bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-ND 4.0 International license.

Title

A standardized framework for representation of ancestry data in genomics studies

Authors

1. Joannella Morales[1*],
2. Emily H. Bowler[1],
3. Annalisa Buniello[1],
4. Maria Cerezo[1],
5. Peggy Hall[2],
6. Laura W. Harris[1],
7. Emma Hastings[1],
8. Heather A. Junkins[2],
9. Cinzia Malangone[1],
10. Aoife C. McMahon[1],
11. Annalisa Milano[1],
12. Danielle Welter[1],
13. Tony Burdett[1],
14. Fiona Cunningham[1]
15. Paul Flicek[1],
16. Helen Parkinson[1]
17. Lucia A. Hindorff[2],
18. Jacqueline A. L. MacArthur[1*],

[1]*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK*
[2]*Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA*

Author information
Lucia A. Hindorff and Jacqueline A.L. MacArthur share joint last authorship of this manuscript.

Abstract

The accurate characterization of ancestry is essential to interpret and integrate human genomics data and for individuals from all ancestral backgrounds to benefit from advances in the field. However, there are no established guidelines for the consistent, unambiguous description of ancestry. To fill this gap and increase standardization, we developed a framework that is applicable to all human genomics studies and resources. In this report we describe the framework and its use to curate all 2,854 NHGRI-EBI GWAS Catalog publications. We demonstrate the broader relevance through its application to populations in projects such as HapMap and 1000 Genomes. We outline recommendations for authors on the implementation of our method and urge that, wherever possible, ancestry be determined using genomic methods. Finally, we present an analysis of the ancestry of individuals, studies and associations included in the Catalog. While the known

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-ND 4.0 International license.

bias towards inclusion of European ancestry individuals persists, African and Hispanic or Latin American ancestry populations contribute disproportionately more associations than expected. We thus encourage the scientific community to target future GWAS and other discovery studies to under-represented groups, which, in addition to being intrinsically merited, may also be more effective at identifying new associations. Widespread adoption of the framework presented here will enable improved analysis, interpretation and integration of data and ultimately, further our understanding of disease.

Text

The past 15 years has seen a dramatic growth in the field of genomics, with numerous efforts focused on understanding the etiology of common human disease and translating this to advances in the clinic. Genome-wide association studies (GWAS), in particular, are now a well-established mechanism to identify links between genetic variation and human disease[1]. The NHGRI-EBI GWAS Catalog[2] (www.ebi.ac.uk/gwas), one of the largest repositories of summary GWAS data, contains over 2,800 publications and 37,000 associations as of April 2017. The Catalog is indispensable for researching existing findings on common diseases, enabling further investigations to identify causal variants, understand disease mechanisms and establish targets for treatment[3–6].

Essential to the interpretation, integration and application of genomic data is the accurate description of the ancestry of the samples studied. However, there are currently no established guidelines for the determination, characterization and classification of ancestral background information, leading to ambiguity and inconsistency when describing populations. This lack of precision has an impact beyond the interpretation and application of results. It can also bias scientific endeavor towards well-described cohorts, perpetuating a cycle of disadvantage for underrepresented communities. The need for genetic studies in more ancestrally diverse populations has been repeatedly articulated[7], most recently by Popejoy and Fullerton[8]. The benefit of including diverse populations extends throughout the translational research spectrum, from GWAS discovery efforts to genomic medicine, for which variant interpretation can be greatly aided by ancestrally diverse sequence information[9,10]. Although inclusion efforts are improving over time, it is challenging to assess the status of such efforts without a defined way of representing ancestry data. In an effort to fill these gaps, we developed a framework to systematically describe and represent detailed ancestry information, with an emphasis on the use of terminology and classification that reflects genetically-determined ancestry. Our framework is applicable to all genomics studies and resources involving human subjects. Its widespread adoption will enable improved analysis, interpretation and integration of data and ultimately, further our understanding of the genetic architecture of disease.

In this article we first describe the framework and its application to the curation, access and visualization of detailed ancestry data for all GWAS Catalog studies. Second, we demonstrate the applicability of our methodology

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

beyond the Catalog by mapping the HapMap Project[11] and 1000 Genomes phase 3 populations[12] to our framework-derived ancestry categories. Third, we outline a set of recommendations for the application of our method to reporting ancestry data in publications. Finally, we present the distribution of ancestral backgrounds within the participants, studies and associations from the GWAS Catalog.

Results

Standardized ancestry representation

The purpose of the framework is to enable the generation of a comprehensive and standardized description of the ancestry of study samples. We recommend that, whenever possible, authors use genomic techniques to determine the ancestry of study participants. Box 1 outlines several approaches currently in use for this purpose. Authors should avoid relying on self-reporting as the main source of information, given that it is a subjective measure, and often not accurately representative of the underlying genetic background. Additionally, software to assess and control for ancestry is readily available and computationally feasible[13–17].

Application of this framework, for example, by curators to samples reported in publications relies on manual extraction of author-reported data, with precedence given to information determined using genomic methods. When the information provided by authors is limited or ambiguous, curators may take into account country of recruitment demographics and peer-reviewed population genetics publications.

Our method involves capturing data in two forms: (1) a free-text description including detailed information about the genealogy of the samples and (2) a structured description generated by mapping the free-text description to defined categories and country identifiers from a list of controlled terms.

The free-text description should be detailed and accurately represent each distinct group analyzed in a specific study. We note that language used by different authors to describe the same population often varies. This heterogeneity of terms likely reflects the use of self-reported descriptors, which may vary depending on the cohort, country of recruitment and/or other factors. Wherever possible, use genetically-confirmed ancestry descriptors, such as Aboriginal Australian[18,19]. Otherwise, use terms that place the samples in context with other populations with known or inferred genetic relatedness, for example by clustering with known reference populations in principle component analysis. In general, terms that pertain to an individual's ethno-cultural background should be avoided, unless this provides additional information regarding the genealogy of the samples. In such cases a descriptor that accurately reflects the underlying genetics should also be provided, such as "Ashekenazi Jewish European ancestry individuals". Particular care should be taken to note if a sample derives from founder or genetically isolated populations. Given their homogeneity and reduced genetic variation, these populations are especially well-suited for GWAS[20] and are increasingly used as sample sources. When describing isolates, the broader

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

genetic background within which this population clusters should also be indicated (e.g. “Old Order Amish population isolate individuals of European ancestry”). If authors are not aware of the ancestry of participants or cannot share it due to confidentiality concerns, we suggest noting this in the publication and avoiding the use of ancestry-related terminology when describing the samples.

Each free-text ancestry description should be mapped to one ancestry category chosen from a list of 17 categories (Table 1) representing distinct regional population groupings. The individuals included in each category are expected to have genetic variation representative of, or known relatedness to, the population in these regions, excluding recent migrations. The mappings of free-text descriptions to the ancestry categories should be carefully considered for each study. We recommend authors determine whether the genetic variation of the study samples is representative of a population with known genetic variation, according to the categories listed in Table 1, and indicate the chosen category in the publication. Accuracy will increase as reference data sets are refined and/or expanded to include additional populations. In the absence of any genetically-determined ancestry information, category assignment should primarily rely on genealogy, rather than purely on geographical location. For descriptors that refer to a population from a country with a homogenous demographic composition, such as Japanese when referring to individuals recruited in Japan, the corresponding category (East Asian) is straightforward. However, for descriptors related to countries with limited published information pertaining to genetic genealogy, such as Azerbaijan, or those with more ancestral diversity, such as Singapore, the distribution of samples to categories is more challenging. In these cases, the United Nations regional and sub-regional groupings (http://unstats.un.org/unsd/methods/m49/m49regin.htm) and the CIA World Factbook (https://www.cia.gov/library/publications/resources/the-world-factbook/index.htm) may be consulted to obtain geographical data and country-specific population information, respectively (Supplementary Table 1). The Factbook is a regularly updated, comprehensive compendium of worldwide demographic data, covering all countries and territories of the world. The descriptors listed therein, while biased by census constructs and not necessarily genetic in nature, are often used by individuals when self-reporting and may allow the mapping of a descriptor to a category. We recommend the Factbook only be consulted in cases where the only known information is the country of recruitment of participants. We expect that as increased care is taken in publications to accurately report ancestry data, reliance on this resource will decrease. Peer-reviewed population genetic studies that report on the genetic background of a given population may also be consulted. This is particularly helpful in cases where the sample cohort is described by authors using geographical or ethno-cultural terms, such as Scandinavian or Punjabi Sikh, or if a study sample could be mapped to several categories due to admixture, such as in the case of Brazilian ancestry. When considering data from these secondary sources, precedence should be given to author-provided data from the original publication, as authors likely have the most accurate information about the samples included in the published study.

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

The structured description also includes country of recruitment (Figure 1 and Supplementary Figure 2a) and country of origin information. Authors should provide information about the country where the samples were collected, avoiding its use as a substitute for the ancestry descriptor. Curators should not infer country of recruitment from an ancestry or cohort descriptor. Country of origin should only be recorded if the country of origin of the study participant's grandparents or the genealogy of the participants dating several generations is known.

Application of the framework

To ensure consistent application of the framework by GWAS Catalog curators, we created a set of detailed extraction guidelines (Supplementary Note). Ancestry data have been manually curated for all GWAS Catalog studies, encompassing over 2,800 publications and 3,700 GWAS studies as of April 2017. Following our recommendations, described above and in Box 2, free-text descriptors were generated for all samples, based on the language provided by authors. These were then each mapped to the corresponding ancestry category from the list in Table 1. Supplementary Table 2 shows free-text descriptors currently in use in the GWAS Catalog, along with the mapped ancestry category. Specific examples to illustrate the application of the framework to Catalog samples can be found in Supplementary Table 3. All curated ancestry data is available from the GWAS Catalog website (Figure 1, Supplementary Fig. 1 and Supplementary Fig. 2a and 2b; www.ebi.ac.uk/gwas) and via download (www.ebi.ac.uk/gwas/api/search/downloads/ancestry). For a subset of Catalog studies, we have curated information describing the method(s) and sources of data utilized by authors when determining the ancestry of study samples. A comparison of the first 100 publications in the Catalog (2005 – 2008) to the first 100 publications of 2016 demonstrated an increase in the number of publications that use genomic methods to determine ancestry (25% in 2005-2008 compared to 57% in 2016) and a decrease in the number of publications that do not provide any ancestry information (15% compared to 3%). However, the number of publications that do not report the source or method of ancestry determination remained fairly constant (17% compared to 18%; Supplementary Fig. 3). In the future, to further encourage the use of genetic methods of ancestry ascertainment, the GWAS Catalog will support the capture and display of this information for every study, starting with 2017 publications and onwards.

To facilitate the access of curated ancestry data, we are developing an ancestry-specific ontology based on our framework. We have defined synonyms and established hierarchical relationships between all curated terms and categories, so that, in the future, when the ontology is integrated into the Catalog's search interface, users are able to perform more powerful and precise ancestry-related queries[21]. This will also be of benefit to other resources wishing to formalize ancestry data. The ancestry ontology can be browsed and downloaded at http://www.ebi.ac.uk/ols/ontologies/ancestro (manuscript in preparation).

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

While this framework was initially devised for immediate application to the GWAS Catalog, it is designed to be relevant to any study or resource involving human subjects. To demonstrate this, we mapped the HapMap Project[11] phase 3 and 1000 Genomes Project[12] phase 3 populations to our categories (Supplementary Table 4). This effort is the first step towards the incorporation of ancestry-specific linkage disequilibrium and population genetics data. It also facilitates the integration of Catalog data with other studies involving these populations.

Analysis of structured GWAS Catalog ancestry data

Several authors[7,8] have reviewed the ancestry distribution in the Catalog, but focused exclusively on the free-text descriptions. Here we present the first analyses using our curated structured data. The use of categories facilitates searching, thus allowing a refined and more coherent review of the data. Similar to previous reports, we found that the majority (77%) of individuals in the Catalog are exclusively of European background (Figure 2a). The second largest group includes individuals of Asian descent (13%), with East Asians comprising 11% of the Catalog's samples. The disproportionate focus on Europeans was more prevalent in the earlier years of the Catalog (86% of individuals in studies published between 2005 and 2010; 74% between 2011 and 2015, Figure 3). The reduced number of European ancestry participants added to the Catalog in the last 5 years correlates with an increase in Asian (7.3% to 14.85%, 2-fold increase), African (0.8% to 3.3%, 4-fold increase), Hispanic/Latin American (0.1% to 0.8%, 6-fold increase) and Middle Eastern (0.01% to 0.05%, 5-fold increase) participants. Though the proportion of Hispanic/Latin Americans exhibited the largest increase, when considering the absolute number of individuals, the largest increase, by far, came from Asian populations; Asian ancestry individuals increased from almost 900,000 in the first 5 years to 6 million added to the Catalog in the last 5 years, compared to an increase of 300,000 Hispanic/Latin Americans.

Interestingly, when we focused our analysis on the number of associations identified in each ancestry category, we noted a different distribution to the ancestry distribution of individuals (Figure 2c). This disparity is particularly pronounced for studies including African or Hispanic/Latin American samples; African ancestries contribute 3% of individuals but 8% of associations, while Hispanic/Latin Americans contribute <1% of individuals compared to 5% of associations. The opposite effect was seen in Europeans, with 52% of associations compared to 77% of individuals. In addition, we also observed a disproportionate number of associations contributed by the "Multiple ancestries" category, likely reflecting the Catalog's inclusion of trans-ethnic meta-analyses and replication efforts in diverse ancestries.

A review of the traits with the largest number of studies in the Catalog presented the same European bias, with 57-80% of studies, depending on the trait, carried out in European ancestry individuals, followed by East Asians (7-28% of studies) (Supplementary Fig. 4). In studies that analyzed multiple ancestries, the vast majority (> 90%) include European ancestry individuals, regardless of the trait. This trend mirrors what we observe when considering all traits included in the Catalog (Figure 2b). The traits that display the largest

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

proportion of ancestral diversity are anthropometric traits, such as body mass index (BMI) and height, and common diseases, including Type 2 Diabetes and cardiovascular disease (Supplementary Fig. 4).

Discussion

Several reports have been published urging the scientific community to ensure that individuals from all ancestry backgrounds benefit from advances in the field of genomics[7,8]. However, it is difficult to track progress in this area without standardized guidelines and metrics. Our proposed framework addresses both challenges. It lays out a mechanism for the generation of consistent and comprehensive ancestry descriptions and this, in turn, facilitates the tracking of ancestry data over time.

Our analyses confirm the persistent bias towards inclusion of European ancestry individuals, which is more pronounced during the early years of the Catalog. This disproportionate focus on European ancestry populations arose from the availability of large, homogeneous cohorts assembled as initial GWAS began. At the time, these cohorts consisted of mainly European ancestry individuals. While we demonstrate a trend towards ancestral diversity with respect to the samples added to the Catalog in the last 5 years, it is important to note that this is largely attributable to an increase in the number of East Asian ancestry individuals. This increase is decidedly a step in the right direction, but we believe falls short of the goal; ancestry distribution in the Catalog does not reflect global population demographics or disease burden. Notably, all “non-European, non-Asian” individuals combined only amount to 4% of the total number of samples in the Catalog, with Hispanic or Latin American, Middle Eastern, Native American and Oceanian populations contributing less than 1% each (Figure 2a).

In addition to confirming published observations, we performed more robust analyses, going beyond proportions of individuals to proportions of associations and traits, and tracking change over time. Our analysis of approximately 33,000 associations noted a disproportionately larger number of associations derived from African and Hispanic or Latin American populations, many of which have significant African admixture[22], than is expected based on the ancestry-specific distribution of individuals. A higher degree of genetic diversity and reduced linkage disequilibrium (LD) in African populations has been described previously[23]. This likely offers an explanation for the results obtained in our analysis, for two reasons. First, shorter LD blocks in African populations facilitate the separation of nearby but independent signals in a way that is more challenging in European populations, in whom LD blocks tend to be longer. Second, the inclusion of larger numbers of African ancestry populations allows for the identification of population-specific variants. Together, these observations suggest that utilizing samples from diverse populations for genomic studies may be advantageous and yield increased and more comprehensive results. Of the commonly studied traits, the largest diversity of backgrounds was found for common anthropometric traits, heart disease, and type 2 diabetes. This is perhaps not surprising considering that metrics for these traits are easy to

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

obtain, and the two diseases are among the top ten causes of death around the world, according to the World Health Organization (http://www.who.int/mediacentre/factsheets/fs310/en/). It is also consistent with the observation that diseases for which global disease burden is substantial tend to lead to increased funding and research infrastructure. Further efforts are required to make sure that diseases that disproportionately affect underrepresented ancestral backgrounds are given proper attention and are analysed in suitable cohorts.

There are limitations to our analysis. First, considering that some cohorts have been included in numerous GWAS, it is highly likely that some individuals are represented multiple times in the Catalog. The impact of this is the skewing of results towards commonly-used or publicly available cohorts, which are perhaps likely to be of European ancestry. Another limitation stems from our criteria for inclusion of associations. Since we only include SNPs with a p-value < $1x10^{-5}$ and only the "index" SNP at each locus, our analysis does not take into account all associations. To address this and make the Catalog more comprehensive, we now make available published summary statistics. This will ensure that future analyses of Catalog data are less biased. Finally, we were unable to assign a category to associations identified in studies that include multiple ancestries. This may be a factor contributing to the reduced number of associations derived from European populations, since the vast majority of multiple ancestry studies include Europeans (Figure 2b).

The analysis of diverse ancestries is advantageous from a scientific perspective. No one population contains all human variants[12], and alleles that are rare in one population may be common in a different population and thus easier to detect. Studies of diverse populations may also aid in fine mapping of existing signals or in identifying population-specific functional variation[12, 24]. Taking population-specific LD patterns into account when designing genotyping arrays will likely facilitate identification of causal variants. Variant interpretation for genomic medicine in ancestrally diverse or admixed populations relies on the availability of non-European allele frequencies, with potentially serious clinical consequences if such data are not available[9]. Finally, disease burden of common or complex diseases (e.g., cardiovascular disease or cancer) disproportionately impacts non-European populations. While we are encouraged by the trend we have seen in recent years towards increased diversity, we note that there are still very clear gaps as some groups continue to be underserved or ignored. We strongly urge the scientific community to expand their efforts to assemble and analyze cohorts, including especially underrepresented communities.

Our analysis not only serves to highlight these important gaps, but also validates the need for this framework as a methodology to improve the description of ancestry in publications. Approximately 5% of individuals in the Catalog (2005 - 2015) are currently mapped to the category "Not reported" due to a lack of adequate information in the publication. Although confidentiality concerns certainly contribute to this, this large proportion of uncharacterized samples supports the notion that guidelines for the reporting of ancestry data are an absolute necessity. For this reason, we offer

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

recommendations to increase standardization of ancestry reporting, with an emphasis on genetic determination of ancestry, in publications (Box 2). We encourage implementation by authors reporting ancestry data and by editors reviewing publications that include human subjects.

There are challenges inherent to both the design of the framework and its application. First, we recognize the sensitivities surrounding the concepts of race, ethnicity and ancestry, and that these terms are often used interchangeably without making a distinction between physical appearance, cultural traditions and genetic variation. This conflation can often be observed in censuses and other demographic tools, influencing how individuals and communities describe their background. As a result, self-reported data may be subjective and may not align with the underlying genetics. The United States Census, for example, defines ancestry as “one’s ethnic origin or descent, "roots," or heritage, or the place of birth of the person or the person’s parents or ancestors before their arrival in the United States”, and recognizes that census classifications “should not be interpreted scientifically”. Specifically, the definitions for “White”, “Black” and “Hispanic or Latino” are problematic from a genetic perspective. For example, the Census allows individuals of Central Asian and Middle Eastern background to self-identify as “white”, even though these populations are known to cluster, in genetic analyses, independently from European ancestry populations. Similarly, the Census allows individuals of Sub-Saharan African descent to self-identify as African American, conflating these two categories. For all the reasons mentioned above, we here recommend that authors move away from relying solely on self-reported information, and instead use genomic mechanisms to determine and describe the ancestry of participants. Box 2 outlines methods currently in use for this purpose. We are aware that when analyzing some cohorts, such as subsets of individuals with electronic medical records, the only data accessible is self-reported. However, we note that the trend in publications is towards genetic determination of ancestry and/or genetic confirmation of self-reported data followed by removal of outliers, if necessary. For instance, our assessment of 100 GWAS studies published in 2016 reveals that the ancestry of at least one study cohort was genetically-determined in 57% of publications, more than double the 25% observed in 100 studies published in 2005 – 2008 (Supplementary Fig. 3).

Another challenge stems from the process of assigning descriptors to one of the categories in Table 1. Given the use of self-reporting and curator inferences due to imprecise characterizations in publications, our categories are not perfect genetic groupings. They should not be taken as definitive or authoritative scientifically-determined global ancestral classifications. Rather, our categories represent regional population groupings that include individuals with distinct and well-defined patterns of genetic variation as well as individuals with known or inferred relatedness to the populations in that grouping. We note that our categories were generated for immediate application to the GWAS Catalog and are a reflection of the information in publications curated and included in it. They are not exhaustive or static; we envision that as more cohorts from diverse populations are analyzed, there might arise a need to create additional categories or sub-categories. Also, as

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

the community continues to move towards the genetic determination of ancestry, our categories will become more precise and granular over time. We recognize, however, that even when using genomic methods to determine ancestry, attempting to classify individuals with significant admixture or belonging to under-studied populations can be challenging. Accurate genetic classification requires well-defined reference populations, such as those included in the HapMap and 1000 Genomes Projects, or informative genetic markers that allow populations to be distinguished. We note that these are lacking for some groups (e.g. Greater Middle Eastern populations) and we thus encourage increased efforts to fill this gap.

Genome-wide association studies have been enormously successful. However, the lack of clarity regarding the ancestry of samples and the lack of studies including diverse ancestral backgrounds raises questions about the interpretation and generalizability of results across populations. The framework we have developed aims to address these challenges. Its widespread adoption will enable the scientific community to investigate the generalizability of trait-associations across diverse populations, to identify associations unique to specific ancestries, to identify novel variants with clinical implications, and to help pinpoint causative variants, thus increasing our understanding of common diseases.

Methods

GWAS Catalog data curation

Details of GWAS publication identification, GWAS Catalog eligibility criteria and curation methods can be found on the GWAS Catalog website www.ebi.ac.uk/gwas/docs/methods. Extracted information encompasses publication information, study cohort information, including ancestry, and SNP-trait association results. Curation of ancestry data from the literature was performed following detailed extraction guidelines (Ancestry Extraction Guidelines in Supplementary Note).

1000 Genomes and HapMap Project population ancestry assignment

Information describing the 1000 Genomes[12] phase 3 and HapMap Project[11] phase 3 populations was taken from the Coriell Institute website (https://catalog.coriell.org/). Ancestry information, including ancestry category, country of recruitment, country of origin and additional information, was assigned to each population following the GWAS Catalog ancestry extraction guidelines (Supplementary Note).

GWAS Catalog ancestry analysis

To determine the distribution of individuals, associations and traits by ancestry category, we first downloaded all Catalog data in tabular form from http://www.ebi.ac.uk/gwas/docs/file-downloads. All data (gwas-catalog-associations_ontology-annotated.tsv, gwas-catalog-ancestry.tsv, gwas-catalog-studies_ontology-associated.tsv, gwas-efo-trait-mappings.tsv) included in these analyses was curated from GWA studies published between 2005 and the end of 2015, with a release date of October 25 2016. The data

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

can be found on the Catalog's FTP site (ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2016/10/25/).

Assessment ancestry determination methods in a subset of the GWAS Catalog

We selected the first 100 publications included in the Catalog (approximately covering the period between March 2005 to January 2008), and for comparison, the first 100 publications from 2016. For each publication, the method was assessed and classified into one of the following: 1. Self-reported, 2. Genetically determined, 3. Ancestry stated without method, 4. Inferred from limited ancestry-related information (e.g. country information), 5. No ancestry information reported and 6. Mixed method (when a combination of methods was utilized to describe the study samples). Publications classified as "Genetically determined" includes those where the author had clearly identified the genetic ancestry or admixture of the population, for example by using methods such as those described in Box 1. It also includes those that confirmed self-reported information or defined samples based on self-reports but then excluded genetic outliers. Publications where no ancestry was stated, but curators inferred an ancestry based on country information are included in the fourth classification. In many cases authors used a statistical method to assess or control for ancestry or population stratification, without assigning individuals to a particular category, for example using a continuous axis of genetic variation from PCA to compute the association statistic. However, since this did not add any information that curators could use to assign a population ancestry to the study, it was not included under category 2.

References

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-1006 (2014).
2. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
3. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–386 (2015).
4. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89,** 607–618 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

5. Pal, L. R. & Moult, J. Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. *J. Mol. Biol.* **427,** 2271–2289 (2015).
6. Mullen, J., Cockell, S. J., Woollard, P. & Wipat, A. An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. *PloS One* **11,** e0155811 (2016).
7. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet. TIG* **25,** 489–494 (2009).
8. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).
9. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375,** 655–665 (2016).
10. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6,** 91 (2014).
11. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
12. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
13. Porras-Hurtado, L. *et al.* An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front. Genet.* **4,** 98 (2013).
14. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
15. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

16. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

17. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).

18. Huoponen, K., Schurr, T. G., Chen, Y. & Wallace, D. C. Mitochondrial DNA variation in an aboriginal Australian population: evidence for genetic isolation and regional differentiation. *Hum. Immunol.* **62,** 954–969 (2001).

19. Nagle, N. *et al.* Antiquity and diversity of aboriginal Australian Y-chromosomes. *Am. J. Phys. Anthropol.* **159,** 367–381 (2016).

20. Cronin, S. *et al.* A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum. Mol. Genet.* **17,** 768–774 (2008).

21. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25,** 1251–1255 (2007).

22. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. *Curr. Opin. Genet. Dev.* **41,** 106–114 (2016).

23. Campbell, M. C., Hirbo, J. B., Townsend, J. P. & Tishkoff, S. A. The peopling of the African continent and the diaspora into the new world. *Curr. Opin. Genet. Dev.* **29,** 120–132 (2014).

24. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet. EJHG* **24,** 1330–1336 (2016).

25. Martínez-Cruz, B. *et al.* In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur. J. Hum. Genet. EJHG* **19,** 216–223 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

26. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345,** 1255832 (2014).

27. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48,** 1071–1076 (2016).

28. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet.* **11,** e1005602 (2015).

29. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).

30. Kayser, M. The human genetic history of Oceania: near and remote views of dispersal. *Curr. Biol. CB* **20,** R194-201 (2010).

Acknowledgements.

Research reported in this publication was supported by the National Human Genome Research Institute and the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers U41-HG007823 and U41-HG006104. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by the European Molecular Biology Laboratory. L.A.H., P.H. and H.J. are employees of the National Human Genome Research Institute. The authors wish to thank all GWAS Catalog users and authors of studies included in the Catalog. We also thank Chris Gignoux for his expert review of the genomic methods of ancestry determination discussed in this manuscript and to Teri Manolio for valuable discussion.

Author contributions.

J.M., J.A.L.M., P.H. H.A.J. and L.A.H. conceived this study and developed the ancestry framework. J.M., J.A.L.M., E.H.B., A.B., M.C., P.H., L.W.H., H.A.J., A.C.M., A.M. and L.A.H. performed curation of ancestry data of GWAS Catalog publications. J.M., J.A.L.M., M.C., T.B. and L.A.H. analyzed the distribution of ancestry categories in the Catalog and interpreted the data. L.W.H, J.A.L.M., L.H. and J.M. assessed the methods of ancestry determination utilized in GWAS Catalog studies and interpreted the data. A.C.M. and J.M. generated the figures. J.M., J.A.L.M and L.W.H. generated the Tables. E.H., D.W., C.M. and T.B. developed the GWAS Catalog curation and search interfaces. D.W. created the ancestry ontology, with contributions

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

from J.M., J.A.L.M. and E.H.B. All authors contributed to the final manuscript, with J.M., J.A.L.M. and L.A.H. playing the key roles.

Competing financial interests. PF is a member of the Scientific Advisory Board of Omicia, Inc.

Materials & Correspondence. Correspondence to Joannella Morales and Jacqueline A. L. MacArthur.

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Table 1 – Ancestry categories. These represent distinct regional population groupings used in this framework. A full list of GWAS Catalog sample descriptions assigned to each category can be found in supplementary table 2.

Ancestry category	Definition	Examples of free-text descriptions included in category
Aboriginal Australian	Includes individuals who either self-report or have been described by authors as Australian Aboriginal. These are expected to be descendents of early human migration into Australia from Eastern Asia and can be distinguished from other Asian populations by mtDNA and Y chromosome variation[18, 19].	Martu Australian Aboriginal
African American or Afro-Caribbean	Includes individuals who either self-report or have been described by authors as African American or Afro-Caribbean. This category also includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes and/or HapMap ACB or ASW populations. We note that there is likely to be significant admixture with European ancestry populations.	African American, African Caribbean
African unspecified	Includes individuals that either self-report or have been described as African, but there was not sufficient information to allow classification as African American, Afro-Caribbean or Sub-Saharan African.	African, non-Hispanic black
Asian unspecified	Includes individuals that either self-report or have been described as Asian but there was not sufficient information to allow classification as East Asian, Central Asian, South Asian or South-East Asian.	Asian, Asian American
Central Asian	Includes individuals who either self-report or have been described by authors as Central Asian[25]. We note that there does not appear to be a suitable reference population for this population and efforts are required to fill this gap.	Silk Road (founder/genetic isolate)
Circumpolar peoples	Includes native populations of Alaska, Siberia, and the Aleutian Archipelago[26]. This category does not include all native	Alaska Native

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

	populations within the Arctic circle for example the Finnish Saami who are descended from Europeans and are therefore included within the European ancestry category.	
East Asian	Includes individuals who either self-report or have been described by authors as East Asian or one of the sub-populations from this region (e.g Chinese). This category also includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes and/or HapMap CDX, CHB, CHS and JPT populations.	Chinese, Japanese, Korean
European	Includes individuals who either self-report or have been described by authors as European, Caucasian, White or one of the sub-populations from this region (e.g Dutch). This category also includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes and/or HapMap CEU, FIN, GBR, IBS and TSI populations.	Spanish, Swedish
Greater Middle Eastern (Middle Eastern, North African or Persian)	Includes individuals who self-report or were described by authors as Middle Eastern, North African, Persian or one of the sub-populations from this region (e.g. Saudi Arabian)[27]. We note there is heterogeneity in this category with different degrees of admixture as well as levels of genetic isolation. We note that there does not appear to be a suitable reference population for this category and efforts are required to fill this gap.	Tunisian, Arab, Iranian
Hispanic or Latin American	Includes individuals who either self-report or are described by authors as Hispanic, Latino, Latin American or one of the sub-populations from this region. This category includes individuals with known admixture of primarily European, African and Native American ancestries, though some may have also a degree of Asian (e.g. Peru). We also note that the levels of admixture vary depending on the country, with Caribbean countries carrying higher levels of African admixture when compared to South American countries, for example. This category also includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes and/or HapMap CLM, MXL, PEL and PUR populations[22, 28].	Brazilian, Mexican

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Native American	Includes indigenous individuals of North, Central and South America, descended from the original human migration into the Americas from Siberia[29]. We note that there does not appear to be a suitable reference population for this category and efforts are required to fill this gap.	Pima Indian, Plains American Indian
Not Reported	Includes individuals for which no ancestry or country of recruitment information is available.	
Oceanian	Includes individuals that either self-report or have been described by authors as Oceanian or one of the sub-populations from this region (e.g. Native Hawaiian)[30]. We note that there does not appear to be a suitable reference population for this category and efforts are required to fill this gap.	Solomon Islander, Micronesian
Other	Includes individuals where an ancestry descriptor is known but insufficient information is available to allow assignment to one of the other categories.	Surinamese, Russian
South Asian	Includes individuals who either self-report or have been described by authors as South Asian or one of the sub-populations from this region (e.g Asian Indian). This category also includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes and/or HapMap BEB, GIH, ITU, PJL and STU populations.	Bangladeshi, Sri Lankan Sinhalese
South East Asian	Includes individuals who either self-report or have been described by authors as South East Asian or one of the sub-populations from this region (e.g Vietnamese). This category also includes includes individuals who genetically cluster with reference populations from this region, for example 1000 Genomes KHV population. We note that East Asian and South East Asian populations are often conflated. However, recent studies indicate a unique genetic background for South East Asian populations.	Thai, Malay
Sub-Saharan African	Includes individuals who either self-report or have been described by authors as Sub-Saharan African or one of the sub-populations from this region (e.g. Yoruban). This category also includes individuals who genetically cluster with reference populations from this region for example 1000 Genomes	Yoruban, Gambian

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

	and/or HapMap ESN, LWK, GWD, MSL, MKK and YRI populations.	

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Box 1 - Genomic methods of ancestry determination

Approach	Representative Method/Software	Description
Mixture model	STRUCTURE, ADMIXTURE	STRUCTURE[13] analyses differences in the distribution of genetic variants amongst populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation, either allowing for admixture or not. ADMIXTURE[14] is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets designed for efficiency in large GWAS datasets. It uses the same statistical model as STRUCTURE but using a different optimization algorithm.
Principal components analysis	EIGENSTRAT	PCA is a multivariate method used to infer continuous axes of genetic variation (eigenvectors) that maximize the variance explained in a small number of dimensions, whilst describing as much of the variability between individuals as possible. GWAS data can be analysed alone or combined with that from reference samples at the same SNPs. Populations can then be identified and outliers removed if necessary. PCA can also be used to correct for population stratification by creating sets of matched cases and controls; alternatively this information can be included in ancestry-adjusted association analyses such as multiple regression[15].
Multi-dimensional scaling	PLINK	Multi-dimensional scaling (MDS), a related multivariate statistical technique, can also be used to estimate axes of genetic variation. The MDS method detects meaningful underlying dimensions that explain observed genetic distance, e.g., pairwise identity-by-state (IBS) distance, among individuals rather than Euclidean distance in PCA[16].
Mixed effects models	EMMAX	These methods effectively control for stratification within a population and are a popular alternative to PCA for this purpose (reflecting data structure not ancestry per se).

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

		The mixed effects model method models population structure and cryptic relatedness as random effects, while can taking into account fixed effects, such as age and gender[17].

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Box 2 - Recommendations for authors reporting ancestry data in publications.

These recommendations were generated by expert curators following a detailed review of all 2,854 GWAS publications included in the Catalog.

1. Preferentially use genomic methods to assess the ancestry of samples included in the GWAS Catalog. See Box 1 for a description of commonly used methods.
2. Indicate whether the background of participants was self-reported, determined by genomic methods or a combination of both. If genetically determined, indicate the analytical procedure utilized.
3. Provide detailed information for each distinct group of samples,
    a. Ancestry descriptors should be as granular as possible (e.g. Yoruban instead of Sub-Saharan African, Japanese instead of Asian)
    b. Avoid using country or citizenship as a substitute for ancestry
    c. Avoid using geographic descriptors that are part of a cohort name as a substitute for ancestry (e.g. TwinsUK cannot be assumed to be European ancestry).
    d. If a population self-identifies using sociocultural descriptors (e.g. Old Order Amish), clearly state the genetic ancestry within which this sub-population falls.
    e. If samples were derived from an isolated or founder population with limited genetic heterogeneity, clearly state the genetic ancestry within which this sub-population falls.
    f. If available, genetic genealogy or ancestry of grandparents or parents should be included
4. Assign an ancestry category for each distinct group of samples. See Table 1 for a list of ancestry categories. Refer to Supplementary Table 2 for a list of descriptors in use in the Catalog with their category assignments.
5. Provide the sample size for each distinct group of samples included in the analysis.
6. Provide country of recruitment.
7. If ancestry information is not available due to confidentiality, or any other, concerns note this in the publication.

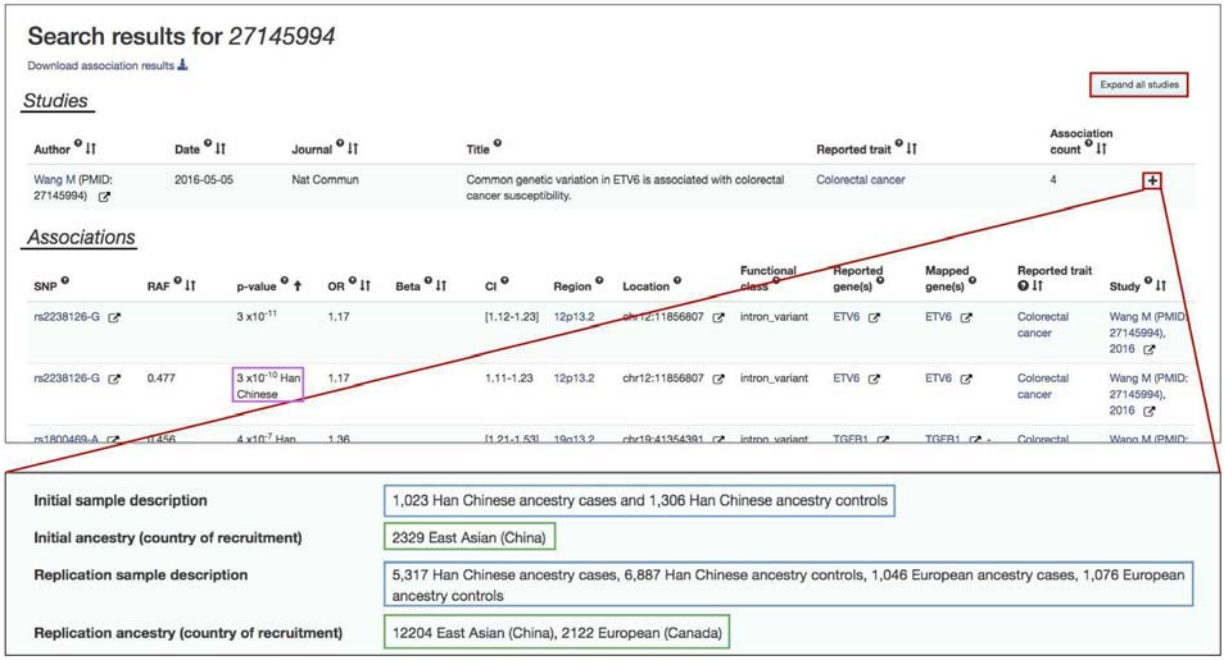
bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figures

1. Figure 1 – Representation of ancestry data in the GWAS Catalog search interface
2. Figure 2 – Ancestry category distribution in the GWAS Catalog
    a. Figure 2a - Distribution of individuals by ancestry category
    b. Figure 2b - Distribution of studies by ancestry category
    c. Figure 2c - Distribution of associations by ancestry category
3. Figure 3 – Distribution of individuals in the 913 studies published between 2005 – 2010 compared to the distribution of individuals in the 2,354 studies published between 2011 – 2015.

bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

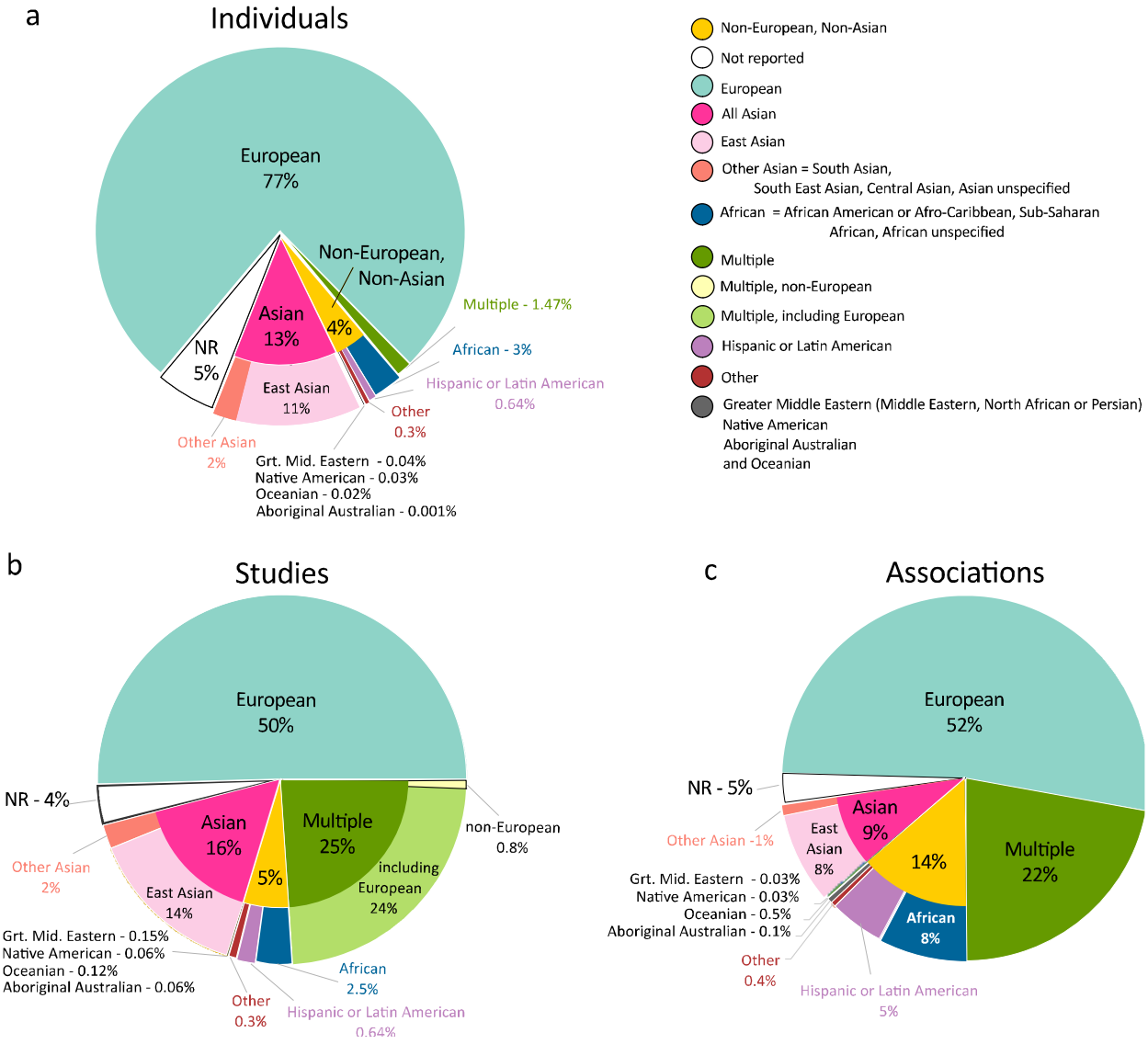
Figure 1 – Representation of ancestry data in the GWAS Catalog search interface (www.ebi.ac.uk/gwas). Ancestry-related data is found in the Studies and Associations tables (underlined in black) when searching the Catalog. This figure shows the results of a search for PubMed Identifier 27145994. The sample description can be found in the Studies table, either by pressing "Expand all Studies" or the "+" on the study of interest (highlighted in red). Sample ancestry is captured in 2 forms: (1) free text description (highlighted in blue) and (2) structured description (highlighted in green). The latter follows the format: sample size, broad category, (country of recruitment). In cases where multiple ancestries are included in a study, the ancestry associated with a particular association is found as an annotation in the p-value column in the Associations table (highlighted in pink).



bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Figure 2. This figure summarizes the distribution of ancestry categories in percentages, of individuals (panel a), studies (panel b) and associations (panel c). The largest category in all panels is European (aqua). At the level of individuals (a), the largest non-European category is Asian (bright pink), with East Asian (light pink) accounting for the majority. Non-European, Non-Asian categories together (yellow) comprise 4% of individuals, and there are 5% (white) of samples for which an ancestry category could not be specified. Panel c demonstrates the disproportionate contribution of associations from African (blue) and Hispanic/Latin American (purple) categories, when compared to the percentage of individuals (a, blue, purple, respectively) and studies (b, blue, purple, respectively).

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/129395; this version posted April 21, 2017. 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-ND 4.0 International license.

Figure 3. This figure displays the distribution of individuals in percentages, included in the 913 studies published between 2005 – 2010 compared to the distribution of individuals included in the 2,354 studies published between 2011 – 2015.