Quantifying multi-ethnic representation in genetic studies of high mortality diseases

Rong Chen, PhD[1], Joel T. Dudley PhD[1], David Ruau PhD[1], Atul J. Butte, MD, PhD[1]
[1] Division of Systems Medicine, Department of Pediatrics,
Stanford University School of Medicine, Stanford, CA

Abstract

Most GWASs were performed using study populations with Caucasian ethnicity or ancestry, and findings from one ethnic subpopulation might not always translate to another. We curated 4,573 genetic studies on 763 human diseases and identified 3,461 disease-susceptible SNPs with genome-wide significance; only 10% of these had been validated in at least two different ethnic populations. SNPs for autoimmune diseases demonstrated the lowest percentage of cross-ethnicity validation. We used the mortality data from the Center for Disease Control and Prevention and identified 19 diseases killing over 10,000 Americans per year that were still lacking publications of even a single cross-ethnic SNP. Fifteen of these diseases had never been studied in large GWAS in non-Caucasian populations, including chronic liver diseases and cirrhosis, leukemia, and non-Hodgkin's lymphoma. Our results demonstrate that diseases killing most Americans are still lacking genetic studies across ethnicities.

Introduction

Genome-wide association studies (GWAS) compare 500,000 to 2 million single nucleotide polymorphisms (SNPs) across diseased and healthy individuals to identify SNPs distinguishing these two groups. In the last decade, over 1000 GWASs had been conducted, resulting in the discovery of thousands of SNPs to be associated with hundreds of disease traits [1,2]. A recent study suggested that 96% of subjects included in current GWASs were from European descendants, or known as Caucasian [3]. However, findings from one ethnic population may not always translate to another. An open question is how many of these disease-associated SNPs discovered from Caucasian populations can be generalized to diverse ethnic groups. The identification of cross-ethnic SNPs is not only critical for our understanding of the genetic basis of disease risk across subpopulations, but also as a mean to efficiently filter out thousands of false positives in current GWASs [4].

Due to limited funding, there is a disparity in the numbers of genetic studies conducted on different diseases. A recent study showed that current levels of NIH disease-specific research funding correlated only modestly with US disease burden, and this correlation had not improved in the last decade [5]. It is important to study diseases with high mortality and morbidity but it is also obviously critical to make sure that diverse ethnic populations representative of the broader US population can benefit from new genome-based findings. The 2010 US census found that majority of US babies were now classified as non-white [6]. The non-Hispanic white segment of the population is expected to decrease to 52.5% in the United States by 2050, with 22.5% Hispanic, 15.7% Black, 10.3% Asian, and 1.1% American Indian comprising the remainder of the national population [7]. For genetic studies to benefit all Americans studies will need to incorporate participants from all of these relevant subpopulations, as SNPs found to be significant across ethnic subpopulations are often the most relevant.

To address this problem, we performed a systematic evaluation of 4,573 reported genetic studies on 763 diseases to identify cross-ethnic SNPs. Cross-ethnic SNPs were defined as being validated across two out of six population groups with $p<5\times10^{-8}$ in each study, including African, Caucasian, Chinese, Indian Asian, Japanese, and South American. We then cross-referenced these findings with the mortality data from the Centers for Disease Control and Prevention (CDC) to identify those specific diseases that kill over 10,000 Americans per year, but are still lacking cross-ethnic SNPs.

Methods

As described previously [1,8], we built Varimed, a manually curated database of human disease-SNP associations from the full text, figures, tables, and supplemental materials of 4,573 human genetics papers. Papers were retrieved for curation using Medical Subject Heading (MeSH) terms for human genetic studies, such as "Polymorphism, Single Nucleotide", "Genetic Predisposition to Disease". For each paper, we recorded more than 100 features including studied broad and narrow phenotypes, studied populations and ethnicities, number of patients in the case

and control groups, p-values, and disease-susceptible risk alleles. With the breadth of this curation process, we believed that we covered the majority of papers relating polymorphisms to human diseases.

We evaluated the genetic findings for each of the 763 diseases covered in the Varimed database, by counting the number of independent cross-ethnic SNPs that had been validated with $p<5\times10^{-8}$ in two or more different population groups, such as African, Caucasian, Chinese, Indian Asian, Japanese, and South American. As an example, we illustrated the method to identify two independent cross-ethnic SNPs from 135 published genetic studies on Rheumatoid Arthritis (RA) in **Figure 1**. Starting from 1,112 SNPs reported as associated with RA, we found 321 SNPs with $p<5\times10^{-8}$. We identified two SNPs that had been replicated in at least two different subpopulations. We defined a pair of SNPs as being in linkage disequilibrium (LD) when their LD $R^2>0.3$ in CEU HapMap data or their genomic distance was within 37,000 base pairs of each other, in the cases when LD data were unavailable. We used 37,000 base pairs as the cutoff because it was the average genomic distance between SNP pairs in LD R^2 within 0.3 and 0.4 in the CEU in HapMap. Then, for each pair of SNPs in LD, we removed the one SNP with fewer number of replicated populations and studies, leading two independent cross-ethnic SNPs. Thus, for RA we identified 135 independent SNPs, 12 independent replicated SNPs, and two independent cross-ethnic SNPs. Similarly, we identified the number of cross-ethnic SNPs for each of 763 diseases in Varimed.

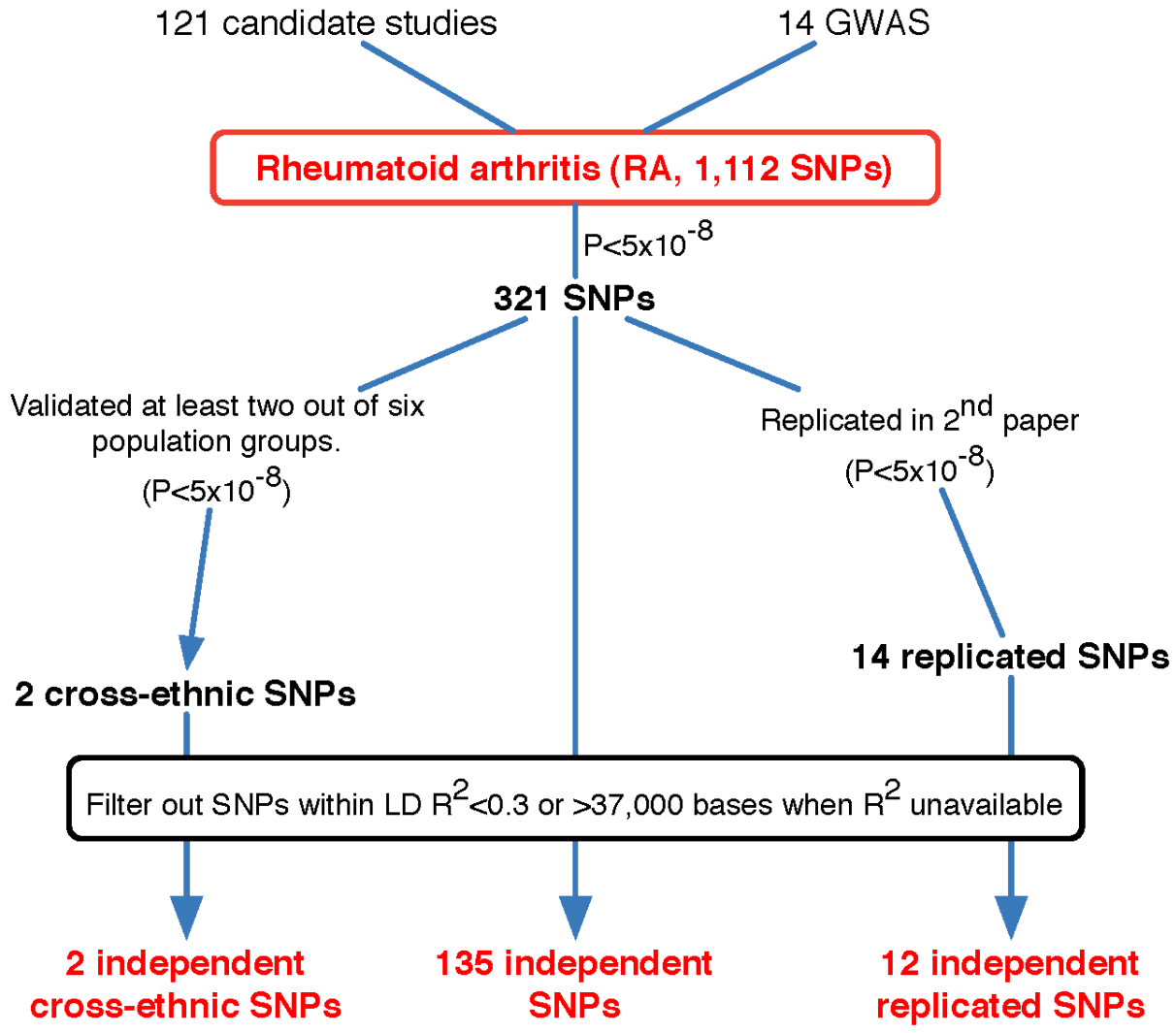

Figure 1: Identification of two independent cross-ethnic SNPs from 135 human genetic studies on rheumatoid arthritis.

We downloaded the number of deaths in the United States in 2009 for 113 causes from the CDC[9], and recorded the ICD-10 codes, disease names, and the mortality data. We then mapped disease names in Varimed and the ICD-10 codes from the CDC using Unified Medical Language System (UMLS) [10] through a three-step approach. First, we identified the Concept Unique IDs (CUIs) for 763 disease names in Varimed using exact match, normalization, and the removal of qualifiers [11]. Then, we mapped each CUI to ICD-10 codes in CDC using three different methods. 1) We tried to directly identify a corresponding ICD-10 code for each CUI, and checked whether it matched or was a child-concept of ICD-10 codes from the CDC. 2) We attempted to identify the parent CUIs for each CUI using the UMLS-Query Perl module[12], searched for ICD-10 codes for these parent CUI , and then checked whether these parent ICD-10 codes matched or were child-concepts of ICD10 codes from the CDC. 3) We manually identified the ICD-10 code for each disease name in Varimed using the ICD-10 Data Service[13], and checked whether it matched or was a child-concept of ICD codes from the CDC. Finally, we manually examined the matches of diseases between Varimed and CDC to validate that these matches were correct. We successfully mapped Varimed diseases to 69 diseases from the CDC.

For each disease category from the CDC, we identified the corresponding diseases in Varimed, and evaluated the genetic findings on these diseases using 1) number of genetic papers, including both GWAS and candidate studies, 2) number of published GWAS papers, 3) number of published GWAS papers with cohort size larger than 1000, 4) number of SNPs with $p<5\times10^{-8}$, 5) number of replicated SNPs, and 6) number of cross-ethnic SNPs. We also distinguished the number of published papers on non-Caucasian populations. Finally, we mapped mortality with genetic findings to identify diseases that killed more than 10,000 American in 2009 but were still lacking cross-ethnic SNPs.

Results

We evaluated the reported genetic findings for 763 diseases using the number of independent cross-ethnic SNPs that had been validated in two or more population groups with $p < 5\times10^{-8}$. We identified diseases that did not have any cross-ethnic SNPs as the diseases with unmet genetic need. We first identified diseases in need for validation by comparing the number of independent SNPs against the number of independent cross-ethnic SNPs across 763 diseases in Varimed. Then, we identified diseases that killed more than 10,000 Americans in 2009 but were still lacking cross-ethnic SNPs as the diseases with unmet medical-genetic need.

Autoimmune diseases were lacking SNPs validated in diverse population groups

We identified 3,461 distinct SNPs validated to be associated with 183 diseases in a single ethnic population with $p<5\times10^{-8}$. Only 398 of them had been validated as cross-ethnic SNPs, increasing the risk of 42 diseases. On average, 10% SNPs with $p<5\times10^{-8}$ for each disease had been validated as cross-ethnic.

We plotted the number of independent cross-ethnic SNPs against the number of independent SNPs for each disease. Surprisingly, these diseases were clearly separated into two distinct classes by the 10% average line (**Figure 2**). Most autoimmune diseases were below the average line with at most two independent cross-ethnic SNPs, while having more than 20 independent SNPs. Only systemic lupus erythematosus deviates from this pattern, with 80 known SNPs with $p<5\times10^{-8}$ and 25 cross-ethnic SNPs. These 25 SNPs were validated from 64 genetic studies on non-Caucasian populations, including six GWAS with cohort size >1000. By contrast, other autoimmune diseases had only been validated in maximum two GWAS with cohort size >1000 on non-Caucasian populations. Validation of these SNPs for these autoimmune diseases in non-Caucasian populations would likely lead to many more cross-ethnic SNPs being identified, but these studies have either not yet been performed or not yet been published.

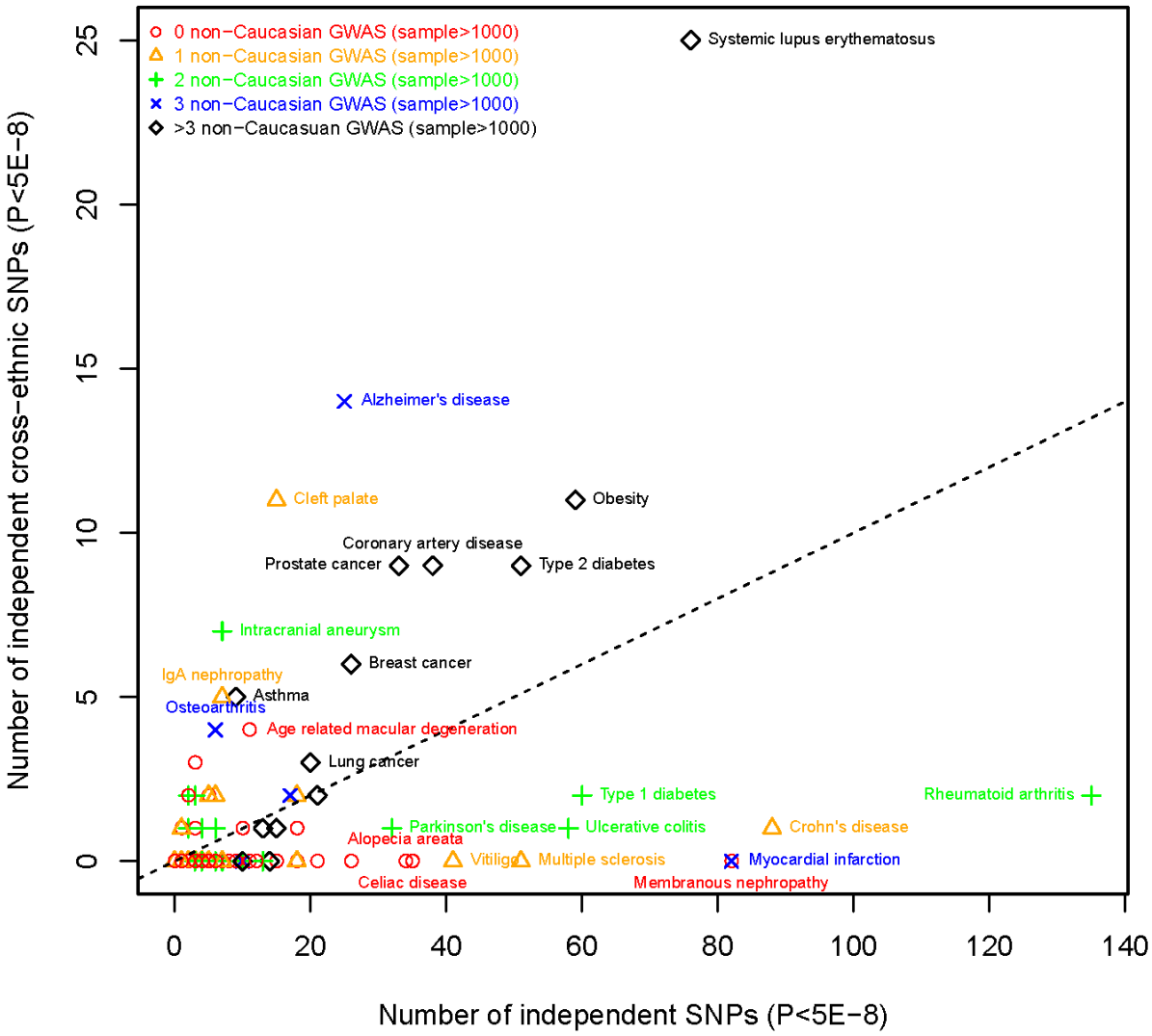

Figure 2: Autoimmune diseases were lacking SNPs validated in diverse population groups. The dotted line represents that only 10% SNPs with $P<5\times10^{-8}$ had been validated as cross-ethnic SNPs, averaged across 763 diseases.

Mortality is only modestly associated with the number of known cross-ethnic SNPs

We used UMLS to map 69 diseases with 2009 mortality data from the CDC to diseases in Varimed, our curated database of published human disease-associated SNPs. Most unmatched diseases are non-genetic causes of death, such as injury, accidents, suicide, and infections. Interestingly, the mortality was only modestly associated with the number of known cross-ethnic SNPs (Pearson coefficient r=0.46, $p=8\times10^{-5}$, **Figure 3**). For example, chronic lower respiratory disease led to the deaths of 137,082 Americans in 2009, and this condition has 5 known cross-ethnic SNPs. Alzheimer's disease killed 78,889 Americans, but has 14 known cross-ethnic SNPs.

However, there were still five diseases that killed more than 100,000 Americans and had fewer cross-ethnic SNPs than the regression line. These five diseases were myocardial infarction (I21-I22), all other forms of heart disease (I26-I28, I34-I38, I42-I49, I51), cerebrovascular diseases (I60-I69), malignant neoplasms of trachea, bronchus, and lung (C33-C34), and ischemic heart disease (ICD10 codes: I20-I25). In other words, these five disease categories had fewer known cross-ethnic disease-associated SNPs than expected.

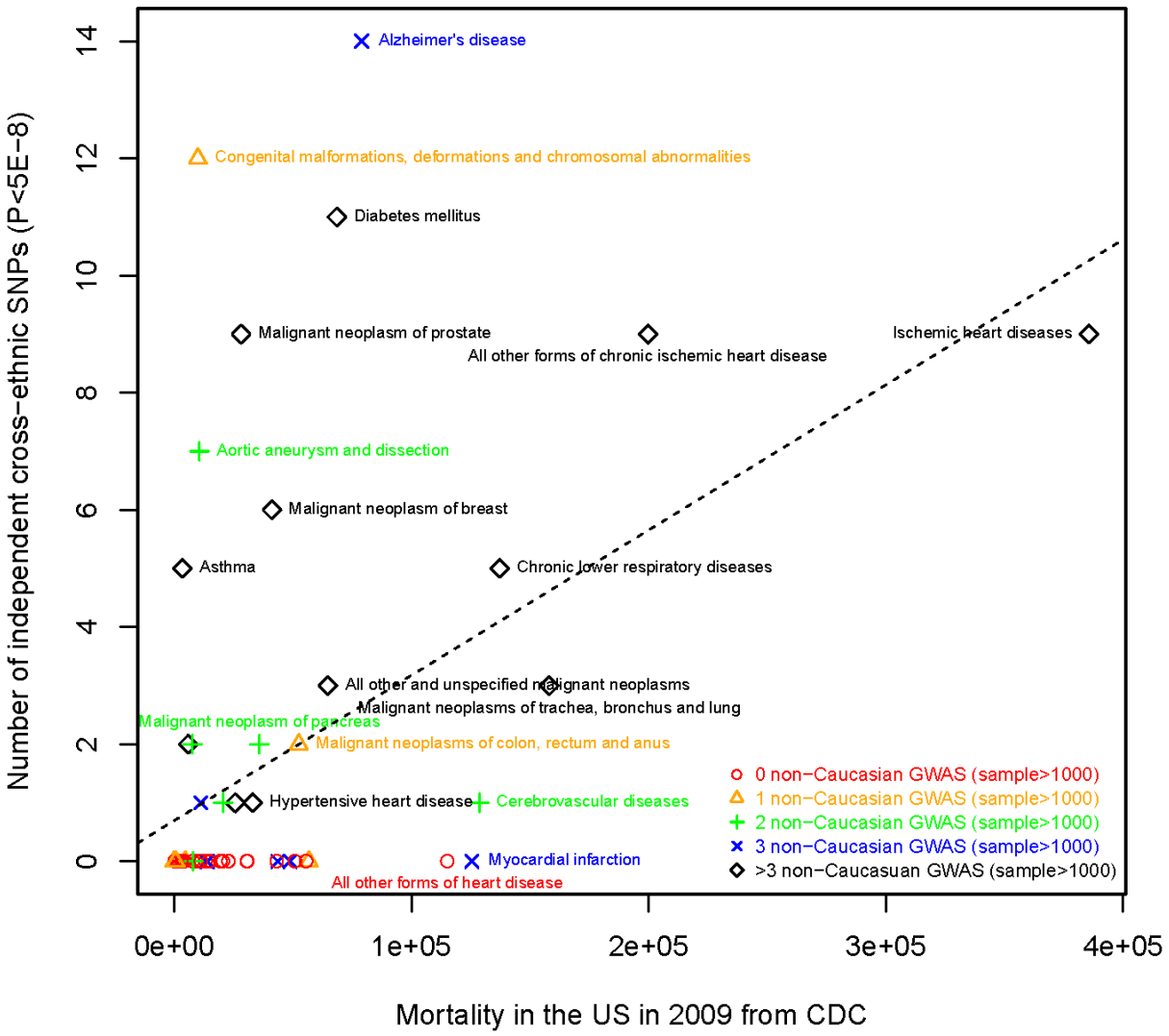

Figure 3: Mortality is only modestly associated with the number of known cross-ethnic SNPs (Pearson coefficient r=0.46, p=8×10^{-5})

Nineteen diseases that killed more than 10,000 Americans in 2009 were still lacking publication of a single cross-ethnic SNP

Surprisingly, we identified 19 diseases that killed more than 10,000 Americans in 2009, but were still lacking even a single cross-ethnic SNP. These 19 diseases were separated into two classes. The first class includes four diseases that had been studied on non-Caucasian in GWAS with cohort size >1000 but these studies did not yield a single cross-ethnic SNP (**Table 1**), including myocardial infarction, heart failure, renal failure, and esophageal cancer. The second class includes 15 diseases that had apparently never been studied using GWAS with cohort size >1000 on non-Caucasian populations, such as chronic liver disease, leukemia, and non-Hodgkin's lymphoma (**Table 2**).

In the first class of diseases, myocardial infarction and renal failure had 15 and 83 SNPs, respectively, with p<5×10^{-8} but none validated in a different subpopulation. For example, myocardial infarction has been studied in 11 GWAS studies and 80 candidate-gene studies in Caucasian, and three GWAS studies and 28 candidate-gene studies in non-Caucasian populations. Despite the fact that 82 distinct SNPs have been discovered to associate with myocardial infarction, only one has been replicated in a 2nd paper, and none validated in a alternative subpopulation with

$p<5\times10^{-8}$. Our results suggested that these 82 SNPs could very likely be false positives, and myocardial infarction might be too complex of a disease phenotype to be investigated by current GWAS approaches.

Table 1: Diseases that killed over 10,000 Americans in 2009 and had been studied in GWAS in non-Caucasian populations were still lacking known cross-ethnic SNPs

Disease categories[&]	Mortality[#]	Count of papers in PubMed[$]			Counts of SNPs ($p<5\times10^{-8}$)		
		All	GWAS	GWAS with cohort size >1000	All	Replicated	Cross-ethnic
Myocardial infarction (I21-I22)	125,361	91(31)	11(3)	11(3)	82	1	0
Heart failure (I50)	56,752	10(4)	3(1)	3(1)	1	0	0
Renal failure (N17-N19)	43,628	50(22)	12(4)	9(3)	15	3	0
Malignant neoplasm of esophagus (C15)	13,916	15(10)	3(3)	3(3)	10	1	0

[&]ICD-10 codes are listed in the parenthesis
[#]Number of people killed in the US in 2009 from Centers for Disease Control and Prevention (CDC)
[$]Numbers of papers on non-Caucasian populations are listed in the parenthesis

We further identified 15 diseases that killed over 10,000 Americans but for which we could not find any published GWAS study with cohort size >1000 involving non-Caucasian populations (**Table 2**). There was only a single unbiased GWAS study within these 15 diseases: a study on 50 childhood acute lymphoblastic leukemia versus 50 controls in Korean [14]. Research funding for GWAS studies on these 14 diseases in non-Caucasian populations will highly likely yield cross-ethnic SNPs.

Finally, different from most other diseases lacking studies in non-Caucasian populations, esophageal cancer kill 13,916 Americans per year but has never been studied on populations other than Chinese and Japanese in a single unbiased GWAS study with cohort size >1000 (**Table 1**). Indeed, most studies on esophageal cancer were conducted outside the United States. Therefore we recommend additional genetic studies of esophageal cancer that incorporate ethnic populations representative of the broader US population.

Table 2: Top ten diseases that killed over 10,000 American in 2009 and had never been studied in large GWAS in non-Caucasian populations

Disease categories[&]	Mortality[#]	Count of papers in PubMed[$]			Counts of SNPs ($p<5\times10^{-8}$)		
		All	GWAS	GWAS cohort size >1000	All	Replicated	Cross-ethnic
All other forms of heart disease (I26-I28,I34-I38,I42-I49,I51) (Ventricular fibrillation, Pulmonary embolism, Atrioventricular block, Sudden cardiac arrest, Dilated cardiomyopathy, Pulmonary hypertension, Hypertrophic cardiomyopathy)	114,971	13(1)	6(0)	2(0)	6	0	0
Pneumonia (J12-J18) (Pneumonia susceptibility)	50,774	1(0)	0(0)	0(0)	0	0	0
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99) (Psoriatic arthritis, Sudden infant	43,076	20(3)	4(0)	3(0)	6	0	0

death syndrome, Oral lichen planus, Syncope, Pemphigus vulgaris, Headache)							
Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98) (Acute respiratory distress syndrome)	30,655	2(0)	0(0)	0(0)	0	0	0
Chronic liver disease and cirrhosis (K70,K73-K74)	30,444	1(1)	0(0)	0(0)	7	0	0
Leukemia(C91-C95)	22,697	27(3)	9(1)	7(0)	13	4	0
Non-Hodgkin's lymphoma (C82-C85)	20,361	11(0)	3(0)	3(0)	2	0	0
Malignant neoplasms of liver and intrahepatic bile ducts (C22) (Liver cancer)	19,311	2(2)	0(0)	0(0)	0	0	0
Alcoholic liver disease (K70) (Alcoholic cirrhosis, Alcoholic liver disease)	15,107	1(1)	0(0)	0(0)	7	0	0
In situ neoplasms, benign neoplasms and neoplasms of uncertain or unknown behavior (D00-D48) (Meningioma, Uterine leiomyoma, Polycythemia vera)	14,616	4(0)	0(0)	0(0)	0	0	0

[&]ICD-10 codes and individual disease names are listed in the parenthesis.
[#]Number of people killed in the US in 2009 from Centers for Disease Control and Prevention (CDC)
[$]Numbers of papers on non-Caucasian populations are listed in the parenthesis

Conclusion

We curated 4,573 genetic studies reporting on 763 human diseases and identified 398 cross-ethnic SNPs (DNA variants that have been validated in two diverse population groups with $p < 5\times10^{-8}$) associated with the increased risk of 42 diseases. We have mapped these diseases to 69 diseases represented in the mortality data of the CDC, and identified 19 diseases which killed over 10,000 Americans in 2009 but were still lacking cross-ethnic SNP.

These 19 diseases are organized into three categories. Fifteen of them had never been studied in any GWAS in non-Caucasian populations, although together they killed 425,942 Americans in 2009 alone. These 15 diseases include chronic liver diseases and cirrhosis, leukemia, non-Hodgkin's lymphoma, and others (**Table 2**). Another category contains myocardial infarction, renal failure, and heart failure which had been studied in diverse population groups but were still lacking the publication of a single cross-ethnic SNP, indicating that alternative techniques are likely needed to identify potential cross-ethnic SNPs (for example, case-control studies involving whole genome sequencing or the inclusion of environmental factors). A third category involves esophageal cancer, which killed 13,916 Americans in 2009 but had never been studied using GWAS on populations other than Chinese and Japanese, including Caucasian.

Our results indicate that diseases killing well over half million Americans every year are still lacking cross-ethnic SNPs, and in a diverse country like the United States of America, genetic disease association studies in diverse population groups still need to be performed.

Acknowledgements

The authors would like to acknowledge Shai Shen-Orr and Alex A. Morgan from Stanford University for insightful suggestions. We thank Alex Skrenchuk and Boris Oskotsky from Stanford University for computer support. R.C., J.T.D., D.R., A.J.B. were funded by the Lucile Packard Foundation for Children's Health, the Hewlett Packard Foundation, the National Institute of General Medical Sciences (R01 GM079719), the National Library of Medicine (R01 LM009719), and the Howard Hughes Medical Institute.

Author Contributions

A.J.B. conceived of the study. R.C. designed and conducted the study and wrote the manuscript. J.T.D., D.R., A.J.B. revised the manuscript. D.R. modified the figure.

References

1. Chen R, Davydov EV, Sirota M, Butte AJ (2010) Non-Synonymous and Synonymous Coding SNPs Show Similar Likelihood and Effect Size of Human Disease Association. PLoS One 5: e13574.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (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.
3. Bustamante CD, Burchard EG, De la Vega FM (2011) Genomics for the world. Nature 475: 163-165.
4. Saccone NL, Saccone SF, Goate AM, Grucza RA, Hinrichs AL, et al. (2008) In search of causal variants: refining disease association signals using cross-population contrasts. BMC Genet 9: 58.
5. Gillum LA, Gouveia C, Dorsey ER, Pletcher M, Mathers CD, et al. (2011) NIH disease funding levels and burden of disease. PLoS One 6: e16837.
6. (2011) Majority Of U.S. Babies Are Non-White For First Time, Census Finds The Huffington Post.
7. http://www.census.gov/population/www/pop-profile/natproj.html.
8. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. (2010) Clinical assessment incorporating a personal genome. Lancet 375: 1525-1535.
9. http://www.cdc.gov/nchs/fastats/deaths.htm.
10. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32: D267-270.
11. Bodenreider O (2000) Using UMLS semantics for classification purposes. Proc AMIA Symp: 86-90.
12. Shah NH, Muse MA (2008) UMLS-Query: a perl module for querying the UMLS. AMIA Annu Symp Proc: 652-656.
13. http://www.icd10data.com/.
14. Han S, Lee KM, Park SK, Lee JE, Ahn HS, et al. (2010) Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. Leuk Res 34: 1271-1274.