome Atlas Research Network. et al. *Nat. Genet.* **45**, 1113–1120 (2013).
14. GTEx Consortium. *Nat. Genet.* **45**, 580–585 (2013).
15. Bader, G. D., Cary, M. P. & Sander, C. *Nucleic Acids Res.* **34**, D504–D506 (2006).
16. Schmiedel, B. J. et al. *Cell* **175**, 1701–1715.e16 (2018).
17. Traglia, M. et al. *Genetics* **205**, 979–992 (2017).
18. Desachy, G. et al. *Mol. Psychiatry* **20**, 170–175 (2015).
19. Li, C. H., Haider, S., Shiah, Y.-J., Thai, K. & Boutros, P. C. *Cancer Res.* **78**, 5527–5537 (2018).
20. Gershoni, M. & Pietrokovski, S. *BMC Biol.* **15**, 7 (2017).
21. Lopes-Ramos, C. M. et al. *Cell Rep.* **31**, 107795 (2020).
22. Issler, O. et al. *Neuron* **106**, 912–926.e5 (2020).
23. Glass, K. et al. *BMC Syst. Biol.* **8**, 118 (2014).
24. Rawlik, K., Canela-Xandri, O. & Tenesa, A. *Genome Biol.* **17**, 166 (2016).
25. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. *Nat. Rev. Genet.* **20**, 173–190 (2019).
26. Liu, J., Morgan, M., Hutchison, K. & Calhoun, V. D. *PLoS One* **5**, e10028 (2010).
27. van Dongen, J. et al. *Nat. Commun.* **7**, 11115 (2016).
28. McCarthy, M. M. & Nugent, B. M. *J. Neuroendocrinol.* **25**, 1133–1140 (2013).

29. McCarthy, M. M. et al. *J. Neurosci.* **29**, 12815–12823 (2009).
30. van Nas, A. et al. *Endocrinology* **150**, 1235–1249 (2009).
31. Lopes-Ramos, C. M. et al. *Cancer Res.* **79**, 2084 (2019).
32. Klein, S. L. & Flanagan, K. L. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
33. Castro, A. et al. *Nat. Commun.* **11**, 4128 (2020).
34. Brazma, A. et al. *Nat. Genet.* **29**, 365–371 (2001).
35. Ashburner, M. et al. *Nat. Genet.* **25**, 25–29 (2000).
36. Kanehisa, M. & Goto, S. *Nucleic Acids Res.* **28**, 27–30 (2000).
37. Fabregat, A. et al. *Nucleic Acids Res.* **46**, D649–D655 (2018). D1.
38. Slenter, D. N. et al. *Nucleic Acids Res.* **46**, D661–D667 (2018). D1.
39. Thomas, P. D. et al. *Genome Res.* **13**, 2129–2141 (2003).

Acknowledgements
K.R.S. gratefully acknowledges grant funding from the NIH (R01NS060752, R01CA164371, U54CA143970, U54CA193489, U01CA220378, U54CA210180, U01CA250481), the James S. McDonnell Foundation, the Ben & Catherine Ivy Foundation, the Zicarelli Foundation, the Arizona Biomedical Research Commission and the Mayo Clinic. M.M.M. gratefully acknowledges grant support from NIH R01DA039062 and R01MH091424. J.B.R. gratefully acknowledges grant funding from NCI (R01CA174737, P01CA245705).

Author contributions
K.M.B. drafted the manuscript and M.M.M., J.B.R. and K.R.S. oversaw the completion of the work.

Competing interests
K.R.S. is a co-founder of Precision Oncology Insights, Inc.

Additional information
**Peer review information** *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work.



# Diversity in immunogenomics: the value and the challenge

Immunogenomics studies have been largely limited to individuals of European ancestry, restricting the ability to identify variation in human adaptive immune responses across populations. Inclusion of a greater diversity of individuals in immunogenomics studies will substantially enhance our understanding of human immunology.

Kerui Peng, Yana Safonova, Mikhail Shugay, Alice B. Popejoy, Oscar L. Rodriguez, Felix Breden, Petter Brodin, Amanda M. Burkhardt, Carlos Bustamante, Van-Mai Cao-Lormeau, Martin M. Corcoran, Darragh Duffy, Macarena Fuentes-Guajardo, Ricardo Fujita, Victor Greiff, Vanessa D. Jönsson, Xiao Liu, Lluis Quintana-Murci, Maura Rossetti, Jianming Xie, Gur Yaari, Wei Zhang, Malak S. Abedalthagafi, Khalid O. Adekoya, Rahaman A. Ahmed, Wei-Chiao Chang, Clive Gray, Yusuke Nakamura, William D. Lees, Purvesh Khatri, Houda Alachkar, Cathrine Scheepers, Corey T. Watson, Gunilla B. Karlsson Hedestam and Serghei Mangul

## Current state of diversity in genomics studies

Genomic studies have mainly used samples from individuals of European ancestry, at the expense of learning from the largest and most genetically diverse populations. For example, 78% of individuals included in genome-wide association studies (GWAS) reported in the GWAS Catalog (https://www.ebi.ac.uk/gwas/home) through January 2019 are of European descent[1], while Asian populations account for 59.5% of the world population based on the Population Reference Bureau's World Population Data Sheet (https://www.prb.org/datasheets/). Though this is partially due to inadequate sampling of non-European populations, researchers tend to exclude data from minority groups when conducting statistical analyses[2] even when diverse datasets are available. The limited inclusion of samples from diverse populations hinders the equitable advancement of genomic medicine as a result of persistent uncertainty with respect to the genetic etiology of disease across populations, as well as differential rates of adverse drug events, treatment outcomes and other health disparities.

In recent years there has been an increased awareness of the limited generalizability of findings across populations and the benefits for the discovery and interpretation of gene–trait associations brought about by the inclusion of diverse populations in genomic studies. This has motivated the inclusion of diverse, multiethnic populations in large-scale genomic studies. For example, whole-genome sequencing in individuals of African descent[3] and whole-exome sequencing in a southern African population[4] have improved understanding of genetic variation in under-represented populations. Additional efforts have been made to establish reference genome datasets for research in diverse populations; these include the GenomeAsia 100K Project, Human Heredity and Health in Africa (H3Africa) initiative, Taiwan Biobank, Population Architecture Using Genomics and Epidemiology (PAGE) Consortium, Trans-Omics for Precision Medicine (TOPMed) program, Clinical Sequencing Evidence-Generating Research (CSER) consortium, Human Genome Reference Program (HGRP) and All of Us Research Program. However, the field of immunogenomics, especially that related to adaptive immune receptors, has yet to benefit from a similar growth in diversity.

## The need for diversity in immunogenomics

Central to immunity are the repertoires of T cell receptors (TCRs), immunoglobulins, human leukocyte antigens (HLAs) and killer cell immunoglobulin-like receptors (KIRs). Thus, analyses of the loci that encode these molecules are critical to immunogenomics studies.

T cells and B cells recognize antigens through their TCRs and immunoglobulins, which are formed through the process of V(D)J (variable, diversity and joining region) recombination. Capturing the vast diversity of recombined, expressed TCR and immunoglobulin repertoires was not possible until the development of