Supplementary information to: Genomics is failing on diversity (Comment in *Nature* 538, 161–164; 2016)

Alice B. Popejoy & Stephanie M. Fullerton

Introduction

Biomedical research is intertwined with genomics, as investigations of genetic causes for disease unravel the biological mechanisms responsible for wellness and illness. The genome-wide association study (GWAS) has been the primary tool for this discovery phase, and despite its limitation to detecting primarily common variation linked to common diseases, it has been very powerful for elucidating previously opaque biological mechanisms. Thousands of significant associations between genetic variants and phenotypes, or biological traits, have been found through GWAS that have contributed to foundational knowledge in the genetic architecture of disease pathways. Unfortunately, GWAS have been performed using samples from individuals of primarily European descent, and as such have missed a substantial portion of the genetic variation that is present in the human population. In order to develop a comprehensive understanding of the ways that genes and environments interact to determine risk for disease and impact drug responses or adverse drug reactions, it will be necessary to study individuals from a wide variety of social and ancestral backgrounds, living in diverse settings and communities.

A. Querying the GWAS Catalog

To query the GWAS Catalog sample descriptions, A.B. Popejoy wrote a Python script to parse the list and description of studies that can be downloaded from the GWAS Catalog website: https://www.ebi.ac.uk/gwas/docs/downloads. The parser identifies the number of independent publications in the Catalog, and reports the distribution of ancestral populations sampled across all studies. It also produces lists of the number of individuals from each unique study population, and can be tailored to produce lists of PubMed ID numbers of studies with different ancestral populations. The program and documentation are available upon request, but may require manual updates as additional populations are added to the Catalog, as it was designed to parse particular phrases and patterns of phrases included in the current sample descriptions.

B. Sample Descriptions and Ancestry Categories

As sample descriptions in the Catalog are not standardized, decisions needed to be made about how to categorize each sample with respect to broader ancestry bins for the aggregate analysis. These decisions were based on ancestry of origin and/or geographic proximity. For example, Afrikaners from South Africa are included in the "European ancestry" bin because of their European (Dutch) ancestral origins, which is discordant from their current residence on the continent of Africa. Singapore and Malaysia are included in the "Asian ancestry" bin while the Philippines and other island nations are included in the "Pacific Islander" bin based on the geographic connection to or separation from the continental landmass of Asia. A complete list of sample descriptions from the GWAS Catalog and their ancestral bins is provided below.

"African ancestry"

African American
African American ancestry
African ancestry
African cases
African Caribbean ancestry
Afro-Caribbean
Afro-Caribbean cases
Afro-Caribbean controls
Afro-Caribbean individuals
Black cases
Black child cases
Black controls
Black individuals
Gambian ancestry
Malawian ancestry
Moroccan ancestry
Nigerian ancestry
North African ancestry
Seychelles female individuals
Seychelles male individuals
Seychellois ancestry
Sub-Saharan African
Tanzanian ancestry
Tunisian ancestry
West African ancestry
Yoruban ancestry

"Arab/Middle Eastern"

Afghanistan ancestry
Arab ancestry
Arab-Israeli ancestry
Arab-Israeli founder
Iranian ancestry
Jewish Israeli cases
Jewish Israeli controls
Jewish-Israeli ancestry
Lebanese ancestry
Israeli/Arab controls
Israeli/Arab cases
Middle Eastern ancestry
Middle Eastern Arab ancestry
Pakistani ancestry
Saudi Arabian ancestry
Turkish ancestry cases
Turkish ancestry
Turkish cases
Turkish controls
Turkish uveitis cases

"Asian ancestry"

Asian ancestry
Bangladeshi ancestry
Chinese ancestry
Dai Chinese ancestry
Dravidian ancestry
East Asian ancestry
East Asian cases
Han Chinese ancestry
Han Chinese cases
Han Chinese controls
Han Chinese individuals
Hong Kong Chinese ancestry
Hui Chinese ancestry
Indian ancestry
Indian Asian ancestry
Japanese ancestry cases
Japanese ancestry
Japanese controls
Japanese ancestry
Jingpo Chinese ancestry
Korean ancestry
Malay ancestry
Malaysian ancestry
Malaysian Chinese ancestry
Mongolian ancestry
Nepalese ancestry
North Indian ancestry
Oriental ancestry
Punjabi Sikh ancestry
She Chinese ancestry
Silk Road individuals
Singaporean ancestry
Singaporean Chinese ancestry
South Asian ancestry
South East Asian ancestry
South Indian
Southern Indian ancestry
Sri Lankan Sinhalese ancestry
Taiwanese ancestry
Taiwanese
Thai ancestry
Thai-Chinese ancestry
Tibetan ancestry
Uighur cases
Uighur controls
Uygur Chinese ancestry
Uygur-Kazakh Chinese
Vietnamese ancestry
Vietnamese-Korean ancestry

"European ancestry"

Amish cases
Amish controls
Amish individuals
Ashkenazi Jewish cases
Ashkenazi Jewish controls
Ashkenazi
Bulgarian ancestry

Carlantino individuals
Carlantino female individuals
Caucasian Eastern Mediterranean ancestry
Cilento individuals
Erasmus Rucphen Family individuals
Erasmus Ruchpen individuals
European
European American cases
European ancestry
European ancestry individuals
European ancestry cases
European ancestry controls
European cases
European child controls
European controls
European individuals
Finland founder cases
Finland founder controls
Finnish Saami individuals
French Canadian individuals
Friuli Venezia Giulia individuals
Hutterite adult individuals
Hutterite individuals
Italian isolated population individuals
Jewish cases
Jewish controls
Korculan individuals
Korkula individuals
Korkulan individuals
Northern Finnish founder individuals
Old Order Amish individuals
Orcadian female individuals
Orcadian individuals
Romanian founder cases
Russian ancestry
Sardinian cases
Sardinian controls
Sardinian individuals
Sorbian individuals
South African Afrikaner ancestry
Southern European ancestry
Talana adult individuals
Talana individuals
Tatar ancestry
Val Borbera individuals
Vis individuals
Western European ancestry

"Hispanic and Latin American"

Brazilian ancestry
Brazilian individuals
Caribbean Hispanic cases
Caribbean Hispanic controls
Costa Rican ancestry
Dominican Republic ancestry
Hispanic ancestry
Hispanic and unknown ancestry
Hispanic asthmatic individuals
Hispanic cases
Hispanic child cases
Hispanic child controls
Hispanic controls
Hispanic female individuals
Hispanic incident cases
Hispanic individuals
Hispanic male individuals
Hispanic newborn cases
Hispanic newborn controls
Hispanic prevalent cases
Latin American cases
Latin American controls
Latin American individuals
Latino cases
Latino child cases
Latino controls
Latino current smoker
Latino female cases
Latino female controls
Latino individuals
Latino male cases
Latino male controls
Mexican American
Mexican ancestry
Surinamese ancestry

"Native peoples"

American Indian ancestry
Bashkir ancestry
Cape Verdian cases
Martu Australian Aboriginal ancestry
Native Hawaiian ancestry
Native American ancestry
Pima Indian ancestry
Plains American Indian ancestry

"South Pacific Islander"

Filipino ancestry
Filipino female individuals
Filipino male individuals
Kosraen individuals
Micronesian ancestry
Oceania ancestry
Papua New Guinean ancestry
Solomon Islander ancestry

C. Replicating Need & Goldstein (2009)

As in the 2009 analysis conducted by David Goldstein and Anna Need, the numbers and proportions of samples presented here need to be understood not as true individual samples, but as potential replicates from the same cohorts and databases (sampling with replacement, and a growing population to choose from over time). It is possible that a large portion of the GWAS conducted simply re-sampled the same individuals from a few large databases for different studies. As such, the numbers of individuals reported in both 2009 and 2016 represent the number of times any individual from a particular ancestral background was sampled, not the number of times independent individuals were sampled from those populations. As a result of re-sampling the same individuals, there may be statistical problems with results across studies investigating the same phenotypes, but this multiple comparisons issue is outside the scope of this paper.

Table S1. Number and proportion of GWAS participants by ancestral group in 2016 and as reported by Need & Goldstein (2009).

	African ancestry only	Arab & Mid.E.[a] only	Asian ancestry only	European & Jewish Only	Hispanic & L.A.[b] only	Native Peoples only	South Pacific Isl.[c] only	Mixed & multiple groups[d]	Total*
2016									
Group-based No. Studies	58 (2.31%)	3 (0.12%)	349 (13.90%)	1461 (58.18%)	19 (7.57%)	5 (0.20%)	9 (0.36%)	484 (19.28%)	2388 (95.10%)*
Total Samples	1045224 (3.08%)	27040 (0.08%)	4795132 (14.13%)	27435555 (80.82%)	184265 (0.54%)	17929 (0.05%)	94043 (0.28%)	347928 (1.02%)	33947116 (100%)
2009								**Mixed[d] Ancestry**	
Group-based No. Studies	1 (0.27%)	0 (0.00%)	26 (6.97%)	322 (86.33%)	3 (0.80%)	2 (0.54%)	1 (0.27%)	11 (2.95%)	366 (98.12%)*
Total No. Samples	9840 (0.57%)	0 (0.00%)	52877 (3.18%)	1581776 (96.37%)	1019 (0.06%)	1102 (0.06%)	2622 (0.15%)	92437 (5.32%)	1741673 (105.7%)*

a Arab and Middle Eastern; **b** Hispanic and Latin American; **c** South Pacific Islander; **d** Mixed group-based studies refer to studies conducted on multiple single-ancestry groups, including original and replication samples; mixed ancestry samples refer to individuals with mixed (more than one) ancestry.
* Totals are >100% due to samples overlapping ancestral categories, and <100% because studies are compared to the total pool, including those without ancestry information.

Table S1. Legend Details.
In Need & Goldstein's (2009) report, the "Mixed" category describes samples with multiple ethnicities (multi-ethnic individuals), so the total percentage of samples across ancestral categories is greater than 100%. In this [2016] analysis, the "Mixed & multiple groups" category has two meanings: for GWAS (studies), it refers to studies that were done using multiple different ancestral populations and studies that included multi-ethnic individuals. Mixed ancestry samples refer to the proportion of samples that were of mixed ancestry (mirroring the 2009 analysis). Studies in the "Mixed & multiple groups" category sampled multiple single-ethnicity populations, which often included Europeans; and among the multi-ethnic populations sampled, most individuals had European mixed with some other ancestry.

In both analyses (2009, 2016) the studies without ancestry information are excluded from the table and thus the total percentages (<100%) reflect the true total number of studies (2,511 for 2016 and 373 for 2009), not the subset of studies with ancestry information. Inclusion of samples without ancestry information in the total comparison pool could bias the resulting percentages if one particular ancestral group happens to be over-represented in the group of samples without this information. Thus, studies and samples without ancestry information were excluded in the pie chart graphic and table.