bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

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/iscb-diversity-manuscript@b372983 on September 22, 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, name-origin, country of affiliation and race/ethnicity of 412 researchers who had been recognized by the *International Society for Computational Biology* (75 fellows and 337 keynote speakers) with over 170,000 researchers who had been the last authors on computational biology papers between 1993 and 2019. The proportion of female fellows and keynote speakers was similar to that of the field overall. However, names of East Asian origin have been persistently underrepresented among fellows and keynote speakers. Moreover, fellows and keynote speakers with an affiliation in the United States were overrepresented by a factor of 2.1; almost two thirds of this excess was accounted for by a deficit of 101 fellows and keynote authors from China, India, France and Italy. Within the US, we found an excess of white fellows and keynote speakers and a depletion of Asian 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:

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

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 minoritized communities. Furthermore, even though minoritized groups innovate at higher rates, their novel contributions are discounted [1]. One first step to address this systemic 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]. Another recent analysis found that minority scientists tend to apply for awards on topics with lower success rates [11]. 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.

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 authors for more than 170,000 computational biology publications 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

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 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 authors of bioinformatics and computational biology manuscripts. We scraped PubMed for manuscripts written in English from 1993–2019 with the MeSH term "computational

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

biology". We downloaded the metadata of manuscripts published in these journals from PubMed, which provided 173,735 articles for evaluation. For each article, we extracted its last author's fore name and last name for analysis.

In a previous version of this work, instead of last authors, we examined corresponding authors of articles in three well-recognized computational biology journals, and the results were consistent with our findings here [12].

Similar gender proportion between ISCB's honorees and the field

We observed a gradual increase of the proportion of predicted female authors, arriving at 28.8% in 2019 (Fig. 1, left). In recent years, ISCB Fellows and keynote speakers appear to have similar gender proportion compared to the population of authors published in computational biology and bioinformatics journals (Fig. 1, right); however, it has not yet reached parity. Examining each honor category, we observed an increasing trend of honorees who were women, 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. We sought to examine whether or not there was a difference in the proportion of female names between authors and honorees. A multiple linear regression of this proportion for the groups and year did not reveal a significant difference (p = 0.558).

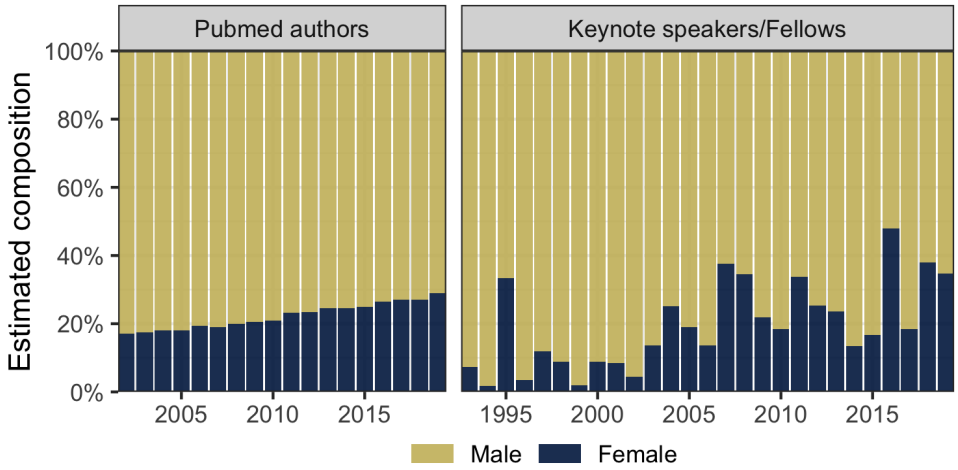

Figure 1: Estimated composition of gender prediction over the years of all Pubmed computational biology and bioinformatics journal authors (left), and all ISCB Fellows and keynote speakers (right). Male proportion (yellow) was computed as the average of the probability of being male of last authors (weight accordingly) or ISCB honorees each year. Female proportion (blue) was the complement of the male proportion. ISCB Fellows and keynote speakers appear have similar gender proportion compared to that of PubMed authors, but the proportion has not reached parity.

Honorees with Celtic/English names are overrepresented while honorees with East Asian names are underrepresented

We applied our Wiki2019-LSTM model to both our computational biology honorees dataset and our dataset of last authors. We found that the proportion of authors with Celtic/English names had decreased (Fig. 2A, left). Among keynote speakers and fellows, we found that the majority of honorees are predicted to have Celtic/English or European names (Fig. 2A, right). When we directly compared honoree composition with PubMed, we observed discrepancies between the two groups (Fig. 2B). A name coming from the group of honorees has significantly higher probability of being Celtic/English ($\beta_{\text{Celtic/English}}$ = 0.11924, $p < 10^{-9}$) and lower probability of being East Asian ($\beta_{\text{East Asian}}$ = -0.14791, $p < 10^{-9}$). The two groups of scientists did not have a significant association with names predicted to be European and in Other categories (p = 0.39475 and p = 0.48625, respectively).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

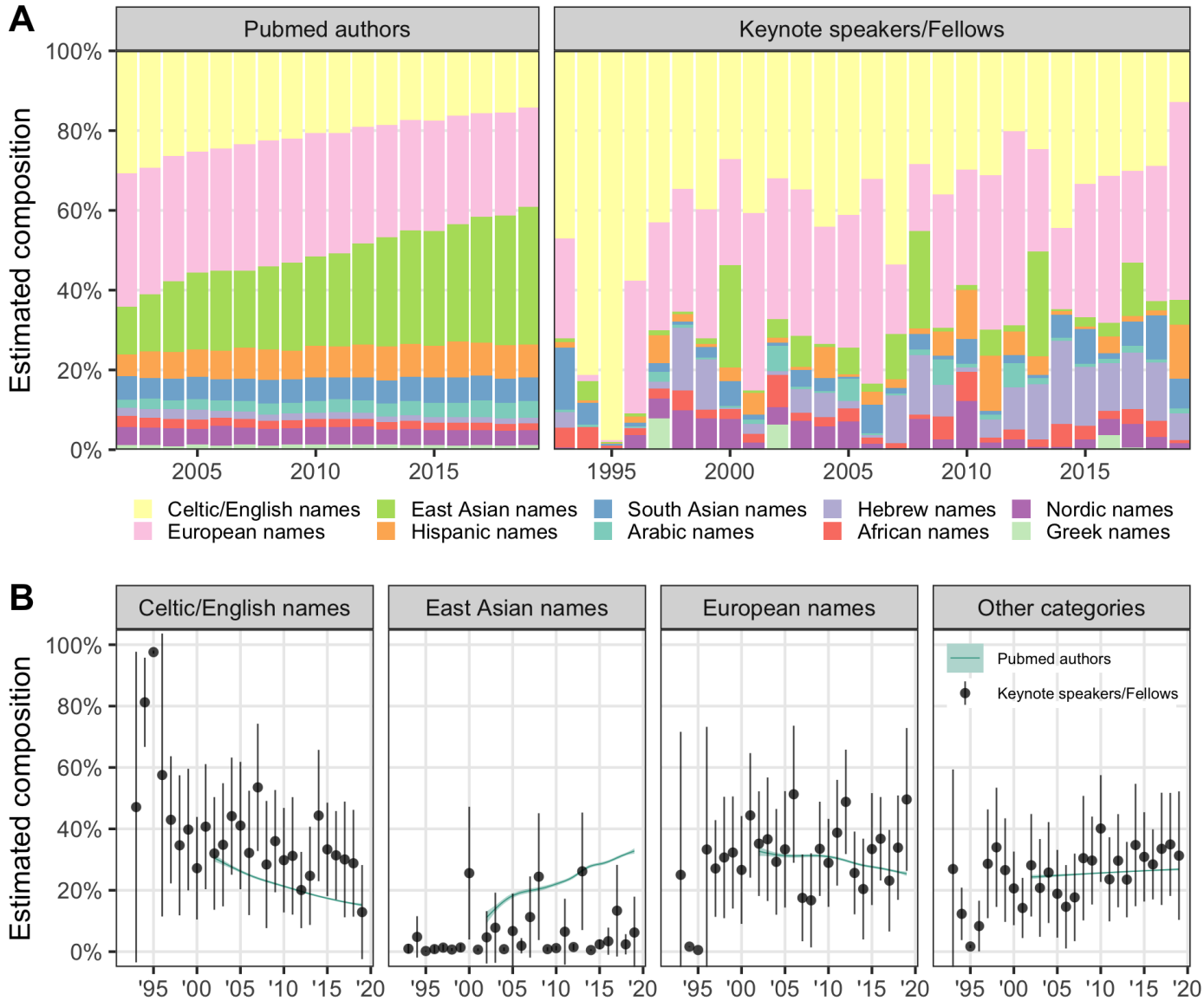

Figure 2: (A) Estimated composition of name origin prediction over the years of all Pubmed computational biology and bioinformatics journal authors (left), and all ISCB Fellows and keynote speakers (right). (B) For each region, the mean predicted probability of Pubmed articles is shown as teal GAM curve, and the mean probability and 95% confidence interval of the ISCB honoree predictions are shown as dark circles and vertical lines. 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).

We reached similar conclusion when applying the Wiki2019-LSTM model to the name origins of only US-affiliated scientists. We note that the US was not included in the training of the Wiki2019-LSTM model (see Methods). We found an overrepresentation of honorees with Celtic/English names ($\beta_{\text{Celtic/English}}$ = 0.053, p = 0.031), a smaller overrepresentation of honorees with European names (β_{European} = 0.046, p = 0.042) and substantial underrepresentation of honorees with East Asian names ($\beta_{\text{East Asian}}$ = -0.010, p = 3.6×10^{-5}). No statistically significant difference was observed between the proportion of honorees and authors with names in Other categories (see Table 1, p = 0.95). Please see analysis notebook for more details.

Overrepresentation of US-affiliated honorees

We analyzed the countries of affiliation between last authors and ISCB honorees. For each country, we report a value of $\log_2$ enrichment (LOE) and its 95% confidence intervals. The full table with all countries and their corresponding enrichment can be browsed interactively in the corresponding analysis notebook. A positive value of LOE indicates a higher proportion of honorees affiliated with that country

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

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 (153 speakers more than expected, LOE = 1.1, 95% CI (0.9, 1.2)) and Israel (14 speakers more than expected, LOR = 2.5, 95% CI (1.7, 3.1)), and an underrepresentation of honorees affiliated with those in China, France, Italy, India, the Netherlands, Korea, Brazil and Taiwan (Fig. 3).

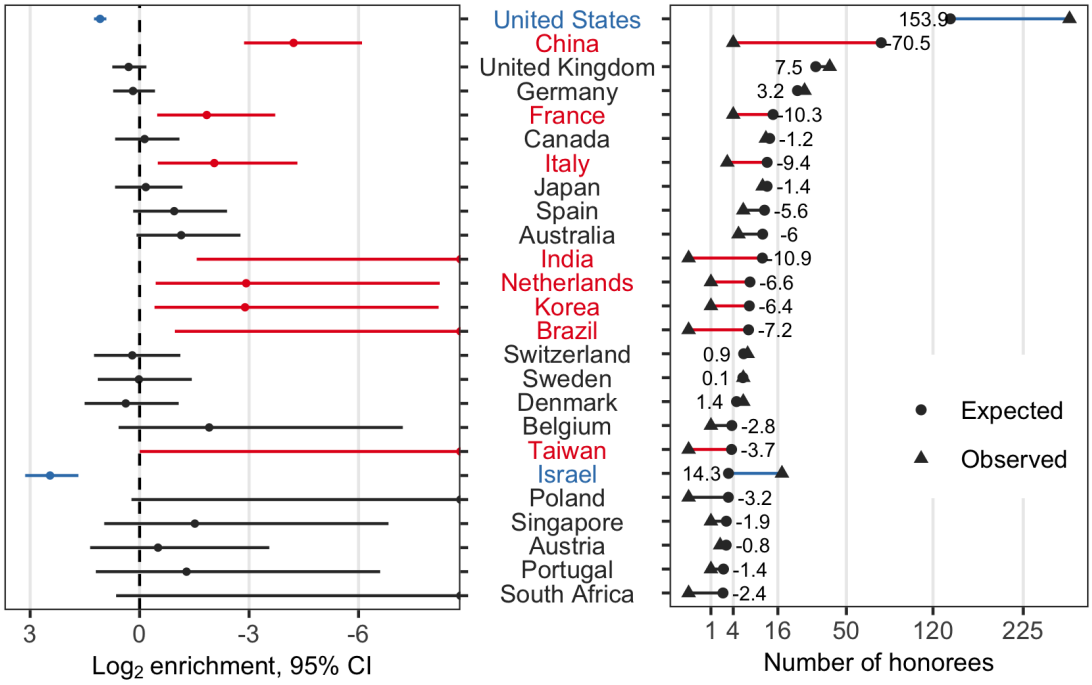

Figure 3: Each country's $\log_2$ enrichment (LOE) and its 95% confidence interval (left), and the absolute difference between observed (triangle) and expected (circle) number of honorees (right). Positive value of LOE indicates a higher proportion of honorees affiliated with that country compared to authors. Countries are ordered based on the proportion of authors in the field. 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, India, the Netherlands, Korea, Brazil and Taiwan.

Overrepresentation of white and underrepresentation of Asian honorees among US-affiliated scientists

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 last authors of computational biology papers whose last names associated with selecting Asian as a race/ethnicity category in the US census (Fig. 4A left). Compared to Pubmed authors, ISCB honorees with US affiliations have a higher proportion of individuals whose last names were associated with selecting white as a race/ethnicity category in the US census (Fig. 4A). Specifically, a name coming from the group of honorees has significantly higher probability of being white (β_{white} = 0.0975, $p = 1.4\times10^{-5}$) and lower probability of being Asian, (β_{Asian} = -0.1122, $p = 1.7\times10^{-6}$). The two groups of scientists did not have a significant association with names predicted to be in Other categories (p = 0.16662). Separating honoree results by honor category did not reveal any clear differences among the categories (see notebook.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

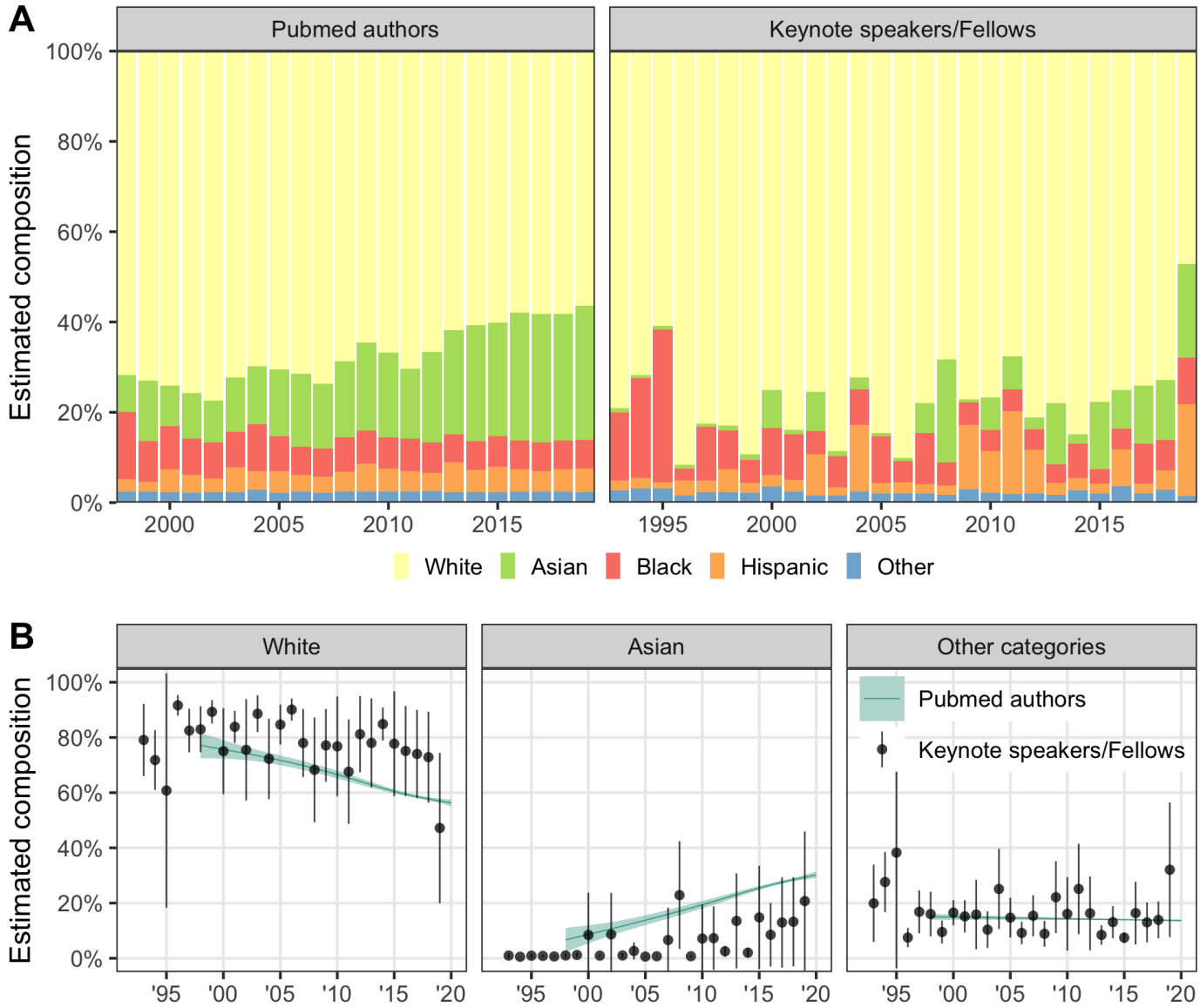

Figure 4: Estimated composition of census-based race/ethnicity prediction over the years of (A) all Pubmed computational biology and bioinformatics journal authors(left) and all ISCB Fellows and keynote speakers (right), (B) For each race/ethnicity category, the mean predicted probability of Pubmed articles is shown as teal generalized additive model (GAM) curve, and the mean probability and 95% confidence interval of the ISCB honoree predictions are shown as dark circles and vertical lines. The large difference between the means and the lack of overlaping of the confidence intervals in most years indicates an overrepresentation of white and underrepresentation of Asian honorees as compared to authors.

Improvements to Honoree Diversity Subsequent to Our Primary Analysis

While our study was primarily designed to assess the diversity of honorees, the findings raise important questions about what can be done to address the disparities. By publishing our original findings on the biorxiv preprint server, we could begin to answer these questions by examining changes subsequent to our initial report. We released version 1.0 of our manuscript on 2020-01-30. Early indications suggest ISCB is now 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. These new results suggest: 1) deserving honorees who were members of under-recognized groups existed but had not been recognized until now, and 2) simply examining honoree distribution's alignment with the field can trigger changes that address issues of unequal representation. However, we note that this analysis dealt only with more senior scientists (the last authors on scientific manuscripts) in the context of honors and that many years of changed honoree

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

distributions will be required for the set of honored scientists to better reflect the field's senior author contributions.

Discussion

A major challenge that we faced in carrying out this work was to narrow down geographic origins for some groups of names. Some name origin groups, such as Hispanic names, are geographically disparate. We were unable to construct a classifier that could distinguish between names from Iberian countries (Spain and Portugal) from those in Latin America in the group of Hispanic names. Discrepancies in representation between these groups are thus undetectable by our classifier. Honoree counts of those with Hispanic names are influenced from Spain as well as Latin America. In such cases, our analyses may substantially understate the extent to which minoritized scientists are underrepresented among honorees and authors.

Biases in authorship practices may also result in our underestimation of the composition of minoritized scientists within the field. We estimated the composition of the field using corresponding author status, but in neuroscience [13] and other disciplines [14] women are underrepresented among such authors. Such an effect would cause us to underestimate the number of women in the field. Though this effect has been studied with respect to gender, we are not aware of similar work examining race, ethnicity, or name origins.

We acknowledged that our supervised learning approaches are neither error free nor bias free. Because wru was trained on the US census to make predictions, many of the missing predictions are on names not observed in the US census. Although the underestimation of the proportion of these names could not be quantified in the list of authors and honorees, we complemented this race/ethnicity prediction method with an additional name origin analysis. By integrating different methods and preserving uncertainty by analyzing prediction probabilities rather than applying a hard assignment for each prediction, we hope to alleviate these biases and discover insightful findings that correctly reflect the current representation diversity at conferences.

Focusing on an international society and meetings, we measured honor and authorship rates worldwide. In this setting, we observe disparities by name groups. Because invitation and honor patterns could be driven by biases associated with name groups, geography, or other factors, we cross-referenced name group predictions with author affiliations to disentangle the relationship between geographic regions, name groups and invitation probabilities. We found that disparities persisted even within the group of honorees with a US affiliation.

An important questions to ask when measuring representation is what the right level of representation is. We suggest that considering equity may be more appropriate than strictly diversity. For example, we found similar representation of women between authors and honorees, which suggests honoree diversity is similar to that of authors. However, if fewer women are in the field because of systemic factors that inhibit their participation, we would not have reached equity. In addition to holding fewer corresponding authorship positions, on average, female scientists of different disciplines are cited less often [15], invited by journals to submit papers less often [14], suggested as reviewers less often [18], and receive significantly worse review scores [16]. Meanwhile, a review of women's underrepresentation in math-intensive fields argued that today's underrepresentation is not explained by historic forms of discrimination but factors surrounding fertility decisions and lifestyle choices, whether freely made or constrained by biology and society [19]. A recent analysis of gender inequality across different disciplines showed that, although both gender groups have equivalent annual productivity, women scientists have higher dropout rates throughout their scientific careers [20].

We found that ISCB's honorees and keynote speakers, though not yet reaching gender parity, appear to have similar gender proportion to the field as a whole. On the other hand, honorees include significantly

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

fewer people of color than the field as a whole, and Asian scientists are dramatically underrepresented among honorees. Societies' honoree selection process failing to reflect the diversity of the field can play a part in why minoritized scientists' innovations are discounted [1]. Although 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.

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 can boost young students' interests in STEM [19] and, for example, lead to higher persistence for undergraduate women in geoscience [21]. Efforts are underway to create Wikipedia entries for more female [22] and black, Asian, and minority scientists [23], which can help early-career scientists identify role models. Societies can contribute toward equity if they design policies to honor scientists in ways that counter these biases such as ensuring diversity in the selection committees.

The central role that scientists play in evaluating each other and each other's findings makes equity critical. Even many nominally objective methods of assessing excellence (e.g., h-index, grant funding obtained, number of high-impact peer-reviewed publications, and 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 [11]. 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, contributions from scientists of color are underrecognized.

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 AAAI 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 [25]. 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 [26], 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 [27] 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

(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.

Last author extraction

We assumed that, in the list of authors for a specific paper, last authors (often research advisors) would be most likely to be invited for keynotes or to be honored as Fellows. Therefore, we utilized PubMed to retrieve last author names to assess the composition of the field. PubMed is a search engine resource 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), within which we were able to extract author first and last names and their order using the E-Utilities APIs. To automate and generalize these tasks, we created the pubmedpy Python package.

From PubMed, we compiled a catalog of 174,397 journal articles that were published from 1993 through 2019, written in English, and tagged with the MeSH term "computational biology", which is equivalent to "bioinformatics" and includes categories such as genomics and systems biology (via PubMed's term explosion to include subterms). Excluding 163 articles with no author information and years with less than 200 articles/year, we analyzed 173,735 articles from 1998–2019.

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.

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. To avoid mislabeling, we only mapped the affiliation to a country if geotext identified 2 or more mentions of that country. For example, in the affiliation string "Laboratory of Computational and Quantitative Biology, Paris, France", geotext detected 2 mentions of places in France: Paris, France. In this case, we assign France to this affiliation.

This country extraction method accommodates multiple countries. Although ideally each affiliation record would refer to one and only one research organization, sometimes journals deposit multiple affiliations in a single structured affiliation. In these cases, we assigned multiple countries to the article. For more details on this approach, please consult the accompanying notebook and label dataset.

When geotext did not return results, we use the geopy approach, which returns a single country for an affiliation when successful. Its `geocoders.Nominatim` function converts names / addresses to geographic coordinates using the OpenStreetMap's Nomatim service. With this method, we split a textual affiliation by punctuation into a list of strings and iterate backward through this list until we found a Nomatim search result. For the above affiliation, the search order would be "France", "Paris", and "Laboratory of Computational and Quantitative Biology". Since Nomatim would return a match for the first term "France" (matched to France), the search would stop before getting to "Paris", and "France" would be assigned to this affiliation.

Our ability to assign countries to authors was largely driven by the availability of affiliations. The country-assignment-rate for last authors from PubMed records was approximately 47%. This reflects the varying availability of affiliation metadata by journal.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

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 [28]. 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 one name. Of 173,735 last authors, 1,002 were missing a fore name in the raw paper metadata and 11,427 had fore names consisting of only initials. Specifically, the metadata for most papers before 2002 (2,566 out of 2,601 papers) only have initials for first and/or middle author names. Without gender predictions for these names, we consider only articles from 2002 on when comparing gender compositions between two groups. Of the remaining authors, genderize.io failed to predict gender for 9,845 of these fore names. We note that approximately 42% of these NA predictions are hyphenated names, which is likely because they are more unique and thus are more difficult to find predictions for. 82% 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 [29] 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 [30] produces reasonable predictions, but its international representation in the data curated from Wikipedia in 2009 [31] 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 looked for nationality mention in the first sentence in the body of the text. In 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:

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

> ... 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.

Second, if this information is not available in the first sentence of the main text, 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.

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 [32]. 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 5. NamePrism excluded the US, Canada and Australia because these countries have been populated by a mix of immigrant groups [32].

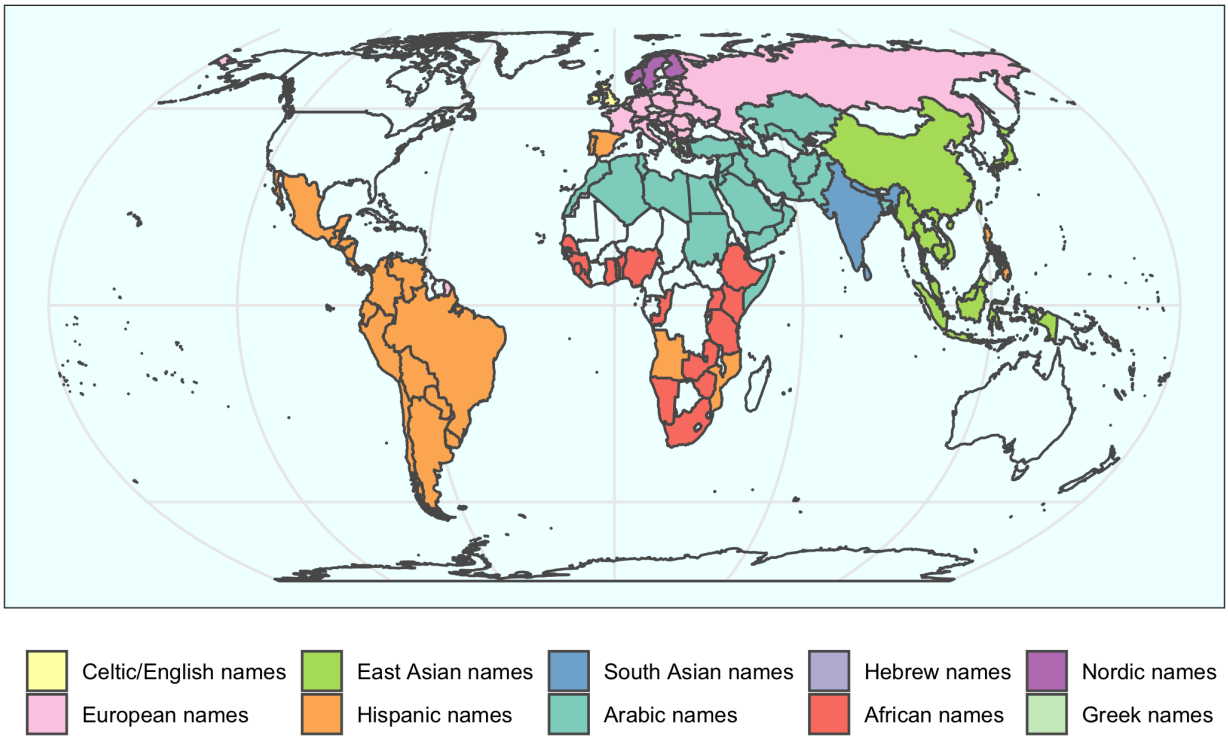

Figure 5: NamePrism groups countries by name similarity. We used this grouping but renamed the groups to focus on the linguistic patterns based on name etymology identified by NamePrism.

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. 6C) 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 readibility.

Predicting Name Origin Groups with LSTM Neural Networks and Wikipedia

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

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 Mgodi, Magambo Phillip Kimuda, Probhonjon Baruah, Adaoha E C Ihekwaba
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. 6A) 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. 6B). 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. 6C).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

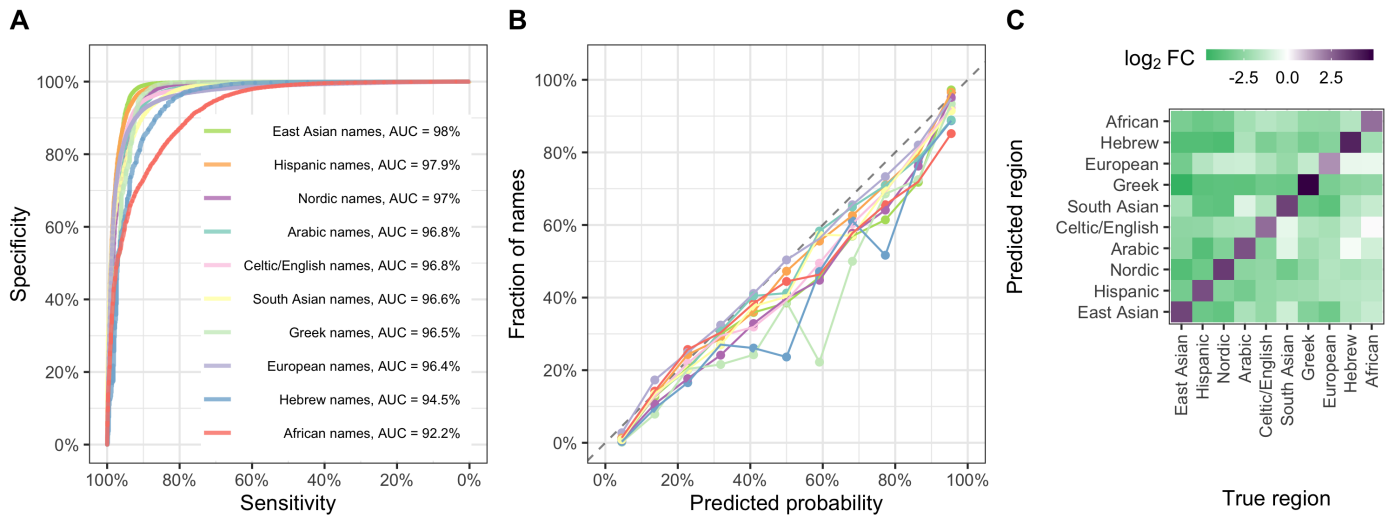

Figure 6: The Wiki2019-LSTM model performs well on the testing dataset. The area under the ROC curve is above 92% for each category, showing strong performance across origin categories (A). A calibration curve, computed with the caret R package, shows consistency between the predicted probabilities (midpoints 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.

Applying Wiki2019-LSTM on the author and honoree datasets, we obtained name origin estimates for all honorees' and authors' name, except the 12,429 that did not have fore names (see breakdown in the Estimation of Gender section above). Once again, because the large majority of author fore names prior to 2002 were recorded with initials only, predictions were not possible, and we excluded 1998–2001 when comparing name origin compositions between two groups.

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 last 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 [33].

Estimation of Race and Ethnicity within the US

The underlying data used for race and ethnicity 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 [34], and ethnicity categories include Hispanic/Latino or Not Hispanic/Latino [35]. We made race and ethnicity predictions of the surnames of US-affiliated authors and honorees using the R package wru, which implements methods described in Imai and Khanna [36]. wru uses similar race/ethnicity categories as in the Census but groups American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander to form the Other category. In the case of names that were not observed in the census, wru's *predict_race()* 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

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 26,771 US-affiliated last authors, 5,020 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.

Statistical analysis

We estimated the levels of representation by performing the following regression of the prediction probability on the group of scientists while controlling for year:

$$prob = \beta_0 + \beta_1 g + \beta_2 y + \epsilon.$$

The dependent variable $prob$ is the prediction probablity of a demographic variable (gender, race/ethnicity, and name origin) for names of scientists in g (honoree or author) during year y. $\epsilon \sim N(0, \sigma^2)$ accounts for random variation.

Data and Resource Availability

This manuscript was written openly on GitHub using Manubot [37]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

References

1. **The Diversity–Innovation Paradox in Science**
   Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, Daniel A. McFarland
   Proceedings of the National Academy of Sciences (2020-04-28) https://doi.org/ggskr7
   DOI: 10.1073/pnas.1915378117 · PMID: 32291335 · PMCID: PMC7196824

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 · PMID: 25411977 · PMCID: 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

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 · PMID: 28418385 · PMCID: 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 · PMID: 24399856 · PMCID: 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 · PMID: 30977853 · PMCID: 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

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 · PMID: 21852498 · PMCID: 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 · PMID: 27306969 · PMCID: PMC4965301

11. **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 · PMID: 31633016 · PMCID: PMC6785250

12. **Analysis of ISCB honorees and keynotes reveals disparities**
    Trang T. Le, Daniel S. Himmelstein, Ariel A. Hippen Anderson, Matthew R. Gazzara, Casey S. Greene
    bioRxiv (2020-04-14) https://www.biorxiv.org/content/10.1101/2020.04.14.927251v1
    DOI: 10.1101/2020.04.14.927251

13. **Persistent Underrepresentation of Women's Science in High Profile Journals**
    Yiqin Alicia Shen, Jason M. Webster, Yuichi Shoda, Ione Fine

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

bioRxiv (2018-03-08) https://doi.org/cmh5
DOI: 10.1101/275362

14. **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 · PMID: 29672508 · PMCID: PMC5908072

15. **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 (2020-08-28) https://arxiv.org/abs/2001.01002
DOI: 10.1038/s41593-020-0658-y

16. **Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution**
Charles W. Fox, C. E. Timothy Paine
Ecology and Evolution (2019-03-04) https://doi.org/gfwjjb
DOI: 10.1002/ece3.4993 · PMID: 30962913 · PMCID: PMC6434606

17. **Quantitative evaluation of gender bias in astronomical publications from citation counts**
Neven Caplar, Sandro Tacchella, Simon Birrer
Nature Astronomy (2017-05-26) https://doi.org/b8rn
DOI: 10.1038/s41550-017-0141

18. **Journals invite too few women to referee**
Jory Lerback, Brooks Hanson
Nature (2017-01-26) https://doi.org/gf4jjz
DOI: 10.1038/541455a · PMID: 28128272

19. **Understanding current causes of women's underrepresentation in science**
S. J. Ceci, W. M. Williams
Proceedings of the National Academy of Sciences (2011-02-07) https://doi.org/dsrczk
DOI: 10.1073/pnas.1014871108 · PMID: 21300892 · PMCID: PMC3044353

20. **Historical comparison of gender inequality in scientific careers across countries and disciplines**
Junming Huang, Alexander J. Gates, Roberta Sinatra, Albert-László Barabási
Proceedings of the National Academy of Sciences (2020-03-03) https://doi.org/ggk89f
DOI: 10.1073/pnas.1914221117 · PMID: 32071248 · PMCID: PMC7060730

21. **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

22. **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

23. **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

24. **Author-Reviewer Homophily in Peer Review**
Dakota Murray, Kyle Siler, Vincent Larivière, Wei Mun Chan, Andrew M. Collings, Jennifer Raymond, Cassidy R. Sugimoto
bioRxiv (2019-08-04) https://doi.org/ddzr
DOI: 10.1101/400515

25. **Research in Computational Molecular Biology**
Lecture Notes in Computer Science
(2016) https://doi.org/gghp87
DOI: 10.1007/978-3-319-31957-5

26. **Research in Computational Molecular Biology**
Lecture Notes in Computer Science
(2011) https://doi.org/dvwjgg
DOI: 10.1007/978-3-642-20036-6

27. **Research in Computational Molecular Biology**
Lecture Notes in Computer Science

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.14.927251; this version posted September 22, 2020. 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.

(2015) https://doi.org/gghwjh
DOI: 10.1007/978-3-319-16706-0

28. **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

29. **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

30. **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

31. **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

32. **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

33. **Statistics in epidemiology: methods, techniques, and applications**
Hardeo Sahai, Anwer Khurshid
CRC Press (1996)
ISBN: 9780849394447

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

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

36. **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

37. **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 · PMID: 31233491 · PMCID: PMC6611653