bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gender disparity in computational biology research publications

Kevin S. Bonham and Melanie I. Stefan

Abstract

While women are generally underrepresented in STEM fields, there are noticeable differences between fields. For instance, the gender ratio in biology is more balanced than in computer science. We were interested in how this difference is reflected in the interdisciplinary field of computational/quantitative biology. To this end, we examined the proportion of female authors in publications from the PubMed and arXiv databases. There are fewer female authors on research papers in computational biology, as compared to biology in general. This is true across authorship positions, year, and journal impact factor. A comparison with arXiv shows that quantitative biology papers have a higher ratio of female authors than computer science papers, placing computational biology in between its two parent fields in terms of gender representation. Both in biology and in computational biology, a female last author increases the probability of other authors on the paper being female, pointing to a potential role of female PIs in influencing the gender balance.

Intro

There is ample literature on the underrepresentation of women in STEM fields and the biases contributing to it. Those biases, though often subtle, are pervasive in several ways: they are often held and perpetuated by both men and women, and they are apparent across all aspects of academic and scientific practice. Undergraduate students show bias in favor of men both when rating their peers[1] and their professors[2]. Professors, in turn, are more likely to respond to e-mail from prospective students who are male[3]. They also show gender bias when hiring staff and deciding on a starting salary[4].

When looking at research output in the form of publication and impact, the story is complex: Women tend to publish less than men[5], are underrepresented in the more prestigious first and last author positions, and publish fewer single-author papers[6]. Articles authored by women are cited less frequently[5], which might in part be due to men citing their own work more often than women do[7]. Inferring bias in these studies is difficult, since the cause of the disparity between male and female authorship cannot be readily determined. At the same time, when stories of scientific discoveries are told, gender biases are readily identified: Work by female scientists is more likely to be attributed to a male colleague[8], and biographies of successful female scientists perpetuate gender stereotypes[9]. Finally, the way in which evidence for gender bias is received is in itself biased: Male scientists are less likely to accept studies that point to the existence of gender bias than are their female colleagues[10].

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Although gender imbalance seems to be universal across all aspects of the scientific enterprise, there are also more nuanced effects. In particular, not all disciplines are equally affected. For instance, in the biosciences over half of PhD recipients are now women, while in computer science, it is less than 20%[11]. This raises an intriguing question, namely how do the effects of gender persist in interdisciplinary fields where the parent fields are discordant for female representation?

To this end, we are interested in the gender balance in computational biology and how it compares to other areas of biology, since computational biology is a relatively young field at the disciplinary intersection between biology and computer science. We examined authorship on publications from Pubmed published between 1997 and 2014 and compared computational biology to biology in general. We found that in computational biology, there are fewer female authors overall, and fewer female authors in first and last authorship positions than in all biological fields combined. This is true across all years, though the gender gap has been narrowing, both in computational biology and in biology overall. A comparison to computer science papers shows that computational biology stands between biology and computer science in terms of gender equality.

Results and Discussion

In order to determine if there is a difference in the gender of authors in computational biology compared to biology as a whole, we used data from Pubmed, a database of biology and biomedical publications administered by the US National Library of Medicine. Pubmed uses Medical Subject Heading (MeSH) terms to classify individual papers by subject. The MeSH term “Computational Biology” is a subset of “Biology” and was introduced in 1997, so we restricted our analysis to primary articles published after this date (see Supplementary Figure 1A-B, Materials and Methods).

To determine the gender of authors, we used the web service Gender-API.com, which curates a database of first names and associated genders from government records as well as social media profiles. Gender-API searches provide information on the likely gender as well as confidence in the estimate based on the number of times a name appears in the database. We used bootstrap analysis to estimate the probability (P_{female}) that an author in a particular dataset is female as well as a 95% confidence interval (see Materials and Methods)

We validated this method by comparing it to a set of 2155 known author:gender pairs from the biomedical literature provided by Filardo et. al.[12] Filardo and colleagues manually determined the genders of the first authors for over 3000 papers by searching for authors’ photographs on institutional web pages or social media profiles like LinkedIn. We used our computational inference of gender for a subset of this data, which suggested P_{female} = 0.373, 95% CI = [0.350, 0.396], compared to the known gender, which gave P_{female} = 0.360 (using the same bootstrap analysis, 95% CI = [0.342, 0.379]) (Supplementary Figure 1C). Because Gender-API does not have information for a subset of names (43% of unique names in our Biology dataset,

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

representing 26.6% of authors), these names are excluded from our analysis. In order to ensure that this was not skewing our analysis, we also determined the P_{female} in Filardo et al.'s known gender dataset excluding those authors, giving P_{female} = 0.381, 95% CI = [.354, 0.409]. Together, these results suggest that our method of automatically assigning gender using Gender-API gives comparable result to hand-curated gender assignment, and that excluding names without clear gender information does not lead us to underestimate the proportion of women in our dataset.

We began by analyzing the gender representation in primary publications from 1997 to 2014. Consistent with previous publications, women were substantially less likely to be in senior author positions than first author positions in publications labeled with the Biology (Bio) MeSH term (Last author, P_{female} = 0.245, 95% CI = [0.243, 0.247], First author, P_{female} = 0.376, 95% CI = [0.373, 0.378]). (Figure 1A, Table 1). We observed the same trend in papers labeled with the computational biology (comp) MeSH term, though the P_{female} at every author position was 4-6 percentage points lower. An analysis of publications by year suggests that the gender gaps in both biology and computational biology are narrowing, but by less than 1 percentage point per year (for bio, ΔP_{female} = 0.0035 / year, 95% CI = [0.0030, 0.0040]. For comp, ΔP_{female} = 0.0049 / year, 95%CI = [0.0041, 0.0057]). However, the discrepancy between biology and computational biology has been consistent over time (Figure 1B).

Table 1			**95% CI**	
Dataset	**Position**	**Mean**	**Lower**	**Upper**
bio	first	0.376	0.373	0.378
	second	0.379	0.376	0.381
	other	0.368	0.367	0.370
	penultimate	0.279	0.277	0.282
	last	0.245	0.243	0.247
comp	first	0.316	0.312	0.320
	second	0.322	0.317	0.327
	other	0.331	0.328	0.333
	penultimate	0.236	0.231	0.241
	last	0.207	0.203	0.211

One possible explanation for the difference in male and female authorship position might be a difference in role models or mentors. If true, we would expect studies with a female principal investigator to be more likely to attract female collaborators. Conventionally in biology, the last author on a publication is the principal investigator on the project. Therefore, we looked at two

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

subsets of our data: publications with a female last author ($P_{female} > 0.8$) and those with a male last author ($P_{female} < 0.2$). We found that women were substantially more likely to be authors at every other position if the paper had a female last author than if the last author was male (Figure 1C, Table 2). It is possible that female trainees are be more likely to pursue computational biology if they have a mentor that is also female. Since women are less likely to be senior authors, this might reduce the proportion of women overall.

Table 2		Male Last Author			Female Last Author		
			95% CI			**95% CI**	
Dataset	**Position**	**Mean**	**Lower**	**Upper**	**Mean**	**Lower**	**Upper**
bio	first	0.362	0.359	0.365	0.478	0.472	0.484
	second	0.359	0.357	0.362	0.46	0.454	0.466
	other	0.355	0.353	0.357	0.425	0.421	0.428
	penultimate	0.259	0.256	0.263	0.336	0.330	0.343
comp	first	0.305	0.300	0.311	0.390	0.378	0.402
	second	0.306	0.300	0.312	0.379	0.366	0.392
	other	0.321	0.318	0.324	0.368	0.361	0.376
	penultimate	0.223	0.218	0.229	0.263	0.249	0.277

Though MeSH terms enable sorting a large number of papers regardless of where they are published, the assignment of these terms is a manual process and may not be comprehensive for all publications. As another way to qualitatively examine gender differences in publishing, we examined different journals, since some journals specialize in computational papers, while others are more general. We looked at the 123 journals that had at least 1000 authors in our bio dataset, and determined P_{female} for each journal separately (Figure 2A). Of these journals, 21 (14%) have titles indicative of computational biology or bioinformatics, and these journals have substantially lower representation of female authors. The 3 journals with the lowest female representation and 6 out of the bottom 10 are all journals focused on studies using computational methods. Only 4 computational biology/bioinformatics journals are above the median of female representation.

One possible explanation might be that women are less likely to publish in high-impact journals, so we considered the possibility that the differences in the gender of authors that we observe could be the result of differences in impact factor between papers published in biology versus computational biology publications. We therefore compared the P_{female} of authors in each journal with that journal's 2014 impact factor (Figure 2B). There is a marginal but significant ($p = 0.0025$) negative correlation between impact factor and gender for the biology dataset ($r^2 = 0.06$, 95% CI = [0.15, 0.0078]) , consistent with previous studies. By contrast, there is no significant correlation between impact factor and P_{female} in computational biology publications. Further, for

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

journals that have articles labeled with the computational biology MeSH term, the P_{female} for those articles is the same or lower than that for all biology publications in the same journal.

Taken together, these data suggest that the authors of computational biology papers are less likely to be women than the authors of biology papers generally. However, since Pubmed does not index computer science publications, we cannot compare computational biology to computer science. Instead, we turned to data from arXiv, a preprint repository for academic papers used frequently by quantitative fields like mathematics and physics. These preprint records cannot be directly compared to peer-reviewed publications indexed on pubmed, but a "quantitative biology" section was added to arXiv in 2003. There are relatively few papers preprints prior to 2007, so we compared preprints in "quantitative biology" to those in "computer science" from 2007-2016.

Women were more likely to be authors in quantitative biology than in computer science in first, second, and middle author positions (Figure 3A, Table 3), though the conventions for determining author order are not necessarily the same in computer science. Nevertheless, women had higher representation in quantitative biology than in computer science for all years except 2009 (Figure 3B). Interestingly, there is a slight but significant ($p < 0.05$) increase in the proportion of female authors over time in quantitative biology, while there's no significant increase in computer science preprints.

Table 3			95% CI	
Dataset	**Position**	**Mean**	**Lower**	**Upper**
arxivbio	first	0.184	0.178	0.190
	second	0.210	0.200	0.219
	other	0.265	0.253	0.276
	penultimate	0.196	0.183	0.209
	last	0.148	0.141	0.155
arxivcs	first	0.157	0.155	0.160
	second	0.175	0.172	0.179
	other	0.188	0.182	0.195
	penultimate	0.175	0.170	0.181
	last	0.155	0.153	0.158

Together, our results suggest that computational biology lies between biology in general and computer science when it comes to gender representation in publications. This is perhaps not surprising given the interdisciplinary nature of computational biology. Compared to biology in general, computational biology papers have fewer female authors, and this is consistent across all authorship positions. Importantly, this difference is not due to a difference in impact factor between computational biology and general biology papers.

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Articles with a female last author tend to have more female authors in other positions and this is true for both biology in general and computational biology. Since the last author position is most often occupied by the principal investigator of the study, this suggests that having a woman as principal investigator has a positive influence on the participation of women. This could be because female PIs are more likely to recognise contributions by female staff members, or because they are more likely to attract female co-workers and collaborators. The publication data cannot differentiate between those two (and other) explanations, but points to the important role that women in senior positions may play as role models for trainees.

Since biology attracts more women than computer science, we suspect that many women initially decide to study biology and later become interested in computational biology. If this is the case, understanding what factors influence the field of study will provide useful insight when designing interventions to help narrow the gender gap in computer science and computational biology.

Materials and Methods

Datasets

Biology pubs 1997-2014 (`bio`)

This dataset[13] contains all English language publications under the MeSH term "Biology" published between 1997 and 2014, excluding many non-primary sources. Downloaded 12 February, 2016. Search term: `("Biology"[Mesh]) NOT (Review[ptyp] OR Comment[ptyp] OR Editorial[ptyp] OR Letter[ptyp] OR Case Reports[ptyp] OR News[ptyp] OR "Biography" [Publication Type]) AND ("1997/01/01"[PDAT] : "2014/12/31"[PDAT]) AND english[language]`

Computational Biology pubs 1997-2014 (`comp`)

Same as above[13], except using MeSH term "Computational Biology". Only uses papers where this is is a major term. Date range was selected because this MeSH term was introduced in 1997. Downloaded 12 February, 2016. Search term: `("Computational Biology"[Majr]) NOT (Review[ptyp] OR Comment[ptyp] OR Editorial[ptyp] OR Letter[ptyp] OR Case Reports[ptyp] OR News[ptyp] OR "Biography" [Publication Type]) AND ("1997/01/01"[PDAT] : "2014/12/31"[PDAT]) AND english[language]`

Med Papers

Subset of author and gender data from Filardo et.al[12]. This dataset did not contain author first names or unique publication identifiers. We searched pubmed for the title, author and

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

publication date, and were able to identify 2155/3153 publications to analyze. Publications with no matching search results or with multiple matching search results were excluded.

arXiv Quantitative Biology (`q-bio`)

This dataset[14] contains all preprints with the label "q-bio" from 2003 (when the section was introduced) to 2014. Downloaded on 10 June, 2016.

arXiv CS (`cs`)

This dataset[14] contains all preprints with the label "cs" from 2003 to 2014. Downloaded on 10 June, 2016.

Gender Inference

Genders were determined using Gender-API (http://gender-api.com), which compares first names to a database compiled from government sources as well as from crawling social media profiles and returns a gender probability and a measure of confidence based on the number of times the name appears in the database. The API was queried with the 74,760 unique first names in the dataset (24 May, 2016).

Mean gender probabilities were determined using bootstrap analysis, excluding names for which no gender information was available (~26.6% of authors). Error bars represent 95% confidence intervals. Code and further explanation can be found on github[15].

Author positions were assigned based on the number of total authors. In papers with 5 or more authors, all authors besides first, second, last and penultimate were designated "other." Papers with 3 authors were assigned only first, second and last, and papers with two authors were assigned only first and last.

Figures

Figure 1

(A) Mean probability that an author in a given position is female for primary articles indexed in Pubmed with the MeSH term Biology (black) or Computational Biology (grey). The bio dataset is inclusive of papers in the comp dataset. Error bars represent 95% confidence intervals.
(B) Mean probability that an author is female for publications in a given year. Error bars represent 95% confidence intervals.
(C) Mean probability that the first (F), second (S), penultimate (P) or other (O) author is female for publications where the last author is male ($P_{female} < 0.2$) or female ($P_{female} > 0.8$). Papers where the gender of the last author was uncertain or could not be determined were excluded. Error bars represent 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2

(A) Mean probability that an author is female for every journal that had at least 1000 authors in our dataset. Grey bars represent journals that have the words "Bioinformatics," "Computational," "Computer," "System(s)," or "-omic(s)" in their title. Vertical line represents the median for female author representation. See also Supplementary Table 1.

(B) Mean probability that an author is female for articles in the "Bio" dataset (black dot) or in the "Comp" dataset (open square) for each journal that had at least 1000 authors plotted against the journals' 2014 impact factor. Journals that had computational biology articles are included in both datasets. Pearson's correlation was calculated for each dataset independently. Bio: $r = -0.246$, 95% CI = [-0.391, -0.088], $p = 0.0025$. Comp: $r = -0.052$, 95% CI = [-0.4458, 0.3592], $p = 0.8106$.

Figure 3

(A) Mean probability that an author in a given position is female for all preprints in the arXiv quantitative biology (black) or computer science (grey) categories between 2007 and 2014. Error bars represent 95% confidence intervals.

(B) Mean probability of authors being female in arXiv preprints in a given year. Error bars represent 95% confidence intervals. Slopes were determined using linear regression. The slope for q-bio is slightly positive ($p < 0.5$), but the slope for cs is not.

Supplementary Figure 1

(A) Number of primary publications per year indexed under the "Biology" MeSH term.

(B) Number of primary publications per year indexed with "Computational Biology" as a major MeSH term.

(C) Comparison of computational gender inference (black) with known genders (white) for the dataset from Filardo et. al.[12]. Grey represents the known proportion of female authors when excluding names for which the gender could not be computationally inferred. Error bars represent 95% confidence intervals.

Supplementary Table 1

https://docs.google.com/spreadsheets/d/1HlRZLd3J-naNdxGNdoUo6Xl11U2tGJZh9GYTP9SxZ0M/edit#gid=1769952756

Acknowledgements

Markus Perl for the free use of Gender-API - contact@gender-api.com
Casper Strømgren for the free use of genderize.io info@genderize.io
Giovanni Filardo for sharing data from ref.12

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Johanna Gutlerner, Marshall Thomas, Diane Lam, and other members of the Curriculum Fellows Program (CFP) at Harvard Medical School (HMS) for helpful feedback and discussions. The HMS CFP and Educational Laboratory for resources and mentorship.

1. Grunspan, D. Z. *et al.* Males Under-Estimate Academic Performance of Their Female Peers in Undergraduate Biology Classrooms. *PLoS One* **11,** e0148405 (2016).
2. MacNell, L., Driscoll, A. & Hunt, A. N. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education* **40,** 291–303 (2014).
3. Milkman, K. L., Akinola, M. & Chugh, D. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psychol.* **100,** 1678–1712 (2015).
4. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* **109,** 16474–16479 (2012).
5. Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: global gender disparities in science. *Nature* **504,** 211–213 (2013).
6. West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The role of gender in scholarly authorship. *PLoS One* **8,** e66212 (2013).
7. King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J. & West, J. D. Men set their own cites high: Gender and self-citation across fields and over time. (2016).
8. Rossiter, M. W. The Matthew Matilda Effect in Science. *Soc. Stud. Sci.* **23,** 325–341 (1993).9. Fara, P. Women in science: Weird sisters? *Nature* **495,** 43–44 (2013).
10. Handley, I. M., Brown, E. R., Moss-Racusin, C. A. & Smith, J. L. Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13201–13206 (2015).
11. National Science Foundation, National Center for Science and Engineering Statistics.

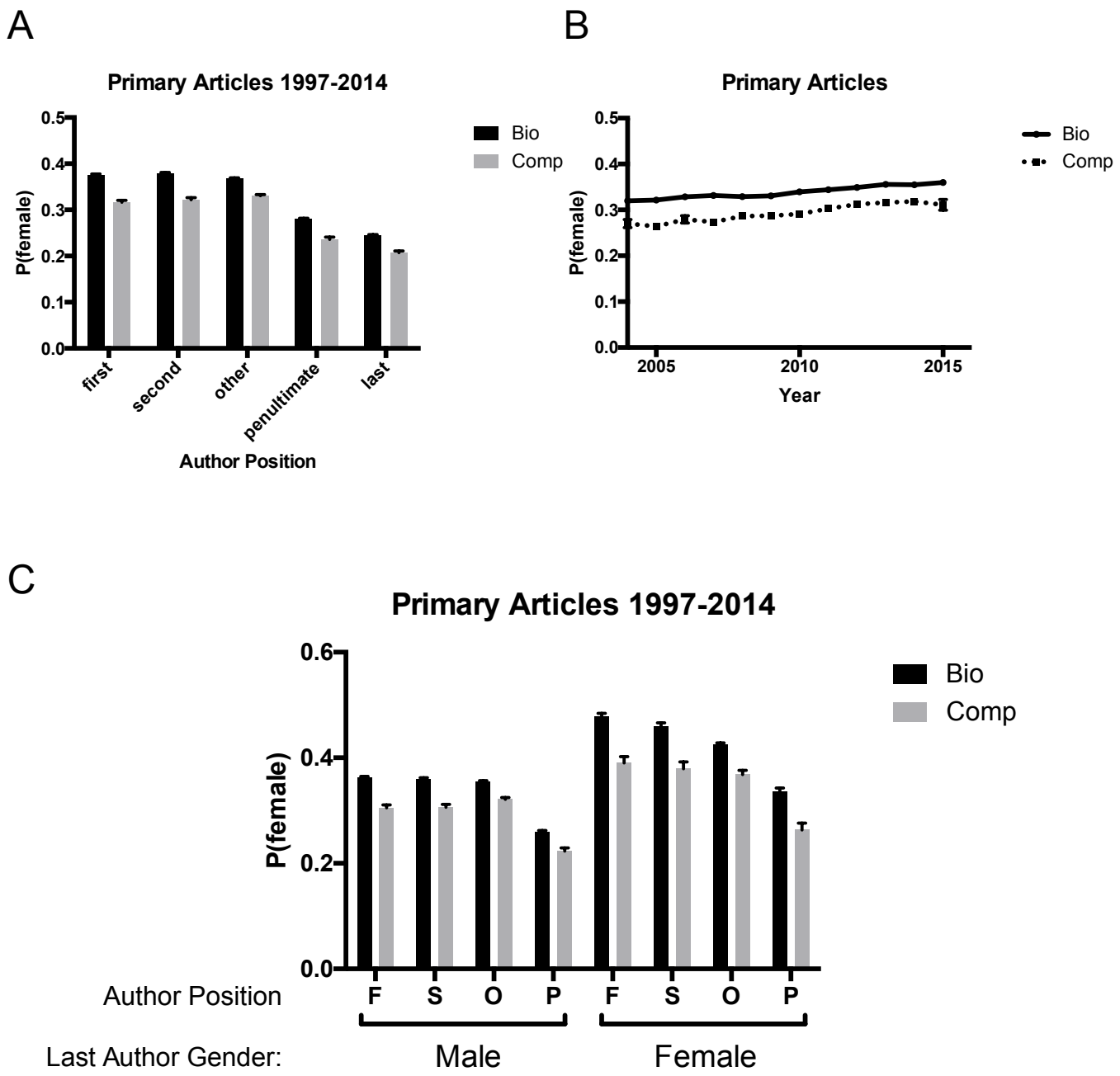
bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015. (2015).

12. Filardo, G. *et al.* Trends and comparison of female first authorship in high impact medical journals: observational study (1994-2014). *BMJ* **352,** i847 (2016).
13. Bonham, K. S. & Stefan, M. Biology and Computational Biology Papers in Pubmed, 1997-2014. (2016). doi:10.5281/zenodo.58990
14. Bonham, K. S. & Stefan, M. Preprints from arXiv.org in cs and q-bio. (2016). doi:10.5281/zenodo.60088
15. Bonham, K. & Stefan, M. *gender-comp-bio: Pre-publication release*. (Zenodo, 2016). doi:10.5281/zenodo.60090

bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

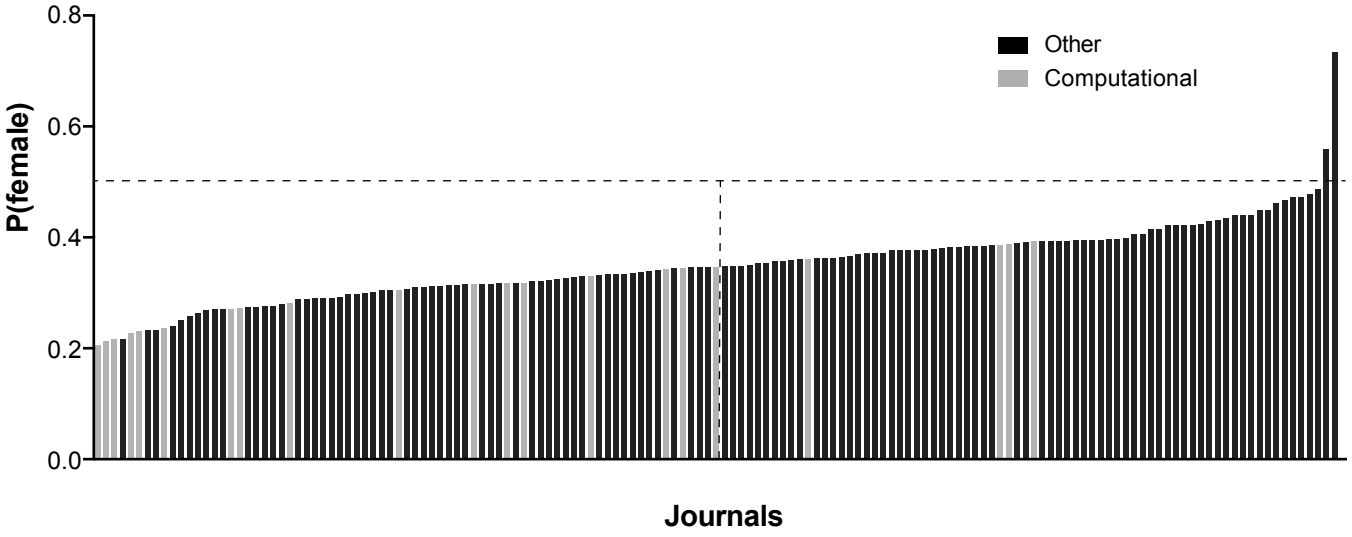
Figure 1



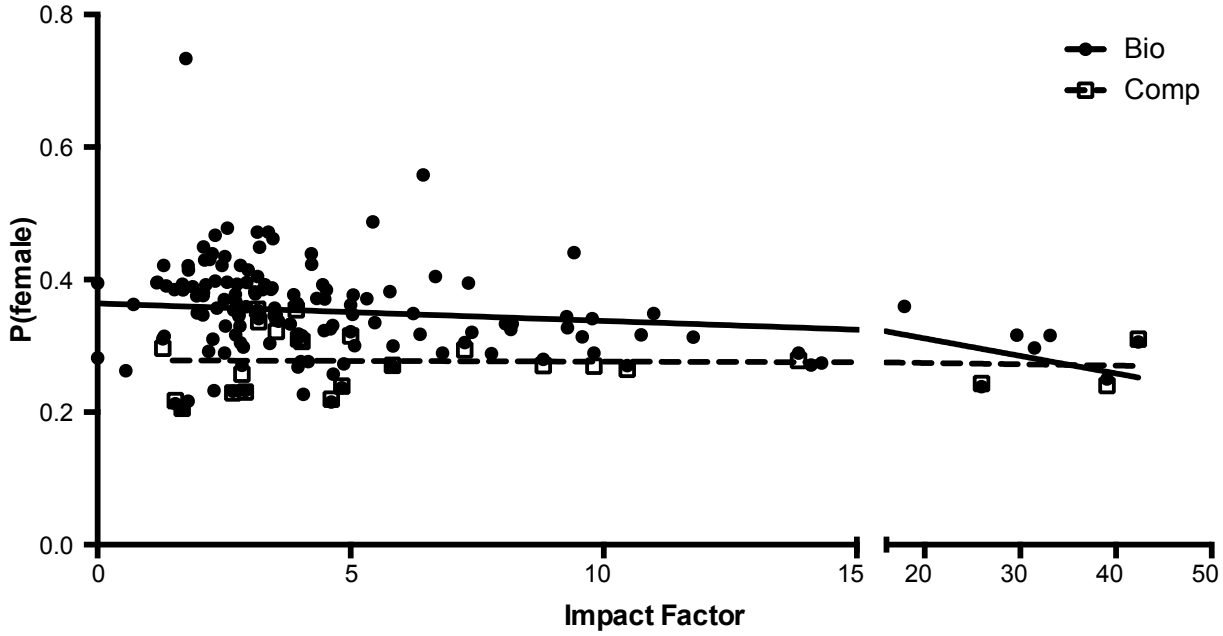
bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2

A



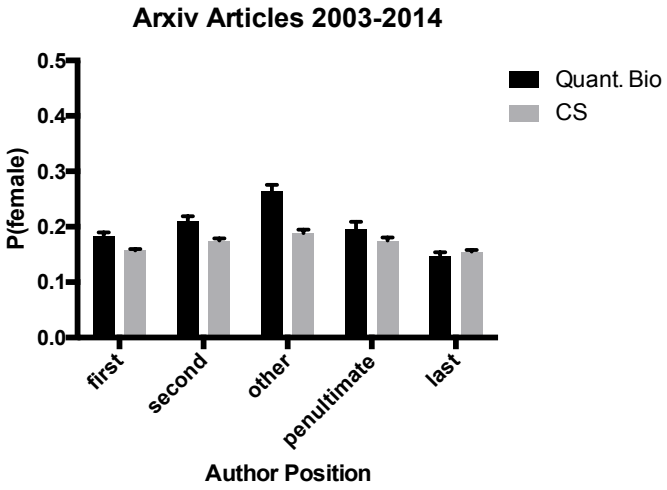
B



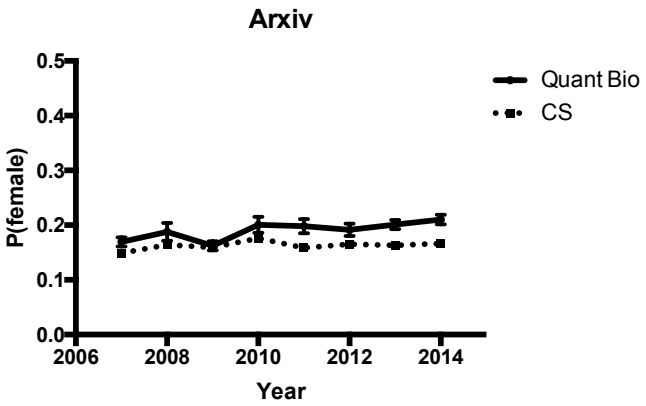
bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

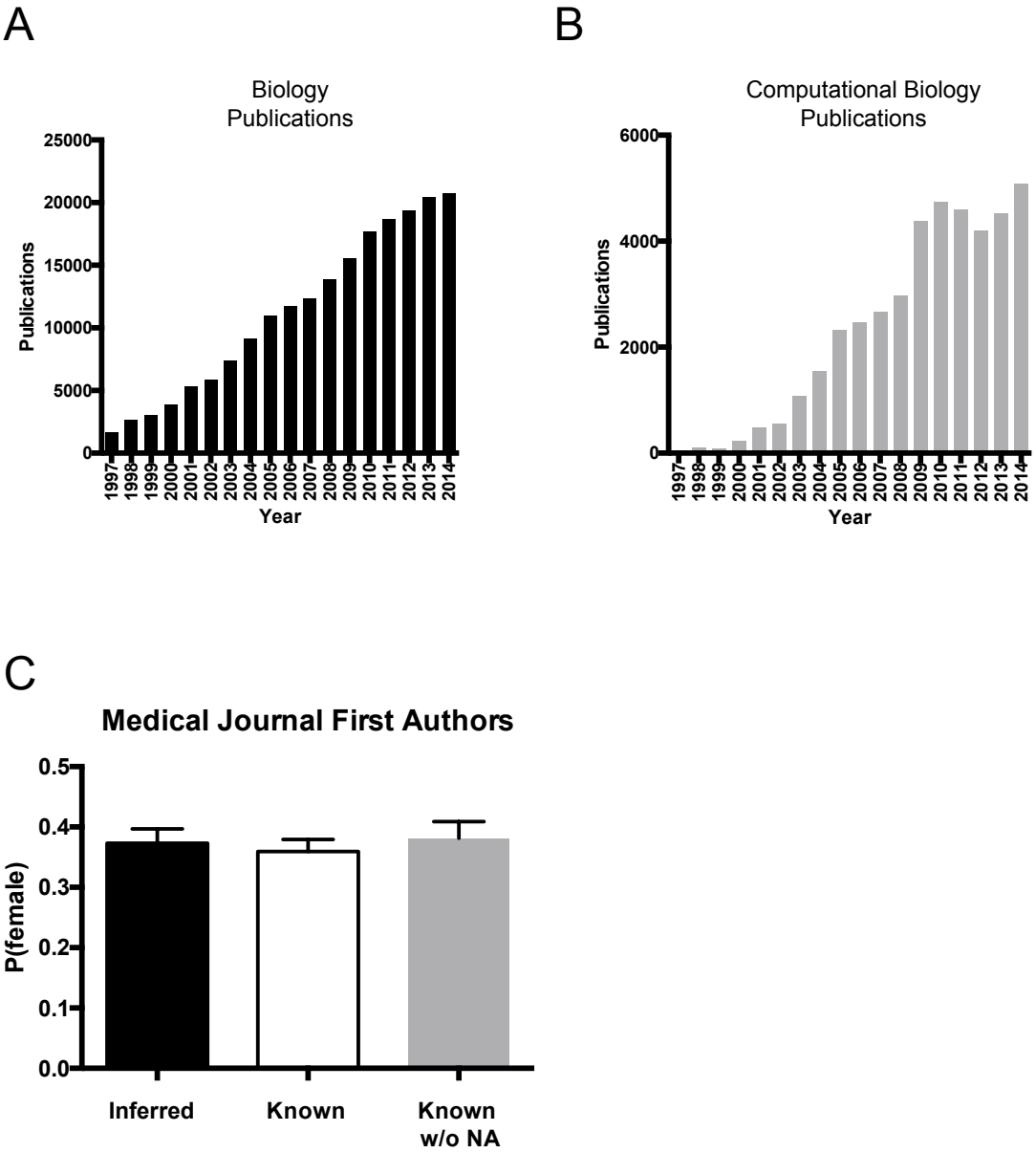
Figure 3

A



B



bioRxiv preprint doi: https://doi.org/10.1101/070631; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 1