Lymphocyte Count Derived Polygenic Score and Interindividual Variability in CD4 T-cell Recovery in Response to Antiretroviral Therapy

Kathleen M. Cardone[1], Scott Dudek[1], Karl Keat[2], Yuki Bradford[1], Zinhle Cindi[1,7], Eric S. Daar[3], Roy Gulick[4], Sharon A. Riddler[5], Jeffrey L. Lennox[6], Phumla Sinxadi[7], David W. Haas[8,9], Marylyn D. Ritchie[1,10]
[1]*Department of Genetics,* [2]*Genomics and Computational Biology Graduate Program*
University of Pennsylvania, Philadelphia, PA, USA
[3]*Lundquist Institute at Harbor-UCLA Medical Center, Torrance, CA, USA*
[4]*Weill Cornell Medicine New York, New York, NY, USA*
[5]*University of Pittsburgh, Pittsburgh, PA, USA*
[6]*Emory University School of Medicine, Atlanta, GA, USA*
[7]*Division of Clinical Pharmacology, Department of Medicine*
University of Cape Town, Cape Town, South Africa
[8]*Vanderbilt University Medical Center, Nashville, TN, USA*
[9]*Meharry Medical College, Nashville, TN, USA*
[10]*Institute for Biomedical Informatics*
University of Pennsylvania, Philadelphia, PA, USA
**Email:* marylyn@pennmedicine.upenn.edu

Access to safe and effective antiretroviral therapy (ART) is a cornerstone in the global response to the HIV pandemic. Among people living with HIV, there is considerable interindividual variability in absolute CD4 T-cell recovery following initiation of virally suppressive ART. The contribution of host genetics to this variability is not well understood. We explored the contribution of a polygenic score which was derived from large, publicly available summary statistics for absolute lymphocyte count from individuals in the general population (PGS_{lymph}) due to a lack of publicly available summary statistics for CD4 T-cell count. We explored associations with baseline CD4 T-cell count prior to ART initiation (n=4959) and change from baseline to week 48 on ART (n=3274) among treatment-naïve participants in prospective, randomized ART studies of the AIDS Clinical Trials Group. We separately examined an African-ancestry-derived and a European-ancestry-derived PGS_{lymph}, and evaluated their performance across all participants, and also in the African and European ancestral groups separately. Multivariate models that included PGS_{lymph}, baseline plasma HIV-1 RNA, age, sex, and 15 principal components (PCs) of genetic similarity explained ~26-27% of variability in baseline CD4 T-cell count, but PGS_{lymph} accounted for <1% of this variability. Models that also included baseline CD4 T-cell count explained ~7-9% of variability in CD4 T-cell count increase on ART, but PGS_{lymph} accounted for <1% of this variability. In univariate analyses, PGS_{lymph} was not significantly associated with baseline or change in CD4 T-cell count. Among individuals of African ancestry, the African PGS_{lymph} term in the multivariate model was significantly associated with change in CD4 T-cell count while not significant in the univariate model. When applied to lymphocyte count in a general medical biobank population (Penn Medicine BioBank), PGS_{lymph} explained ~6-10% of variability in multivariate models (including age, sex, and PCs) but only ~1% in univariate models. In summary, a lymphocyte count PGS derived from the general population was not consistently associated with CD4 T-cell recovery on ART. Nonetheless, adjusting for clinical covariates is quite important when estimating such polygenic effects.

Keywords: HIV; Polygenic Scores; Lymphocyte Count; CD4 T-Cell Count; Pharmacogenomics

1. Introduction

1.1. *Incomplete CD4 T-Cell Recovery in Response to Antiretroviral Therapy*

Human immunodeficiency virus type 1 (HIV-1) is a global health challenge, with 38.4 million individuals worldwide living with HIV[1], including nearly 1.2 million in the United States[2]. This virus depletes CD4 T lymphocytes (hereafter referred to as CD4 cells), a critical component of the immune system[3]. Effective antiretroviral therapy (ART) controls viral replication, improves health and prevents transmission[4]. With viral load reduction, CD4 cell counts may return to normal levels, but in many individuals this is not achieved[5–7]. Understanding the etiology of CD4 cell recovery is important because individuals with lower CD4 cell counts may be at increased risk for non-AIDS conditions such as hepatic cirrhosis, cardiovascular disease, kidney disease, and cancer[8].

The etiology of incomplete CD4 cell recovery has not been fully elucidated, but many biological, demographic, treatment, and genetic factors have been associated[9]. Individuals who begin ART with CD4 cell counts <200 cells/mm^3 are less likely to achieve normal CD4 cell counts >500 cells/mm[5–7]. Other biological factors associated with this treatment response include higher body mass index (BMI), lower naïve/memory CD4$^+$ cell ratio, lower CD4/CD8 cell ratios, and other immunological factors[9]. Demographic factors have also been associated with poor CD4 cell recovery including older age, male sex, and Eastern African ancestry, as well as specific ART regimens[9,10]. Additionally, variants that influence the absorption, distribution, metabolism, and elimination of ART may also play a role[11]. Genes with single nucleotide polymorphisms (SNPs) reported to be associated with CD4 cell recovery on ART have included *IL-2*, *IL-2Rβ*, *IL-2Rγ*, *IL-15*, *IL-15Rα*, *TRAIL*, *Bim*, *TNF-α*, and *IFN-γ*[12]. One particular SNP (rs6897932) in *IL7RA* was associated with a faster CD4 cell count increase in individuals of both European and African ancestry, but another SNP in this gene (rs3194051) was only associated with this response in individuals of African ancestry[13,14]. Another study suggested that differences in *CCR5* genotype and *CCL3L1* dosage were associated with the extent and rate of CD4 cell recovery[15]. Additionally, *HLA-Bw4* homozygosity was associated with impaired CD4 cell recovery[16]. Particular mitochondrial DNA haplogroups were associated with CD4 cell recovery in individuals of European and African ancestry[17,18]. More recently, whole exome sequencing associated 41 genes with CD4 cell response in females[19].

Although multiple genes and SNPs have been associated with poor CD4 cell count recovery on ART, these explain a small fraction of the variance. Previous studies considered effects of SNPs individually, which fails to consider whether combinations of many SNPs may explain a larger portion of the variance. Many conditions are polygenic (e.g., coronary artery disease), meaning that many genes and variants have impact[20]. It is conceivable that CD4 cell recovery on ART is also polygenic, so it is worth exploring whether polygenic scores may explain a larger portion of the genetic variance, which has never been investigated for this treatment response. Furthermore, understanding the

pharmacogenomic underpinnings of treatment response has the potential to better individualize therapy[21].

1.2 *Polygenic Scores May Predict Complex Treatment Responses*

One way to assess the contribution of many variants in combination is by applying Polygenic Scores (PGS), which are the mathematical, cumulative aggregation of risk derived from the total contribution of numerous variants in the genome[22]. PGS effectively predict phenotypes such as schizophrenia[23–27], bipolar disorder[23,28,29], breast cancer[30–33], type 2 diabetes[30,34,35], coronary artery disease [30,34,36], and atrial fibrillation[30,34,37]. Given their success in other disease areas, it is plausible that PGS could predict poor CD4 cell recovery in response to ART.

When using PGS, it is important to consider the potential for ancestral health disparity. Across many phenotypes, PGS is more predictive for individuals of European ancestry because this population has more readily available summary statistics from large genome-wide association studies (GWAS)[38]. An ultimate goal of PGS is clinical implementation so that patients can be informed of their genetic risk for disease[38]. However, clinical implementation could create a larger health disparity whereby individuals of European ancestry may more readily benefit from these risk prediction models[38]. Thus, it is important to improve risk prediction for global populations. This is particularly important for HIV given its global distribution of prevalence, particularly in Africa. We hope to better predict genetic risk in individuals of African ancestry by generating a PGS based on summary statistics generated in a dataset of individuals largely of African ancestry, in addition to a PGS generated in a dataset of individuals largely of European ancestry. Additionally, we plan to use PRScsx, a method that more effectively predicts polygenic risk in global populations[39].

In this study, we assess whether the PGS generated from a general population is predictive of CD4 cell recovery in persons living with HIV (PWH). A similar approach used a body mass index PGS generated from a general population to study ART-associated weight gain[40]. As there are no large GWAS studies of CD4 cell count, either in the general population or in PWH, we generate statistical power by using summary statistics on total lymphocyte count from a general population, for which large sample sizes are publicly available. Finally, the principle of predicting phenotypic effects in a population affected by a health condition by using genetics from the general population was effective in one study that found that variants associated with cardiac QRS duration in individuals without cardiac diseases were also associated with arrhythmia and atrial fibrillation[41]. We assess whether this same principle applies to treatment response by testing whether the genetic underpinnings of lymphocyte count in a general population predicts CD4 cell recovery in PWH. We hypothesize that cumulative genetic variants that affect total lymphocyte count also affect recovery of the CD4 T cell subset in response to ART (i.e., that a lymphocyte count PGS [PGS_{lymph}] generated from the general population will be associated with CD4 cell recovery on ART). We also hypothesize that PGS_{lymph} will be associated with CD4 cell counts prior to initiating ART.

2. Methods

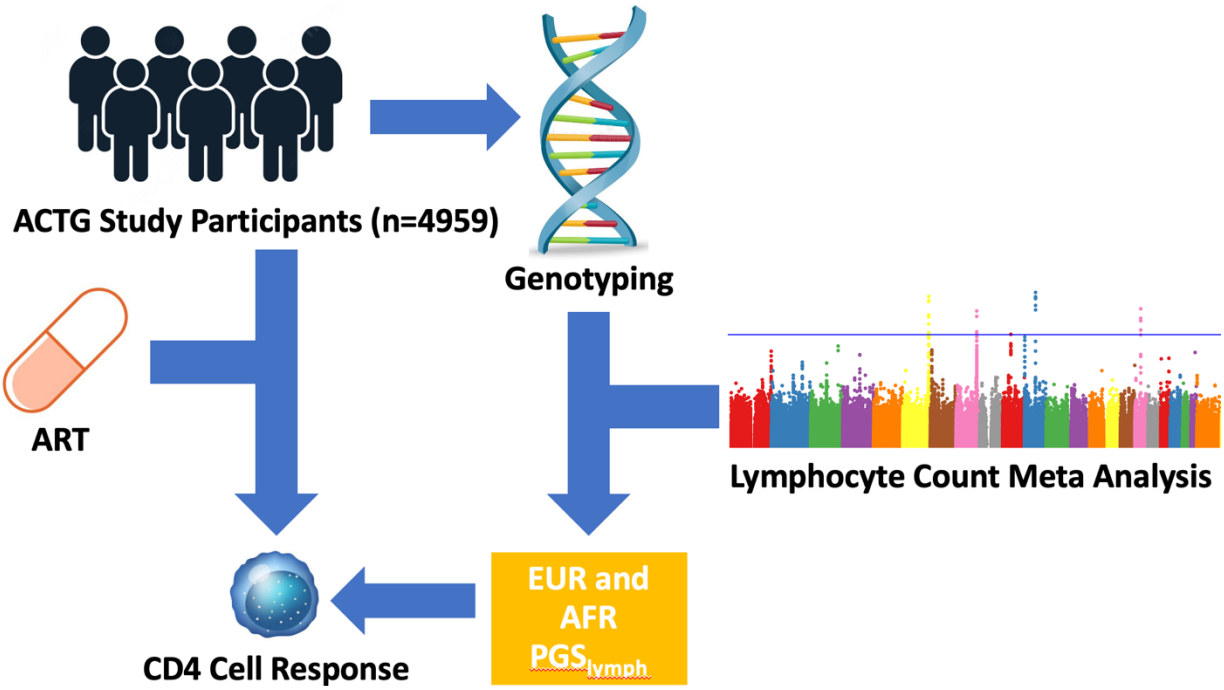

Figure 1: Study Overview: EUR and AFR PGS_{lymph} were trained using lymphocyte count GWAS summary statistics. Both PGS_{lymph} were applied to individuals in the AIDS Clinical Trials Group (ACTG) to assess its predictability of CD4 cell response to ART.

2.1 *Data and Study Participants*

2.1.1 *Lymphocyte Count Meta Analysis*

We used publicly available summary statistics from a published meta-analysis of existing GWAS for lymphocyte count in populations of European and African ancestry in the general population[42]. The meta-analysis included 524,923 individuals of European ancestry with 47,264,266 SNPs, and 13,477 individuals of African ancestry with 34,121,887 SNPs[42]. The European ancestry summary statistics were subset to 1,120,498 SNPs that were present on the European linkage disequilibrium (LD) panels and the African ancestry summary statistics were subset to 1,225,091 SNPs that were present on the African LD reference panels.

2.1.2 *AIDS Clinical Trials Group*

Participants were ART-naïve individuals who had initiated ART in prospective, randomized clinical trials of the AIDS Clinical Trials Group (ACTG), and had consented to genetic research and provided DNA under ACTG protocol A5128[43]. Data were generated by conducting a retrospective analysis of

this cohort. Individuals had initiated ART in the United States in studies ACTG384, A5095 (NCT00013520), A5142 (NCT00050895), A5202 (NCT00118898), and A5257 (NCT25285539)[44–47]. All participants provided written, informed consent for genetic testing. Drug class components of regimens were randomly assigned except for nucleoside reverse transcriptase inhibitor (NRTI) choice in A5142. Included individuals had the following data: imputed genotype, sex, genetically inferred ancestry (GIA), lymphocyte count or CD4 cell count data. Additional eligibility criteria included HIV-1 RNA <400 copies/mL at week 48 on ART.

2.1.3 *Penn Medicine BioBank*

The Penn Medicine BioBank (PMBB) is an electronic health record (EHR)-linked biobank research program at the University of Pennsylvania[48]. PMBB participants included in this study provided consent for research including access to their medical records, blood sample collection, and generation of genetic data[48]. Individuals with both imputed genotype data from PMBB v2.0 and with lymphocyte count data were included in PGS analysis as a positive control. Included individuals had the following data: imputed genotype, lymphocyte count, sex, and GIA.

2.2 *Genotyping and Quality Control*

2.2.1 *AIDS Clinical Trials Group*

DNA extracted from whole blood was labeled with coded identifiers and genotyped in seven phases. Phases 1-3 were genotyped at the Broad Institute (Phases 1 and 2 with HumanHap650Yv3_A, and Phase 3 with Human1M-Duov3_B). Phases 4-7 were genotyped at the Vanderbilt Technologies for Advanced Genomics (VANTAGE) facility (Phase 4 using the Human Core Exome chip, phase 5 with the HumanOmni2.5Exome-8-v1.1_A1 chip, Phase 6 with the HumanOmni25-8v1-2_A1 chip, and phase 7 with the Illumina Infinium Multi-Ethnic Global BeadChip (MEGA^{EX}).

Post-genotype quality control procedures utilizing PLINK v1.9[49] were conducted by Vanderbilt Technologies for Advanced Genomics Analysis and Research Design (VANGARD). Prior to imputation, samples with genotyping efficiency < 99% or with discordance between genotype sex and reported sex were removed. After completing these quality control procedures, each genotyping phase was imputed separately utilizing the TOPMed reference panel, which was parallelized by chromosome to increase computational efficiency[50]. During the imputation process, liftOver was used to transform genotype data to genome build 38[50]. After imputation, PLINK was used to merge the seven imputed datasets, and variants with imputation R^2 scores < 0.3, genotyping call rates < 95%, or minor allele frequency (MAF) < 0.05 were dropped[49]. GIA was determined using principal component analysis (PCA) with 1000 Genomes as the reference, subsequently assigning each participant to one of six

superpopulations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS), and Other.

2.2.2 *Penn Medicine BioBank*

DNA was extracted from blood samples. Approximately 80% of samples were genotyped by the Regeneron Genomics Center (RGC) using an Illumina Global Screening Array v.2.0 (GSAv2)[48], while the remaining 20% were genotyped by the Center for Applied Genomics (CAG) at the Children's Hospital of Philadelphia using the GSAv1 and GSAv2 genotyping array[48].

Prior to imputation, sample level quality control was conducted[48]. Using PLINK v1.9, variants with genotyping call rates < 95%, individuals with sample call rates < 90%, and individuals with discordance between reported sex and genotype sex were dropped[48]. Autosomes were imputed utilizing a TOPMed version R2 genome build 38 reference panel[48,50]. After imputation, variants with imputation R^2 scores < 0.3, genotype call rate < 99%, MAF < 1%, and/or were multi-allelic were dropped using PLINK v1.9[48]. Individuals with sample call rate < 99% or discordant sex information were also dropped[48]. PCA was done to identify GIA using 1000 Genomes as the reference and subsequently separated individuals into six superpopulations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS), Other[48].

2.3 *Polygenic Score Calculation*

The PGS_{lymph} was constructed using PRScsx (version released on July 29 2021), which integrates summary statistics and LD panels across genetically diverse populations to better predict polygenic risk in global populations[39]. 1000 Genomes phase 3 LD reference panels were used in the calculation[51]. Summary statistics from the lymphocyte count meta-analysis were used to train the PGS_{lymph}[42]. The PGS_{lymph} was applied to ACTG study participants with CD4 cell count data using PLINK2 "--score" function[49]. As positive controls, the PGS_{lymph} was also applied to individuals with lymphocyte count data in ACTG as well as individuals with lymphocyte count data in PMBB.

2.4 *Statistical Analysis*

The results were analyzed to assess model predictability across all ancestries combined, and in European and African ancestries separately. Linear regressions were calculated, and performance was assessed with an R^2 value generated from a multivariate linear regression between the phenotype of interest and the PGS_{lymph}. Additionally, performance of individual covariates was assessed with effect sizes generated from these regressions. We used a p-value threshold of 0.05 to assess significance. Regressions were calculated in individuals of European and African ancestry only, as well as individuals of all superpopulations combined. PGS_{lymph} was applied to two different cohorts, ACTG

and PMBB. In ACTG, the predictability of the PGS_{lymph} for three different phenotypes was assessed: the square root (SQRT) of CD4 cell count at study entry prior to ART (baseline), change in CD4 cell count from study entry to 48 weeks of ART (a measure of treatment response), and inverse normal lymphocyte count prior to ART (a control variable). We performed two regressions for each phenotype, one without correcting for any covariates, and one correcting for age, sex, principal components (PC) of genetic similarity 1-15, as well as log_{10}-HIV-1 RNA (a measure of viral load). Additionally, we adjusted for SQRT of baseline CD4 cell count in regression models between PGS_{lymph} and change in CD4 cell count on ART. In addition to these regressions, we also evaluated interactions between the PGS_{lymph} and age, sex, viral load, and baseline CD4 cell count to identify whether PGS_{lymph} interacts with any covariate. In PMBB, the predictability of PGS_{lymph} for inverse normal lymphocyte count was assessed as a positive control and to understand predictability in a general medical biobank population. Similarly, two regressions were performed, one without correcting for covariates, and one correcting for age, sex, and PC1-15. These results were visualized using SynthesisView[52].

3. Results

Table 1: ACTG Participant Demographics at Baseline

	Lymphocyte Count Data	Baseline CD4 Cell Count Data	On-Treatment CD4 Cell Count Data
Total, N	4680	4959	3274
European ancestry, n (%)	1835 (39.2%)	1958 (39.4%)	1319 (40.3%)
African ancestry, n (%)	1721 (36.8%)	1826 (36.8%)	1154 (35.2%)
Male/Female, n (%)	3824/856 (81.7%/18.3%)	4051/908 (81.7%/18.3%)	2715/559 (82.9%/17.1%)
Age, mean (range)	37.9 (17.0-77.0)	38.0 (17.0-77.0)	38.2 (17.0-76.0)

Table 2: PMBB Demographics

	Lymphocyte Count Data
Total, N	37211
European ancestry, n (%)	25330 (68.1%)
African ancestry, n (%)	10217 (27.5%)
Male/Female, n (%)	18215/18996 (49.0%/51.0%)
Mean Age (Range)	55.6 (13.9-101.7)

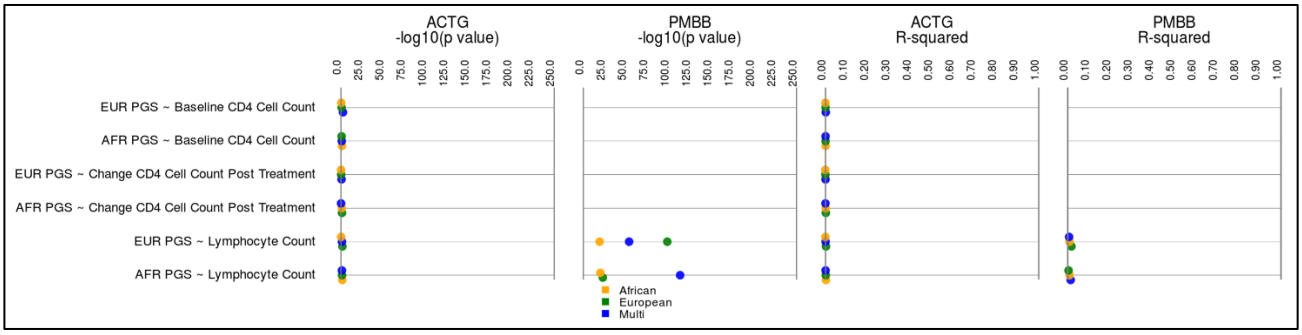

Figure 2: Summary of Regression Results Between PGS_{lymph} and Phenotype without Controlling for Covariates (Age, Sex, PC1-15, log_{10}HIV-1 RNA (viral load), and SQRT of baseline CD4 cell count)

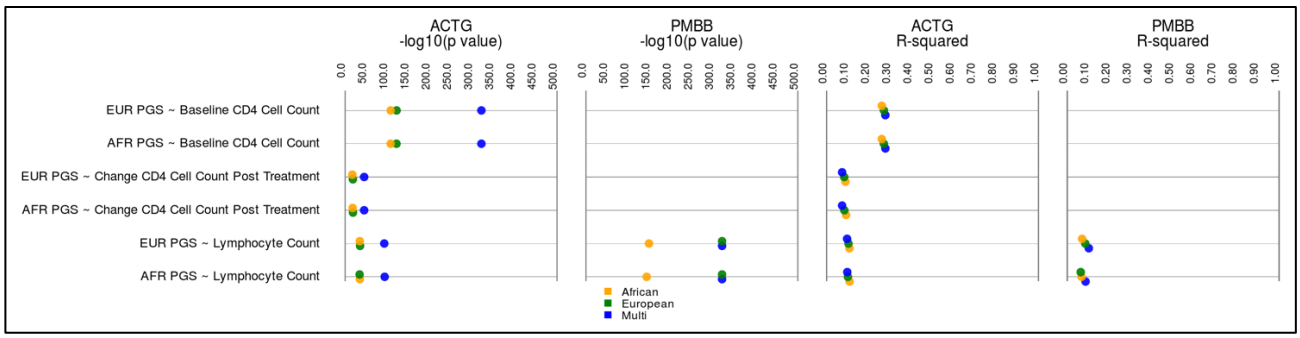

Figure 3: Summary of Regression Results Between PGS_{lymph} and Phenotype While Controlling for Covariates (Age, Sex, PC1-15, log_{10}HIV-1 RNA (viral load), and SQRT of baseline CD4 cell count)

4. Discussion

A lymphocyte count PGS trained in the general population did not effectively predict baseline CD4 cell count or change in CD4 cell count in response to ART, leading to rejection of our hypothesis that poor CD4 cell recovery in response to ART is dependent on each individual's overall genetic predisposition to this outcome. When running regressions without correcting for covariates, R^2 values were low across all ancestry groups and most regressions were not statistically significant (Figure 2, Supplementary Table 1. In contrast, clinical covariates were predictive of these phenotypes. When correcting for covariates, performance of the model improved markedly. Baseline regressions performed modestly ($R^2 = 0.278$) while on-treatment regressions were not very predictive ($R^2 = 0.073$), although all values were statistically significant (Figure 3, Supplementary Table 2). However, because the PGS_{lymph} itself was not highly predictive, the success of this model was mostly due to the contribution of covariates. Additionally, when including covariates in the model, the model including the African PGS_{lymph} better predicted change in CD4 cell count on-treatment in individuals of African ancestry than the model including the European PGS_{lymph} (R^2 was greater by 0.003) (Figure 3, Supplementary Table 2). This is the only case where we see improved performance by an AFR PGS_{lymph} compared to a EUR PGS_{lymph}. Interestingly, when considering effects of individual covariates in this model, the influence of the AFR

PGS_{lymph} is significant (p = 0.044) in individuals of African ancestry with an effect size of -2.062 (Supplementary Table 3). In comparison to other covariates, this effect size is minimal, but suggests that the AFR PGS_{lymph} is playing a role. Furthermore, this shows that our methods improved PGS_{lymph} performance in individuals of African ancestry, which was likely because of a combination of a PGS_{lymph} based on African ancestry summary statistics and utilizing PRScsx for calculation.

In univariate analyses, lymphocyte count PGS did not effectively predict baseline lymphocyte count in ACTG participants. R^2 values were also low and insignificant (Figure 2, Supplementary Table 5). Performance improved when including covariates in this model, as R^2 values rose to ~0.10 and regressions became statistically significant (Figure 3, Supplementary Table 6). Within the covariate models, the influence of the EUR PGS_{lymph} is significant in individuals of European ancestry (p = 0.018) with a minimal effect size of 0.025 (Supplementary Table 7). However, as the effect size is small, though significant, the EUR PGS_{lymph} is not adding much to this model. Still, this significant effect is exhibited as the R^2 value of the EUR PGS_{lymph} covariate model in individuals of European ancestry (0.103) is slightly higher than the R^2 value of the AFR PGS_{lymph} covariate model in individuals of European ancestry (0.101) (Figure 3, Supplementary Table 6). Additionally, in the multivariate model, the influence of the AFR PGS_{lymph} is significant in the multi-ancestry group (p=8.7e-3) with an effect size of 8.3e-3 (Supplementary Table 9). Although this evidently did not have a large impact on the model, the effects of this are still present as the R^2 value of the AFR PGS_{lymph} covariate model in the multi-ancestry group (0.098) is slightly higher than the R^2 value of the EUR PGS_{lymph} covariate model in the multi-ancestry group (0.097) (Figure 3, Supplementary Table 6). Also, it is interesting that the R^2 value did not increase as high as in CD4 cell count regressions, perhaps because viral load was the greatest contributing covariate (viral load had the lowest p-value of all variables in all CD4 cell count regressions), and total lymphocyte counts are not greatly affected by viral load, in contrast to CD4 cell counts[53] (Supplementary Table 3).

Although this model did not perform well in PWH, it performed slightly better when applied to a general medical biobank population. The PGS_{lymph} best predicted lymphocyte count in a general medical biobank population. Regressions were highly statistically significant, likely due to a large sample size (~37,000 individuals). In the univariate model, the African PGS_{lymph} applied to the multi-ancestry group and the European PGS_{lymph} applied to the European population had the highest R^2 values (~0.01) (Figure 2, Supplementary Table 11). It is interesting that these regressions had the highest R^2 values, as these are the only ACTG lymphocyte count regressions that had a significant contribution from PGS_{lymph} in the multivariable model. Seeing these patterns across the general population and PWH shows that the AFR PGS_{lymph} performs best in a multi-ancestry group and the EUR PGS_{lymph} performs best in individuals of European ancestry. When controlling for covariates, performance of the model increased. R^2 values rose to ~0.06-0.10 and p-values dropped even lower (Figure 3, Supplementary Table 12). This mirrors the impact of covariates seen in PWH. The effect size of the EUR PGS_{lymph} was ~0.01 in all ancestry groups (Supplementary Table 13). It is interesting that without covariates, the EUR PGS_{lymph} in individuals of European ancestry was the only regression mirroring this effect size (Figure

2, Supplementary Table 11). The effect size of the AFR PGS_{lymph} was much lower, ~-5e-3 (Supplementary Table 14). This effect size was mirrored in the AFR PGS_{lymph} regressions without covariates in European and African ancestry, as the R^2 values were also low (~$3e^{-3}$ or $8e^{-3}$), but interestingly the R^2 value was higher when the AFR PGS_{lymph} was applied to the multi-ancestry group (~0.01) (Figure 2, Supplementary Table 11).

Although these results showed that PGS_{lymph} itself is not predictive of this treatment response, some results show that in combination with covariates, the impact of PGS_{lymph} can become significant, suggesting a possible synergistic effect between PGS_{lymph} and clinical covariates in the model. In the regressions between AFR PGS_{lymph} and change in CD4 cell count in individuals of African ancestry, the impact of the PGS_{lymph} was insignificant, but when including clinical covariates in the regression, the impact of the PGS_{lymph} became significant (Supplementary Table 3). However, the AFR PGS_{lymph} did not significantly interact with any covariates, eliminating the possibility of a synergistic effect (Supplementary Table 4). Additionally, in the regressions between the AFR PGS_{lymph} and baseline lymphocyte count in PWH of all ancestry groups, as well as in the regressions between the EUR PGS_{lymph} and baseline lymphocyte count in individuals of European ancestry, the same patterns were observed (Supplementary Table 7, Supplementary Table 9). Similarly, the AFR PGS_{lymph} did not significantly interact with any covariates, but the EUR PGS_{lymph} significantly interacted with age (Supplementary Table 8, Supplementary Table 10). Thus, it is possible that in PWH, there are synergistic effects between the EUR PGS_{lymph} and covariates, thus leading the PGS_{lymph} to become significant. These findings highlight the importance of including clinical covariates in PGS analyses, not only because the covariates themselves very predictive of treatment response, but also because they seem to interact with the PGS_{lymph} in some way. Another explanation for this observation is that covariates with strong effects overshadow the effects of PGS_{lymph} when not controlled for. Covariates such as viral load have such high significance and large effect sizes, that the effects of smaller impact variables such as PGS_{lymph} are not seen unless these covariates were controlled for. Thus, it is important to consider clinical covariates when implementing PGS in a clinical setting.

This study had several limitations. First, the sample size of the African ancestry summary statistics that were used to generate the African PGS_{lymph} were small (~13,000 individuals), which is due to the lack of availability of lymphocyte count summary statistics for individuals of African ancestry. To improve these results, more lymphocyte count GWAS data are needed in future studies, as it is possible that the AFR PGS_{lymph} could have performed better with a larger base sample size. Additionally, the ACTG sample size was modest (~4600 individuals) which was subset to even smaller groups when stratified by ancestry. It is possible that associations with PGS_{lymph} may have become statistically significant with a larger sample size. Subsequent work in this area could investigate whether this model is predictive of other drug response traits, specifically other ART treatment responses.

Polygenic scores have the potential to leverage large, publicly available datasets to find novel genetic discoveries in pharmacogenomic cohorts. This study utilized a novel method to predict CD4 cell recovery in response to ART and illustrated the importance of including clinical covariates in a

PGS model. As more associations or lack thereof are found, we continue to narrow down the biological underpinnings of responses to ART including suboptimal CD4 cell recovery.

5. Acknowledgements

The authors are grateful to the many persons living with HIV who volunteered for ACTG protocols ACTG384, A5095, A5142, A5202 and A5257. In addition, they acknowledge the contributions of study teams and site staff for these protocols. We thank Paul J. McLaren, PhD (Public Health Agency of Canada, Winnipeg, Canada) for prior involvement and collaborations that used these genome-wide genotype data. Study drugs were provided by Bristol-Myers Squibb, Inc., Gilead Sciences, Inc., GlaxoSmithKline, Inc.. The clinical trials were A5095 (NCT00013520), A5142 (NCT00050895), and A5202 (NCT00118898).

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number UM1 AI068634, UM1 AI068636 and UM1 AI106701. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research reported in this publication was also supported in part by grants funded by the National Center for Research Resources and the National Center for Advancing Translational Sciences.

Grant support included TR000124 (to E.S.D.); AI110527, AI077505, TR000445, and AI069439 (to D.W.H.); AI077505 (to K.M.C, S.D., and M.D.R). This work was supported by the Tennessee Center for AIDS Research (P30) AI110527.

Clinical research sites that participated in ACTG protocols A5095, A5142 and/or A5202, and collected DNA under protocol A5128 were supported by the following grants from the National Institutes of Health (NIH): A1069412, A1069423, A1069424, A1069503, AI025859, AI025868, AI027658, AI027661, AI027666, AI027675, AI032782, AI034853, AI038858, AI045008, AI046370, AI046376, AI050409, AI050410, AI050410, AI058740, AI060354, AI068636, AI069412, AI069415, AI069418, AI069419, AI069423, AI069424, AI069428, AI069432, AI069432, AI069434, AI069439, AI069447, AI069450, AI069452, AI069465, AI069467, AI069470, AI069471, AI069472,AI069474, AI069477, AI069481, AI069484, AI069494, AI069495, AI069496, AI069501, AI069501, AI069502, AI069503, AI069511, AI069513, AI069532, AI069534, AI069556, AI072626, AI073961, RR000046, RR000425, RR023561, RR024156, RR024160, RR024996, RR025008, RR025747, RR025777, RR025780, TR000004, TR000058, TR000124, TR000170, TR000439, TR000445, TR000457, TR001079, TR001082, TR001111, and TR024160.

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the

National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878.

6. Supplementary Tables

All supplemental data can be found at: https://ritchielab.org/files/PSB_supplemental_data/PSB_LymphocytePGSHIV_Cardone_2023_Supplementary.pdf

References

1. HIV. https://www.who.int/data/gho/data/themes/hiv-aids#cms.
2. Basic Statistics | HIV Basics | HIV/AIDS | CDC. https://www.cdc.gov/hiv/basics/statistics.html (2022).
3. Vidya Vijayan, K. K., Karthigeyan, K. P., Tripathi, S. P. & Hanna, L. E. Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections. *Front. Immunol.* **8**, 580 (2017).
4. Menéndez-Arias, L. & Delgado, R. Update and latest advances in antiretroviral therapy. *Trends Pharmacol. Sci.* **43**, 16–29 (2022).
5. Kelley, C. F. *et al.* Incomplete Peripheral CD4 $^{+}$ Cell Count Restoration in HIV-Infected Patients Receiving Long-Term Antiretroviral Treatment. *Clin. Infect. Dis.* **48**, 787–794 (2009).
6. Lok, J. J. *et al.* Long-term increase in CD4+ T-cell counts during combination antiretroviral therapy for HIV-1 infection. *AIDS* **24**, 1867–1876 (2010).
7. Moore, R. D. & Keruly, J. C. CD4+ Cell Count 6 Years after Commencement of Highly Active Antiretroviral Therapy in Persons with Sustained Virologic Suppression. *Clin. Infect. Dis.* **44**, 441–446 (2007).
8. Baker, J. V. *et al.* CD4+ count and risk of non-AIDS diseases following initial treatment for HIV infection. *AIDS* **22**, 841–848 (2008).

9. Yang, X. *et al.* Incomplete immune reconstitution in HIV/AIDS patients on antiretroviral therapy: Challenges of immunological non-responders. *J. Leukoc. Biol.* **107**, 597–612 (2020).
10. Geng, E. H. *et al.* CD4+ T cell recovery during suppression of HIV replication: an international comparison of the immunological efficacy of antiretroviral therapy in North America, Asia and Africa. *Int. J. Epidemiol.* **44**, 251–263 (2015).
11. Haas, D. W. & Tarr, P. E. Perspectives on pharmacogenomics of antiretroviral medications and HIV-associated comorbidities: *Curr. Opin. HIV AIDS* **10**, 116–122 (2015).
12. Haas, D. W. *et al.* Immunogenetics of CD4 Lymphocyte Count Recovery during Antiretroviral Therapy: An AIDS Clinical Trials Group Study. *J. Infect. Dis.* **194**, 1098–1107 (2006).
13. Hartling, H. J. *et al.* Polymorphism in interleukin-7 receptor α gene is associated with faster CD4+ T-cell recovery after initiation of combination antiretroviral therapy. *AIDS* **28**, 1739–1748 (2014).
14. Rajasuriar, R. *et al.* The role of SNPs in the α-chain of the IL-7R gene in CD4+ T-cell recovery in HIV-infected African patients receiving suppressive cART. *Genes Immun.* **13**, 83–93 (2012).
15. Ahuja, S. K. *et al.* CCL3L1-CCR5 genotype influences durability of immune recovery during antiretroviral therapy of HIV-1–infected individuals. *Nat. Med.* **14**, 413–420 (2008).
16. Rauch, A. *et al. HLA-Bw4* Homozygosity Is Associated with an Impaired CD4 T Cell Recovery after Initiation of Antiretroviral Therapy. *Clin. Infect. Dis.* **46**, 1921–1925 (2008).
17. Grady, B. J. *et al.* Mitochondrial Genomics and CD4 T-Cell Count Recovery After Antiretroviral Therapy Initiation in AIDS Clinical Trials Group Study 384. *JAIDS J. Acquir. Immune Defic. Syndr.* **58**, 363–370 (2011).

18. Guzmán-Fulgencio, M. *et al.* European mitochondrial haplogroups are associated with CD4+ T cell recovery in HIV-infected patients on combination antiretroviral therapy. *J. Antimicrob. Chemother.* **68**, 2349–2357 (2013).
19. Greenblatt, R. *et al.* Genetic and clinical predictors of CD4 lymphocyte recovery during suppressive antiretroviral therapy: Whole exome sequencing and antiretroviral therapy response phenotypes. *PLOS ONE* **14**, e0219201 (2019).
20. Polygenic Risk Scores. *Genome.gov* https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores (2022).
21. National Institute of General Medical Sciences. *National Institute of General Medical Sciences (NIGMS)* https://nigms.nih.gov/.
22. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
23. Calafato, M. S. *et al.* Use of schizophrenia and bipolar disorder polygenic scores to identify psychotic disorders. *Br. J. Psychiatry J. Ment. Sci.* **213**, 535–541 (2018).
24. Jonas, K. G. *et al.* Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. *Transl. Psychiatry* **9**, 300 (2019).
25. Zheutlin, A. B. *et al.* Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems. *Am. J. Psychiatry* **176**, 846–855 (2019).
26. Alnæs, D. *et al.* Brain Heterogeneity in Schizophrenia and Its Association With Polygenic Risk. *JAMA Psychiatry* **76**, 739–748 (2019).

27. Mas-Bermejo, P. *et al.* Schizophrenia polygenic risk score in psychosis proneness. *Eur. Arch. Psychiatry Clin. Neurosci.* (2023) doi:10.1007/s00406-023-01633-7.
28. Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V. & Zammit, S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: A systematic review. *J. Affect. Disord.* **234**, 148–155 (2018).
29. Hasseris, S. *et al.* Polygenic Risk and Episode Polarity Among Individuals With Bipolar Disorder. *Am. J. Psychiatry* **180**, 200–208 (2023).
30. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
31. Li, H. *et al.* Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. *Genet. Med.* **19**, 30–35 (2017).
32. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
33. Roberts, E., Howell, S. & Evans, D. G. Polygenic risk scores and breast cancer risk prediction. *Breast Edinb. Scotl.* **67**, 71–77 (2023).
34. O'Sullivan, J. W., Ashley, E. A. & Elliott, P. M. Polygenic risk scores for the prediction of cardiometabolic disease. *Eur. Heart J.* **44**, 89–99 (2023).
35. McCarthy, M. I. & Mahajan, A. The value of genetic risk scores in precision medicine for diabetes. *Expert Rev. Precis. Med. Drug Dev.* **3**, 279–281 (2018).

36. Rao, A. S. & Knowles, J. W. Polygenic risk scores in coronary artery disease. *Curr. Opin. Cardiol.* **34**, 435–440 (2019).

37. Pulit, S. L. *et al.* Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. *Neurol. Genet.* **4**, e293 (2018).

38. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

39. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).

40. Keat, K. *et al.* Leveraging Multi-Ancestry Polygenic Risk Scores for Body Mass Index to Predict Antiretroviral Therapy-Induced Weight Gain. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **28**, 233–244 (2023).

41. Ritchie, M. D. *et al.* Genome- and Phenome-Wide Analyses of Cardiac Conduction Identifies Markers of Arrhythmia Risk. *Circulation* **127**, 1377–1385 (2013).

42. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

43. Haas, D. W. *et al.* A Multi-Investigator/Institutional DNA Bank for AIDS-Related Human Genetic Studies: AACTG Protocol A5128. *HIV Clin. Trials* **4**, 287–300 (2003).

44. Daar, E. S. Atazanavir Plus Ritonavir or Efavirenz as Part of a 3-Drug Regimen for Initial Treatment of HIV-1: A Randomized Trial. *Ann. Intern. Med.* **154**, 445 (2011).

45. Gulick, R. M. Three- vs Four-Drug Antiretroviral Regimens for the Initial Treatment of HIV-1 InfectionA Randomized Controlled Trial. *JAMA* **296**, 769 (2006).

46. Gulick, R. M. *et al.* Triple-Nucleoside Regimens versus Efavirenz-Containing Regimens for the Initial Treatment of HIV-1 Infection. *N. Engl. J. Med.* **350**, 1850–1861 (2004).

47. Riddler, S. A. *et al.* Class-Sparing Regimens for Initial Treatment of HIV-1 Infection. *N. Engl. J. Med.* **358**, 2095–2106 (2008).

48. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).

49. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

50. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

51. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

52. Pendergrass, S. A., Dudek, S. M., Crawford, D. C. & Ritchie, M. D. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Min.* **3**, 10 (2010).

53. 2.14 How CD4 and viral load are related | Training manual | HIV i-Base. https://i-base.info/ttfa/section-2/14-how-cd4-and-viral-load-are-related/.