

Analysis of ISCB honorees and keynotes reveals disparities

A DOI-citable version of this manuscript is available at <https://doi.org/10.1101/2020.04.14.927251>.

This version of the manuscript [contains changes](#) subsequent to the [version 1.0 release](#).

This manuscript ([permalink](#)) was automatically generated from [greenelab/isb-diversity-manuscript@deaa7e2](#) on May 3, 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 · Funded by the Gordon and Betty Moore Foundation ([GBMF4552](#))

- **Ariel A. Hippen Anderson**

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

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

- **Matthew R. Gazzara**

 [0000-0001-7710-4551](#) ·  [mrgazzara](#) ·  [MR_Gazzara](#)

Department of Genetics, 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 · Funded by the Gordon and Betty Moore Foundation ([GBMF4552](#))

Abstract

Delivering a keynote talk at a conference organized by a scientific society, or being named as a fellow by such a society, indicates that a scientist is held in high regard by their colleagues. To explore if the distribution of such indicators of esteem in the field of bioinformatics reflects the composition of this field, we compared the gender, country of affiliation, race/ethnicity and name-origin of 412 researchers who had been recognized by the *International Society for Computational Biology* (75 fellows and 337 keynote speakers) with 34,050 researchers who had been the corresponding authors on papers in three leading bioinformatics journals between 1993 and 2019. The proportion of female fellows and keynote speakers was similar to that of the field overall. However, fellows and keynote speakers with an affiliation in the United States were over-represented by a factor of 1.6; moreover, almost half of this excess was accounted for by a deficit of 41 fellows and keynote authors from China, France and Italy. Furthermore, within the US we found an excess of white fellows and keynote speakers, and a depletion of Asian fellows and keynote speakers. Globally, names of East Asian origin have been persistently underrepresented among fellows and keynote speakers

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, a recent analysis found that minority scientists tend to apply for awards on topics with lower success rates [1]. This finding might be the result of minority scientists selecting topics in more poorly funded areas. Alternatively, reviewing scientists may not recognize the scientific importance of these topics, which may be of particular interest to minority scientists. One way to address this issue is to directly examine peer recognition in different scientific 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 name origin 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 that aim to have a global reach: [Intelligent Systems for Molecular Biology](#) (ISMB) and [Research in Computational Molecular Biology](#) (RECOMB).

We used multiple methods to predict the gender, race/ethnicity, and name origins 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 name origin. 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 non-white speakers and honorees. The lack of Asian scientists among keynote speakers and Fellows was particularly pronounced when compared against the field-specific background.

Results

Curated Honorees and Literature-derived Potential Honorees

We curated a dataset of ISCB honorees that included 412 honorees who were keynote speakers at international ISCB-associated conferences (ISMB and RECOMB) 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 1993–2019. 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 Orongo 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 almost 30,000 articles for evaluation. However, although PubMed provides author order, it does not provide corresponding author information. To determine corresponding authors for an article, we used the PMC corresponding author information when it was available (20,696 articles) and the PubMed last author as a fallback when corresponding author information was missing (9,053 articles).

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.

We observed a slow increase of the proportion of predicted female authors, arriving at just over 20% in 2019 (Fig. 1, left). We observe very similar trend within each journal, but estimated female proportion has increased the most in *BMC Bioinformatics* (see [notebook](#)). In recent years, 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. 1, right); however, it has not yet reached parity. Further, taking all the years together, a Welch two-sample t-test did not reveal any statistically significant difference in the mean probability of ISCB honorees predicted to be female compared to that of authors ($t_{418} = 0.753$, $p = 0.226$). We observed an increasing trend of honorees who were women in each honor category, especially in the group of ISCB Fellows (see [notebook](#)), which markedly increased after 2015. Through 2019, there were a number of years when meetings or ISCB Fellow classes have a high probability of recognizing only male honorees and none that appeared to have exclusively female honorees.

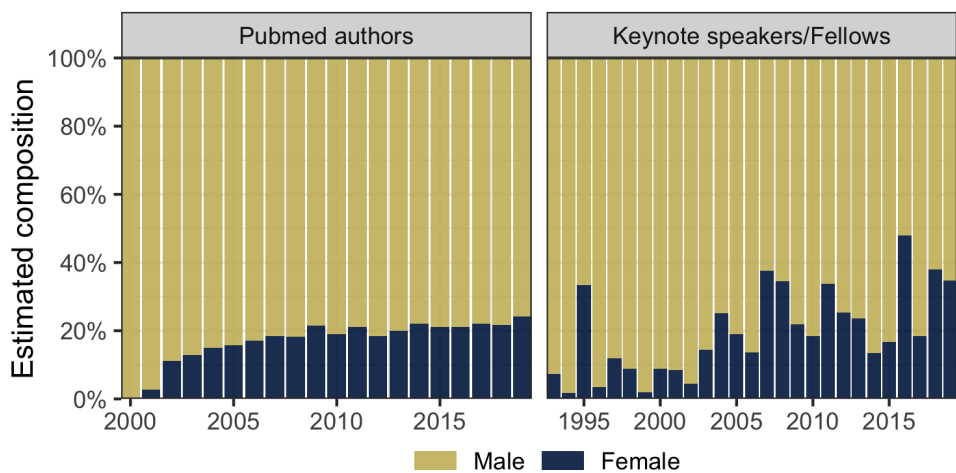


Figure 1: ISCB Fellows and keynote speakers appear more evenly split between men and women than PubMed authors in recent years, but the proportion has not reached parity. Estimated composition of gender prediction over the years of all Pubmed computational biology and bioinformatics journal authors (left), and all ISCB Fellows and keynote speakers (right) was computed as the average of prediction probabilities of Pubmed articles or ISCB honorees each year.

Predicting Name Origin Groups with LSTM Neural Networks and Wikipedia

Table 1 shows the size of the training set for each of the name origin groups as well as a few examples of PubMed author names that had at least 90% prediction probability in that group. We refer to this dataset as Wiki2019 (available online in [annotated_names.tsv](#)).

Table 1: Predicting name-origin groups of names trained on Wikipedia’s living people. The table lists the 10 groups and the number of living people for each region that the LSTM was trained on. Example names shows actual author names that received a high prediction for each region. Full information about which countries comprised each region can be found in the online dataset [country_to_region.tsv](#) .

Group	Training Size	Example Names
Celtic/English names	154,890	Adam O Hebb, Oliver G Pybus, David W Ritchie, James WJ Anderson, James W MacDonald, Robert Clarke
European names	78,157	Tracey M Filzen, Jos H Beijnen, Caroline Louis-Jeune, Christian Lorenzi, Boris Vassilev, Verena Heinrich
Hispanic names	66,931	Ramón Latorre, Antonio J Jimeno-Yepes, Felipe A Simão, Paulo S L de Oliveira, Juan Carlos Rodríguez-Manzaneque, Natalia Acevedo-Luna
East Asian names	54,588	Heejoon Chae, Wenchao Jiang, Haizhou Liu, Miho Uchida, Wenxuan Zhang, Jiali Feng
Arabic names	31,418	Hamidreza Chitsaz, Farzad Sangi, Habib Motieghader, Berke Ç Toptas, Ali Aliyari, Bülent Arman Aksoy
Nordic names	28,978	Cecilia M Lindgren, Ellen Larsen, Jesper R Gådin, Janne H Korhonen, Johan Åqvist, Jens Nilsson
South Asian names	20,025	Amitabh Chak, Matthew G Seetin, Matrika Gupta, Sumudu P Leelananda, VS Kumar Kolli, Swanand Gore
African names	17,826	Timothy Kinyanjui, Jammbe Z Musoro, Nyaradzo M Mgodì, Magambo Phillip Kimuda, Probhonjon Baruah, Adaoha E C Ihekwa
Hebrew names	4,549	Alexander J Sadovsky, Boris Shraiman, Gil Goldshlager, Eytan Adar, Aviva Peleg, Nir Esterman
Greek names	4,138	Gianni Panagiotou, Themis Lazaridis, Eleni Mijalis, Nikolaos Tsiantis, Konstantinos A Kyritsis, Dimitris E Messinis

We next aimed to predict the name origin groups of honorees and authors. We constructed a training dataset with more than 700,000 name-nationality pairs by parsing the English-language Wikipedia. We trained a LSTM neural network on n-grams to predict name groups. We found similar performance across 1, 2, and 3-grams; however, because the classifier required fewer epochs to train with 3-grams, we used this length in the model that we term Wiki2019-LSTM. Our Wiki2019-LSTM returns, for each given name, a probability of that name originating from each of the specified 10 groups. We observed a multiclass area under the receiver operating characteristic curve (AUC) score of 95.9% for the classifier, indicating that the classifier can recapitulate name origins with high sensitivity and specificity. For each individual group, the high AUC (above 92%, Fig. 2A) suggests that our classifier was sufficient for use in a broad-scale examination of disparities. We also observed that the model was well calibrated (Fig. 2B). We also examined potential systematic errors between pairs of name origin groupings with a confusion heatmap and did not find off-diagonal enrichment for any pairing (Fig. 2C).

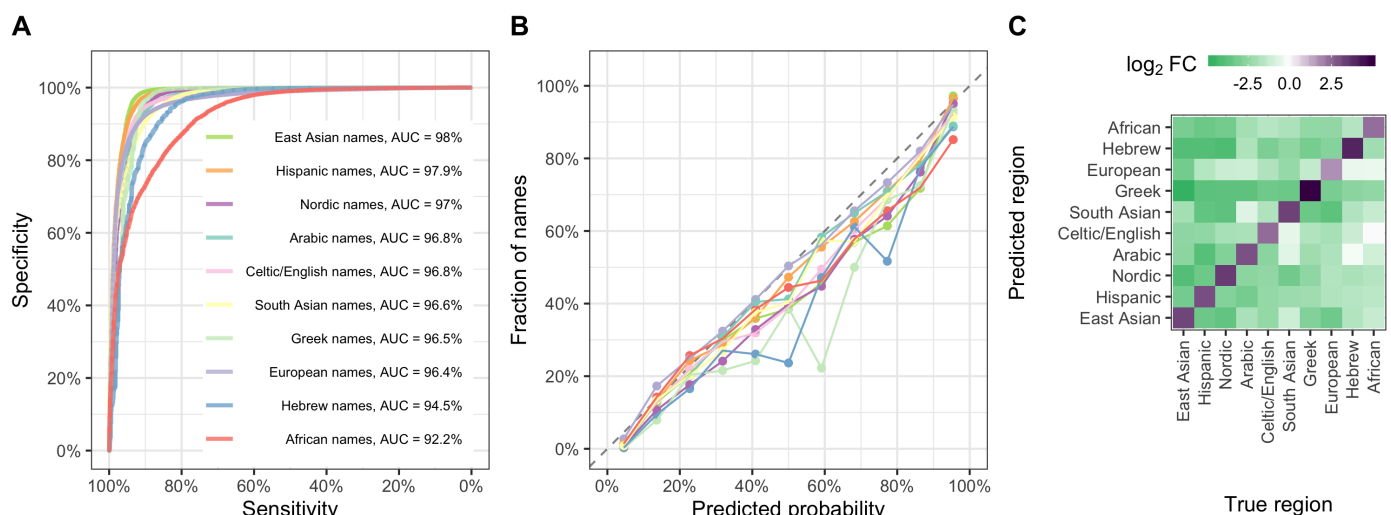


Figure 2: The Wiki2019-LSTM model performs well on the testing dataset. The area under the ROC curve is above 92% for each category, showing strong performance across origin categories (A). A calibration curve, computed with the caret R package, shows consistency between the predicted probabilities (midpoints of each fixed-width bin) and the observed fraction of names in each bin (B). Heatmap showing whether names from a given group (x-axis) received higher (purple) or lower (green) predictions for each group (y-axis) than would be expected by group prevalence alone (C). The values represent \log_2 fold change between the average predicted probability and the prevalence of the corresponding predicted group in the testing dataset (null). Scaling by group prevalence accounts for the imbalance of groups in the testing dataset. In all cases, the classifier predicts the true groups above the expected null probability (matrix diagonals are all purple). For off-diagonal cells, darker green indicates a lower mean prediction compared to the null. For example, the classifier does not often mistake East Asian names as Greek, but is more prone to mistaking South Asian names as Celtic/English.

Assessing the Name Origin 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 found that the proportion of authors with Celtic/English names had decreased (Fig. 3A, left), particularly for papers published in *Bioinformatics* and *BMC Bioinformatics* (see [notebook](#)). Among keynote speakers and fellows, we found that the majority of honorees are predicted to have Celtic/English or European names (Fig. 3A, right). When we directly compared honoree composition with PubMed, we observed discrepancies between the two groups, namely a large overrepresentation of keynote speakers with Celtic/English names ($t_{418} = 9.01$, $p < 10^{-16}$) and a substantial underrepresentation of keynote speakers with East Asian names ($t_{454} = -15.8$, $p < 10^{-16}$) (Fig. 3B). No statistically significant difference was observed between the proportion of honorees and authors with European names ($t_{425} = 0.0282$, $p = 0.489$) or in other categories (see Table 1, $t_{425} = -0.176$, $p = 0.860$).



Figure 3: Compared to the name collection of Pubmed authors, honorees with Celtic/English names are overrepresented while honorees with East Asian names are underrepresented. No statistically significant difference was observed between the proportion of honorees and authors with European names or in other categories (see Table 1). Estimated composition of name origin prediction over the years of (A, left) all Pubmed computational biology and bioinformatics journal authors, and (A, right) all ISCB Fellows and keynote speakers was computed as the average of prediction probabilities of Pubmed articles or ISCB honorees each year. (B) For each region, the mean predicted probability of Pubmed articles is shown as teal LOESS curve, and the mean probability and 95% confidence interval of the ISCB honoree predictions are shown as dark circles and vertical lines.

Affiliation Analysis

We analyzed the countries of affiliation between corresponding authors and ISCB honorees. For each country, we report a value of \log_2 enrichment (LOE) and its 95% confidence intervals (Table 2). A positive value of LOE indicates a higher proportion of honorees affiliated with that country compared to authors. A LOE value of 1 represents a one-fold enrichment (i.e., observed number of honorees is twice as much as expected). In the 20 countries with the most publications, we found an overrepresentation of honorees affiliated with institutions and companies in the US (97 speakers more than expected, LOE = 0.6, 95% CI (0.5, 0.8)) and Israel (12 speakers more than expected, LOR = 1.6, 95% CI (0.9, 2.3)), and an underrepresentation of honorees affiliated with those in China, France, Italy, the Netherlands, Taiwan, and India (Fig. 4).

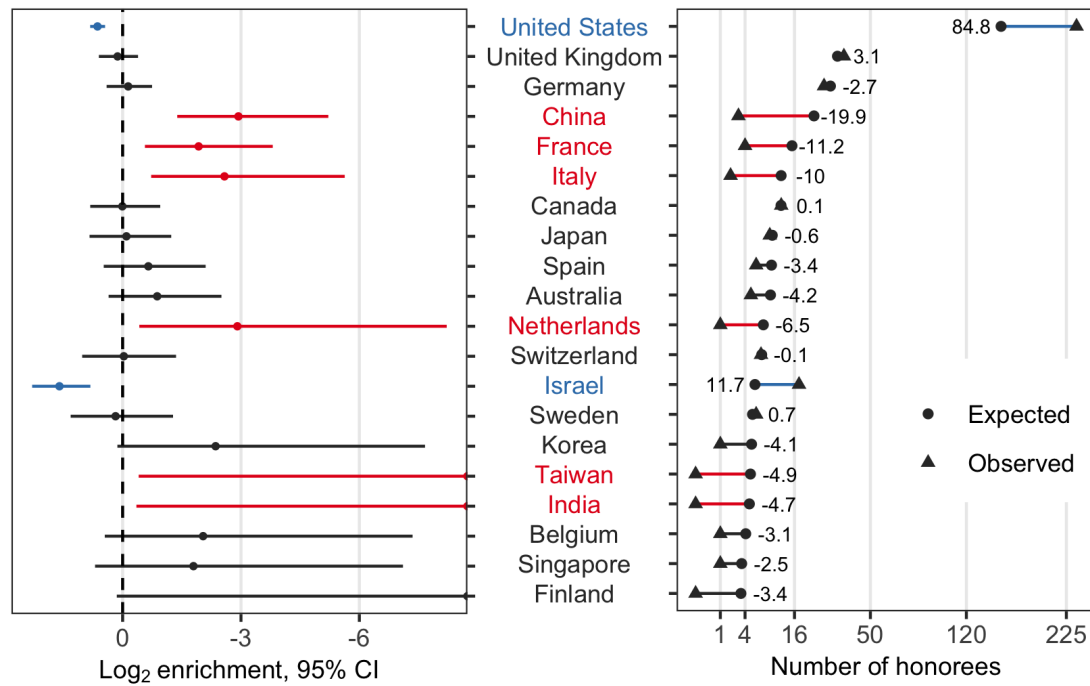


Figure 4: The overrepresentation of honorees affiliated with institutions and companies in the US and Israel contrasts the underrepresentation of honorees affiliated with those in China, France, Italy, the Netherlands, Taiwan, and India. Each country's log₂ enrichment (LOE) and its 95% confidence interval are displayed on the left. Observed (triangle) and expected (circle) number of honorees and their differences (observed - expected) are shown in square-root scale on the right. Countries are ordered based on the proportion of authors in the field.

Table 2: Enrichment and depletion in proportion of ISCB honorees compared to Pubmed corresponding authors of 20 countries with the most publications. The table lists the countries and their corresponding enrichment, which we computed by dividing the observed proportion of honorees by expected proportion of honorees. The expected proportion was calculated using corresponding author proportions. A positive Log₂(Enrichment) indicated a higher proportion of honorees than corresponding authors affiliated with that country. The full table with all countries can be browsed interactively in the corresponding [analysis notebook](#).

Country	Author proportion	Observed	Expected	Observed - Expected	Enrichment	Log ₂ (Enrichment)	95% Confidence Interval
United States	38.76%	237.5	152.7	84.8	1.6	0.6	(0.4, 0.8)
United Kingdom	8.36%	36.0	32.9	3.1	1.1	0.1	(-0.4, 0.6)
Germany	7.55%	27.0	29.7	-2.7	0.9	-0.1	(-0.7, 0.4)
China	5.82%	3.0	22.9	-19.9	0.1	-2.9	(-5.2, -1.4)
France	3.86%	4.0	15.2	-11.2	0.3	-1.9	(-3.8, -0.6)
Italy	3.04%	2.0	12.0	-10.0	0.2	-2.6	(-5.6, -0.7)
Canada	3.03%	12.0	11.9	0.1	1.0	0.0	(-1, 0.8)
Japan	2.44%	9.0	9.6	-0.6	0.9	-0.1	(-1.2, 0.8)
Spain	2.39%	6.0	9.4	-3.4	0.6	-0.7	(-2.1, 0.5)
Australia	2.33%	5.0	9.2	-4.2	0.5	-0.9	(-2.5, 0.4)
Netherlands	1.91%	1.0	7.5	-6.5	0.1	-2.9	(-8.2, -0.4)
Switzerland	1.81%	7.0	7.1	-0.1	1.0	-0.0	(-1.4, 1)
Israel	1.46%	17.5	5.8	11.7	3.0	1.6	(0.8, 2.3)
Sweden	1.34%	6.0	5.3	0.7	1.1	0.2	(-1.3, 1.3)

Country	Author proportion	Observed	Expected	Observed - Expected	Enrichment	Log ₂ (Enrichment)	95% Confidence Interval
Korea	1.30%	1.0	5.1	-4.1	0.2	-2.4	(-7.7, 0.1)
Taiwan	1.25%	0.0	4.9	-4.9	0.0	-Inf	(-Inf, -0.4)
India	1.20%	0.0	4.7	-4.7	0.0	-Inf	(-Inf, -0.3)
Belgium	1.04%	1.0	4.1	-3.1	0.2	-2.0	(-7.3, 0.5)
Singapore	0.88%	1.0	3.5	-2.5	0.3	-1.8	(-7.1, 0.7)
Finland	0.85%	0.0	3.4	-3.4	0.0	-Inf	(-Inf, 0.1)

Assessing the Racial and Ethnic Diversity of US-affiliated Authors and Honorees

We predicted the race and ethnicity of US-affiliated 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. 5A). This was primarily driven by publications in *Bioinformatics* and *BMC Bioinformatics* (Fig. 5B, top). We did not observe a corresponding increase at *PLOS Computational Biology* (Fig. 5B, bottom). Compared to Pubmed authors, ISCB honorees with US affiliations have a higher proportion of individuals whose last names we associated with selecting white as a race/ethnicity category in the US census (Fig. 5C vs. A). Separating honoree results by honor category did not reveal any clear differences (Fig. 5D).

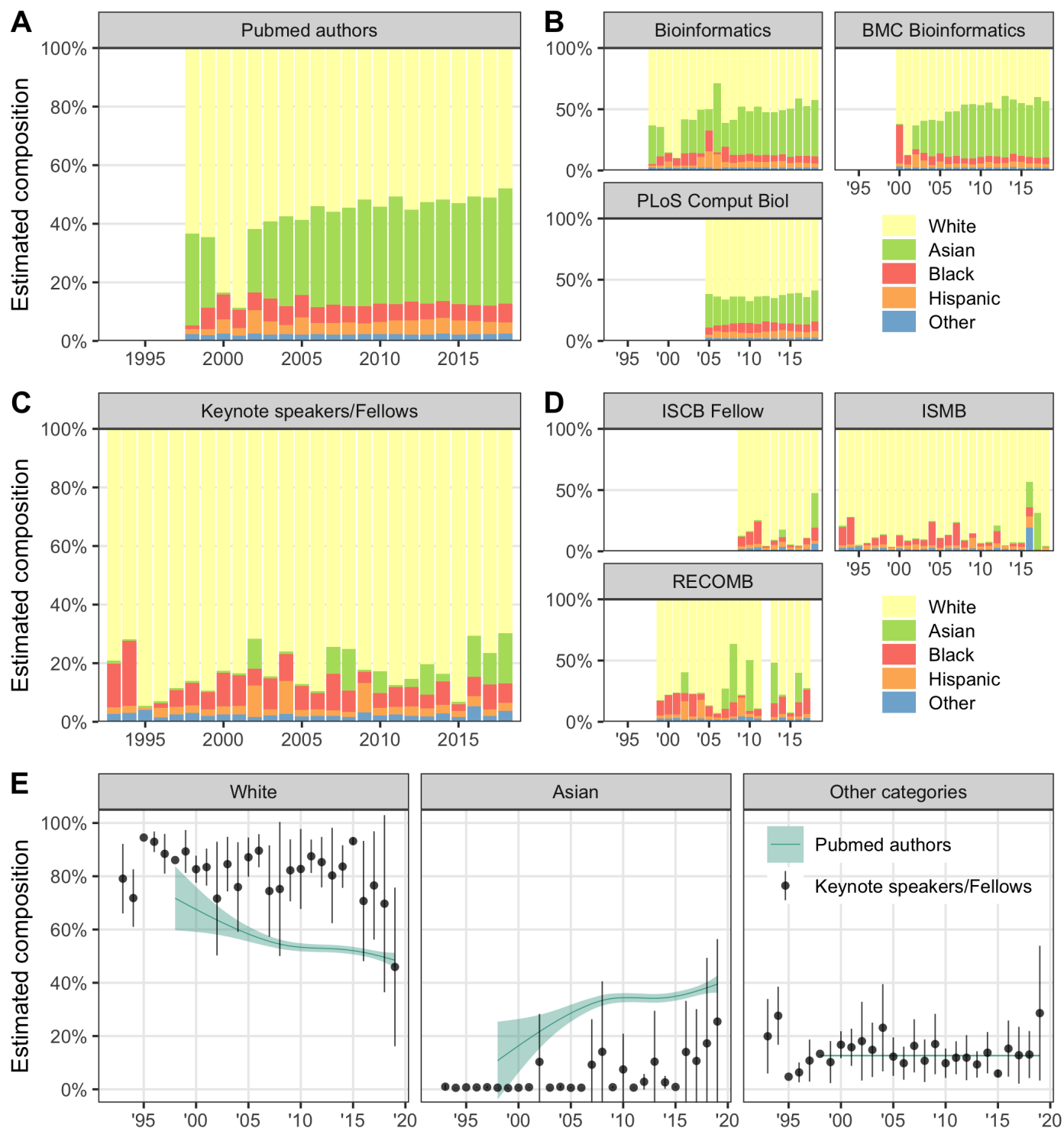


Figure 5: We find an overrepresentation of white and underrepresentation of Asian honorees as compared to authors. Estimated composition of census-based race/ethnicity prediction over the years of (A) all Pubmed computational biology and bioinformatics journal authors, (B) authors in each journal, (C) all ISCB Fellows and keynote speakers, and (D) ISCB honorees in each honor category was computed as the average of prediction probabilities of Pubmed articles or ISCB honorees each year. For each race/ethnicity category, the mean predicted probability of Pubmed articles is shown as teal LOESS curve, and the mean probability and 95% confidence interval of the ISCB honoree predictions are shown as dark circles and vertical lines (E).

We directly compared honoree and author results from 1993 to 2019 for the predicted proportion of white, Asian, and other categories (Fig. 5E). We found that, over the years, white honorees have been significantly overrepresented ($t_{218} = 14.8$, $p < 10^{-16}$) and Asian honorees have been significantly underrepresented ($t_{236} = -18.8$, $p < 10^{-16}$). A Welch two-sample t-test did not reveal any statistically significant difference in the mean probability of ISCB speakers predicted to be in Other categories compared to authors ($t_{203} = 1.79$, $p = 0.0747$).

Assessing the Name Origin Diversity of US-affiliated Authors and Honorees

We applied our Wiki2019-LSTM model to US-affiliated authors and honorees and found similar pattern as when we considered all authors and honorees. We note that the US was not included in the training of the Wiki2019-LSTM model (see Methods). Specifically, the proportion of authors with Celtic/English names had decreased (Fig. 6A, left), particularly for papers published in *Bioinformatics* and *BMC Bioinformatics* (see [notebook](#)). Meanwhile, the majority of honorees are predicted to have Celtic/English or European names (Fig. 6A, right). When we directly compared honoree composition with PubMed, we observed discrepancies between the two groups, namely a large overrepresentation of honorees with Celtic/English names ($t_{247} = 6.46$, $p < 10^{-9}$), a smaller overrepresentation of honorees with European names ($t_{250} = 2.54$, $p = 0.0059$), and a substantial underrepresentation of honorees with East Asian names ($t_{329} = -20.6$, $p < 10^{-16}$) (Fig. 6B).

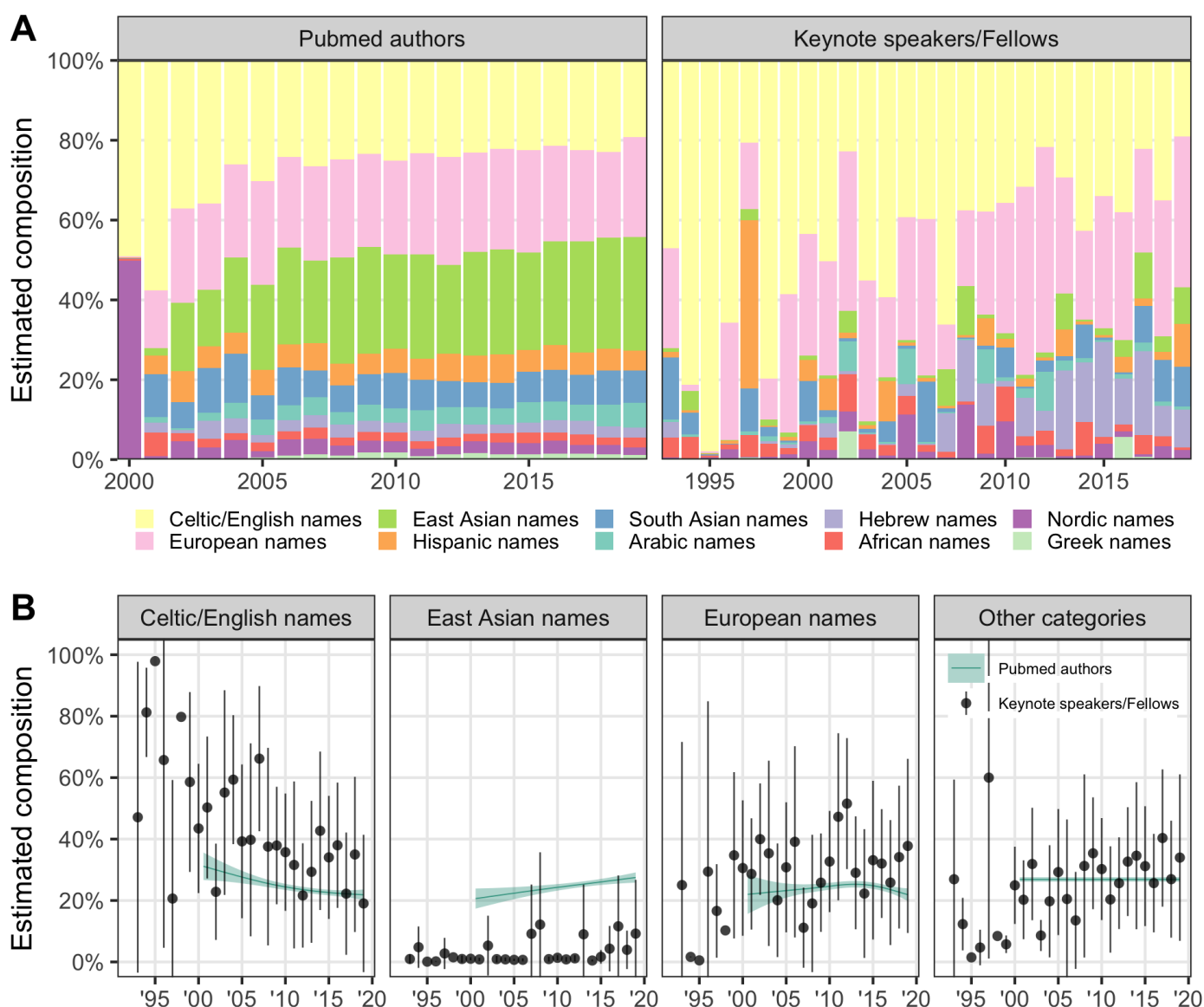


Figure 6: Compared to the name collection of US-affiliated Pubmed authors, US-affiliated honorees with Celtic/English and European names are overrepresented while US-affiliated honorees with East Asian names are underrepresented. No statistically significant difference was observed between the proportion of honorees and authors with names in other categories (see Table 1, $t_{252} = -0.152$, $p = 0.879$). Estimated composition of name origin prediction over the years of (A, left) US-affiliated journal authors, and (A, right) all US-affiliated Fellows and keynote speakers was computed as the average of prediction probabilities of US-affiliated corresponding authors or ISCB honorees each year. (B) For each region, the mean predicted probability of US-affiliated corresponding authors is shown as teal LOESS curve, and the mean probability and 95% confidence interval of the US-affiliated ISCB honoree predictions are shown as dark circles and vertical lines.

Reception & Feedback

We released [version 1.0](#) of our manuscript on 2020-01-30. We submitted this version to the 28th annual ISMB conference planned for July 2020 in Montreal. The study received public feedback almost immediately. Specifically, a community member [pointed out](#) problems with the country-group names we imported from NamePrism.

On 2020-02-20, ISCB released a [statement on equity and inclusion](#) that alluded to our study, but did not directly name or link to it. The statement contained the following paragraph:

We continually revisit our procedures for progress towards equity and inclusion and assess their efficacy. We do not think that automated procedures of assessment are appropriate for this purpose. Algorithms performing such categorization suffer from inherent limitations that distort the notion of balance we are striving for. Specifically, categorization of disjoint regional, ethnic or gender groups of scientists, while easing classification, also functions divisively.

Our ISMB submission was reviewed by 5 peer reviewers and [rejected](#) on 2020-03-13. We thank the reviewers for their input, and have made our [responses public](#)).

Much of the feedback centered around the inherent imperfections of automated detection of race or nationality. Therefore, we expanded our assessment to include country of affiliation, which can be detected with high fidelity from PubMed metadata. Adding affiliation information allowed us to examine disparities within a single country and to rule out geographic disparities as the sole factor leading to an underrepresentation of Asian honorees.

We appreciate the feedback we have received. Early indications suggest ISCB may be heeding the findings of our analysis and increasing the diversity of honorees. In 2020, among 12 ISCB Fellows and 5 ISMB keynote speakers, the mean predicted probability of each honoree having an East Asian name is 33%, higher than any estimate in previous years (see [notebook](#)). The set of honorees also includes the first ISCB Fellow from China. Compared to past years, the 2020 honorees appear to better reflect the diversity of scientists in the computational biology field.

Materials and Methods

Honoree Curation

From [ISCB's webpage listing ISCB Distinguished Fellows](#), we found recipients listed by their full names 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 identified major ISCB-associated conferences as those designated flagship (ISMB) or those that had been held on many continents (RECOMB). To identify **ISMB Keynote Speakers**, we examined the webpage for each ISMB meeting. The invited speakers at ISMB before 2001 were listed in the Preface pages of each year's proceedings, which were archived in the [ISMB collection](#) of the AAI digital library. We found full names of all keynote speakers for the years 1993–2019.

For the 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 included 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 above-mentioned tables, we excluded this speaker from our dataset.

Name processing

When extracting honoree names, we began with the full name as provided on the site. Because our prediction methods required separated first and last names, 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. For metadata from PubMed and PMC where first (fore) and last names are coded separately, we applied the same cleaning steps. We created [functions to simplify names](#) in the pubmedpy Python package to support standardized fore and last name processing.

Corresponding author extraction

We assumed that, in the list of authors for a specific paper, corresponding authors (often research advisors) would be 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 corresponding authors per publication.

We evaluated two resources for extracting corresponding authors from papers: [PubMed](#) and [PubMed Central](#) (PMC). Both resources are provided by the US 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 first and last names and their order within a record. 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 fell back to PubMed data (9,053 articles). Almost all articles without PMC data were from *Bioinformatics* because it is 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 inversely weighted by the number of corresponding authors per paper.

Countries of Affiliations

Publications often provide affiliation lists for authors, which generally associate authors with research organizations and their corresponding physical addresses. We implemented affiliation extraction in the pubmedpy Python package for both PubMed and PMC XML records. These methods extract a sequence of textual affiliations for each author. While ideally each affiliation record would refer to one and only one research organization, sometimes journals deposit multiple affiliations in a single structured affiliation. For example, we extracted the following composite affiliation for all authors of [PMC4147893](#):

‘Multimodal Computing and Interaction’, Saarland University & Department for Computational Biology and Applied Computing, Max Planck Institute for Informatics, Saarbrücken, 66123 Saarland, Germany, Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, 15206 PA, USA, Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany, Université Pierre et Marie Curie, UMR7238, CNRS-UPMC, Paris, France and CNRS, UMR7238, Laboratory of Computational and Quantitative Biology, Paris, France.

We designed a method for extracting countries from affiliations that accommodated multiple countries. We relied on two Python utilities to extract countries from text: [geotext](#) and [geopy.geocoders.Nominatim](#). The first, [geotext](#), used regular expressions to find mentions of places from the [GeoNames database](#). In the above text, [geotext](#) detected four mentions of places in Germany: Saarland, Saarbrücken, Saarland, Germany. Anytime [geotext](#) identified 2 or more mentions of a country, we labeled the affiliation as including that country.

[geopy.geocoders.Nominatim](#) converts names / addresses to geographic coordinates using the OpenStreetMap’s [Nominatim](#) service. We split textual affiliations by punctuation and found the first segment, in reverse order, that returned any Nominatim search results. For the above affiliation, the search order was “France”, “Paris”, “Laboratory of Computational and Quantitative Biology”, etcetera. Since searching “France” returns a match by Nominatim, the following queries would not be made. When a match was found, we extracted the country containing the location. This approach returns a single country for an affiliation when successful. When labeling affiliations with countries, we only used these values when [geotext](#) did not return results or had ambiguity amongst countries without multiple matches. For more details on this approach, consult the accompanying [notebook](#) and [label dataset](#).

Our ability to assign countries to authors was largely driven by the availability of affiliations. In our catalog of corresponding authors, 95.7% of authors from PubMed Central records were assigned at least one country. The country-assignment-rate for corresponding authors from PubMed records was only 32.2%. This reflects the varying availability of affiliation metadata by journal. Countries were assigned to all corresponding authors for 99.9% of articles in *BMC Bioinformatics* 98.7% in *PLOS Computational Biology*, and 49.6% in *Bioinformatics*. Country assignment for corresponding authors in *Bioinformatics* were unavailable for over half of all articles annually from 1998–2013. *Bioinformatics* affiliation metadata improved starting in 2014, and the country-assignment-rate has hovered around ~95% for the last few years (see [notebook](#)).

For ISCB honorees, during the curation process, if an honoree was listed with their affiliation at the time, we recorded this affiliation for analysis. For ISCB Fellows, we used the affiliation listed on the ISCB page.

Because we could not find affiliations for the 1997 and 1998 RECOMB keynote speakers' listed for these years, they were left blank. If an author or speaker had more than one affiliation, each was inversely weighted by the number of affiliations that individual had.

Estimation of Gender

We predicted the gender of honorees and authors using the <https://genderize.io> API, which was trained on over 100 million name-gender pairings collected from the web and is one of the three widely-used gender inference services that provide gender classifications with over 98% accuracy [14]. We used author and honoree first names to retrieve predictions from genderize.io. The predictions represent the probability of an honoree or author being male or female. We used the estimated probabilities and did not convert to a hard group assignment. For example, a query to <https://genderize.io> on January 26, 2020 for "Casey" returns a probability of male of 0.74 and a probability of female of 0.26, which we would add for an author with this first name. Because of the limitations of considering gender as a binary trait and inferring it from first names, we only consider predictions in aggregate and not as individual values for specific scientists.

Of 412 ISCB honorees, genderize.io fails to provide gender predictions for three names. Of 34,005 corresponding authors, 45 were missing a fore name altogether in the raw paper metadata and 1,466 had fore names consisting of only initials. Of the remaining authors, genderize.io failed to predict gender for 1,578 of these fore names. We note that approximately 52% of these NA predictions are hyphenated names, which is likely because they are more unique and thus are more difficult to find predictions for. 79% of these names were predicted to be of Asian origin by last name (see the race/ethnicity prediction model below). This bias of NA predictions toward non-English names has been previously observed [15] and may have a minor influence on the final estimate of gender compositions.

Estimation of Name Origin Groups

We developed a model to predict geographical origins of names. The existing Python package *ethnicolr* [16] produces reasonable predictions, but its international representation in the data curated from Wikipedia in 2009 [17] is still limited. For instance, 76% of the names in *ethnicolr*'s Wikipedia dataset are European in origin.

To address these limitations in *ethnicolr*, we built a similar classifier, a Long Short-term Memory (LSTM) neural network, to infer the region of origin from patterns in the sequences of letters in full names. We applied this model on an updated, approximately 4.5 times larger training dataset called Wiki2019 (described below). We tested multiple character sequence lengths and, based on this comparison, selected tri-characters for the primary results described in this work. We trained our prediction model on 80% of the Wiki2019 dataset and evaluated its performance using the remaining 20%. This model, which we term Wiki2019-LSTM, is available in the online file [LSTM.h5](#).

To generate a training dataset for name origin prediction that reflects a modern naming landscape, we scraped the English Wikipedia's category of [Living People](#). This category, which contained approximately 930,000 pages at the time of processing in November 2019, is regularly curated and allowed us to avoid pages related to non-persons. For each Wikipedia page, we used two strategies to find a full birth name and location context for that person. First, we used information from the personal details sidebar; the information in this sidebar varied widely but often 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 ..." This structure comes from editor [guidance on biography articles](#) and is designed to capture:

... the country of which the person is a citizen, national or permanent resident, or if the person is notable mainly for past events, the country where the person was a citizen, national or permanent resident when the person became notable.

We used regular expressions to parse out the person's name from this structure and checked that the expression after "is a" matched a list of nationalities. We were able to define a name and nationality for 708,493 people by using the union of these strategies. This process produced country labels that were more fine-grained than the broader patterns that we sought to examine among honorees and authors. We initially grouped names by continent, but later decided to model our categorization after the hierarchical taxonomy used by [NamePrism](#) [18]. The NamePrism taxonomy was derived from name-country pairs by producing an embedding of names by Twitter contact patterns and then grouping countries using the similarity of names from those countries. The countries associated with each grouping are shown in Fig 7. NamePrism excluded the US, Canada and Australia because these countries have been populated by a mix of immigrant groups [18].

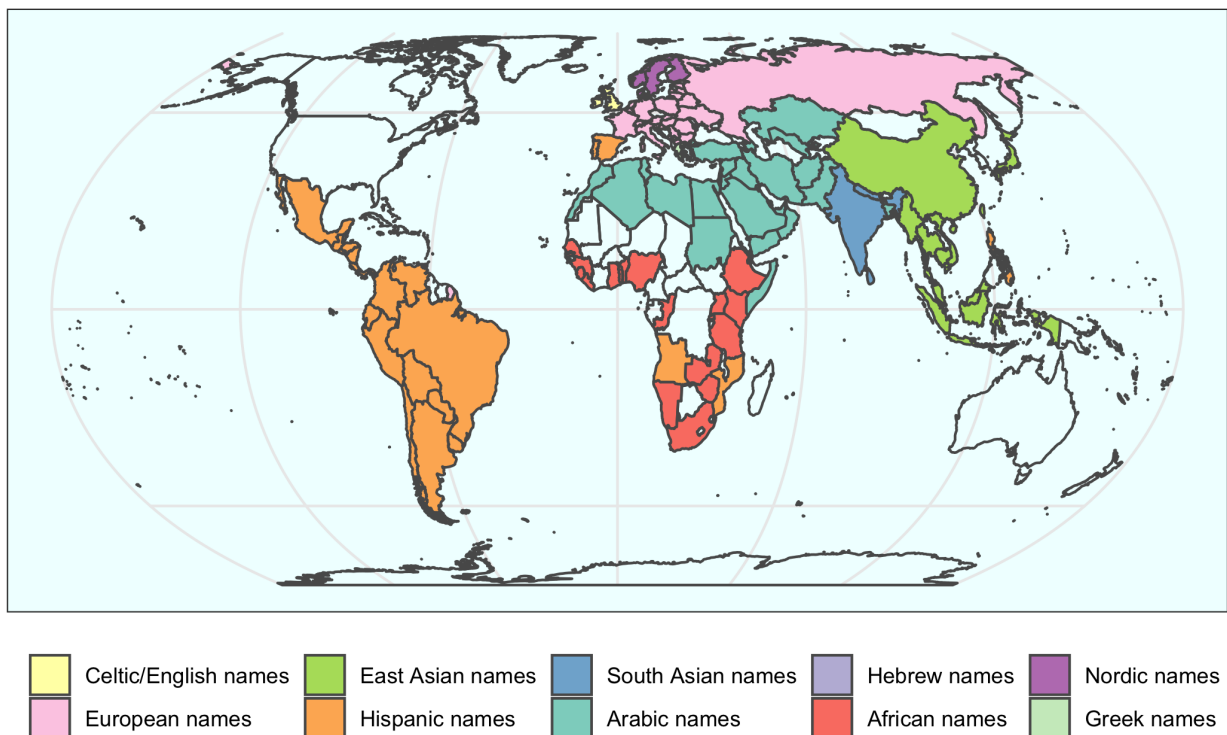


Figure 7: NamePrism groups countries by name similarity. We used this grouping but renamed the groups to focus on the linguistic patterns based on name etymology identified by NamePrism.

In an earlier version of this manuscript, we also used category names derived from NamePrism, but a reader [pointed out](#) the titles of the groupings were problematic; therefore, in this version, we renamed these groupings to reflect that the NamePrism approach primarily identifies groups based on linguistic patterns from name etymology rather than religious or racial similarities. We note that our mapping from nationality to name origins was not without error. For example, a scientist of Israeli nationality may not bear a Hebrew name. These mismatches were assessed via the heatmap of the model performance (Fig. 2C) and complemented by the affiliation analysis below. An alternative approach is to assign arbitrary names to these groups such as via letter coding (e.g., A, B, C, etc.), but we did not choose this strategy because ten arbitrary letters for ten groups can greatly reduce the paper's readability.

Affiliation Analysis

For each country, we computed the expected number of honorees by multiplying the proportion of authors whose affiliations were in that country with the total number of honorees. We then performed an enrichment analysis to examine the difference in country affiliation proportions between ISCB

honorees and field-specific corresponding authors. We calculated each country's enrichment by dividing the observed proportion of honorees by the expected proportion of honorees. The 95% confidence interval of the \log_2 enrichment was estimated using the Poisson model method [19].

Estimation of Race and Ethnicity within the US

We predicted the race and ethnicity of US-affiliated honorees and authors using the R package `wru`. `wru` implements methods described in Imai and Khanna [20] to predict race and ethnicity using surname and location information. The underlying data used for prediction are derived from the US Census, in which an individual's race and ethnicity are based on their self-identification with one or more groups. Specifically, the race categories include White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other race, and Two or more races [21], and ethnicity categories include Hispanic/Latino or Not Hispanic/Latino [22]. `wru` uses similar race/ethnicity categories but groups American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander to form the Other category.

We used only the surname of author or honoree to make predictions via the `predict_race()` function from `wru`. However, in the case of names that were not observed in the census, the function outputs the average demographic distribution from the census, which may produce misleading results. To avoid this suboptimal imputation, we modified the function to return a status denoting that results were inconclusive (NA) 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 US.

Of 239 US-affiliated ISCB honoree entries, `wru` fails to provide race/ethnicity predictions for 45 names. Of 11,545 US-affiliated corresponding authors, 2,352 had a last name for which `wru` did not provide predictions. One limitation of `wru` and other methods that infer race, ethnicity, or nationality from last names is the potentially inaccurate prediction for scientists who changed their last name during marriage, a practice more common among women than men.

Data and Resource Availability

This manuscript was written [openly on GitHub](#) using Manubot [23]. The Manubot HTML version is available under a Creative Commons Attribution (CC BY 4.0) License at <https://greenelab.github.io/iscb-diversity-manuscript/>. Our analysis of authors and ISCB-associated honorees is available under CC BY 4.0 at <https://github.com/greenelab/iscb-diversity>, with source code also distributed under a BSD 3-Clause License. Rendered Python and R notebooks from this repository are browsable at <https://greenelab.github.io/iscb-diversity/>. Our analysis of PubMed, PubMed Central, and author names relies on the Python `pubmedpy` package, developed as part of this project and available under a Blue Oak Model License 1.0 at <https://github.com/dhimmel/pubmedpy> and on [PyPI](#). Our Wikipedia name dataset is dedicated to the public domain under CC0 License at <https://github.com/greenelab/wiki-nationality-estimate>, with source code to construct the dataset available 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 (2019-10-09) <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 (2017-05-01) <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 (2015-02-24) <https://arxiv.org/abs/1502.06326>
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. Ruzycki, 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**
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**
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**
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. Comparison and benchmark of name-to-gender inference services

Lucía Santamaría, Helena Mihaljević
PeerJ Computer Science (2018-07-16) <https://doi.org/ggpbn6>
DOI: [10.7717/peerj-cs.156](https://doi.org/10.7717/peerj-cs.156)

15. Gender Prediction Methods Based on First Names with genderizeR

Kamil Wais
The R Journal (2016) <https://doi.org/gf4zqx>
DOI: [10.32614/rj-2016-002](https://doi.org/10.32614/rj-2016-002)

16. Predicting Race and Ethnicity From the Sequence of Characters in a Name

Gaurav Sood, Suriyan Laohaprapanon
arXiv (2018-07-31) <https://arxiv.org/abs/1805.02109>

17. Name-ethnicity classification from open sources

Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, Steven Skiena
Association for Computing Machinery (ACM) (2009) <https://doi.org/fs3pr8>
DOI: [10.1145/1557019.1557032](https://doi.org/10.1145/1557019.1557032)

18. Nationality Classification Using Name Embeddings

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, Steven Skiena
Association for Computing Machinery (ACM) (2017) <https://doi.org/ggjc78>
DOI: [10.1145/3132847.3133008](https://doi.org/10.1145/3132847.3133008)

19. Statistics in epidemiology: methods, techniques, and applications

Hardeo Sahai, Anwer Khurshid
CRC Press (1996)
ISBN: [9780849394447](https://www.isbn-international.org/product/9780849394447)

20. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records

Kosuke Imai, Kabir Khanna
Political Analysis (2017-01-04) <https://doi.org/f8ntmv>
DOI: [10.1093/pan/mpw001](https://doi.org/10.1093/pan/mpw001)

21. <https://data.census.gov/cedsci/deeplinks?url=https://factfinder.census.gov/help/en/race.htm>

22. Census Bureau News(2001-04-05) <https://web.archive.org/web/20010405061504/http://www.census.gov/Press-Release/www/2001/raceqandas.html>

23. Open collaborative writing with Manubot

Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) <https://doi.org/c7np>
DOI: [10.1371/journal.pcbi.1007128](https://doi.org/10.1371/journal.pcbi.1007128) · PMID: [31233491](https://pubmed.ncbi.nlm.nih.gov/31233491/) · PMCID: [PMC6611653](https://pubmed.ncbi.nlm.nih.gov/PMC6611653/)