Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky415

GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide

Aroon T. Chande[1,2,3], Lu Wang[1,3], Lavanya Rishishwar[1,2,3], Andrew B. Conley[2,3], Emily T. Norris[1,2,3], Augusto Valderrama-Aguirre[3,4] and I. King Jordan[1,2,3,*]

[1]School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA 30332, USA, [2]IHRC-Georgia Tech Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA, [3]PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia and [4]Biomedical Research Institute, Faculty of Health, Universidad Libre-Seccional Cali. Cali, Valle del Cauca, Colombia

Received January 31, 2018; Revised May 01, 2018; Editorial Decision May 02, 2018; Accepted May 03, 2018

ABSTRACT

Human populations from around the world show striking phenotypic variation across a wide variety of traits. Genome-wide association studies (GWAS) are used to uncover genetic variants that influence the expression of heritable human traits; accordingly, population-specific distributions of GWAS-implicated variants may shed light on the genetic basis of human phenotypic diversity. With this in mind, we developed the GlobAl Distribution of GEnetic Traits web server (GADGET http://gadget.biosci.gatech.edu). The GADGET web server provides users with a dynamic visual platform for exploring the relationship between worldwide genetic diversity and the genetic architecture underlying numerous human phenotypes. GADGET integrates trait-implicated single nucleotide polymorphisms (SNPs) from GWAS, with population genetic data from the 1000 Genomes Project, to calculate genome-wide polygenic trait scores (PTS) for 818 phenotypes in 2504 individual genomes. Population-specific distributions of PTS are shown for 26 human populations across 5 continental population groups, with traits ordered based on the extent of variation observed among populations. Users of GADGET can also upload custom trait SNP sets to visualize global PTS distributions for their own traits of interest.

INTRODUCTION

All human traits that have been measured thus far show evidence for some amount of heritability (1). The expression of heritable traits is influenced, to varying degrees, by the presence of specific genetic variants. Since the frequencies of most genetic variants are known to vary among human populations, heritable traits may be expected to differ across populations as well. Indeed, human populations around the world show tremendous variation for a wide variety of heritable traits.

Genome-wide association studies (GWAS) can shed light on the genetic architecture underlying heritable human traits. Over the last ten or so years, numerous GWAS studies have been used to discover thousands of genetic variants that influence the expression of hundreds of human traits, including anthropomorphic, behavioral, and health-related phenotypes (2). Exploration of the distribution of GWAS implicated variants across global populations has the potential to yield insight into the genetic basis of human phenotypic variation.

Heritable human traits are complex and polygenic; they are influenced by the action of genetic variants at multiple loci throughout the genome, along with environmental factors. Recently, genome-wide polygenic trait scores (PTS) have emerged as a powerful tool for predicting individuals' phenotypes based on the numbers of effect (risk) alleles encoded in their genomes (3–5). PTS can be computed by summing the numbers of effect alleles encoded in an individual genome, and scores can be weighted by considering allele effect sizes. In the case of health-related phenotypes, PTS are often referred to as genetic risk scores, reflecting the predicted health risk to individuals entailed by the presence of disease-implicated variants in their genomes. We reasoned that calculation of PTS for different human population groups could be used to shed light on population-specific variation for heritable human traits. To this end, we developed the GlobAl Distribution of GEnetic Traits (GADGET) web server, providing users with an intuitive tool for exploring the relationship between worldwide genetic diversity and the genomic architecture underlying a wide variety of human phenotypes (Figure 1). The GADGET web server allows users to explore the population-specific distributions of pre-computed PTS for >800 human traits across 26 global populations. Users also have the op-

*To whom correspondence should be addressed. Tel: +1 404 385 2224; Email: king.jordan@biology.gatech.edu

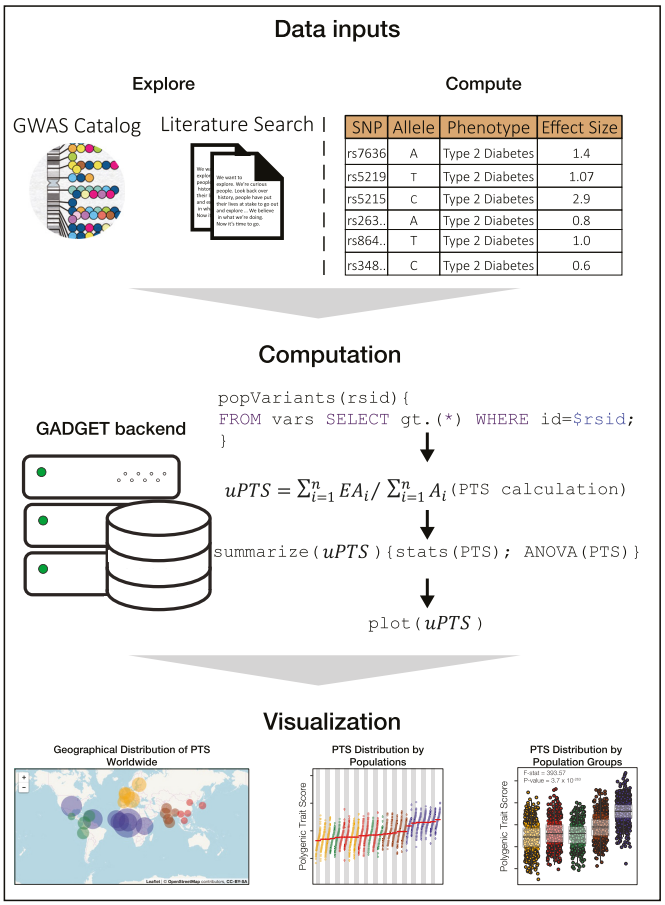

Figure 1. Schematic overview of the GADGET web server workflow. Curated (Explore mode) or user-provided (compute mode) trait SNP sets are used to calculate genome-wide PTS. Population-specific PTS distributions are shown for 26 global populations organized into 5 continental (super) population groups. Among population PTS variation is quantified with ANOVA (*F*-statistics and *P*-values provided).

tion to upload custom SNP sets in order to assess the global distribution of PTS for their own traits of interest.

It should be noted that GADGET is intended as a tool for researchers to explore population-specific distributions of genetic variants that have been associated with a wide variety of human traits. Users of the site should treat the results with caution, as the interpretation of PTS across populations can be complicated by a number of factors (6). In this sense, the PTS distributions returned by GADGET can perhaps best be considered as working hypotheses, rather than definitive assertions of population-specific differences in genetic traits.

IMPLEMENTATION

Platform

The GlobAl Distribution of GEnetic Traits (GADGET) web server frontend and visualizations are built using the R programming language (https://www.r-project.org) with the Shiny web development package (http://shiny.rstudio.com). The user-editable spreadsheet is provided by rhandsontable (https://github.com/jrowen/rhandsontable), and shinyjs (https://github.com/daattali/shinyjs) provides the R-to-javascript interface for custom tab interactions and input validation. GADGET results are visualized with ggplot2 (http://ggplot2.org) (box plots and strip charts) and leaflet.js (http://leafletjs.com) (maps). The computation backend consists of an SQL database along with companion Perl and Python scripts for data validation and PTS score calculation. In an effort to simplify deployment of the Explore module, pre-computed results are exported to flat-files. GADGET supports sharing of results via the 'Share table' function.

Polygenic trait SNP data

The GADGET web server allows users to (1) explore the global distributions of pre-computed genome-wide polygenic trait scores (PTS) for >800 phenotypes or (2) to compute worldwide PTS for user-defined traits of interest. The pre-computed PTS are based on trait-specific sets of SNPs taken from the NHGRI-EBI GWAS Catalog, whereas the user-defined PTS are derived using user-supplied sets of SNPs corresponding to traits of interest. Both pre-computed and user-defined PTS are calculated using genome sequence variant data for 2504 individual genomes from 26 global populations characterized as part of the 1000 Genomes Project (1KGP).

Polygenic trait and SNP (effect allele) information used to compute PTS were taken from the NHGRI-EBI GWAS Catalog version 1.0.1 (2), and trait descriptors were organized into functionally coherent trait categories using the EBI Experimental Factor Ontology (7) November 2017 data release. We rely on the NHGRI-EBI GWAS Catalog study eligibility criteria and SNP reporting methods (https://www.ebi.ac.uk/gwas/docs/methods) for the purpose of curating SNPs associated with specific traits; accordingly, SNPs are incorporated into our trait-specific SNP sets if they show a genome-wide association *P*-value <1.0e−5. The trait-associated SNPs curated from the NHGRI-EBI GWAS Catalog are all based on dbSNP build 150 and the human genome assembly version GRCh38.p10. For each trait-associated SNP, we record the effect allele reported by the GWAS catalog for the purposes of PTS score calculation, and SNP effect alleles are all indexed to the positive strand.

Two classes of trait-specific SNP sets were curated for subsequent PTS calculation: (i) individual trait SNP sets parsed directly from the NHGRI-EBI GWAS Catalog annotations and (ii) higher order trait SNP sets organized according to the EBI Experimental Factor Ontology. For the individual trait SNP sets, text strings from the NHGRI-EBI GWAS Catalog were first normalized in order to import the data into an SQL database. Primary SNP sets were created by querying the database for each unique entry in the 'DISEASE/TRAIT' column. Resulting subsets were then filtered to remove interaction terms, duplicate variants, and multi-allelic variants that do not have defined effect alleles. Effect and non-effect alleles were swapped as needed to be consistent within a set (e.g. all effect alleles in the 'Type 2 Diabetes' set should increase risk of diabetes). Finally, trait SNP sets with fewer than three variants were removed after the other filters were applied.

For the higher order trait SNP sets, the EBI Experimental Factor Ontology was parsed to obtain terminal nodes and subterminal (internal) nodes with high numbers of connected terminal nodes. First, the terminal node trait descriptors were used to create trait SNP sets from the filtered NHGRI-EBI GWAS Catalog annotations described in the previous paragraph. The majority of the resulting trait SNP sets were identical to the existing individual trait SNP sets taken directly from the NHGRI-EBI GWAS Catalog. The resulting trait SNP sets that differed from the existing sets were retained as additional sets for subsequent PTS calculation. Second, trait descriptors from the highly connected subterminal nodes were used to create higher order, functionally coherent trait SNP sets from the filtered NHGRI-EBI GWAS Catalog annotations. The rationale for this approach was to maximize the ability to calculate PTS for the traits reported in the NHGRI-EBI GWAS Catalog. For example, there are a number of GWAS studies for which the NHGRI-EBI GWAS Catalog only reports a single SNP for a given trait, and thus do not yield sufficient information for PTS calculation. Combining identical or related traits into more densely populated higher order SNP sets allows for greater trait coverage of the GWAS Catalog for the purpose of PTS calculations.

All of the trait descriptors parsed from the Experimental Factor Ontology were hierarchically organized into a custom ontology containing 818 discrete traits, a browsable visualization of which is available online at: https://gadget.biosci.gatech.edu/ontology.html. This custom ontology was created to yield a simplified and more intuitive organizational scheme for human phenotypes, which we used to classify our trait SNP sets into 11 functionally related categories for visualizing PTS results: aging, brain health and disorders, cancer, diabetes, general health, heart and health disorders, immune disease and disorders, miscellaneous, obesity, pulmonology, and reproductive health.

Individual and population-specific SNP variant data

The GADGET web server uses publicly available genotype data from the 1000 Genomes Project (1KGP) Phase 3 data release (8) to compute genome-wide PTS for 2504 individuals from 26 worldwide populations, which are organized into five continental (or super) population groups according to the 1KGP scheme: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). SNP variant genotype data for these individuals were downloaded as VCF files from the 1KGP website at: http://www.internationalgenome.org/data. The VCF files were processed to remove SNP variants with >5% missingness, and the remaining variants were annotated with SnpEff (9). A customized version of the Gemini v0.20.0 application (10) was used to import the resulting filtered and annotated VCF files into a SQLite3 database. The same database was populated with trait and GWAS citation information for all of the trait SNP sets created as previously described. The resulting combined database is queryable by chromosomal position, rsID, gene symbol, trait name and PMID.

Genome-wide polygenic trait scores

Genome-wide PTS are calculated for individual genome sequences from the processed 1KGP SNP variant data using the curated trait SNP sets described previously (Explore mode) or with user-supplied SNP sets that correspond to traits of interest (Compute mode). In the Explore mode, unweighted PTS (*uPTS*) are calculated as the normalized sums of the number of effect alleles found in the genome for all trait-associated SNPs:

$$uPTS = \sum_{i=1}^{n} EA_i / \sum_{i=1}^{n} A_i \quad (1)$$

where $EA_i \in \{0, 1, 2\}$ are homozygous absent, heterozygous, and homozygote present effect alleles for each trait-associated SNP i, and $A_i \in \{0, 1, 2\}$ are the total number of alleles with basecalls at each SNP i. PTS are only computed for cases where there are at least three SNP positions with basecalls, i.e. when $\sum_{i=1}^{n} A_i \geq 6$, thereby eliminating the possibility of any division by zero error. In the Compute mode, PTS can be computed for user-supplied SNP sets as either unweighted or weighted sums of the number of effect alleles. Weighted PTS (*wPTS*) employ effect size estimates, either odds ratios or β-values, to weight the numbers of observed effect alleles for each trait associated SNP:

$$wPTS = \sum_{i=1}^{n} (EA_i \times es_i) / \sum_{i=1}^{n} A_i \quad (2)$$

where es_i is the SNP-specific effect size estimate.

GADGET USE CASES

Explore mode

In the Explore mode of the GADGET web server, users can visualize the global distributions of genome-wide PTS for 818 polygenic traits organized into 11 phenotypic categories. For each trait, unweighted PTS are calculated for the 2504 individual genomes from the 1KGP, and population-specific PTS distributions are shown for 26 global populations organized into five continental (super) population groups. The resulting population-specific PTS are visualized as scaled circles on a global map as well as population-specific box plots. The area (A) of the circle for each population (i) is computed as: $A_i = \pi r^2$, where $r = 10 \times (2^{PTS_i / \max PTS})$. The among population variance of trait-specific PTS is measured using ANOVA, for the five continental population groups, with F-statistics, P-values and false discovery rate q-values reported in the trait table. A detailed summary of population-specific PTS values along with the results of the ANOVA analyses are provided in the 'Summary statistics' field.

Figure 2 shows an example of the Explore mode output for the trait diisocyanate-induced asthma, which shows the most extreme population-specific PTS distributions for any of the pre-computed traits. Diisocyanates are chemical building blocks used to make a wide array of polyurethane products and represent a ubiquitous environmental contaminant. They are a leading cause of workplace respiratory problems and representative of a large class of environmental triggers for respiratory distress (11,12). Accordingly, diisocyanate-induced asthma has been investigated by

Figure 2. Example output for the GADGET Explore mode. Users can explore global PTS distributions for 818 traits organized into 11 phenotypic categories. The summary table shows traits in descending order of their ANOVA *F*-statistics, measuring the extent of among population PTS variation, alongside their statistical significance values (*P* and *q*). Example results are shown for the highlighted trait diisocyanate-induced asthma. Scaled circles are used to represent population-specific PTS values on a global map. Box-plot PTS distributions are shown for all 26 global populations and for the 5 continental (super) population groups. Users have the option to view all the SNPs and effect alleles used to compute PTS for the displayed trait.

Figure 3. Example output for the GADGET Compute mode. Users can supply their own trait SNP sets for PTS calculation and global PTS distribution visualization. An example trait SNP table, for acute kidney disease, is shown here. This trait is broken down into three phenotypes based on the ancestry-origin of the GWAS SNPs used for PTS calculation. PTS are calculated for all phenotypes, and users can explore each phenotype individually. As with the pre-computed PTS shown in the explore mode, PTS calculated from user-supplied SNP sets are visualized on a global map and as population-specific box plots. ANOVA statistics are shown on the plot for the five continental (super) population groups.

GWAS in an effort to elucidate the genetic architecture of environmentally triggered asthma (13). Results generated by the GADGET web server show that individuals from African populations have far higher genetic risk for environmentally triggered asthma than any other population group, as measured by their diisocyanate-induced asthma PTS. The East Asian and Admixed American population groups, which show similar diisocyanate-induced asthma PTS distributions, have the next highest genetic risk profiles for this trait, whereas European populations show uniformly low diisocyanate-induced asthma PTS.

These PTS distributions reflect known health disparities for asthma, underscoring the potential utility of comparing PTS across global population groups for investigating the genetic basis of population-specific health outcomes. The results are consistent with previous work showing a relationship between African genetic ancestry and asthma risk in African Americans (14). Furthermore, in the United States, African-Americans have the highest prevalence of environmentally triggered asthma followed by Hispanics and East Asians, with European Americans showing relatively low levels of asthma (https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=15) (15).

Compute mode

In the Compute mode of the GADGET web server, users can supply their own SNP sets in order to analyze global PTS distributions for their traits of interest. The required fields for user-supplied trait SNP tables are: rsIDs, the identity of the effect allele, trait name, and effect size estimates. PTS for the 1KGP individuals and populations can be computed as unweighted or weighted, and users can supply SNP sets for one or more traits of interest in a single file. The SNP set input file format requirements are specified on the website along with an example SNP table that can be downloaded and/or run on the server.

Figure 3 shows the Compute mode output for acute kidney disease based on the example SNP table that is found on the website. These SNPs were curated from a trans-ethnic meta-analysis of five acute kidney disease GWAS, wherein SNP effects were inferred separately for African, European and Native American ancestry groups (16). The example input SNP table for this trait considers SNP effects

separately for African-American (AfrAm), American Indian (AmInd), and European American (EurAm) GWAS-implicated SNPs, following the convention of the original paper, as can be seen in phenotype column labels. Once the PTS are computed for the three distinct SNP sets, users can toggle among the results for each set using the dropdown menu ('Choose a phenotype to explore'). In this way, the extent to which PTS are influenced by the population ancestry of the study subjects in the GWAS can be assessed.

DISCUSSION

Methods for calculating PTS

There are a number of different factors that need to be considered when choosing the specific set of SNPs to be used in PTS calculation for any given trait (BioRxiv: https://www.biorxiv.org/content/early/2017/02/05/106062). The most fundamental decision relates to the number of SNPs to include in a trait set. At the extreme ends of the spectrum, there is the top-SNP approach, whereby only SNPs that reach genome-wide significance are used for either unweighted or weighted score calculation, versus the all-SNP approach, whereby effect sizes are used to weight the phenotypic contributions of all the SNPs that were genotyped in a given study. Between these two extremes, PTS calculation methods can use different GWAS *P*-value thresholds to determine whether SNPs should be included in a trait set. The approach that the GADGET web server uses to calculate PTS can be considered as a soft version of the top-SNP approach, since it employs a fairly stringent *P*-value threshold of 10^{-5}, which is nevertheless far more inclusive than the standard GWAS genome-wide significance threshold of 10^{-8}. Our approach is also distinguished by the fact that it sometimes combines SNPs from multiple GWAS into single trait sets. We have found that this approach provides additional resolution for PTS calculation, based in part on the use of larger numbers of SNPs for PTS calculation. Since the effect sizes between multiple studies may not be directly comparable, the pre-computed PTS reported in the server's explore mode are calculated via the unweighted approach. The option for users to supply their own SNP sets provides more flexibility for the computation of PTS, both with respect to the number of SNPs that can be used as well as the weighting scheme.

Genetic ancestry effects on PTS calculation

The vast majority of GWAS have been conducted in populations with European ancestry (17,18), and the extent to which GWAS-implicated variants replicate across populations remains a matter of contention (19). On the one hand, a number of trans-ethnic studies have shown that the majority of GWAS implicated variants replicate across populations (20–22). This is even true for traits such as type 2 diabetes (23,24), which shows highly population-specific PTS distributions (25,26). Furthermore, while the same tag SNPs may not reach genome-wide significance in distinct populations, the haplotypes that they mark are often found to replicate among populations. Nevertheless, even SNPs that replicate between populations can differ with respect to population-specific effect size and explanatory power. Furthermore, a recent study showed that the effects of demographic history on allele frequencies can reduce the accuracy of PTS calculated among divergent populations; for example, PTS for the highly heritable trait height were found to be unreliable across populations (6). Even GWAS variants that do replicate across populations can show substantial heterogeneity with respect to effect sizes in different populations (BioRxiv: https://www.biorxiv.org/content/early/2017/09/15/188094).

In any case, the results reported by our web server should be interpreted with caution in light of the fact that population-specific PTS will inevitably be generated from SNPs implicated by GWAS on subjects with distinct ancestries. Thus, the PTS distributions that we show may best be considered as hypotheses that can be used to stimulate and guide further investigations. It is also worth noting that, as we illustrated in the example for acute kidney disease, the Compute utility provided on our webserver, whereby users provide their own SNP sets for traits of interest, provides one way to explore whether and how the ancestry of GWAS study subjects influences population-specific distributions of PTS. In addition, the comparison of unweighted and weighted scores for user-supplied SNP sets can be used to evaluate the effect of ancestry-specific effect size estimates on PTS population differences.

Conclusion: Interpreting PTS differences across populations

As mentioned previously, the meaning of PTS differences across human populations very much remains an area of active investigation and there are numerous possible interpretations for such results. It is important to keep these alternative explanations in mind when interpreting the worldwide PTS distributions generated by the GADGET server. Some of the possible explanations for PTS differences among global populations are: (i) the genetic predisposition to the trait differs among populations, (ii) the top SNPs used for the analysis differ among populations, but the overall genetic predisposition for the trait would balance out if additional SNPs were included in the PTS calculation, (iii) the apparent population differences in genetic predisposition for any given trait could disappear due heterogeneous effect sizes among populations, (iv) observed population differences in PTS could be due to stochastic effects related to demographic factors (e.g. genetic drift). These are just some of the possible explanations; the list is by no means exhaustive. In addition, problems with the original GWAS studies or issues with accuracy of the GWAS database used to generate trait-associated SNP sets could also cause problems with global PTS distributions. In light of these caveats, PTS results generated by GADGET should be treated with caution.

Finally, users are cautioned not to use GADGET to draw conclusions regarding the genetic basis of racial differences. GADGET allows for the interrogation of PTS differences across human population groups characterized as part of the 1KGP, which are defined by geographic origin and distinguished by genetic ancestry. We make no attempt to delineate racial groups from these populations, and the extent

to which racial classifications accurately reflect genetic ancestry remains a matter of contention (27–30).

ACKNOWLEDGEMENTS

The authors would like to acknowledge three anonymous reviewers for their helpful comments.

FUNDING

IHRC-Georgia Tech Applied Bioinformatics Laboratory [RF383]; Georgia Tech Bioinformatics Graduate Program. Funding for open access charge: IHRC, Inc.
Conflict of interest statement. None declared.

REFERENCES

1. Polderman,T.J., Benyamin,B., de Leeuw,C.A., Sullivan,P.F., van Bochoven,A., Visscher,P.M. and Posthuma,D. (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.*, **47**, 702–709.
2. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
3. Chatterjee,N., Shi,J. and Garcia-Closas,M. (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, **17**, 392–406.
4. Chatterjee,N., Wheeler,B., Sampson,J., Hartge,P., Chanock,S.J. and Park,J.H. (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, **45**, 400–405.
5. International Schizophrenia, C., Purcell,S.M., Wray,N.R., Stone,J.L., Visscher,P.M., O'Donovan,M.C., Sullivan,P.F. and Sklar,P. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
6. Martin,A.R., Gignoux,C.R., Walters,R.K., Wojcik,G.L., Neale,B.M., Gravel,S., Daly,M.J., Bustamante,C.D. and Kenny,E.E. (2017) Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, **100**, 635–649.
7. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
8. The,1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
9. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, **6**, 80–92.
10. Paila,U., Chapman,B.A., Kirchner,R. and Quinlan,A.R. (2013) GEMINI: Integrative exploration of genetic variation and genome annotations. *PLOS Comput. Biol.*, **9**, e1003153.
11. (2000) From the centers for disease control and prevention. Availability of work-related lung disease surveillance report, 1999. *JAMA*, **283**, 1955.
12. McDonald,J.C., Chen,Y., Zekveld,C. and Cherry,N.M. (2005) Incidence by occupation and industry of acute work related respiratory diseases in the UK, 1992–2001. *Occup. Environ. Med.*, **62**, 836–842.
13. Yucesoy,B., Kaufman,K.M., Lummus,Z.L., Weirauch,M.T., Zhang,G., Cartier,A., Boulet,L.P., Sastre,J., Quirce,S., Tarlo,S.M. *et al.* (2015) Genome-Wide Association Study identifies novel loci associated with Diisocyanate-Induced occupational asthma. *Toxicol. Sci.*, **146**, 192–201.
14. Flores,C., Ma,S.F., Pino-Yanes,M., Wade,M.S., Perez-Mendez,L., Kittles,R.A., Wang,D., Papaiahgari,S., Ford,J.G., Kumar,R. *et al.* (2012) African ancestry is associated with asthma risk in African Americans. *PLoS One*, **7**, e26807.
15. Bryant-Stephens,T. (2009) Asthma disparities in urban environments. *J. Allergy Clin. Immunol.*, **123**, 1199–1206.
16. Iyengar,S.K., Sedor,J.R., Freedman,B.I., Kao,W.H., Kretzler,M., Keller,B.J., Abboud,H.E., Adler,S.G., Best,L.G., Bowden,D.W. *et al.* (2015) Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: family investigation of nephropathy and diabetes (FIND). *PLoS Genet.*, **11**, e1005352.
17. Popejoy,A.B. and Fullerton,S.M. (2016) Genomics is failing on diversity. *Nature*, **538**, 161–164.
18. Need,A.C. and Goldstein,D.B. (2009) Next generation disparities in human genomics: concerns and remedies. *Trends Genet.*, **25**, 489–494.
19. Bustamante,C.D., Burchard,E.G. and De la Vega,F.M. (2011) Genomics for the world. *Nature*, **475**, 163–165.
20. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
21. Carlson,C.S., Matise,T.C., North,K.E., Haiman,C.A., Fesinmeyer,M.D., Buyske,S., Schumacher,F.R., Peters,U., Franceschini,N., Ritchie,M.D. *et al.* (2013) Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.*, **11**, e1001661.
22. Li,Y.R. and Keating,B.J. (2014) Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.*, **6**, 91.
23. Qi,Q., Stilp,A.M., Sofer,T., Moon,J.Y., Hidalgo,B., Szpiro,A.A., Wang,T., Ng,M.C.Y., Guo,X., Consortium, M.E.-a.o.t.D.i.A.A. *et al.* (2017) Genetics of type 2 diabetes in U.S. Hispanic/Latino Individuals: Results from the hispanic community health Study/Study of latinos (HCHS/SOL). *Diabetes*, **66**, 1419–1425.
24. Waters,K.M., Stram,D.O., Hassanein,M.T., Le Marchand,L., Wilkens,L.R., Maskarinec,G., Monroe,K.R., Kolonel,L.N., Altshuler,D., Henderson,B.E. *et al.* (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet.*, **6**, e1001078.
25. Chen,R., Corona,E., Sikora,M., Dudley,J.T., Morgan,A.A., Moreno-Estrada,A., Nilsen,G.B., Ruau,D., Lincoln,S.E., Bustamante,C.D. *et al.* (2012) Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLoS Genet.*, **8**, e1002621.
26. Chande,A.T., Rowell,J., Rishishwar,L., Conley,A.B., Norris,E.T., Valderrama-Aguirre,A., Medina-Rivas,M.A. and Jordan,I.K. (2017) Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia. *Sci. Rep.*, **7**, 17127.
27. Banda,Y., Kvale,M.N., Hoffmann,T.J., Hesselson,S.E., Ranatunga,D., Tang,H., Sabatti,C., Croen,L.A., Dispensa,B.P., Henderson,M. *et al.* (2015) Characterizing Race/Ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics*, **200**, 1285–1295.
28. Burchard,E.G., Ziv,E., Coyle,N., Gomez,S.L., Tang,H., Karter,A.J., Mountain,J.L., Perez-Stable,E.J., Sheppard,D. and Risch,N. (2003) The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.*, **348**, 1170–1175.
29. Yudell,M., Roberts,D., DeSalle,R. and Tishkoff,S. (2016) Science and society. Taking race out of human genetics. *Science*, **351**, 564–565.
30. Reich,D.(2018) *Who We Are and How We Got Here: Ancient DNA and the new science of the human past*. Oxford University Press.