

Manuscript Title

This manuscript ([permalink](#)) was automatically generated from [greenelab/iscb-diversity-manuscript@dd48cda](#) on January 23, 2020.

Authors

- **Trang T. Le**

 [0000-0003-3737-6565](#) ·  [trang1618](#) ·  [trang1618](#)

Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania

- **Daniel S. Himmelstein**

 [0000-0002-3012-7446](#) ·  [dhimmel](#) ·  [dhimmel](#)

Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania

- **Ariel A. Hippen Anderson**

 [0000-0001-9336-6543](#) ·  [arielah](#) ·  [AHippenAnderson](#)

Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania

- **Casey S. Greene**

 [0000-0001-8713-9213](#) ·  [cgreene](#) ·  [greenescientist](#)

Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

Abstract

Professional societies and the conferences that they manage provide an important venue for the dissemination of scientific knowledge. Being invited to deliver a keynote at an international society meeting or named a fellow of such a society is a major recognition. We sought to understand the extent to which such recognitions reflected the composition of their corresponding field. We collected keynote speaker invitations for the international meetings held by the International Society for Computational Biology as well as the names of Fellows. We compared these individuals with last and corresponding author contributions in the society's partner journals. We used multiple methods to estimate the race, ethnicity, gender, and nationality of authors and the recipients of these honors. Individuals from certain ancestries and countries appear to be under-recognized among honorees.

Introduction

Scientists' roles in society include identifying important topics of study, undertaking an investigation of those topics, and disseminating their findings broadly. The scientific enterprise is largely self-governing: scientists act as peer reviewers on papers and grants, comprise hiring committees in academia, make tenure decisions, and select which applicants will be admitted to doctoral programs. A lack of diversity in science could lead to pernicious biases that hamper the extent to which scientific findings are relevant to minority communities. For example, finding that minority scientists tend to apply for awards on topics with lower success rates [1] could be interpreted either as minority scientists select topics in more poorly funded areas or that majority scientists score topics of particular interest to minority scientists as less worthy of funding. Consequently, it is important to examine peer recognition in fields.

Gender bias among conference speakers has been recognized as an area that can be improved with targeted interventions [2,3,4,5]. Having more female organizers on conference committees is associated with having more female speakers [6]. At medical conferences in the US and Canada, the proportion of female speakers is increasing at a modest rate [7]. Gender bias appears to also influence funding decisions: an examination of scoring of proposals in Canada found that reviewers asked to assess the science produced a smaller gender gap in scoring than reviewers asked to assess the applicant [8]. Challenges extend beyond gender: an analysis of awards at the NIH found that proposals by Asian, black or African-American applicants were less likely to be funded than those by white applicants [9]. There are also potential interaction effects between gender and race or ethnicity that may particularly affect women of color's efforts to gain NIH funding [10].

We sought to understand the extent to which honors and high-profile speaking invitations were distributed equitably among gender, race/ethnicity, and nationality groups by an international society and its associated meetings. As computational biologists, we focused on the [International Society for Computational Biology](#) (ISCB), its honorary Fellows as well as its affiliated international meetings: [Intelligent Systems for Molecular Biology](#) (ISMB), [Research in Computational Molecular Biology](#) (RECOMB), and [Pacific Symposium on Biocomputing](#) (PSB).

We used multiple methods to predict the gender, race/ethnicity, and nationality of honorees. Existing methods were relatively US-centric because most of the data was derived in whole or in part from the US Census. We scraped more than 700,000 entries from English-language Wikipedia that contained nationality information to complement these existing methods and built multiple machine learning classifiers to predict nationality. We also examined the last and corresponding authors for publications in ISCB partner journals to establish a field-specific baseline using the same metrics. The results were consistent across all approaches: we found a dearth of minority speakers and honorees. The lack of Asian speakers was particularly pronounced when compared against the field-specific background.

Materials and Methods

Honoree Curation

ISCB Fellows Recipients

We examined the ISCB [webpage of ISCB Fellows](#). We found recipients listed for the years 2009-2019. We gleaned the full name of the fellow as well as the year in which they received the honor. We used the name as provided on the site. For certain methods we were required to split the full name into first and last names. In this case we chose the first non-initial name as the first name and the final name as the last name. We did not consider a hyphen to be a name separator: for hyphenated names, all components were included.

ISMB Keynote Speakers

We examined the webpage for each ISMB meeting. We were able to successfully find pages with keynote speakers for the years 2002-2019. We gleaned the full name of each keynote speaker as well as the year in which they delivered a keynote. We used the name as provided on the site. We split names into first and last names as described for ISCB Fellows.

PSB Keynote Speakers

We examined the webpage for each PSB meeting. We were able to successfully find pages with keynote speakers for the years 1999-2020. We gleaned the full name of each keynote speaker as well as the year in which they delivered a keynote. We used the name as provided on the site. We split names into first and last names as described for ISCB Fellows.

RECOMB Keynote Speakers

We examined the webpage for each RECOMB meeting. We found conference webpages with keynote speakers for 1999, 2000, 2001, 2004, 2007, 2008, and 2010-2019. We were able to fill in the missing years using information from the RECOMB 2016 proceedings, which summarizes the first 20 years of the RECOMB conference [11]. This volume has two tables of keynote speakers from 1997-2006 (Table 14, page XXVII) and 2007-2016 (Table 4, page 8). Using these tables to verify the conference speaker lists, we arrived at two special instances of inclusion/exclusion. Although Jun Wang was not included in these tables, we were able to confirm that he was a keynote speaker in 2011 with the RECOMB 2011 proceedings [12], and thus we include this speaker in the dataset. Marian Walhout was invited as a keynote speaker but had to [cancel](#) the talk due to other obligations. Because her name was neither mentioned in the 2015 proceedings [13] nor in the earlier tables, we exclude this speaker from our dataset. For other keynote speakers, we gleaned their full name as well as the year in which they delivered a keynote. We used the name as provided on the site. We split names into first and last names as described for ISCB Fellows.

Corresponding author extraction

We assumed that research advisors in the field would be those most likely to be invited for keynotes or to be honored as Fellows. Therefore, we collected corresponding author names to assess the composition of the field, weighted by the number of publications.

We evaluated two resources for extracting corresponding authors from papers: [PubMed](#) and [PubMed Central](#) (PMC). Both resources are provided by the U.S. National Library of Medicine and index scholarly articles. PubMed contains a record for every article published in journals it indexes (30 million records total circa 2020) and provides abstracts but not fulltext. PMC, which provides fulltext

access, does not contain every article from every journal (5.9 million records total circa 2020). In general, open access journals will deposit their entire catalog to PMC (e.g. *BMC Bioinformatics* & *PLOS Computational Biology*), while toll access journals (e.g. *Bioinformatics*) will only deposit articles when funders require it. Since PMC requires publishers to submit fulltext articles in a structured XML format, the machine-readability and breadth of metadata in PMC is often superior to PubMed.

Of PMC's 5.9 million fulltext articles, only 2.7 million are part of the "[Open Access Subset](#)" which allows for downloading the structured fulltext as opposed to just viewing the article online. However, authorship information does not require full text records. We were able to download structured frontmatter (rather than fulltext) records from PMC's [OAI-PMH service](#), so we were not limited to just the Open Access Subset. For PubMed, we used the E-Utilities APIs. For PubMed records, we were able to extract author names (first and last) and order. For PMC, we were able to extract these fields as well as whether each author was a corresponding author. To automate and generalize these tasks, we created the [pubmedpy](#) Python package.

We selected three journals to represent the field of bioinformatics and computational biology, including two ISCB Partner Journals (*PLOS Computational Biology* and *Bioinformatics*) and one field-specific journal that is not a partner (*BMC Bioinformatics*). From PubMed, we compiled a catalog of 29,755 journal articles published from when each journal was established through 2019. We were able to retrieve authorship information for all but 6 of these articles using PubMed or PubMed Central.

To determine corresponding authors for an article, we relied on PMC data if available (20,696 articles) and otherwise fallback to PubMed data (9,053 articles). Almost all articles without PMC data were from *Bioinformatics*, since its a "selective deposit" rather than "full participation" journal in PMC.

We performed further analysis on PMC authors to learn more about corresponding author practices. First, we developed and evaluated a method to infer a corresponding author when the coded corresponding status was not available. For papers with multiple authors and at least one corresponding author, the first author was corresponding 43% of the time, whereas the last author was corresponding 62% of the time. Therefore, we assumed the last author was corresponding when coded corresponding author status was not available (120 articles from PMC and all articles from PubMed).

Second, we investigated the number of corresponding authors for PMC articles. 81% of these articles had a single corresponding author. 1.7% had no corresponding authors. Of these, many were editorials (e.g. [PMC1183510](#), the announcement of *PLOS Computational Biology*). A very small number of papers had over 10 corresponding authors. Some of these instances were true outliers, like [PMC5001208](#) with 21 corresponding authors. Others like [PMC3509495](#) were incorrect, due to upstream errors. To not give undue influence to papers with multiple corresponding authors, subsequent analyses on corresponding authors are weighted by 1 divided by the number of corresponding authors per paper.

Estimation of Gender

We predicted the gender of honorees and authors using the <https://genderize.io> API that produces predictions trained on over 100 million name-gender pairings collected from the web. We used author and honoree first names to retrieve predictions from genderize.io. The predictions, which consider gender as a binary trait, represent the probability of an honoree or author being male or female.

Estimation of Race and Ethnicity

We predicted the race and ethnicity of honorees and authors using the R package `wru`. `wru` implements methods described in Imai and Khanna [14] to predict race and ethnicity using surname and location information. The underlying data used for prediction are derived from the United States Census. We used only the surname of author or honoree to make predictions via the `predict_race()` function. However, in the case of names that were not observed in the census, the function's behavior was to use the average demographic distribution from the census. We modified the function to return a status denoting that results were inconclusive instead. This prediction represents the probability of an honoree or author selecting a certain race or ethnicity on a census form if they lived within the United States.

Estimation of Nationality

To complement `wru`'s race and ethnicity estimation, we applied the Python package `ethnicolr` developed by Sood and Laohaprapanon [15]. `ethnicolr` also uses United States Census data, but also adds in data curated from Wikipedia in 2009 [16]. While `ethnicolr` overcomes some of the limitations of `wru`, its international representation is still limited. For instance, 76% of the names in `ethnicolr`'s Wikipedia dataset are European in origin, and the dataset contains remarkably fewer Asian, African, and Middle Eastern names compared to that of `wru`.

A number of the limitations of `ethnicolr` are related to the source data. Since Wikipedia has grown substantially since 2009, we created an updated dataset and set of predictors based on the `ethnicolr` methodology, which are described below.

Constructing a Name-to-Nationality Dataset

To generate a training dataset for nationality prediction, we scraped Wikipedia's category of [Living People](#) (approximately 930,000 pages when we did our data scrape in November 2019). This category is regularly curated and allowed us to avoid pages related to non-persons and to reflect a modern naming landscape. For each Wikipedia page, we used two strategies to find a full birth name and nationality for that person. First, we pulled information from the personal details sidebar; the information in this sidebar varied widely but usually contained a full name and a place of birth. Second, in the body of the text of most English-language biographical Wikipedia pages, the first sentence usually begins with, for example, "John Edward Smith (born 1 January 1970) is an American novelist known for ..." We used regular expressions to parse out the person's name from this structure, as well as checking that the expression after "is a" matched a list of possible nationalities. Using a union of these strategies, we were able to define a name and nationality for 708,493 people via their Wikipedia pages. Our Wikipedia-based process returned only a nationality or country of origin, which was more fine-grained than the broader regional patterns that we sought to examine among ISCB honorees and bioinformatics authors. We initially annotated based on continent, but later decided to model our categorization after the hierarchical nationality taxonomy used by NamePrism [17]. After several iterations, we settled on the following categories: Hispanic (including Latin America and Iberia), African, Israeli, Muslim, SouthAsian, EastAsian, European (non-British, non-Iberian), and CelticEnglish (including United States, Canada, and Australia). We make this dataset, which we term Wiki2019, available at https://github.com/greenelab/wiki-nationality-estimate/master/data/annotated_names.tsv.

Nationality Prediction with LSTM Neural Networks

We trained a Long Short-term Memory (LSTM) neural network, the same kind of classifier used by `ethnicolr`. This classifier learned to recognize patterns in the sequences of letters in a name to infer region of origin. We evaluated multiple lengths of letters and, based on this comparison, used tri-characters for the primary results described in this work. To evaluate our classifier, we used unseen examples from `ethnicolr`'s Wiki2009 dataset as ground truth to test our best model (LSTM, 3-grams).

Our prediction model, which we term Wiki2019-LSTM is available at <https://github.com/greenelab/wiki-nationality-estimate/master/models/LSTM.h5>.

Results

Curated Honorees and Literature-derived Potential Honorees

We curated a dataset of ISCB honorees that included 411 honorees who were keynote speakers at international ISCB-associated conferences (ISMB, RECOMB, and PSB) as well as ISCB Fellows. The ISCB Fellows set contained the complete set of fellows named (2009-2019). Keynote speakers were available for ISMB for all years from 2002-2019. Keynote speakers from PSB were available for all years from 1999-2020. Keynote speakers for RECOMB were available for all years from 1997-2019. We included individuals who were honored multiple times as separate entries. For example, Christine Orengo was a keynote speaker at RECOMB 2004 and became an ISCB Fellow in 2016, and thus was counted twice in this list.

We sought to compare this dataset with a background distribution of potential speakers, which we considered to be last or senior authors of bioinformatics and computational biology manuscripts. We used those published in [Bioinformatics](#), [BMC Bioinformatics](#), and [PLOS Computational Biology](#) as a set of bioinformatics and computational biology manuscripts. We downloaded the metadata of manuscripts published in these journals from PubMed, which provided more than 30000 articles for evaluation. However, although PubMed provides author order, it does not provide corresponding author information.

We downloaded article information from PubMed Central (PMC), which provides corresponding author information to directly measure the fraction of manuscripts for which last authors were also corresponding authors in these journals. Of the 21411 articles in PMC's, the last author was also a corresponding author for 17401 of them. Because the concordance was high and the PubMed set was more complete, we used the PubMed last author set for all subsequent analyses.

Assessing Gender Diversity of Authors and Honorees

Although Bioinformatics was established in 1998 and BMC Bioinformatics in 2000, the metadata for these journal papers before 2002 only have initials for first and/or middle author names. Therefore, without first and middle names, we do not have author gender predictions before this year.

There is a slow increase of the proportion of predicted female authors, arriving at just over 20% on average in 2019 (Fig. [1A](#)). We observe very similar trend within each journal, but estimated female proportion has increased the least in *PLOS Computational Biology* (Supplementary Fig. S1). ISCB fellows and keynote speakers appear to be more evenly split between men and women compared to the population of authors published in computational biology and bioinformatics journals (Fig. [1B](#)); however, it has not yet reached parity. The overall increasing female proportion trend seems to be present in each honor category, especially in the group of ISCB Fellows (Fig. [1C](#)), which seems to have markedly increased after 2015. PSB has only one or two keynotes per year, so its estimated proportion of female speakers varies considerably (Fig. [1C](#)). However, the 2020 PSB keynotes were the only set of honorees across our analysis for which nearly all the probability was ascribed to female speakers.

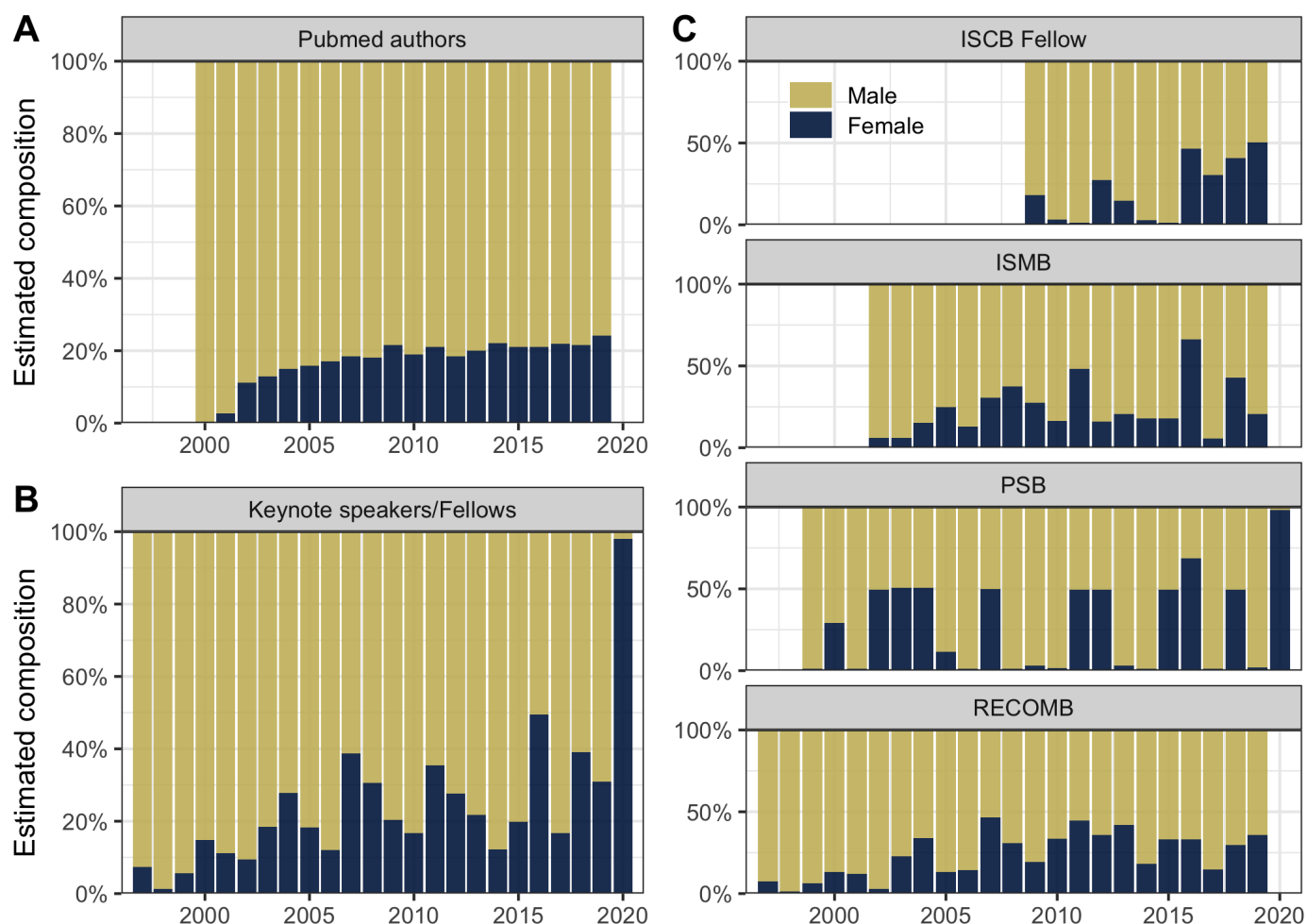


Figure 1: ISCB fellows and keynote speakers appear more evenly split between men and women compared to Pubmed authors, but the proportion has not reached parity. Estimated composition of gender prediction over the years of (A) all Pubmed computational biology and bioinformatics journal authors, (B) all ISCB fellows and keynote speakers, and (C) ISCB honorees in each honor category.

Assessing the Racial and Ethnic Diversity of Authors and Honorees

We predicted the race and ethnicity of authors and honorees using *wru*, which is based on US census data. We found that an increasing proportion of authors in computational biology and bioinformatics journals had last names associated with selecting Asian as a race/ethnicity category in the US census (Fig. 2A). This was primarily driven by publications in *Bioinformatics* and *BMC Bioinformatics* (Fig. 2B, top). We did not observe an increase at *PLOS Computational Biology* (Fig. 2B, bottom). A higher proportion of individuals who had last names associated with selecting white as a race/ethnicity category in the US census were ISCB honorees (Fig. 2C) than authors (Fig. 2A). Separating honoree results by honor category did not reveal any clear differences (Fig. 2D).

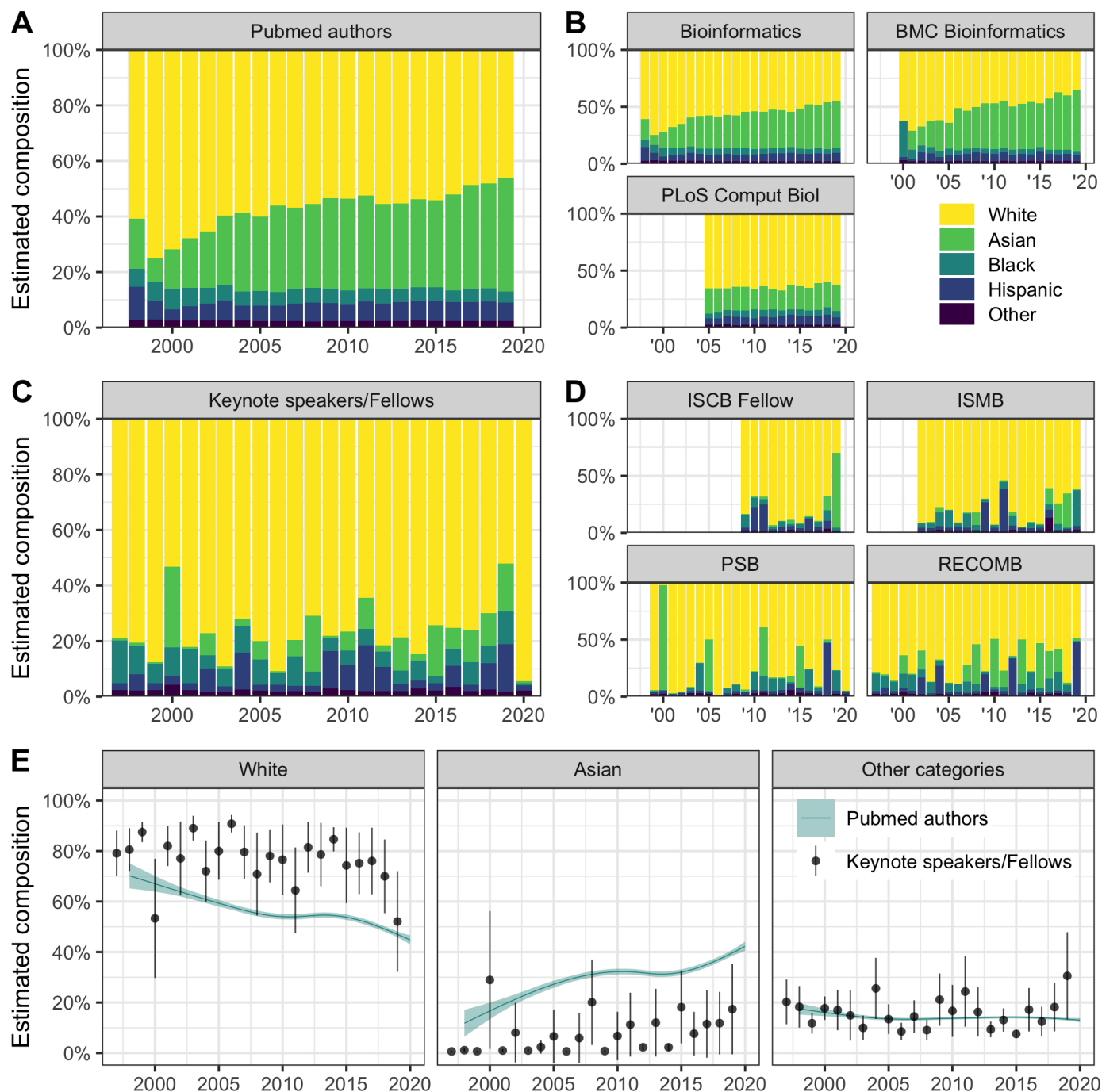


Figure 2: We find an overrepresentation of white and underrepresentation of Asian honorees as compared to authors. Estimated composition of census-based race/ethnicity prediction over the years of (A) all Pubmed computational biology and bioinformatics journal authors, (B) authors in each journal, (C) all ISCB Fellows and keynote speakers, (D) ISCB honorees in each honor category and (E) the composition's statistics of Pubmed authors (blue LOESS curve) and ISCB honorees (mean and 95% confidence interval, vertical) for each race/ethnicity category.

We directly compared honoree and author results from 1997 to 2020 for the predicted proportion of white, Asian, and other categories (Fig. 2E). We find that white honorees have been significantly overrepresented and Asian honorees have been significantly underrepresented in most years. Though we estimate the fraction of non-white and non-Asian authors to be relatively similar to the estimated honoree rate, we note that both are represented at levels substantially lower than in the US population. The proportion of Hispanic authors and honorees at these venues may also be influenced by authors from Spain instead of Latin America and is likely to understate the extent to which minoritized scientists are underrepresented among honorees and authors.

Predicting Nationality with LSTM Neural Networks and Wikipedia

We achieved a high overall accuracy on this test data with similar precision-recall values over the classes as we saw for our Wiki2019 dataset's test set, giving confidence in our model. We also tested the accuracy of the LSTM trained on the ethnicolr Wiki2009 dataset in predicting the labels for unseen names in our Wiki2019 dataset. Importantly, this model did not achieve as high an accuracy as the model trained on Wiki2019 and tested on either dataset, suggesting our new, larger dataset was an improvement.

Assessing the Nationality Diversity of Authors and Honorees

We applied our Wiki2019-LSTM model to both our computational biology honorees dataset and our dataset of corresponding authors. We saw proportional discrepancies between the two groups, namely a large overrepresentation of CelticEnglish (including American) keynote speakers and a large underrepresentation of East Asian keynote speakers. The proportion of keynote speakers of African, Muslim, and Hispanic origin were also slightly smaller than the field of computational biology as a whole, as represented by corresponding authors in major journals. Notably, our regional classifier did not distinguish between names from Iberia (Spain and Portugal) and names from Spanish and Portuguese-speaking countries in Latin America. Discrepancies in representation between these groups are thus undetectable by our classifier, but anecdotal evidence indicates that Iberian keynote speakers may make up a disproportionate amount of the Hispanic group.

Conclusions

We estimate the composition of the field using corresponding author status, but in neuroscience [18] and other disciplines [19] women are underrepresented among such authors. Such an effect would cause us to underestimate the number of women in the field. In addition to holding fewer corresponding authorship positions, female scientists are cited less often [20], invited by journals to submit papers less often [19] and suggested as reviewers less often [21]. Societies, both through their honorees and the individuals who deliver keynotes at their meetings, can play a positive role in improving the presence of female STEM role models, which, for example, may lead to higher persistence for undergraduate women in geoscience [22]. Efforts are underway to create Wikipedia entries for more female [23] and black, Asian, and minority scientists [24], which can help early-career scientists identify role models. We find that ISCB's honorees and keynote speakers, though not yet reaching gender parity, appear to be more evenly split between men and women than the field as a whole. On the other hand, we find that honorees include significantly few people of color than the field as a whole, and that Asian scientists are dramatically under-represented among honorees.

The central role that scientists play in evaluating each other and each other's findings makes diversity particularly critical. Even many nominally objective methods of assessing excellence (h-index, grant funding obtained, number of high-impact peer-reviewed publications, total number of peer-reviewed publications) are subject to the bias of peers during review. These could be affected by explicit biases, implicit biases, or pernicious biases in which a reviewer might consider a path of inquiry, as opposed to an individual, to be more or less meritorious based on the reviewer's own background [1]. Our efforts to measure the diversity of honorees in an international society suggests that, while a focus on gender parity may be improving some aspects of diversity among honorees, scientists of color do not appear to be recognized at levels consistent with their membership among the pool of potential honorees.

Data and Resource Availability

A manubot instance that hosts version history for this manuscript is available under the Creative Commons Attribution License at <https://github.com/greenelab/iscb-diversity-manuscript>. Our analysis of authors and ISCB-associated honorees is available under the Creative Commons Attribution

License at <https://github.com/greenelab/iscb-diversity>, with source code also distributed under a BSD 3-Clause License.

References

1. **Topic choice contributes to the lower rate of NIH awards to African-American/black scientists**
Travis A. Hoppe, Aviva Litovitz, Kristine A. Willis, Rebecca A. Meseroll, Matthew J. Perkins, B. Ian Hutchins, Alison F. Davis, Michael S. Lauer, Hannah A. Valantine, James M. Anderson, George M. Santangelo
Science Advances (2019-10) <https://doi.org/gghp8t>
DOI: [10.1126/sciadv.aaw7238](https://doi.org/10.1126/sciadv.aaw7238) · PMID: [31633016](https://pubmed.ncbi.nlm.nih.gov/31633016/) · PMCID: [PMC6785250](https://pubmed.ncbi.nlm.nih.gov/PMC6785250/)
2. **Ten Simple Rules to Achieve Conference Speaker Gender Balance**
Jennifer L. Martin
PLoS Computational Biology (2014-11-20) <https://doi.org/gf853n>
DOI: [10.1371/journal.pcbi.1003903](https://doi.org/10.1371/journal.pcbi.1003903) · PMID: [25411977](https://pubmed.ncbi.nlm.nih.gov/25411977/) · PMCID: [PMC4238945](https://pubmed.ncbi.nlm.nih.gov/PMC4238945/)
3. **How scientists are fighting against gender bias in conference speaker lineups**
Katie Langin
Science (2019-02-11) <https://doi.org/gghp8v>
DOI: [10.1126/science.caredit.aaw9742](https://doi.org/10.1126/science.caredit.aaw9742)
4. **Speaking out about gender imbalance in invited speakers improves diversity**
Robyn S Klein, Rhonda Voskuhl, Benjamin M Segal, Bonnie N Dittel, Thomas E Lane, John R Bethea, Monica J Carson, Carol Colton, Susanna Rosi, Aileen Anderson, ... Anne H Cross
Nature Immunology (2017-05) <https://doi.org/gghp8s>
DOI: [10.1038/ni.3707](https://doi.org/10.1038/ni.3707) · PMID: [28418385](https://pubmed.ncbi.nlm.nih.gov/28418385/) · PMCID: [PMC5775963](https://pubmed.ncbi.nlm.nih.gov/PMC5775963/)
5. **Addressing the underrepresentation of women in mathematics conferences**
Greg Martin
arXiv (2015-02-23) <https://arxiv.org/abs/1502.06326v1>
6. **The Presence of Female Conveners Correlates with a Higher Proportion of Female Speakers at Scientific Symposia**
Arturo Casadevall, Jo Handelsman
mBio (2014-01-07) <https://doi.org/qsh>
DOI: [10.1128/mbio.00846-13](https://doi.org/10.1128/mbio.00846-13) · PMID: [24399856](https://pubmed.ncbi.nlm.nih.gov/24399856/) · PMCID: [PMC3884059](https://pubmed.ncbi.nlm.nih.gov/PMC3884059/)
7. **Trends in the Proportion of Female Speakers at Medical Conferences in the United States and in Canada, 2007 to 2017**
Shannon M. Ruzyski, Sarah Fletcher, Madalene Earp, Aleem Bharwani, Kirstie C. Lithgow
JAMA Network Open (2019-04-12) <https://doi.org/gghp8r>
DOI: [10.1001/jamanetworkopen.2019.2103](https://doi.org/10.1001/jamanetworkopen.2019.2103) · PMID: [30977853](https://pubmed.ncbi.nlm.nih.gov/30977853/) · PMCID: [PMC6481599](https://pubmed.ncbi.nlm.nih.gov/PMC6481599/)
8. **Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency**
Holly O Witteman, Michael Hendricks, Sharon Straus, Cara Tannenbaum
The Lancet (2019-02) <https://doi.org/djc5>
DOI: [10.1016/s0140-6736\(18\)32611-4](https://doi.org/10.1016/s0140-6736(18)32611-4)
9. **Race, Ethnicity, and NIH Research Awards**
D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, R. Kington
Science (2011-08-18) <https://doi.org/csf8j8>
DOI: [10.1126/science.1196783](https://doi.org/10.1126/science.1196783) · PMID: [21852498](https://pubmed.ncbi.nlm.nih.gov/21852498/) · PMCID: [PMC3412416](https://pubmed.ncbi.nlm.nih.gov/PMC3412416/)

10. **Gender, Race/Ethnicity, and National Institutes of Health R01 Research Awards**
 Donna K. Ginther, Shulamit Kahn, Walter T. Schaffer
Academic Medicine (2016-08) <https://doi.org/f8zzw4>
 DOI: [10.1097/acm.0000000000001278](https://doi.org/10.1097/acm.0000000000001278) · PMID: [27306969](https://pubmed.ncbi.nlm.nih.gov/27306969/) · PMCID: [PMC4965301](https://pubmed.ncbi.nlm.nih.gov/PMC4965301/)
11. **Research in Computational Molecular Biology**
 Mona Singh (editor)
Lecture Notes in Computer Science (2016) <https://doi.org/gghp87>
 DOI: [10.1007/978-3-319-31957-5](https://doi.org/10.1007/978-3-319-31957-5)
12. **Research in Computational Molecular Biology**
 Vineet Bafna, S. Cenk Sahinalp (editors)
Lecture Notes in Computer Science (2011) <https://doi.org/dvwjgg>
 DOI: [10.1007/978-3-642-20036-6](https://doi.org/10.1007/978-3-642-20036-6)
13. **Research in Computational Molecular Biology**
 Teresa M. Przytycka (editor)
Lecture Notes in Computer Science (2015) <https://doi.org/gghwjh>
 DOI: [10.1007/978-3-319-16706-0](https://doi.org/10.1007/978-3-319-16706-0)
14. **Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records**
 Kosuke Imai, Kabir Khanna
Political Analysis (2016) <https://doi.org/f8ntmv>
 DOI: [10.1093/pan/mpw001](https://doi.org/10.1093/pan/mpw001)
15. **Predicting Race and Ethnicity From the Sequence of Characters in a Name**
 Gaurav Sood, Suriyan Laohaprapanon
arXiv (2018-05-05) <https://arxiv.org/abs/1805.02109v1>
16. **Name-ethnicity classification from open sources**
 Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, Steven Skiena
Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09 (2009) <https://doi.org/fs3pr8>
 DOI: [10.1145/1557019.1557032](https://doi.org/10.1145/1557019.1557032)
17. **Nationality Classification Using Name Embeddings**
 Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, Steven Skiena
Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17 (2017) <https://doi.org/ggjc78>
 DOI: [10.1145/3132847.3133008](https://doi.org/10.1145/3132847.3133008)
18. **Persistent Underrepresentation of Women's Science in High Profile Journals**
 Yiqin Alicia Shen, Jason M. Webster, Yuichi Shoda, Ione Fine
Cold Spring Harbor Laboratory (2018-03-02) <https://doi.org/cmh5>
 DOI: [10.1101/275362](https://doi.org/10.1101/275362)
19. **The gender gap in science: How long until women are equally represented?**
 Luke Holman, Devi Stuart-Fox, Cindy E. Hauser
PLOS Biology (2018-04-19) <https://doi.org/gdb9db>
 DOI: [10.1371/journal.pbio.2004956](https://doi.org/10.1371/journal.pbio.2004956) · PMID: [29672508](https://pubmed.ncbi.nlm.nih.gov/29672508/) · PMCID: [PMC5908072](https://pubmed.ncbi.nlm.nih.gov/PMC5908072/)

20. The extent and drivers of gender imbalance in neuroscience reference lists

Jordan D. Dworkin, Kristin A. Linn, Erin G. Teich, Perry Zurn, Russell T. Shinohara, Danielle S. Bassett
arXiv (2020-01-03) <https://arxiv.org/abs/2001.01002v1>

21. Journals invite too few women to referee

Jory Lerback, Brooks Hanson
Nature (2017-01) <https://doi.org/gf4jjz>
DOI: [10.1038/541455a](https://doi.org/10.1038/541455a) · PMID: [28128272](https://pubmed.ncbi.nlm.nih.gov/28128272/)

22. Role modeling is a viable retention strategy for undergraduate women in the geosciences

Paul R. Hernandez, Brittany Bloodhart, Amanda S. Adams, Rebecca T. Barnes, Melissa Burt, Sandra M. Clinton, Wenyi Du, Elaine Godfrey, Heather Henderson, Ilana B. Pollack, Emily V. Fischer
Geosphere (2018-10-31) <https://doi.org/gghp9d>
DOI: [10.1130/ges01659.1](https://doi.org/10.1130/ges01659.1)

23. Why we're editing women scientists onto Wikipedia

Jess Wade, Maryam Zaringhalam
Nature (2018-08-14) <https://doi.org/gdz52z>
DOI: [10.1038/d41586-018-05947-8](https://doi.org/10.1038/d41586-018-05947-8)

24. Why we're creating Wikipedia profiles for BAME scientists

Nicola O'Reilly
Nature (2019-03-07) <https://doi.org/gfwhcr>
DOI: [10.1038/d41586-019-00812-8](https://doi.org/10.1038/d41586-019-00812-8)