1	DISPARITIES IN ABO BLOOD TYPE DETERMINATION ACROSS DIVERSE ANCESTRIES:
2	A SYSTEMATIC REVIEW AND VALIDATION IN THE ALL OF US RESEARCH PROGRAM
3	
4	Kiana L. Martinez, ¹ Andrew Klein, ¹ Jennifer R. Martin, ^{1,2} Chinwuwanuju U. Sampson, ¹ Jason B.
5	Giles, ¹ Madison L. Beck, ¹ Krupa Bhakta, ¹ Gino Quatraro, ¹ Juvie Farol, ³ Jason H. Karnes ^{1,5,*}
6	
7	¹ Department of Pharmacy Practice and Science, The University of Arizona R. Ken Coit College
8	of Pharmacy, Tucson, AZ
9	² The University of Arizona Health Sciences Library, The University of Arizona, Tucson AZ
10	³ Department of Clinical and Translational Science, The University of Arizona College of
11	Medicine, Tucson AZ
12	⁴ Department of Biomedical informatics, Vanderbilt University Medical Center, Nashville, TN
13	
14	*Correspondence to:
15	
16	Jason H. Karnes, PharmD, PhD
17	Associate Professor
18	University of Arizona R. Ken Coit College of Pharmacy
19	1295 N Martin AVE
20	Tucson, AZ 85721
21	520-626-1447
22	karnes@arizona.edu
23	
24	
25	
26	

2

27 Abstract

28

29 Background:

- 30 ABO blood types have widespread clinical use and robust associations with cardiovascular
- 31 disease. Many studies determine ABO blood types using tag single nucleotide polymorphisms
- 32 (tSNPs) to characterize functional variation. However, tSNPs with low linkage disequilibrium
- 33 (LD) may promote misinference of ABO blood types, particularly in diverse populations.

34 Methods:

- 35 Bibliographic databases were searched for studies (2005-2022) using tSNPs to determine ABO
- 36 alleles in accordance with PRISMA 2020 guidelines. We calculated linkage between tSNPs and
- 37 functional variants across inferred continental ancestry groups from 1000 Genomes (AFR, AMR,
- EAS, EUR). We compared r^2 across ancestry and assessed real-world consequences by
- 39 comparing tSNP-derived blood types to serology in a large, diverse population from the All of Us
- 40 Research Program (AoURP).

41 **Results:**

- 42 We observe a lack of phasing and frequent use of inappropriate tSNPs in blood type
- determination, particularly for O alleles. Linkage between functional variants and O allele tSNPs
- 44 was significantly lower in African (median $r^2=0.443$) compared to East Asian ($r^2=0.946$,
- p=1.1x10-5) and European ($r^2=0.869$, p=0.023). In AoURP, discordance between tSNP-derived
- 46 blood types and serology was high across all SNPs in African ancestry individuals and linkage
- 47 was strongly correlated with discordance across all ancestries (ρ =-0.90, p=3.08×10-23).

48 **Conclusion:**

- 49 We observe common use of inappropriate tSNPs to determine ABO blood type, particularly for
- 50 O alleles and with some tSNPs mistyping up to 58% of individuals. Our results highlight the lack
- of transferability of tSNPs across ancestries and potential exacerbation of disparities in genomic
- 52 research for underrepresented populations.

3

53 Introduction

54 Patients with European ancestry make up 12% of the world's population, but account for approximately 81% of individuals in genomic studies^{1,2}. This lack of representation in genomic 55 56 studies has led to multiple racial and ethnic disparities in genomic research. Among such 57 disparities is the assumption that tag single-nucleotide polymorphisms (tSNPs) are appropriate 58 proxies for functional variation across ancestry groups. Because of differences in minor allele 59 frequency (MAF) and linkage disequilibrium (LD) structure, tSNPs are not always portable 60 across ancestry groups and inappropriate use of tSNPs may lead to reduced statistical power and erroneous associations^{3,4}. This practice may exacerbate existing disparities in genomic 61 research for underrepresented populations, especially when important clinical traits are inferred 62 63 based on tSNP variation.

64 ABO blood types have widespread clinical use, are among the first biomarkers tested in 65 the newborn, and are robustly associated with myriad phenotypes, such as hemolytic disease of 66 the newborn transfusion and solid organ transplant-related complications, and thromboembolic disease⁵⁻⁹. Individuals with non-O blood types have been shown to have increased risk for 67 68 coronary artery disease, myocardial infarction, deep vein thrombosis, and stroke¹⁰⁻¹³. The four 69 ABO blood groups (A, B, AB, and O) are largely the result of a small number of functional variants that produce combinations of ABO alleles (A, B, and O)¹⁴⁻¹⁶. While serological typing is 70 71 considered the gold standard for determining ABO blood types, studies often determine ABO blood type using genetic data when serological typing results are not collected or not 72 available^{17,18}. These studies often use tSNPs to determine ABO blood type because important 73 74 functional variation in ABO includes single base pair insertion/deletion polymorphisms that are not easily or accurately characterized on arrays or with genome-wide imputation¹⁹. Additionally, 75 76 most ABO tSNPs used to identify blood groups were originally identified in European and East Asian populations^{20,21}, leading to the possibility that tSNPs are used inappropriately in diverse 77 78 populations⁴.

4

79 Many diverse populations, including those with admixed American and African ancestry, are included in studies that infer ABO blood groups based on genetic variants²²⁻²⁵. However, 80 many of the commonly utilized tSNPs in ABO are unlikely to accurately capture functional 81 82 variation in these populations due to differences in LD patterns and MAFs^{21,26}. In this study, our 83 primary aim is to evaluate the portability and suitability of tSNPs used to determine ABO alleles 84 and blood types across diverse populations in published literature. We aimed to evaluate 85 inconsistencies in racial/ethnic representation and bioinformatic methodologies across 86 published studies investigating genotype-derived ABO blood groups. Finally, we investigated 87 the real-world consequences of utilization of tSNPs from existing studies through examination of LD patterns, genotype-derived ABO blood groups, and ABO serological test results from the 88 89 diverse participants in the All of Us Research Program (AoURP).

90

91 Methodology

92 In the systematic review portion of this study, we captured all recent publications 93 (publication dates 2005-2022) that used single-nucleotide polymorphisms (SNPs) to determine 94 ABO alleles (A, B, and O). ABO alleles are derived from pre-defined functional variants and are used to determine ABO blood types: OO (O blood type); AA, AO (A blood type); BB, BO (B 95 blood type); AB (AB blood type). We identified and recorded SNPs, either functional variants or 96 97 tSNPs, that were used to determine ABO alleles. As studies did not report LD between their utilized tSNPs and functional variants within their population, we calculated LD using r^2 in 98 reference populations provided by the 1000 Genomes Project (1KGP)²⁷. Because studies also 99 100 rarely reported genetically-derived ancestry for their populations, we inferred continental ancestry from 1KGP superpopulations (African ancestry [AFR], Admixed American ancestry 101 102 [AMR], East Asian ancestry [EAS], European ancestry [EUR], and South Asian ancestry [SAS]) to estimate r² between functional variants and tSNPs. To evaluate the real-life consequences of 103 104 utilizing potentially inappropriate tSNPs in place of functional variants, we compared ABO blood

5

types and alleles derived from tSNPs utilized in existing studies, functional variants, and ABO
serology test results in the diverse population with whole genome sequencing from the
AoURP²⁸. We calculated discordance between these different derivations of ABO blood types
and alleles across ancestries in the AoURP. Approval for human subjects research was not
required since all data used in this study was publicly available. Research conducted within the
AoURP is considered non-human subjects research since all data were de-identified and
available for public access.

112 <u>Systematic Review Search Strategy, Inclusion Criteria, Data Extraction, and Derived Data</u>

We conducted a systematic search for peer-reviewed publications (2005-2022) that used 113 114 SNPs to determine ABO blood types (Figure 1). The results of this review is reported in 115 accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 statement and checklist²⁹. Detailed descriptions of the search strategy, 116 117 inclusion criteria, data extraction, and derived data are presented in the Supplemental Materials (See Supplemental Methods and Tables S1-S6). We utilized the reference management tool 118 119 EndNote 20 (Clarivate, Philadelphia, PA, USA) for the initial compilation of references based on 120 our search strategy, and then used DistillerSR (DistillerSR, Ottawa, Canada) as our literature 121 reviewer software that enabled us to efficiently screen references and extract data. The 122 following bibliographic databases were searched: PubMed/Medline (National Library of Medicine), Embase (Elsevier), Scopus (Elsevier), Ovid/Medline (Wolters Kluwer), and CINHAL 123 124 Plus with Full Text (Ebsco). In addition, Dissertations and Theses Global (ProQuest) were also 125 searched as well as Google Scholar. These searches were conducted within 1 week in August 126 2022. Key words including MeSH (Medical Subject Heading) terms and free-text words were searched in titles, abstracts, and text. The full complete queries conducted in each database are 127 128 provided in the Supplementary Materials (Table S2).

129 The publications meeting the following criteria were included for the initial title and 130 abstract evaluation: 1) available in the English language, 2) the organism being studied is

6

131 human, and 3) is a primary article. In a secondary evaluation we screened the full text and 132 included only peer-reviewed articles that used single SNPs to derive ABO alleles from genomic 133 data. Major exclusion criteria were: 1) clinical trials, 2) reviews, and 3) meta-analysis. The 134 screening forms are available in Table S3 and Table S4. The bibliographic information 135 associated with each reference, such as reference title, authors, and journal, were automatically 136 extracted by DistillerSR. The primary data extracted included: 1) SNPs used for each ABO 137 allele, 2) presence of a haplotype analysis, 3) cohort size, 4) cohort population description, 5) 138 primary phenotype of interest, and 6) sequencing/genotyping platform (**Table S4**). We ensured 139 consistency of the collected data by having our data independently extracted by two different reviewers with disagreement between reviewers being resolved by consensus. 140 To better understand the range of phenotypes assessed in conjunction with ABO blood 141 142 type, we categorized the primary phenotype of the studies into twelve categories 143 (hematological, cardiovascular, oncology, infections disease, digestive system/gastrointestinal, 144 inflammatory, immune system, hepatic, neurological, metabolic, miscellaneous, multiple 145 phenotypes). Because genotyping technology influences the functional ABO variation that could 146 be determined, we also collected sequencing and genotyping platforms used to determine the 147 SNPs and ABO alleles. We categorized these platforms into four groups: 1) sequencing, 2) 148 genotyping, 3) PCR/specific target, and 4) multiple genotyping/sequencing platforms. Definitions 149 for these categories are available in the **Supplemental Materials**. Since the use of phased genetic data and imputation of haplotype structure is required to accurately determine ABO 150 151 alleles (**Table S5**), we also identified how many studies used a haplotype analysis to assign 152 functional variants or tSNPs to individual copies of chromosome nine. 153 Inferred Continental Ancestry for Assessment of LD 154 As our aim was to evaluate tSNPs across diverse populations, we were required to

determine LD between tSNPs and functional variants, as well as describe reported populations
 in terms of continental ancestry. Since studies using tSNPs did not report population-specific LD

7

values for linkage between functional variants and tSNPs, we estimated LD using r² from 157 representative populations from 1KGP²⁷. We used LDLink³⁰, specifically the LDpop³¹ module, to 158 calculate r² specific to super- and sub-populations as defined by 1KGP²⁷. For our primary 159 160 analysis, r² was calculated for each of the 1KGP superpopulations (AFR, AMR, EAS, EUR, or 161 SAS) that corresponded to each study population as described below. In studies where ABO functional variation was used, r² was determined to be 1. In studies where more than one SNP 162 was used and the population for which the SNP was used was not described, we assumed 163 studies used the SNP with the highest r^2 value for their population. 164

For each study in the systematic review, we inferred continental ancestry as described 165 by the International Genome Sample Resource (IGSR) for their 1KGP²⁷ phase three collection. 166 The vast majority of studies included in this systematic review did not describe genetically-167 168 derived ancestry for their participants. For example, many studies described their populations using racial terminology such as "White" or "Caucasian" or in terms of the population's 169 170 geography such as being "Japanese" or "Finnish". Thus, in scenarios where continental 171 ancestry was not provided, we inferred continental ancestry based on these other descriptors 172 and categorized them into AFR, AMR, EAS, EUR, or SAS.

173 <u>Statistical Analysis for Systematic Review</u>

We visually examined LD structure of the reported SNPs across populations using 174 LDBlockShow³² and data from 1KGP on GRCh38²⁷. R² was calculated between functional 175 variants and all reported tSNPs for each superpopulation. A tSNP was only included in our 176 population-specific analysis if that tSNP was actually used to derive ABO blood types in an 177 existing study including a population with similar ancestry. To examine within and across 178 population variation of LD, we calculated r² between all reported tSNPs and functional variants 179 180 across the 26 subpopulations described in 1KGP. Values were then grouped by 181 superpopulation to evaluate variation within and across superpopulations. The functional variant for O vs non-O alleles was defined as rs8176719 and the functional variants for A vs B alleles 182

8

were defined as rs7853989, rs8176743, rs8176746, and rs8176747. For our primary analyses, 183 we calculated a single r² value for each population in each study between the study's reported 184 tSNP and the functional variant (rs8176719 for O vs non-O and rs8176746 for A vs B) using the 185 186 matching 1KGP superpopulation. We first conducted a Kruskal Wallis test for differences in median r² across superpopulations for each tSNP using R³³. Then we performed pairwise 187 188 comparisons across superpopulations for each tSNP with Dunn's test of multiple comparisons 189 using rank sums. Our primary analysis excluded studies that used ABO allele-determining functional variants, however a secondary analysis was also performed that included studies 190 using functional variants. We excluded a single study with inferred South Asian (SAS) 191 continental ancestry due to the small sample size, restricting our inferred continental ancestries 192 of interest to AFR, AMR, EAS and EUR. Studies that had populations with unclear ancestry 193 were also excluded. We calculated the coefficient of variation for r² to evaluate variance across 194 195 superpopulations in a kernel density plot.

196 Derivation of Population and Data for All of Us Research Program Analysis

197 To evaluate real-world consequences of utilizing tSNPs in place of functional variants or serology, we utilized a subset of AoURP²⁸. The AoURP, sponsored by the National Institutes of 198 199 Health (NIH), is a longitudinal cohort study with an aim to advance precision medicine by 200 collecting electronic health records (EHR), participation-provided information (PPI), and 201 genomic data for at least one million US residents. AoURP EHR short-read whole genome sequence (srWGS) data was accessed via the version 7 curated data repository (CDR v7) 202 203 within the Controlled Tier. Participants that were included in this cohort were recruited between 204 2018 and 2022. AoURP participants with both ABO blood types derived from serology and srWGS data were included in our analyses. Serology was obtained from the EHR domain "Labs 205 206 & Measurements" under the logical observation identifiers names and codes (LOINC) code 882-1 (Abo and Rh group [Type] In Blood) and 883-9 (ABO group [Type] in Blood). For individuals 207

9

with multiple ABO serology results, we excluded any individuals whose ABO blood type differedupon repeat testing.

We used computed ancestry provided by the AoURP consortium based on srWGS ^{34,35}. 210 211 The ancestry categories are based on those described in gnomAD³⁶, Human Genome Diversity Project³⁷, and 1KGP²⁷: African (AFR), Latino/Native American/Admixed American (AMR), East 212 213 Asian (EAS), European (EUR), Middle Eastern (MID), and South Asian (SAS). A random forest 214 classifier was trained on a set of the HGDP and 1000 Genomes samples, and 16 principal 215 components (PCs) of the training sample genotypes were generated. AoURP samples were projected into the PC analysis space of the training data, and the classifier was applied to 216 determine ancestry. 217

218 Derivation of ABO Blood Type and Alleles in the AoURP

219 srWGS was accessed in the Hail MatrixTable format from the v7 genomics CDR. We 220 extracted SNPs of interest, including tSNPs utilized in existing studies and functional ABO 221 variants from the AoURP curated subset files "srWGS: ACAF Threshold" that is comprised of srWGS SNP and insertion-deletion (indel) variants that are frequent in the AoURP computed 222 ancestry subpopulations. Using SHAPEIT5³⁸, we phased a ~1 million base pair locus that 223 224 encompassed the ABO gene with the 1KGP phase three collection used as the reference. 225 ABO alleles were derived using functional variants and then tSNPs. For deriving ABO alleles 226 from functional variants, we utilized a 2-SNP approach. Individuals were first assessed on whether they had O or non-O blood type alleles using the functional variant rs8176719 with T 227 228 engendering O alleles and TC engendering non-O alleles. Those with non-O alleles were further 229 assessed to determine if they had A or B alleles using the functional variant rs8176746 with G 230 engendering A alleles and T engendering B alleles. Blood types were inferred based on whether 231 an individual had the following alleles: O/O (O), A/O or A/A (A), B/O or B/B (B), A/B (AB). 232 We then derived ABO alleles using tSNPs from existing studies. We first derived only O vs non-O alleles using the O vs non-O tSNPs: rs505922 (C>T), rs657152 (A>C), rs8176704 (G>A), 233

10

234 rs687289 (A>G), rs612169 (G>A), rs529565 (C>T), rs8176693 (C>T), rs514659 (C>A), and 235 rs8176645 (A>T). In each scenario the reference allele engendered the non-O allele, and the 236 alternative allele engendered the O allele. Those with non-O/non-O or non-O/O alleles were 237 considered to have a non-O blood type, and those with O/O alleles were considered to have an 238 O blood type. To assess A vs B tSNPs, we utilized a 2-SNP approach. The functional variant 239 rs8176719 was first used to determine O vs non-O, then A vs B alleles were derived using the A 240 vs B tSNPs: rs8176749 (C>T), rs8176672 (C>T), rs8176741 (G>A), rs8176722 (C>A), 241 rs8176720 (T>C). In each scenario the reference allele engendered the A allele and the alternative allele engendered the B allele. Blood types were subsequently determined using the 242 alleles as previously described. 243 244 Statistical Analysis for All of Us Research Program Cohort 245 LD between the functional variants and tSNPs was calculated using PLINK in the 246 AoURP cohort overall (ALL) and in each ancestry-specific population (AMR, AFR, EUR, SAS, 247 EAS, MID). Similarly, we calculated discordance between ABO blood types derived from

functional variants versus tSNPs. Additionally, we calculated discordance between ABO blood
 types derived from SNPs (both functional variants and tSNPs), and those captured by serology.

To evaluate the relationship between LD and discordance, we measured the relationship

251 between the discordance observed between the O vs non-O functional variant and tSNPs, and

the r² calculated between the functional variant and tSNPs with Spearman's correlation. Data

from the All of Us Research Program is accessible only through the Researcher Workbench

254 (https://workbench.researchallofus.org) as stipulated in the informed consent of participants in

the program. This data use agreement prohibits investigators from providing row level data on

AllofUs participants and thus providing a de-identified dataset is not possible for this manuscript.

The code used for this demonstration project is available within the Researcher Workbench.

258

259 Results

11

260 <u>Characteristics of the Studies for Systematic Review</u>

261 The initial search retrieved 2,693 publications from seven databases and resulted in 136 262 publications that met criteria for inclusion in this review (Figure 1; Table S6). the mean and 263 median year of publication were both 2016 (Table 1; Figure S1). Included studies were 264 published in a variety of journals (**Table S7**) and focused on a wide variety of phenotypes with the majority of studies focused on hematological (n=28, 20.6%), cardiovascular (n=28, 20.6%), 265 266 or oncology phenotypes (n=21, 15.4%) (Figure S2). The majority of platforms used fell under 267 the genotyping category (n=72, 52,9%), followed by PCR/specific target (n=46, 33,8%). sequencing (n=5, 3.7%), and multiple platforms (n=13, 9.6%) (Table 1; Figure S3). A total of 56 268 (41.2%) of the studies did not describe use of a haplotype analysis, 52 (38.2%) did describe a 269 270 haplotype analysis, and it was unclear for 28 (20.6%) of the studies (Table 1). Most studies 271 (n=100, 73.5%) were done in populations that would be described as a single population group, 272 AFR (n=7), AMR (n=3), EAS (n=30), EUR (n=59), and SAS (n=1). A total of 17 (12.5%) of the 273 studies were done in multiple populations and 19 (14.0%) of the studies had an unclear 274 population breakdown (Table 1; Table S8). Regardless of whether the study was done in a 275 single or multiple populations, most studies were done in EUR populations (n=74, 56.9%). 276 A total of 11 SNPs were used to determine O vs non-O blood type alleles and a total of 9 277 SNPs were used to determine A vs B blood type alleles (Figure S4). Some studies used more 278 than one SNP (n=9 [6.6%] to determine O vs non-O and n=20 [14.7%] to determine A vs B) while others used a single SNP (n=115 [84.6%] to determine O vs non-O and n=87 [64.0%] to 279 determine A vs B) (Table S9). Out of the 136 studies, 67 (49.3%) used the functional variant 280 (rs8176719) to determine the O blood type allele and 90 (66.2%) studies used at least one of 281 the functional variants (rs7853989, rs8176743, rs8176746, rs8176747) to determine the A or B 282 283 blood type allele. Linkage and Suitability Across Inferred Ancestry Groups for O vs non-O tSNPs Used by Studies 284

285 <u>in Systematic Review</u>

12

286	tSNPs used to determine the O allele were in highest LD with the functional variant in
287	the EAS population with median r^2 =0.95 (range 0.39-1) (Figure 2). No tSNP was in perfect LD
288	with the O allele functional variant rs8176719 across all populations. In the AFR population,
289	tSNPs used in prior studies displayed extremely low LD with the functional variant with median
290	r ² =0.36 (range 0.14-0.47) (Figure 2B). LDpop was unable to retrieve values for the tSNP
291	rs72238104 used by one study in the EAS population ³⁹ . We observed a wide range of LD
292	between the reported tSNPs and the functional variant(s) across superpopulations, especially in
293	the AFR group (Figure 3). All tSNPs had significantly different medians across
294	superpopulations (Table S10). Most notably, for seven out of the nine total tSNPs, the AFR
295	population exhibited significantly lower r ² values when compared to EAS and SAS populations
296	(Table S11).
297	In the primary analysis for the O vs non-O SNPs, when a study used a tSNP we
298	compared r ² values for the reported tSNPs to the functional variant rs8176719 using inferred
299	ancestry groups tSNPs had median r^2 =0.443 (range 0.387-0.448) in AFR, median r^2 =0.869
300	(range 0.853-0.975) in EUR, and median r^2 =0.946 (range 0.946-0.976) in EAS (Figure 4A). The
301	single study that used a tSNP in an AMR population had r^2 of 0.820. Tag SNPs used by studies
302	had significantly lower r^2 in AFR (median r^2 =0.443) than in EAS (median r^2 =0.946; p=1.1x10 ⁻⁵)
303	and EUR (median r^2 =0.869; p=0.023). We also observed that tSNPs performed worse in EUR
304	(median r^2 = 0.869) populations compared to EAS populations (median r^2 =0.946; p=1.6x10 ⁻³).
305	Other pairwise comparisons of median r ² between populations were not statistically significant
306	(Table S11).
307	Linkage and suitability Across Inferred Ancestry Groups of A vs B tSNPs Used by Studies in
308	Systematic Review
309	The four functional variants that differentiate between the A and B blood type alleles
310	(rs7853989, rs8176743, rs8176746, rs8176747) were in very high LD with one another only in

AFR, EAS, and SAS with range of r^2 =0.96-1 (**Figure S5**). In AMR and EUR, the functional

13

312	variant rs7853989 had lower LD with the other three functional variants with r^2 =0.72 in AMR and
313	r^2 =0.71 in EUR. In EAS, all the tSNPs were in high LD with one another with the exception of
314	rs8176720, which was also in low LD with the other reported SNPs (r^2 range 0.05-0.36 across
315	all populations). Within and across superpopulations, we observed a wide range of LD between
316	the reported tSNPs and the functional variant(s) (Figures S5-S7). Using rs8176746 as the A vs
317	B representative functional variant, four (rs7853989, rs8176672, rs8176720, rs8176722) out of
318	the seven tSNPs had significantly different medians across superpopulations (Table S10).
319	Pairwise comparisons between superpopulations were further examined for these 4 tSNPS, but
320	no one population consistently performed worse than the others (Table S11).
321	With respect to A vs B SNPs, the number of studies that used one of the functional
322	variants (rs7853989, rs8176743, rs8176746, rs8176747) vs tSNPs were: AFR (78.6%; 21.4%),
323	AMR (100%; 0%), EAS (92.9%; 7.1%), and EUR (75%, 25%). The two studies in AFR
324	populations that used A vs B tSNPs both had $r^2=1$ (Figure 4B). For studies in EAS populations,
325	one study used a tSNP with r^2 = 0.9936 and one with a r^2 =0.2566. For studies using an EUR
326	population, the r ² range was 0.1816-1 with a median of 1. No studies using an AMR population
327	used a tSNP to infer A or B blood type alleles (Figure 4B). Pairwise comparisons of median r^2
328	between populations were not statistically significant (Table S12).
329	Linkage and Suitability of tSNPs from Studies in Systematic Review, Evaluated in the All of Us
330	Research Program
331	A total of 10,771 individuals were identified with both srWGS and ABO blood type
332	serology results in the AoURP, comprised of individuals with AFR (n=2,000; 18.6%), AMR
333	(n=3,794; 35.2%), EAS (n=302; 2.8%), EUR (n=4,496; 41.7%), MID (n=47; 0.4%) and SAS
334	(n=132; 1.2%) ancestry (Table 2). A majority of the cohort was female (n=7,774; 71.9%) and
335	had a mean age of 53. Nearly half of the cohort had the O blood type (n=5,274; 49.0%) with the
336	O blood type being most prevalent in individuals with AFR and AMR ancestry (n=1,004; 50.2%

337 and n=2,179; 57.4%).

14

338	The median r^2 for O vs non-O tSNPs (n=9) versus the functional variant was r^2 =0.75
339	(range 0.18-0.83) in the complete cohort (ALL) (Figure S8). Compared to the other ancestry
340	groups, all the tSNPs were observed to have very low LD in the AFR ancestry group with
341	median r ² =0.44 (range 0.13-0.56). LD was consistently low across all ancestry groups for
342	rs8176704 with median r^2 =0.12 (range 0.001-0.13) and rs8176693 with median r^2 =0.19 (range
343	0.14-0.39). The median r^2 for A vs B tSNPs (n=5) versus the functional variant was r^2 =0.84
344	(range 0.15-0.998) in ALL. LD was consistently low across all ancestry groups for rs8176720
345	with median r^2 = 0.194 (range 0.076-0.305). Heatmaps of discordance between blood types
346	derived from functional variants versus tSNPs are provided in the Supplementary Figures S9
347	and S10 .
348	Performance of Genotype-Derived ABO Blood Types in the AoURP
349	We assessed discordance between O and non-O blood types derived from the functional

350 variant rs8176719 and tSNPs, and those derived from serology in the complete cohort and 351 across ancestry groups (Figure 5A). The tSNPs rs8176704 and rs8176693 displayed the most 352 discordance across all ancestry groups with a median of 0.43 (range 0.38-0.58) and 0.46 (range 353 0.40-0.58) respectively. Compared to the other ancestry groups, discordance was particularly 354 high across all SNPs in AFR with a median of 0.19 (range 0.01-0.51). In almost all cases, O and 355 non-O blood types derived from the functional variant observed less discordance with serology 356 than O and non-O blood types derived from tSNPs. In 10 cases, the tSNPs observed less 357 discordance than the functional variant: rs505922 in EUR and SAS, rs687289 in EUR and SAS, rs612169 in EUR and SAS, rs529565 in EUR and SAS, and rs514659 in EUR and SAS. The 358 359 majority of the discordance observed in these case occurred in the same 120 EUR participants and ≤ 20 EAS participants across the tSNPs. Additionally, LD (r^2) significantly correlated with 360 discordance with Spearman's ρ =-0.90 and p-value=3.08x10⁻²³ (Figure 5B). 361

362

363 Discussion

15

364 In the absence of serology, ABO blood types and alleles are often determined using 365 genomic variants and frequently derived in the pursuit of investigating associations or risk with cardiovascular disease^{9,40-42}. Here we present the results of a systematic review and analysis 366 that aimed to understand if studies are using appropriate tSNPs for their population to determine 367 ABO alleles, and if tSNPs are transferable across different populations. We found that many 368 tSNPs used to differentiate between O vs non-O alleles had low r² with the functional variant, 369 370 particularly for underrepresented populations such as AFR and AMR populations. This non-371 transferability of O vs non-O tSNPs was further exemplified by the disparate LD structures we observed in the ABO gene across continental populations. Not only did we observe highly 372 variable LD structure for reported SNPs across continental populations, but we also observed a 373 374 wide range of linkage between the reported tSNPs and the functional variant(s) within 375 populations. Furthermore, in a large and diverse cohort from AoURP, we found that O and non-376 O blood types derived from tSNPs had generally high discordance with serology, particularly in 377 the population with African ancestry. The observed discordance was highly inversely correlated 378 with LD.

The lack of diversity in genetic research has been well cited^{2,28,43-45}, and the issue is 379 compounded when findings done in one population are erroneously assumed to translate 380 directly to another population⁴⁶⁻⁴⁸. Genetic studies have been disproportionately performed in 381 cohorts of European ancestry with low representation of AFR and AMR individuals^{43,46} and our 382 383 systematic review reflects this trend. Despite the research bias towards European populations, our study shows that ABO tSNPs generally performed the best in EAS populations. This likely 384 reflects the early ABO genetic research performed primarily in Japanese cohorts^{14,49}. 385 Interestingly, we observed reductions in LD and blood group accuracy for tSNPs in EUR 386 387 compared to EAS populations. Our results suggest that many of the tSNPs that were originally identified in EAS populations by early ABO researchers have been used without appropriate 388 389 vetting in more diverse populations.

16

390 Acknowledging that tSNPs are not always transferable across populations also has broad implications for genomic research, which claims to be making efforts towards increasing 391 cohort diversity⁵⁰. While many studies established a tSNP linkage threshold of $r^2 \ge 0.8^{4,51,52}$, often 392 for genotyping and array purposes⁵³⁻⁵⁵, other studies acknowledge that the most conservative 393 approach to selecting tSNPs is to select tSNPs that are a perfect proxy ($r^2=1.0$) for the SNP of 394 interest⁵². Additionally, discordance of ABO blood alleles derived from functional variants and 395 396 tSNPs increases linearly as LD decreases in a diverse population, with a clear delineation 397 observed at $r^2=0.9$. Our results suggest that a conservative and very high linkage ($r^2>0.9$) threshold is needed to maintain blood type accuracy in the population of interest. As seen in this 398 399 study, tSNPs used for inferring O vs non-O alleles are not transferable across populations 400 particularly in AFR and AMR populations. Even within superpopulations (AFR, AMR, EAS, 401 EUR), tSNPs exhibit wildly varied linkage. A more rigorous approach would be to forgo the use 402 of tSNPs and instead use functional variants, although we recognize that many genotyping 403 methods cannot capture indels, such as the functional variant that differentiates between O vs non-O blood type alleles (rs8176719). Imputation of indels may be unreliable in some 404 405 populations since accuracy is dependent on the matching of LD structure between a cohort of interest and a reference population⁵⁶. In many cases, no appropriate reference population or 406 very small reference populations are available for underrepresented continental ancestry 407 groups, and this may lead to reduced imputation accuracy ⁵⁷⁻⁵⁹. Furthermore, many programs for 408 phasing and imputation are not able to handle indels ^{60,61}. 409

To evaluate the real-world consequences of utilizing tSNPs in place of functional variants, we utilized a large and diverse subset of the AoURP ²⁸ cohort to observe discordance between ABO blood types derived from functional variants, tSNPs, and serology. Discordance was generally high between O and non-O blood types derived from the functional variant and those derived from serology in participants with AFR ancestry. This finding was unsurprising as LD calculated between the functional variant and tSNPs in AFR populations from 1KGP and

17

AoURP were consistently low. Using tSNPs to derive ABO blood groups would have an
increased likelihood of being inaccurate in under-represented populations such as those with
African ancestry. Our data indicates that, depending on the tSNP used to derive O vs non-O
blood types and the population of interest, as many as 58 out of a 100 people could be
mistyped. Previously observed associations in ABO and the corresponding conclusions based
on these potentially inaccurate blood types are likely to further exacerbate the already prevalent
health disparities in genomic research.

423 We also observed inconsistent and inadequate documentation provided by publications. 424 Very few studies described their population in terms of continental genetic ancestry and instead used only racial or ethnic terms which makes evaluating the appropriateness of a tSNP 425 426 challenging. For a large portion of the studies, it was unclear whether the study conducted a 427 haplotype analysis. This is relevant because, to determine ABO alleles, phased data and a 428 haplotype analysis must be performed. Alarmingly, we were able to confirm that only 38.2% of 429 the studies conducted a phasing or haplotype analysis, the lack of which would likely result in 430 reduced accuracy of ABO blood group assignment.

431 This study has several limitations worthy of mention. Many studies only described their 432 populations using racial or ethnic terms, and we were forced to infer continental genetic 433 ancestry based off this poor proxy. We acknowledge the limitations of this approach and that, 434 while categories such as race and ethnicity have historically been used as proxies for genetic 435 ancestry due to positive correlation, race and ethnicity are socially-constructed concepts that should not be conflated with the biologically based groupings derived from genetic ancestry⁶². 436 Similarly, even the populations described by the most diverse publicly available dataset, the 437 1KGP²⁷, did not describe an appropriate continental ancestry population associated with Middle 438 439 Eastern individuals. This reflects a larger issue that much work is needed to establish reference 440 files that are more comprehensive and global. Secondly, some studies used more than one SNP to determine O vs non-O or A vs B blood type alleles but did not describe in which 441

18

442 scenario each SNP was used. Thus, we were compelled to select the best performing SNP and 443 our resulting analyses may reflect more optimistic results than the reality. In the systematic 444 review portion of this study, we were not able to identify individuals that overlapped in multiple 445 studies and adjust this total count in our analyses. Lastly, we were unable to account for 446 complex haplotypes, rare O-determining variants, or non-ABO variation conferring changes in blood type serology.⁶³⁻⁶⁵. For instance, individuals of the Bombay and para-Bombay phenotypes 447 448 appear to have functional A and B ABO alleles, but lack functional alleles in FUT1 and FUT2, 449 which encode precursor substrates for A and B transferases. Consequently, we are unable to 450 evaluate whether the few instances where blood types derived from tSNPs displayed less discordance with serology than with blood types derived from the functional variant was due to 451 452 inaccuracies from serological typing, problems with sequence accuracy known to be associated with indels^{66,67}, or rare functional variation in ABO or related genes. 453

454

455 **Conclusion**

456 ABO blood types have widespread clinical use and are robustly associated with 457 cardiovascular disease. Many studies opted to use functional variants to determine ABO blood 458 alleles, however, as observed in this study, there is potential for misinference of ABO blood 459 types due to using inappropriate tSNPs thus potentially leading to erroneous conclusions. We 460 observed that LD structure among tSNPs used in studies to infer ABO alleles varied substantially across populations. Even within superpopulations (AFR, AMR, EAS, EUR, SAS), 461 linkage between a tSNP and a functional variant varied widely. Many studies that used tSNPs 462 for determining O vs non-O alleles that had low r² particularly in AFR and AMR populations. 463 Furthermore, high discordance was observed between O and non-O blood types derived from 464 465 tSNPs versus serology, particularly in the AFR population from AoURP. Our results not only 466 have important implications for how ABO alleles are inferred from genomic data, but also highlight that tSNPs are not always transferable across continental populations and may result 467

19

- in furthering the genomic research and health disparities⁴⁵ already seen in underrepresented
- 469 populations.
- 470

471 Availability of data and materials

- 472 Data from the *All of Us* Research Program is accessible only through the Researcher
- 473 Workbench (https://workbench.researchallofus.org/workspaces/aou-rw-
- 474 a90a6fe4/duplicateofabophewas) as stipulated in the informed consent of participants in the
- 475 program. This data use agreement prohibits investigators from providing row level data on All of
- 476 Us participants and thus providing a de-identified dataset is not possible for this manuscript. The
- 477 code used for this demonstration project is available within the Researcher Workbench.
- 478 Datasets not generated from the All of Us Research Program and the associated code are
- 479 available on GitHub (https://github.com/Karnes-Lab/ABO_systematic_review).
- 480

481 Acknowledgements

- 482 The authors would like to thank the participants of the *All of Us* Research Program. Research
- reported in this publication was supported by the National Institutes of Health's National Heart,
- Lung, and Blood Institute (grants R01 HL156993, R01 HL158686, R21HL172036) and the
- 485 Office of the Director (grant OT2OD036485).
- 486

487 Author contributions

- 488 Conceived the study, J.H.K., K.L.M., J.B.G, J.R.M.; databases searches for systematic review,
- J.R.M., K.L.M.; data curation for systematic review, K.L.M., C.U.S., M.L.B, K.B., G.Q., J.F.; data
- 490 analysis, K.L.M., A.K.; analysis support, C.U.S.; visualization, K.L.M., A.K., wrote the paper,
- 491 K.L.M, J.H.K., A.K., with input from all authors.
- 492

493 **Declaration of interests**

20

494 The authors declare no competing interests.

495

496 **References**

- 1. Popejoy AB, Crooks KR, Fullerton SM, et al. Clinical Genetics Lacks Standard
- 498 Definitions and Protocols for the Collection and Use of Diversity Measures. Am J Hum Genet.
- 499 Jul 2 2020;107(1):72-82. doi:10.1016/j.ajhg.2020.05.005
- 2. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-
- 501 164. doi:10.1038/538161a
- 3. Cousin E, Genin E, Mace S, et al. Association studies in candidate genes: strategies to
- select SNPs to be tested. *Hum Hered*. 2003;56(4):151-9. doi:10.1159/000073200
- 4. Altshuler D, de Bakker PIW, Burtt NP, et al. Transferability of tag SNPs in genetic
- association studies in multiple populations. *Nature genetics*. 2006;38(11):1298-1303.
- 506 doi:10.1038/ng1899
- 507 5. Grundbacher FJ. The Etiology of ABO Hemolytic Disease of the Newborn. *Transfusion*
- 508 (Philadelphia, Pa). 1980;20(5):563-568. doi:10.1046/j.1537-2995.1980.20581034512.x
- 6. Welsby IJ, Phillips-Bute B, Mathew JP, et al. ABO blood group influences transfusion
- and survival after cardiac surgery. *Journal of thrombosis and thrombolysis*. 2014;38(3):402-408.
- 511 doi:10.1007/s11239-013-1045-2
- 512 7. Sazama K. Reports of 355 transfusion-associated deaths: 1976 through 1985.
- 513 Transfusion (Philadelphia, Pa). 1990;30(7):583-590. doi:10.1046/j.1537-
- 514 2995.1990.30790385515.x
- 515 8. Mugabe B, Thomas D, Bolton-Maggs P, Cohen H. Serious Hazards of Transfusion
- 516 (SHOT): Its Implications for Intensive Care. Journal of the Intensive Care Society.
- 517 2013;14(3):215-219. doi:10.1177/175114371301400308

518	9.	Groot HE, Villegas Sierra LE, Said MA, Lipsic E, Karper JC, van der Harst P. Genetically
-----	----	--

- 519 Determined ABO Blood Group and its Associations With Health and Disease. Arteriosclerosis,
- *thrombosis, and vascular biology.* 2020;40(3):830-838. doi:10.1161/atvbaha.119.313658
- 521 10. Chen Z, Yang S-H, Xu H, Li J-J. ABO blood group system and the coronary artery
- disease: an updated systematic review and meta-analysis. Scientific Reports. 2016/03/18
- 523 2016;6(1):23250. doi:10.1038/srep23250
- 524 11. Vasan SK, Rostgaard K, Majeed A, et al. ABO Blood Group and Risk of
- 525 Thromboembolic and Arterial Disease: A Study of 1.5 Million Blood Donors. *Circulation*. Apr 12
- 526 2016;133(15):1449-57; discussion 1457. doi:10.1161/circulationaha.115.017563
- 527 12. Abegaz SB. Human ABO Blood Groups and Their Associations with Different Diseases.
- 528 BioMed research international. 2021;2021:6629060-6629060. doi:10.1155/2021/6629060
- 13. Ward SE, O'Sullivan JM, O'Donnell JS. The relationship between ABO blood group, von
- 530 Willebrand factor, and primary hemostasis. *Blood*. 2020;136(25):2864-2874.
- 531 doi:10.1182/blood.2020005843
- 14. Yamamoto F-i, Clausen H, White T, Marken J, Hakomori S-i. Molecular genetic basis of
- the histo-blood group ABO system. *Nature (London)*. 1990;345(6272):229-233.
- 534 doi:10.1038/345229a0
- 15. Yamamoto F-i, McNeill PD, Hakomori S-i. Genomic organization of human histo-blood
- 536 group ABO genes. *Glycobiology (Oxford)*. 1995;5(1):51-58. doi:10.1093/glycob/5.1.51
- 537 16. Carlton VEH, Ireland JS, Useche F, Faham M. Functional single nucleotide
- polymorphism-based association studies. *Human Genomics*. 2006;2(6):391-402.
- 539 doi:10.1186/1479-7364-2-6-391
- 17. Lee SH, Park G, Yang YG, Lee SG, Kim SW. Rapid ABO genotyping using whole blood
- without DNA purification. *Korean J Lab Med*. Jun 2009;29(3):231-7.
- 542 doi:10.3343/kjlm.2009.29.3.231

- Seltsam A, Doescher A. Sequence-Based Typing of Human Blood Groups. *Transfusion medicine and hemotherapy*. 2009;36(3):204-212. doi:10.1159/000217322
- 545 19. Gorakshakar A, Gogri H, Ghosh K. Evolution of technology for molecular genotyping in
- 546 blood group systems. Indian J Med Res. Sep 2017;146(3):305-315.
- 547 doi:10.4103/ijmr.IJMR_914_16
- 548 20. Ni X, Bai C, Nie C, et al. Identification and replication of novel genetic variants of ABO
- 549 gene to reduce the incidence of diseases and promote longevity by modulating lipid
- 550 homeostasis. *Aging (Albany NY)*. Nov 22 2021;13(22):24655-24674.
- 551 doi:10.18632/aging.203700
- 552 21. Melzer D, Perry JRB, Hernandez D, et al. A genome-wide association study identifies
- protein quantitative trait loci (pQTLs). *PLoS genetics*. 2008;4(5):e1000072-e1000072.
- 554 doi:10.1371/journal.pgen.1000072
- 555 22. Greer JB, Larusch J, Brand RE, O'Connell MR, Yadav D, Whitcomb DC. ABO blood
- group and chronic pancreatitis risk in the NAPS2 cohort. *Pancreas*. 2011;40(8):1188-1194.
- 557 doi:10.1097/MPA.0b013e3182232975
- 558 23. Ohira T, Cushman M, Tsai MY, et al. ABO blood group, other risk factors and incidence
- of venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology
- 560 (LITE). Journal of thrombosis and haemostasis. 2007;5(7):1455-1461. doi:10.1111/j.1538-
- 561 7836.2007.02579.x
- 562 24. Song J, Chen F, Campos M, et al. Quantitative influence of ABO blood groups on factor
- 563 VIII and its ratio to von willebrand factor, novel observations from an ARIC study of 11,673
- subjects. *PloS one*. 2015;10(8):e0132626-e0132626. doi:10.1371/journal.pone.0132626
- 565 25. Zakai NA, Judd SE, Alexander K, et al. ABO blood type and stroke risk: The RE asons
- 566 for Geographic and Racial Differences in Stroke Study. Journal of Thrombosis and
- 567 *Haemostasis*. 2014;12(4):564-570.

23

- 568 26. Stram DO. Tag SNP selection for association studies. *Genetic epidemiology*.
- 569 2004;27(4):365-374. doi:10.1002/gepi.20028
- 570 27. Altshuler DM, Albers CA, Abecasis GR, et al. A global reference for human genetic
- 571 variation. *Nature (London)*. 2015;526(7571):68-74. doi:10.1038/nature15393
- 572 28. Denny JC, Rutter JL, Goldstein DB, et al. The "All of Us" Research Program. *The New*
- 573 England journal of medicine. 2019;381(7):668-676. doi:10.1056/NEJMsr1809937
- 29. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated
- 575 guideline for reporting systematic reviews. *BMJ (Online)*. 2021;372doi:10.1136/bmj.n71
- 576 30. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-
- 577 specific haplotype structure and linking correlated alleles of possible functional variants.
- 578 *Bioinformatics*. 2015;31(21):3555-3557. doi:10.1093/bioinformatics/btv402
- 579 31. Alexander TA, Machiela MJ. LDpop: an interactive online tool to calculate and visualize
- 580 geographic LD patterns. BMC bioinformatics. 2020;21(1):14-14. doi:10.1186/s12859-020-3340-
- 581 1
- 582 32. Dong S-S, He W-M, Ji J-J, Zhang C, Guo Y, Yang T-L. LDBlockShow: a fast and
- 583 convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call
- format files. *Briefings in bioinformatics*. 2021;22(4)doi:10.1093/bib/bbaa227
- 585 33. R: A Language and Environment for Statistical Computing 2023. https://www.R-

586 project.org

- 587 34. Program AoUR. Genomic Research Data Quality Report: All of Us Curated Data
- 588 Repository (CDR) release C2022Q4R9. 2023. <u>https://support.researchallofus.org/hc/en-</u>
- 589 <u>us/articles/4617899955092</u>
- 590 35. Wang X, Ryu J, Kim J, et al. Common and rare variants associated with cardiometabolic
- 591 traits across 98,622 whole-genome sequences in the All of Us research program. Journal of
- 592 human genetics. 2023;68(8):565-570. doi:10.1038/s10038-023-01147-z

- 593 36. Karczewski KJ, Cummings BB, Laricchia KM, et al. The mutational constraint spectrum
- ⁵⁹⁴ quantified from variation in 141,456 humans. *Nature (London)*. 2020;581(7809):434-443.
- 595 doi:10.1038/s41586-020-2308-7
- 596 37. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future.
- 597 Nature Reviews Genetics. 2005;6(4):333-340. doi:10.1038/nrg1596
- 598 38. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of
- 599 whole-genome and whole-exome sequencing data in the UK Biobank. *Nature genetics*.
- 600 2023;55(7):1243-1249. doi:10.1038/s41588-023-01415-w
- 39. Masuda M, Okuda K, Ikeda DD, Hishigaki H, Fujiwara T. Interaction of genetic markers
- associated with serum alkaline phosphatase levels in the Japanese population. *Hum Genome*
- 603 Var. 2015;2:15019. doi:10.1038/hgv.2015.19
- 40. Bruzelius M, Strawbridge RJ, Trégouët D-A, et al. Influence of coronary artery disease-
- associated genetic variants on risk of venous thromboembolism. *Thrombosis research*.
- 606 2014;134(2):426-432.
- 41. Li S, Schooling CM. A phenome-wide association study of ABO blood groups. *BMC*
- 608 Medicine. 2020/11/17 2020;18(1):334. doi:10.1186/s12916-020-01795-4
- 42. Zakai NA, Judd SE, Alexander K, et al. ABO blood type and stroke risk: the REasons for
- 610 Geographic And Racial Differences in Stroke Study. *Journal of thrombosis and haemostasis*.
- 611 2014;12(4):564-570. doi:10.1111/jth.12507
- 43. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and
- remedies. *Trends in genetics*. 2009;25(11):489-494. doi:10.1016/j.tig.2009.09.012
- 44. Bustamante CD, De La Vega FM, Burchard EG. Genomics for the world. *Nature*
- 615 *(London).* 2011;475(7355):163-165. doi:10.1038/475163a
- 45. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack of diversity in genomic
- databases is a barrier to translating precision medicine research into practice. *Health Affairs*.
- 618 2018;37(5):780-785. doi:10.1377/hlthaff.2017.1595

19 40. Martin AR, Rahai M, Rahalahi T, Okada T, Neale DM, Daly MJ. Chilical use of cu	artin AR, Kanai M, Kamatani Y, Okada Y, Neale BM,	aly MJ. Clinical use of curre
---	---	-------------------------------

- polygenic risk scores may exacerbate health disparities. *Nature genetics*. 2019;51(4):584-591.
- 621 doi:10.1038/s41588-019-0379-x
- 47. Carlson CS, Matise TC, North KE, et al. Generalization and Dilution of Association
- Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study.
- 624 PLoS Biology. 2013;11(9):e1001661-e1001661. doi:10.1371/journal.pbio.1001661
- 48. Duncan L, Shen H, Gelaye B, et al. Analysis of polygenic risk score usage and
- performance in diverse human populations. *Nature communications*. 2019;10(1):3328-9.
- 627 doi:10.1038/s41467-019-11112-0
- 49. Yamamoto F. A historical overview of advances in molecular genetic/genomic studies of
- the ABO blood group system. *Glycoconjugate Journal*. 2021:1-12.
- 50. Green ED, Gunter C, Biesecker LG, et al. Strategic vision for improving human health at
- 631 The Forefront of Genomics. *Nature (London)*. 2020;586(7831):683-692. doi:10.1038/s41586-
- 632 **020-2817-4**
- 51. Verlouw JAM, Clemens E, de Vries JH, et al. A comparison of genotyping arrays.
- 634 European journal of human genetics : EJHG. 2021;29(11):1611-1624. doi:10.1038/s41431-021-
- 635 **00917-7**
- 52. Daly MJ, Altshuler D, de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB. Efficiency and
- power in genetic association studies. *Nature genetics*. 2005;37(11):1217-1223.
- 638 doi:10.1038/ng1669
- 53. Hudson TJ, Lander ES, Schaffner SF, Daly MJ, Rioux JD. High-resolution haplotype
- 640 structure in the human genome. *Nature genetics*. 2001;29(2):229-232. doi:10.1038/ng1001-229
- 54. Gabriel SB, Schaffner SF, Nguyen H, et al. The Structure of Haplotype Blocks in the
- Human Genome. Science (American Association for the Advancement of Science).
- 643 2002;296(5576):2225-2229. doi:10.1126/science.1069424

- 55. Patil N, Berno AJ, Hinds DA, et al. Blocks of Limited Haplotype Diversity Revealed by
- High-Resolution Scanning of Human Chromosome 21. Science (American Association for the
- 646 Advancement of Science). 2001;294(5547):1719-1723. doi:10.1126/science.1065573
- 56. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum
- 648 *Genet*. 2009;10:387-406. doi:10.1146/annurev.genom.9.081307.164242
- 57. Schurz H, Müller SJ, Van Helden PD, et al. Evaluating the accuracy of imputation
- 650 methods in a five-way admixed population. *Frontiers in genetics*. 2019;10:34-34.
- 651 doi:10.3389/fgene.2019.00034
- 52 58. Nelson SC, Doheny KF, Pugh EW, et al. Imputation-based genomic coverage
- assessments of current human genotyping arrays. G3 (Bethesda). Oct 3 2013;3(10):1795-807.
- 654 doi:10.1534/g3.113.007161
- 55 59. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide Association Studies in
- Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*.
- 657 2019;179(3):589-603. doi:10.1016/j.cell.2019.08.051
- 658 60. Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster genotype imputation.
- 659 Bioinformatics. 2015;31(5):782-784. doi:10.1093/bioinformatics/btu704
- 660 61. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and
- 661 methods. *Nature genetics*. 2016;48(10):1284-1287. doi:10.1038/ng.3656
- 662 62. Fuentes A, Ackermann RR, Athreya S, et al. AAPA Statement on Race and Racism.
- 663 American journal of physical anthropology. 2019;169(3):400-402. doi:10.1002/ajpa.23882
- 664 63. Chopra G, Kataria M, Batra A, kaur G, Kumar R. Detection of a weaker subgroup of A in
- ABO blood group system. Asian journal of transfusion science. 2022;16(1):132-134.
- 666 doi:10.4103/ajts.ajts_66_21
- 667 64. Mizuno N, Ohmori T, Sekiguchi K, et al. Alleles Responsible for ABO Phenotype-
- 668 Genotype Discrepancy and Alleles in Individuals with a Weak Expression of A or B Antigens.
- 669 Journal of forensic sciences. 2004;49(1):1-8. doi:10.1520/jfs2003073

27

670	65.	Poskitt TR. Fortwenaler	Jr HP. A Studv of Weak	Subaroups of Blood	Group A with an
		,			

- 671 Antiglobulin-Latex Test. *Transfusion (Philadelphia, Pa)*. 1974;14(2):158-166.
- 672 doi:10.1111/j.1537-2995.1974.tb04510.x
- 673 66. Narzisi G, Schatz MC. The challenge of small-scale repeats for indel discovery. *Frontiers*
- *in bioengineering and biotechnology*. 2015;3:8-8. doi:10.3389/fbioe.2015.00008
- 675 67. Fang H, Wu Y, Narzisi G, et al. Reducing INDEL calling errors in whole genome and
- 676 exome sequencing data. *Genome medicine*. 2014;6(10):89-89. doi:10.1186/s13073-014-0089-z
- 677

678 Figure descriptions

679

680 Figure 1. Study selection process per PRISMA 2020 guidelines.

681

Figure 2. LD heatmap for SNPs that differentiate O vs non-O alleles across 1000

683 Genomes Project-defined superpopulations. This figure represents a 16.59 kb region of the ABO gene starting at 133.258 Mb and ending at 133.274 Mb. This region encompasses the 684 locations of the O vs non-O functional variant and tag SNPs. The blue block represents the 685 686 noncoding regions and the pink lines represent exons. The functional variant is highlighted in red. Linkage dissociation blocks representing the r² values of the O vs non-O functional variant 687 688 and tSNPs are shown beneath the section of the ABO gene. Lighter color blocks represent 689 lower r² values. A) Represents r² values across all continental populations. B) - F) Represents r² 690 values across specific continental populations. AFR: African ancestry, AMR: Admixed American 691 ancestry, EAS: East Asian ancestry, EUR: European ancestry, SAS: South Asian ancestry. 692

Figure 3. Linkage disequilibrium of O versus non-O tSNPs with functional variants across
 inferred continental ancestry groups. A) r² between O vs non-O functional variant

28

60F	
695	(rs8176719) and tag SNPs calculated across subpopulations and plotted per superpopulation
696	(AFR: African ancestry, AMR: Admixed American ancestry, EAS: East Asian ancestry, EUR:
697	European ancestry, SAS: South Asian ancestry). B) The coefficient of variation for r^2 between
698	the O vs non-O functional variant (rs8176719) and tSNPs calculated for each subpopulation
699	population and plotted as a kernel density estimate plot across superpopulations.
700	
701	Figure 4. Linkage disequilibrium between tSNPs and functional variants for O vs. non-O
701 702	Figure 4. Linkage disequilibrium between tSNPs and functional variants for O vs. non-O (A) and A vs. B (B) by inferred continental ancestries for each study. r ² was calculated for
701 702 703	Figure 4. Linkage disequilibrium between tSNPs and functional variants for O vs. non-O (A) and A vs. B (B) by inferred continental ancestries for each study. r ² was calculated for each SNP used to determine ABO alleles for each study included in this systematic review that
701 702 703 704	Figure 4. Linkage disequilibrium between tSNPs and functional variants for O vs. non-O (A) and A vs. B (B) by inferred continental ancestries for each study. r ² was calculated for each SNP used to determine ABO alleles for each study included in this systematic review that used a tSNP. r ² was calculated using the reference population (AFR: African ancestry, AMR:
701 702 703 704 705	Figure 4. Linkage disequilibrium between tSNPs and functional variants for O vs. non-O (A) and A vs. B (B) by inferred continental ancestries for each study. r ² was calculated for each SNP used to determine ABO alleles for each study included in this systematic review that used a tSNP. r ² was calculated using the reference population (AFR: African ancestry, AMR: Admixed American ancestry, EAS: East Asian ancestry, EUR: European ancestry) that best

that study is included more than once as each population may have a different r^2 value. For

studies using more than one SNP for a specific population, we assumed they were using the

SNP with the highest r^2 value for their population. **A)** rs8176719 was the functional variant used

to calculate r^2 which differentiates O vs non-O alleles and **B**) rs8176746 was selected as the

representative functional variant which differentiates A vs B alleles and was used to calculate r^2 .

No studies using an AMR population used a tSNP to derive A vs B blood type alleles.

713

Figure 5. Discordance between O vs non-O blood types from tSNPs and serology. A)

Discordance between ABO O and non-O blood types derived from the O vs non-O functional
 variant (rs8176719) and tSNPs and those derived from serology are displayed as a heatmap.

717 Discordance (proportion of non-matches in a cohort of 10,771) was calculated across all

ancestry groups (ALL) and calculated for each inferred continental ancestry group separately

- 719 (AFR: African ancestry, AMR: Admixed American ancestry, EAS: East Asian ancestry, EUR:
- European ancestry, MID: Middle Eastern ancestry, SAS: South Asian ancestry). B) Discordance

29

- 721 between blood types derived from O vs non-O tSNPs and those derived from serology was
- plotted along with LD (r²) between the O vs non-O functional variant and tSNPs. A Spearman's
- 723 correlation coefficient was calculated (-0.90, p-value = 1.51×10^{-38}).
- 724 Tables
- 725

Table 1. Summary of characteristics of systematic review studies.

	AFR	AMR	EAS	EUR	SAS	Multiple	Unclear
Total studies, n	7	3	30	59	1	17	19
Total participants, n*	11096	2173	215613	1139015	646	1643744	1089870
Mean Year of Study	2015	2015	2016	2016	2021	2016	2017
Sequencing/genotyping							
platform, n (%)							
sequencing	2 (28.6)	0 (0)	0 (0)	1 (1.7)	0 (0)	0 (0)	2 (10.5)
genotyping	4 (57.1)	0 (0)	12 (40.0)	36 (61.0) ⁻	l (100.0)	10 (58.8)	9 (47.4)
PCR/specific target	1 (14.3)	3 (100.0)	14 (46.7)	17 (28.8)	0 (0)	5 (29.4)	6 (31.6)
multiple platforms	0 (0)	0 (0)	4 (13.3)	5 (8.5)	0 (0)	2 (11.8)	2 (10.5)
Performed haplotype							
analysis, n (%)							
yes	4 (57.1)	1 (33.3)	5 (16.7)	27 (45.8)	l (100.0)	10 (58.8)	4 (21.2)
no	2 (28.6)	2 (66.7)	19 (63.3)	21 (35.6)	0 (0)	3 (17.6)	9 (47.4)
unclear	1 (14.3)	0 (0)	6 (20.0)	11 (18.6)	0 (0)	4 (23.5)	6 (31.6)

Each study is categorized as having participants from only one of the inferred continental ancestry groups (AFR: African ancestry, AMR: Admixed American ancestry, EAS: East Asian ancestry, EUR: European ancestry, SAS: South Asian ancestry) or having participants from multiple inferred continental ancestry groups (Multiple) or it was unclear which group

30

participants fell in (Unclear). Total number of studies is provided (n) as well as proportion within the same inferred continental ancestry group (%) unless otherwise specified.

*Total participants count may be incomplete due to some studies having unclear cohort size. Additionally, this count does not adjust for overlapping participants, i.e. when different studies use the same cohort.

726			
727			
728			
729			
730			
731			
732			
733			
734			
735			
736			
737			
738			
739			
740			
741			
742			
743			
744			
745			

746

747

Table 2. Demographic	characteristi	cs for All of L	Js cohort.					
	AFR	AMR	EAS	EUR	MID	SAS		
Total, n	2000	3794	302	4496	47	132		
Mean Age (Standard								
Deviation)	50 (16.6)	45 (16.1)	53 (15.9)	61 (16.8)	50 (17.5)	52 (16.2)		
Sex (female), n (%)	1535 (76.8)	3066 (80.8)	208 (68.9)	2822 (62.8)	30 (63.8)	83 (62.9)		
Blood Type, n (%)								
C	D 1004 (50.2)	2179 (57.4)	122 (40.4)	1901 (42.3)	18 (38.3)	52 (39.4)		
/	A 553 (27.7)	1104 (29.1)	91 (30.1)	1875 (41.7)	22 (46.8)	38 (28.8)		
E	3 384 (19.2)	437 (11.5)	73 (24.2)	561 (12.5)	≤20 (10.6)	32 (24.2)		
A	3 59 (3.0)	74 (2.0)	≤20 (5.3)	159 (3.5)	≤20 (4.3)	≤20 (7.6)		
Each participant is est	timated to hav	ve AFR (Afric	an), AMR (A	Admixed Ame	erican), EAS	(East		
Asian). EUR (Europea	Asian), FUR (European), MID (Middle Fastern), or SAS (South Asian) ancestry. Total number of							

participants is provided (n) as well as proportion within the same ancestry group (%) unless otherwise specified. Aggregate statistics corresponding to fewer than 20 participants are suppressed in compliance with AoURP data and statistics dissemination policy.

It is made available ander a set by the the title international income

Identification of studies via databases and registers











EUR









Inferred Continental Ancestry

Inferred Continental Ancestry





0

0.5

0.4