Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades

Yen Ji Julia Byeon,[1,3] Rezarta Islamaj,[2] Lana Yeganova,[2] W. John Wilbur,[2] Zhiyong Lu,[2] Lawrence C. Brody,[4,*] and Vence L. Bonham[3,*]

Summary

To inform continuous and rigorous reflection about the description of human populations in genomics research, this study investigates the historical and contemporary use of the terms "ancestry," "ethnicity," "race," and other population labels in *The American Journal of Human Genetics* from 1949 to 2018. We characterize these terms' frequency of use and assess their odds of co-occurrence with a set of social and genetic topical terms. Throughout *The Journal*'s 70-year history, "ancestry" and "ethnicity" have increased in use, appearing in 33% and 26% of articles in 2009–2018, while the use of "race" has decreased, occurring in 4% of articles in 2009–2018. Although its overall use has declined, the odds of "race" appearing in the presence of "ethnicity" has increased relative to the odds of occurring in its absence. Forms of population descriptors "Caucasian" and "Negro" have largely disappeared from *The Journal* (<1% of articles in 2009–2018). Conversely, the continental labels "African," "Asian," and "European" have increased in use and appear in 18%, 14%, and 42% of articles from 2009–2018, respectively. Decreasing uses of the terms "race," "Caucasian," and "Negro" are indicative of a transition away from the field's history of explicitly biological race science; at the same time, the increasing use of "ancestry," "ethnicity," and continental labels should serve to motivate ongoing reflection as the terminology used to describe genetic variation continues to evolve.

Introduction

The field of human genetics has struggled since its inception with the task of conceptualizing and describing geographic and population-based genetic variation. First thought of as hierarchical and unequal taxonomic types, then reframed as isolates that differ in allele frequency,[1] and now in terms of genetic ancestry,[2] the idea of the "population" in human genetics has continuously evolved since the field's earliest decades. Today, advances in genomics continue to spur discussions about how the field can accurately describe human genetic diversity.[3] Central to these discussions is how it will reconcile its legacy of scientific racism.[4] We use this phrase to refer both to the historical practice of studying races as distinct biological groups and more broadly to the incorrect conceptualization of racial difference as biological in ways that contribute to social stratification and inequity.

Today, three concepts take center stage in these discussions, each of which brings its own challenges: ancestry, ethnicity, and race. Racial and ethnic group membership is used as a covariate in genomic studies to account for confounding related to genetic ancestry or social determinants of health. For example, geneticists may address confounding due to genetic ancestry by stratifying analyses by racial or ethnic categories or improve power to detect genetic associations by including a race or ethnicity variable that accounts for variation due to social stratification.[5] Although the field has made progress in rejecting the idea of racial and ethnic categories as discrete biological units, the continuing use of race and ethnicity as proxies for genetic ancestry remains scientifically and socially problematic.[6] Ancestry, more specifically, genetic ancestry, has been described as information about the ancestors or populations from whom one has inherited genetic material.[7] Although ancestry may lend itself to a quantitative description of human genetic variation, a unified definition of this concept has yet to be developed, and even a precise definition of the "populations" from whom one has inherited genetic material remains elusive.[7,8]

Given the complexity of these concepts and their underlying histories, there is a lack of consensus in the field on how ancestry, ethnicity, and race should be understood. This is reflected in the increasingly heterogeneous ways that the concepts are employed in clinical research and practice.[9,10] Members of the genetics community have called for consensus on how these data should and should not be used[6] as well as called on the National Institutes of Health to support the National Academy of the Sciences, Engineering, and Medicine in developing a consensus statement on best practices for characterizing human genetic diversity in research.[11,12] Others have proposed standardized systems for annotating populations[13] and expressed optimism that advances in genetic technologies may allow the field to move past the use of race and ethnicity.[3,14]

An important component of ongoing efforts to establish consensus in this area of human genetics is knowledge about the social and historical paths through which the

[1]Department of Sociology, Princeton University, Princeton, NJ 08544, USA; [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; [3]Social and Behavioral Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; [4]Division of Genomics and Society, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: bonhamv@mail.nih.gov (V.L.B.), lbrody@mail.nih.gov (L.C.B.)
https://doi.org/10.1016/j.ajhg.2021.10.008.

Table 1. Percentage of 200-token segments and articles containing terms used for analysis, 1949–2018

Term	Articles, n (%)	Even segments, n (%)
Admixture, admixtures, admix, admixed	981 (8.5)	2,969 (2.5)
African, Africans (excluding African American)	1,116 (9.6)	3,398 (2.9)
Allele, alleles, allelic	7,984 (68.9)	37,219 (31.5)
Ancestry, ancestries, ancestral, ancestrally	2,351 (20.3)	6,799 (5.8)
Asian, Asians (excluding Asian American)	950 (8.2)	2,755 (2.3)
Behavior, behaviors, behavioral	1,608 (13.9)	2,928 (2.5)
Caucasian, Caucasians, Caucasoid, Caucasoids	1,391 (12.0)	3,381 (2.9)
Diversity, diverse	2,114 (18.2)	4,527 (3.8)
Environment, environments, environmental, environmentally	2,843 (24.5)	5,966 (5.1)
Ethnicity, ethnicities, ethnic, ethnically	2,208 (19.1)	4,344 (3.7)
European, Europeans (excluding European American)	2,637 (22.8)	6,545 (5.5)
Frequency, frequencies	7,769 (67.0)	28,741 (24.2)
Geography, geographies, geographic, geographically	1,131 (9.8)	2,438 (2.1)
Haplotype, haplotypes, haplotypic	3,720 (32.1)	15,489 (13.1)
Hispanic, Hispanics	397 (3.4)	883 (0.7)
Language, languages, linguistic, linguistically	870 (7.5)	1,992 (1.7)
Latino, Latinos, Latina, Latinas, Latinx	67 (0.6)	181 (0.2)
Linkage, linkages	5,605 (48.4)	21,950 (18.6)
Locus, loci	7,111 (61.4)	31,446 (26.7)
Negro, Negroes, Negroid, Negroids	373 (3.2)	1,121 (1.0)
Population, populations	7,572 (65.3)	31,899 (27.0)
Race, races, racial, racially	852 (7.4)	1,691 (1.6)
Religion, religions, religious, religiously	247 (2.1)	386 (0.3)
Social, socially	1,038 (9.0)	1,898 (1.6)
Socioeconomic, socioeconomically	215 (1.9)	353 (0.3)
Total	11,590 (100.0)	117,986 (100.0)

field has come to its current understanding of ancestry, ethnicity, and race. To this end, we investigated how the frequency of the terms "ancestry," "ethnicity," "race," and other population labels have changed over the 70-year publication history of *The American Journal of Human Genetics* (1949–2018). Additionally, in order to assess the evolving context in which the three concepts were used, we tested for non-random term co-occurrences between "ancestry," "ethnicity," and "race" and a predetermined set of social, genetic, and population terms from 1949 to 2018. In doing so, we aim to push for continuous and rigorous reflection surrounding the use of these population concepts in human genetics.

Material and methods

Data

We obtained digital versions of the full text of every document published in *The Journal* from its founding in 1949 up to 2018. These files were held by the National Library of Medicine (NLM) at the National Institutes of Health and obtained for purposes of research with permission from the American Society of Human Genetics. We sought matches for all articles in this archive to a PMID in MedLine and/or a PMCID in PubMed Central. Articles without a PMCID or PMID, which comprised book reviews, abstract books, disciplinary announcements, indexes, and tables of contents, were not included in the dataset. The majority of the remaining articles were scientific research articles (11,360), and a small minority (275) were award speeches and other communications. Of the 11,635 articles included in analysis, 6,750 were in the form of extracted text from optical character recognition (OCR) versions of scanned journal pages. For the remaining 4,885 articles, the full text was readily available in XML format.

After removing the references sections of articles, all text was converted to the ASCII character set. For text obtained from OCR versions of PDF files, words broken over line breaks were repaired as described in the supplemental materials and methods. Punctuation, numerical tokens, and single-character terms were removed. Finally, we ensured that any occurrence of the term

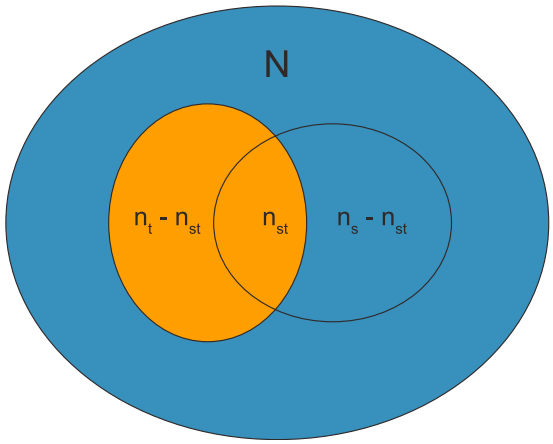

Figure 1. Visualization of parameters of random variable *y*, n_t, n_{st}, and n_s

Visualization of parameters of random variable *y*, n_t, n_{st}, and n_s, where *t* is the target term, *s* is the co-occurring word of interest, and *N* is the total number of 200-token segments in a given range of years. n_t represents number of segments containing *t*, n_s represents number of segments containing *s*, and n_{st} represents number of segments containing both.

"race" in the dataset referred solely to the population concept, as the term could also be a part of an author name (e.g., Robert Race), an abbreviation for a molecular biology method (rapid amplification of cDNA ends), or the word in the sense of a competition (e.g., "race to the finish line"). Informed by term associations, manual review, and orthographic characteristics of the term "race," we converted "race" to "xace" wherever it was not used in the population sense.

Selection of terms for analysis

We preselected 25 terms for which to calculate frequencies of use and odds of co-occurrence. In addition to "ancestry," "ethnicity," and "race," we examined 15 topical terms that have been related to these concepts ("admixture," "allele," "behavior," "diversity," "environment," "frequency," "geography," "haplotype," "language," "linkage," "locus," "population," "religion," "social," "socioeconomic") and seven population descriptors ("African," "Asian," "Caucasian," "European," "Hispanic," "Latina/o/x," "Negro"). The population descriptors "African," "Asian," and "European" refer to these specific forms of the descriptors and exclude uses of "African American," "Asian American," and "European American." Terms were selected for their relevance to ancestry, ethnicity, and race as well as specificity of meaning (e.g., "culture" was not chosen because it could also refer to cell cultures; "Black" and "White" could refer to the colors, as in "the black arrows indicate…" or "white blood cell"). We expanded each selected term to include alternate forms of the word with the same stem; for example, instances of "ancestral," "ancestries," and "ancestrally" were all counted as uses of "ancestry." Table 1 lists the 25 terms, their alternate forms, and their frequencies.

In order to investigate the ideas associated with and relationships between "ancestry," "ethnicity," and "race" in *The Journal*, we determined and compared co-occurrence patterns between (1) pairs of "ancestry," "ethnicity," and "race" (i.e., "ancestry" + "ethnicity," "ancestry" + "race," "ethnicity" + "race"), (2) 15 topical terms and each of "ancestry," "ethnicity," and "race" (i.e., each topical term + "ancestry," each topical term + "ethnicity," each topical term + "race" for 45 comparisons total), and (3) seven population descriptors and each of "ancestry," "ethnicity," and "race" (i.e., each population descriptor + "ancestry," each population descriptor + "ethnicity," each population descriptor + "race" for 21 comparisons total).

Measuring term co-occurrence

Using co-occurrence as a measure of relatedness between a pair of words is guided by a fundamental distributional hypothesis in linguistics stating that the meaning of words is determined by the contexts in which they occur or the words with which they occur.[15] This hypothesis informs statistical methods such as language modeling,[16] word embeddings,[17,18] and word similarity measures.[19,20] Given a pair of words, we analyze whether they co-occur more than expected by chance and interpret this as evidence that they have a semantic relationship.

Co-occurrence refers to how often a pair of words appear together in the same texts or documents. Ideally, these documents would all be of the same length, as longer documents are *a priori* more likely to contain a given word than shorter documents. To eliminate this bias, we split documents into disjoint text windows or segments of 200 words (space separated tokens) and defined the co-occurrence between two terms as the number of text windows in which both terms occur.[20] An advantage of this partitioning is that the relatively small size of the segments implies a closer relationship between words that co-occur. Using smaller pieces of text also increases the sensitivity of statistical testing to determine whether terms co-occur at a higher frequency than predicted by random mixing. Although paragraphs could also be used as text segments, a substantial proportion of the text we analyzed was obtained by OCR applied to scanned text, making it difficult to identify paragraph boundaries.

When partitioning documents into 200-token segments, it is possible that our two terms of interest become separated by the border between adjacent segments. To account for terms that are in close proximity but separated by the border of adjacent segments, we started a new segment after every 100 tokens, such that each had a 100-token overlap with adjacent segments. To eliminate double-counting of co-occurrences caused by the overlaps, we numbered the segments and computed results separately for even- and odd-numbered segments. We report results by using even segments. Results computed from odd segments were not substantially different and are reported in Table S3. Yearly counts of articles and segments are shown in Figure S1.

Figure 1 illustrates our method for assessing whether two terms co-occurred more often than expected by chance in a given set of segments. In our analyses, we considered either the even or odd segments from 10-year intervals at a time, incrementing the decades by one year in order to identify temporal trends (e.g., 1949–1958, 1950–1959, 1951–1960, …, 2008–2017, 2009–2018). The outside blue oval represents the set of all 200-token segments in a decade. Term *t* splits the space into two subsets of segments: those that contain the term (orange oval) and the rest. Further, consider term *s* and assume that it co-occurs with *t* in n_{st} segments. The p value is the probability that the observed or greater overlap between the two terms would happen by chance as determined by the size of the space of segments and the number of segments containing *s* and the number containing *t*. Mathematically, a random variable *y* representing the overlap between the target term *t* and

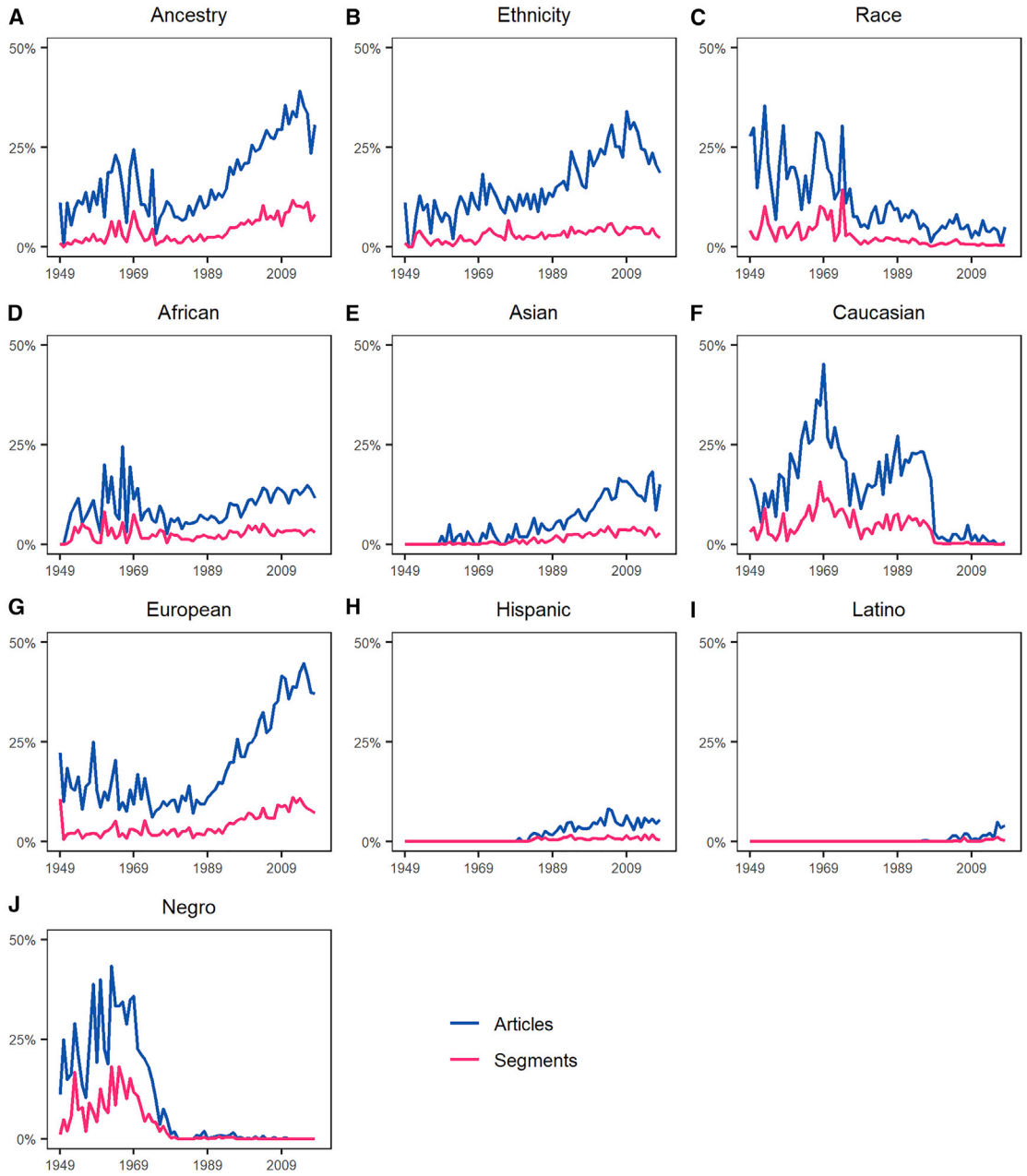

Figure 2. Percentage of 200-token segments containing "ancestry," "ethnicity," "race," and other population descriptors
(A–J) Yearly percentage of 200-token segments (pink) and articles (blue) containing "ancestry" (A), "ethnicity" (B), "race" (C), "African" (D), "Asian" (E), "Caucasian" (F), "European" (G), "Hispanic" (H), "Latina/o/x" (I), and "Negro" (J) from 1949–2018.

term s is a hypergeometric random variable with parameters n_s, n_t, n_{st} and the probability function:[21]

$$P(y) = \binom{n_t}{y}\binom{N-n_t}{n_s-y}\Big/\binom{N}{n_s}$$

The mean of this distribution is the expected co-occurrence on a random basis and is given by $n_s n_t/N$. We compute the p value, i.e., the probability of the observed or a greater co-occurrence frequency arising by chance as the following:

$$p\ value = \sum_{y=n_{st}}^{\min(n_s,\, n_t)} P(y)$$

This calculation is equivalent to representing the presence or absence of two terms in a two-by-two contingency table and conducting a one-sided Fisher's exact test.[19]

We measured the "effect size" as the odds ratio (OR)[22]— the ratio of the odds of term t occurring in a segment where term s is present—to its odds of being present in the absence of s. The OR is given by $\frac{n_{st}(N-n_s-n_t+n_{st})}{(n_s-n_{st})(n_t-n_{st})}$, and N represents the total number of even or odd number of segments in a given decade and n the number of segments with s, t, or both (Figure 1). 95% confidence intervals (CIs) for the ORs are given by a well-known formula.[22] As p values and CIs were calculated with different distributions, ORs whose CIs include "1" can be statistically significant.

Table 2. Percentage of 200-token segments and articles containing population terms in 1949–58 and 2009–18

Term	Decade	Segments, % (n)	Articles, % (n)
Ancestry	1949–58	1% (48)	10% (31)
	2009–18	9% (2,585)	33% (721)
Ethnicity	1949–58	2% (56)	8% (25)
	2009–18	4% (1,150)	26% (571)
Race	1949–58	5% (149)	22% (69)
	2009–18	<1% (151)	4% (89)
African	1949–58	2% (74)	7% (23)
	2009–18	3% (905)	13% (288)
Asian	1949–58	0% (0)	0% (0)
	2009–18	3% (935)	14% (310)
Caucasian	1949–58	4% (131)	12% (38)
	2009–18	<1% (25)	<1% (22)
European	1949–58	2% (76)	15% (48)
	2009–18	9% (2,500)	40% (881)
Hispanic	1949–58	0% (0)	0% (0)
	2009–18	1% (287)	5% (112)
Latina/o/x	1949–58	0% (0)	0% (0)
	2009–18	<1% (128)	2% (45)
Negro	1949–58	7% (217)	21% (65)
	2009–18	<1% (2)	<1% (1)

Results

Frequency of use

The proportion of articles containing the population terms "ancestry," "ethnicity," "race," "African," "Asian," "Caucasian," "European," "Hispanic," "Latina/o/x," and "Negro" were calculated for each year from 1949–2018. The percent of articles that include "race" has declined since 1949 (Figure 2C), appearing in 22% of articles from 1949–58 and 5% in 2009–18 (Table 2). Conversely, the percent using "ancestry" and "ethnicity" have increased (Figures 2A and 2B). "Ancestry" increased in use from 10% of articles in 1949–58 to 33% in 2009–18. "Ethnicity" appeared in 8% of articles in 1949–58 and 26% in 2009–18 (Table 2). The continental terms "African," "Asian," and "European" have also increased in use, while the terms "Caucasian" and "Negro" have declined (Figures 2D–2J, Table 2). The proportion of articles containing the remaining 15 topical terms are shown in Figure S2. Yearly frequencies from 1949–2018, for both even and odd segments, are reported in Table S1.

Co-occurrence patterns

Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated for pairs of "ancestry," "ethnicity," and "race" for overlapping decades from 1949–2018. ORs have increased over time between "race" and "ethnicity," from 4.7 (CI 2.3, 9.5) in 1949–58 to 23.9 (CI 17.2, 33.1) in 2009–18. Ratios between "ancestry" and "race" and between "ancestry" and "ethnicity" have remained comparatively constant (Figure 3, Table 3). ORs were also generated between 15 topical terms and each of "ancestry," "ethnicity," and "race" (Figure 4, Figure S3) and between the seven population descriptors and each of "ancestry," "ethnicity," and "race" (Figure 5). All ORs, their 95% CIs, and p values for the observed co-occurrences between pairs of terms are given in Table S2 for even segments and Table S3 for odd segments.

Discussion

We have described the evolving usage of population terms in the 70-year publication history of *The American Journal of Human Genetics*. We find that from 1949–2018, the term "race" has declined in use, while increasing in co-occurrence with "ethnicity." At the same time, the use of "ancestry" and "ethnicity" has increased. We also describe changes in the use of specific population descriptors that may align with societal trends in their use outside of genetics.

The use of the term "race" in *The Journal* has consistently declined since 1949, while that of "ancestry" and

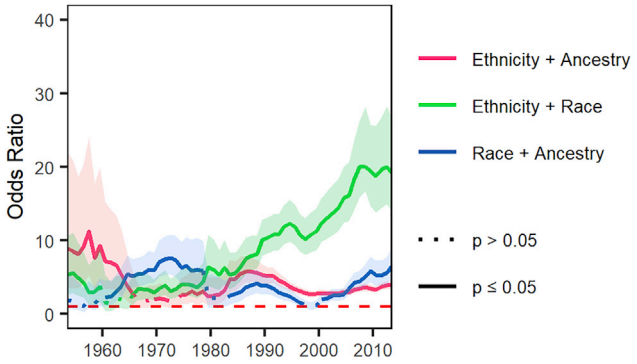

Figure 3. Odds ratios between "ancestry," "ethnicity," and "race"

ORs between "ethnicity" and "ancestry" (pink), "race" and "ancestry" (blue), and "race" and "ethnicity" (green) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (e.g., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

"ethnicity" have increased. These findings are consistent with those of Popejoy et al.'s survey of clinical geneticists in which participants reported ancestry, followed by ethnicity then race, as important to clinical variant interpretation and ordering genetic tests.[23] We hypothesize that as the field grows more cognizant of historical and ongoing debates about the use of race in genetics, ancestry and ethnicity may increasingly be perceived as more scientifically valid, historically neutral, or practically useful. This is not without its own criticisms, as we will discuss further below. We also found an increase in the odds ratio between "race" and "ethnicity" throughout the history of *The Journal*. This may be attributable to the increasing use of combined phrases such as "race/ethnicity" and "race and/or ethnicity," which have emerged as the distinction between the two concepts has become more ambiguous.

Furthermore, we report temporal changes in the use of specific population descriptors, adding support to the long-standing wisdom that population labels are not based on immutable biological order but shift in tandem with social context.[24] Along with the finding above that the use of "race" has declined, the labels "Caucasian"[25] and "Negro" have declined in *The Journal* over the past several decades. These terms, particularly in the form of "Caucasoid" and "Negroid," were used by 19th century race scientists and later by 20th century geneticists to refer to pseudoscientific biological race groups. "Hispanic" and "Latina/o/x" first appeared in *The Journal* in 1980 and 1996, respectively. Each of these changes in the use of population descriptors took place in a broader social context. For example, the decline of the term "Negro" can be connected not only to the discrediting of the idea of a "Negroid race" on scientific terms but also to African-descent Americans' efforts to reject or claim social identifiers in contexts outside of genetics.[26–28] Similarly, the adoption of "Hispanic" and "Latina/o/x" in genetics did not originate from within the field but from a convergence of commercial, activist, and government interests in creating a panethnic, institutionally recognized category from the diverse range of Latin American nationalities in the US.[29]

Some of the shifts described in this paper may signal constructive change. For example, the term "Caucasian," which has declined in use in *The Journal*, has been criticized for its historical connections to racist taxonomies and lack of scientific justification.[30] However, areas remain for continued investigation and critical reflection. For example, although the term "race" has declined, commentary in this area has pushed not necessarily for the complete removal of race from genetic and biomedical research but for a refocusing on racism and race as a social category with biological consequences.[4,11,12] Moreover, as numerous scholars have discussed, practices that racialize populations can persist in the sciences without explicit use of the term "race."[31–34] The continental population terms "African," "Asian," and "European," which we have shown are increasing in use in *The Journal*, have been critiqued for their resemblance to historical racial taxonomies and their inability to capture immense within-group heterogeneity.[8,35]

Table 3. Odds ratios between "ancestry," "ethnicity," and "race"

Term 1	Term 2	Decade	Odds ratio (95% CI)	p value
Ancestry	ethnicity	1949–58	8.2 (3.4, 20.1)	2.2×10^{-4}
		2009–18	3.9 (3.3, 4.4)	1.4×10^{-65}
Ancestry	race	1949–58	4.2 (2.0, 9.2)	1.5×10^{-3}
		2009–18	6.6 (4.8, 9.2)	1.3×10^{-23}
Ethnicity	race	1949–58	4.7 (2.3, 9.5)	2.0×10^{-4}
		2009–18	23.9 (17.2, 33.1)	1.3×10^{-60}

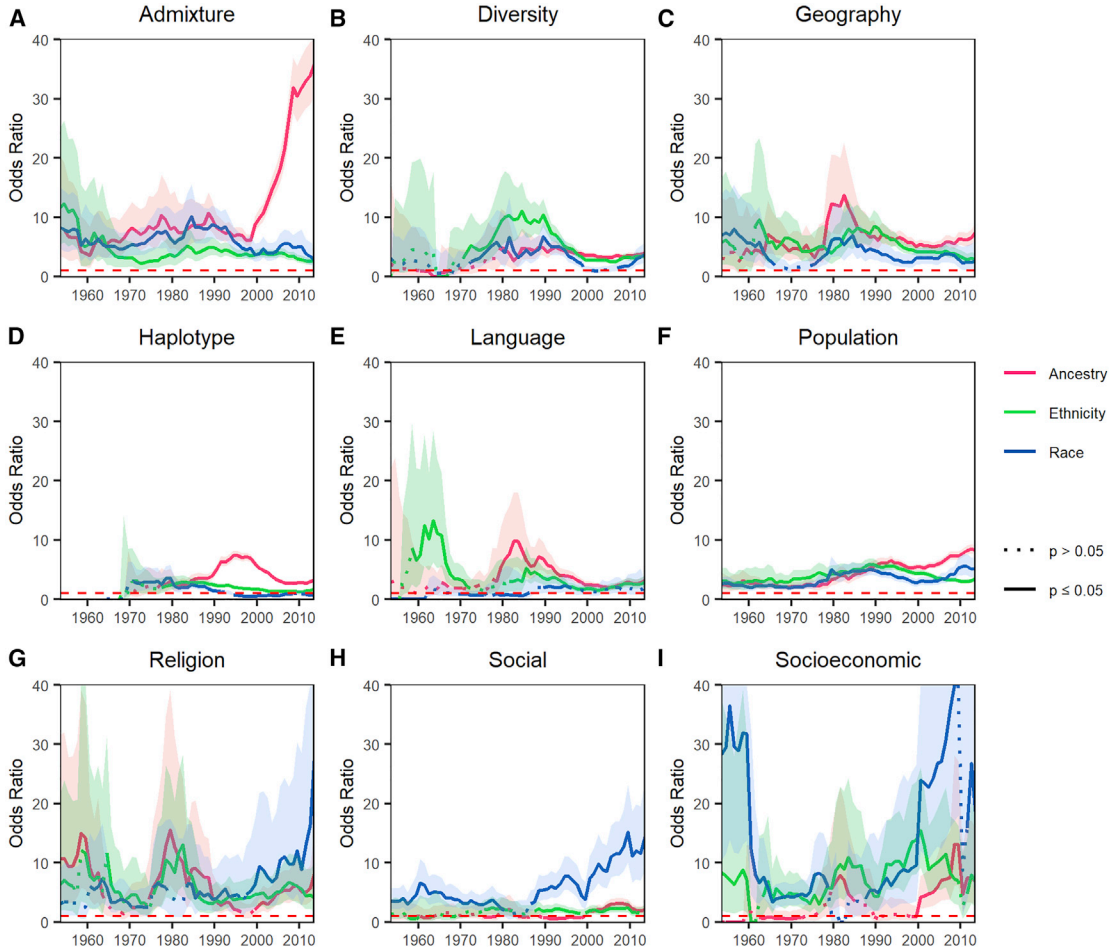

Figure 4. Odds ratios between "ancestry," "ethnicity," "race," and select topical terms
(A–I) ORs between "ancestry" (pink), "ethnicity" (green), and "race" (blue) and "admixture" (A), "diversity" (B), "geography" (C), "haplotype" (D), "language" (E), "population" (F), "religion" (G), "social" (H), and "socioeconomic" (I) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (e.g., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

This study has several limitations. First, we examined a single journal, and the trends we describe may not generalize to other contexts in the field. However, our analysis of the entire corpus of a single journal may be a strength relative to other studies of biomedical corpora, which tend to be limited to abstracts because of data availability. Second, we pre-selected a set of terms that we chose not to alter throughout the course of our analyses. As a result, we were limited in our ability to explore or discover new aspects of ancestry, ethnicity, and race that may deviate from our current biases about the concepts. We also could not examine many relevant descriptors such as "Black," "White," and "Native American," as these terms were either confounded by other meanings in the text or did not have high enough frequency in the dataset to conduct statistical analyses. Third, odds ratios were sensitive to the amount of data available, meaning that time periods with limited amounts of text or term uses were prone to large, not necessarily meaningful, fluctuations. Finally, although quantitative analyses of text are unique in their ability to detect patterns that are difficult through manual review, we recognize these methods' limited ability to provide insight into how our terms and concepts of interest were used qualitatively.

Nonetheless, our research has documented and quantitated historical changes in the use of population concepts in the entirety of *The Journal*'s text corpus. Our results can serve to motivate ongoing reflection as the concepts and population group labels used to study global genetic variation continues to evolve. Such reflection is critical to the field's ability to accurately describe human genetic variation and adopt new genomic methods in a way that is attentive to its troubled history with race.

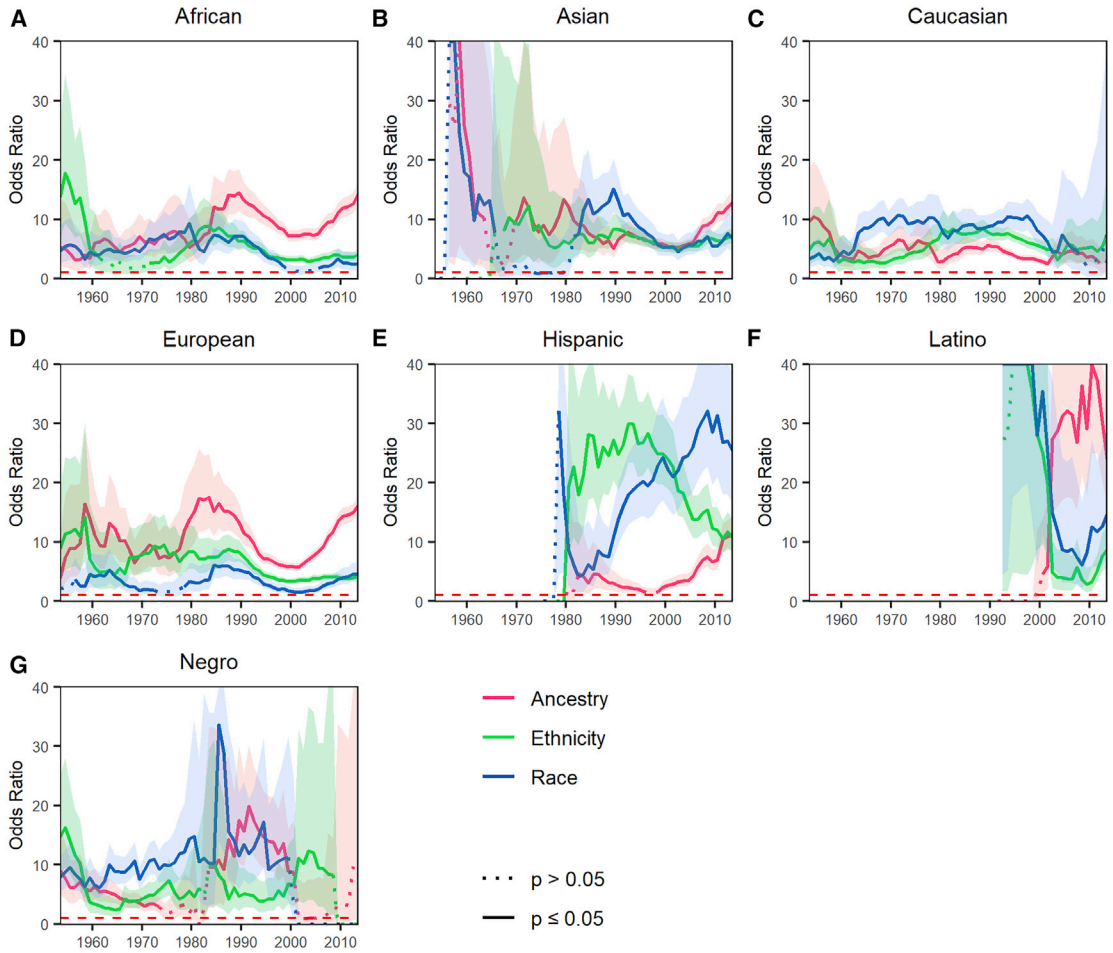

Figure 5. Odds ratios between "ancestry," "ethnicity," "race," and population descriptors

(A–G) ORs between "ancestry" (pink), "ethnicity" (green), and "race" (blue) and "African" (A), "Asian" (B), "Caucasian" (C), "European" (D), "Hispanic" (E), "Latina/o/x" (F), and "Negro" (G) in overlapping decades from 1949–2018 (e.g., 1949–59, 1950–60, 1951–61 …). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (i.e., values at 1954 represent co-occurrence ratios for 1949–59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence intervals. For ease of viewing, ORs and upper confidence intervals are cut if they exceeded 40 (see Tables S2 and S3 for untruncated values). The horizontal red dashed line marks an OR of 1.0.

Data and code availability

The code generated during this study and the data are available upon request. Please contact the corresponding authors for the data.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.10.008.

Acknowledgments

We would like to acknowledge the American Society of Human Genetics for sharing the *AJHG* archive for this research, Natalie Xie for contributions to data cleaning and the development of a graphical interface for data visualization, and Won Kim for assistance in data processing. This work was supported by the Intramural Research Programs of the National Center for Biotechnology Information and the National Human Genome Research Institute, National Institutes of Health.

Declaration of interests

The authors declare no competing interests.

Received: January 30, 2021
Accepted: October 20, 2021
Published: December 2, 2021

References

1. Veronika Lipphardt (2013). From "Races" to "Isolates" and "Endogamous Communities": Human Genetics and the Notion of Human Diversity in the 1950s. In Human Heredity

in the Twentieth Century, B. Gausemeier, S. Müller-Wille, and E. Ramsden, eds. (Pickering and Chatto), pp. 55–68.
2. Fujimura, J.H., and Rajagopalan, R. (2011). Different differences: the use of 'genetic ancestry' versus race in biomedical human genetic research. Soc. Stud. Sci. *41*, 5–30.
3. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692.
4. Brothers, K.B., Bennett, R.L., and Cho, M.K. (2021). Taking an antiracist posture in scientific publications in human genetics and genomics. Genet. Med. *23*, 1004–1007.
5. Khan, A., Mchugh, C., Conomos, M.P., Gogarten, S.M., and Nelson, S.C. (2020). Guidelines on the use and reporting of race, ethnicity, and ancestry in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. arxiv, 2108.07858. https://arxiv.org/ftp/arxiv/papers/2108/2108.07858.pdf.
6. Bonham, V.L., Green, E.D., and Pérez-Stable, E.J. (2018). Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. JAMA *320*, 1533–1534.
7. Mathieson, I., and Scally, A. (2020). What is ancestry? PLoS Genet. *16*, e1008624.
8. Weiss, K.M., and Long, J.C. (2009). Non-Darwinian estimation: my ancestors, my genes' ancestors. Genome Res. *19*, 703–710.
9. Popejoy, A.B., Ritter, D.I., Crooks, K., Currey, E., Fullerton, S.M., Hindorff, L.A., Koenig, B., Ramos, E.M., Sorokin, E.P., Wand, H., et al. (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum. Mutat. *39*, 1713–1720.
10. Panofsky, A., and Bliss, C. (2017). Ambiguity and Scientific Authority: Population Classification in Genomic Science. Am. Sociol. Rev. *82*, 59–87.
11. Yudell, M., Roberts, D., DeSalle, R., Tishkoff, S.; and 70 signatories (2020). NIH must confront the use of race in science. Science *369*, 1313–1314.
12. Yudell, M., Roberts, D., DeSalle, R., and Tishkoff, S. (2016). Taking race out of human genetics. Science *351*, 564–565.
13. Huddart, R., Fohner, A.E., Whirl-Carrillo, M., Wojcik, G.L., Gignoux, C.R., Popejoy, A.B., Bustamante, C.D., Altman, R.B., and Klein, T.E. (2019). Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. Clin. Pharmacol. Ther. *105*, 1256–1262.
14. Bonham, V.L., Callier, S.L., and Royal, C.D. (2016). Will Precision Medicine Move Us beyond Race? N. Engl. J. Med. *374*, 2003–2005.
15. Harris, Z.S. (1954). Distributional Structure. Distrib. Struct. WORD *10*, 146–162.
16. Schütze, H., Manning, C.D., and Raghavan, P. (2008). Introduction to information retrieval (Cambridge University Press Cambridge).
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Red Hook, NY, USA: Curran Associates Inc.), pp. 3111–3119.
18. Li, Y., and Yang, T. (2018). Word Embedding for Understanding Natural Language: A Survey (Cham: Springer), pp. 83–104.
19. Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics *19*, 61–74.
20. Terra, E., and Clarke, C.L.A. (2003). Frequency estimates for statistical word similarity measures (Association for Computational Linguistics), pp. 165–172.
21. Larson, H. (1982). Introduction to Probability Theory and Statistical Inference (New York: Wiley).
22. Tenny, S., and Hoffman, M.R. (2020). Odds Ratio (Treasure Island, FL: StatPearls Publishing).
23. Popejoy, A.B., Crooks, K.R., Fullerton, S.M., Hindorff, L.A., Hooker, G.W., Koenig, B.A., Pino, N., Ramos, E.M., Ritter, D.I., Wand, H., et al. (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am. J. Hum. Genet. *107*, 72–82.
24. Morning, A. (2014). Race and its Categories in Historical Perspective (Brooklyn Hist. Soc. Crossing Borders, Bridg. Gener).
25. Mukhopadhyay, C.C. (2018). Getting rid of the word "caucasian.". In Privilege (Routledge), pp. 231–236.
26. Grant, R., and Orr, M. (1996). Language, Race and Politics: From "Black" to "African-American.". Polit. Soc. *24*, 137–152.
27. Martin, B.L. (1991). From Negro to Black to African American: The Power of Names and Naming. Polit. Sci. Q. *106*, 83–107.
28. Smith, T.W. (1992). Changing Racial Labels: From "Colored" to "Negro" to "Black" to "African American.". Public Opin. Q. *56*, 496–514.
29. Mora, G.C. (2014). Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American (University of Chicago Press).
30. Popejoy, A.B. (2021). Too many scientists still say Caucasian. Nature *596*, 463.
31. Roberts, D. (2011). Redefining Race in Genetic Terms. In Fatal Invention: How Science, Politics, and Big Business Re-Create Race in the Twenty-First Century (The New Press), pp. 58–80.
32. Fullwiley, D. (2008). The biologistical construction of race: 'admixture' technology and the new genetic medicine. Soc. Stud. Sci. *38*, 695–735.
33. Fujimura, J.H., and Rajagopalan, R.M. (2020). Race, ethnicity, ancestry, and genomics in Hawai'i: Discourses and practices. Hist. Stud. Nat. Sci. *50*, 596–623.
34. Duello, T.M., Rivedal, S., Wickland, C., and Weller, A. (2021). Race and Genetics vs. 'Race' in Genetics: A Systematic Review of the Use of African Ancestry in Genetic Studies (Evol. Med. Public Heal).
35. Rajagopalan, R., and Fujimura, J. (2012). Making history via DNA, making DNA from history: Deconstructing the race-disease connection in admixture mapping. In Genetics and the Unsettled Past: The Collision of DNA, Race, and History, K. Wailoo, A. Nelson, and C. Lee, eds. (Rutgers University Press), pp. 143–163.

The American Journal of Human Genetics, Volume *108*

Supplemental information

Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades

Yen Ji Julia Byeon, Rezarta Islamaj, Lana Yeganova, W. John Wilbur, Zhiyong Lu, Lawrence C. Brody, and Vence L. Bonham

Repairing of words broken over line breaks

Text obtained from OCR versions of scanned PDFs contained words broken over lines with hyphens. In order to detect and repair these broken words, we compared the text in our files with text from the PubMed database. PubMed contains over 31 million records, of which over 17 million have abstracts. Because PubMed does not break words by inserting hyphens at the end of lines, we can compare the AJHG text with the text from PubMed records to detect broken words. For this purpose, we computed the frequencies of all single words and two-word pairs that did not involve stop words. If tokens A and B appeared separated by a hyphen in AJHG text, we determined the frequency in PubMed of A, B, AB, and A B. If AB was a stop word, had a frequency of more than 10,000, had a frequency more than ten times the frequency of the least frequent of A or B, or was more frequent than A B, we substituted AB as the correct term. Otherwise, we left the hyphenated form.

Figure S1. Yearly number of articles and 200-token segments

Number of even 200-token segments (pink) and articles (blue) in corpus by year, from 1949-2018.

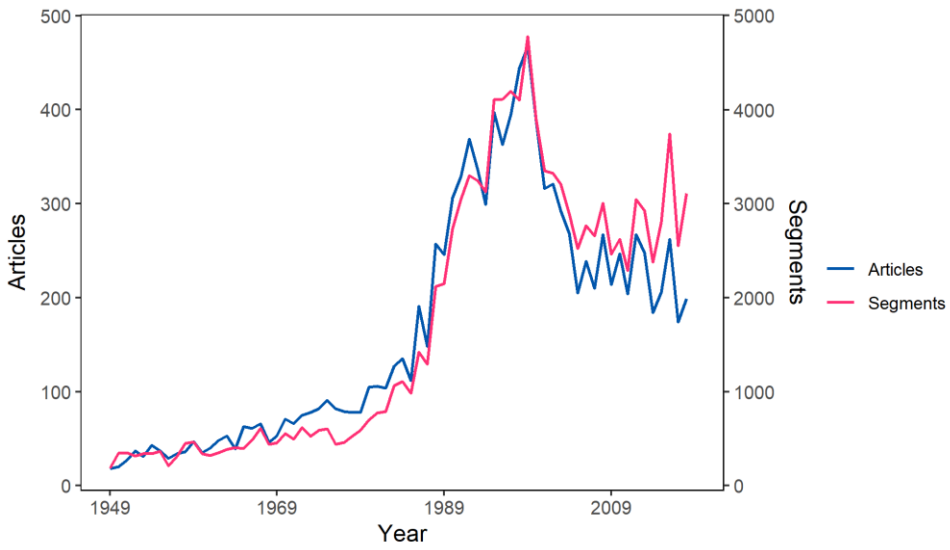

Figure S2. Frequency of fifteen topical terms

Yearly percentage of even 200-token segments (pink) and articles (blue) containing admixture (A), allele (B), behavior (C), diversity (D), environment (E), frequency (F), geography (G), haplotype (H), language (I), linkage (J), locus (K), population (L), religion (M), social (N), socioeconomic (O), from 1949-2018.

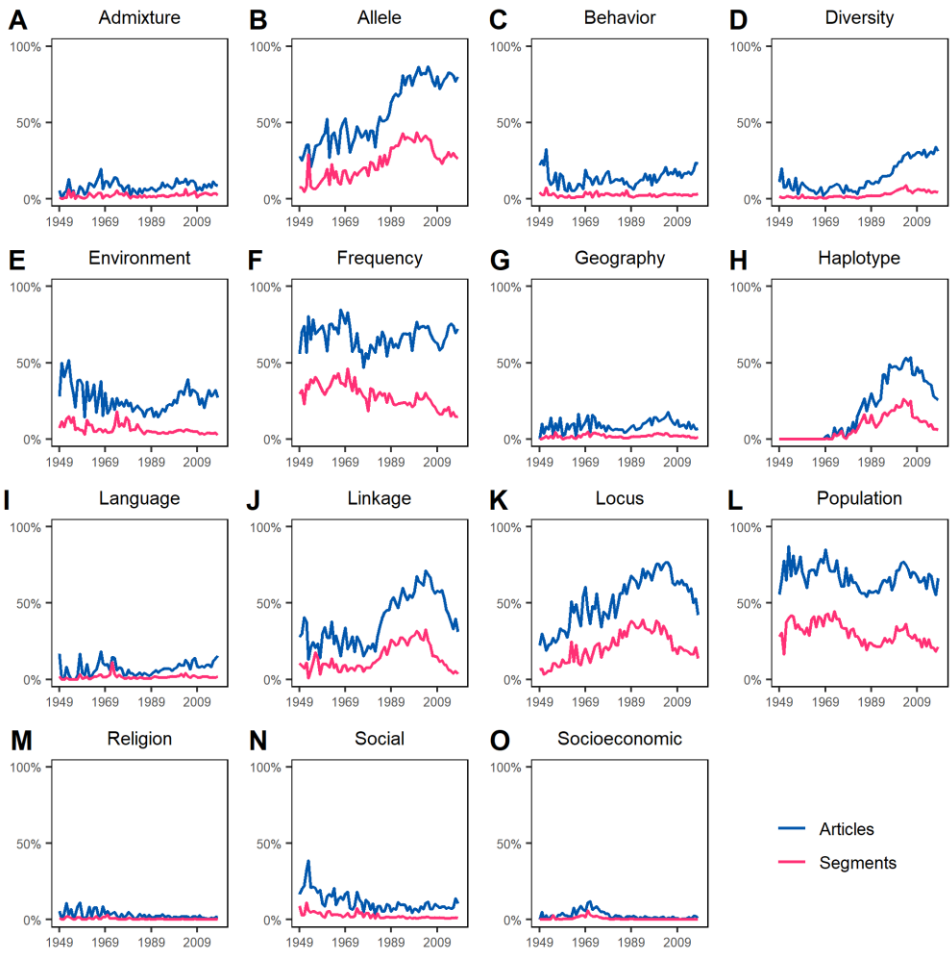

Figure S3. Odds ratios between ancestry, ethnicity, race and topical terms

ORs between ancestry (pink), ethnicity (green), race (blue) and allele (A), behavior (B), environment (C), frequency (D), linkage (E), and locus (F) in overlapping decades from 1949- 2018 (e.g. 1949-59, 1950-60, 1951-61…). Each point on the line graph represents the value of the ratio for which the corresponding year is the midpoint (e.g. values at 1954 represent co-occurrence ratios for 1949-59). Solid line segments indicate decades where the number of co-occurrences was significantly greater than expected by chance, with $p \leq 0.05$. Dotted line segments indicate that the number of co-occurrences was not significantly greater than expected by chance. Shaded regions surrounding a curve and of the same color indicate 95% confidence bounds. The horizontal red dashed line marks an OR of 1.0.