

1 Title

2 A standardized framework for representation of ancestry data in genomics
3 studies, with application to the NHGRI-EBI GWAS Catalog

4

5 Authors

- 6 1. Joannella Morales^{1*}, jmorales@ebi.ac.uk
- 7 2. Emily H. Bowler¹, ehbowler13@gmail.com
- 8 3. Annalisa Buniello¹, buniello@ebi.ac.uk
- 9 4. Maria Cerezo¹, mcerezo@ebi.ac.uk
- 10 5. Peggy Hall², Peggy.Hall@nih.gov
- 11 6. Laura W. Harris¹, ljwh@ebi.ac.uk
- 12 7. Emma Hastings¹, emmakhastings@gmail.com
- 13 8. Heather A. Junkins², junkinsh@mail.nih.gov
- 14 9. Cinzia Malangone¹, cinzia@ebi.ac.uk
- 15 10. Aoife C. McMahon¹, aoifem@ebi.ac.uk
- 16 11. Annalisa Milano¹, annalisa@ebi.ac.uk
- 17 12. Danielle Welter¹, dwelter@ebi.ac.uk
- 18 13. Tony Burdett¹, tburdett@ebi.ac.uk
- 19 14. Fiona Cunningham¹, fiona@ebi.ac.uk
- 20 15. Paul Flicek¹, flicek@ebi.ac.uk
- 21 16. Helen Parkinson¹, parkinson@ebi.ac.uk
- 22 17. Lucia A. Hindorff², hindorffl@mail.nih.gov
- 23 18. Jacqueline A. L. MacArthur^{1*}, jalm@ebi.ac.uk

24

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*

29

30 **Author information**

31 Joannella Morales and Jacqueline MacArthur are corresponding authors.
32 Lucia A. Hindorff and Jacqueline A.L. MacArthur share joint last authorship of
33 this manuscript.

34

35 **Abstract**

36

37 **Background**

38 The accurate description of ancestry is essential to interpret and integrate
39 human genomics data, and to ensure that advances in the field of genomics
40 benefit individuals from all ancestral backgrounds. However, there are no
41 established guidelines for the consistent, unambiguous and standardized
42 description of ancestry. To fill this gap, we provide a framework, designed for
43 the representation of ancestry in GWAS data, but with wider application to
44 studies and resources involving human subjects.

45

46 **Result**

47 Here we describe our framework and its application to the representation of
48 ancestry data in a widely-used publically available genomics resource, the
49 NHGRI-EBI GWAS Catalog. We present the first analyses of GWAS data

50 using our ancestry categories, demonstrating the validity of the framework to
 51 facilitate the tracking of ancestry in big data sets. We exhibit the broader
 52 relevance and integration potential of our method by its usage to describe the
 53 well-established HapMap and 1000 Genomes reference populations. Finally,
 54 to encourage adoption, we outline recommendations for authors to implement
 55 when describing samples.

56

57 Conclusions

58 While the known bias towards inclusion of European ancestry individuals in
 59 GWA studies persists, African and Hispanic or Latin American ancestry
 60 populations contribute a disproportionately high number of associations,
 61 suggesting that analyses including these groups may be more effective at
 62 identifying new associations. We believe the widespread adoption of our
 63 framework will increase standardization of ancestry data, thus enabling
 64 improved analysis, interpretation and integration of human genomics data and
 65 furthering our understanding of disease.

66

67 **Keywords**

68 Genomics - Genome wide association studies – GWAS Catalog – Ancestry –
 69 diversity – Population Genetics

70

71 **Background**

72

73 The past 15 years have seen a dramatic growth in the field of genomics, with
 74 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[1–3]. The NHGRI-EBI GWAS Catalog[2,4], one of the largest repositories of summary GWAS data, contains over 3,000 publications and 43,000 SNP-trait associations as of July 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[5–8].

84

Essential to the interpretation and application of genomic data is the accurate description of the ancestry of the samples studied. Levels of genetic diversity and patterns of linkage disequilibrium (LD) vary by ancestry, with important implications for the experimental design, the generalizability of results and the identification of causal variants. Standardized ancestry representation is necessary to integrate data from different sources for further analysis and to enable robust search functionalities in bioinformatics resources. There are currently no established guidelines for the characterization and classification of ancestral background information. This has led to ambiguity and inconsistency, along with challenges in accessing and integrating data.

95

The need for genetic studies in more ancestrally diverse populations has been repeatedly articulated[9], most recently by Popejoy and Fullerton[10] and by Manolio, et al.[11]. The benefit of including diverse populations extends throughout the translational research spectrum, from GWAS discovery efforts

100 to genomic medicine, for which variant interpretation can be greatly aided by
101 ancestrally diverse sequence information[12,13]. Although inclusion efforts
102 are improving over time, it is challenging to assess the status of such efforts,
103 or to implement improved approaches, without a standardized way of
104 representing ancestry data.

105

106 To fill these gaps, we here provide a framework to systematically describe
107 and represent detailed ancestry information. Our method is driven by an
108 analysis of data we manually curated from over 3,000 GWAS publications.
109 We developed it for immediate application to the GWAS Catalog to make
110 curated data accessible, searchable and compatible with other genomics
111 data. However, the framework is broadly applicable to studies and resources
112 involving human subjects, and its widespread adoption will enable improved
113 analysis, interpretation and integration of data, ultimately furthering our
114 understanding of the genetic architecture of disease in different human
115 populations.

116

117 **Results**

118 Standardized ancestry framework

119 We developed the framework to enable the generation of comprehensive and
120 standardized representation of ancestry information for samples included in
121 GWAS. We had several motivations: 1) our observation, from curating
122 thousands of publications, that ancestry data is often poorly represented,
123 inconsistent or even completely absent, 2) the requirement to describe,
124 represent and store ancestry data in a manner that allows robust searching

125 and visualization of data in the GWAS Catalog, 3) the increased interest in
126 ancestry from the scientific community and 4) the need to analyze the
127 ancestry of samples, assess diversity and generate metrics that would allow
128 the community to identify and address gaps in this area.

129

130 Our framework involves representing ancestry data in two forms: (1) a
131 detailed sample description and (2) an ancestry category from a controlled list
132 (Table 1). Detailed descriptions aim to capture accurate, informative and
133 comprehensive information regarding the ancestry or genealogy of the
134 samples. Ancestry categories are used to establish hierarchical relationships
135 between groups and populations. We believe this dual representation of
136 ancestry is both informative and useful. The two types of descriptions
137 complement each other. The detailed description is granular and
138 heterogeneous, whereas the use of a limited number of categories reduces
139 complexity, facilitating data representation, integration, searchability and
140 further analyses. The framework also allows for country information to be
141 recorded providing additional detail on sample demographics.

142

143 Application of the framework to the GWAS Catalog and other resources

144 Application of this framework by curators to samples reported in publications
145 relies on manual interpretation and extraction of author-reported data. To
146 ensure consistent application by GWAS Catalog curators, we created a set of
147 comprehensive data extraction guidelines (Supplementary Note). When the
148 information provided by authors is limited or ambiguous, we consider country

149 of recruitment demographics and peer-reviewed population genetics
150 publications.

151

152 We have now generated detailed descriptions and assigned ancestry
153 categories to 3,000 publications, representing 4,000 separate GWA studies
154 and 83 million individuals, as of July 2017. A full list of all detailed descriptions
155 currently included in the Catalog is provided in Supplementary Table 1.
156 Examples that illustrate how the framework was applied to Catalog samples
157 can be found in Supplementary Table 2. All curated ancestry data is available
158 from the GWAS Catalog website[4] (Figure 1) and via download[14].

159

160 Detailed description

161 The detailed description aims to accurately represent the ancestry or
162 genealogy of each distinct group analyzed in a specific study in detail, as
163 reported by the author. Information about the homogeneity of the samples,
164 including whether the cohort is admixed or taken from a founder or isolated
165 population, is included. In the GWAS Catalog, the majority of the detailed
166 descriptions include terms that describe the location of participants' ancestors
167 over the past few generations ("French", "Japanese"), while admixed
168 populations are primarily described using ethnic descriptors ("Hispanic").
169 Isolated populations are described using either location or ethnicity terms in
170 addition to being described explicitly as genetically isolated ("Old Order Amish
171 (founder or genetic isolate) population", "Norfolk Island (founder or genetic
172 isolate) population").

173

174 Ancestry categories

175 Ancestry category assignment from the list presented in Table 1 requires
176 careful consideration. When clearly stated, author-reported categories are
177 extracted, with precedence given to genetically-inferred data. If a category is
178 not stated, curators infer the category based on the detailed description for
179 the sample, which, as noted above, represents author-provided information.

180

181 In the absence of any ancestry data, the category “Not Reported” is assigned,
182 unless geographical location of sample recruitment is stated. In such
183 instances, curators infer ancestry from external sources, such as the United
184 Nations[15] and The World Factbook[16]. Selecting a category for samples
185 that derive from a country with a homogenous demographic composition,
186 such as Japan, is straightforward. However, for samples from populations with
187 limited known genetic genealogy, such as Azerbaijan, or for samples recruited
188 in countries with ancestral diversity, such as Singapore, assigning a category
189 is more challenging. These sources are particularly useful to obtain
190 geographical and country-specific population information. The World Factbook
191 is a regularly updated, comprehensive compendium of worldwide
192 demographic data, covering all countries and territories of the world. However,
193 since it does not necessarily provide ancestry data, the World Factbook is
194 consulted when the only known information is the country of recruitment of
195 samples. We expect that as increased care is taken to accurately report
196 ancestry data, reliance on this resource will decrease. Peer-reviewed
197 population genetic studies that characterize the genetic background of a given
198 population may also be consulted. This is particularly helpful in cases where

the sample cohort self-reported or is described using geographical or ethno-cultural terms, such as “Scandinavian” or “Punjabi Sikh”. Supplementary Table 3 provides a list of countries for which external sources were consulted. If the ancestry data provided in publications does not allow the resolution of samples into ancestrally distinct sets, more than one category may be selected from the list in Table 1, such as in the Catalog entry for Jiang R *et al.* [17][18].

206

Country information

Country of recruitment (Figure 1) and country of origin provides additional demographic information and is extracted for each distinct sample set. Country of origin or recruitment is author-reported and not inferred from ancestry data. An exception is made for occasions when authors combine country of recruitment with an ancestry description (“Singaporean Chinese”). In these cases, we infer the country of recruitment (“Singapore”) although it is not explicitly stated. Country of origin is defined as the country of origin of the study participant’s grandparents or as the genealogy of the participants dating several generations.

217

Wider application of the framework

The HapMap[19] and 1000 Genomes[20] projects have delivered a comprehensive survey of human genetic variation across worldwide populations. The application of our method to the ancestry representation of these reference populations therefore provides huge integration potential and demonstrates the relevance of our framework beyond GWAS publications.

224 For all populations, we assigned ancestry category, country of recruitment,
225 country of origin and a detailed description, if provided by each project
226 (Supplementary table 4). We are developing an integrated tool to provide
227 access to population genetic and linkage disequilibrium data from these
228 reference populations. This will be available in the near future.

229

230 To facilitate the application of our approach to databases and resources, we
231 have developed an ancestry-specific ontology based on our framework. We
232 have defined terms, identified synonyms and established hierarchical
233 relationships between all curated terms and categories. Integration of the
234 ontology into any search interface will enable users to perform more powerful
235 and precise ancestry-related queries[21]. We aim to integrate it into the
236 GWAS Catalog website in the near future. The ancestry ontology[22] can be
237 browsed and downloaded (manuscript in preparation).

238

239 Application of framework to assess changes in diversity

240 Several members of the community have called for greater efforts to increase
241 diversity in genomics studies[10]. However, it is difficult to assess progress or
242 identify concrete areas for improvement without a method to easily track
243 changes in ancestry data. Our framework addresses this challenge by
244 requiring the use of specific categories when describing ancestry. This
245 process establishes hierarchical relationships between populations, thus
246 facilitating reproducible tracking of changes in standardized categories over
247 time. A number of authors[9,10] have reviewed the ancestry distribution in the
248 Catalog, but focused exclusively on the detailed descriptions, which are

249 heterogeneous as they are based on the authors' language. Here we present
250 the first analyses using our ancestry categories and demonstrate the validity
251 of our framework to facilitate the tracking of ancestry in big data sets.

252

253 As expected, similar to previous reports[10], we found that the majority (78%)
254 of individuals in the Catalog are exclusively of European ancestry (Figure 2a).
255 The second largest group includes individuals of Asian descent (11%), with
256 East Asians comprising 9% of the Catalog's samples. The disproportionate
257 focus on Europeans was more prevalent in the earlier years of the Catalog
258 (86% of individuals in studies published between 2005 and 2010; 76%
259 between 2011 and 2016, Figure 3). The reduced proportion of European
260 ancestry individuals added to the Catalog in the last 5 years reflects an
261 increase in Asian (7.5% to 11.4%, 1.5-fold increase), African (0.8% to 2.8%,
262 3.5-fold increase), Hispanic or Latin American (0.1% to 1.2%, 9-fold increase)
263 and Middle Eastern (0.01% to 0.08%, 7-fold increase) samples. Though the
264 proportion of Hispanic or Latin Americans exhibited the largest increase, when
265 considering the absolute number of individuals, the largest increase, by far,
266 came from Asian populations; Asian ancestry individuals increased from
267 almost 900,000 in the first 5 years to over 7 million added to the Catalog in the
268 last 5 years, compared to an increase of 721,000 Hispanic or Latin
269 Americans.

270

271 A similar trend is observed when analyzing the ancestry distribution in
272 independent GWA studies. Approximately 50% of all studies are performed on
273 exclusively European ancestry individuals, with an additional 24% of studies

including some samples of European descent (Figure 2b). A more granular analysis of the traits with the largest number of GWAS 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. 3). In studies that analyzed multiple ancestries, the vast majority (> 90%) include some European ancestry individuals, regardless of the trait. The traits that display the largest 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. 3).

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 or Latin American samples; African ancestries contribute 2.4% of individuals but 7% of associations, while Hispanic/Latin Americans contribute 1.3% of individuals compared to 4.3% of associations. The opposite effect was seen in Europeans, with 54% of associations compared to 78% 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.

Recommendations to authors

299 The analysis of the over 3,000 GWAS publications revealed inconsistent and
 300 ambiguous reporting of ancestry data, with a significant percentage of studies
 301 (~ 4%) not reporting any ancestry information at all. Given that there are no
 302 established guidelines for the description of ancestry, and in an effort to assist
 303 the community as it seeks to improve in this area, we here provide a set of
 304 specific recommendations for authors, also summarized in Box 1. We believe
 305 implementation of these recommendations will improve the quality of reporting
 306 and have a positive impact on the interpretation of published results, data re-
 307 use and reproducibility.

308

309 We recommend that authors make every effort to generate a detailed
 310 description for each distinct set of individuals included in their studies. Authors
 311 should also assess whether the genetic diversity of each distinct set is
 312 representative of one of the known populations listed and defined in Table 1,
 313 and indicate the corresponding category in the publication. If authors have no
 314 knowledge about the ancestry of the participants, are not able to determine it
 315 or cannot share it due to confidentiality concerns, we suggest noting this
 316 explicitly in the publication.

317

318 Where possible authors should provide genetically inferred, in addition to self-
 319 reported, ancestry information as the latter is often not an accurate
 320 representation of the underlying genetic background. Software to assess and
 321 control for ancestry is readily available and computationally feasible[23–28].
 322 Encouragingly the proportion of studies that assess ancestry by genetic
 323 methods has increased, from 25% in the early days of GWAS (the first 100

324 eligible publications, 2005-2008) to 57% in 2016 (the first 100 eligible
325 publications; Supplementary Fig. 4). Thus, we suggest authors consider using
326 the methods listed in Supplementary Box 1 when inferring the ancestry of
327 samples.

328

329 In general, terms that pertain to an individual's ethno-cultural background
330 should be avoided, unless this provides additional information regarding the
331 genealogy of the samples. In such cases a descriptor that accurately reflects
332 the underlying genetics should also be provided. For example, when
333 describing "Punjabi Sikh" participants, every effort should be made to assess
334 the genetic background and to indicate this in the publication, for example by
335 stating "Punjabi Sikh South Asian ancestry individuals" rather than simply
336 "Punjabi Sikh" or "Sikh".

337

338 Particular care should be taken to note if a sample derives from founder or
339 genetically isolated population; given their homogeneity and reduced genetic
340 variation, these populations are especially well-suited for GWAS[29] and are
341 increasingly used as sample sources. To reduce ambiguity, when describing
342 isolates, the broader genetic background within which the population clusters
343 should also be indicated. For example, Old Order Amish participants should
344 be described as "Old Order Amish population isolate individuals of European
345 ancestry". While describing admixed populations can be challenging due to
346 varying levels of admixture, every effort should be made to explicitly note
347 whether the population is admixed and the ancestral backgrounds that
348 contribute to admixture. For example "Hispanics/Latinos are ethnically

heterogeneous, with admixture of European, West African, and Amerindian
ancestral populations”, as stated in Hodonsky 2017[30].

Discussion

Summary

In this article we describe a framework for the standardized representation of
ancestry data from genomics studies. Our method provides structure to
unstructured data, enabling robust searching across large datasets and
integration across resources. We have established validity by application to
the over 3,000 GWAS publications currently in the GWAS Catalog and
performed a detailed analysis. These data represent over 83 million
individuals from diverse ancestral backgrounds, and, as such, it is likely one
of the largest and most widely-used repositories of curated ancestry data. We
have demonstrated relevance to, and integration potential with, other data
types and studies by using our method to describe well-characterized
reference populations, such as the HapMap and 1000 Genomes project
populations. This will greatly facilitate integration of studies involving these
populations with data included in the Catalog, and, indeed, with any other
resource that implements our framework. We display the utility of the ancestry
categories to simplify the tracking of efforts towards diversity, allowing the
identification of gaps and highlighting specific areas for improvement.
Interestingly, in addition to confirming known biases, our category-based
analyses revealed that African and Hispanic or Latin American ancestry
populations contribute a disproportionately high number of associations,

suggesting that analyses including these groups may be more effective at identifying new associations. Finally, stemming from our extensive manual review of publications, we note a lack of current standards with regard to ancestry reporting and offer recommendations to authors to implement when describing their samples. This, we believe, will increase consistency and reduce ambiguity, facilitating the interpretation of results.

Limitations to the framework

There are challenges inherent to both the design of the framework and its application. 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. The United States Census, for example, defines “White” as “a person having origins in any of the original peoples of Europe, the Middle East, or North Africa”[31], even though Middle Eastern and North African populations are known to cluster, in genetic analyses, independently from European ancestry populations. We thus recommend that authors continue to move away from relying solely on self-reported information and, as much as possible, also use genomic mechanisms to infer and describe the ancestry of participants.

We are aware that classifying the global human population is a challenging endeavor and the subject of numerous publications by expert population

399 geneticists. Due to human evolution and patterns of migration, the ancestry of
 400 a particular population is complex and its definition is dependent on time. It is
 401 generally accepted that all modern human populations originated in Africa,
 402 and, due to the relatively small amount of genetic variation between
 403 populations from disparate geographical locations[20], genetic diversity
 404 among modern populations may be more suitably described as a continuum.
 405 However, it is both possible and useful to generate informative groupings.
 406 Reference populations, such as those included in the HapMap and 1000
 407 Genomes Projects, or ancestry informative markers[32] that allow populations
 408 to be distinguished, have been characterized, and methods have been
 409 developed to adjust for population stratification and separate samples into
 410 clusters. In fact, these analyses between and within populations have
 411 demonstrated that clusters identified through genomic methods generally
 412 align with geographical and regional groupings [33]. Taking this into account,
 413 we designed our framework such that our categories closely resemble
 414 classifications currently in use by the community and defined by experts.

415

416 We do not view our categories as exhaustive or static. We envision that as
 417 more cohorts from diverse populations are characterized, there might arise a
 418 need to create additional categories or sub-categories. In addition,
 419 anticipating that admixture is likely to increase in the future, due to migration,
 420 for example, we also created categories to represent known (for example,
 421 “Hispanic or Latin American”) and emerging (for example, “Other admixed
 422 ancestries”) admixed groups. We recognize that classification of admixed
 423 samples is particularly challenging. The degree and type of admixture may

424 vary within the population, and the accuracy of classification requires well-
425 defined reference samples, which are lacking for some groups. As the
426 community moves towards genetically-inferred ancestry descriptions, our
427 categories are likely to become more precise and granular over time.

428

429 We believe there are benefits to utilizing categories. Practically, categories
430 facilitate data integration and allow robust searches, which is an essential
431 component of databases such as the GWAS Catalog. Also, the use of
432 categories can be useful when carrying out further analyses. For example,
433 querying the generalizability of results, identifying ancestry-specific
434 associations or utilizing linkage disequilibrium information from a reference
435 population to identify independent signals. We did not set out to define novel
436 or authoritative global ancestral classifications. Rather, we aim to formalize
437 and encourage the use of existing classifications to increase standardization
438 and improve resource functionality, ultimately enabling more robust scientific
439 analyses.

440

441 Assessing diversity in genomics

442 Several reports have been published urging the scientific community to
443 ensure that individuals from all ancestry backgrounds benefit from advances
444 in the field of genomics[9,10]. Our proposed framework lays out a mechanism
445 for the generation of consistent and comprehensive ancestry descriptions and
446 this, in turn, facilitates the establishment of metrics and ultimately, the tracking
447 of ancestry data over time.

448

449 Our analysis of individuals in GWA studies, using ancestry categories rather
 450 than Catalog detailed descriptions, as carried out in previous studies, confirms
 451 the persistent bias towards inclusion of European ancestry samples and the
 452 modest trend towards increased diversity. Our more robust analysis of
 453 approximately 43,000 associations found a disproportionately larger
 454 proportion of associations derived from African and Hispanic or Latin
 455 American populations, many of which have significant African admixture[34],
 456 than is expected based on the proportion of individuals. We suggest that the
 457 higher degree of genetic diversity and reduced linkage disequilibrium (LD) in
 458 African populations[35] offers an explanation for this result. Shorter LD blocks
 459 in African populations facilitate the separation of nearby but independent
 460 signals in a way that is more challenging in European populations, in which
 461 LD blocks tend to be longer. Additionally, the inclusion of larger numbers of
 462 individuals from African or Hispanic or Latin American populations allows for
 463 the identification of variants specific to these ancestries. Together, these
 464 observations suggest that utilizing samples from diverse populations for
 465 genomic studies may be advantageous and yield increased and more
 466 comprehensive results.

467
 468 There are limitations to our analysis. First, considering that some cohorts
 469 have been included in numerous GWAS, it is highly likely that some
 470 individuals are represented multiple times in the Catalog. The impact of this is
 471 the skewing of results towards commonly-used or publicly available cohorts,
 472 which are perhaps likely to be of European or Asian ancestry. Another
 473 limitation stems from our criteria for inclusion of associations. Since we only

474 include variants with a p-value $< 1 \times 10^{-5}$ and only the “index” variant at each
 475 locus, our analysis does not take into account all associations. To address
 476 this and make the Catalog more comprehensive, we now also include in the
 477 Catalog published summary statistics[36]. Finally, we were unable to assign a
 478 category to associations identified in studies that include multiple ancestries.
 479 This may be a factor contributing to the reduced number of associations
 480 derived from European populations, since the vast majority of multiple
 481 ancestry studies include Europeans (Figure 2b).

482

483 The analysis of diverse ancestries is advantageous from a scientific
 484 perspective. No one population contains all human variants[37], and alleles
 485 that are rare in one population may be common in a different population and
 486 thus easier to detect. Studies of diverse populations may also aid in fine
 487 mapping of existing signals or in identifying population-specific functional
 488 variation[37,38]. Variant interpretation for genomic medicine in ancestrally
 489 diverse or admixed populations relies on the availability of non-European
 490 allele frequencies, with potentially serious clinical consequences if such data
 491 are not available[12]. Finally, disease burden of common or complex diseases
 492 (e.g., cardiovascular disease or cancer) disproportionately impacts non-
 493 European populations. Of the commonly studied traits, the largest diversity of
 494 backgrounds was found for common anthropometric traits, heart disease, and
 495 type 2 diabetes. This is perhaps not surprising considering that metrics for
 496 these traits are easy to obtain, and the two diseases are among the top ten
 497 causes of death around the world, according to the World Health
 498 Organization[39]. 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. 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.

Recommendations to authors

Our analysis also validates the need for a framework to improve the description of ancestry. Approximately 5.8% of individuals in the Catalog (2005 - 2016) are currently labeled with 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 recommendations to increase standardization of ancestry reporting, with an emphasis on genetically-inferred ancestry, in publications (Box 1). We encourage implementation by authors reporting ancestry data and by editors reviewing publications that include human subjects.

Conclusions

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

524 interpretation and generalizability of results across populations. The
525 framework we provide aims to address these challenges. It improves
526 standardization of ancestry, increases integration of data, supports the
527 assessment of diversity in large sets and facilitates further analyses. Its
528 widespread adoption will enable the scientific community to investigate the
529 generalizability of trait-associations across diverse populations, to identify
530 associations unique to specific ancestries, to identify novel variants with
531 clinical implications, and to help pinpoint causative variants, thus increasing
532 our understanding of common diseases.

533

534 **Methods**

535

536 GWAS Catalog data curation

537 Details of GWAS publication identification, GWAS Catalog eligibility criteria
538 and curation methods can be found on the GWAS Catalog website[40].
539 Extracted information encompasses publication information, study cohort
540 information, including ancestry, and SNP-trait association results. Curation of
541 ancestry data from the literature was performed according to Ancestry
542 Extraction Guidelines outlined in the Supplementary Note.

543

544 1000 Genomes and HapMap Project population ancestry assignment

545 Information describing the 1000 Genomes[20] phase 3 and HapMap
546 Project[19] phase 3 populations was taken from the Coriell Institute
547 website[41]. Ancestry information, including ancestry category, country of
548 recruitment, country of origin and additional information, was assigned to each

549 population following the GWAS Catalog ancestry extraction guidelines
550 (Supplementary Note).

551

552 GWAS Catalog ancestry analysis

553 To determine the distribution of individuals, associations and traits by ancestry
554 category, we first downloaded all Catalog data in tabular form[14]. All data
555 (gwas-catalog-associations_ontology-annotated.tsv, gwas-catalog-
556 ancestry.tsv, gwas-catalog-studies_ontology-associated.tsv, gwas-efo-trait-
557 mappings.tsv) included in these analyses were curated from GWA studies
558 published between 2005 and the end of 2015, with a release date of July 18
559 2017. The data can be found on the Catalog's FTP site[42].

560

561 Analysis of ancestry assessment methods in a subset of the GWAS Catalog

562 We selected the first 100 publications included in the Catalog (approximately
563 covering the period between March 2005 to January 2008), and for
564 comparison, the first 100 publications from 2016. For each publication, the
565 method was assessed and classified into one of the following: 1. Self-
566 reported, 2. Genetically assessed, 3. Ancestry stated without method, 4.
567 Inferred from limited ancestry-related information (e.g. country information), 5.
568 No ancestry information reported and 6. Mixed method (when a combination
569 of methods was utilized to describe the study samples). Publications classified
570 as "Genetically assessed" includes those where the author had clearly
571 identified the genetic ancestry or admixture of the population, for example by
572 using methods such as those described in Supplementary Box 1. It also
573 includes those that confirmed self-reported information or defined samples

574 based on self-reports but then excluded genetic outliers. Publications where
 575 no ancestry was stated, but curators inferred an ancestry based on country
 576 information are included in the fourth classification. In many cases authors
 577 used a statistical method to assess or control for ancestry or population
 578 stratification, without assigning individuals to a particular category, for
 579 example using a continuous axis of genetic variation from PCA to compute
 580 the association statistic. However, since this did not add any information that
 581 curators could use to assign a population ancestry to the study, it was not
 582 included under category 2.

583

584 **Declarations.**

585 Ethics approval and consent to participate

586 Not Applicable

587

588 Consent for publication

589 Not Applicable

590

591 Availability of data and materials

592 The datasets generated and/or analyzed during the current study are
 593 available on the NHGRI-EBI GWAS Catalog search interface[4] and in
 594 spreadsheet form[14].

595

596 Competing interests

597 PF is a member of the Scientific Advisory Board of Omicia, Inc.

598

599 Funding

600 Research reported in this publication was supported by the National
601 Human Genome Research Institute and the National Institute of General
602 Medical Sciences of the National Institutes of Health under Award Numbers
603 U41-HG007823 and U41-HG006104. The content is solely the responsibility
604 of the authors and does not necessarily represent the official views of the
605 National Institutes of Health. This research was also supported by the
606 European Molecular Biology Laboratory. L.A.H., P.H. and H.J. are employees
607 of the National Human Genome Research Institute.

608

609 Authors' contributions

610 J.M., J.A.L.M., P.H. H.A.J. and L.A.H. conceived this study and
611 developed the ancestry framework. J.M., J.A.L.M., E.H.B., A.B., M.C., P.H.,
612 L.W.H., H.A.J., A.C.M., A.M. and L.A.H. performed curation of ancestry data
613 of GWAS Catalog publications. J.M., J.A.L.M., M.C., T.B. and L.A.H. analyzed
614 the distribution of ancestry categories in the Catalog and interpreted the data.
615 L.W.H, J.A.L.M., L.H. and J.M. assessed the methods of ancestry
616 determination utilized in GWAS Catalog studies and interpreted the data.
617 A.C.M. and J.M. generated the figures. J.M., J.A.L.M and L.W.H. generated
618 the Tables. E.H., D.W., C.M. and T.B. developed the GWAS Catalog curation
619 and search interfaces. D.W. created the ancestry ontology, with contributions
620 from J.M., J.A.L.M. and E.H.B. All authors contributed to the final manuscript,
621 with J.M., J.A.L.M. and L.A.H. playing the key roles.

622 Acknowledgements

623 The authors wish to thank all GWAS Catalog users and authors of
624 studies included in the Catalog. We also thank Chris Gignoux for his expert
625 review of the genomic methods of ancestry determination discussed in this
626 manuscript and to Teri Manolio for valuable discussion.

627

628 **References**

629

630 1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al.
631 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J.
632 Hum. Genet. 2017;101:5–22.

633 2. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new
634 NHGRI-EBI Catalog of published genome-wide association studies (GWAS
635 Catalog). Nucleic Acids Res. 2017;45:D896–901.

636 3. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The
637 NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic
638 Acids Res. 2014;42:D1001-1006.

639 4. GWAS Catalog [Internet]. [cited 2017 Aug 4]. Available from:

640 <http://www.ebi.ac.uk/gwas/>

641 5. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR,
642 Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and
643 evidence for colocalization of causal variants with lymphoid gene enhancers.
644 Nat. Genet. 2015;47:381–6.

- 645 6. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L,
646 Manolio T, et al. Abundant pleiotropy in human complex diseases and traits.
647 Am. J. Hum. Genet. 2011;89:607–18.
- 648 7. Pal LR, Moulton J. Genetic Basis of Common Human Disease: Insight into the
649 Role of Missense SNPs from Genome-Wide Association Studies. J. Mol. Biol.
650 2015;427:2271–89.
- 651 8. Mullen J, Cockell SJ, Woollard P, Wipat A. An Integrated Data Driven
652 Approach to Drug Repositioning Using Gene-Disease Associations. PLoS
653 One. 2016;11:e0155811.
- 654 9. Need AC, Goldstein DB. Next generation disparities in human genomics:
655 concerns and remedies. Trends Genet. 2009;25:489–94.
- 656 10. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature.
657 2016;538:161–4.
- 658 11. Manolio TA. In Retrospect: A decade of shared genomic associations.
659 Nature. 2017;546:360–1.
- 660 12. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et
661 al. Genetic Misdiagnoses and the Potential for Health Disparities. N. Engl. J.
662 Med. 2016;375:655–65.
- 663 13. Li YR, Keating BJ. Trans-ethnic genome-wide association studies:
664 advantages and challenges of mapping in diverse populations. Genome Med.
665 2014;6:91.

- 666 14. GWAS Catalog [Internet]. [cited 2017 Aug 4]. Available from:
667 <http://www.ebi.ac.uk/gwas/docs/file-downloads>
- 668 15. UNSD — Methodology [Internet]. [cited 2017 Aug 4]. Available from:
669 <https://unstats.un.org/unsd/methodology/m49/>
- 670 16. The World Factbook — Central Intelligence Agency [Internet]. [cited 2017
671 Aug 4]. Available from: [https://www.cia.gov/library/publications/resources/the-](https://www.cia.gov/library/publications/resources/the-world-factbook/index.html)
672 [world-factbook/index.html](https://www.cia.gov/library/publications/resources/the-world-factbook/index.html)
- 673 17. GWAS Catalog [Internet]. [cited 2017 Aug 14]. Available from:
674 <http://www.ebi.ac.uk/gwas/search?query=22391508>
- 675 18. Jiang R, French JE, Stober VP, Kang-Sickel J-CC, Zou F, Nylander-
676 French LA. Single-Nucleotide Polymorphisms Associated with Skin Naphthyl-
677 Keratin Adduct Levels in Workers Exposed to Naphthalene. *Environ. Health*
678 *Perspect.* 2012;120:857–64.
- 679 19. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L,
680 Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation
681 in diverse human populations. *Nature.* 2010;467:52–8.
- 682 20. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM,
683 Garrison EP, Kang HM, et al. A global reference for human genetic variation.
684 *Nature.* 2015;526:68–74.
- 685 21. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The
686 OBO Foundry: coordinated evolution of ontologies to support biomedical data
687 integration. *Nat. Biotechnol.* 2007;25:1251–5.

- 688 22. Ancestry Ontology [Internet]. [cited 2017 Aug 4]. Available from:
689 <http://www.ebi.ac.uk/ols/ontologies/ancestro>
- 690 23. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu MV.
691 An overview of STRUCTURE: applications, parameter settings, and
692 supporting software. *Front. Genet.* 2013;4:98.
- 693 24. Alexander DH, Novembre J, Lange K. Fast model-based estimation of
694 ancestry in unrelated individuals. *Genome Res.* 2009;19:1655–64.
- 695 25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D.
696 Principal components analysis corrects for stratification in genome-wide
697 association studies. *Nat. Genet.* 2006;38:904–9.
- 698 26. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al.
699 Variance component model to account for sample structure in genome-wide
700 association studies. *Nat. Genet.* 2010;42:348–54.
- 701 27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et
702 al. PLINK: a tool set for whole-genome association and population-based
703 linkage analyses. *Am. J. Hum. Genet.* 2007;81:559–75.
- 704 28. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to
705 population stratification in genome-wide association studies. *Nat. Rev. Genet.*
706 2010;11:459–63.
- 707 29. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, Hernandez DG,
708 et al. A genome-wide association study of sporadic ALS in a homogenous
709 Irish population. *Hum. Mol. Genet.* 2008;17:768–74.

- 710 30. Hodonsky CJ, Jain D, Schick UM, Morrison JV, Brown L, McHugh CP, et
711 al. Genome-wide association study of red blood cell traits in
712 Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos.
713 PLoS Genet. 2017;13:e1006760.
- 714 31. Bureau UC. About [Internet]. Available from:
715 <https://www.census.gov/topics/population/race/about.html>
- 716 32. Paschou P, Lewis J, Javed A, Drineas P. Ancestry informative markers for
717 fine-scale individual assignment to worldwide populations. J. Med. Genet.
718 2010;47:835–47.
- 719 33. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al.
720 Genes mirror geography within Europe. Nature. 2008;456:98–101.
- 721 34. Adhikari K, Mendoza-Revilla J, Chacón-Duque JC, Fuentes-Guajardo M,
722 Ruiz-Linares A. Admixture in Latin America. Curr. Opin. Genet. Dev.
723 2016;41:106–14.
- 724 35. Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. The peopling of the
725 African continent and the diaspora into the new world. Curr. Opin. Genet. Dev.
726 2014;29:120–32.
- 727 36. GWAS Catalog [Internet]. [cited 2017 Aug 4]. Available from:
728 <http://www.ebi.ac.uk/gwas/downloads/summary-statistics>
- 729 37. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM,
730 Garrison EP, Kang HM, et al. A global reference for human genetic variation.
731 Nature. 2015;526:68–74.

- 732 38. Asimit JL, Hatzikotoulas K, McCarthy M, Morris AP, Zeggini E. Trans-
733 ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet. EJHG*.
734 2016;24:1330–6.
- 735 39. WHO | The top 10 causes of death [Internet]. WHO. Available from:
736 <http://www.who.int/mediacentre/factsheets/fs310/en/>
- 737 40. GWAS Catalog [Internet]. [cited 2017 Aug 4]. Available from:
738 <https://www.ebi.ac.uk/gwas/docs/methods>
- 739 41. Coriell Biorepository [Internet]. [cited 2017 Aug 4]. Available from:
740 <https://catalog.coriell.org/>
- 741 42. Index of /pub/databases/gwas/releases/2017/07/18/ [Internet]. [cited 2017
742 Aug 4]. Available from:
743 <ftp://ftp.ebi.ac.uk/pub/databases/gwas/releases/2017/07/18/>
- 744 43. Huoponen K, Schurr TG, Chen Y, Wallace DC. Mitochondrial DNA
745 variation in an aboriginal Australian population: evidence for genetic isolation
746 and regional differentiation. *Hum. Immunol.* 2001;62:954–69.
- 747 44. Nagle N, Ballantyne KN, van Oven M, Tyler-Smith C, Xue Y, Taylor D, et
748 al. Antiquity and diversity of aboriginal Australian Y-chromosomes. *Am. J.*
749 *Phys. Anthropol.* 2016;159:367–81.
- 750 45. Martínez-Cruz B, Vitalis R, Ségurel L, Austerlitz F, Georges M, Théry S, et
751 al. In the heartland of Eurasia: the multilocus genetic landscape of Central
752 Asian populations. *Eur. J. Hum. Genet. EJHG.* 2011;19:216–23.

- 753 46. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al.
 754 Characterization of Greater Middle Eastern genetic variation for enhanced
 755 disease gene discovery. *Nat. Genet.* 2016;48:1071–6.
- 756 47. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E,
 757 Ortiz-Tello P, et al. Genomic Insights into the Ancestry and Demographic
 758 History of South America. *PLoS Genet.* 2015;11:e1005602.
- 759 48. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al.
 760 Reconstructing Native American population history. *Nature.* 2012;488:370–4.
- 761 49. Kayser M. The human genetic history of Oceania: near and remote views
 762 of dispersal. *Curr. Biol. CB.* 2010;20:R194-201.

763

764

765 **Acknowledgements.**

766 Research reported in this publication was supported by the National Human
 767 Genome Research Institute and the National Institute of General Medical
 768 Sciences of the National Institutes of Health under Award Numbers U41-
 769 HG007823 and U41-HG006104. The content is solely the responsibility of the
 770 authors and does not necessarily represent the official views of the National
 771 Institutes of Health. This research was also supported by the European
 772 Molecular Biology Laboratory. L.A.H., P.H. and H.J. are employees of the
 773 National Human Genome Research Institute. The authors wish to thank all
 774 GWAS Catalog users and authors of studies included in the Catalog. We also
 775 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.

778

779

780 **Figures, tables and additional files**

781

782 **Table 1. Ancestry categories.** Distinct regional population groupings used in this
783 framework. They are assigned to cohorts with distinct and well-defined patterns of
784 genetic variation, in addition to individuals with inferred relatedness to these cohorts.
785 A full list of GWAS Catalog sample descriptions assigned to each category can be
786 found in supplementary table 2.

Ancestry category	Definition	Examples of detailed descriptions for samples included in the category
Aboriginal Australian	Includes individuals who either self-report or have been described by authors as Australian Aboriginal. These are expected to be descendants of early human migration into Australia from Eastern Asia and can be distinguished from other Asian populations by mtDNA and Y chromosome variation[43,44].	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[45]. 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)

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)[46]. 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	Tunisian, Arab, Iranian

	category and efforts are required to fill this gap.	
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[34,47].	Brazilian, Mexican
Native American	Includes indigenous individuals of North, Central and South America, descended from the original human migration into the Americas from Siberia[48]. We note that there does not appear to be a suitable reference population for this category and	Pima Indian, Plains American Indian

	efforts are required to fill this gap.	
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)[49]. 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
Other admixed ancestry	Includes individuals who either self-report or have been described by authors as admixed and do not fit the definition of the other admixed categories already defined ("African American or Afro-Caribbean" or "Hispanic or Latin American").	
South Asian	Includes individuals who either self-report or have been described by authors as South	Bangladeshi, Sri Lankan

	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, PJI and STU populations.	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 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

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

787

788

789 **Box 1. Recommendations for authors reporting ancestry data in**

790 **publications.** These recommendations were generated by expert curators

791 following a detailed review of the over 3,000 GWAS publications included in

792 the Catalog.

793

794 1. Preferentially use genomic methods to assess the ancestry of samples

795 included in the GWAS Catalog. See Box 1 for a description of

796 commonly used methods.

797 2. Indicate whether the background of participants was self-reported,

798 inferred by genomic methods or a combination of both. If genetically

799 inferred, indicate the analytical procedure utilized.

800 3. Provide detailed information for each distinct group of samples,

801 a. Ancestry descriptors should be as granular as possible (e.g.

802 Yoruban instead of Sub-Saharan African, Japanese instead of

803 Asian).

804 b. Avoid using country or citizenship as a substitute for ancestry.

805 c. Avoid using geographic descriptors that are part of a cohort

806 name as a substitute for ancestry (e.g. TwinsUK cannot be

807 assumed to be European ancestry).

808 d. If a population self-identifies using sociocultural descriptors (e.g.

809 Old Order Amish), clearly state the genetic ancestry within which

810 this sub-population falls.

811 e. If samples were derived from an isolated or founder population

812 with limited genetic heterogeneity, clearly state the genetic

- 813 ancestry within which this sub-population falls.
- 814 f. If available, genetic genealogy or ancestry of grandparents or
- 815 parents should be included.
- 816 4. Assign an ancestry category for each distinct group of samples. See
- 817 Table 1 for a list of ancestry categories. Refer to Supplementary Table
- 818 1 for a list of descriptors in use in the Catalog with their category
- 819 assignments.
- 820 5. Provide the sample size for each distinct group of samples included in
- 821 the analysis.
- 822 6. Provide country of recruitment.
- 823 7. If ancestry information is not available due to confidentiality, or any
- 824 other concerns, note this in the publication.
- 825
- 826

827

828

829 **Figures**

830 1. Figure 1 – Representation of ancestry data in the GWAS Catalog

831 search interface

832 2. Figure 2 – Ancestry category distribution in the GWAS Catalog

833 a. Figure 2a - Distribution of individuals by ancestry category

834 b. Figure 2b - Distribution of studies by ancestry category

835 c. Figure 2c - Distribution of associations by ancestry category

836 3. Figure 3 – Distribution of individuals in GWAS Catalog studies

837 published between 2005 – 2010 compared to 2011 – 2016.

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) Detailed description (highlighted in blue) and (2) Ancestry category (highlighted in green). The latter follows the format: sample size, 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).

Search results for 27145994

[Download association results](#)

[Expand all studies](#)

Studies

Author	Date	Journal	Title	Reported trait	Association count
Wang M (PMID: 27145994)	2016-05-05	Nat Commun	Common genetic variation in ETV6 is associated with colorectal cancer susceptibility.	Colorectal cancer	4

Associations

SNP	RAF	p-value	OR	Beta	CI	Region	Location	Functional class	Reported gene(s)	Mapped gene(s)	Reported trait	Study
rs2238126-G		3 x 10 ⁻¹¹	1.17		[1.12-1.23]	12p13.2	chr12:11856807	intron_variant	ETV6	ETV6	Colorectal cancer	Wang M (PMID: 27145994), 2016
rs2238126-G	0.477	3 x 10 ⁻¹⁰ Han Chinese	1.17		1.11-1.23	12p13.2	chr12:11856807	intron_variant	ETV6	ETV6	Colorectal cancer	Wang M (PMID: 27145994), 2016
rs1800468-A	0.1456	4 x 10 ⁻⁷ Han	1.36		1.21-1.53	19p13.2	chr19:41354391	intron_variant	TGFB1	TGFB1	Colorectal	Wang M (PMID: 27145994), 2016

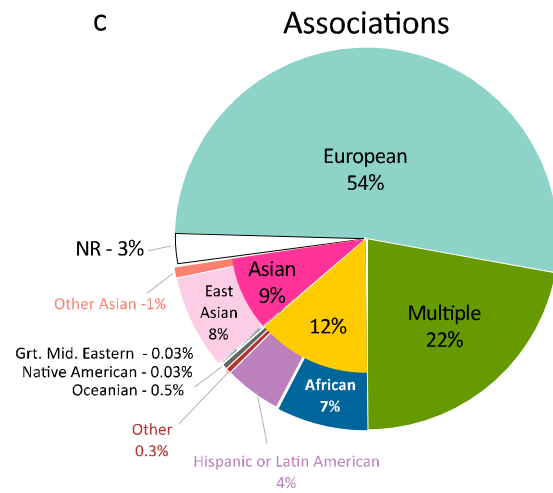
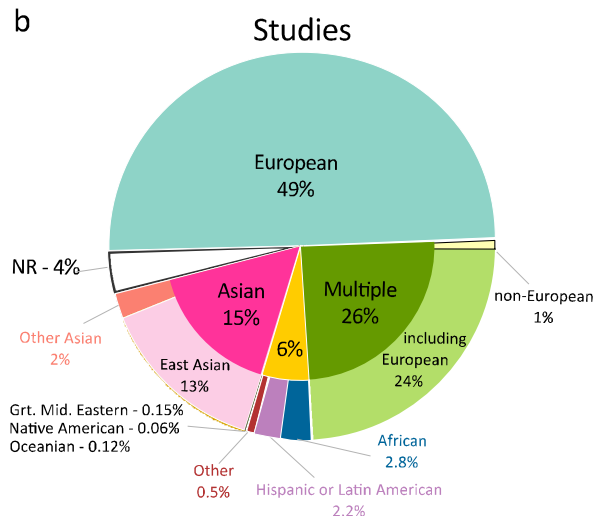
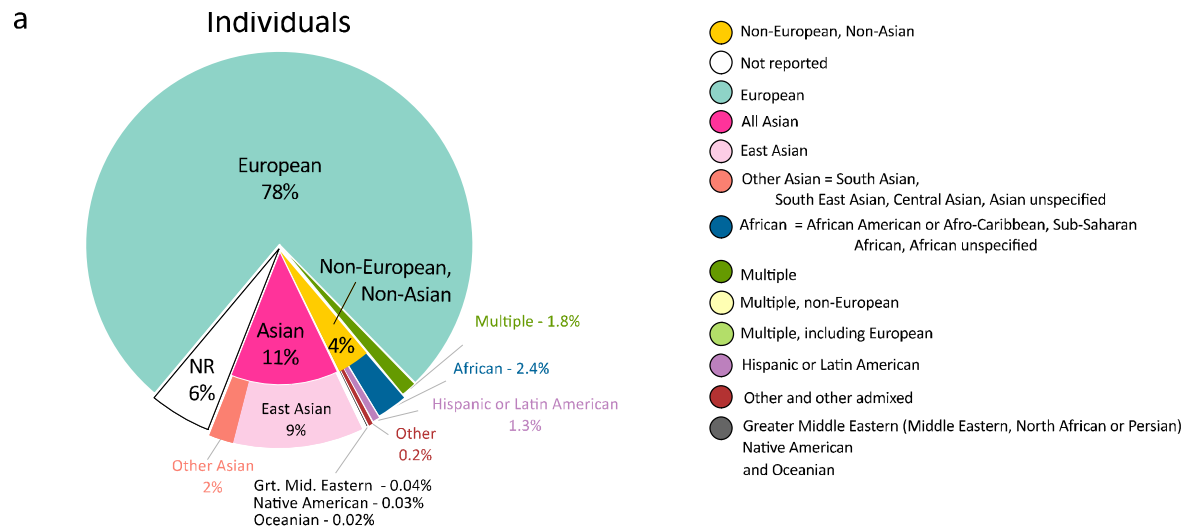
Initial sample description 1,023 Han Chinese ancestry cases and 1,306 Han Chinese ancestry controls

Initial ancestry (country of recruitment) 2329 East Asian (China)

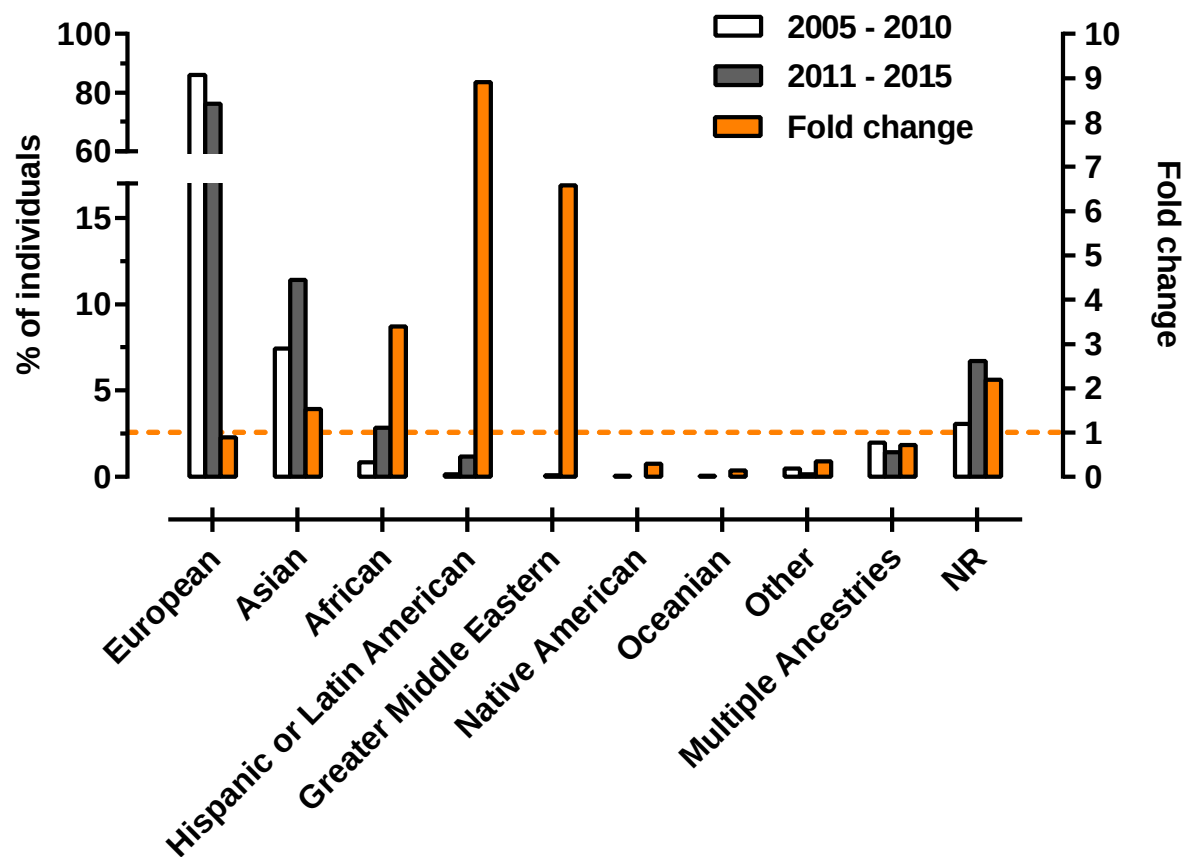
Replication sample description 5,317 Han Chinese ancestry cases, 6,887 Han Chinese ancestry controls, 1,046 European ancestry cases, 1,076 European ancestry controls

Replication ancestry (country of recruitment) 12204 East Asian (China), 2122 European (Canada)

Figure 2. Ancestry category distribution in the GWAS Catalog. This figure summarizes the distribution of ancestry categories in percentages, of individuals (N=83,200,468; panel a), studies (N= 4,100; panel b) and associations (N=43,919; 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 6% (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).



1 **Figure 3. Distribution of individuals in GWAS Catalog studies published**
2 **between 2005 – 2010 compared to 2011 – 2016.** This figure displays the
3 distribution of individuals in percentages, included in the 915 studies
4 published between 2005 – 2010 compared to the distribution of individuals
5 included in the 2,905 studies published between 2011 – 2016.
6



7

8