GENE EXPRESSION DIFFERENCES BY RACE AND GENETIC ANCESTRY IN B-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

Freddy Barragan[1,2], Lauren J. Mills[2,3], Andrew R. Raduski[2], Erin L. Marcotte[2,3], Logan G. Spector[2,3], Kelsey E. Grinde[1], Lindsay A Williams[2,3]

[1] Department of Mathematics, Statistics, and Computer Science, Macalester College; [2] Division of Epidemiology and Clinical Research, Department of Pediatrics, University of Minnesota; [3] Masonic Cancer Center, University of Minnesota

INTRODUCTION

There are substantial disparities in B-cell acute lymphoblastic leukemia (B-ALL) survival and clinical presentation by race/ethnicity.

African American (AA) children have lower incidence rates of B-ALL, but present with more severe mutations and have a higher risk of death than European American (EA) and Latinx (LAT) children.

Previous studies have measured the effect of social determinants of health by race/ethnicity, but they do not wholly explain the observed disparities in outcome.

PURPOSE

These analyses aim to understand how gene expression and survival in children with B-ALL differ by genetic ancestry.

METHODOLOGY

Comprehensive gene expression and genotype data on children and young adults ages 0-23 diagnosed with B-ALL (N=286; N=306) were accessed via the National Cancer Institute's TARGET database and St. Jude's PeCan cloud.

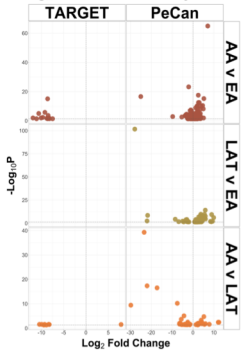
Using RFMix, we inferred genetic ancestry in our TARGET samples at the global and local levels using SNP data. Then, using K-Means clustering we grouped samples into 4 ancestry clusters of predominantly African (AFR), Amerindian (AMR), East Asian (EAS), and European (EUR) ancestry. Finally, we used KNN classification to infer race/ethnicity in cases with missing data.

Using `limma` and `DESeq2` with Benjamini-Hochberg (BH) corrections for multiple testing, we conducted differential expression analyses to identify genes whose expression differs by race/ethnicity and genes whose expression is associated with increasing proportions of AFR, AMR, and EUR ancestry, respectively (BH-p < 0.05).

Kaplan Meier curves and Log-Rank p-values were used to identify significant differences in survival between ancestry clusters. Cox proportional hazard models were then used to measure the association between risk of death, gene expression in DE genes, and increasing proportions of AFR, AMR, and EUR ancestry.

GENE EXPRESSION VARIES BETWEEN RACIAL GROUPS

Log_2FC of DE Genes by Race



`limma` uses linear models and Empirical Bayes to relate predictors to the expression of a set of genes.

After adjusting for multiple testing, we identified 69, 0, and 290 significant (BH-p < 0.05) DE genes when comparing AA with EA, LAT with EA, and LAT with AA, while controlling for subtype and phase of data collection in TARGET.

After adjusting for multiple testing, we identified 374, 198, and 50 significant (BH-p < 0.05) DE genes when comparing AA with EA, LAT with EA, and LAT with AA while controlling for subtype in PeCan.

TARGET

Comparison	Reference	LogFC < 0	LogFC > 0
AA v. EA	AA	13	0
LAT v. EA	EA	0	0
LAT v. AA	AA	7	1

PeCan

Comparison	Reference	LogFC < 0	LogFC > 0
AA v. EA	EA	112	262
LAT v. EA	EA	14	184
LAT v. AA	AA	30	20

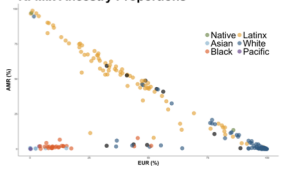
HOWEVER, REPORTED RACE IS CRUDE & SOCIALLY CONSTRUCTED

SNPs are single base pair variants. Some of these SNPs are more common in some geographic communities than others.

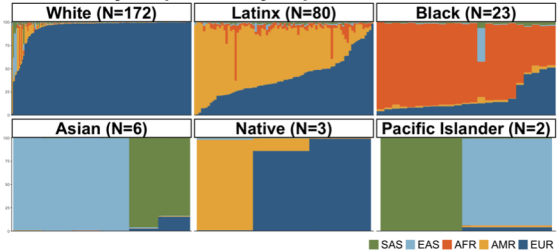
RFMix classifies a person's local ancestry depending on the posterior probability of an ancestry at a particular SNP in the genome.

Ancestral variations within racial groups can be incredibly dramatic. Therefore, race is a poor proxy for genetic ancestry.

RFMix Ancestry Proportions



Ancestry Proportions by Reported Race in TARGET



EXPRESSION OF GENES INVOLVED IN CANCER & TRANSCRIPTION REGULATION MAY BE ASSOCIATED WITH INCREASED PROPORTIONS OF AFR, AMR, & EUR ANCESTRY

After adjusting for multiple testing, we identified 10, 2, and 7 genes whose expression is significantly associated (BH-p < 0.05) with increasing proportions of AFR, AMR, and EUR ancestry, when controlling for subtype and phase of TARGET collection.

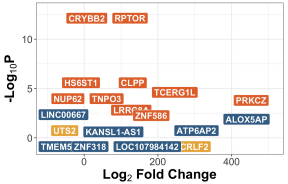
Newly identified DE genes for AMR ancestry indicate ancestry proportions may improve sensitivity of DE analysis in admixed communities.

17 of these DE genes have been identified as tumor suppressors, prognostic biomarkers, or transcription regulators in other studies for adult cancers.

TARGET

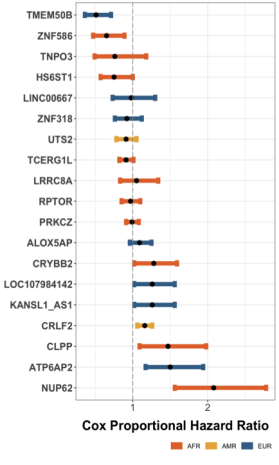
Ancestry	Number of DE Genes
AFR	10
AMR	2
EUR	7

Log_2FC of DE Genes for a 10% Ancestry Difference



9 OF THESE DE GENES MAY BE PROGNOSTIC MARKER CANDIDATES

Risk of Death by Log_2 Mean Expression



We found that for 9 of 19 DE genes, there was a statistically significant association between risk of death and expression, when controlling for relevant covariates including race, age, sex, subtype, WBC, CNS & testicular involvement.

Of these 9 genes, 4 were associated with increased proportions of AFR ancestry, 3 with EUR ancestry, and 1 with AMR ancestry.

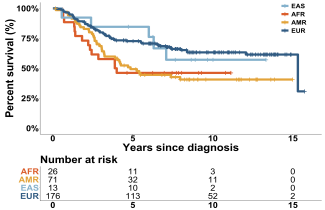
2 of these genes were associated with decreased risk of death, while 7 DE genes were associated with an increased risk of death (HRs 1.16-2.08).

Gene	HR (95% CI)	P-Value
TMEM50B	0.51 (0.36, 0.71)	<0.001
CRLF2	1.16 (1.08, 1.24)	<0.001
NUP62	2.08 (1.56, 2.78)	<0.001
ATP6AP2	1.50 (1.17, 1.94)	0.002
ZNF586	0.65 (0.47, 0.89)	0.006
CLPP	1.47 (1.09, 1.98)	0.012
CRYBB2	1.28 (1.02, 1.59)	0.031
KANSL1-AS1	1.26 (1.02, 1.56)	0.034
LOC107984142	1.26 (1.02, 1.56)	0.034

INCREASING PROPORTIONS OF AFR & AMR ANCESTRY ARE ASSOCIATED WITH HIGHER RISKS OF DEATH

Comparison	Ancestry	Hazard Ratio (95% CI)	P-Value
AA v. EUR	AFR	1.15 (1.05, 1.25)	0.002
EUR v. LAT	AMR	1.12 (1.05, 1.20)	<0.001
AA v. LAT	AFR	1.07 (0.96, 1.19)	0.21

B-ALL Survival (p = 0.0037)



CONCLUSIONS

Using RFMix to infer genetic ancestry, we've shown that ancestry varies highly within the socially defined categories of race.

We have identified 19 genes whose expression is significantly associated with increases in genetic ancestry. Promisingly, 17 of these 19 genes have also been implicated as genes of interest in other cancer studies, while 7 of these also show associations with increased risk of death.

Further, we've found that 10% increases in AFR and AMR ancestry are associated with a 1.15x and 1.12x greater risks of death, when controlling for relevant covariates.

However, future work is needed to understand the causal mechanisms of these genes and the true effect of ancestry, if any.

IMPACT

Genetic ancestry may help explain the observed differences in B-ALL incidence, survival, and clinical characteristics. Our findings highlight the need for detailed genetic, demographic, and clinical data to help epidemiologists better understand how diseases affect communities differently. In addition, our work highlights the necessity of developing statistical methods suited for highly admixed communities.

FUNDING

Research reported in this publication was supported by the Macalester Mann-Hill Fellowship (FB). The content is solely the responsibility of the authors and does not necessarily represent the official views of Macalester College or the University of Minnesota.